Discovery of a class of giant virus relatives displaying unusual functional traits and prevalent within plankton: the *Mirusviricetes*

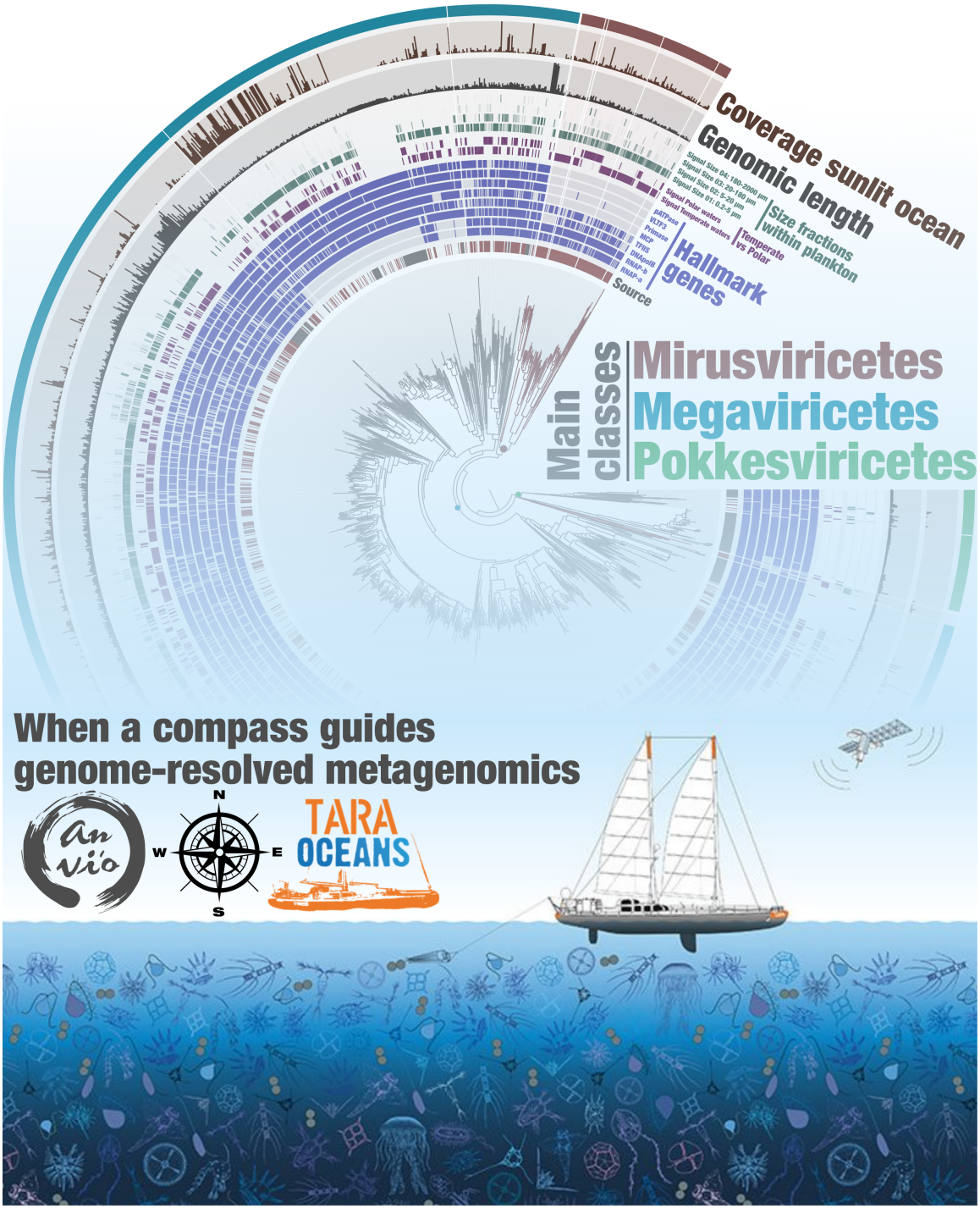
Morgan Gaïa[1,2#], Lingjie Meng[3#], Eric Pelletier[1,2], Patrick Forterre[4,5], Chiara Vanni[6], Antonio Fernendez-Guerra[7], Olivier Jaillon[1,2], Patrick Wincker[1,2], Hiroyuki Ogata[3] and Tom O. Delmont*[1,2]

[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
Co-first authors
* Corresponding author: tomodelmont@gmail.com

Abstract: Large and giant DNA viruses of the phylum *Nucleocytoviricota* have a profound influence on the ecology and evolution of planktonic eukaryotes. Recently, various *Nucleocytoviricota* genomes have been characterized from environmental metagenomes based on the occurrence of hallmark genes identified from cultures. However, lineages diverging from the culture genomics functional principles have been overlooked thus far. Here, we developed a phylogeny-guided genome-resolved metagenomic framework using a single hallmark gene as compass, a subunit of DNA-dependent RNA polymerase encoded by most *Nucleocytoviricota*. We applied this method to large metagenomic data sets from the surface of five oceans and two seas and characterized 697 non-redundant *Nucleocytoviricota* genomes up to 1.45 Mbp in length. This database expands the known diversity of the class *Megaviricetes* and revealed two additional putative classes we named *Proculviricetes* and *Mirusviricetes*. Critically, the diverse and prevalent *Mirusviricetes* population genomes seemingly lack several hallmark genes, in particular those related to viral particle morphogenesis. Instead, they share various genes of known (e.g., TATA-binding proteins, histones, proteases and viral rhodopsins) and unknown functions rarely detected if not entirely missing in other characterized *Nucleocytoviricota* classes. Phylogenomics, comparative genomics, functional trends and the signal among planktonic cellular size fractions point to *Mirusviricetes* being a major, functionally divergent class of large DNA viruses that actively infect eukaryotes in the sunlit ocean using an enigmatic functional life style. Finally, we built a comprehensive marine genomic database for *Nucleocytoviricota* by combining multiple environmental surveys that might contribute to future endeavors exploring the ecology and evolution of plankton.

Key words: *Nucleocytoviricota*, NCLDVs, *Tara* Oceans, population genomics, compass binning, biodiversity, ecology, evolution, plankton.

Cover: This survey took advantage of the environmental DNA sequencing and bioinformatics legacies of ***Tara* Oceans** and **anvi'o** to discover new classes of **giant virus relatives** in the sunlit ocean by means of "**Phylogeny-guided Genome-Resolved Metagenomics**".



Planktonic illustration was designed by Noan Le Bescot (Ternog design)

Introduction:

The Nucleocytoplasmic Large DNA virus (NCLDV) assemblage groups related families of large and giant double-stranded DNA viruses infecting eukaryotes[1]. They now correspond to the phylum *Nucleocytoviricota* within the *Varidnaviria* realm[2]. Some of them can build large viral particles and maintain genomes that can reach multiple megabases in length[3–5]. The large diversity of marine NCLDVs suggests they play a critical role in regulating the occurrence and blooming activity of planktonic microbial eukaryotes in the sunlit ocean[6–8], as already known for a few cultured NCLDVs[9–11], likely through complex ecological interactions that have yet to be fully comprehended. NCLDVs also impact the evolution of their hosts via lateral gene transfers, since for instance, up to 10% of genes were identified as originating from NCLDVs among the green algae[12]. They could have also been driving the evolution of eukaryotes (including the emergence of complex proto-eukaryotic traits) through a long-standing coevolution[13–15].

Genomic characterizations using co-cultures with the eukaryotic hosts exposed a restricted set of hallmark genes occurring in most NCLDV lineages[14,16,17]. Those genes represent key functional principles of NCLDVs related to DNA and RNA processing as well as viral particle morphogenesis. The relatively congruent phylogenetic signal of these genes provided the means to delineate evolutionary boundaries of the phylum *Nucleocytoviricota*[14,17]. More recently, genome-resolved metagenomics was applied to target NCLDVs in various environments using the occurrence of these hallmark genes for guidance[6,7]. These culture-independent surveys were most successful in marine systems and led to a considerable expansion of the genomic landscape of NCLDVs. The phylum *Nucleocytoviricota* currently includes two formally described classes (*Pokkesviricetes* and the more diverse *Megaviricetes*) that encompass at least six orders and 32 families that often lack culture representatives[2,18].

Previous environmental surveys, however, only considered *Nucleocytoviricota* population genomes containing a majority of the hallmark genes (an indicator for high genomic completion), and by doing so overlooked putative NCLDV lineages substantially diverging from the functional principles defined from the legacy of culture genomics. Here, we analyzed large metagenomic data sets from the sunlit ocean using an original phylogeny-guided genome-resolved metagenomic framework that employs a single hallmark gene as compass to survey abundant marine NCLDV lineages. We characterized and manually curated hundreds of NCLDV population genomes up to 1.45 Mbp in length. This database further expands the known diversity of the class *Megaviricetes* and exposes two additional classes we named *Proculviricetes* (6 population genomes only detected in high latitudes) and *Mirusviricetes* (111 population genomes that occur globally in the sunlit ocean). Critically, *Mirusviricetes* population genomes lack some hallmark genes related to DNA processing and, more surprisingly, viral particle morphogenesis. Instead, they contain various core genes of known and unknown function missing or rarely

detected in the other NCLDV lineages. Our investigations point to *Mirusviricetes* being a functionally divergent class of NCLDVs that actively infect eukaryotes in the sunlit ocean using an enigmatic lifestyle.

Results:

Diversity of DNA–dependent RNA polymerase B genes in the sunlit ocean

We performed a comprehensive search for the DNA–dependent RNA polymerase B subunit (RNApolB) genes from the euphotic zone of polar, temperate, and tropical oceans using 798 metagenomes derived from the *Tara* Oceans expeditions[19]. 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 already used as inputs for 11 metagenomic co-assemblies (~12 million contigs longer than 2,500 nucleotides were produced) using geographically bounded samples to characterize archaeal, bacterial and eukaryotic metagenome-assembled genomes (MAGs)[20,21]. Here, we recovered 7,591 RNApolB genes from these 11 co-assemblies using a dedicated Hidden Markov Model (HMM) and built a non-redundant protein database of 2,728 RNApolB sequences >800 amino acids in length with similarity <90% (Table S2). We included RNApolB sequences from 262 reference genomes of known archaea, bacteria, eukaryotes and *Nucleocytoviricota*[14] for perspective to guide global taxonomic affiliations. The phylogenetic analysis of sequences (Figure S1) recapitulated previously observed trends, such as the monophyly of the Bacteria and the three eukaryotic nuclear RNApolB clades[14], or else the considerable diversity of *Imitervirales* order (class *Megaviricetes*) already revealed by means of gene markers[22] and genome-resolved metagenomics[6,7]. The phylogenetic tree also included several deep-branching lineages with no representatives among the references, which we dubbed "RNApolB new lineages". With a positioning clearly disconnected from the three domains of life, we hypothesized these RNApolB genes may correspond to previously unknown *Nucleocytoviricota* lineages. Centered on this RNApolB compass, we performed a phylogeny-guided genome-resolved metagenomic survey and explored the genomic landscape of the new lineages.

A phylogeny-guided genome-resolved metagenomic survey of NCLDVs

The ~12 million contigs were previously organized into a constrained set of 2,550 metagenomic blocks[20,21] based on their sequence composition and differential coverage across metagenomes (CONCOCT algorithm[23]). By design, most blocks contain multiple unrelated genomic contents. Contigs containing RNApolB genes corresponding to the new lineages were located in 79 blocks (Table S3). We then used HMMs for eight hallmark genes and 149 additional orthologous groups often found in reference NCLDVs[14] to guide the characterization of MAGs from those blocks from within the anvi'o interactive interface[24,25]. Importantly, MAGs with a

limited number of hallmark genes were considered as long as they contained the focal RNApolB gene. MAGs containing multiple hallmark genes but not the RNApolB gene were also considered. We found that many of the blocks contained multiple NCLDV MAGs. Finally, we also characterized NCLDV MAGs from 57 additional blocks that contained multiple RNApolB genes corresponding to known NCLDV lineages mostly affiliated to the class *Megaviricetes* (Tables S2 and S3). In total, we recovered 697 non-redundant (average nucleotide identify <98%) NCLDV MAGs ranging from 51 kbp to 1.45 Mbp in length (average of 254 kbp). Finally, we manually curated each NCLDV MAG (visual inspection in the context of genomic and environmental information), as previously done for the three domains of life within plankton[20,21,26].

In line with previous NCLDV environmental genomic surveys, MAGs from the class *Megaviricetes* contained a majority of the eight hallmark genes we considered here (Table S4). In contrast, most MAGs corresponding to the RNApolB new lineages were systematically missing four of these markers: the major capsid protein (MCP), the D5-like primase-helicase (primase), the Poxvirus Late Transcription Factor 3 (VLTF3), and the packaging ATPase (pATPase) (Table S4). Remaining hallmark genes (the two DNA-dependent RNA polymerase largest subunits, RNApolA and RNApolB; the family B DNA polymerase, DNApolB; and the Transcription Factor II-S, TFIIS) were commonly found in those MAGs, however, the lack of signal for half the hallmark genes defined from culture genomics excluded them from the previous surveys. This demonstrates the utility of single gene-based, phylogeny-guided genome-resolved metagenomics, which in our survey pointed to MAGs containing intriguing RNApolB genes. We subsequently created a comprehensive marine NCLDV genomic resource that includes the newly identified MAGs in order to assess the evolutionary relationship between the RNApolB new lineages and phylum *Nucleocytoviricota*.

Discovery of two NCLDV classes: the *Mirusviricetes* and *Proculviricetes*

We incorporated environmental genomes characterized from previous metagenomic surveys and created a genomic resource containing 1,593 non-redundant (average nucleotide identify <98%) marine NCLDV MAGs, along with 224 reference NCLDV genomes from culture and cell sorting (Table S4). This genomic database built around marine NCLDVs has a total volume of 580 Mbp and contains ~0.6 million genes. It includes 236 MAGs from Moniruzzaman et al.[7], 659 MAGs from Schulz et al.[6] and the 697 manually curated MAGs from our survey. We determined the biogeography of these NCLDVs in the sunlit ocean by mapping ~300 billion metagenomic reads from 939 *Tara* Oceans filters encompassing nine planktonic size fractions ranging from 0.2 µm to 2000 µm (Table S1). 85.4% of the marine MAGs and 8.5% of the reference genomes (many of which are not marine) were detected in at least one of the *Tara* Oceans metagenomes (Table S5).

A Genomic statistics

Clade	Number of MAGs	Av. length (kbp)	Av. number of contigs	Av. N50 (kbp)	Av. GC-content	Av. number of genes	Av. gene length (aa)	Av. number of KEGG functions	Av. number of gene clusters	Total coverage (939 metagenomes)	Oceanic regions: Polar: Arctic Ocean	Polar: Southern Ocean	Temperate: Indian Ocean	Temperate: Atlantic Ocean	Temperate: Pacific Ocean	Temperate: Med. Sea	Temperate: Red Sea	Size fractions: Size 01 0.2-5 µm	Size 02 3-20 µm	Size 03 20-180 µm	Size 04 180-2000 µm
M1	42	169	18	24	43%	183	282	31	103	2,332x	7%	5%	11%	11%	42%	22%	1%	80%	13%	4%	3%
M2	40	253	29	16	41%	264	292	43	160	2,484x	41%	10%	5%	9%	13%	19%	3%	72%	19%	3%	6%
M3	17	174	17	20	40%	172	306	16	95	1,106x	0.3%	0%	6%	29%	40%	18%	6%	83%	15%	1%	0.4%
M4	10	149	14	24	53%	137	320	13	39	421x	30%	0%	10%	3%	30%	27%	0%	71%	15%	3%	10%
M5	2	158	18	12	44%	159	282	24	44	134x	100%	0%	0%	0%	0%	0%	0%	80%	20%	0.5%	0%
Proculviri.	6	147	26	7	47%	149	283	21	41	222x	83%	17%	0%	0%	0%	0%	0%	10%	76%	11%	3%

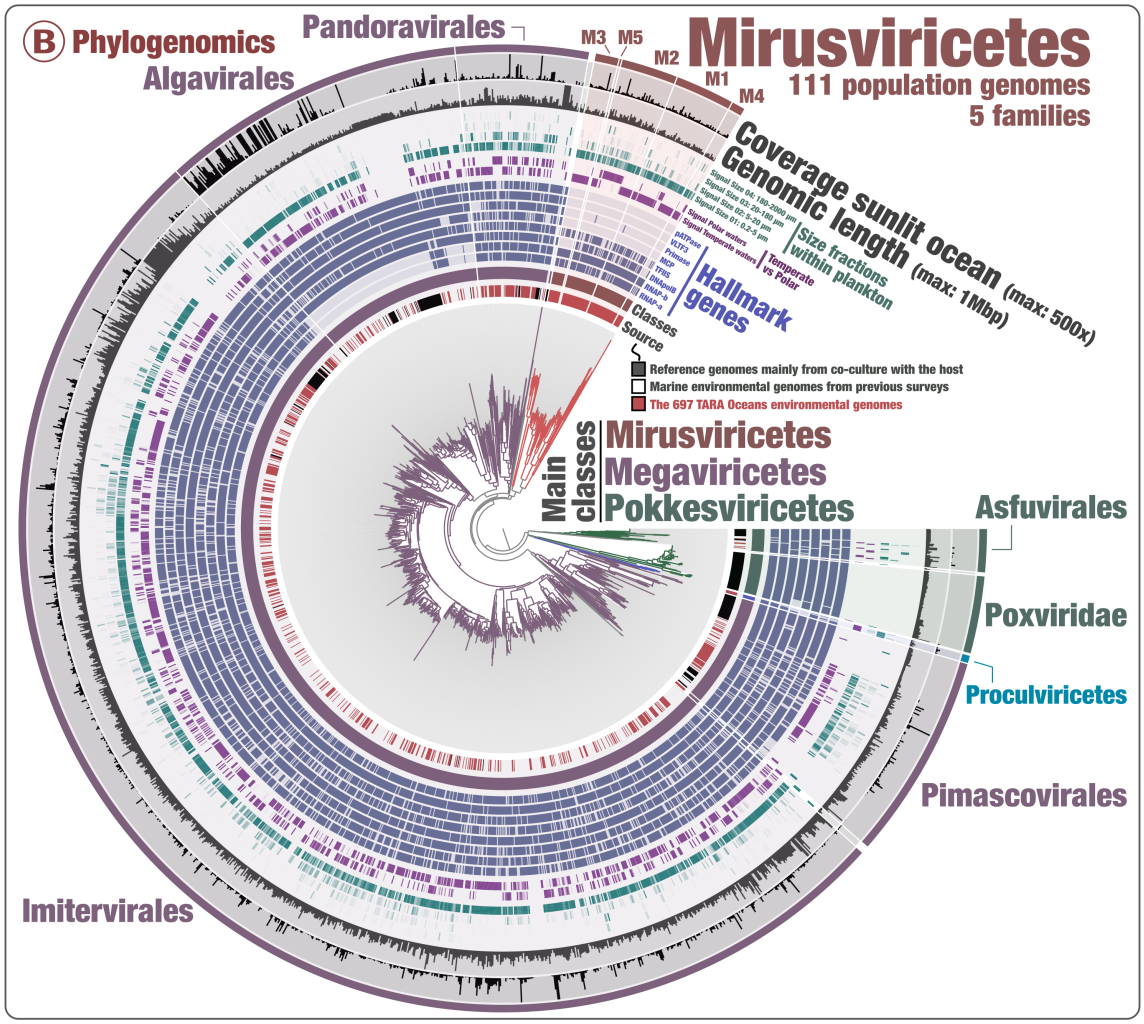


Figure 1: Evolutionary history of abundant and widespread marine viruses within the phylum *Nucleocytoviricota*. Top panel displays genomic and environmental statistics for the five *Mirusviricetes* families as well as *Proculvicetes*. Bottom panel displays a maximum-likelihood phylogenetic tree built from the NCLDV marine genomic database (1,756 genomes) based on a concatenation of four hallmark genes (3,712 amino acid positions) using the PMSF mixture model and rooted between the *Pokkesviricetes* and other *Nucleocytoviricota*. The tree was decorated with layers of complementary information and visualized with anvi'o.

We identified and manually curated the RNApolA, RNApolB, DNApolB and TFIIS sequences from the marine NCLDV genomic database. Hallmark genes found multiple times in a given genome were scrutinized in the context of phylogenetic

signal for each hallmark gene (see methods). On one side, we identified putative contaminants (mostly in MAGs from automatic binning) that were removed for downstream analyses. On the other side, we identified paralogs (i.e., duplication events) for all four hallmark genes that in the most striking case of RNApolB included a majority of clades within the diverse *Imitervirales* order[22,27] (Figure S2, Table S4). This could explain why the RNApolB gene performs relatively poorly as a phylogenetic marker for the new diversity of *Nucleocytoviricota* when duplicates are not considered[18]. The paralog sequences were systematically branching as sister clades, suggesting the duplication event took place before the diversification of the related taxonomic clades. We also identified the putative acquisition of a distant TFIIS in a small subclade of *Algavirales* leading to this hallmark gene being in multi-copy in the corresponding MAGs. From this manual curation effort, it became evident that cases of multi-copy hallmark genes (contaminants, paralogs and gene acquisitions) need to be properly resolved prior to studying the evolutionary history of the phylum *Nucleocytoviricota*.

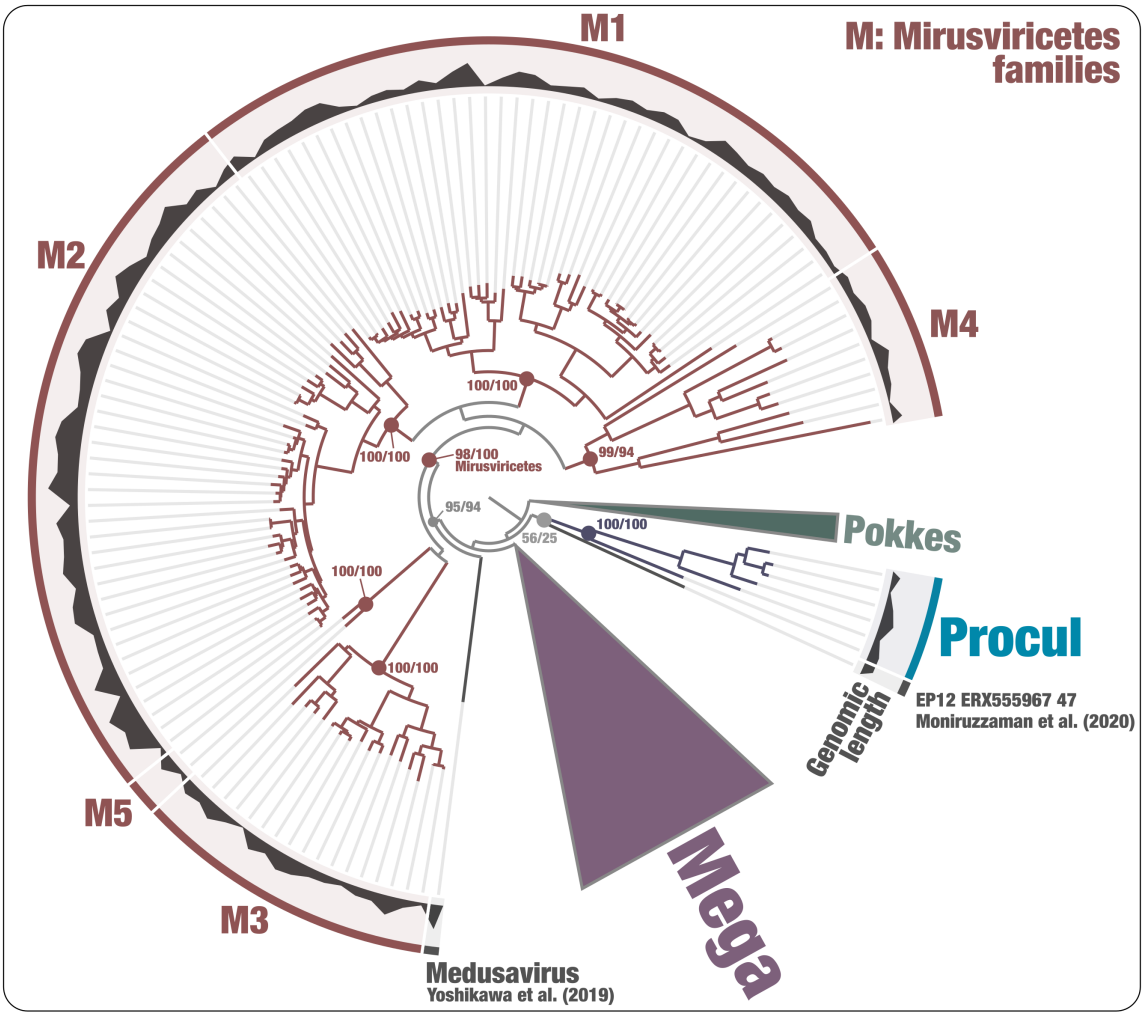


Figure 2: Evolutionary history of *Proculviricetes* and the five families of *Mirusviricetes*. The maximum-likelihood phylogenetic tree (same as in figure 1) was built from the NCLDV marine genomic database (1,756 genomes) based on a concatenation of four hallmark genes (3,712 amino

acid positions) using the PMSF mixture model and rooted between the *Pokkesviricetes* and other *Nucleocytoviricota*. 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). The *Megaviricetes* and *Pokkesviricetes* were collapsed to better visualize patterns for the two new putative NCLDV classes.

We then performed a phylogenetic analysis using a concatenation of the four curated, duplicate-free hallmark genes by selecting the paralog clade with the shortest branch for each duplication event. As a first insight, all the MAGs characterized from our survey appear to fall within the scope of *Nucleocytoviricota* (Figure 1). A majority of MAGs further expanded the recently revealed genomic diversity of *Megaviricetes*. For instance, we reconstructed MAGs related to previously described subclades of *Imitervirales*[6,7], but we also identified a relatively large clade within the *Pimascovirales*, potentially sister group to *Iridoviridae*, as well as a few MAGs basal to this order (see the "Source layer" in Figure 1). But more interestingly, the RNApolB new lineages formed two highly supported monophyletic deep-branching clades that are distinct from both *Megaviricetes* and *Pokkesviricetes* (support values in figure 2). We propose naming these classes *Proculviricetes* (6 MAGs only detected in high latitudes) and *Mirusviricetes* (111 MAGs). Only *Mirusviricetes* is depleted of HMM hits for the MCP, primase, VLTF3, and pATPase. This class is organized into five distinct clades (putative families) we labeled M1, M2, M3 (also lacking the DNApol B), M4 and M5 (Figures 1 and 2, Table S4). The *Mirusviricetes* MAGs have a length ranging from 53 Kbp to 438 Kbp (average of 198 Kbp) and were characterized from all the oceanic regions (Table S5), indicating that this class is prevalent within plankton in the sunlit ocean.

The lack of HMM hits for four hallmark genes within *Mirusviricetes* represents an important enigma given our current understanding of the functional basis of NCLDVs. Especially, lack of signal for both the MCP (building block of the viral particle[28]) and pATPase (motor that pumps DNA into the viral particle[29]) raises questions regarding the infection cycle of *Mirusviricetes*. One possibility is that *Mirusviricetes* are not *bona fide* viruses but large plasmids such as those previously described in some Bacteria and Archaea (e.g., Streptomyces contain giant linear plasmids[30]). Alternatively, these viruses might encode novel families of capsid proteins, as previously observed in *Pandoraviridae*[2,31], and novel type of packaging ATPases. This second possibility appears more likely since biogeographic signal across the planktonic cellular size fractions clearly suggests that *Mirusviricetes* occur as both free-living and intracellular genomes (or attached viral particles), similar to what is observed for other marine NCLDVs that regularly infect eukaryotic plankton (Figure 1, panel B). For instance, MAGs from family M4 were substantially detected in the 3-20 µm and 180-2000 µm size fractions in addition to the viral particle size fraction (0.2-3 µm), suggesting that the corresponding viruses infected eukaryotes with a broad cellular size range during the *Tara* Oceans expeditions. In addition, the *Mirusviricetes* MAGs were relatively abundant, with a cumulative coverage totalizing 6,479x (i.e., the 111 *Mirusviricetes* MAGs were cumulatively sequenced more than 6,000 times among the metagenomes considered), suggesting they infect relatively abundant marine eukaryotes in both temperate and polar waters using a functional

lifestyle that substantially diverges compared to *Megaviricetes* and *Pokkesviricetes* (Figure 1). We subsequently investigated the gene content of the marine NCLDV genomic database in both the known and unknown coding sequence space to better understand some of the core functional properties of *Mirusviricetes* as compared to *Megaviricetes*, *Pokkesviricetes* and *Proculviricetes*.

Mirusviricetes is a functionally divergent class of NCLDVs

We used AGNOSTOS[32] to characterize 29,414 non-singleton gene cluster communities sharing remote homologies (thereafter called gene clusters) from the marine NCLDV genomic database (Table S6). Clustering of NCLDVs based on the occurrence of these gene clusters emphasized the complex functional history of the *Megaviricetes* and *Pokkesviricetes* classes, with some clades (e.g., the *Imitervirales* and *Algavirales*) split into multiple groups (Figure 3). In contrast, *Mirusviricetes* MAGs branched together, with the five families being organized into distinct groups that largely recapitulated phylogenomic signals (e.g., M1 and M2 are sister clades). Overall, the occurrence of gene clusters agnostic of any functional inferences indicates that *Mirusviricetes* MAGs share important genomic traits distinct from those occurring in other NCLDV lineages. This comparative genomic analysis strongly supports the monophyly of *Mirusviricetes* as observed from the phylogenomic analysis of the four conserved hallmark genes.

We then focused our attention on the most widespread gene clusters within the marine NCLDV genomic database. 35 gene clusters occurred in more than 50% of the 111 *Mirusviricetes* MAGs, representing core genes for this newly identified class (Table S6). Critically, most of them appear unique to this class and 24 lack any functional annotation (Figure 3, panel A). Among the 11 *Mirusviricetes* core genes with a functional annotation, five were detected in only <10% of the other NCLDVs. They correspond to trypsin and peptidase M16 (proteases), viral rhodopsin (ion-conducting pathway), TATA-binding protein (transcription), and a histone (DNA stability) (Figure 3, panel B). In contrast, 20 out of the 28 gene clusters occurring in more than 50% of the other NCLDVs within the database have a functional annotation. They exposed additional NCLDV core gene clusters highly depleted if not entirely missing within the *Mirusviricetes*: RNApol5, RNA pol10, ribonuclease 3 and the SWIB protein (transcription), ResIII helicase and D5 NTPase (DNA replication), and dNK (kinase activity) (Figure 3, panel B). As a possible limitation, genes sharing similar functional properties might be placed in different clusters due to extensive sequence divergences (further investigations are underway). Yet, AGNOSTOS results were often coherent with the HMM hits (e.g., see VLTF3 and pATPase in the Figure 2, panel B), suggesting that gene clusters with remote homologies properly encapsulated the diversity of at least some of the NCLDV core functions.

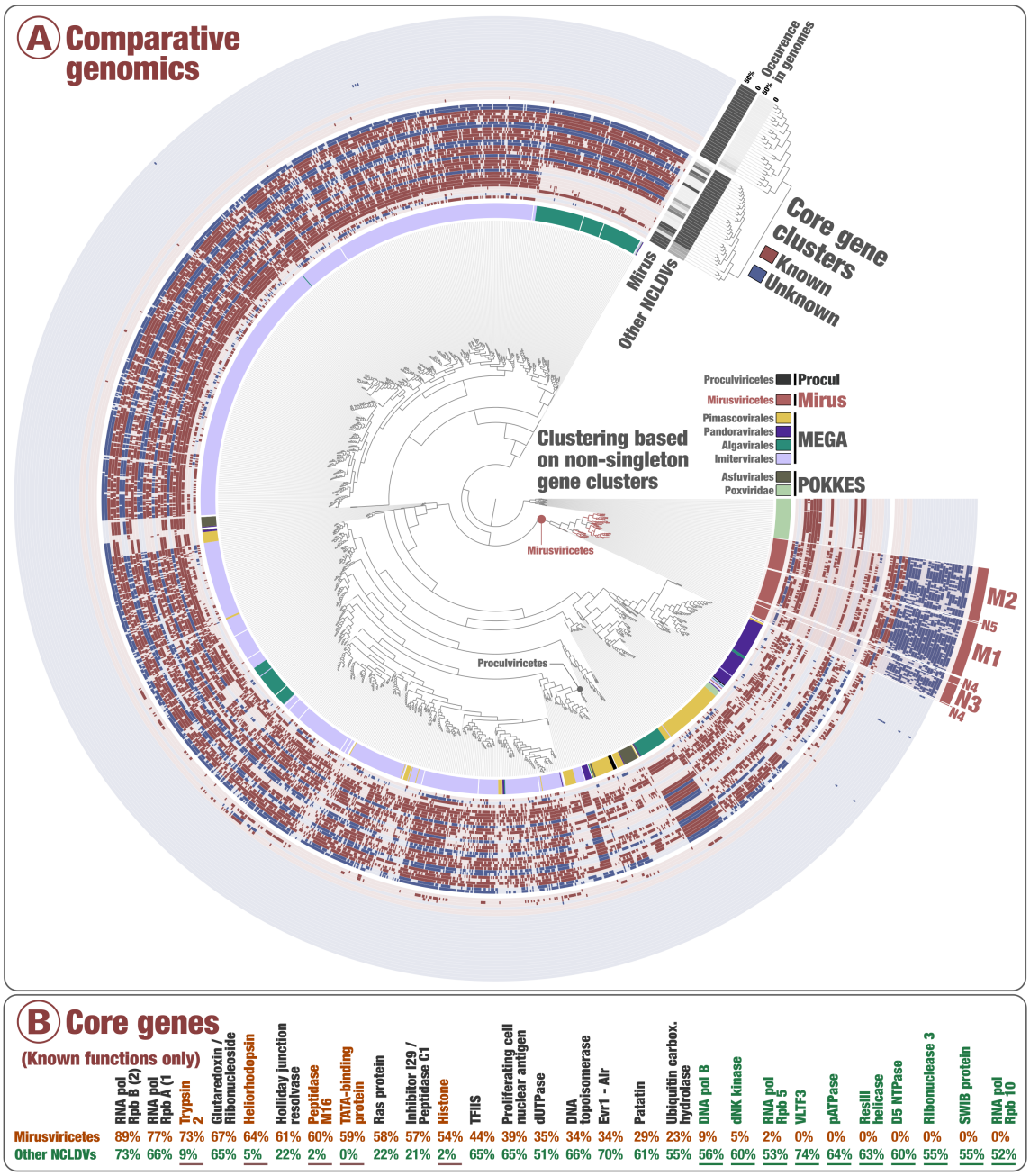


Figure 3: Functional clustering of abundant and widespread marine viruses within the phylum *Nucleocytoviricota*. In panel A, the inner tree is a clustering of NCLDV genomes >100kbp in length based on the occurrence of all the non-singleton gene clusters (Euclidean distance), rooted with the *Chordopoxvirinae*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

Besides core genes, *Mirusviricete* MAGs showed significant enrichment in several functions compared with NCLDVs in other lineages (Table S7). Those functions notably include to genesis and trafficking of membrane vesicles, such as charged

multivesicular body protein 4A/B and 5, synaptobrevin homologs, Arf/Sar family proteins, RAB family proteins, and TBC1 domain family member 6 (exocytosis and actin organization)[33–38]. In addition, cell adhesion related proteins, such as stabilin-2 and fibulin 1/2[38,39], were also enriched in *Mirusviricete*. Together, this group of viruses may be capable of actively reprogramming intracellular membrane trafficking. The phage-induced integrase XerD was also detected in few *Mirusviricete* MAGs, suggesting the existence of the lysogenetic phase of the life cycle for some of these viruses[40].

Overall, comparative genomic and functional results provided strong lines of evidence indicating that *Mirusviricetes* display a unique functional lifestyle within plankton - as compared to the other NCLDVs - that remains largely enigmatic due to the predominance of core genes of unknown functions. Those are prime candidates for functional equivalents of the MCP and pATPase. Further analyses (3D structure, phylogenies, occurrences beyond the scope of NCLDVs) are ongoing to better understand the functioning of *Mirusviricetes* within plankton.

Discussion

Our phylogeny-guided genome-resolved metagenomic survey of plankton in the surface of five oceans and two seas has allowed us to go beyond the known diversity of large and giant DNA viruses. Especially, we identified and characterized the genomic landscape of two new putative classes within the phylum *Nucleocytoviricota*: the *Proculviricetes* (recovery of a few population genomes from the Arctic and Southern Oceans) and the more prevalent *Mirusviricetes*. Since *Nucleocytoviricota* is otherwise formally organized into two classes (the *Megaviricetes* and *Pokkesviricetes*)[2,18], the recovery of more than 100 population genomes for *Mirusviricetes*, organized into five distinct subclades potentially corresponding to families is a noticeable expansion of this prominent phylum. Furthermore, the lack of signal for several hallmark genes related to DNA replication and more surprisingly viral particle morphogenesis suggests that the class *Mirusviricetes* may substantially diverge from the functional principles defined by means of culture genomics. Thus, this environmental genomic survey contributes to our global understanding of the diversity and functioning of *Nucleocytoviricota* within plankton.

Biogeographic signal demonstrated that *Mirusviricetes* 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 what was observed for other *Nucleocytoviricota* lineages (e.g., *Imitervirales* and *Pandoravirales* orders), strongly indicating that the five *Mirusviricetes* families regularly infect yet to be identified eukaryotic plankton, possibly with a broad cellular size range. These results suggest that *Mirusviricetes* viruses influence the ecology of key marine eukaryotes, albeit probably not to the same extent as the more diverse and abundant class of *Megaviricetes*. Moving forward, the comprehensive marine

genomic database for *Nucleocytoviricota* we built, which includes *Mirusviricetes* and *Proculviricetes*, may help solve many hypotheses that orbit around these large and giant DNA viruses, from their potential roles in the emergence and evolution of eukaryotes[14,15,41–44], to their regulation of ecologically critical marine eukaryotes[45,46].

Lack of signal for multiple hallmark genes in *Mirusviricetes* is an enigma. In particular, the MCP and pATPase genes correspond to the core of the virion morphogenesis module[28,29] and are conserved in the entire *Varidnaviria* realm that includes viruses infecting the three domains of life[2,47]. The large conservation of these two genes suggests that they are critical for the life style of these viruses. The few known exceptions, lacking either the MCP such as Pandoraviruses[48], or the pATPase such as Pithovirus-like viruses[4,49,50] are known for their different viral particle shapes and seem to have developed alternative strategies for protecting their genomes[51]. So far, the life style of *Mirusviricetes*, the only class of *Nucleocytoviricota* seemingly lacking both MCP and pATPase, hence remains unclear, especially regarding their ability to form viral particles. However, the presence of a histone (potentially related to the packaging of DNA in the viral particle[52]) among the core functions of *Mirusviricetes* suggests they have acquired a novel type of particle synthesis pathway, and do form viral particles. Critically, our comparative genomic analysis of the phylum *Nucleocytoviricota* exposed several core genes of unknown function for *Mirusviricetes* entirely missing in the other classes. These genes are prime candidates for the formation of a viral particle. Yet, it remains possible that *Mirusviricetes* do not build their own particles and/or use a substantially different infection strategy. The presence of a putative integrase in few *Mirusviricetes* population genomes is also reminiscent of the virophages, related members of the Varidnaviria realm able to integrate their hosts' genomes using a specific lifestyle[53].

The discovery of two classes within the phylum *Nucleocytoviricota* infecting marine eukaryotes is a reminder that we have not yet grasped the full diversity and complexity of plankton and related viruses in the sunlit ocean. This critical interface considerably influences global climate and is directly impacted by human-induced environmental changes, stressing the need to better understand its fundamental functional mechanisms. Here, we demonstrated that phylogeny-guided genome-resolved metagenomics could play a role, alongside more conventional culture-dependent and independent approaches, to characterize previously unknown genomic lineages. We anticipate that similar endeavors, empowered by expeditions such as *Tara* Oceans[19] and using a variety of hallmark genes as compass will shed light on other important genomic components of plankton in years to come, contributing to our basic understanding of the ecology and evolution of complex, dynamic, but also fragile marine ecosystems.

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[20,21]. Those 0.28 trillion reads were used as inputs for 11 metagenomic co-assemblies using MEGAHIT[54] v1.1.1, and the contig header names were simplified in the resulting assembly outputs using anvi'o[24] 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[55] 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[56] (minimum identity of 95%) and stored the recruited reads as BAM files using samtools[57], (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[23] 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 [58] 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[59] 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[14]: The sequences were aligned with MAFFT[60] 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[61] Plus option with IQ-TREE[62] 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[58] 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[14] (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[14] 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[63], 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.[6] (n=743) and Moniruzzaman et al.[7] (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[64]. We determined the average nucleotide identity (ANI) of each pair of NCLDV MAGs using the dnadiff tool from the MUMmer package[65] 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[55] 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[66] v2.10.1) and phylogenetic reconstructions, as previously described for eukaryotic hallmark genes[20]. 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[60] 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[62] v1.6.2 was used for the phylogenetic reconstructions, with the ModelFinder[61] 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)[67] and ultrafast bootstrap approximation (UFBoot[68]). 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.

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[62] v1.6.2 with the best fitting model according to the BIC from the ModelFinder[61] 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[69] (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)[67] and ultrafast bootstrap approximation (UFBoot[68]). 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.[6], Moniruzzaman et al.[7] and Aylward et al.[18].

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 [26]. The number of recruited reads below this cut-off was set to 0 before determining vertical coverage and percent of recruited reads.

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[20].

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[32]. AGNOSTOS produces groups of genes with low functional entropy as shown in Vanni et al. 2020[32] and Delmont et al. 2020[20] 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[70], RefSeq[71], UniRef90[72], NCVOGs[73], and NCBI nr database using Diamond[74] v2.0.6 with a cut-off E-value 1×10^{-5}. A recently published GVOG database[18] was also used in annotation using hmmer[58] 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[75] v2.1.5. Finally, tRNAscan-SE[76] v2.0.7 predicted 7,734 tRNAs.

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 *Mirusviricetes* 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 two classes of NCLDV. The latin adjective "**Mirus**" (strange, surprising, extraordinary) was selected to describe the large class of NCLDVs lacking signal for several hallmark genes: the ***Mirusviricetes***. The latin adverb "**Procul**" (away, at distance, far off) was selected to describe the 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.17694365. This link provides access to (1) the RNApolB genes reconstructed from the *Tara* Oceans assemblies, 2) individual FASTA files for the 1,593 non-redundant marine giant virus MAGs (including the 697 manually curated MAGs from our survey) and 224 reference giant virus genomes contained in the NCLDV marine genomic database, (3) genes and proteins found in the NCLDV marine genomic database (4) the phylogenetic tree corresponding to figures 1 and 2 with associated metadata and their anvi'o PROFILE databases, (5) and all the supplemental tables.

Contributions:

Tom O. Delmont conducted the study. Tom O. Delmont, Morgan Gaïa, Lingjie Meng, Chiara Vanni and Eric Pelletier 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. Chiara Vanni produced gene cluster with remote homologies. 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), and The Kyoto University Foundation, and the Collaborative Research Program of the Institute for Chemical Research, Kyoto University (2021-29, 2020-28). 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:

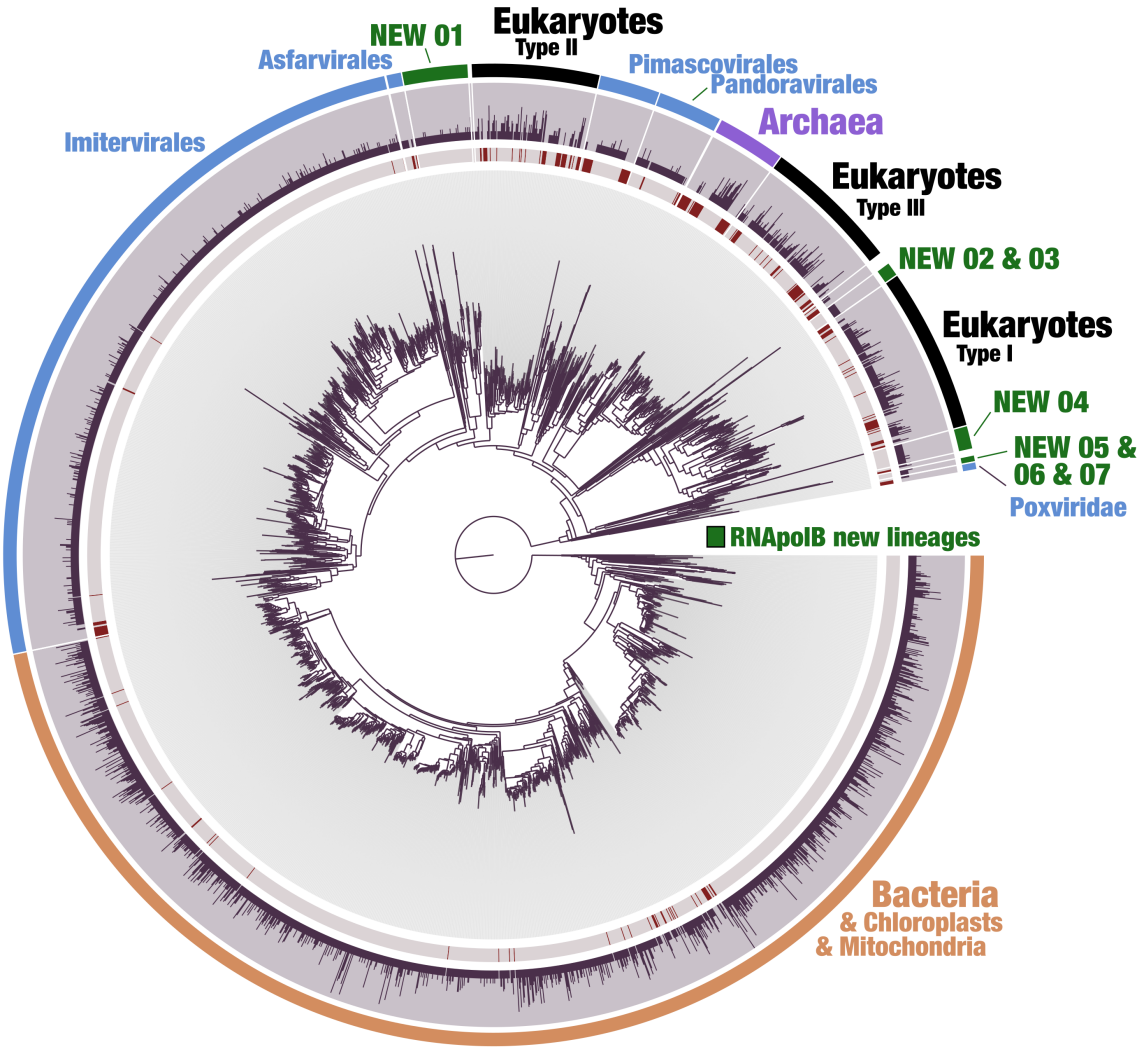


Figure S1: Evolutionary diversity of the DNA-dependent RNA polymerase B in the sunlit ocean. The maximum-likelihood phylogenetic tree is based on 2,728 RNApolB sequences more than 800 amino acids in length with similarity <90% (gray color) identified from 11 large marine metagenomic co-assemblies. This 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.

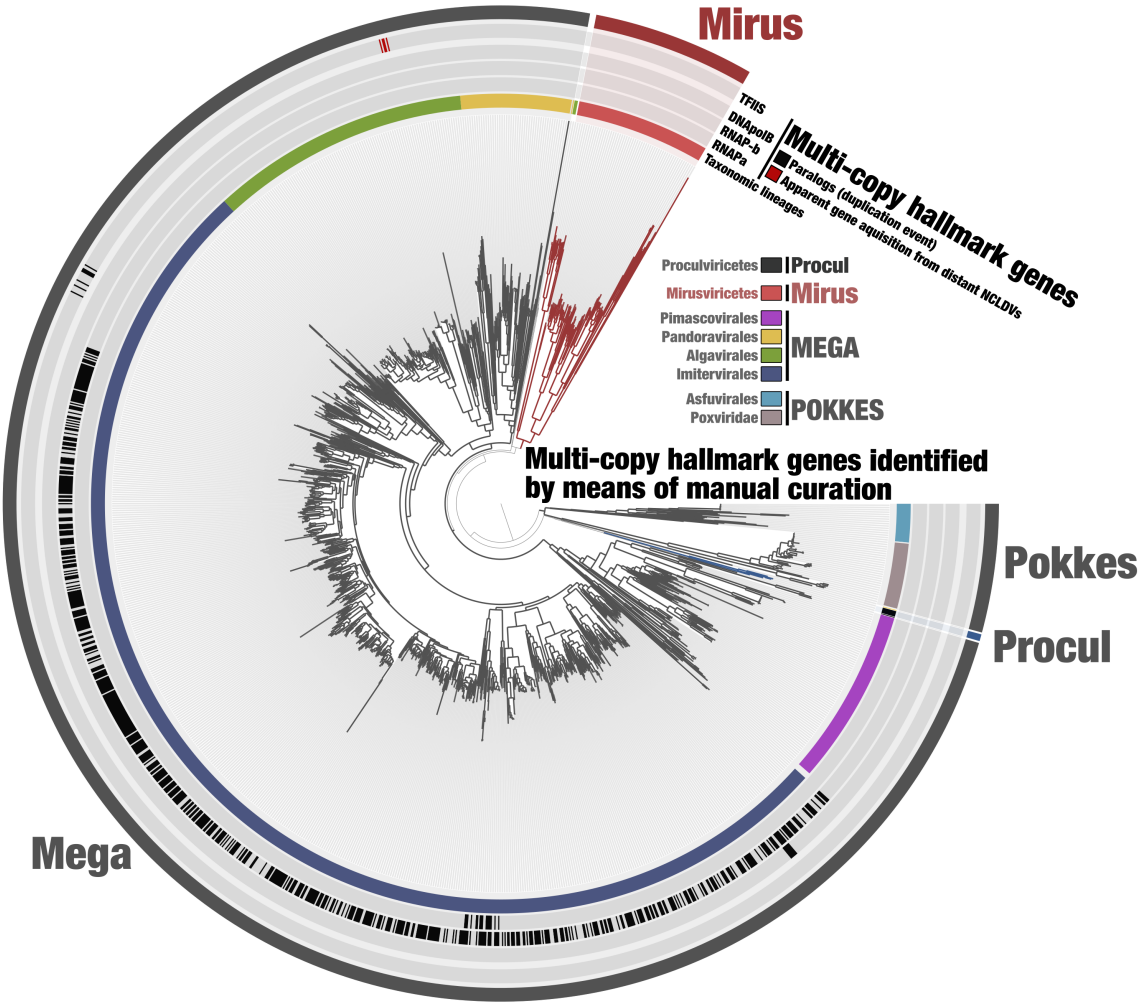


Figure S2: A phylogennetic perspective on hallmark gene markers found in multiple copy in some clades of NCLDVs. The phylogenomic tree (same as in figures 1 and 2) was built from the NCLDV marine genomic database based on a concatenation of four hallmark genes using the PMSF mixture model. Genomes only containing the TFIIS hallmark gene were excluded from this analysis.

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: Metagenomic blocks containing DNA-dependent RNA polymerase subunit B genes of interest and used for genome-resolved metagenomics

Table S4: Genomic and environmental statistics for the marine NCLDV genomic database.

Table S5: Biogeographic signal for the marine NCLDV genomic database.

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

Table S7: Functional annotations for genes found in the marine NCLDV genomic database. Statistics are included.

References:

1. Koonin, E. V & Yutin, N. Nucleo-cytoplasmic Large DNA Viruses (NCLDV) of Eukaryotes. *eLS* (2012). doi:10.1002/9780470015902.A0023268
2. Koonin, E. V. *et al.* Global Organization and Proposed Megataxonomy of the Virus World. *Microbiol. Mol. Biol. Rev.* **84**, (2020).
3. Wilhelm, S. W. *et al.* A Student's Guide to Giant Viruses Infecting Small Eukaryotes: From Acanthamoeba to Zooxanthellae. *Viruses 2017, Vol. 9, Page 46* **9**, 46 (2017).
4. Legendre, M. *et al.* Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses with a pandoravirus morphology. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 4274–4279 (2014).
5. Philippe, N. *et al.* Pandoraviruses: Amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. *Science (80-.).* (2013). doi:10.1126/science.1239181
6. Schulz, F. *et al.* Giant virus diversity and host interactions through global metagenomics. *Nature* (2020). doi:10.1038/s41586-020-1957-x
7. Moniruzzaman, M., Martinez-Gutierrez, C. A., Weinheimer, A. R. & Aylward, F. O. Dynamic genome evolution and complex virocell metabolism of globally-distributed giant viruses. *Nat. Commun.* (2020). doi:10.1038/s41467-020-15507-2
8. Endo, H. *et al.* Biogeography of marine giant viruses reveals their interplay with eukaryotes and ecological functions. *Nat. Ecol. Evol. 2020 412* **4**, 1639–1649 (2020).
9. Bratbak, G., Wilson, W. & Heldal, M. Viral control of Emiliania huxleyi blooms? *J. Mar. Syst.* **9**, 75–81 (1996).
10. Moniruzzaman, M. *et al.* Diversity and dynamics of algal Megaviridae members during a harmful brown tide caused by the pelagophyte, Aureococcus anophagefferens. *FEMS Microbiol. Ecol.* **92**, (2016).
11. Tarutani, K., Nagasaki, K. & Yamaguchi, M. Viral impacts on total abundance and clonal composition of the harmful bloom-forming phytoplankton Heterosigma akashiwo. *Appl. Environ. Microbiol.* **66**, 4916–4920 (2000).
12. Moniruzzaman, M., Weinheimer, A. R., Martinez-Gutierrez, C. A. & Aylward, F. O. Widespread endogenization of giant viruses shapes genomes of green algae. *Nature* (2020). doi:10.1038/s41586-020-2924-2
13. Bell, P. J. L. The viral eukaryogenesis hypothesis: A key role for viruses in the emergence of eukaryotes from a prokaryotic world environment. in *Annals of the New York Academy of Sciences* (2009). doi:10.1111/j.1749-6632.2009.04994.x
14. 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.* (2019). doi:10.1073/pnas.1912006116

15. Takemura, M. Medusavirus Ancestor in a Proto-Eukaryotic Cell: Updating the Hypothesis for the Viral Origin of the Nucleus. *Front. Microbiol.* **11**, 2169 (2020).
16. Yutin, N. & Koonin, E. V. Hidden evolutionary complexity of Nucleo-Cytoplasmic Large DNA viruses of eukaryotes. *Virol. J.* **9**, 1–18 (2012).
17. Koonin, E. V. & Yutin, N. Evolution of the Large Nucleocytoplasmic DNA Viruses of Eukaryotes and Convergent Origins of Viral Gigantism. *Adv. Virus Res.* **103**, 167–202 (2019).
18. Aylward, F. O., Moniruzzaman, M., Ha, A. D. & Koonin, E. V. A phylogenomic framework for charting the diversity and evolution of giant viruses. *PLOS Biol.* **19**, e3001430 (2021).
19. Sunagawa, S. *et al.* Tara Oceans: towards global ocean ecosystems biology. *Nat. Rev. Microbiol.* 1–18 (2020). doi:10.1038/s41579-020-0364-5
20. 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
21. Delmont, T. O. *et al.* Heterotrophic bacterial diazotrophs are more abundant than their cyanobacterial counterparts in metagenomes covering most of the sunlit ocean. *ISME J. 2021* 1–10 (2021). doi:10.1038/s41396-021-01135-1
22. Mihara, T. *et al.* Taxon Richness of 'Megaviridae' Exceeds those of Bacteria and Archaea in the Ocean. *Microbes Environ.* **33**, 162–171 (2018).
23. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**, 1144–1146 (2014).
24. Eren, A. M. *et al.* Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **3**, e1319 (2015).
25. Eren, A. M. *et al.* Community-led, integrated, reproducible multi-omics with anvi'o. *Nature Microbiology* (2021). doi:10.1038/s41564-020-00834-3
26. Delmont, T. O. *et al.* Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. *Nat. Microbiol.* 1 (2018). doi:10.1038/s41564-018-0176-9
27. Gallot-Lavallée, L., Blanc, G. & Claverie, J.-M. Comparative Genomics of Chrysochromulina Ericina Virus and Other Microalga-Infecting Large DNA Viruses Highlights Their Intricate Evolutionary Relationship with the Established Mimiviridae Family. *J. Virol.* **91**, (2017).
28. Kirnbauer, R., Booy, F., Cheng, N., Lowy, D. R. & Schiller, J. T. Papillomavirus L1 major capsid protein self-assembles into virus-like particles that are highly immunogenic. *Proc. Natl. Acad. Sci.* **89**, 12180–12184 (1992).
29. Hilbert, B. J. *et al.* Structure and mechanism of the ATPase that powers viral genome packaging. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E3792–E3799 (2015).
30. Kinashi, H. Giant linear plasmids in Streptomyces: a treasure trove of antibiotic biosynthetic clusters. *J. Antibiot. 2011 641* **64**, 19–25 (2010).
31. Philippe, N. *et al.* Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. *Science* **341**, 281–286 (2013).
32. Vanni, C. *et al.* Unifying the known and unknown microbial coding sequence space. *bioRxiv* 2020.06.30.180448 (2021). doi:10.1101/2020.06.30.180448
33. Strack, B., Calistri, A., Craig, S., Popova, E. & Göttlinger, H. G. AIP1/ALIX is a

binding partner for HIV-1 p6 and EIAV p9 functioning in virus budding. *Cell* **114**, 689–699 (2003).
34. McNew, J. A. *et al.* Ykt6p, a prenylated SNARE essential for endoplasmic reticulum-Golgi transport. *J. Biol. Chem.* **272**, 17776–17783 (1997).
35. Pasqualato, S., Renault, L. & Cherfils, J. Arf, Arl, Arp and Sar proteins: a family of GTP-binding proteins with a structural device for 'front-back' communication. *EMBO Rep.* **3**, 1035–1041 (2002).
36. Matsuto, M., Kano, F. & Murata, M. Reconstitution of the targeting of Rab6A to the Golgi apparatus in semi-intact HeLa cells: A role of BICD2 in stabilizing Rab6A on Golgi membranes and a concerted role of Rab6A/BICD2 interactions in Golgi-to-ER retrograde transport. *Biochim. Biophys. Acta* **1853**, 2592–2609 (2015).
37. Gao, X. D. *et al.* The GAP activity of Msb3p and Msb4p for the Rab GTPase Sec4p is required for efficient exocytosis and actin organization. *J. Cell Biol.* **162**, 635–646 (2003).
38. Adachi, H. & Tsujimoto, M. FEEL-1, a novel scavenger receptor with in vitro bacteria-binding and angiogenesis-modulating activities. *J. Biol. Chem.* **277**, 34264–34270 (2002).
39. Twal, W. O. *et al.* Fibulin-1 suppression of fibronectin-regulated cell adhesion and motility. *J. Cell Sci.* **114**, 4587–4598 (2001).
40. Canavessi, A. M. O., Harms, J., Gatti, N. D. L. & Splitter, G. A. The role of integrase/recombinase xerD and monofunctional biosynthesis peptidoglycan transglycosylase genes in the pathogenicity of Brucella abortus infection in vitro and in vivo. *Microb. Pathog.* **37**, 241–251 (2004).
41. Forterre, P. & Gaïa, M. Giant viruses and the origin of modern eukaryotes. *Current Opinion in Microbiology* (2016). doi:10.1016/j.mib.2016.02.001
42. 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
43. Livingstone Bell, P. J. Viral eukaryogenesis: Was the ancestor of the nucleus a complex DNA virus? *J. Mol. Evol.* (2001). doi:10.1007/s002390010215
44. Takemura, M. Poxviruses and the origin of the eukaryotic nucleus. *J. Mol. Evol.* (2001). doi:10.1007/s002390010171
45. Kaneko, H. *et al.* Eukaryotic virus composition can predict the efficiency of carbon export in the global ocean. *iScience* **24**, 102002 (2021).
46. Laber, C. P. *et al.* Coccolithovirus facilitation of carbon export in the North Atlantic. *Nat. Microbiol. 2018 35* **3**, 537–547 (2018).
47. 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. *Front. Microbiol.* **12**, (2021).
48. Legendre, M. *et al.* Diversity and evolution of the emerging Pandoraviridae family. *Nat. Commun.* (2018). doi:10.1038/s41467-018-04698-4
49. Andreani, J. *et al.* Cedratvirus, a Double-Cork Structured Giant Virus, is a Distant Relative of Pithoviruses. *Viruses* **8**, (2016).
50. Andreani, J. *et al.* Orpheovirus IHUMI-LCC2: A new virus among the giant viruses. *Front. Microbiol.* **8**, (2018).

51. Krupovic, M., Yutin, N. & Koonin, E. Evolution of a major virion protein of the giant pandoraviruses from an inactivated bacterial glycoside hydrolase. *Virus Evol.* **6**, (2020).
52. Valencia-Sánchez, M. I. *et al.* The structure of a virus-encoded nucleosome. *Nat. Struct. Mol. Biol. 2021 285* **28**, 413–417 (2021).
53. Duponchel, S. & Fischer, M. G. Viva lavidaviruses! Five features of virophages that parasitize giant DNA viruses. *PLOS Pathog.* **15**, e1007592 (2019).
54. 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).
55. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
56. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
57. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
58. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
59. 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).
60. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* (2013). doi:10.1093/molbev/mst010
61. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A. & Jermiin, L. S. ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat. Methods* (2017). doi:10.1038/nmeth.4285
62. 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. *Mol. Biol. Evol.* (2015). doi:10.1093/molbev/msu300
63. Delmont, T. O. & Eren, A. M. Identifying contamination with advanced visualization and analysis practices: metagenomic approaches for eukaryotic genome assemblies. *PeerJ* **4**, e1839 (2016).
64. Needham, D. M. *et al.* Targeted metagenomic recovery of four divergent viruses reveals shared and distinctive characteristics of giant viruses of marine eukaryotes. *Philos. Trans. R. Soc. B* **374**, (2019).
65. 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).
66. 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).
67. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
68. Hoang, D. T., Chernomor, O., Von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.*

(2018). doi:10.1093/molbev/msx281
69. 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).
70. Mihara, T. *et al.* Linking Virus Genomes with Host Taxonomy. *Viruses* **8**, (2016).
71. 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).
72. 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).
73. 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).
74. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
75. 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 Res.* (2019). doi:10.1093/nar/gky1085
76. 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).