Mirusviruses provide a missing link in the evolution of giant viruses Plankton-infecting relatives of herpesviruses clarify the evolutionary trajectory of giant viruses

Morgan Gaïa, Lingjie Meng, Eric Pelletier, Patrick Forterre, Chiara Vanni, Antonio Fernandez-Guerra, Olivier Jaillon, Patrick Wincker, Hiroyuki Ogata, Mart Krupovic, et al.

Plankton-infecting relatives of herpesviruses clarify the evolutionary trajectory of giant viruses

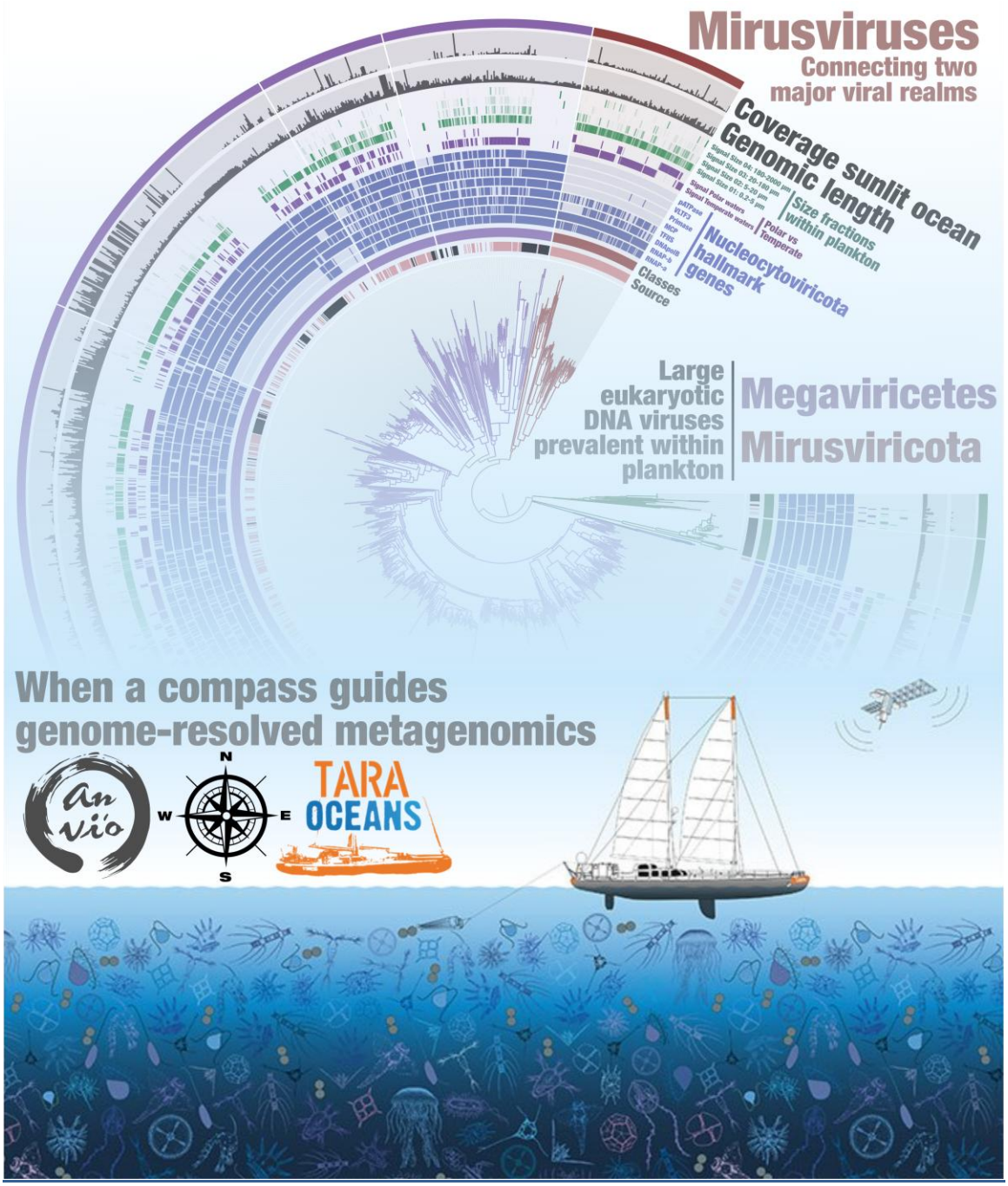
Morgan Gaïa[1,2], Lingjie Meng[3], Eric Pelletier[1,2], Patrick Forterre[4,5], Chiara Vanni[6], Antonio Fernandez-Guerra[7], Olivier Jaillon[1,2], Patrick Wincker[1,2], Hiroyuki Ogata[3], Mart Krupovic[8], and Tom O. Delmont[1,2]

[1] Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, 91057 Evry, France.
[2] Research Federation for the study of Global Ocean systems ecology and evolution, FR2022/Tara GOsee, Paris, France.
[3] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, 611-0011, Japan
[4] Institut de Biologie Intégrative de la Cellule (I2BC), CNRS, Université Paris-Saclay, 91198 Gif sur Yvette, France
[5] Institut Pasteur, Département de Microbiologie, 25 rue du Docteur Roux, 75017, Paris, France
[6] MARUM center for marine environmental sciences, University of Bremen, Germany
[7] Lundbeck Foundation GeoGenetics Centre, GLOBE Institute, University of Copenhagen, Copenhagen, Denmark
[8] Institut Pasteur, Université Paris Cité, CNRS UMR6047, Archaeal Virology Unit, Paris, France

Abstract: DNA viruses have a major influence on the ecology and evolution of cellular organisms, but their overall diversity and evolutionary trajectories remain elusive. Here, we performed a phylogeny-guided genome-resolved metagenomic survey of the sunlit oceans which exposed a major, previously undescribed clade of large eukaryotic DNA viruses. Viruses in this clade encode a virion morphogenesis module characteristic of the realm *Duplodnaviria*, which until now included *Caudoviricetes* (tailed phages) and *Herpesvirales* (animal-infecting viruses). The conservation of the morphogenetic module firmly places the new clade within *Duplodnaviria* as a phylum-level group, which we name '*Mirusviricota*'. Remarkably, a larger fraction of 'Mirusviricota' genes, including hallmark informational markers, have homologs in giant eukaryotic DNA viruses from the realm *Varidnaviria*. We suggest that '*Mirusviricota*' played a key role in the evolution of eukaryotic DNA viruses and provides missing links in the evolution of both animal herpesviruses from tailed prokaryotic viruses and giant varidnaviruses from smaller relatives. Our results are supported by more than 100 manually curated environmental genomes, including a near-complete contiguous genome of 432 Kbp. The prevalence of '*Mirusviricota*' viruses within plankton and their classification into several phylogenetically diverse clades emphasizes their potential impact in marine ecosystems. Finally, genomic and environmental data indicate that 'Mirusviricota' harbor a rich, functionally diverse gene repertoire, including multiple heliorhodopsins, which modulates virus-host interactions in the sunlit oceans.

Key words: *Mirusviricota*, mirusviruses, *Tara* Oceans, plankton, metagenomics, DNA viruses, Giant viruses, eukaryotes, ecology, evolution, major capsid protein, phylogenomics

Cover: The environmental DNA sequencing and bioinformatics legacies of ***Tara* Oceans** and **anvi'o** exposed a putative new phylum dubbed "***Mirusviricota***" by means of **phylogeny-guided genome-Resolved metagenomics**. Unexpectedly, the mirusviruses are plankton-infecting relatives of herpesviruses that clarify the evolutionary trajectory of prominent large and giant viruses from another realm.



Modified from the original planktonic illustration of Noan Le Bescot

Introduction:

Most double-stranded DNA viruses are classified into two major realms: *Duplodnaviria* and *Varidnaviria*. *Duplodnaviria* comprises tailed bacteriophages and related archaeal viruses of the class *Caudoviricetes* and eukaryotic viruses of the order *Herpesvirales*. *Varidnaviria* includes large and giant eukaryotic DNA viruses from the phylum *Nucleocytoviricota* as well as smaller eukaryotic, bacterial and archaeal viruses with tailless icosahedral capsid[1]. The two realms were established based on the non-homologous sets of virion morphogenesis genes (virion module), including the structurally unrelated major capsid proteins (MCP) with the 'double jelly-roll' and HK97 folds in *Varidnaviria* and *Duplodnaviria*, respectively[1]. Both realms are represented across all domains of life, with the respective ancestors thought to date back to the last universal cellular ancestor[2]. Within *Duplodnaviria*, bacterial and archaeal members of the *Caudoviricetes* display a continuous range of genome sizes, from ~10 kb to >700 kb, whereas herpesviruses, restricted to animal hosts, are more uniform, with genomes in the range of 100-300 kb. Herpesviruses likely evolved from bacteriophages, but the lack of related viruses outside of the animal kingdom raises questions regarding their exact evolutionary trajectory[3]. Members of the *Varidnaviria* also display a wide range of genome sizes, from ~10 kb to >2 Mb, but there is a discontinuity in the complexity between large and giant viruses of the *Nucleocytoviricota* and the rest of varidnaviruses with genomes <50 kb. It has been suggested that *Nucleocytoviricota* have evolved from a smaller varidnavirus ancestor[4–6], but the complexification entailing acquisition of multiple informational genes (informational module), including those for transcription, such as the multisubunit DNA-dependent RNA polymerase, remains to be fully understood.

The *Caudoviricetes* and *Nucleocytoviricota* viruses are prevalent in the sunlit ocean where they play a critical role in regulating the community composition and blooming activity of plankton[7–9]. Metagenomic assemblies and genome-resolved metagenomics were used to target these viruses within plankton, complementing culture-dependent virus isolation efforts and leading to our improved appreciation of their genomic diversity[7,8,10–12]. These environmental surveys emphasized the prevalence and functional complexity of DNA viruses in the sunlit ocean and supported the evolutionary demarcation between the virion and informational modules of *Varidnaviria* and *Duplodnaviria* realms. Previous efforts to uncover the diversity of large eukaryotic DNA viruses in marine metagenomes largely relied on identification of a nearly full set of *Nucleocytoviricota* hallmark genes[7,8]. Although this approach ensures high genomic completion and quality of the identified viral genomes, it is by default unsuitable for the discovery of novel marine DNA viruses outside of the *Nucleocytoviricota*.

Here, we performed a phylogeny-guided genome-resolved metagenomic survey of planktonic DNA viruses containing a DNA-dependent RNA polymerase gene. The survey covers nearly 300 billion metagenomic reads from surface ocean samples of

the *Tara* Oceans expeditions[13–15]. We characterized and manually curated >500 *Nucleocytoviricota* population genomes, expanding the known diversity of the class *Megaviricetes* and recovering a putative new class. But most notably, our survey led to the discovery of a putative new phylum of large DNA viruses infecting marine eukaryotes, dubbed '*Mirusviricota*' (>100 population genomes up to >400 kb in size). These viruses are diverse and prevalent in the sunlit ocean where they likely play a central role in the ecology of abundant planktonic eukaryotes. Members of '*Mirusviricota*' contain a virion module characteristic of the realm *Duplodnaviria* (including a distinct HK97-fold MCP, a possible evolutionary intermediate between the MCPs of *Caudoviricetes* and *Herpesvirales* based on protein structure analyses). Yet, these viruses also contain a large fraction of genes more closely related to those of *Nucleocytoviricota*, including key informational hallmark markers. A near-complete mirusvirus genome supports these results. The atypical chimeric attributes of '*Mirusviricota*' provide important insights into the evolution of DNA viruses, clarifying the evolution of animal herpesviruses from tailed bacteriophages and giant varidnaviruses from smaller relatives.

Results:

Environmental genomic recovery of a new clade of large DNA viruses

DNA–dependent RNA polymerase subunits A (RNApolA) and B (RNApolB) are evolutionarily informative gene markers that are universally conserved in the three cellular domains and occur in most *Nucleocytoviricota* viruses[5,16]. Here, we performed a comprehensive search for RNApolB genes from the euphotic zone of polar, temperate, and tropical oceans using 798 metagenomes derived from the *Tara* Oceans expeditions[13]. They correspond to surface waters and deep chlorophyll maximum (DCM) layers from 143 stations covering the Pacific, Atlantic, Indian, Arctic, and Southern Oceans, as well as the Mediterranean and Red Seas, encompassing eight plankton size fractions ranging from 0.8 µm to 2000 µm (Table S1). These 280 billion reads were used as inputs for 11 metagenomic co-assemblies (~12 million contigs longer than 2,500 nucleotides) using geographically related samples, and cellular metagenome-assembled genomes (MAGs) were subsequently characterized[14,15]. We recovered RNApolB genes from these co-assemblies using a broad-spectrum profile hidden Markov model (HMM) and subsequently built a database of more than 2,500 non-redundant protein sequences (similarity <90%; Table S2). Phylogenetic signal for these sequences recapitulated previously observed trends, such as monophyletic groups for Bacteria, Archaea and the three eukaryotic nuclear RNApolB clades[5], as well as the considerable diversity of *Nucleocytoviricota* already revealed by means of gene markers[17] and genome-resolved metagenomics[7,8] (Figure S1). The phylogenetic analysis revealed previously undescribed deep-branching lineages with no representatives among known viruses, which we dubbed Mirus (Latin for surprising, strange) once sequential increases in phylogenetic resolution supported their monophyly. With a positioning clearly disconnected from the three domains of life, we hypothesized

that the Mirus RNApolB clades may correspond to previously unknown lineages of double-stranded DNA viruses.

We performed a two-step phylogeny-guided genome-resolved metagenomic survey of the ~12 million contigs. First, we used the RNApolB genes corresponding to *Nucleocytoviricota* and Mirus as a compass to identify metagenomic assembly subsets of interest. Then, the sequence composition and differential coverage of contigs across metagenomes were computed independently for each subset, as previously done for the cellular MAGs[15,18], in order to manually delineate the genomic context of RNApolB genes (see Table S3 and supplemental information for more details). We successfully characterized and curated 587 non-redundant *Nucleocytoviricota* MAGs (average nucleotide identify <98%) up to 1.45 Mbp in length (average of ~270 Kbp), as well as 111 non-redundant Mirus MAGs up to 438 kbp in length (average of ~200 Kbp). The Mirus MAGs were identified in all the oceanic regions (Table S5), emphasizing their prevalence in the sunlit ocean. This database was supplemented with MAGs from previous metagenomic surveys[7,8]. The final non-redundant genomic resource contained 1,482 marine *Nucleocytoviricota* MAGs, 224 reference *Nucleocytoviricota* genomes from culture and cell sorting, and the 111 Mirus MAGs (Table S4). This genomic database, enriched in large and giant marine eukaryotic DNA viruses, has a total volume of 580 Mbp and contains ~0.6 million genes. It was used to map 300 billion metagenomic reads from nine *Tara* Oceans plankton size fractions (0.2 µm to 2000 µm)[19]. In subsequent sections, we used this database (thereafter called Global Ocean Eukaryotic Viral [GOEV] database) to contextualize the evolutionary history, functional lifestyle and environmental distribution of newly identified *Nucleocytoviricota* and Mirus MAGs.

Evolutionary links between mirusviruses and the *Nucleocytoviricota* phylum

The newly assembled *Nucleocytoviricota* MAGs contain most of the hallmark genes corresponding to the virion and informational modules characteristic of this viral phylum[4,5] (Table S4) and further expand the known diversity of the class *Megaviricetes*. We also identified one putative new class-level group, '*Proculviricetes*' (several MAGs exclusively detected in the Arctic and Southern Oceans) (Figure 1). The Mirus MAGs also contained key genes evolutionarily related to the *Nucleocytoviricota* informational module, including RNApolA and RNApolB, family B DNA polymerase (DNApolB), and the transcription factor II-S (TFIIS; Figures 1 and S2). For instance, the RNApolA and DNApolB of most Mirus MAGs branched as a sister clade to the order '*Pandoravirales*'. Notably, however, the Mirus MAGs were devoid of identifiable homologs of the *Nucleocytoviricota* virion module, including the hallmark double jelly-roll MCP (Figure 1). Phylogenomic inferences derived from our manually curated database of the four informational gene markers (see Figure 1 and supplemental information) indicate that Mirus represents a new viral clade related to *Nucleocytoviricota* and organized into seven distinct lineages, M1 to M7 (from the most to least populated), with the latter being represented by a single MAG (Figure 2, Table S4).

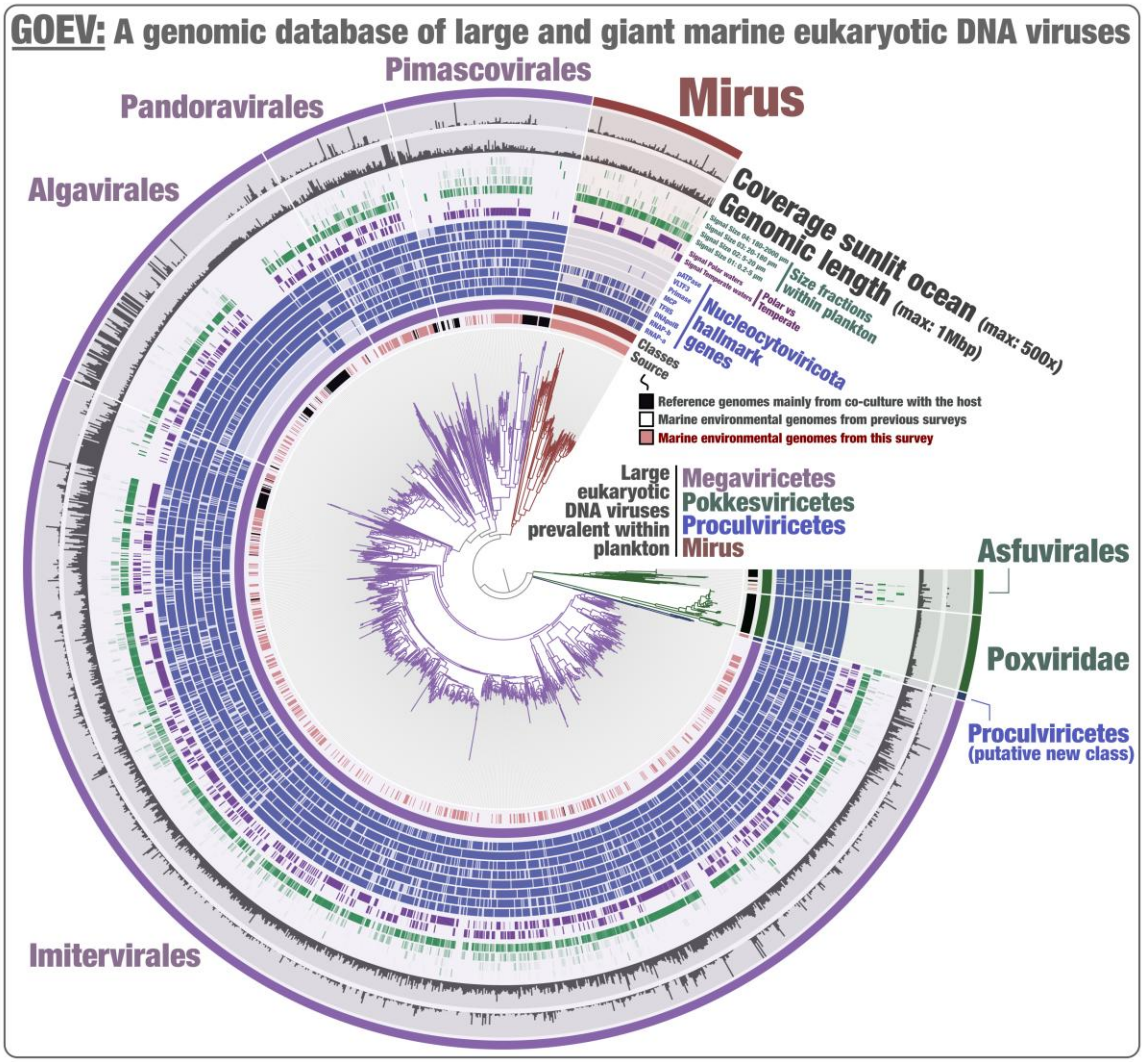


Figure 1: Evolutionary relationship between *Nucleocytoviricota* and mirusviruses. The figure displays a maximum-likelihood phylogenetic tree built from the GOEV database (1,722 genomes) based on a concatenation of manually curated RNApolA, RNApolB, DNApolB and TFIIS genes (3,715 amino acid positions) using the PMSF mixture model and rooted between the *Pokkesviricetes* and the rest. The tree was decorated with layers of complementary information and visualized with anvi'o.

The Mirus lineages were detected across multiple cellular size fractions. Their distribution suggests they occur as rather large extracellular particles (abundant in the 0.2-3 µm size fraction) and in relation to planktonic eukaryotes (spikes of detection in the larger cellular size fractions), in line with what can be observed for the *Nucleocytoviricota* viruses (Figures 1 and 2). For instance, MAGs from lineages M6 and M4 were enriched in the 3-20 µm and 180-2000 µm size fractions, respectively, in addition to the viral particle size fraction (0.2-3 µm). Collectively, these results suggest that the Mirus MAGs represent viruses of diverse marine eukaryotes with a broad cellular size range (Figure 2A). In addition, mirusviruses were relatively abundant, with a cumulative coverage of 6,479x (i.e., the 111 Mirus MAGs were cumulatively sequenced more than 6,000 times among the considered metagenomes), suggesting that the corresponding viruses infect relatively abundant

marine eukaryotes in both temperate and polar waters. Phylogeny-guided host predictions[18] based on co-occurrence patterns using Mirus and eukaryotic MAGs[13] revealed two orders of phytoplankton, Chloropicales (green algae) and Phaeocystales (haptophytes), which showed significant clade-level associations with Mirus subclade M2 (Figure S3 and Table S5). Furthermore, at the individual MAG level, some mirusviruses in other subclades showed a high pairwise association with eukaryotes affiliated to other lineages, such as diatoms and the heterotrophic MAST-4 (Table S5). Together with the large phylogenetic diversity of Mirus MAGs, these results suggest a broad host range of mirusviruses.

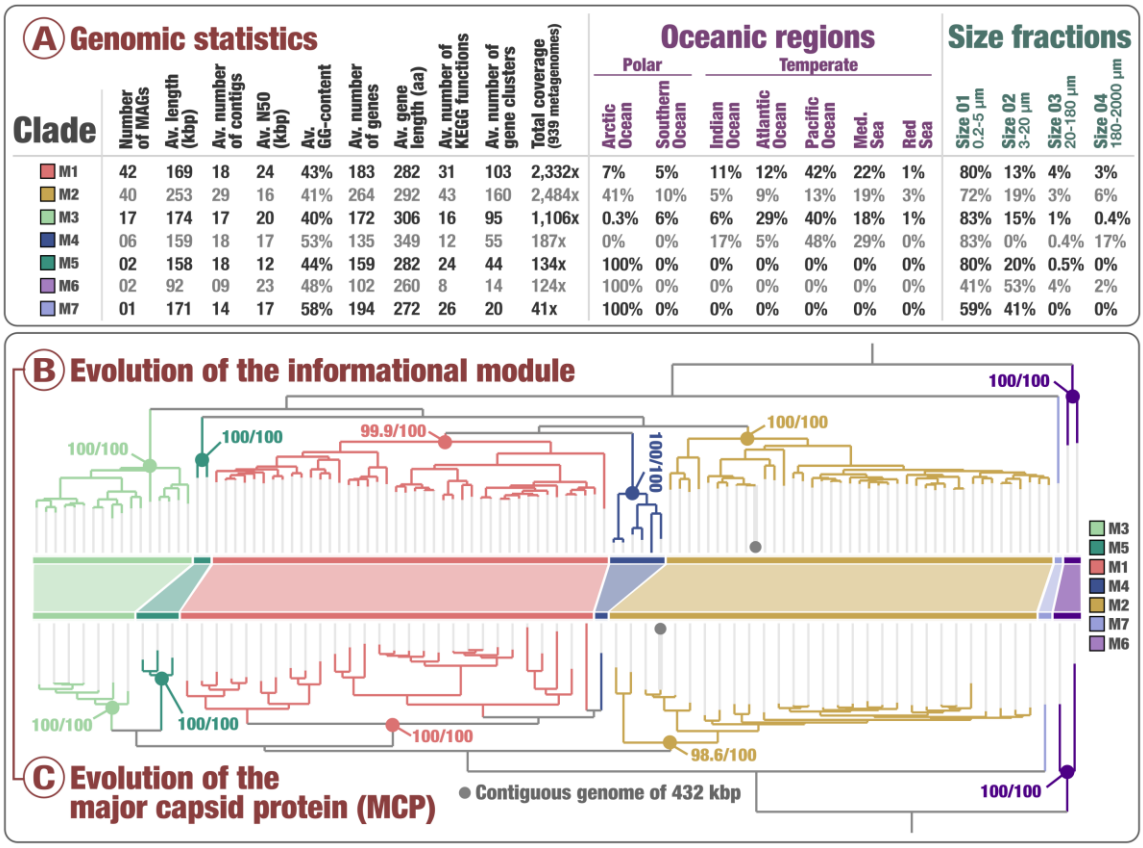


Figure 2: Genomic statistics and evolution of mirusviruses. Panel A displays genomic and environmental statistics for the seven 'Mirusviricota' clades. Panel B displays a maximum-likelihood phylogenetic tree built from the 'Mirusviricota' MAGs based on a concatenation of four hallmark informational genes (RNApolA, RNApolB, DNApolB, TFIIS; 3,715 amino acid positions) using the LG+F+R7 model. Panel C displays a maximum-likelihood phylogenetic tree built from the 'Mirusviricota' MAGs based on the Major Capsid Protein (701 amino acid positions) using the LG+R6 model. Both trees were rooted between clade M6 and other clades. Values at nodes represent branch supports (out of 100) calculated by the SH-like aLTR (1,000 replicates; left score) and ultrafast bootstrap approximations (1,000 replicates; right score).

Mirusviruses are functionally divergent from *Nucleocytoviricota* clades

We used AGNOSTOS[20] to characterize 29,414 non-singleton gene clusters from the GOEV database (Table S6). Clustering of *Nucleocytoviricota* and Mirus genomes

based on the occurrence of these gene clusters emphasized the complex functional makeup of the *Megaviricetes* and *Pokkesviricetes* classes, with some clades (e.g., the *Imitervirales* and *Algavirales*) split into multiple groups (Figure 3). In contrast, the Mirus MAGs clustered together, with the lineages being organized into distinct groups that largely recapitulated phylogenomic signals. Overall, the distribution of gene clusters, which is agnostic of any functional inferences, indicates that Mirus MAGs share genomic traits sufficiently distinct from those occurring in the known *Nucleocytoviricota* lineages. This comparative genomic analysis strongly supports the monophyly of the Mirus MAGs, corroborating the inferences from the phylogenomic analysis of the informational module.

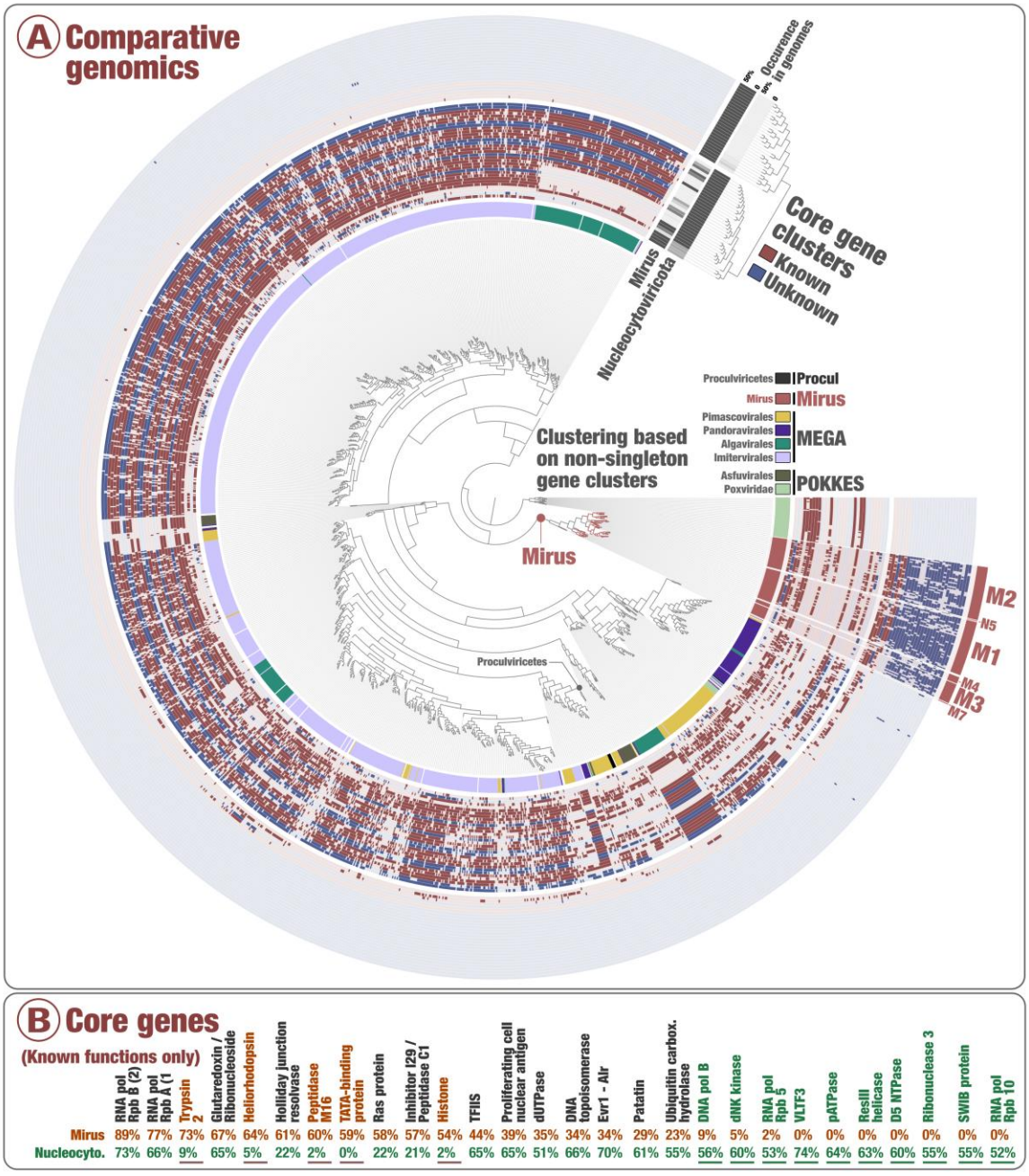


Figure 3: Functional clustering of abundant and widespread marine viruses within mirusviruses and *Nucleocytoviricota*. In panel A, the inner tree is a clustering of 'Mirusviricota' and *Nucleocytoviricota* genomes >100kbp in length based on the occurrence of all the non-singleton gene clusters (Euclidean distance), rooted with the *Chordopoxvirinae* subfamily of *Poxviridae* genomes. Layers of information display the main taxonomy of *Nucleocytoviricota* as well as the occurrence of 60 gene clusters detected in at least 50% of 'Mirusviricota' or *Nucleocytoviricota*. The 60 gene clusters are clustered based on their occurrence (absence/presence) across the genomes. Panel B displays the occurrence of gene clusters of known functions detected in at least 50% of 'Mirusviricota' or *Nucleocytoviricota* genomes.

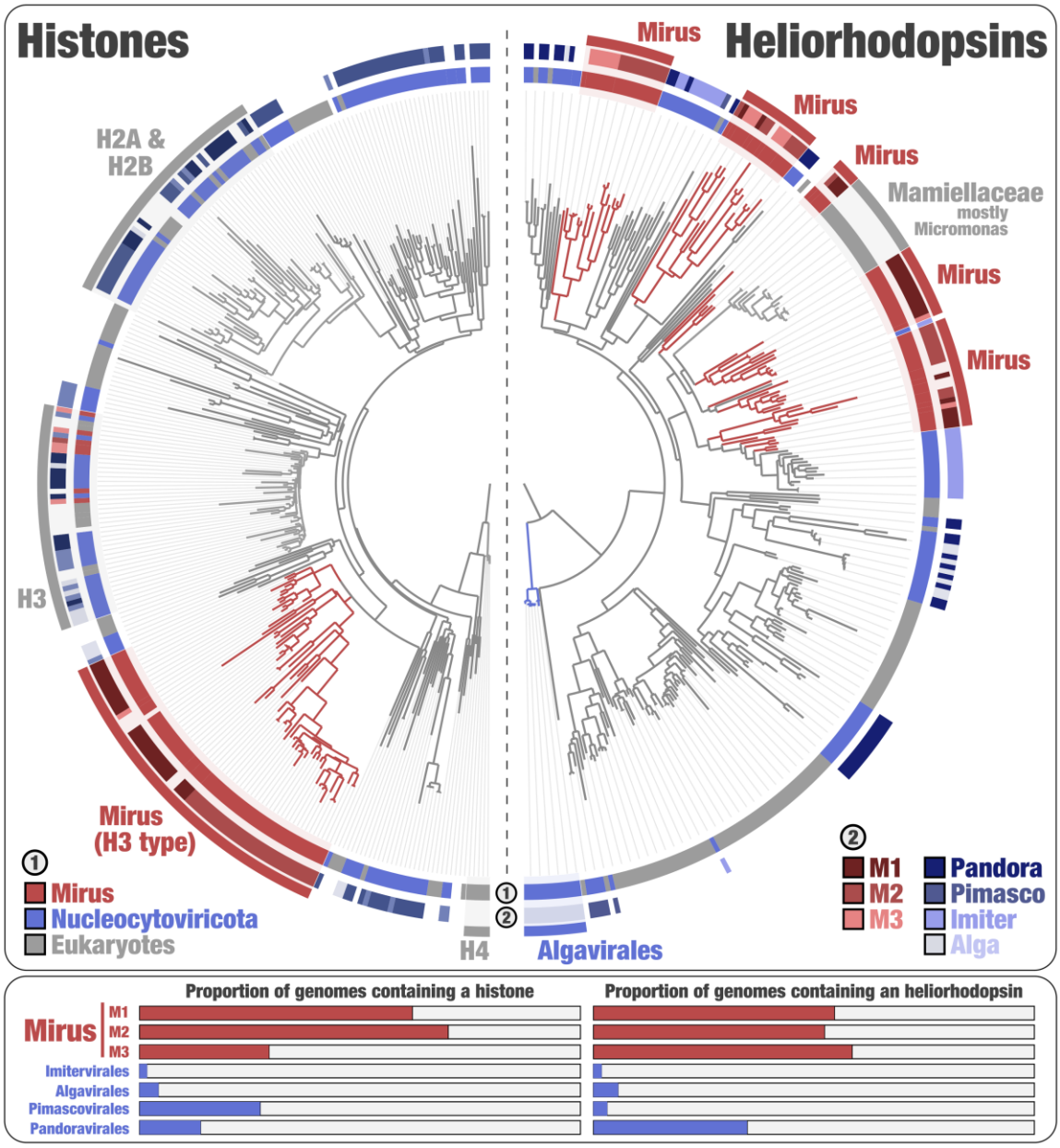


Figure 4: The phylogeny of histones and heliorhodopsins. Left panel displays a maximum-likelihood phylogenetic of histones occurring in MAGs of eukaryotes, *Nucleocytoviricota* and mirusviruses. Mainly eukaryotic clades distant from H2-H3-H4 were excluded in order to focus on the more restrained viral signal. Right panel displays a maximum-likelihood phylogenetic tree of heliorhodopsins occurring in eukaryotes, *Nucleocytoviricota* and mirusviruses.

Among the most widespread gene clusters within the viral genomic database, thirty-five gene clusters occurred in more than 50% of the 111 Mirus MAGs, representing some of the core genes for this newly identified clade (Table S6). Critically, most of these clusters appear to be unique to this clade and 24 lacked any functional annotation (Figure 3A). Among the 11 Mirus core genes with a functional annotation, five were detected in <10% of the known *Nucleocytoviricota* lineages. They correspond to trypsin and M16-family peptidase (proteases), viral rhodopsin (ion-conducting pathway), TATA-binding protein (transcription), and a histone (DNA stability) (Figure 3B). Phylogenetic trees of histones and viral rhodopsins show several monophyletic clades of Mirus sequences, suggesting that they are relatively distant from the corresponding proteins of both *Nucleocytoviricota* and marine planktonic eukaryotes (Figure 4). Notably, heliorhodopsins encoded by *Micromonas* populations (prominent green algae widespread in the sunlit ocean) might have been acquired from mirusviruses through horizontal gene transfer. In contrast to Mirus, 20 out of the 28 gene clusters shared by more than 50% of the *Nucleocytoviricota* genomes within the GOEV database are functionally annotated. Analysis of these core gene clusters showed that they are highly depleted if not entirely missing in Mirus: RNA pol5, RNA pol10, ribonuclease 3 and the SWIB protein (transcription), ResIII helicase and D5 NTPase (DNA replication), and deoxyribonucleoside kinase (nucleotide metabolism) (Figure 3B). Overall, comparative genomics and the functional inferences strongly suggest that Mirus viruses display a unique lifestyle within plankton with a specific functional gene complement - as compared to the known *Nucleocytoviricota* lineages - that remains largely enigmatic due to the predominance of core genes for which functional predictions are challenging.

Mirus is a putative phylum ('*Mirusviricota*') within the realm *Duplodnaviria*

Given that major virion morphogenesis proteins are typically conserved across members of a particular virus lineage, we focused on the core genes defined in the comparative genomic analysis for identification of the 'missing' virion module of Mirus. We compared HMMs of the Mirus core proteins to a comprehensive HMM database of viral and cellular proteins. This analysis provided no hits for either the double jelly-roll MCP or the FtsK-like genome packaging ATPase, two virion module markers characteristic of the realm *Varidnaviria*. Instead, this comparison provided a remote hit to a HK97-fold MCP encoded by a *Caudoviricetes* bacteriophage. The corresponding function occurs in 82% of the Mirus MAGs, mostly in single copy. The hit covers 68% of the template MCP (268/351 aa; YP_004306755), but encompasses only an N-terminal fragment of the considerably longer (>800 aa) query Mirus HMM. To validate the result of the profile-profile comparison and to understand the discrepancy between the phage and Mirus protein sizes, we predicted the 3D structure of the corresponding Mirus protein using AlphaFold2[21,22] and RoseTTAFold[23]. Both programs produced similar models with a readily recognizable HK97-fold domain (Figure 5), corresponding to the N-terminal region recognized in the profile-profile comparisons. Notably, in herpesvirus MCPs, the HK97-fold

domain, referred to as the 'floor' domain, responsible for capsid shell formation, is embellished with a 'tower' domain which projects away from the surface of the assembled capsid[24]. The 'tower' domain is an insertion within the A-subdomain of the core HK97-fold[24,25]. Remarkably, the Mirus protein contained an insertion within the A-subdomain, equivalent to the 'tower' domain of herpesvirus MCPs (Figure 5). Such 'tower' domains have not been thus far described for any member of the *Caudoviricetes*, including the so-called jumbo phages (i.e., phages with very large genome[26]), explaining the limited coverage of the hit in profile-profile comparisons[18]. The size of the Mirus MCP 'tower' suggests that this protein could represent an evolutionary intermediate between *Caudoviricetes* (no tower) and *Herpesvirales* (larger tower) MCPs. Protein divergences prevented any resolutive phylogenetic analyses to compare the HK97-fold MCP of *Herpesvirales*, *Caudoviricetes* and Mirus. Yet, phylogenetic inferences of the Mirus MCP alone perfectly recapitulates the seven clades initially identified based on the informational module (Figure 2, panel C), indicative of coevolution of the two functional modules.

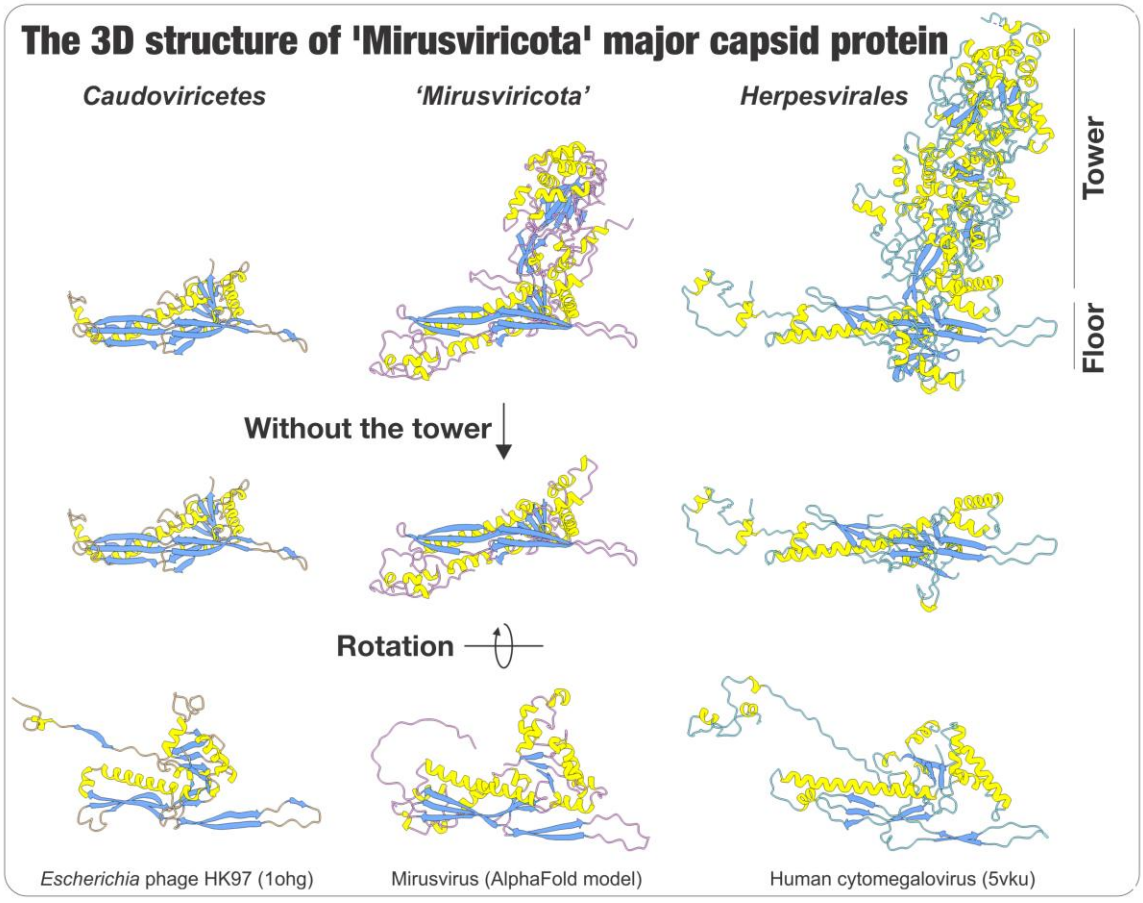


Figure 5: 3D structure of the major capsid protein (MCP). The figure displays MCP 3D structures for *Escherichia* phage HK97 (*Caudoviricetes*), a reference genome for the mirusviruses (estimated using Alphafold), and the human cytomegalovirus (*Herpesvirales*).

Consistent with the identification of the HK97-fold MCP, further profile-profile comparisons uncovered all other components of the *Duplodnaviria*-specific virion module, including the terminase (ATPase-nuclease, key component of the DNA packaging machine), portal protein and the capsid maturation protease (Pfam PF04586, known as assemblin in herpesviruses). The presence of the entire virion module genes in mirusviruses strongly suggests that they are *bona fide* large DNA viruses capable of forming virus particles similar to those of viruses in the realm *Duplodnaviria*.

Analysis of the gene content of *Nucleocytoviricota*, *Herpesvirales* and mirusviruses indicates that DNApolB is a rare hallmark core gene shared by the three groups of eukaryote-infecting DNA viruses. Phylogenetic inferences of this marker using not only the viral genomic database but also a wide range of eukaryotic and additional viral lineages[27] supported the evolutionary distance of mirusviruses, being distinct from all other known clades of double-stranded DNA viruses (Figure 6). The Mirus DNApolB was positioned in the vicinity of *Herpesviridae, Megaviricetes* and the Zeta group of eukaryotes (the closest *Nucleocytoviricota* relative being '*Pandoravirales*'), distant from both *Pokkesviricetes* and *Caudoviricetes*.

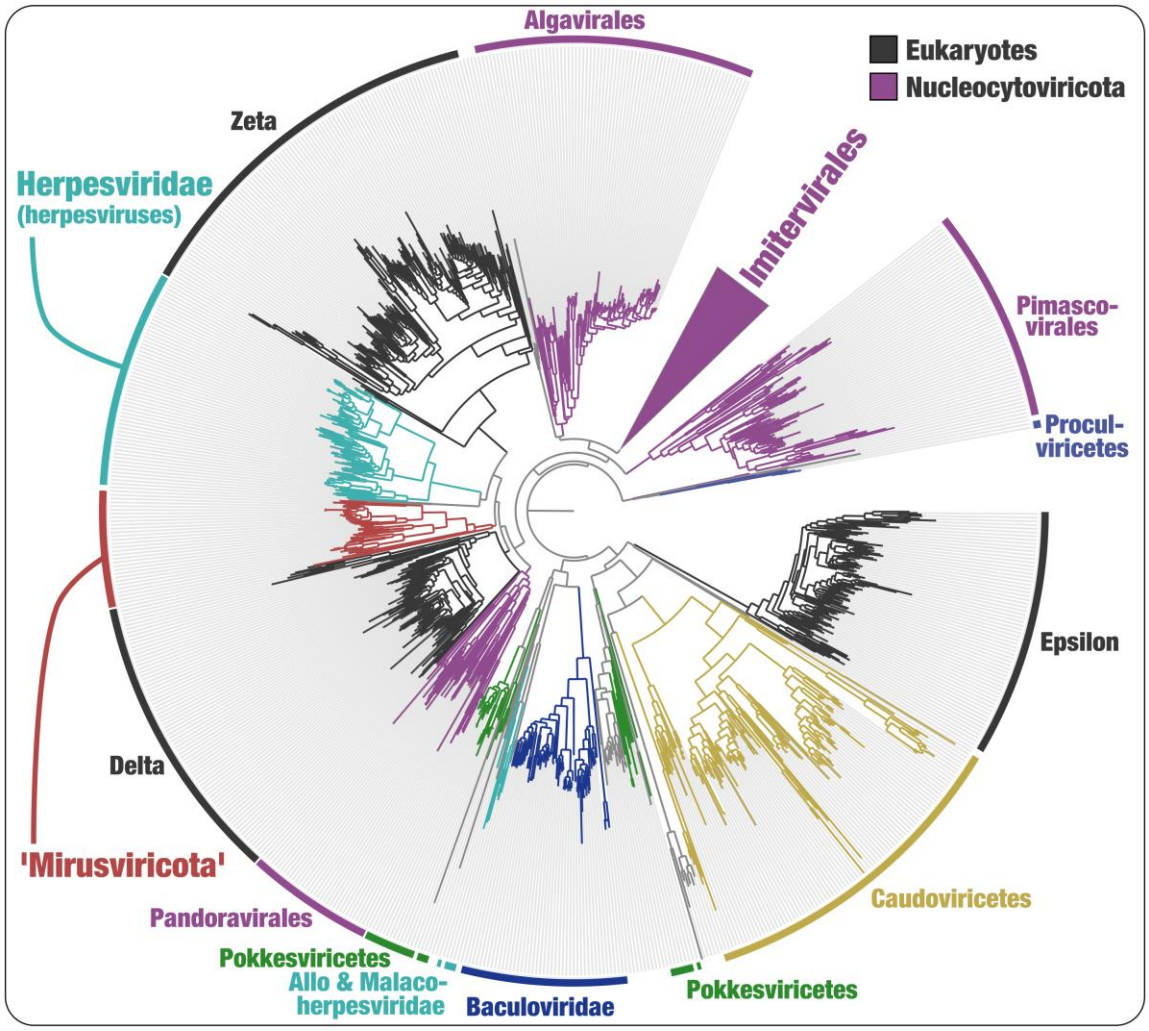


Figure 6: Phylogeny of the DNApolB hallmark genes. The figure displays a maximum-likelihood phylogenetic tree of DNA-polymerase B-family sequences (1,080 sites, 2,213 sequences) from the database described herein, *Duplodnaviria* and *Baculoviridae* sequences from the NCBI viral genomic database, and eukaryotic and viral sequences from Kazlauskas et al.[27]. Eukaryotic Epsilon-type and related sequences were used as outgroup.

Taken together, the (i) considerable genetic distances between the virion modules of mirusviruses, *Caudoviricetes* and *Herpesvirales* (preventing robust and informative phylogenetic inferences), (ii) distinct 3D structure of the mirusvirus MCP, (iii) and DNApolB phylogenetic inferences, indicate that mirusviruses represent a putative new phylum within the *Duplodnaviria*, which we dubbed '*Mirusviricota*'. These viruses would hence represent a third clade within realm *Duplodnaviria*, next to the well-known *Caudoviricetes* (phylum *Uroviricota*) and *Herpesvirales* (phylum *Peploviricota*) lineages. To our knowledge, '*Mirusviricota*' represents the first eukaryote-infecting lineage of *Duplodnaviria* found to be widespread and abundant within plankton in the sunlit ocean.

The chimeric nature of 'Mirusviricota' links two major realms of DNA viruses

On the one hand, mirusviruses belong to the realm *Duplodnaviria* based on their virion module (Figures 2 and 5). On the other hand, hallmark informational markers RNApolA, RNApolB and TFIIS (Figure S2), that are all missing in known *Herpesvirales* lineages, display high sequence similarities to the corresponding proteins encoded by members of the phylum *Nucleocytoviricota*. Such atypical distribution of the 'Mirusviricota' attributes places this virus group in between the two major realms of DNA viruses, *Duplodnaviria* (virion module) and *Varidnaviria* (informational module and other core functions). To further validate the genomic content of mirusviruses and to exclude the possibility of artificial chimerism, we created an HMM for the newly identified '*Mirusviricota'* MCP and used it as bait to search for complete genomes in additional databases. First, we only found two '*Mirusviricota'* MCPs in a comprehensive viral genomic resource from the <0.2 μm size fraction of the surface oceans (GOV2[10]), suggesting that virions in this clade are usually >0.2 μm in size. We subsequently screened the '*Mirusviricota'* MCP in a database containing hundreds of metagenomic assemblies from the 0.2-3 μm size fraction of the surface oceans[28]. We found a contiguous '*Mirusviricota*' genome in the Mediterranean Sea with a length of 431.5 kb, just 6 kb shorter than the longest '*Mirusviricota'* MAG. This near-complete contiguous genome contains all marker genes of '*Mirusviricota'* and is affiliated to the clade M2 based on both the information module and MCP phylogenies (see Figure 2, panels B and C).

We compared the 355 genes found in the near-complete contiguous genome to two comprehensive genomic databases corresponding to the realms *Duplodnaviria* (748,546 genes and 57,295 HMMs) and *Varidnaviria* (269,523 genes and 16,689 HMMs)[29] and found 86 significant hits (HMM searches, e-value <e-6). Only six of them had better matches within the *Duplodnaviria* database (hits to phages and jumbo phages) and included the terminase protein (best match for a Vibriophage). The remaining 80 genes had better matches within the *Varidnaviria* (hits to *Nucleocytoviricota* only) and occurred relatively homogeneously across the genome (Figure 7). These include the RNApolA, RNApolB, DNApolB, topoisomerase II, TATA-binding protein, histone, multiple heliorhodopsins, Ras-related GTPases, cell surface receptor, ubiquitin, and trypsin. While the evolutionary trajectories of the corresponding genes remain uncertain, the shared gene content emphasizes a strong functional connectivity between mirusviruses and the large and giant DNA viruses within the realm *Varidnaviria*.

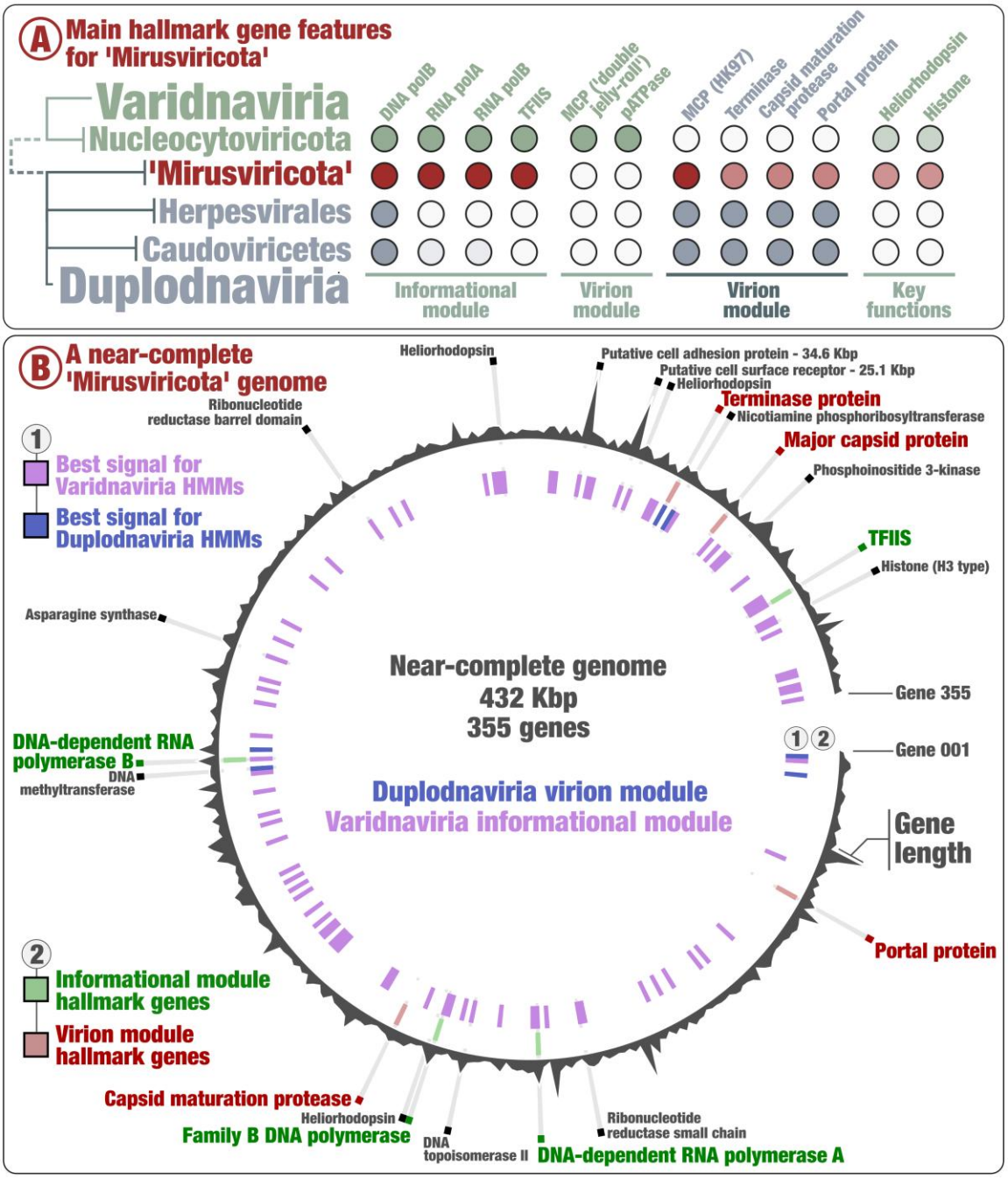


Figure 7: A near-complete genome for '*Mirusviricota*'. The figure displays the length of 355 genes found in a near complete genome of 'Mirusviricota' (clade M2), along with their link to two viral domains (gene versus HMM signal). The figure also highlights hallmark genes for the informational and particle modules of the virus.

The near-complete contiguous genome also included two unusually long genes. The first one, 34.6 kb in length, is longer than any gene in the *Duplodnaviria* and *Nucleocytoviricota* genomic databases and might represent a record holder in terms of length for a viral gene. NCBI blast provided a weak hit for a YadA-like family protein (cell attachment) used by pathogenic bacteria to invade eukaryotic cells. The second one is 25.1 kb in length and NCBI blast provided a weak hit for a Hyalin

Repeat domain also involved in bacterial cell adhesion. Looking at gene lengths more broadly, we also found genes >20 Kbp in *Nucleocytoviricota* (up to 30.7 kb) and *Caudoviricetes* (up to 25,7 kb) but not in *Herpesvirales* (longest gene of 13.9 kb only). Thus, mirusviruses contain unusually long genes, a trait shared with certain members of the *Nucleocytoviricota* and *Caudoviricetes*, which may represent common strategies for invading planktonic cells.

Overall, the near-complete contiguous '*Mirusviricota*' genome perfectly recapitulated the chimeric attributes of this putative phylum as initially observed based on >100 manually curated MAGs. Sequence similarities between proteins indicated that key functionalities within and beyond the informational module connect '*Mirusviricota*' to *Nucleocytoviricota*. Since 'Mirusviricota' belongs to the *Duplodnaviria* based on the presence of this realm-specific virion module, we demonstrated that the genomic content of this putative new phylum fills a critical evolutionary gap between the two major realms of double-stranded DNA viruses.

Discussion

Our phylogeny-guided genome-resolved metagenomic survey of plankton in the surface of five oceans and two seas exposed a major clade of large eukaryotic DNA viruses, with genomes up to >400 Kbp in length, that are diverse and prevalent in the sunlit ocean. This clade, dubbed '*Mirusviricota'*, corresponds to a putative new phylum within the realm *Duplodnaviria* that until now included the bacteria- and archaea-infecting *Caudoviricetes* and animal-infecting *Herpesvirales*. The '*Mirusviricota*' phylum is organized into at least seven sub-clades that might correspond to distinct families. Although both mirusviruses and *Herpesvirales* include the eukaryote-infecting duplodnaviruses, they display very different genomic features. Most notably, mirusviruses substantially deviates from all other previously characterized groups of DNA viruses, with the virion morphogenesis module (a basis for highest-rank viral taxonomy) affiliated to the realm *Duplodnaviria* and the informational module closely related to that of large and giant viruses within the realm *Varidnaviria*. These chimeric attributes were recapitulated in a near-complete contiguous genome of 431.5 Kbp and could explain why previous environmental surveys did not identify this putative phylum despite its considerable metagenomic signal. '*Mirusviricota*' has a substantial functional overlap with large and giant eukaryotic varidnaviruses that goes well beyond the informational module. The discovery of '*Mirusviricota*' is a reminder that we have not yet grasped the full ecological and evolutionary complexity of even the most abundant double-stranded DNA viruses in key ecosystems such as the surface of our oceans and seas.

To our knowledge, *'Mirusviricota'* is the first identified eukaryotic virus clade within the *Duplodnaviria* realm that is widespread within plankton. Biogeographic signal computed by means of genome-wide metagenomic read recruitments indicates that mirusviruses are prevalent and relatively abundant in various regions of the sunlit oceans, from pole to pole. In addition, distribution patterns across planktonic

cellular size fractions mirrored those observed for the *Nucleocytoviricota* lineages, strongly suggesting that mirusviruses actively infect eukaryotic plankton, possibly with a broad distribution of the cellular sizes. Viral-host predictions using the eukaryotic environmental genomic data characterized from *Tara* Oceans[15] linked some mirusviruses to several eukaryotic lineages including key phototrophs (e.g., haptophytes) that now await experimental validations. In addition, the strong signal for '*Mirusviricota*' in the 0.2-3 µm size fraction contrasts with a lack of signal for its major capsid protein in a comprehensive viral genomic resource from the <0.2 µm size fraction[10]. This suggests that the virion of '*Mirusviricota'*, which remains to be visualized, is larger than 0.2 µm. Indeed, the virions of alloherpesviruses, which infect fish and have substantially smaller genomes compared to mirusviruses, are up to 0.2 µm in diameter[30]. In terms of functions, comparative genomics indicates that '*Mirusviricota'* viruses have a rather complex lifestyle (wide array of functions), and encode phylogenetically distinct branches of H3 histones (proteins involved in chromatin formation within the eukaryotic cells[31]) and heliorhodopsins (viral rhodopsins, light-gated channels already identified in *Nucleocytoviricota* lineages, for which functional role remains unclear[32,33]). Analysis of the biogeography and functional gene repertoire suggest that mirusviruses influence the ecology of key marine eukaryotes, using a distinct lifestyle that resembles, to some extent, that of marine varidnaviruses. For instance, phylogenetic analysis suggests that a core *Micromonas* heliorhodopsin has originated from a mirusvirus, implicating them as important players in the evolution of planktonic eukaryotes by means of gene flow. Concurrently, the close sequence similarity between mirusvirus and *Micromonas* heliorhodopsins indicates that green algae might serve as hosts for some of the '*Mirusviricota*' lineages. As of now, the prevalence and diversification of '*Mirusviricota*' indicates that we just started exploring the full extent of the genomic diversity of this virus group within plankton, and possibly beyond. Moving forward, the genomics of '*Mirusviricota*' could help solve many hypotheses that orbit around the large and giant eukaryotic DNA viruses, from their potential roles in eukaryogenesis[5,34–38], to their regulation of ecologically critical marine eukaryotes[39,40].

Viruses of the *Herpesvirales* and *Nucleocytoviricota* are descendants of two ancient virus lineages, *Duplodnaviria* and *Varidnaviria*, respectively, with the corresponding ancestors antedating the last universal cellular ancestor (LUCA)[1,2] . Nevertheless, the exact evolutionary trajectories and the identity of the respective most recent common ancestors of these prominent eukaryote-infecting double-stranded DNA viral clades remain elusive, in part due to the lack of known intermediate states. Particularly puzzling is the evolutionary trench between relatively simple varidnaviruses (those infecting Bacteria and Archaea, as well as virophages and *Adenoviridae*), with minimalistic gene repertoires for virion formation and genome replication, and complex eukaryotic varidnaviruses (the *Nucleocytoviricota*) with large and giant genomes encoding multiple functions, including nearly complete genome replication and transcription machineries. Similarly enigmatic is the gap between the ubiquitous *Caudoviricetes*, some of which rival *Nucleocytoviricota* in terms of functional complexity and richness of their gene repertoires, and

Herpesvirales, which are restricted to animal hosts and uniformly lack the transcription machinery and practice nuclear replication. The chimeric attributes of '*Mirusviricota'* might now help to clarify both of these long-standing questions. It is possible that genome complexity observed in the contemporary '*Mirusviricota*' lineages has been attained and inherited from a common ancestor with large *Caudoviricetes*, an ancient virus lineage which is believed to have diversified at the time of LUCA. Indeed, most of the proteins involved in genome replication and nucleotide metabolism conserved in mirusviruses are also widespread in *Caudoviricetes.* The closer sequences similarity of these proteins to homologs in *Nucleocytoviricota* could then suggest that the genes were transferred between ancestors of the two groups after the divergence of '*Mirusviricota*' and *Caudoviricetes.* Such transfer of informational module from a mirusvirus to the ancestor of *Nucleocytoviricota* would explain a sudden transition from small varidnaviruses to complex *Nucleocytoviricota*. It is unlikely that the complexity of the *Nucleocytoviricota* genomes has evolved through gradual accretion of informational genes in a particular lineage of varidnaviruses. The described scenario provides a logical explanation to the emergence of *Nucleocytoviricota*, without invoking unknown and unknowable extinct virus lineages. Beside the core functions responsible for virion formation and information processing, another important component of both *Nucleocytoviricota* and '*Mirusviricota*' genomes, specifically related to eukaryote-specific aspects of virus-host interactions, has likely evolved more recently though horizontal gene transfers or convergent gene acquisition from different sources, facilitated by shared hosts and ecosystems. Finally, the identification of '*Mirusviricota*' extends the presence of duplodnaviruses beyond animals to eukaryotic plankton hosts, strongly suggesting their ancient association with eukaryotes. The presence of the tower domain in both '*Mirusviricota*' and *Herpesvirales* MCPs is suggestive of their common ancestry, rather than independent evolution from distinct *Caudoviricetes* clades. With the shorter size of the tower domain and considering later emergence of animals compared to unicellular eukaryotes, '*Mirusviricota*' viruses might more closely resemble the ancestral state of eukaryotic duplodnaviruses. If so, animal herpesviruses would have evolved by reductive evolution, in particular, by losing the transcription machinery and switching to nuclear replication[1,2].

Given the relatively early stage of viral metagenomics, it is more than likely that other intermediate states between '*Mirusviricota*' and *Herpesvirales* on one hand, and '*Mirusviricota*' and large and giant viruses (the *Nucleocytoviricota*) on the other hand also exist. Their discovery, coupled with phylogenomic analyses and genome-wide 3D structure comparisons will be instrumental in further refinement of our understanding of the evolutionary trajectories of prominent eukaryotic DNA viruses within the realms *Duplodnaviria* and *Varidnaviria*.

Material & methods:

***Tara* Oceans metagenomes.** We analyzed a total of 937 *Tara* Oceans metagenomes available at the EBI under project PRJEB402 (https://www.ebi.ac.uk/ena/browser/view/PRJEB402). Table S1 reports general information (including the number of reads and environmental metadata) for each metagenome.

Constrained automatic binning with CONCOCT. The 798 metagenomes corresponding to size fractions ranging from 0.8 µm to 2 mm were previously organized into 11 'metagenomic sets' based upon their geographic coordinates[14,41]. Those 0.28 trillion reads were used as inputs for 11 metagenomic co-assemblies using MEGAHIT[42] v1.1.1, and the contig header names were simplified in the resulting assembly outputs using anvi'o[43] v6.1. Co-assemblies yielded 78 million contigs longer than 1,000 nucleotides for a total volume of 150.7 Gbp. Constrained automatic binning was performed on each co-assembly output, focusing only on the 11.9 million contigs longer than 2,500 nucleotides. Briefly, (1) anvi'o profiled contigs using Prodigal[44] v2.6.3 with default parameters to identify an initial set of genes, (2) we mapped short reads from the metagenomic set to the contig using BWA v0.7.15[45] (minimum identity of 95%) and stored the recruited reads as BAM files using samtools[46], (3) anvi'o profiled each BAM file to estimate the coverage and detection statistics of each contig, and combined mapping profiles into a merged profile database for each metagenomic set. We then clustered contigs with the automatic binning algorithm CONCOCT[47] by constraining the number of clusters per metagenomic set to a number ranging from 50 to 400 depending on the set (total of 2,550 metagenomic blocks from ~12 million contigs).

Diversity of DNA-dependent RNA polymerase B subunit genes. We used HMMER [48] v3.1b2 to detect genes matching to the DNA-dependent RNA polymerase B subunit (RNApolB) among all 2,550 metagenomic blocks based on a single HMM model. We used CD-HIT[49] to create a non-redundant database of RNApolB genes at the amino acid level with sequence similarity <90% (longest hit was selected for each cluster). Short sequences were excluded. Finally, we included reference RNApolB amino acid sequences from Bacteria, Archaea, Eukarya and giant viruses[5]: The sequences were aligned with MAFFT[50] v7.464 and the FFT-NS-i algorithm with default parameters, and trimmed at >50% gaps with Goalign v0.3.5 (https://www.github.com/evolbioinfo/goalign). We performed a phylogenetic reconstruction using the best fitting model according to the Bayesian Information Criterion (BIC) from the ModelFinder[51] Plus option with IQ-TREE[52] v1.6.2. We visualized and rooted the phylogeny using anvi'o. This tree allowed us to identify new RNApolB clades.

Phylogeny-guided genome-resolved metagenomics. Each metagenomic block containing at least one of the RNApolB genes of interest (see previous section) was manually binned using the anvi'o interactive interface to specifically search for NCLDV MAGs. First, we used HMMER[48] v3.1b2 to identify eight hallmark genes (eight distinct HMM runs within anvi'o) as well 149 additional orthologous groups often found in reference NCLDVs[5] (a single HMM run within anvi'o). The interface

considers the sequence composition, differential coverage, GC-content, and taxonomic signal of each contig, and displayed the eight hallmark genes as individual layers as well 149 additional orthologous groups often found in reference NCLDVs[5] as a single extra layer for guidance. During binning, no restriction was applied in term of number of giant virus core gene markers present, as long as the signal suggested the occurrence of a putative NCLDV MAG. Note that while some metagenomic blocks contained a limited number of NCLDV MAGs, others contained dozens. Finally, we individually refined all the NCLDV MAGs >50kbp in length as outlined in Delmont and Eren[53], and renamed contigs they contained according to their MAG ID.

A non-redundant genomic database of marine NCLDVs. We incorporated into our database marine NCLDV MAGs characterized using automatic binning by Schulz et al.[7] (n=743) and Moniruzzaman et al.[8] (n=444), in part using *Tara* Oceans metagenomes. We also incorporated 235 reference NCLDV genomes mostly characterized by means of cultivation but also cell sorting within plankton[54]. We determined the average nucleotide identity (ANI) of each pair of NCLDV MAGs using the dnadiff tool from the MUMmer package[55] v4.0b2. MAGs were considered redundant when their ANI was >98% (minimum alignment of >25% of the smaller MAG in each comparison). Manually curated MAGs were selected to represent a group of redundant MAGs. For groups lacking manually curated MAGs, the longest MAG was selected. This analysis provided a non-redundant genomic database of 1,593 marine MAGs plus 224 reference genomes. We created a single CONTIGs database for this set of NCLDV genomes using anvi'o. Prodigal[44] was used to identify genes.

Curation of hallmark genes. The amino-acid sequence datasets for RNApolA, RNApolB, DNApolB, and TFIIS were manually curated through BLASTp alignments (BLAST[56] v2.10.1) and phylogenetic reconstructions, as previously described for eukaryotic hallmark genes[41]. Briefly, multiple sequences for a single hallmark gene within the same MAG were inspected based on their position in a corresponding single-protein phylogenetic tree performed with the same protocol as described above ("Diversity of DNA-dependent RNA polymerase B subunit genes" section). The genome's multiple sequences were then aligned with BLASTp to their closest reference sequence, and to each other. In case of important overlap with >95% identity (likely corresponding to a recent duplication event), only the longest sequence was conserved; in case of clear split, the sequences were fused and accordingly labeled for further inspection. Finally, RNApolA and RNApolB sequences shorter than 200 aa were also removed, as DNApolB sequences shorter than 100 aa, and TFIIS sequences shorter than 25 aa. This step created a set of curated hallmark genes.

Alignments, trimming, and single-protein phylogenetic analyses. For each of the four curated hallmark genes, the sequences were aligned with MAFFT[50] v7.464 and the FFT-NS-i algorithm with default parameters. Sites with more than 50% gaps were trimmed using Goalign v0.3.5 (https://www.github.com/evolbioinfo/goalign).

IQ-TREE[52] v1.6.2 was used for the phylogenetic reconstructions, with the ModelFinder[51] Plus option to determine the best fitting model according to BIC. Supports were computed from 1,000 replicates for the Shimodaira-Hasegawa (SH)-like approximation likelihood ratio (aLRT)[57] and ultrafast bootstrap approximation (UFBoot[58]). As per IQ-TREE manual, supports were deemed good when SH-like aLRT >= 80% and UFBoot >= 95%. Anvi'o v7.1 was used to visualize and root the phylogenetic trees.

Resolving hallmark genes occurring multiple times. We manually inspected all the duplicated sequences (hallmark genes detected multiple times in the same genome) that remained after the curation step, in the context of the individual phylogenetic trees (see previous section). First, duplicates were treated as putative contaminations based on major individual (i.e. not conserved within a clade) incongruences with the position of the corresponding genome in the other single-protein trees. The putative contaminants were easily identified and removed. Second, we identified hallmark gene paralogs encapsulating entire clades and/or subclades (Fig S2), suggesting that the duplication event occurred before the diversification of the concerned viral clades. This is notably the case for the majority of *Imitervirales*, which have two paralogs of the RNApolB. These paralogs were conserved for single-protein trees, but only the paralog clades with the shortest branch were conserved for congruence inspection and concatenation. Finally, we also detected a small clade of *Algavirales* viruses containing a homolog of TFIIS branching distantly from the ordinary TFIIS type, suggesting a gene acquisition. These sequences were not included in subsequent analyses. This step created a set of curated and duplicate-free hallmark genes.

Identification and modeling of the Mirus major capsid protein. The putative major capsid protein of Mirus as well as the other morphogenetic module proteins were identified using HHsearch against the publicly available Pfam v35, PDB70, and UniProt/Swiss-Prot viral protein databases[59,60]. The candidate MCP was then modeled using the AlphaFold2[21,22] and RoseTTAFold[23]. The resulting 3D models were then compared to the major capsid protein structures of phage HK97 and human cytomegalovirus and visualized using ChimeraX[61].

Supermatrix phylogenetic analysis of NCLDV genomes. Concatenations of the four aligned and trimmed curated and duplicated-free hallmark genes (methods as described above) were performed in order to increase the resolution of the phylogenetic tree. Genomes only containing TFIIS out of the four hallmark genes were excluded. For the remaining MAGs and reference genomes, missing sequences were replaced with gaps. Ambiguous genomes, determined based on the presence of major and isolated (i.e. not a clade pattern) incongruences within single and concatenated proteins trees, as well as on frequent long branches and unstable positions in taxon sampling inferences, were removed. The concatenated phylogenetic trees were reconstructed using IQ-TREE[52] v1.6.2 with the best fitting model according to the BIC from the ModelFinder[51] Plus option. The resulting tree was then used as a guide tree for a phylogenetic reconstruction based on the site-

specific frequency PMSF mixture model[62] (LG+C30+F+R10). For the concatenated trees, supports were computed from 1,000 replicates for the Shimodaira-Hasegawa (SH)-like approximation likelihood ratio (aLRT)[57] and ultrafast bootstrap approximation (UFBoot[58]). As per IQ-TREE manual, supports were deemed good when SH-like aLRT >= 80% and UFBoot >= 95%. Anvi'o v7.1 was used to visualize and root the phylogenetic trees.

Taxonomic inference of NCLDV MAGs. We determined the taxonomy of NCLDV MAGs based on the phylogenetic analysis results, using guidance from the reference genomes as well as previous taxonomical inferences by Schulz et al.[7], Moniruzzaman et al.[8] and Aylward et al.[16].

Biogeography of NCLDV genomes. We performed a mapping of all metagenomes to calculate the mean coverage and detection of the marine NCLDV genomic database. Briefly, we used BWA v0.7.15 (minimum identity of 90%) and a FASTA file containing the 1,593 MAGs and 224 reference genomes to recruit short reads from all 937 metagenomes. We considered MAGs were detected in a given filter when >25% of their length was covered by reads to minimize non-specific read recruitments [63]. The number of recruited reads below this cut-off was set to 0 before determining vertical coverage and percent of recruited reads.

Viral-host predictions. Ecological network analysis was performed using a relative abundance matrix of '*Mirusviricota*' MAGs in the pico-size fractions (0.22-3 µm) and relative abundances of eukaryotic genomes in the following five size fractions: 0.8-5 µm, 3-20 µm, 20-180 µm, 180-2,000 µm, and 0.8-2000 µm. To create the input files for network inference, we combined the Mirus matrix with each of the eukaryotic matrices (corresponding to different size fractions), and only the samples represented by both virus and eukaryotes were placed in new files. Relative abundances in the newly generated matrices were normalized using centered log-ratio (*clr*) transformation to Mirus and eukaryotes separately. Only MAGs observed in at least three samples were considered. Ecological network was built using FlashWeave[64] v0.15.0 sensitive model with Julia v1.3.1. A threshold to determine the statistical significance was set to alpha = 0.01. To compare the performance of FlashWeave to naive correlations, we calculated the spearman rank correlation coefficient using the package Scipy v1.3.1. To build a global Mirus-eukaryote interactome, we pooled associations from the five size fractions by keeping the best positive or negative associations of each genome pair (i.e., the edges with the highest absolute weights). We used a phylogeny-guided filtering approach, Taxon Interaction Mapper (TIM)[18], to predict the host using the global virus-eukaryote interactome. All the virus-eukaryote associations were mapped on the '*Mirusviricota*' informational module phylogenetic tree to calculate the significance of the enrichment of specific associations using TIM. TIM provided a list of nodes in the viral tree and associated NCBI taxonomies (order, class, and phylum) of eukaryotes that show significant enrichment in the leaves under the nodes.

Cosmopolitan score. Using metagenomes from the Station subset 1 (n=757; excludes the 0.8-2000 µm size fraction lacking in the first leg of the *Tara* Oceans expeditions), NCLDV MAGs and reference genomes were assigned a "cosmopolitan score" based on their detection across 119 stations, as previously computed for eukaryotic MAGs[39].

AGNOSTOS functional aggregation inference. AGNOSTOS partitioned protein coding genes from the marine NCLDV genomic database in groups connected by remote homologies, and categorized those groups as members of the known or unknown coding sequence space based on the workflow described in Vanni et al. 2020[20]. AGNOSTOS produces groups of genes with low functional entropy as shown in Vanni et al. 2020[20] and Delmont et al. 2020[41] allowing us to provide functional annotation (Pfam domain architectures) for some of the gene clusters using remote homology methods.

Functional inferences of NCLDV genomes. Genes from the marine NCLDV genomic database were BLASTP against Virus-Host DB[65], RefSeq[66], UniRef90[67], NCVOGs[68], and NCBI nr database using Diamond[69] v2.0.6 with a cut-off E-value 1×10^{-5}. A recently published GVOG database[16] was also used in annotation using hmmer[48] v3.2.1 search with E-value 1×10^{-3} as a significant threshold. In addition, KEGG Orthology (KO) and functional categories were assigned with the Eggnog-Mapper[70] v2.1.5. Finally, tRNAscan-SE[71] v2.0.7 predicted 7,734 tRNAs.

Realm assignation of genes. Two in-house HMM databases were created as follows. First, all coding sequences (CDS) labeled as "*Nucleocytoviricota*" were removed from the *Varidnaviria* CDS dataset (N= 53,776) in the Virus-Host Database (VHDB[29], May 2022). To this dataset, *Tara* Ocean NCLDV MAGs (all were manually curated) and 235 reference NCLDV genomes were integrated. The final *Nucleocytoviricota* protein database contained 269,523 CDS. Similarly, we replaced all *Herpesvirales* CDS in the VHDB *Duplodnaviria* CDS dataset with *Herpesvirales* protein sequences downloaded from NCBI in April 2022. Additionally, a marine *Caudovirales* database including jumbo phage environmental genomes[12,72] was integrated into the *Duplodnaviria* proteins. The final *Duplodnaviria* protein database contained 748,546 proteins. Proteins in the two databases were independently clustered at 30% sequence identity (-c 0.4 --cov-mode 5), using Linclust in MMseqs[73] v13-45111. Gene clusters with fewer than 3 genes were removed, and the remaining gene clusters were aligned using [50] v 7.487. HMM files (N= 16,689 and 57,259 for *Varidnaviria* and *Duplodnaviria*, respectively) were created using the hmmhuild in HMMER3[74] v3.2.1. All proteins in the near-complete '*Mirusviricota*' genome were searched against the two custom HMM databases using the hmmsearch with a cut-off E-value 1×10^{-6}.

Statistical analyses. A "greater" Fisher's exact test was employed to identify KO functions as well as gene clusters with remote homologies that are differentially occurring between the 111 'Mirusviricota' MAGs on one side, and all other NCLDVs

in the database on the other side. P-values were corrected using the Benjamini-Hochberg procedure in R, and values <0.05 were considered significant.

Naming of Mirus and Procul. The latin adjective "**Mirus**" (*surprising*, *strange*) was selected to describe the putative new *Duplodnaviria* phylum: the '***Mirusviricota'***. The latin adverb "**Procul**" (away, at distance, far off) was selected to describe the putative new class of NCLDVs discovered from the Arctic and Southern Oceans: the '***Proculviricetes'***.

Data availability. Data our study generated has been made publicly available at https://doi.org/10.6084/m9.figshare.20284713.v1. This link provides access to (1) the RNApolB genes reconstructed from the *Tara* Oceans assemblies (along with references), (2) individual FASTA files for the 1,593 non-redundant marine NCLDV and Mirus MAGs (including the 697 manually curated MAGs from our survey) and 224 reference NCLDV genomes contained in the GOEV database, (3) the GOEV anvi'o CONTIGS database, (4) genes and proteins found in the GOEV database, (5) manually curated hallmark genes and corresponding phylogenies, (6) HMMs for hallmark genes, (7) the supplemental tables, (8) and the supplemental information document.

Contributions:

Tom O. Delmont conducted the study. Morgan Gaïa, Lingjie Meng, Mart Krupovic, Chiara Vanni, Eric Pelletier and Tom O. Delmont performed the primary data analysis. Tom O. Delmont completed the genome-resolved metagenomic analysis. Morgan Gaïa and Tom O. Delmont curated the marker genes and identified the biological duplicates. Morgan Gaïa performed phylogenetic and phylogenomic analyses. Lingjie Meng performed functional analyses and virus-host predictions. Hiroyuki Ogata supervised Lingjie Meng. Chiara Vanni produced gene cluster with remote homologies. Mart Krupovic discovered the major capsid protein of '*Mirusviricota*' and other key genes of the virion module, linking the two realms. Eric Pelletier performed comparative genomic and biogeographic analyses. All the authors contributed to interpreting the data and writing the manuscript.

Conflict of interest:

Authors declare having no conflicts of interest.

Acknowledgments:

Our survey was made possible by two scientific endeavors: the sampling and sequencing efforts by the *Tara* Oceans Project, and the bioinformatics and visualization capabilities afforded by anvi'o. We are indebted to all who contributed to these efforts, as well as other open-source bioinformatics tools for their

commitment to transparency and openness. *Tara* Oceans (which includes the *Tara* Oceans and *Tara* Oceans Polar Circle expeditions) would not exist without the leadership of the *Tara* Oceans Foundation and the continuous support of 23 institutes (https://oceans.taraexpeditions.org/). Some of the computations were performed using the platine, titane and curie HPC machine provided through GENCI grants (t2011076389, t2012076389, t2013036389, t2014036389, t2015036389 and t2016036389). This study was in part supported by Japan Society for the Promotion of Science (JSPS) KAKENHI (18H02279, 22H00384), Research Unit for Development of Global Sustainability, Kyoto University Research Coordination Alliance, and the International Collaborative Research Program of the Institute for Chemical Research, Kyoto University (2022-26, 2021-29, 2020-28). M.K. was supported by grants from the l'Agence Nationale de la Recherche (ANR-20-CE20-0009-02 and ANR-21-CE11-0001-01). Part of computational work was performed at the SuperComputer System, Institute for Chemical Research, Kyoto University. This article is contribution number XXX of *Tara* Oceans.

Supplemental figures:

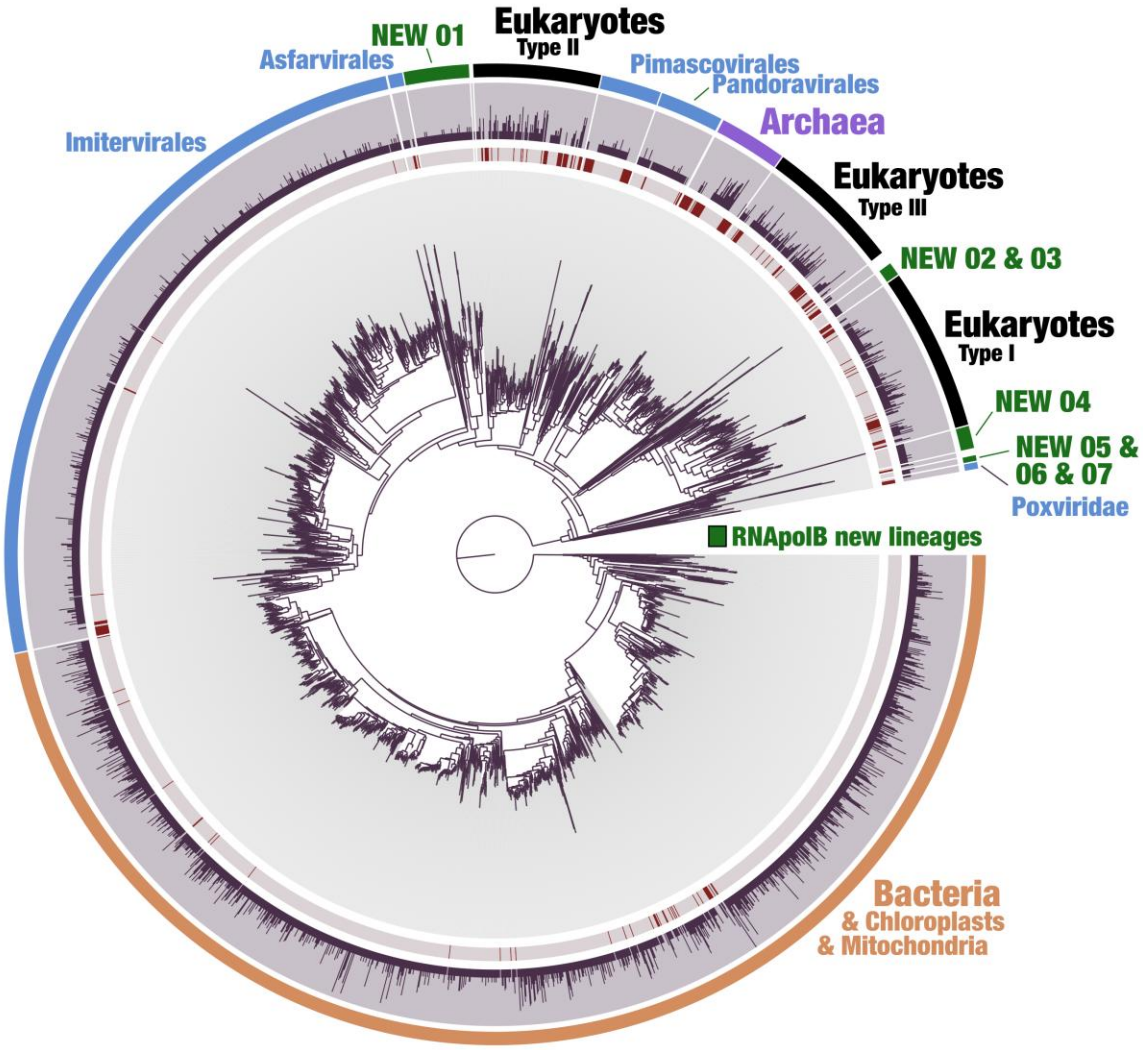


Figure S1: Evolutionary diversity of the DNA-dependent RNA polymerase B in the sunlit ocean. The maximum-likelihood phylogenetic tree is based on 2,728 RNApolB sequences more than 800 amino acids in length with similarity <90% (gray color) identified from 11 large marine metagenomic co-assemblies. This analysis also includes 262 reference RNApolB sequences (red color in the first layer) corresponding to known archaeal, bacterial, eukaryotic and giant virus lineages for perspective. The second layer shows the number of RNApolB sequences from the 11 metagenomic co-assemblies that match to the selected amino acid sequence with identity >90%. Finally, RNApolB new lineages are displayed in green.

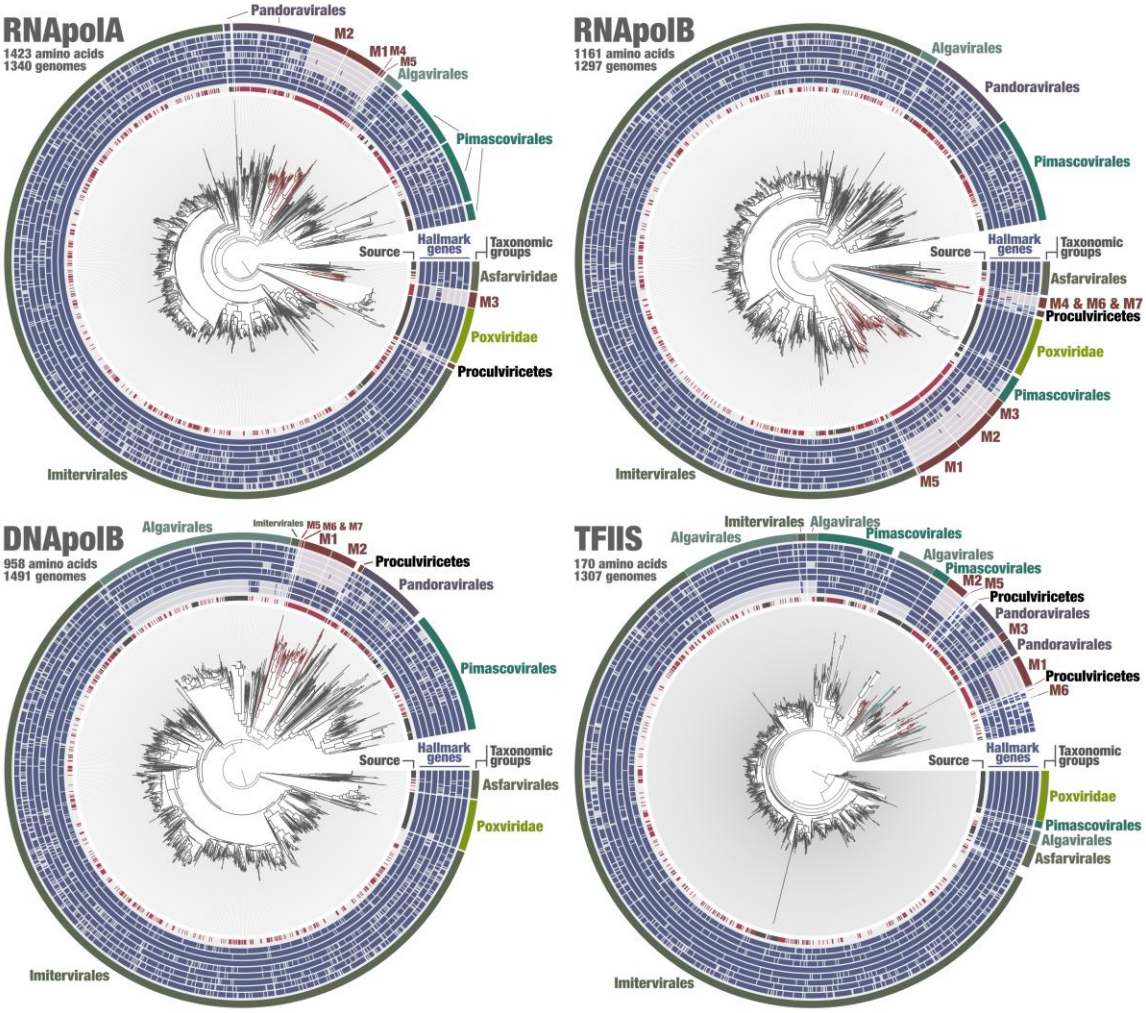


Figure S2: Single-protein phylogenies of the four informational hallmark genes in the GOEV database. Maximum-likelihood phylogenetic trees of the DNA-dependent RNA polymerase subunits A and B, the DNA polymerase B family, and Transcription Factor II-S. In each tree, the branches corresponding to mirusviruses are colored in red. The layers provide additional information, from innermost to outermost: the source (red for the MAGs in our survey, black for references, uncolored for MAGs from other surveys), the presence/absence of hallmark genes in the corresponding genomes, and their taxonomy.

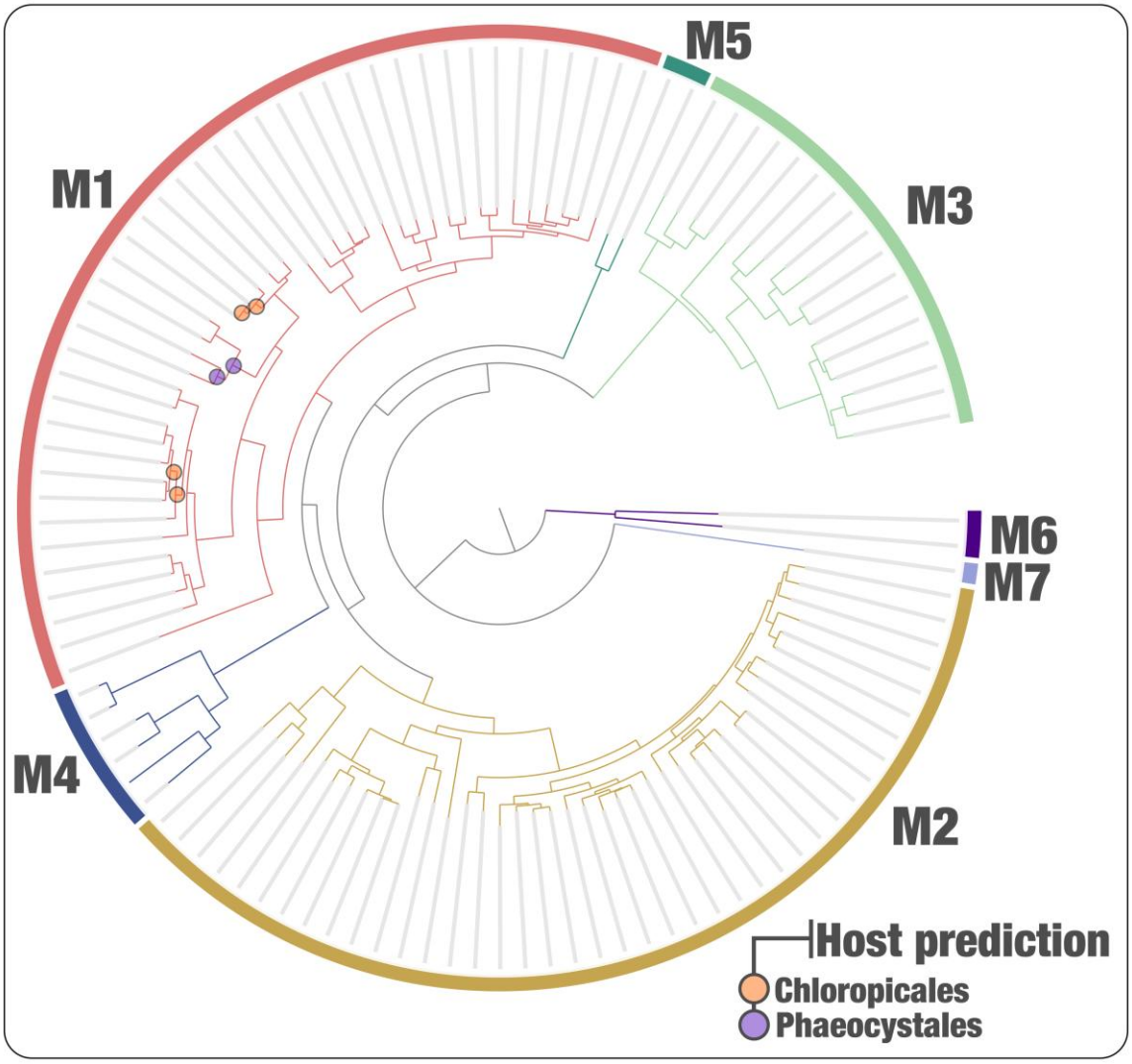


Figure S3: Class-level host predictions for the mirusviruses. The tree corresponds to a phylogeny of Mirus MAGs (informational module, see figure 2), and circles represents clades associated to eukaryotic lineages based on correlation patterns.

Supplemental tables:

Table S1: Description of 937 Tara Oceans metagenomes.

Table S2: DNA-dependent RNA polymerase subunit B genes characterized from the sunlit ocean.

Table S3: Metabins containing DNA-dependent RNA polymerase subunit B genes of interest and used for genome-resolved metagenomics (see supplemental information for more details).

Table S4: Genomic and environmental statistics for the GOEV database.

Table S5: Biogeographic signal for the GOEV database.

Table S6: Occurrence of gene clusters with remote homologies for the GOEV database. Statistics are included.

Table S7: Functional annotations for genes found in the GOEV database. Statistics are included.

References:

1. Koonin, E. v. *et al.* Global Organization and Proposed Megataxonomy of the Virus World. *Microbiology and Molecular Biology Reviews : MMBR* **84**, (2020).
2. Krupovic, M., Dolja, V. v. & Koonin, E. v. The LUCA and its complex virome. *Nature Reviews Microbiology 2020 18:11* **18**, 661–670 (2020).
3. Koonin, E. v., Dolja, V. v. & Krupovic, M. Origins and evolution of viruses of eukaryotes: The ultimate modularity. *Virology* **479–480**, 2–25 (2015).
4. Krupovic, M. & Koonin, E. v. Polintons: a hotbed of eukaryotic virus, transposon and plasmid evolution. *Nat Rev Microbiol* **13**, 105–115 (2015).
5. Guglielmini, J., Woo, A. C., Krupovic, M., Forterre, P. & Gaia, M. Diversification of giant and large eukaryotic dsDNA viruses predated the origin of modern eukaryotes. *Proc Natl Acad Sci U S A* **116**, 19585–19592 (2019).
6. Woo, A. C., Gaia, M., Guglielmini, J., da Cunha, V. & Forterre, P. Phylogeny of the Varidnaviria Morphogenesis Module: Congruence and Incongruence With the Tree of Life and Viral Taxonomy. *Frontiers in Microbiology* **12**, 1708 (2021).
7. Schulz, F. *et al.* Giant virus diversity and host interactions through global metagenomics. *Nature* (2020) doi:10.1038/s41586-020-1957-x.
8. Moniruzzaman, M., Martinez-Gutierrez, C. A., Weinheimer, A. R. & Aylward, F. O. Dynamic genome evolution and complex virocell metabolism of globally-distributed giant viruses. *Nature Communications 2020 11:1* **11**, 1–11 (2020).
9. Endo, H. *et al.* Biogeography of marine giant viruses reveals their interplay with eukaryotes and ecological functions. *Nature Ecology & Evolution 2020 4:12* **4**, 1639–1649 (2020).
10. Gregory, A. C. *et al.* Marine DNA Viral Macro- and Microdiversity from Pole to Pole. *Cell* **177**, 1109-1123.e14 (2019).
11. Al-Shayeb, B. *et al.* Clades of huge phages from across Earth's ecosystems. *Nature 2020 578:7795* **578**, 425–431 (2020).
12. Weinheimer, A. R. & Aylward, F. O. Infection strategy and biogeography distinguish cosmopolitan groups of marine jumbo bacteriophages. *The ISME Journal 2022* 1–11 (2022) doi:10.1038/s41396-022-01214-x.
13. Sunagawa, S. *et al.* Tara Oceans: towards global ocean ecosystems biology. *Nature Reviews Microbiology* 1–18 (2020) doi:10.1038/s41579-020-0364-5.
14. Delmont, T. O. *et al.* Heterotrophic bacterial diazotrophs are more abundant than their cyanobacterial counterparts in metagenomes covering most of the sunlit ocean. *The ISME Journal 2021* 1–10 (2021) doi:10.1038/s41396-021-01135-1.
15. Delmont, T. O. *et al.* Functional repertoire convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean. *Cell Genomics* 100123 (2022) doi:10.1016/J.XGEN.2022.100123.

16. Aylward, F. O., Moniruzzaman, M., Ha, A. D. & Koonin, E. v. A phylogenomic framework for charting the diversity and evolution of giant viruses. *PLOS Biology* **19**, e3001430 (2021).
17. Mihara, T. *et al.* Taxon Richness of "Megaviridae" Exceeds those of Bacteria and Archaea in the Ocean. *Microbes Environ* **33**, 162–171 (2018).
18. Delmont, T. O. Discovery of nondiazotrophic Trichodesmium species abundant and widespread in the open ocean. *Proceedings of the National Academy of Sciences* **118**, e2112355118 (2021).
19. Carradec, Q. *et al.* A global ocean atlas of eukaryotic genes. *Nature Communications 2018 9:1* **9**, 1–13 (2018).
20. Vanni, C. *et al.* Unifying the known and unknown microbial coding sequence space. *Elife* **11**, (2022).
21. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature 2021 596:7873* **596**, 583–589 (2021).
22. Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat Methods* **19**, 679–682 (2022).
23. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
24. Zhang, Y. *et al.* Atomic structure of the human herpesvirus 6B capsid and capsid-associated tegument complexes. *Nat Commun* **10**, (2019).
25. Duda, R. L. & Teschke, C. M. The amazing HK97 fold: versatile results of modest differences. *Curr Opin Virol* **36**, 9–16 (2019).
26. Hua, J. *et al.* Capsids and Genomes of Jumbo-Sized Bacteriophages Reveal the Evolutionary Reach of the HK97 Fold. *mBio* **8**, (2017).
27. Kazlauskas, D., Krupovic, M., Guglielmini, J., Forterre, P. & Venclovas, C. S. Diversity and evolution of B-family DNA polymerases. *Nucleic Acids Research* **48**, 10142 (2020).
28. Paoli, L. *et al.* Biosynthetic potential of the global ocean microbiome. *Nature 2022* 1–8 (2022) doi:10.1038/s41586-022-04862-3.
29. Mihara, T. *et al.* Linking Virus Genomes with Host Taxonomy. *Viruses* **8**, (2016).
30. Miwa, S., Ito, T. & Sano, M. Morphogenesis of koi herpesvirus observed by electron microscopy. *J Fish Dis* **30**, 715–722 (2007).
31. Venkatesh, S. & Workman, J. L. Histone exchange, chromatin structure and the regulation of transcription. *Nature Reviews Molecular Cell Biology 2015 16:3* **16**, 178–189 (2015).
32. Bratanov, D. *et al.* Unique structure and function of viral rhodopsins. *Nature Communications 2019 10:1* **10**, 1–13 (2019).
33. Zabelskii, D. *et al.* Viral rhodopsins 1 are an unique family of light-gated cation channels. *Nature Communications 2020 11:1* **11**, 1–16 (2020).
34. Forterre, P. & Gaïa, M. Giant viruses and the origin of modern eukaryotes. *Current Opinion in Microbiology* Preprint at https://doi.org/10.1016/j.mib.2016.02.001 (2016).
35. Cunha, V. Da *et al.* Giant viruses encode novel types of actins possibly related to the origin of eukaryotic actin: the viractins. *bioRxiv* (2020) doi:10.1101/2020.06.16.150565.

36. Livingstone Bell, P. J. Viral eukaryogenesis: Was the ancestor of the nucleus a complex DNA virus? *Journal of Molecular Evolution* (2001) doi:10.1007/s002390010215.
37. Takemura, M. Medusavirus Ancestor in a Proto-Eukaryotic Cell: Updating the Hypothesis for the Viral Origin of the Nucleus. *Frontiers in Microbiology* **11**, 2169 (2020).
38. Takemura, M. Poxviruses and the origin of the eukaryotic nucleus. *J Mol Evol* (2001) doi:10.1007/s002390010171.
39. Kaneko, H. *et al.* Eukaryotic virus composition can predict the efficiency of carbon export in the global ocean. *iScience* **24**, 102002 (2021).
40. Laber, C. P. *et al.* Coccolithovirus facilitation of carbon export in the North Atlantic. *Nature Microbiology 2018 3:5* **3**, 537–547 (2018).
41. Delmont, T. O. *et al.* Functional repertoire convergence of distantly related eukaryotic plankton lineages revealed by genome-resolved metagenomics. *bioRxiv* 2020.10.15.341214 (2020) doi:10.1101/2020.10.15.341214.
42. Li, D., Liu, C. M., Luo, R., Sadakane, K. & Lam, T. W. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2014).
43. Eren, A. M. *et al.* Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **3**, e1319 (2015).
44. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
45. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
46. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
47. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nature Methods* **11**, 1144–1146 (2014).
48. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput Biol* **7**, e1002195 (2011).
49. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
50. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* **30**, 772–780 (2013).
51. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods 2017 14:6* **14**, 587–589 (2017).
52. Nguyen, L. T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution* **32**, 268–274 (2015).
53. Delmont, T. O. & Eren, A. M. Identifying contamination with advanced visualization and analysis practices: metagenomic approaches for eukaryotic genome assemblies. *PeerJ* **4**, e1839 (2016).

54. Needham, D. M. *et al.* Targeted metagenomic recovery of four divergent viruses reveals shared and distinctive characteristics of giant viruses of marine eukaryotes. *Philosophical Transactions of the Royal Society B* **374**, (2019).
55. Delcher, A. L., Phillippy, A., Carlton, J. & Salzberg, S. L. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res* **30**, 2478–2483 (2002).
56. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403–410 (1990).
57. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Systematic Biology* **59**, 307–321 (2010).
58. Hoang, D. T., Chernomor, O., Von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol Biol Evol* **35**, 518–522 (2018).
59. Gabler, F. *et al.* Protein Sequence Analysis Using the MPI Bioinformatics Toolkit. *Curr Protoc Bioinformatics* **72**, (2020).
60. Steinegger, M. *et al.* HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **20**, (2019).
61. Pettersen, E. F. *et al.* UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci* **30**, 70–82 (2021).
62. Wang, H. C., Minh, B. Q., Susko, E. & Roger, A. J. Modeling Site Heterogeneity with Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation. *Syst Biol* **67**, 216–235 (2018).
63. Delmont, T. O. *et al.* Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. *Nature Microbiology 2018 3:7* **3**, 804–813 (2018).
64. Tackmann, J., Matias Rodrigues, J. F. & von Mering, C. Rapid Inference of Direct Interactions in Large-Scale Ecological Networks from Heterogeneous Microbial Sequencing Data. *Cell Systems* **9**, 286-296.e8 (2019).
65. Mihara, T. *et al.* Linking Virus Genomes with Host Taxonomy. *Viruses* **8**, (2016).
66. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35**, (2007).
67. Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B. & Wu, C. H. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932 (2015).
68. Yutin, N., Wolf, Y. I., Raoult, D. & Koonin, E. V. Eukaryotic large nucleo-cytoplasmic DNA viruses: clusters of orthologous genes and reconstruction of viral genome evolution. *Virol J* **6**, (2009).
69. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**, 59–60 (2015).
70. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research* **47**, D309–D314 (2019).

71. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**, 955–964 (1997).
72. Al-Shayeb, B. *et al.* Clades of huge phages from across Earth's ecosystems. *Nature 2020 578:7795* **578**, 425–431 (2020).
73. Hauser, M., Steinegger, M. & Söding, J. MMseqs software suite for fast and deep clustering and searching of large protein sequence sets. *Bioinformatics* **32**, 1323–1330 (2016).
74. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research* **39**, W29–W37 (2011).

1. Koonin, E. v. *et al.* Global Organization and Proposed Megataxonomy of the Virus World. *Microbiology and Molecular Biology Reviews : MMBR* **84**, (2020).
2. Krupovic, M., Dolja, V. v. & Koonin, E. v. The LUCA and its complex virome. *Nature Reviews Microbiology 2020 18:11* **18**, 661–670 (2020).
3. Koonin, E. v., Dolja, V. v. & Krupovic, M. Origins and evolution of viruses of eukaryotes: The ultimate modularity. *Virology* **479–480**, 2–25 (2015).
4. Krupovic, M. & Koonin, E. v. Polintons: a hotbed of eukaryotic virus, transposon and plasmid evolution. *Nat Rev Microbiol* **13**, 105–115 (2015).
5. Guglielmini, J., Woo, A. C., Krupovic, M., Forterre, P. & Gaia, M. Diversification of giant and large eukaryotic dsDNA viruses predated the origin of modern eukaryotes. *Proc Natl Acad Sci U S A* **116**, 19585–19592 (2019).
6. Woo, A. C., Gaia, M., Guglielmini, J., da Cunha, V. & Forterre, P. Phylogeny of the Varidnaviria Morphogenesis Module: Congruence and Incongruence With the Tree of Life and Viral Taxonomy. *Frontiers in Microbiology* **12**, 1708 (2021).
7. Schulz, F. *et al.* Giant virus diversity and host interactions through global metagenomics. *Nature* (2020) doi:10.1038/s41586-020-1957-x.
8. Moniruzzaman, M., Martinez-Gutierrez, C. A., Weinheimer, A. R. & Aylward, F. O. Dynamic genome evolution and complex virocell metabolism of globally-distributed giant viruses. *Nature Communications 2020 11:1* **11**, 1–11 (2020).
9. Endo, H. *et al.* Biogeography of marine giant viruses reveals their interplay with eukaryotes and ecological functions. *Nature Ecology & Evolution 2020 4:12* **4**, 1639–1649 (2020).
10. Gregory, A. C. *et al.* Marine DNA Viral Macro- and Microdiversity from Pole to Pole. *Cell* **177**, 1109-1123.e14 (2019).
11. Al-Shayeb, B. *et al.* Clades of huge phages from across Earth's ecosystems. *Nature 2020 578:7795* **578**, 425–431 (2020).
12. Weinheimer, A. R. & Aylward, F. O. Infection strategy and biogeography distinguish cosmopolitan groups of marine jumbo bacteriophages. *The ISME Journal 2022* 1–11 (2022) doi:10.1038/s41396-022-01214-x.
13. Sunagawa, S. *et al.* Tara Oceans: towards global ocean ecosystems biology. *Nature Reviews Microbiology* 1–18 (2020) doi:10.1038/s41579-020-0364-5.
14. Delmont, T. O. *et al.* Heterotrophic bacterial diazotrophs are more abundant than their cyanobacterial counterparts in metagenomes covering most of the sunlit ocean. *The ISME Journal 2021* 1–10 (2021) doi:10.1038/s41396-021-01135-1.

15. Delmont, T. O. *et al.* Functional repertoire convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean. *Cell Genomics* 100123 (2022) doi:10.1016/J.XGEN.2022.100123.
16. Aylward, F. O., Moniruzzaman, M., Ha, A. D. & Koonin, E. v. A phylogenomic framework for charting the diversity and evolution of giant viruses. *PLOS Biology* **19**, e3001430 (2021).
17. Mihara, T. *et al.* Taxon Richness of "Megaviridae" Exceeds those of Bacteria and Archaea in the Ocean. *Microbes Environ* **33**, 162–171 (2018).
18. Delmont, T. O. Discovery of nondiazotrophic Trichodesmium species abundant and widespread in the open ocean. *Proceedings of the National Academy of Sciences* **118**, e2112355118 (2021).
19. Carradec, Q. *et al.* A global ocean atlas of eukaryotic genes. *Nature Communications 2018 9:1* **9**, 1–13 (2018).
20. Vanni, C. *et al.* Unifying the known and unknown microbial coding sequence space. *Elife* **11**, (2022).
21. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature 2021 596:7873* **596**, 583–589 (2021).
22. Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat Methods* **19**, 679–682 (2022).
23. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
24. Zhang, Y. *et al.* Atomic structure of the human herpesvirus 6B capsid and capsid-associated tegument complexes. *Nat Commun* **10**, (2019).
25. Duda, R. L. & Teschke, C. M. The amazing HK97 fold: versatile results of modest differences. *Curr Opin Virol* **36**, 9–16 (2019).
26. Hua, J. *et al.* Capsids and Genomes of Jumbo-Sized Bacteriophages Reveal the Evolutionary Reach of the HK97 Fold. *mBio* **8**, (2017).
27. Kazlauskas, D., Krupovic, M., Guglielmini, J., Forterre, P. & Venclovas, C. S. Diversity and evolution of B-family DNA polymerases. *Nucleic Acids Research* **48**, 10142 (2020).
28. Paoli, L. *et al.* Biosynthetic potential of the global ocean microbiome. *Nature 2022* 1–8 (2022) doi:10.1038/s41586-022-04862-3.
29. Mihara, T. *et al.* Linking Virus Genomes with Host Taxonomy. *Viruses* **8**, (2016).
30. Miwa, S., Ito, T. & Sano, M. Morphogenesis of koi herpesvirus observed by electron microscopy. *J Fish Dis* **30**, 715–722 (2007).
31. Venkatesh, S. & Workman, J. L. Histone exchange, chromatin structure and the regulation of transcription. *Nature Reviews Molecular Cell Biology 2015 16:3* **16**, 178–189 (2015).
32. Bratanov, D. *et al.* Unique structure and function of viral rhodopsins. *Nature Communications 2019 10:1* **10**, 1–13 (2019).
33. Zabelskii, D. *et al.* Viral rhodopsins 1 are an unique family of light-gated cation channels. *Nature Communications 2020 11:1* **11**, 1–16 (2020).
34. Forterre, P. & Gaïa, M. Giant viruses and the origin of modern eukaryotes. *Current Opinion in Microbiology* Preprint at https://doi.org/10.1016/j.mib.2016.02.001 (2016).

35. Cunha, V. Da *et al.* Giant viruses encode novel types of actins possibly related to the origin of eukaryotic actin: the viractins. *bioRxiv* (2020) doi:10.1101/2020.06.16.150565.
36. Livingstone Bell, P. J. Viral eukaryogenesis: Was the ancestor of the nucleus a complex DNA virus? *Journal of Molecular Evolution* (2001) doi:10.1007/s002390010215.
37. Takemura, M. Medusavirus Ancestor in a Proto-Eukaryotic Cell: Updating the Hypothesis for the Viral Origin of the Nucleus. *Frontiers in Microbiology* **11**, 2169 (2020).
38. Takemura, M. Poxviruses and the origin of the eukaryotic nucleus. *J Mol Evol* (2001) doi:10.1007/s002390010171.
39. Kaneko, H. *et al.* Eukaryotic virus composition can predict the efficiency of carbon export in the global ocean. *iScience* **24**, 102002 (2021).
40. Laber, C. P. *et al.* Coccolithovirus facilitation of carbon export in the North Atlantic. *Nature Microbiology 2018 3:5* **3**, 537–547 (2018).
41. Delmont, T. O. *et al.* Functional repertoire convergence of distantly related eukaryotic plankton lineages revealed by genome-resolved metagenomics. *bioRxiv* 2020.10.15.341214 (2020) doi:10.1101/2020.10.15.341214.
42. Li, D., Liu, C. M., Luo, R., Sadakane, K. & Lam, T. W. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2014).
43. Eren, A. M. *et al.* Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **3**, e1319 (2015).
44. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
45. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
46. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
47. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nature Methods* **11**, 1144–1146 (2014).
48. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput Biol* **7**, e1002195 (2011).
49. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
50. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* **30**, 772–780 (2013).
51. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods 2017 14:6* **14**, 587–589 (2017).
52. Nguyen, L. T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution* **32**, 268–274 (2015).

53. Delmont, T. O. & Eren, A. M. Identifying contamination with advanced visualization and analysis practices: metagenomic approaches for eukaryotic genome assemblies. *PeerJ* **4**, e1839 (2016).
54. Needham, D. M. *et al.* Targeted metagenomic recovery of four divergent viruses reveals shared and distinctive characteristics of giant viruses of marine eukaryotes. *Philosophical Transactions of the Royal Society B* **374**, (2019).
55. Delcher, A. L., Phillippy, A., Carlton, J. & Salzberg, S. L. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res* **30**, 2478–2483 (2002).
56. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403–410 (1990).
57. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Systematic Biology* **59**, 307–321 (2010).
58. Hoang, D. T., Chernomor, O., Von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol Biol Evol* **35**, 518–522 (2018).
59. Gabler, F. *et al.* Protein Sequence Analysis Using the MPI Bioinformatics Toolkit. *Curr Protoc Bioinformatics* **72**, (2020).
60. Steinegger, M. *et al.* HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **20**, (2019).
61. Pettersen, E. F. *et al.* UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci* **30**, 70–82 (2021).
62. Wang, H. C., Minh, B. Q., Susko, E. & Roger, A. J. Modeling Site Heterogeneity with Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation. *Syst Biol* **67**, 216–235 (2018).
63. Delmont, T. O. *et al.* Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. *Nature Microbiology 2018 3:7* **3**, 804–813 (2018).
64. Tackmann, J., Matias Rodrigues, J. F. & von Mering, C. Rapid Inference of Direct Interactions in Large-Scale Ecological Networks from Heterogeneous Microbial Sequencing Data. *Cell Systems* **9**, 286-296.e8 (2019).
65. Mihara, T. *et al.* Linking Virus Genomes with Host Taxonomy. *Viruses* **8**, (2016).
66. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35**, (2007).
67. Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B. & Wu, C. H. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932 (2015).
68. Yutin, N., Wolf, Y. I., Raoult, D. & Koonin, E. V. Eukaryotic large nucleo-cytoplasmic DNA viruses: clusters of orthologous genes and reconstruction of viral genome evolution. *Virol J* **6**, (2009).
69. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**, 59–60 (2015).

70. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research* **47**, D309–D314 (2019).
71. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**, 955–964 (1997).
72. Al-Shayeb, B. *et al.* Clades of huge phages from across Earth's ecosystems. *Nature 2020 578:7795* **578**, 425–431 (2020).
73. Hauser, M., Steinegger, M. & Söding, J. MMseqs software suite for fast and deep clustering and searching of large protein sequence sets. *Bioinformatics* **32**, 1323–1330 (2016).
74. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research* **39**, W29–W37 (2011).