Unveiling the global diversity and evolution of giant viruses through ocean metagenomics

Lingjie Meng

Table of Contents

Abstract

Giant viruses are a group of eukaryote-infecting double-stranded DNA viruses that possess large functional repertoires. So far, all known giant viruses belong to the phylum *Nucleocytoviricota*. By infecting a wide variety of eukaryotes, giant viruses are abundant and widespread in the ocean, from the Arctic Ocean to the Southern Ocean. Despite their importance in the marine ecosystems, our knowledge of marine giant viral ecology and evolution remains largely limited as the scarcity of isolated virus-host pairs. Compared to fastidious isolation approaches, *in silico* analyses hold the promise of bridging knowledge gaps, to reveal the enigmatic world of marine giant viruses.

Taking advantage of the recent large-scale marine metagenomics census, I explored the ecology and evolution of marine giant viruses using global ocean data (i.e., the *Tara* Oceans), the findings of which are encapsulated in my Ph.D. dissertation. Four major findings are organized into separate chapters in this dissertation as outlined below. 1) The global distribution of marine giant viruses showed various latitudinal diversity gradient patterns along size fractions and lineages, suggesting they have a diverse host range. 2) The co-occurrence-based network host prediction approach was quantitively assessed and improved by a phylogeny-guided mapping method. 3) A distinct boundary between giant viral communities in polar and nonpolar environments was identified. Further, results supported the hypothesis that recurrent evolutionary adaptations across the boundary are likely driven by alterations of viral gene repertoire. 4) '*Mirusviricota*', a group of plankton-infecting DNA giant viruses with remarkable chimeric attributes, was discovered prevalent in the ocean, providing missing links in the evolution of both herpesviruses and giant viruses.

Together, this dissertation provides a comprehensive examination of the ecology, evolution, and methodology on marine giant viruses using metagenomic data, shedding new light on the enigmatic viral world.

Chapter 1 General Introduction

At the end of the nineteenth century, viruses were initially discovered and initially defined as 'filterable agents' causing infectious diseases in animals and plants. They represent a group of 'life form' that exclusively parasitize living cells. Correspondingly, for a long while, viruses were seen as simple and small entities that could not be observed under light microscopes. However, the appearance of a group of double-stranded DNA viruses has substantially changed our understanding of viruses (La Scola et al., 2003). Those viruses possess large genomes (~2.5 Mbp; pandoravirus) and outsized virions (~1500 nm; pithovirus) (some are even larger than typical bacteria) (La Scola et al., 2003; Philippe et al., 2013; Raoult et al., 2004). Therefore, they are referred to as '**Giant Viruses**'. Over the past two decades, the identified diversity of giant viruses has significantly expanded, particularly with the development of *in situ* sequencing techniques (Schulz et al., 2022). The continuously increasing diversity, as well as the gigantism of viral genomes, has sparked widespread interest in understanding their mysterious origins and evolutionary trajectories. Marine habitats, where abundant and diverse viral signals were detected (Schulz et al., 2020), serve as an ideal resource for understanding giant viruses. Giant viruses infect a broad range of marine eukaryotes, from unicellular organisms and macroalgae to animals (Sun et al., 2020). They play important ecological roles through various means, such as host populations regulation, biogeochemical cycle, and genetic exchange. This chapter provides a brief summary of the fundamental knowledge and findings about giant viruses, especially marine giant viruses.

1.1 Diversity and Evolution of Giant Virus

1.1.1 'Giant Virus' in the Dissertation

With the increasing discovery of viruses, understanding viral 'species diversity' has long been a goal but challenge due to several limitations. Firstly, the absence of a universal marker gene makes it difficult to identify and infer the phylogeny for all viruses (Sullivan, 2015). Secondly, the high mutation rates (Duffy et al., 2008), observed in many viruses, can lead to significant genetic divergence, increasing challenges on the taxonomy for close relativeness. Thirdly, many virus-host pairs are difficult-to-culture in laboratory environment (Schulz et al., 2022), making the validation of viral identification and classification challenging. Fourthly, the complex viral physiology and broad range of hosts further add the complexity in understanding the real diversity. In recent years, our understanding on viral diversity has been largely improved through the UViG[1] (Roux et al., 2019, 2021). Meanwhile, a comprehensive hierarchical virus taxonomy was proposed by the International Committee on Virus Taxonomy (ICTV) (Siddell et al., 2023), based on a hybrid viral evolution hypothesis combining 'virus-first' and 'escape' models (details in 1.1.3) (Koonin et al., 2020a; Krupovic, Dolja, et al., 2020; Krupovic et al., 2019). This well accepted system includes six taxa at the highest level, 'Realms'[2], each of which is thought to be monophyletic and share highly conserved specific traits over time (Gorbalenya et al., 2020; Koonin et al., 2020b). Subsequently, among 17 viral phyla, the phylum *Nucleocytoviricota*, which belongs to the Realm *Varidnaviria* and the Kingdom *Bamfordvirae*, is well-known for including all known giant viruses, like Acanthamoeba polyphaga mimivirus (APMV), megavirus chilensis (Arslan et al., 2011), tupanvirus (Abrahão et al., 2018), Cafeteria roenbergensis virus (CroV) (Fischer et al., 2010). Members of *Nucleocytoviricota* are also commonly referred to as NucleoCytoplasmic Large DNA Viruses (NCLDVs). Due to the absence

[1] UViG: Uncultivated virus genome identified directly from environmental meta-omics datasets.

[2] The rank of 'Realm' corresponds to the rank of 'Domain' for cellular life

of a precise definition for 'giant viruses', the terms 'NCLDV' and 'giant virus' (including small nucleocytoviruses) have been used interchangeably for an extended period.

In addition to *Nucleocytoviricota*, a newly identified viral phylum dubbed '*Mirusviricota*' (as the major finding discussed in Chapter 5) also possesses large DNA genomes, approximately 400 kb in size. Mirusviruses encode a set of structural proteins of the virion module[3], represented by the HK97-fold Major Capsid Proteins (HK97-MCPs), similar to phages and herpesviruses, which makes them belong to the Realm *Duplodnaviria* (Iranzo et al., 2016; Krupovic et al., 2019). However, mirusviruseses likely infect marine plankton, so that sharing an overlapped ecological niche with members of *Nucleocytoviricota.* Also, mirusviruseses have a similar function repertoire and a similar information module[4] with with giant viruses under *Nucleocytoviricota*, rather than the viruses under *Duplodnaviria* (Gaïa et al., 2023). Taken together, working on mirusviruses helps to gain a rather comprehensive understanding on the diversity and ecology on giant viruses. So '*Mirusviricota'* was also taken into account for 'Giant Viruses' in the context of this dissertation.

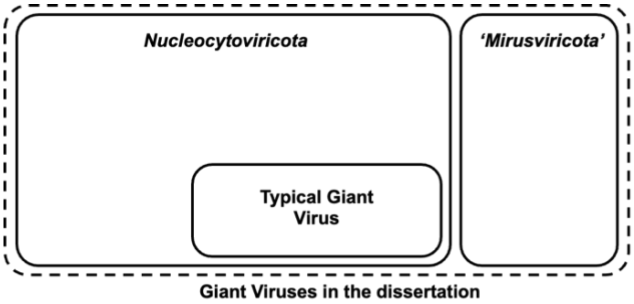


Fig. 1-1 Definition of 'Giant Virus' used in the dissertation. Typically, 'Giant Virus' stands for a group of viruses with large and giant genomes and virions. All known giant viruses belong to the phylum *Nucleocytoviricota*. Because of the similar niche and function repertoire, as well as for clarity and consistency of this dissertation, the relative general definition of 'Giant Virus' was applied that includes both known phylum *Nucleocytoviricota* and novel '*Mirusviricota*', a potential phylum discussed in the Chapter 5.

Overall, this dissertation focuses on the ecology and evolution of marine eukaryotic DNA viruses. For clarity and consistency, the term 'Giant Virus' (GV in some abbreviations) will

[3] Virion module refers to a set of genes responsible for the formation of virions.

[4] Informational module refers to a gene set responsible for the replication and trancription.

collectively refer to members of the *Nucleocytoviricota* and '*Mirusviricota'* phyla (Fig. 1-1). It's worth noting that while some members of *Nucleocytoviricota*, such as iridoviruses and ascoviruses, have large but not typical giant genomes, their genomic information also offer valuable insights into the overarching landscape of viral diversity and ecology.

1.1.2 Diversity of known Giant Virus

Two classes and five orders of the phylum *Nucleocytoviricota* have been approved by the ICTV (Koonin & Yutin, 2019): one class is *Megaviricetes*, containing orders *Pimascovirales*, *Algavirales*, and *Imitervirales*; the other one class is *Pokkesviricetes*, including orders of *Asfuvirales*, *Chitovirales*. Additionally, a recent phylogenomic study supplemented by environmental data has proposed an additional order, '*Pandoravirales'*, under the class *Megaviricetes* (Aylward et al., 2021), which addresses the paraphyletic issues present within original *Algavirales*. This taxonomic system, comprising two classes and six orders, was employed in Chapter 2, Chapter 4 and Chapter 5 of this dissertation. In Chapter 3, due to the absence of a unified taxonomic standard at the time of analysis (2019), a family-level grouping was performed. Members within *Nucleocytoviricota* are very diverse and share a small set of core genes. Three genes are universal in all isolated *Nucleocytoviricota* genomes: DNA polymerase B family (DNApolB), primase-helicase and poxvirus late transcription factor 3 (VLTF3). Two genes for double Jellyroll folded MCP (DJR-MCP) and packaging ATPase are nearly universally conserved (Koonin & Yutin, 2019).

Table 1 Taxonomy system of *Nucleocytoviricota* used in this dissertation

Phylum	Class	Order	Family
Nucleocytoviricota	***Megaviricetes***	***Imitervirales***	***Allomimiviridae; Mesomimiviridae; Schizomimiviridae; Mimiviridae***
		Algavirales	***Phycodnaviridae; 'Prasinoviridae';***

		Pimascovirales	***Ascoviridae; Iridoviridae; Marseillviridae*; *'Pithoviridae'***
		'Pandoraviralesʼ	***'Pandoraviridaeʼ*; *'Coccolithoviridaeʼ* *Mamonoviridae***
	Pokkesviricetes	***Asfuvirales***	***Asfarviridae***
		Chitovirales	***Poxviridae***

Under *Megaviricetes*, the first giant virus, APMV, was identified in 2003 (La Scola et al., 2003). Since then, dozens of relatives of APMV have been isolated and have formed a monophyletic order, *Imitevirales*. *Imitevirales* is the most diversified order (currently including four families (Aylward et al., 2023)) that was found to be ubiquitous and abundant in nature (Hingamp et al., 2013; Y. Li et al., 2018; Mihara et al., 2018). *Imitevirales* infects a wide range of aquatic eukaryotes, including protists and algae (Fischer et al., 2010; La Scola et al., 2003; Sun et al., 2020). The order *Algavirales*, previously known as phycodnaviruses, predominantly infects marine and freshwater algae. Members of this order, such as raphidovirus, play crucial roles in marine ecosystems, acting as regulatory factors in terminating harmful algal blooms (Nagasaki & Yamaguchi, 1997). *Pimascovirales* includes four phylogenomic relatives: *Ascoviridae*, *Iridoviridae*, *Marseileviridae*, and one 30,000-year-old family, *Pithoviridae* (Legendre et al., 2014). In aquatic ecosystems, iridoviruses infect hosts ranging from invertebrates (shrimp) (Tang et al., 2007) to vertebrate animals (fish) (Whittington et al., 2010). The recently proposed *Pandoravirales* further adds to the complexity of the *Megaviricetes* class (Philippe et al., 2013). This order is established based on the phylogenomic closeness between pandoraviruses and coccolithoviruses. The latter infect the marine algae *Emiliania huxleyi*, a species abundant and ecologically significant in oceanic waters (Wilson et al., 2005).

Another Class, *Pokkesviricetes*, represents a divergent group with a long branch (Aylward et al., 2021). Within this class, the order *Chitovirales* is particularly well studied for its inclusion of poxviruses, which include the causative agents of smallpox in humans (Hopkins, 1983). However, poxvirus relatives are not typically found in marine environments. The other order within *Pokkesviricetes*, *Asfuvirales*, is far more relevant to marine ecosystems. This order includes viruses that infect a wide array of aquatic organisms, ranging from abalones (Matsuyama et al., 2020) to single cellular dinoflagellates (Ogata et al., 2009).

There are still several giant virus lineages do not neatly fit into the above *Nucleocytoviricota* orders. For instance, medusavirus that was recently isolated in hot springs, has exhibited a long history of co-evolution with its eukaryotic hosts (Yoshikawa et al., 2019). A family-level taxonomy, *Mamonoviridae*, has been proposed to categorize medusaviruses within the *Nucleocytoviricota* and accepted by ICTV (Zhang et al., 2023). Additionally, a group of shrimp-associated viruses, dubbed mininucleovirus, was suggested to be a distinct family within the *Pimascovirales* group (Subramaniam et al., 2020). Although the taxonomy remains uncertain, a smaller double-stranded DNA virus, yaravirus, has demonstrated phylogenetic relatedness with giant viruses and may act as evolutionary intermediaries between giant *Nucleocytoviricota* and small polinton-like viruses (M Boratto et al., 2020). Together, these fascinating discoveries suggest that the full scale of giant virus diversity has yet to be completely explored.

1.1.3 Evolutionary Scenarios

"So, how did giant viruses achieve such diverse?" Understanding the evolution of giant viruses necessitates background of entire virosphere. In summary, there are three traditional theories about the origins and evolution of viruses: 1) 'Virus-first hypothesis' that viruses evolved from complex molecules of protein and nucleic acid predate the divergence of life; 2) 'Reduction hypothesis' that virus underwent a reductive evolution from cells and is supported by the discovery of giant viruses; 3) 'Escape hypothesis' that viruses originally evolved from selfish genes of larger organisms. As mentioned, the present ICTV taxonomic system relies on a hybrid hypothesis (Siddell et al., 2023). Namely, the informational modules of viruses originated from a primordial genetic pool. During their extensive evolutionary history, the gene pool has undergone numerous changes, often being replaced by genes acquired from their cellular hosts, including those in the virion module (Krupovic et al., 2019). This model is supported by estimates of gene gain and gene loss throughout evolution, which suggest that giant viruses evolved from smaller eukaryotic dsDNA viruses (Koonin & Yutin, 2019).

Similarly, the origin of giant viruses also remains subjects of controversy. Conflicting phylogenetic analyses offer different perspectives on whether they form a monophyletic group or have multiple origins (Iyer et al., 2006). Some studies suggest that giant viruses, like the giant pandoraviruses, emerged during the diversification of modern eukaryotes (Krupovic, Yutin, et al., 2020). However, most of giant viruses were proposed to be ancient and predated the Last Eukaryotic Common Ancestor (LECA). A phylogeny-based study with DNA-dependent RNA polymerase suggested giant virus to exist on the earth and start co-evolution with proto-eukaryotes dating from 2 billion years ago (Guglielmini et al., 2019). Another study also supportes this hypothesis by showing that some giant viruses encode viractins (actin-related genes in viruses) (Da Cunha et al., 2022), which could have been acquired from proto-eukaryotes and possibly reintroduced in the pre-LECA eukaryotic lineage. These differing viewpoints highlight the complexity and ongoing uncertainty in tracing the evolutionary history of giant viruses.

In addition, several mechanisms were suggested to contribute to the gigantism evolution of giant viruses: Horizontal Gene Transfer (HGT) from cellular organisms (Irwin et al., 2022; Kijima et al., 2021), HGT between viruses (Wu et al., 2023) and gene duplication (Machado et al., 2023). Additionally, giant viruses may have contributed to the evolution of eukaryotic organisms. Apart from prolonged co-evolution through gene transfers, the widespread endogenization of giant viruses also potentially alter the evolution of hosts (details in 1.2.3) (Moniruzzaman, Weinheimer, et al., 2020). Another hypothesis is viral eukaryogenesis, suggesting eukaryotic nucleus may have originated from an ancient giant virus (Bell, 2022). This hypothesis is supported by the similarities between certain properties of giant viruses and cellular nucleus, such as mRNA capping and tubulin.

1.1.4 Knowledge Gap in Giant Virus Evolution

The evolutionary history of giant viruses remains a subject of ongoing debate and uncertainty. The hybrid hypothesis for ICTV system suggests that giant viruses evolved from smaller eukaryotic dsDNA viruses. In fact, members of eukaryotic dsDNA viruses under *Varidnaviria* exhibit a wide range of genome sizes, from as small as ~10 kb to over 2 Mb. Within this broad range, a noticeable gap in genome complexity exists between giant viruses, *Nucleocytoviricota*, and the rest of the *Varidnaviria* viruses with genomes smaller than 50 kb. In addition, some critical evolutionary events, like the acquisition of multiple informational genes, set them apart from their smaller *Varidnaviria* counterparts. "When and how did these pivotal evolutionary events happen?" The full understanding of this complexification process remains elusive.

1.2 The Ecology of Marine Giant Viruses

1.2.1 Host Spectrum and Infection Dynamics

Viruses can only thrive in an environment where their hosts exist. Giant viruses are known to infect a broad range of eukaryotes, from unicellular eukaryotes and macroalgae to animals (Sun et al., 2020) (Fig. 1-2). Because of the isolation approach, Amoebae (mainly *Acanthamoeba* and *Vermamoeba*) are known to be infected by many giant viruses and served as either native host or a melting pot among amoeba-associated organisms like bacterial symbionts (Boyer et al., 2009). Apart from amoebae, especially in marine systems, giant viruses can infect many phytoplankton groups, such as Pelagophyceae (Gastrich et al., 2004), Mamiellophyceae (Finke et al., 2017), Dinophyceae (Ogata et al., 2009), and Haptophyte (Wilson et al., 2005). Several other non-photosynthetic eukaryotic lineages, such as Bicoecea (Colson et al., 2011) and Choanoflagellatea (Needham et al., 2019), were also reported as viral hosts in marine environments[5]. *Poxviridae* and *Iridoviridae* exclusively infect animals from small invertebrates to large vertebrates (Tang et al., 2007; Whittington et al., 2010). Together, these studies indicate infectious relationships between giant viruses and a wide range of marine eukaryotes. An interesting hypothesis about host spectrum is that viruses with larger genomes tend to infect a wider range of hosts because of the plasticity and variability (Sun et al., 2020).

Generally, giant viruses follow a similar replication cycle to other viruses, which includes stages such as virus entry, gene expression, replication, assembly, and finally, release of virions. However, the infection strategies vary within *Nucleocytoviricota*. Some giant viruses have the capability to replicate in the host cell cytoplasm with viral factories while some can only replicate in the nucleus, as summarized in a review (Schulz et al., 2022). The diversity of infection strategy could be reflected in the population dynamics of host-virus pairs. The population of many eukaryotic host organisms, particularly marine plankton, is greatly regulated by giant viruses. For example, correlated population dynamics were observed between the *E. huxleyi* and coccolithovirus during the bloom (Martínez Martínez et al., 2007). Conversely, some giant viruses have a long-term coexistence period with their hosts during infection (Blanc-Mathieu et al., 2021). A model between microalgea and prasinoviruses was built that a stable relationship is based on resistant-susceptible switch involved a large deletion on one chromosome (Yau et al., 2020). All in all, giant viruses are able to play an important role in

[5] Choanoflagellatea was identified as the host based on a single-cell sequencing estimation.

controlling the population of plankton in the ocean. Their ecological roles are reflected in several events, like algae blooms (Tarutani et al., 2000) and carbon cycles (Kaneko et al., 2021). Gaining insights into the distributions and dynamics of giant viruses could deepen our understanding of the micro-interactions in the ocean, further revealing the broader ecological roles that giant viruses play.

1.2.2 Biogeography of Giant Virus

Viruses of *Nucleocytoviricota* have been discovered in a wide variety of natural habitats. One determinant factor is that hosts of some giant viruses are ubiquitous. For example, though both infecting amoebae, APMV was isolated from the water of a cooling tower (La Scola et al., 2003), while medusavirus was discovered in hot spring water (Yoshikawa et al., 2019). Another reason is that giant viruses infect diverse range of hosts as introduced. A metagenomic survey conducted on soil samples identified 240 MCP homologues, suggesting that giant viruses and their hosts remain largely underexplored in terrestrial ecosystems (Schulz et al., 2018). The discovery also covers temporal scaled samples, like an ancient pithovirus was isolated from approximately 30,000-year-old Siberian permafrost (Legendre et al., 2014). In a remote organic lake in Antarctic, Organic Lake Phycodnaviruses (OLPVs) were detected along with their potential associated virophages (Yau et al., 2011). Despite the lack of evidence proving that giant viruses, apart from poxviruses, can infect humans, some studies identified the presence of giant viruses in the human-related samples (Colson et al., 2017). Among all these diverse habitats, marine ecosystems is the environment harboring most detected viral signals (Schulz et al., 2020).

Giant viruses exhibit high levels of diversity in oceanic environments (Monier et al., 2008). Utilizing a single marker gene, RNAP, *Imitervirales* was found to have greater diversity and richness compared to bacteria and archaea (Mihara et al., 2018). From one liter of marine water, more than

5,000 Operational Taxonomic Units (OTUs)[6] of *Imitervirales* could be detected (Y. Li et al., 2018). For abundance, giant viruses were found to outnumber eukaryotes (Hingamp et al., 2013). Specifically, *Algalvirales* is the most abundant, followed by the *Imitervirales* group as the second most abundant. Moreover, giant viruses are the active in the ocean, owning most of the detected active viral genes (86%) (Carradec et al., 2018). This metatranscriptomic data shows that giant viruses are actively infecting their eukaryotic hosts. Another DNApolB survey demonstrated that giant viruses are ubiquitously distributed over the oceans of the globe (Endo et al., 2020), while the arctic ocean virome is heterogenous. Giant viruses are not confined to the epipelagic zone as some are found in deep-sea waters, like tupanvirus (Abrahão et al., 2018), and deep-sea sediments (Bäckström et al., 2019). Despite the community of giant viruses in the deep sea remains largely unexplored, it is suspected that some viruses in the deep sea may be vertically transported from surface waters (Endo et al., 2020), and their infectivity might be affected by iron availability (Gilbert et al., 2023).

Collectively, research based on global metagenomic data has provided a comprehensive overview of the distribution of giant viruses. At the genome level, Metagenome-Assembled Genomes (MAGs[7]) of giant viruses were detected in diverse environments, including non-marine saline, terrestrial, and wastewater environments. Same to their diversity, it has been confirmed that the highest abundance of viral signals is found in the ocean (Moniruzzaman, Martinez-Gutierrez, et al., 2020; Schulz et al., 2020). Some studies, based on metagenomic and metatranscriptomic, have analyzed the diversity pattern for certain lineages of giant viruses (Ibarbalz et al., 2019a), characterized by lower biodiversity in polar regions and higher biodiversity in temperate zones.

[6] OTU: an operational definition used to classify groups of closely related viruses.

[7] MAGs: genomes reconstructed from metagenome data.

1.2.3 Ecological Roles of the Giant Viruses

Viruses are the most abudant biological entities on our planet (Chow & Suttle, 2015; Suttle, 2007). They regulate the composition of marine communities and are a major force behind biogeochemical cycles (Suttle, 2007). Giant virus could also regulate host populations though cell lysis (Tarutani et al., 2000). Cell lysis caused by viruses promotes the production of dissolved organic matter and accelerates the recycling of potentially growth-limiting nutrient elements (the 'viral shunt'). Viruses have also been proposed to drive particle aggregation and transfer into the deep sea via the release of sticky, carbon-rich viral lysate (the 'viral shuttle'). Although research on ecological roles of giant viruses is still largely limited, many evidence have suggested giant viruses to play critical roles in marine ecosystems by regulating host populations and may also contribute to the marine carbon pump (Kaneko et al., 2021).

Another potential path through which viruses could impact the environment is Auxiliary Metabolic Genes (AMGs) (Breitbart et al., 2007). AMGs are viral genes with high similarity to host homologues and can potentially alter the metabolic activities of hosts. Giant viruses have large gene repertoire that encode proteins with diverse functions, which potentially possess many AMGs. Some giant viruses (i.e., APMV, CroV, klosneuviruses and tupanviruses) encode several genes involved in protein translation (Abrahão et al., 2018; Fischer et al., 2010; Schulz et al., 2017). Additionally, giant viruses also encode other components for a relatively independent 'life', including carbohydrate metabolism genes, glycolysis and gluconeogenesis, tricarboxylic acid cycle (Moniruzzaman, Martinez-Gutierrez, et al., 2020). These findings suggest that giant viruses may have relatively independent 'life circulation'. Further, some evidence suggested giant viruses isolated from high latitude areas encode enzymes for the biosynthesis of unsaturated fatty acids, which may be part of a strategy to rewire the host fatty acid physiology (Rosenwasser et al., 2014). Together, giant viruses could have ecological impact in the environment through their function repertoire.

Endogenous Viral Elements (EVEs) are whole or large fragments of viral genomes integrated into host DNA, existing in giant viruses called giant EVEs (GEVEs). These GEVEs are usually found in green algae and can span hundreds to thousands of kilobases and contribute significantly to eukaryotic evolution. A 1.5 Mb continuous GEVE was identified in a fungus, *Rhizophagus irregularis* (Zhao et al., 2023). Such integrations of giant viruses found in eukaryotic genomes represent an often-overlooked pathway for introducing new genetic material, which can significantly influence the composition of those eukaryotic genomes (Moniruzzaman, Weinheimer, et al., 2020; Zhao et al., 2023). Moreover, studying giant virus signals in eukaryotic genomes also sheds light on unknown virus-host relationships, as oomycetes where no such viruses have yet been isolated (Hingamp et al., 2013). At an ecological scale, GEVEs potentially alter the ecological niches of their hosts by modifying the evolution of hosts, further emphasizing the ecological importance of viruses.

1.2.4 The Enigma of Giant Virus Ecology

Giant viruses have been shown to be widespread in the ocean, extending even to the frigid Arctic and Antarctic regions. A considerable proportion of unique OTUs were found specifically in the Arctic Ocean (Endo et al., 2020). However, the mechanisms by which these viruses form distinct communities in polar environments remain unknown. Given that polar regions are coldest habitats on earth and cellular organisms have diverse adaptation strategies, like distinct physiological or morphological, for survival, it is worth testing whether giant viruses have developed similar adaptive strategies.

Understanding the ecology of giant viruses involves not only the ambient environment but also their hosts. Given the challenges in experimentally isolating virus-host pairs, predicting hosts of giant viruses is crucial. Although considerable effort has been made to uncover interactions between giant viruses and potential hosts, the reliability of existing prediction approaches, like co-occurrence network methods, has yet to be quantitatively evaluated. This limitation restricts our comprehensive understanding of the ecology of giant viruses. Consequently, more methodological effort are necessary to evaluate and improve the accuracy of host predictions for giant viruses.

1.3 Methodology of giant virus studies

1.3.1 Cultivation-dependent Approaches

The first giant virus, APMV, was isolated through amoeba co-culture in 2003. During the last two decades, many candidate host cells, mainly amoeba, have been predominantly used as hosts for co-cultivating and discovering new giant viruses. The co-cultivation with amoeba has led to the isolation of closely related giant viruses capable of infecting this protist, such as orpheovirus (Andreani et al., 2018), medusavirus (Yoshikawa et al., 2019) and pandoravirus (Philippe et al., 2013). Given the role of giant viruses in pathogenizing protozoa, culture-based approaches of amoeba co-culture has become the gold standard for isolating giant viruses. In addition to viruses that infect amoebas, cultivation technique had been effectively used for isolating viruses of non-amoeba hosts, such as Paramecium bursaria chlorella virus 1 (Meints et al., 1981) and Haptolina ericina virus (Johannessen et al., 2015).

Even in the era of high-throughput sequencing, cultivation-dependent approaches maintain an irreplaceable role. 1) They provide trustable reference genomes that serve as seeds and form initial databases for in silico studies. 2) Cultivation-dependent techniques are effective and reliable for investigating the ecology and evolution of giant viruses, often serving as a cross-validation for computational methods. 3) Cultivation-dependent approaches make physiological experiments feasible by using isolated virus-host pairs, which yield deeper insights into their biological interactions and ecological functions.

1.3.2 Cultivation-independent Approaches

On the other hand, cultivation-independent approaches also offer valuable insights in giant virus studies. Metagenomics refers to perform sequencing genetic material recovered directly from environmental or clinical samples without isolation or cultivation. This approach allows for a comprehensive examination of microbial/viral communities in their natural environments. One cultivation-independent approach, read mapping-based techniques, is an approach mapping metagenomic reads to giant virus reference genomes or genes. Based on the mapping signals, viruses could be detected, and their environmental abundance could be estimated (Deeg et al., 2018; Hingamp et al., 2013). However, these approaches are largely dependent on the quality of the reference genome database and can generate false positive results due to the mis-mapping. The advantage of such approaches is that they are sensitive enough to detect viruses at low levels. Different from mapping-based approaches, marker gene surveys focus on performing a phylogenetic analysis of signature genes, like MCPs (Schulz et al., 2018) and DNApolB (Endo et al., 2020). Although generally less sensitive than read mapping, this method is less prone to errors as it employs phylogenetic analysis to confirm monophyly of clades.

In addition to all approached mentioned, genome-resolved metagenomics involves the reconstruction of MAGs, a technique particularly useful for studying giant viruses given their bacteria like large genomes. For other viruses, the typical absence of universal marker genes often leads to the neglect of viral bins[8] in most microorganism-centric metagenome projects. However,

[8] Bin/Binning: Grouping reads or contigs and assigning them to individual viral genomes.

using certain marker genes, several studies have employed custom workflows to identify giant virus MAGs (Bäckström et al., 2019; Moniruzzaman, Martinez-Gutierrez, et al., 2020; Schulz et al., 2020). Despite the presence of some limitations, such as incompleteness, MAGs still provide valuable insights into the evolutionary history and functional repertoire inference of giant viruses. These studies usually rely on a set of universal conserved genes or lineage-specific core orthologs. However, it is the difficult to explore novel giant virus with seldom or variated marker genes (Schulz et al., 2020). Such a common constraint primarily comes from a over dependence on a known set of core genes, which may overlook potential clues for exploring natural diversity of giant viruses.

1.4 Aims of the present study

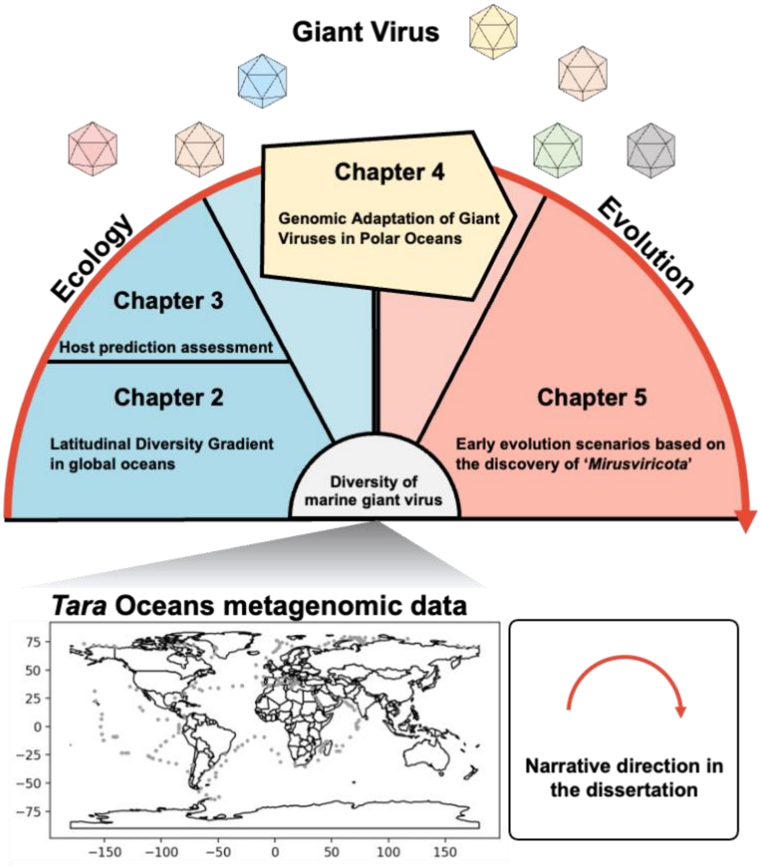

Giant Virus
Ecology
Evolution
Chapter 4
Genomic Adaptation of Giant Viruses in Polar Oceans
Chapter 3
Host prediction assessment
Chapter 2
Latitudinal Diversity Gradient in global oceans
Chapter 5
Early evolution scenarios based on the discovery of '*Mirusviricota*'
Diversity of marine giant virus
Tara Oceans metagenomic data
75
50
25
0
−25
−50
−75
−150
−100
−50
0
50
100
150
Narrative direction in the dissertation

Fig. 2-3 Overview of the four research chapters in this dissertation

This study used metagenomic data from the *Tara* Oceans project to explore giant virus populations on a global scale. Initially, the biogeography of marine giant viruses was investigated (Chapter 2). Then the effectiveness of co-occurrence network-based predictions in identifying interactions between giant viruses and potential hosts was assessed (Chapter 3). Using the same network methodology, a global virus-eukaryote interaction network was constructed, revealing a distinct polar boundary. Following this, a hypothesis that polar environment drives viral evolution was proposed (Chapter 4). Finally, the discovery of '*Mirusviricota*' was presented, offering critical insights into the early evolution of giant viruses (Chapter 5).

Taking advantage of the recent large-scale marine metagenomics census, I aimed to comprehensively explore the ecology and evolution of marine giant viruses using global ocean data (i.e., the *Tara* Oceans), the findings of which are encapsulated in my Ph.D. dissertation. Generally, four major findings are organized into separate chapters in this dissertation (Fig. 1-3).

In Chapter 2, using the abundance profile of a newly generated genome database, the global distribution of marine giant viruses was investigated. The results revealed diverse latitudinal gradient patterns of giant viruses variated across sampling size fractions and viral lineages. These variations likely represent different host ranges for giant viruses, as supported by polar host prediction results based on co-occurrence analyses. Co-occurrence networks have commonly been used for predicting hosts of giant viruses, however, the performance has never been evaluated. To address this, Chapter 3 quantitatively assessed the effectiveness of this approach for predicting interacting partners of giant viruses in marine environments. Chapter 4 introduces the global interaction network of giant viruses and eukaryotic plankton genomes, identifying a distinct boundary between genome communities in polar and nonpolar environments. Further, a hypothesis was proposed that recurrent evolutionary adaptations of giant viruses across the polar boundary are likely driven by alterations of viral gene repertoire. In addition to viral evolution driven by polar environments, Chapter 5 further discusses the early evolution of giant viruses. '*Mirusviricota*', a group of plankton-infecting DNA giant viruses with remarkable chimeric attributes, has been discovered to be prevalent in the ocean, providing crucial missing links in the ancient evolution of both herpesviruses and giant viruses.

Chapter 2 Latitudinal Diversity Gradient of Marine Giant Viruses

2.1 Abstract

Giant viruses play important roles in marine ecosystems by regulating host populations and may also contribute to the marine carbon pump. Therefore, a comprehensive understanding of the biogeography of marine giant viruses is necessary. Utilizing the GOEV database[9], this chapter delves deeply into the biogeography of giant virus genomes. A total of 1,380 giant virus genomes were detected in 928 samples across diverse oceanic environments. Six main groups of marine giant viruses are identified and explored, offering a nuanced view into their genome sizes and distribution patterns in various oceanic size fractions. Giant viruses are widespread in the global ocean, either as free virions or within host cells. Across different size fractions and taxonomic groups, various diversity gradient patterns of viruses were observed. Furthermore, temperature was identified as the predominant factor among variables impacting viral latitudinal distribution. Additionally, several eukaryotes, including diatom as a potential polar specific host, were predicted to be associated with giant viruses, suggesting a broader infection spectrum of giant viruses in the ocean. The study in this chapter provides insights into the intricate relationships among marine viral distributions, further providing the possibility to study their genetic characteristics, and environmental variables.

2.2 Introduction

[9] GOEV database: Global Ocean Eukaryotic Viral database.

Giant viruses, including two viral phyla *Nucleocytoviricota* and '*Mirusviricota*' (defined in 1.1.1 and mainly discussed in Chapter 5), form a group of dsDNA viruses with large genomes and diverse gene repertoires that infect multiple eukaryotic species[10] (Gaïa et al., 2023; Sun et al., 2020). Several studies have demonstrated a high diversity of members of *Nucleocytoviricota*. Remarkably, one *Nucleocytoviricota* lineage, the *Megaviridae* (currently *Imitervirales*), surpasses even the diversity of both marine bacteria and archaea domains. Such evaluation was performed using two marker genes, RNAPa and RNAPb (Mihara et al., 2018). From just a few liters of seawater collected from Osaka Bay, over 5,000 different OTUs of *Imitervirales* were identified, though only 20 of these were known isolations (Li et al., 2018). A comprehensive biogeography study on giant viruses, employing a conserved single-copy gene marker, DNApolB, has demonstrated that *Nucleocytoviricota* viruses are abundant and diverse in the global ocean (Endo et al., 2020). Although gene level makers have shede the light of viral diversity, the information of marine giant virus genomes was not accessible for a long time. Advances in high-throughput sequencing and bioinformatics have facilitated the assembly of thousands of draft genomes or MAGs of *Nucleocytoviricota* from environmental samples (Bäckström et al., 2019; Moniruzzaman, Martinez-Gutierrez, et al., 2020; Schulz et al., 2020). A genome-resolved worldwide distribution and diversity of marine giant viruses became feasible, particularly after the construction of the GOEV database.

Meanwhile, *Nucleocytoviricota* have been found to infect a growing number of marine eukaryotic organisms, not only phytoplankton like haptophytes, chlorophytes, and dinoflagellates but also non-photosynthetic eukaryotes such as bicosoecids and choanoflagellates (See 1.2.1). This broad host range emphasizes their ecological impact on marine ecosystems, where they exert top-down effects on diverse eukaryotic communities. Despite the known knowledge, the extensive phylogenetic

[10] '*Mirusviricota*' is speculated to infect marine picoplankton and nanoplankton.

diversity of giant viruses indicates that we still underestimate their entire host spectrum, which leaves the understanding of these relationships still in far from comprehensive.

In this chapter, the genome-resolved global biogeography of giant viruses, both *Nucleocytoviricota* and '*Mirusviricota*', was discussed. The findings were based on metagenomic data sourced from the *Tara* Oceans project and the GOEV database. This comprehensive dataset spans diverse geographic regions, from pole to pole. It is aimed to uncover diversity patterns of marine giant viruses and predict the interaction virus-eukaryote pairs. This information can deepen our understanding of the distribution and ecological roles of giant viruses across various marine environments.

2.3 Methods

2.3.1 Generation of the GOEV database

Details of the workflow for creating the GOEV database and other method sections are available in the original paper (Gaïa et al., 2023). In brief, 928 *Tara* Oceans metagenomes were co-assembled based on 11 geographical coordinates, employing MEGAHIT (D. Li et al., 2015) and Anvi'o software (Eren et al., 2015). Produced 78 million contigs then underwent constrained automatic binning through CONCOCT (Alneberg et al., 2014), resulting in 2,550 metagenomic blocks (bins). DNA-dependent RNA polymerase B subunit (RNAPb) genes in 2,550 blocks were identified through HMMER (Eddy, 2011) and CD-HIT (W. Li & Godzik, 2006). Then metagenomic blocks containing interesting RNAPb genes, specifically potential viral ones, were manually binned and curated. To ensure the GOEV database have comprehensive representation, MAGs from two previous surveys (Moniruzzaman, Martinez-Gutierrez, et al., 2020; Schulz et al., 2020) and reference *Nucleocytoviricota* genomes were incorporated into the final database. After redundancy removal, the final database consisted of 1,593 marine MAGs and 224 reference genomes. The taxonomic classification of the GOEV database was defined by a phylogenomic analysis using the concatenated genes of RNApolA, RNApolB, DNApolB, and TFIIS.

2.3.2 Reads mapping

A total of 928 *Tara* Oceans metagenomic reads were mapped to the genomes and genes, respectively, to ascertain the mean coverage using BWA with a minimum identity of 90%. 1380 detected viruses were classified into six main taxonomic groups: five orders (i.e., *Algavirales*, *Asfuvirales*, *Imitervirales*, *Pandoravirales*, and *Pimascovirales*) and the newly discovered phylum, *Mirusviricota*. Metagenomes of two layers two depths (Surface and Deep Chlorophyll Maximum) and six different size fractions were used in this study: 0.22–1.6 μm or 0.22–3.0 μm ('Pico'), 0.8–5 μm ('Piconano'), 5–20 μm ('Nano'), 20–200 μm ('Micro'), 200–2,000 μm ('Macro'), and 0.8–2,000 μm ('Broad'). Size fraction below 0.22 μm was excluded due to its low relative abundance and high overlap with the Pico size fraction. MAGs results were retained if at least 25% of the viral genome was covered by reads. Relative abundance of a giant virus in each sample was calculated in Reads Per Kilobase per Million mapped reads (RPKM). Subsampling was not performed prior to biodiversity calculation because, as the pre-study showed, sequencing depths of metagenomes did not significantly influence the abundance pattern or Shannon's diversity of giant virus communities.

2.3.3 Ecological analyses

Most of analyses on viral ecology were conducted using R. The 'vegan' package was utilized to calculate various metrics (Oksanen et al., 2018), including richness, Shannon's index, and Pielou's evenness for each collected sample. To examine compositional variations among samples, Bray-Curtis dissimilarity was used for non-metric multidimensional scaling (NMDS) ordination. The study determined statistical significance between different sample groups, categorized by size fractions and biomes, using Analysis of Similarity (ANOSIM). This was performed with 9,999 permutations and a P-value threshold of 0.01 was set for significance. Data visualization was executed using the 'ggplot2' and 'rgdal' packages (Bivand et al., 2018), which facilitated the creation of various plots and maps of the sampling locations.

2.3.4 Niche calculation

Niche calssfication was performed for size fraction, biomes, and environmental variables using abundance profiles from *Tara* Oceans metagenomes. Details could be checked in the original paper (Meng et al., 2023). Size index served as an indicator of the distribution preference along sampling size of viruses. A larger size index for a virus suggests that its signal was more likely detected in association with larger-bodied organisms. Further methods also included the assignment of genomes to ‘Polar’ or ‘Nonpolar’ biome niches based on mapping signals and statistical significance determination via the Wilcoxon rank-sum test, considering the Benjamini-Hochberg correction. A robust ecological optimum[11] was calculated for MAGs, reflecting the optimal physical conditions concerning different environmental variables. Proportions of RPKM were used in the formation of a weighted vector populated with environmental values, which then facilitated the determination of ecological optimum and tolerance range, implementing a methodological safeguard of requiring a minimum of 10 observations and a 30% non-NA value threshold to avoid deriving misleading ecological optima and tolerance ranges.

2.4 Results

[11] Robust ecological optima represent niche values, and tolerance ranges (realized niche widths) were also calculated.

2.4.1 Biogeography of giant virus genomes

A comprehensive biogeographic analysis of giant virus genomes based on the GOEV database was refined to contain 1,817 genomes. Analyzing the abundance profiles of these 1,817 genomes across *Tara* Oceans samples, 1,380 viral genomes exhibited signals in at least one of 928 samples (Fig 2-1). Despite the plateau observed in rarefaction analyses that amalgamated all *Tara* Oceans samples, the study noted that genomes in micro- and macro-size fractions remained under-sampled (Meng et al., 2023). The detected viruses were taxonomically partitioned into six principal groups: *Algavirales*, *Asfuvirales*, *Imitervirales*, *Pandoravirales*, *Pimascovirales*, and '*Mirusviricota*', each with a unique genomic size and abundance range. Notably, the *Imitervirales* constituted the largest group (i.e., most MAGs/genomes), with some of its members being widely distributed, albeit at low abundance. In

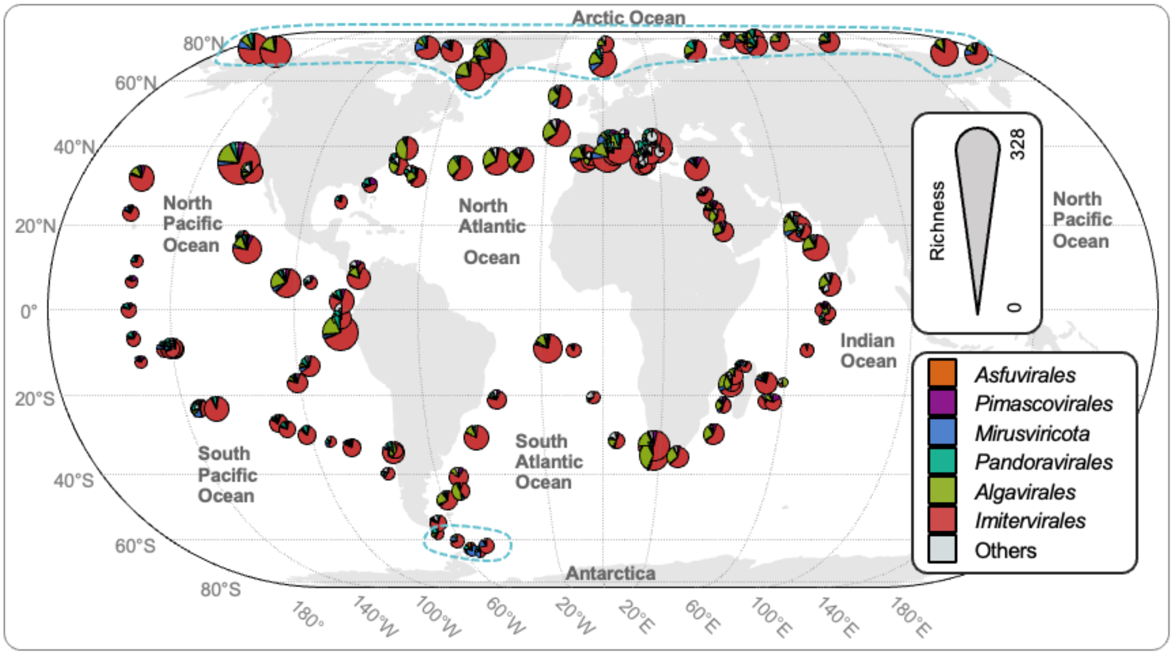


Fig. 2-1 World map demonstrating distribution of marine giant viruses.

Each pie chart represents the richness of viral communities, consisting of six main groups, and the size of the pie chart is proportional to the total richness at the station. The richness of two depths and different size fractions (See 2.3.2) of one sampling station are integrated into one pie chart. Polar samples are highlighted using dashed lines. This figure has been published and modified under the CC-BY license from the paper by L. Meng et al., 2023.

contrast, the second-largest group, algaviruses, exhibited a significantly higher cumulative RPKM compared to other groups but were found in fewer samples than Imitervirales, thereby revealing intriguing patterns and disparities in the biogeography and abundance of these giant virus genomes in oceanic environments (Fig. 2-2).

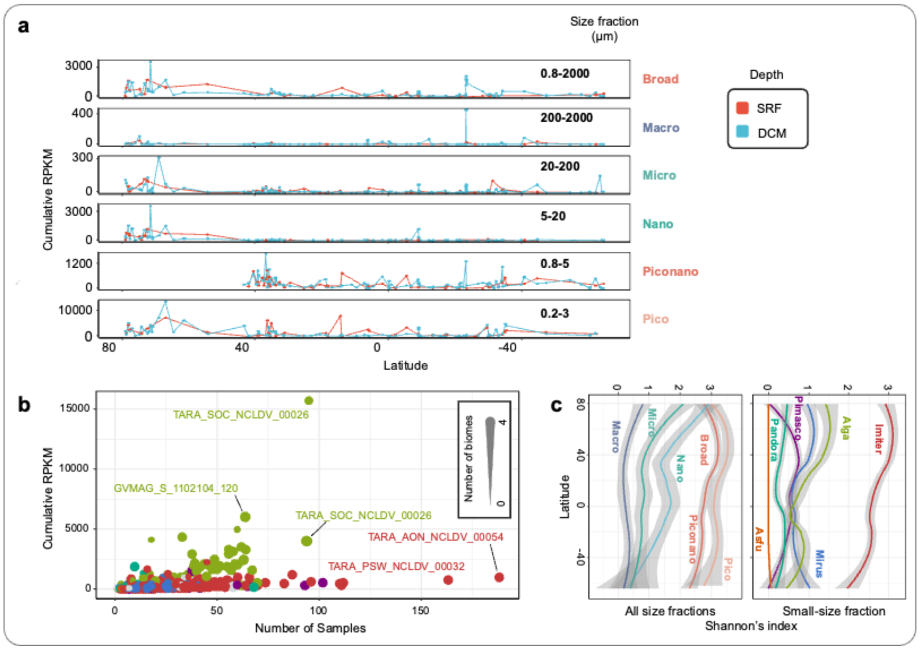


Fig. 2-2 Abundance and diversity of giant virus.

a, Total cumulated abundance in each sample along the latitude. b, Cumulative RPKM of individual viral genomes versus the number of samples in which the respective genomes were observed. c, Locally estimated scatterplot smoothing plots of the latitudinal distributions of viral diversity (Shannon's index). The left panel presents the total diversity of all giant viruses along a latitudinal gradient in different size fractions. The right panel shows the diversity of communities of six main groups in the small-size fractions (namely Pico, Piconano, and Broad size fractions). This figure has been published and modified under the CC-BY license from the paper by L. Meng et al., 2023.

2.4.2 Distribution of giant viruses across size fractions

Giant viruses were detected in various size fractions, across numerous stations (Fig. 2-2). Size index values demonstrated a broad distribution across size fractions for giant viruses, as opposed to individual eukaryotic taxa (Fig. 2-3). Viral signals originating from larger size fractions (e.g., >0.8

μm) could be attributed to viral genomes encapsulated within their host cells, while those emanating from smaller size fractions (e.g., 0.2–3.0 μm) could stem from either free virions or host-encased viral genomes. Infection stage categories were allocated to viral genomes, designated as either ‘virion’ (0.2–3.0 μm) or ‘cellular’ (>0.8 μm), guided by the distribution of RPKM across size fractions. Through this categorization, it was elucidated that 15% of viruses (n = 211) predominated in the ‘virion’ category, and 8% (n = 111) were predominant in the ‘cellular’ category. The proportions of *Imitervirales* and *Algavirales* were comparatively low, while *Pimascovirales*, Pandoravirales, *Asfuvirales*, and ‘*Mirusviricota*’ manifested higher proportions in the ‘cellular’ category relative to the ‘virion’ category, suggesting variable host size spectra or different infection cycles across virus lineages.

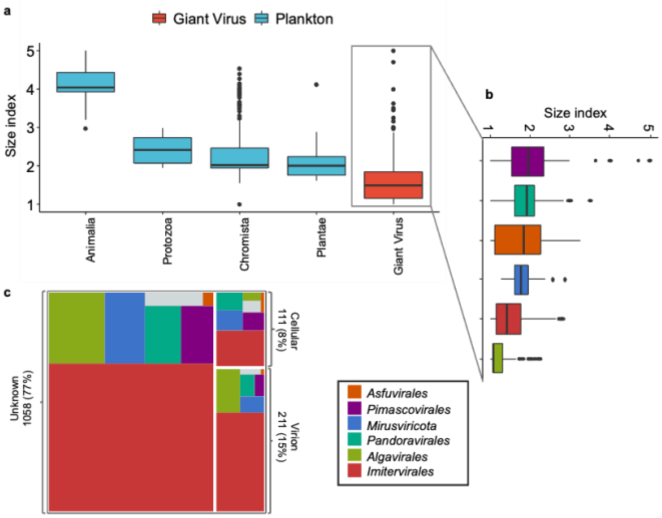


Fig. 2-3 Size distribution of viruses.

a, Boxplots for the size indices of eukaryotic kingdoms and giant viruses. b, Boxplots for the size indices of six virus main groups. c, Treemap diagram showing the number of giant viruses assigned to “virion” or “cellular” size categories. Colours indicate the main groups. This figure has been published and modified under the CC-BY license from the paper by L. Meng et al., 2023.

2.4.3 Temperature-related latitudinal distribution gradual

Latitudinal diversity gradients are typically characterized by relatively low biodiversity in polar regions and higher biodiversity in temperate zones. This pattern is prevalent across a broad range of marine microorganisms. Previous research has identified a similar latitudinal diversity gradient for giant viruses in small size fractions (Ibarbalz et al., 2019a). However, typical latitudinal diversity gradients were not observed for prokaryotic dsDNA viruses or RNA viruses. In our study, varying diversity gradient patterns among viruses of different size fractions and main taxonomic groups was observed (Fig. 2-2). Specifically, a peak in diversity at mid-latitudes was noted in smaller-size virus fractions (namely Pico, Piconano and Broad size fractions), while hotspots of viral diversity were found in the Arctic regions within larger-size fractions (namely Nano, Micro and Macro size fractions). The presence of Arctic diversity hotspots for certain viruses, such as those in larger-size fractions and mirusviruses, might be attributable to their host ranges.

For each genome of giant viruses, a robust ecological optimum was calculated, encompassing variables such as temperature, salinity, latitude, ChlorophyllA, Si, NO2, and PO4. Based on spearman analysis, the strongest correlation was observed between temperature and latitude optima among viral genomes (Spearman rho = -0.886), indicating the latitude niche of giant viruses could be explained by temperature. Strong correlations were also observed between latitude and salinity (rho=-0.432), and latitude and ChlorophyllA (rho=0.579). None of the other variables displayed an absolute rho value exceeding 0.1. Furthermore, neither salinity nor ChlorophyllA distinctly influenced the virus-eukaryote network as temperature impacted. So temperature was determined to be the most fitting variable for elucidating the latitudinal distribution of viruses in this investigation.

2.4.4 Potential Hosts

Using the phylogeny-informed Taxon Interaction Mapper (TIM) method[12] (See 3.3.3), connections between viruses and eukaryotes were established through a clade-to-clade relationship and examining whether leaves (i.e., viral genomes) under a node of the virus tree are enriched with a specific predicted host group (Fig. 2-3) (details of this analysis will be described in the Chapter 3). TIM designated five predicted host taxa to 34 viral clades, encapsulating 6.38% of total viral

[12] TIM: Filtering associations between viruses and potential hosts based on their taxonomies.

genomes. The predictions, which included recognized virus-host relationships such as *Mesomimiviridae* with Phaeocystales and Pelagomonadales, and Prasinovirus with Mamiellales. Furthermore, the exploration of GEVEs, widespread across various eukaryotes, illuminated the impacts of giant viruses on host genome evolution. Insertions of genomes of giant viruses and their satellite viruses (virophages) were detected in marine eukaryotic genomes. Among the five taxa of predicted viral hosts, the diatom order Chaetocerotales presented the most significant number of insertion signals of both giant viruses and virophages (Fig. 2-3), suggesting infection of dsDNA viruses in Chaetocerotales. Genomes of two Chaetocerotales isolates also exhibited a substantial level of GEVE-like signals. Notably, diatoms of Chaetocerotales, being abundant and varied in both the Arctic and Southern Oceans, and with genomes in the marine eukaryotic database exclusively distributed in high latitudinal polar oceans. Host predictions also reflect the diversity of giant viruses in high-latitude regions, despite the *in silico* prediction being constrained by the lack of evidence of host-virus interaction and the current absence of genomes of polar *Chaetocerotales* isolates.

2.4 Discussion

The analyses of giant virus biogeography in this chapter provide insights into studies of viral ecology, revealing variations in their community composition in response to diverse ecological factors. However, most previous metagenomic studies of giant viruses predominantly used data extracted from a pico-size fraction (0.2–3.0 μm). The larger size fractions, which usually include the hosts of giant viruses, were overlooked due to fact that viruses are difficult to be detected with low abundance (Hingamp et al., 2013). The depth of datasets provided by *Tara* Oceans and the GOEV database enabled us in this study to evaluate the abundance of giant viruses in size fractions

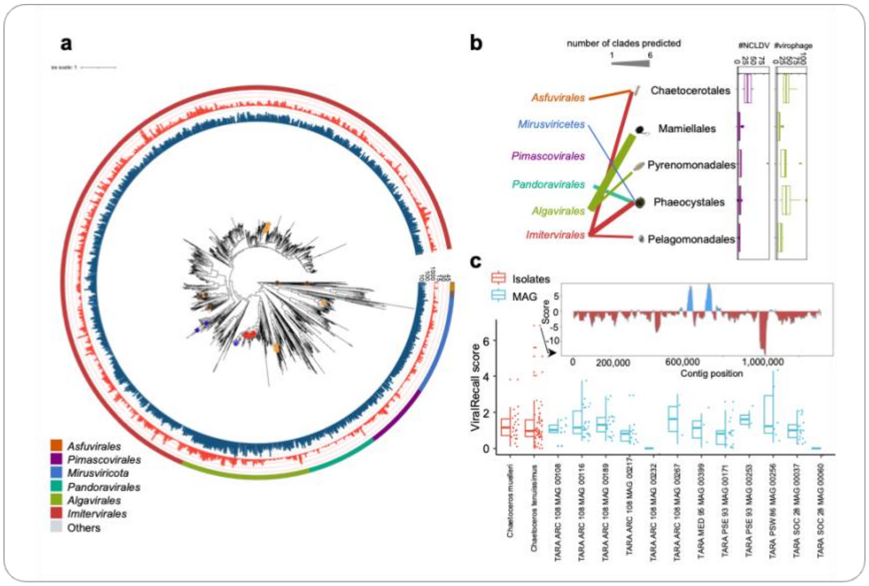


Fig. 2-4 Host prediction of viruses.

a, Phylogenomic tree of giant viruses with circles representing putative host predicted by TIM (See 3.3.3). b, line colours represent the six main groups and line widths are proportional to the number of clades predicted to the associated hosts. Boxplots at right show the number of detected viral signals in eukaryotic genomes. c, ViralRecall scores of 12 Chaetocerotales MAGs and 2 isolates. An example of *Chaetoceros tenuissimus* contig was given above. This figure has been published and modified under the CC-BY license from the paper by L. Meng et al., 2023.

extending up to the meso-fraction (200–2000 μm). For the first time, a comprehensive genome-resolved survey revealed giant virus distribution patterns and variations across size fractions, taxonomic groups and oceanic regions. One observation was the notable prevalence of imiterviruses. However, their widespread distribution did not make them into high abundance, contrasting with algaviruses which, despite being found in significantly fewer samples, exhibited a notably higher cumulative RPKM. Furthermore, the distinguishing of viral genomes into 'virion' and 'cellular' categories hints towards the potential of viral replication strategies and life cycles. For example, the proportions of *Imitervirales* and *Algavirales* genomes assigned to the 'cellular' category compared with those to the 'virion' category were relatively low. On the contrary, such proportions of *Pimascovirales*, *Pandoravirales*, *Asfuvirales* and *Mirusviricota* were relatively high, implying different host size ranges or infection cycles for different groups of viruses.

Temperature is more important than other factors in determining the virus communities, which is consistent with previous marine surveys for bacteria (Sunagawa et al., 2015) and bacterial phages (Gregory et al., 2019). This raises crucial questions about the impact of global temperature changes on tiny life forms in ecosystems. Latitudinal diversity gradients are characterized by relatively low polar and high temperate biodiversity (Hillebrand, 2004) and are widespread across all ranges of marine microorganisms. Previous studies revealed a similar latitudinal diversity gradient for giant viruses (Ibarbalz et al., 2019a), but not for prokaryotic dsDNA viruses (Gregory et al., 2019) and RNA viruses (Dominguez-Huerta et al., 2022). In this study, various diversity gradient patterns were observed among viruses of different size fractions and main taxonomic groups. The reasons underlying the Arctic diversity hotspots for some viruses (e.g., viruses in large-size fractions and mirusviruses) may reflect their host ranges as previously suggested (Ibarbalz et al., 2019b). Although not included in the dissertation, the eukaryotic nodes (i.e., potential hosts) associated with viruses exhibited a pattern that deviates from the typical diversity gradient trend, displaying increasing diversity towards higher latitudinal regions (Meng et al., 2023). Further explorations into the intricate interplays of these environmental parameters, viral biogeography, and host interactions are warranted, may unveil about the overarching mechanisms governing microbial and viral ecologies in our oceans.

Chapter 3 Quantitative Assessment of *Nucleocytoviricota* Host Prediction

3.1 Abstract

Members of viral phylum *Nucleocytoviricota* infect a broad range of eukaryotic hosts. However, the knowledge of their hosts is limited because only a few viruses have been isolated so far. Taking advantage of the recent large-scale marine metagenomics census, *in silico* host prediction approaches

are expected to fill the gap and further expand our knowledge of virus–host relationships for unknown *Nucleocytoviricota*. In this Chapter, marker-based co-occurrence networks of *Nucleocytoviricota* and eukaryotic taxa were recruited to predict virus–host interactions. Using the positive likelihood ratio to assess the performance of host prediction for *Nucleocytoviricota*, several co-occurrence approaches are benchmarked and demonstrated an increase in the odds ratio of predicting true positive relationships four-fold compared with random host predictions. A phylogeny-informed filtering method, Taxon Interaction Mapper (TIM), further refines host predictions from high-dimensional co-occurrence networks. The prediction performance is improved by twelve-fold. Finally, networks of giant viruses and virophages are inferred to corroborate that co-occurrence approaches are effective for predicting interacting partners of *Nucleocytoviricota* in marine environments.

3.2 Introduction

As introduced in the Chapter 1, *Nucleocytoviricota* is a viral phylum infecting a wide spectrum of eukaryotic hosts. One of the most critical issues hindering our understanding on the ecology and evolution of giant viruses is that laboratory isolations only represent a small fraction of the interactions present in the ocean. Earlier studies have illustrated that *Nucleocytoviricota* potentially have the capacity to infect more diverse hosts than presently known. One evidence is the widespread of gene transfer analyses between viruses and eukaryotes (Gallot-Lavallée & Blanc, 2017; Schulz et al., 2020). Such an interaction/coevolution may have initiated even before LECA[13] (Guglielmini et al., 2019). This diversification time was supported by a recent study that some giant viruses encode

[13] LECA: Last Eukaryotic Common Ancestor, which the oldest fossil evidence is about 2 billion years ago.

viractins, which could have been procured from proto-eukaryotes and possibly reintroduced into the pre-LECA eukaryotic lineage (Da Cunha et al., 2022). Together, these findings filled a knowledge gap in *Nucleocytoviricota* biology and host diversity, however, more efforts are needed to illuminate the poorly understood virus–host relationships and the largely mysterious *Nucleocytoviricota* world.

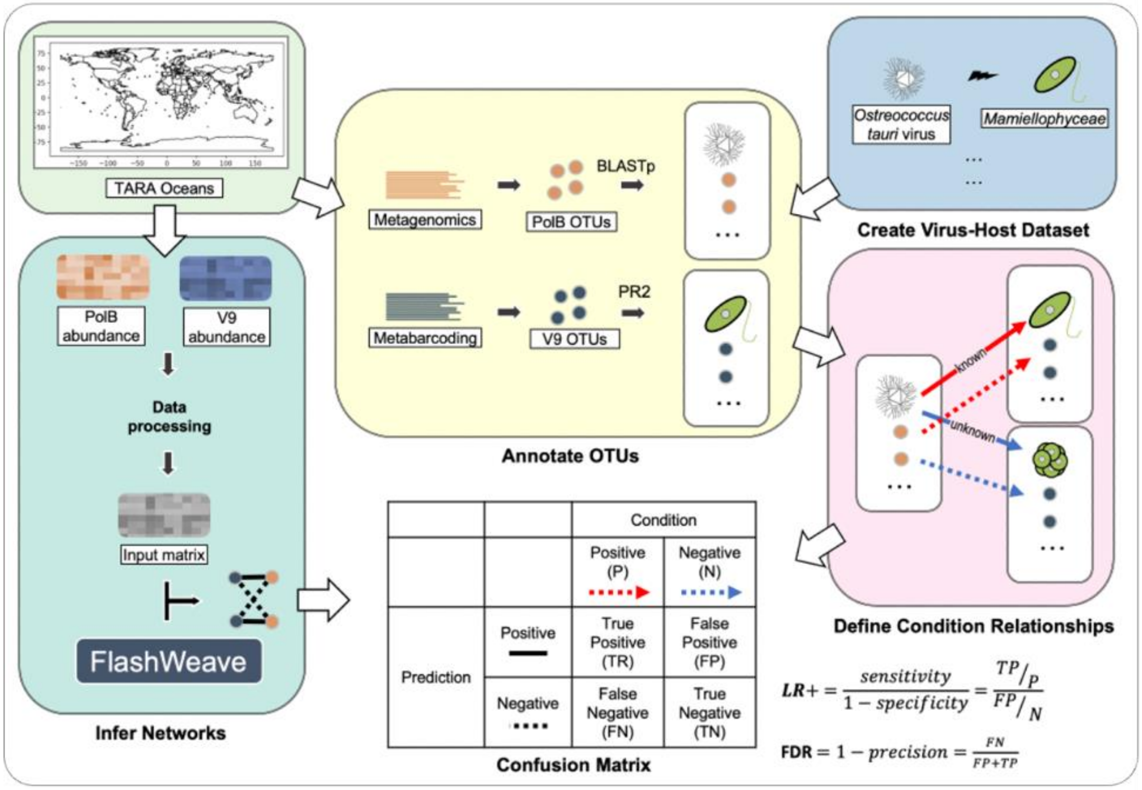
Such efforts include the cultivation-based methods, like co-culture with amoeba. Beyond that, other culture-independent experimental approaches, including a high-throughput host-virus identification using cell sorting (Needham et al., 2019). Metagenomics, especially proficient at assessing a large faction of ecosystem diversity, has been progressively used to explore *Nucleocytoviricota* host range. Some comparative genomics analyses, like the identification of horizontal gene transfer predictions, have significantly broadened the host range of *Nucleocytoviricota* (Schulz et al., 2020). Investigating GEVE in eukaryotic lineages can also prove invaluable for inferring species-specific virus–host associations (Moniruzzaman et al., 2022; Moniruzzaman, Weinheimer, et al., 2020).

Abundance-based analyses are commonly used for predicting host-virus relationships (Hingamp et al., 2013; Moniruzzaman et al., 2017), largely because viruses can only thrive where their hosts are present. This approach has also been extended to predict associations between *Nucleocytoviricota* and their parasites (virophages) (Roux et al., 2017). However, the reliability of the coexistence-based methods has been questioned (Coenen & Weitz, 2018), particularly due to some natural features, like the potential time lags between the dynamics of viruses and their hosts (Martínez Martínez et al., 2007; Tomaru et al., 2004). The effectiveness of co-occurrence network methods in predicting hosts for *Nucleocytoviricota* has not been quantitatively assessed, which hampers the broader application of these methods. Therefore, specialized techniques are needed to accurately test and improve the performance of co-occurrence-based predictions for *Nucleocytoviricota* host identification.

To address the problems given above, this Chapter focuses on the prediction of virus–host relationships between *Nucleocytoviricota* and eukaryotes by constructing co-occurrence networks using different methods. To quantitatively assess the performance of network-based host prediction, the positive likelihood ratio (LR+) using reference data for known virus–host relationships was recruited. Further, a phylogeny-informed filtering method was performed to refine the predictions from high dimensional complexity of networks.

3.3 Methods

3.3.1 Metagenomic and metabarcoding data



		Condition	
		Positive (P)	Negative (N)
Prediction	Positive	True Positive (TR)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

$$LR+ = \frac{sensitivity}{1 - specificity} = \frac{TP/P}{FP/N}$$

$$FDR = 1 - precision = \frac{FN}{FP+TP}$$

Fig. 3-1 Overall workflow for inferring co-occurrence networks and quantitative assessment.

This figure shows the workflow used in this study. The definition of the confusion matrix for quantitative assessment is shown in the table. The LR+ and FDR equations are given at the lower right corner of the plot. This figure has been published and modified under the CC-BY license from the paper by L. Meng et al., 2021.

The microbial metagenomic and eukaryotic metabarcoding data utilized in this study were initially generated from plankton samples collected by the *Tara* Oceans expedition. Given that our research demands paired metagenomic and metabarcoding datasets, data originated from the euphotic zone samples, specifically those from the surface (SRF) and Deep Chlorophyll Maximum (DCM) layers. A simple schematic plot was given in Fig. 3-1. DNApolB served as the marker gene for *Nucleocytoviricota*. A comprehensive total of 6818 DNApolB OTUs were extracted from the metagenomic datasets, notably the second version of the Ocean Microbial Reference Gene Catalog (OM-RGC V2), by pplacer. DNApolB sequences were assigned into seven families (*Mimiviridae*, *Phycodnaviridae*, *Marseilleviridae*, *Ascoviridae*, *Iridoviridae*, *Asfarviridae*, and *Poxviridae*) and two additional viral groups ('Medusavirus' and 'Pithovirus') based on the taxonomic system in 2019. Regarding eukaryotes, metabarcoding data which target the 18S ribosomal RNA gene hypervariable V9 region (V9) were used. Taxonomic annotation of the eukaryotic metabarcoding data was previously conducted by the *Tara* Oceans consortium using an extensive V9_PR2 reference database, stemming from the original Protist Ribosomal Reference (PR2) database. A Centred Log-Ratio (clr) transformation was performed after assigning a pseudo-count of one to all data entries and filtered out MAGs observed in fewer than three samples. The network was reconstructed using FlashWeave software. The details of network inference could be referred to the original papers (Meng et al., 2021).

3.3.2 Network validation

A database of existing *Nucleocytoviricota* host pairs was generated to evaluate virus–host associations. In brief, 69 known virus–host relationships for *Nucleocytoviricota* were manually compiled for the validation, annotating eukaryotic taxonomic groups at the 'Major lineages' level in the updated PR2 database. Utilizing 'Major lineages' was pivotal due to known virus-host relationship deficiencies and their encompassing representation of eukaryotic V9 diversity. Subsequent BLASTp searches from *Tara* Oceans PolB sequences against the virus reference database defined metagenomic DNApolB groups with a 65%

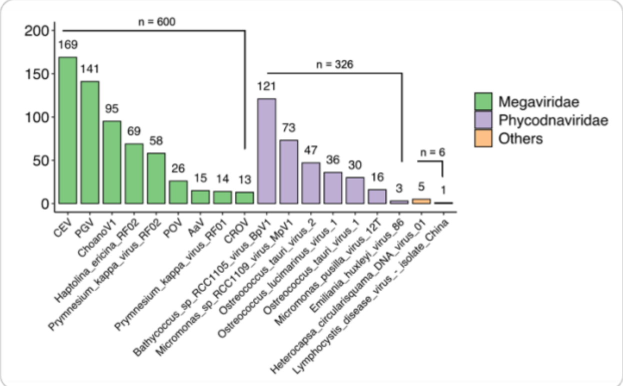


Fig. 3-2 Number of environmental PolB OTUs recruited in the network validation.

The number of marine viral OTUs (N = 932) were grouped to reference viral DNApolB sequences (N = 69). This figure has been published and modified under the CC-BY license from the paper by L. Meng et al., 2021.

sequence identity threshold (tested using the reference data), chosen for its ability to better discern virus host infections within major lineages and provide a superior LR+ (Fig. 3-2).

The Positive Likelihood Ratio (LR+) was employed to estimate predictive accuracy, where an LR+ close to 1 indicates performance akin to random prediction (equation is given in Fig. 3-1). From detected polB OTU and V9 OTU associations, only the top positive or negative associations were retained. Additionally, the False Discovery Rate (FDR) was calculated as a secondary assessment. To compare among five size fractions, abundances in overlapping samples of various sizes (0.8–5 μm, 5–20 μm, 20–180 μm, and 180–2000 μm) were used, ensuring comparable sample numbers across fractions and reducing bias in network topology.

3.3.3 Phylogeny-guided filtering of host predictions and its assessment

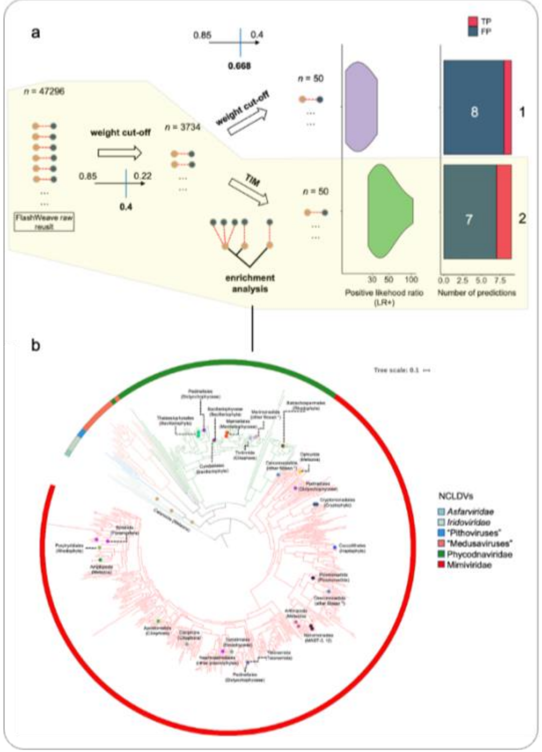


Fig. 3-3 Filtration of FlashWeave results. The process of the filtration of co-occurrence associations using TIM. The number of retained *polB* –V9 pairs are given by *n*, the performance between TIM filtration and a further weight cut-off of 0.668 were compared. The process using TIM is shown in yellow. (B) Phylogenetic tree of viruses and corresponding TIM-based predicted eukaryotic host groups. Predicted hosts were shown with colored circles. This figure has been published and modified under the CC-BY license from the paper by L. Meng et al., 2021.

Host predictions were refined by TIM based on co-occurrence networks. TIM operates on the assumption that viruses and hosts that are evolutionarily related tend to interact, extracting the most plausible virus-host associations from co-occurrence networks. To operate, TIM needs a virus phylogenetic tree and a set of virus-eukaryote connections, then it assesses whether leaves (viral OTUs) under a virus tree node are enriched with a particular predicted eukaryotic group compared to the rest of the tree using Fisher's exact test and Benjamini–Hochberg adjustment (Fig. 3-3).

In practice, a viral phylogenetic tree was reconstructed by excluding all DNApolB sequences absent in the FlashWeave network associations and filtering remaining sequences by amino acid sequence length (≥ 500 aa). Protein alignment was executed using MAFFT-linsi, removing 18 sequences manually due to misalignment with other DNApolB sequences. A total of 501 DNApolB sequences were utilized to construct a maximum likelihood phylogenetic tree with FastTree. Subsequently, the DNApolB–V9 associations were mapped on the tree to gauge the significance of the enrichment of particular associations via TIM, which provided a list of nodes in the viral tree and associated NCBI taxonomies of eukaryotes showing significant enrichment in the leaves under the nodes. The TIM result, visualized with iTOL, was transformed into a network where nodes correlate to major eukaryotic lineages, and network edge weight was determined by the number of tree nodes in each viral family subtree enriched with a specific major eukaryotic lineage. Visualization of the network was achieved with Cytoscape using a prefuse force-directed layout. To evaluate TIM's efficacy in improving prediction, associations predicted by TIM were extracted and their performance compared with raw and weight cut-off results.

3.4 Results

3.4.1 Global Virus-Eukaryote Co-Occurrence Networks

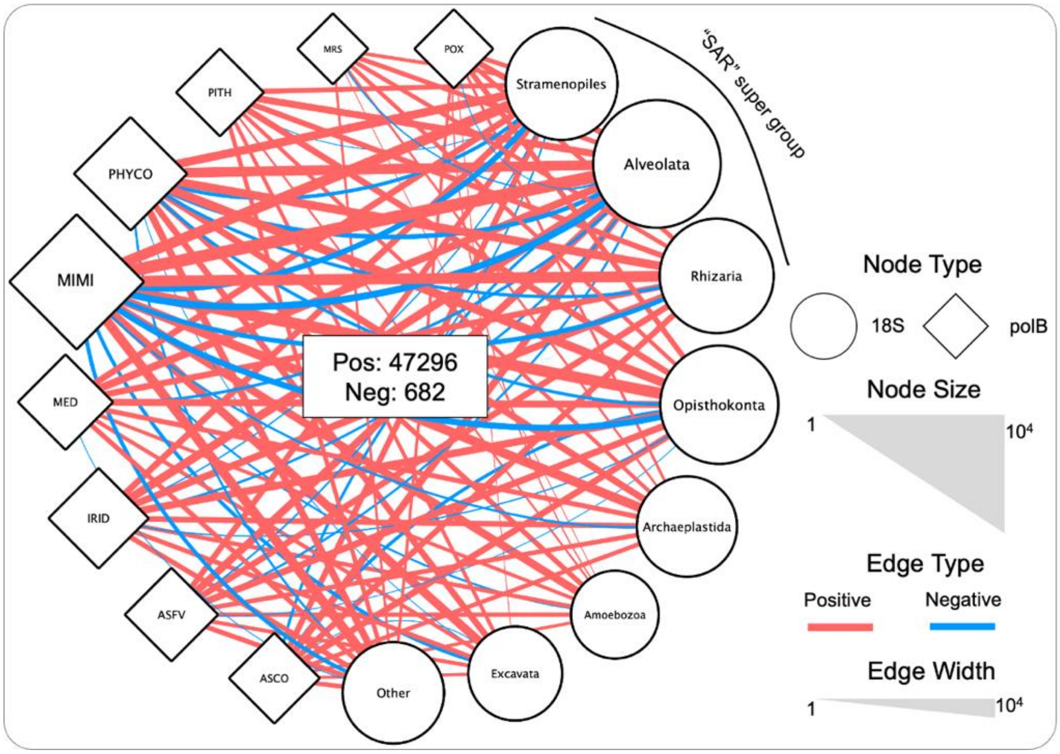


Fig. 3-4 DNApolB–V9 co-occurrence network.

A co-occurrence analysis at the OTU level was performed and constructed the network with pooled DNApolB –V9 associations from five size fraction networks. PolB OTUs were grouped at the family or family-like level, and V9 OTUs were grouped using annotation at high taxonomic ranks. The size of each node indicates the number of OTUs that belong to the group, and the width of each edge indicates the number of associations between two connected groups. Associations with positive weight are shown in red and negative associations are shown in blue. This figure has been published and modified under the CC-BY license from the paper by L. Meng et al., 2021.

Five co-occurrence networks, corresponding to five independent size fractions, comprising 20,148 V9 and 5,234 DNApolB OTUs (nodes), which were connected by 47,978 associations (edges) (Fig. 3-4). Of these, 47,296 associations had positive weights, while 682 held negative weights. Associations involving the family *Mimiviridae* were the most numerous, totaling 36,830, while *Marseilleviridae* formed the fewest associations, having 132 edges with eukaryotes. Taxonomic annotation highlighted Alveolata, Opisthokonta, Rhizaria, and Stramenopiles as major eukaryotic groups linked to *Nucleocytoviricota*. Mimiviridae and Alveolata exhibited the largest number of edges at 16,548.

Virus–eukaryote associations generally decreased with enlarging size fraction. The 0.8–5-µm fraction displayed the highest number of DNApolB–V9 associations (10,647) and the greatest eukaryotic community diversity. However, the 0.8–inf-µm network was largest for edges with positive weights (10,477). Eukaryotic community compositions in the networks, annotated by major lineages, varied across size fractions. In the smallest (0.8–5 µm) and broad size fraction (0.8–inf µm), Marine Alveolate Group II was the dominant eukaryotic lineage associated with *Nucleocytoviricota*, whereas associations with Metazoa and Collodaria rose with increasing size fractions. In the largest network (180–2000 µm), Metazoa contributed to 39.31% of the total DNApolB–V9 edges.

3.4.2 Network validation

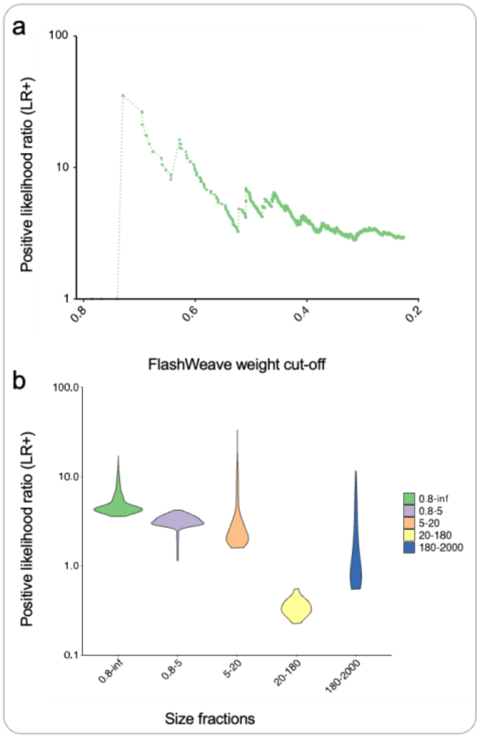


Fig. 3-5 Positive likelihood ratios (LR+) in the virus–host validation.

a. General performance of co-occurrence networks is shown with the LR+, and the LR+ values are plotted by dots and connected by a dashed line along with the association weight. A threshold to determine the statisticaly significant associations was set to alpha < 0.01. b. Performance of each size fraction network is shown with the violin plot by ggplot2 with a bandwidth of 2. Size fractions are presented in µm. This figure has been published and modified under the CC-BY license from the paper by L. Meng et al., 2021.

The quantitative assessment of predicting DNApolB–V9 associations using the LR+ involved defining groups of metagenomic DNApolBs, recruiting 932 OTUs, and contributing 6,191 polB–V9 associations in the FlashWeave networks. Pooled associations from five co-occurrence networks (4,069 associations after removing redundancy) were evaluated for overall performance. Additionally, LR+ was calculated for edges with positive and negative weights, separately, considering potential different infectious patterns. The LR+ of host prediction for positive associations was higher than 1 (LR+ = 1 indicates no change in the likelihood of the condition). The LR+ generally increased with the cut-off for FlashWeave weights, which indicated that condition positive cases are enriched in the edges with higher weights. This result demonstrated that the co-occurrence-based host prediction of giant viruses outperformed random prediction (Fig. 3-5). Nonetheless, the false discovery rate (FDR) was high, which indicated that the predictions contained numerous virus–host edges that were not considered condition positive based on current knowledge. Analysis of the rest of the study was thus confined to positive associations.

Performance comparison across different size fractions indicated that smaller size fraction networks, including the 0.8–inf-μm size fraction, offered superior prediction of virus–host relationships, with 0.8–inf-μm yielding the highest average LR+ among the five size fractions (LR+ = 4.97). Despite the general trend of smaller size fractions achieving higher LR+ values than larger ones, some exceptions were noted between 180–2000 and 20–180 μm. Furthermore, while the 0.8–inf-μm, 0.8–5-μm, and 5–20-μm size fractions all presented LR+ values exceeding 1, the 5–20-μm size fraction outperformed the others in terms of both LR+ and FDR when the weight exceeded 0.8.

Comparison of abundance filtration strategies using Flashweave-S (sensitive model) and FlashWeave-HE (heterogeneous model) yielded no consistent pattern in prediction performance. The networks derived from the Q1 filtration strategy showcased optimum performance using Flashweave-S, though Q1 filtration did not consistently outperform Q2 for Flashweave-HE inferred networks. Generally, Flashweave-S offered superior performance to the HE model across all filtration strategies. Lastly, in a comparison of networks inferred by FlashWeave-S, FastSpar, and Spearman methods, although all three generated a comparable number of positive associations, FlashWeave-S achieved the largest number of true positive predictions.

3.4.3 Host Prediction Improvement

The utilization of a phylogeny-guided host prediction tool, TIM, which filters DNApolB–V9 associations under the assumption that viruses and hosts that are evolutionarily related tend to infect each other, identified 24 eukaryotic taxonomic groups specifically associated with several viral lineages (Fig. 3-6). Comparing the performance of TIM results with raw FlashWeave results involved converting the three primary eukaryotic taxonomic ranks to their associated major lineages and plotting the associations as a network. This network illustrated that three families (*Mimiviridae*, *Phycodnaviridae*, and *Iridoviridae*) had enriched connections in specific eukaryotic lineages. Three known virus-host pairs were identified, such as Haptophyta–*Mimiviridae*, Mamiellophyceae–*Phycodnaviridae*, and Metazoa–*Iridoviridae*.

The TIM-filtered results exhibited a marked improvement in performance, with an average LR+ of TIM-enriched associations of 42.22, surpassing that of raw FlashWeave associations across various weight cut-offs (Fig. 3-6). Moreover, the false discovery rate (FDR) decreased significantly from 0.97 (no cut-off) and 0.95 (weight cut-off of 0.4) to 0.74.

Results exposed diverse putative hosts (spanning 13 lineages) for *Mimiviridae*, including algae, protozoans, and metazoans, with Metazoa having the most enriched nodes connected to *Mimiviridae*. Furthermore, specific eukaryotic lineages like MAST-3,12, Cryptophyta, Foraminifera, and Ciliophora were found to have robust relationships with *Mimiviridae*. Phycodnaviridae had connections with six eukaryotic lineages after TIM filtration, notably Bacillariophyta, 'other filosan (part of filosan Cercozoa)', and Mamiellophyceae. Rhodophyta, Ciliophora, and Dictyochophyceae presented links to both *Mimiviridae* and *Phycodnaviridae*, and an association was also identified between *Iridoviridae* and Metazoa.

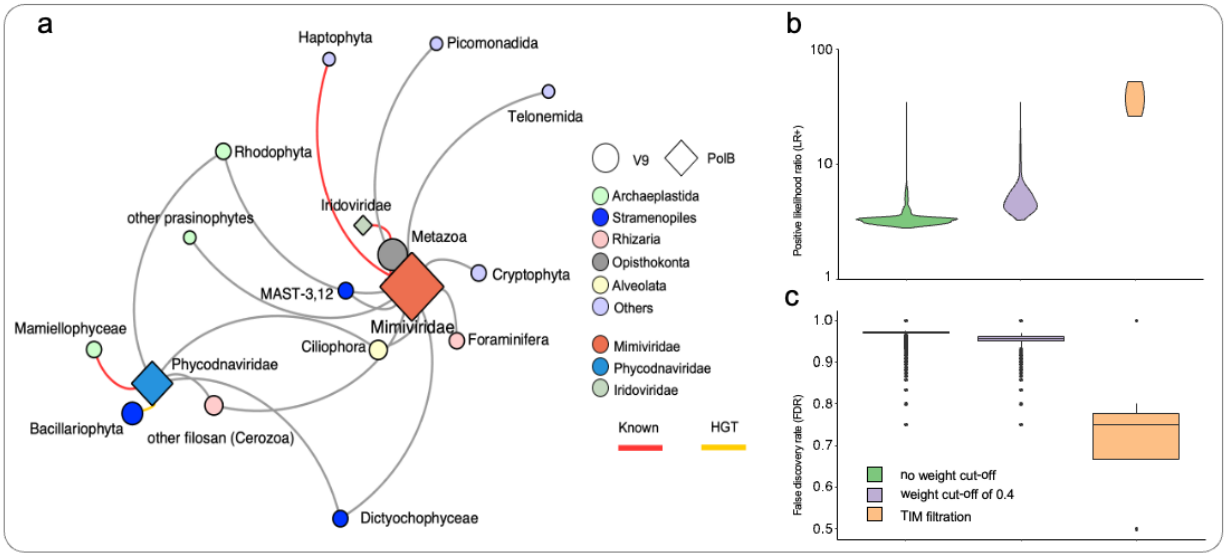


Fig. 3-6 Prediction of virus–host relationships with TIM.

a. Undirected network that shows the relationships between viruses and eukaryotes after TIM filtration. The size of each node indicates the number of predicted interactions of this group. The weight length of network edges as defined by the number of tree nodes enriched in each viral family subtree to specific eukaryotic major lineages in the TIM analysis. Known virus–host relationships are highlighted in red, and the pairs found to have horizontal gene transfer are highlighted in yellow. b. Performance of networks on the host prediction for original FlashWeave results without a weight cut-off, weight cut-off > 0.4, and TIM filtration. (C) FDR of networks for the host prediction with the original FlashWeave results without a weight cut-off, weight cut-off > 0.4, and TIM filtration. This figure has been published and modified under the CC-BY license from the paper by L. Meng et al., 2021.

3.5 Discussion

In this chapter, global ocean co-occurrence networks were constructed to predict the interactions of giant viruses (specifically *Nucleocytoviricota*) with various eukaryotic hosts. These networks were built using gene markers of giant viruses and eukaryotes from marine metagenome and metabarcoding datasets. The result revealed dense network edges representing extensive virus-eukaryote interactions, particularly involving two viral lineages, *Mimiviridae* and *Phycodnaviridae*. In this study, LR+ was calculated to quantitively assess the performance. LR+ is calculated with two relative values, sensitivity and specificity, based on the reference labels. The LR+ of host predictions using *Tara* Oceans metagenomics was higher than 1. For the first time, it is demonstrated that the co-occurrence-based host prediction of giant viruses outperformed random prediction. Moreover, the LR+ increased along with increasing cut-off values for the edge weights. Findings in this chapter indicate that higher weight values outputted from FlashWeave increase the probability of predicting true virus-host pairs.

Among all the predicted associations, true predictions (namely known virus-host pairs in previous literatures) were exclusively limited to positive weight associations, indicating that giant virus and host abundances were positively correlated across the different locations at a global scale. Further, employing a phylogeny-guided filtering method, TIM, improved the predictive performance and revealed several associations consistent with existing knowledge, such as *Phycodnaviridae* and Mamiellophyceae, *Mimiviridae* and Haptophyta, and *Iridoviridae* and Metazoa. In addition, some predicted connections by TIM were divergent from the known dataset, offering insights into possible novel virus-host pairs. Not included in the dissertation but discussed in the original paper, associations between virophages (parasites of giant viruses) and giant viruses were also explored (Meng et al., 2021), supporting a good performance of co-occurrence-based prediction and uncovering potential broader virus-virophage interaction ranges than previously known. The findings shed light on viral ecological interactions, co-evolution, and molecular exchange in the aquatic environment. This could offer a base for further explorations into the mechanistic underpinnings of these interactions and the implications for marine microbial and viral ecology.

Chapter 4: Genomic Adaptation of Giant Viruses in Polar Oceans

4.1 Abstract

Despite the perennially frigid conditions, polar oceans host a high and unique biodiversity. Various organisms display adaptive strategies in this challenging environment, yet the adaptation mechanisms of viruses remain largely unexplored. Viruses of the phyla *Nucleocytoviricota* and '*Mirusviricota*' represent eukaryote-infecting large and giant DNA viruses with functionally diverse repertoire. In this chapter, leveraging the marine giant virus genome database and the corresponding abundance profile, an ecological barrier distinctly separating polar and nonpolar viral communities is identified, attributing temperature as the key factor for dramatic shifts in the virus-host network at the polar/nonpolar boundary. Ancestral niche reconstruction indicates multiple recurrence of adaptation to polar conditions throughout evolutionary history of giant viruses, resulting in modern polar-adapted viruses scattered across phylogenetic trees. Several viral functions are likely specific related to polar adaptation, although most of their homologues are not identified as polar-adaptive genes in eukaryotes. The findings reveal that giant viruses adapt to cold environments by altering their functional repertoire using a strategy distinct from the adaptation stategy of their hosts.

4.2 Introduction

Polar regions are known to be some of the Earth's coldest environments, featuring pronounced seasonal changes in light. Despite these conditions, a large diversity of life forms, from single cullular organisms to large animals, thrives there thanks to the high primary productivity of phytoplankton. Organisms that have adapted to these extreme conditions commonly display unique physiological or morphological traits, enhancing their survival. For example, polar bears have developed specific morphological features, driven by genetic variations that originated in their ancestral gene pools (Alfredo et al., 2020). Additionally, Arctic and Antarctic fishes have evolved antifreeze proteins that allow them to maintain physiological activity in cold waters (DeVries & Cheng, 2005). For unicellular organisms, some psychrophilic bacteria exhibit oxygen-scavenging enzymes or modify their membrane chemistry (Methé et al., 2005).

How do viruses adapt to polar environments? As introduced in Chapter 1, viruses are the most abundant biological entities in the ocean and play critical roles in regulating microbial communities. Some known examples of virus adaptation is by acquiring metabolic genes, like cyanophages in low phosphorous settings often have genes related to phosphorus assimilation (Kelly et al., 2013). Moving to polar regions, recent metagenomic research has found a diverse range of viruses in both the Arctic (Xia et al., 2022) and Antarctic (Yau et al., 2011). The Arctic Ocean, in particular, has an elevated diversity of prokaryotic dsDNA viruses (Gregory et al., 2019). Many genes unique to these viruses appear to be under positive selection, based on mutation rate ratios, suggesting a role in adaptation to Arctic conditions (Gregory et al., 2019). Another study found that a prokaryotic virus reduced its genome size when exposed to lower temperatures in culture conditions (Ogunbunmi et al., 2022). Moreover, it has been observed that closely related viruses can show different infection dynamics in response to temperature variations, indicating that temperature can influence the selection of both viruses and their hosts (Demory et al., 2017, 2021). Giant viruses, with their giant genomes and diverse gene repertoires, present particularly intriguing subjects for studying adaptations to polar environments.

Therefore, this chapter focuses on examining the genes of giant viruses to better understand their adaptability to cold polar oceans. Utilizing data from the extensive *Tara* Oceans research project, the result demonstrates a clear barrier exists in giant virus genomes between polar and non-polar regions. Evolutionary trajectories of these viruses are explored to estimate their adaptive strategies over time. The analysis culminates in identifying specialized gene functions for viruses in polar areas, revealing that their adaptive approach to cold polar environments differs from that of their hosts.

4.3 Methods

4.3.1 Virus–plankton interaction network

Relative abundance of virus and eukaryotic MAGs across various size fractions (i.e., Pico, Piconano, Nano, Micro, Macro) was recruited to investigate the relationships between different viruses and eukaryotes. To ensure computational feasibility and robustness in our subsequent analyses, a Centred Log-Ratio (clr) transformation was performed after assigning a pseudo-count of one to all data entries and filtered out MAGs observed in fewer than three samples. Utilizing FlashWeave software (Tackmann et al., 2019), edges in the network analysis was filtered stringent statistical significance thresholds ($\alpha < 0.01$). Each discovered relationship was assigned with a weight between -1 and +1. The best positive or negative association (i.e., the edges with the highest absolute weights between two genomes) were selected to build the integrated interactome. Visualization of the network was using Cytoscape (Paul Shannon et al., 2003) with a prefuse force-directed layout. Additionally, proteins from linked genome pairs were aligned using the BlastP (Camacho et al., 2009) feature in Diamond software, adhering to a strict E-value cut-off to validate our findings. Robust ecological optima identified in the network were determined based on the methods described in Chapter 2.

4.3.2 Ancestral states estimation and Relative Evolution Divergency

Ancestral states of Nonpolar and Polar viruses were estimated using the function ‘ace’ (Ancestral Character Estimation) in the R package ‘ape’ (Paradis & Schliep, 2019). The input files were a rooted phylogenetic tree based on the four-hallmark gene set (RNApolA, RNApolB, DNApolB and TFIIS). In the tree, only viruses with biome assignments of Polar or Nonpolar were retained, and excluded viruses with 'Unknown' biomes. Parameters of type = 'discrete', method = 'ML', and model = 'ER' (one-parameter equal rates model) were used. The ancestral states were analysed based on a series of likelihood values for Polar and Nonpolar. Relative Evolutionary Divergence (RED) values were calculated using the 'get_reds' function in the package 'castor' (Louca & Doebeli, 2018).

4.3.3 KO enrichment and phylogenetic signal

As described in 2.3.4, viral genomes were categorically assigned a biome niche - 'Polar', 'Nonpolar', or 'Unknown', contingent on presence, absence, and overrepresentation of specific elements in varying biomes ('Biome and size niche' section). The enrichment of a given KO in Polar genomes, across four taxonomic levels, was rigorously assessed utilizing Fisher's exact test, with a significance threshold established at a corrected P-value of 0.05, post-Benjamini-Hochberg (BH) adjustment. Polar-specific KOs were defined under temperature and latitude optima and engaged in comparative analyses between fractions of components recognized as polar-specific KOs and all other pathway fractions. By exploring the phylogenetic signal of functions, it will be clear that whether polar-specific functions are under environmental selections or followed speciation history. For this purpose, two models, the Brownian motion model, and the Lambda model, were compared to evaluate their capacity to explain the trait distribution through a likelihood ratio test, utilizing 'fitContinuous' in R package 'geiger' (Pennell et al., 2014). The P-values to reject the null hypothesis were calculated by assuming chi-squared distribution with 1 d.f. for the likelihood-ratio test statistic and adjusted using the BH procedure. The threshold was set to a corrected *p*-value of 0.05.

4.4 Results

4.4.1 Polar barrier for giant virus genomes

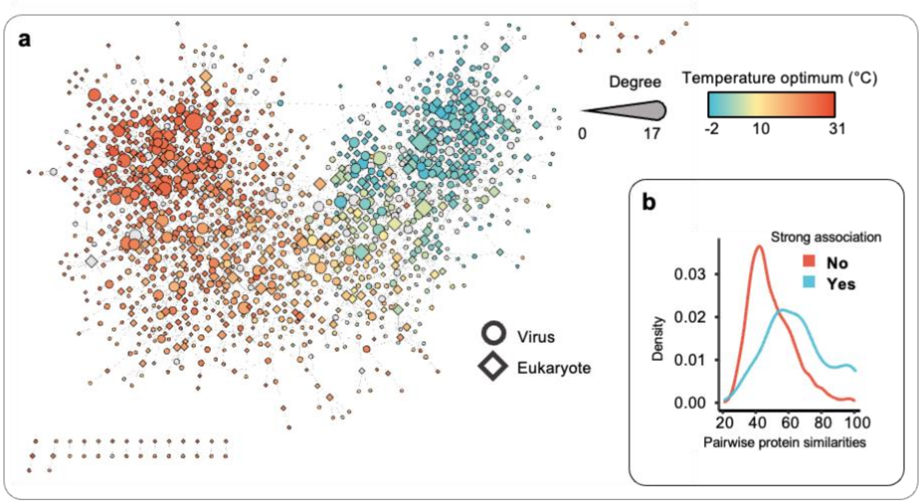


Fig. 4-1 A virus–plankton interaction network.

Five individual networks inferred using input matrices for the relative frequencies of eukaryotes and giant viruses. The best positive or negative association (i.e., the edges with the highest absolute weights between two genomes) were selected to build the integrated interactome. Node colour represents the temperature optima of each genome for viruses and eukaryotes. A total of 1,347 nodes are in the network. b, The distribution of pairwise sequence similarity of proteins (one protein from the eukaryotic genome and one from the viral genome). Blue line indicates the distribution for pairs with a strong virus–eukaryote association in the interactome (edge weight of ≥ 0.4), while the red line is for pairs lacking a strong association. This figure has been published and modified under the CC-BY license from the paper by L. Meng et al., 2023.

A global virus–plankton genome network via co-occurrence analysis was inferred using the GOEV and marine eukaryotic genome data (Fig. 4-1). A total of 2,135 virus–eukaryote associations were identified, with a dominant 91.94% being positive associations. Pairs showcasing strong co-occurrence associations exhibited significantly elevated protein similarities between their genomes, indicating horizontal gene transfers between these pairs. Notably, by estimating robust temperature optima for individual viruses and eukaryotes, it is identified that a strong correlation between the temperature optima and the structure of the virus–eukaryote network. A dramatic structural change in the network at the temperature-dependent polar/nonpolar boundary is the source of the uniqueness of polar viral communities.

Of the 569 genomes detected in polar regions, 262 were exclusive to these regions. Accordingly, biome-based classification of viral communities (Polar, Coastal, Trades, and Westerlies) had significant explanatory power for community variation (ANOSIM, $P < 0.01$). Further, R value of the ANOSIM test increased from 0.4021 to 0.6141 after merging three nonpolar biomes, demonstrating the existence of a clear polar barrier for giant virus communities. Viral communities in the Arctic regions were characterized by relatively high abundances showing peaks in cumulative RPKM plots for different size fractions (Chapter 2).

4.4.2 Recurrent polar adaptations throughout viral evolution

To explore viral adaptation across the polar barrier, ecological niche categories of "Polar" or "Nonpolar" were assigned to individual viral genomes. Out of 1,380 viral genomes, 450 were designated as Polar and 818 as Nonpolar. An additional 111 genomes were labeled "Unknown" due to their unclear distribution patterns (Fig. 4-2). This ecological niche assignment aligned with robust temperature and latitude optima. For example, a lineage of mirusviruses mainly consisted of Polar viruses, with a sub-clade comprising Nonpolar viruses. The classification was supported by the robust temperature optima of each genome. Although limitations like unequal sampling and sequencing depth could influence niche assignments, several instances supported these categorizations, such as the correct Polar designation for *Chrysochromulina ericina* virus, found in high latitude Norwegian coastal waters.

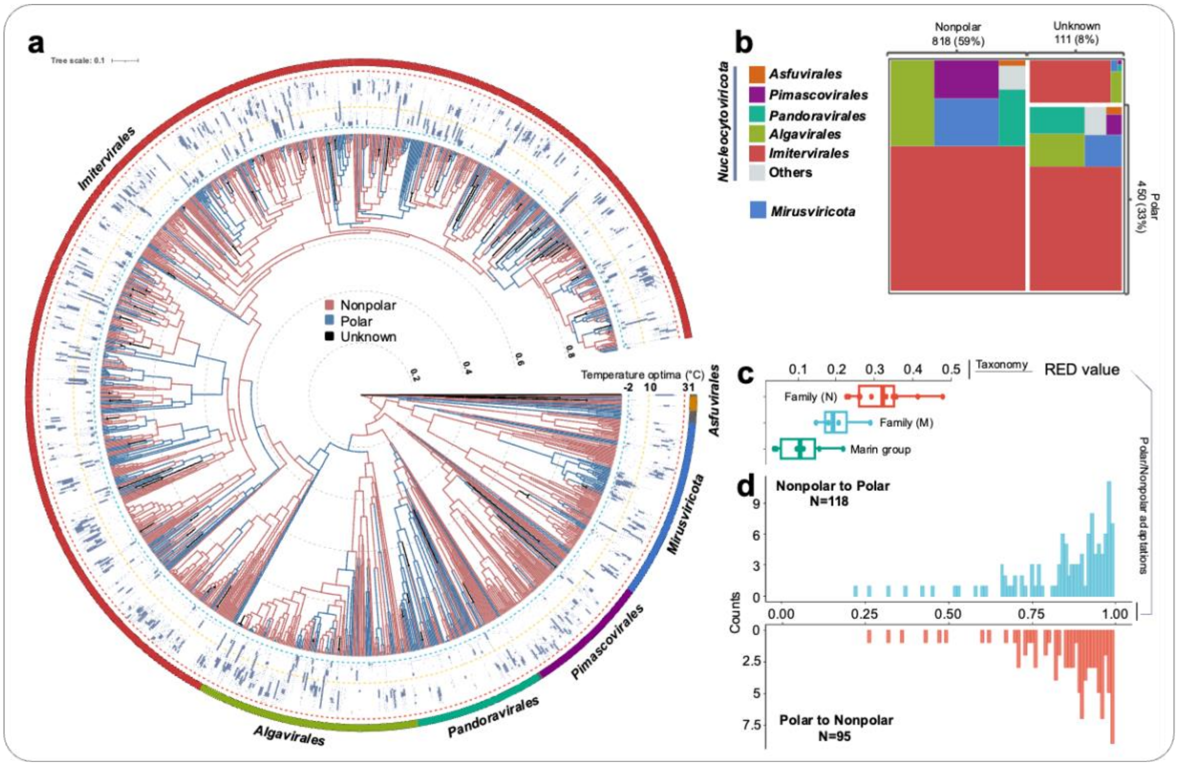


Fig.4-2 Inferred ancestral polar and nonpolar niches for viruses.

a, Ancestral "Polar" and "Nonpolar" states were estimated using the phylogenetic tree based on a one-parameter equal rates model. The outermost layer shows the taxonomy of six main groups. The boxplots in the second layer show the temperature optima of the viral genomes. For each box, n = 10,000 temperature values were analysed as outlined in the methodology section on robust ecological optimum and tolerance. Only polar and nonpolar genomes were included in the tree. b, The treemap diagram shows the number of viruses assigned to Polar, Nonpolar or "Unknown" biomes. Colours indicate the main taxonomic groups. c, Relative Evolutionary Divergence (RED) values for viral main groups (n = 6) and families. N stands for the phylum Nucleocytoviricota (n = 17) and M stands for Mirusviricota (n = 5). d, Histograms of RED values for the nodes at which "polar" or "nonpolar" adaptation events were inferred. RED values of child nodes in adaptation events were shown. This figure has been published and modified under the CC-BY license from the paper by L. Meng et al., 2023.

Next, this study employed a maximum likelihood approach for Polar/Nonpolar state reconstruction for ancestral nodes in the phylogenetic tree. As a result, 118 transitions from Nonpolar to Polar and 95 transitions from Polar to Nonpolar were inferred along the branches of the tree (Fig. 4-2). These niche adaptations occurred repeatedly throughout the evolutionary history of these viruses, which originated from a root inferred to be Nonpolar. However, the data couldn't rule out the possibility of an original polar root due to challenges in establishing the tree's root. Most of the reconstructed niche adaptations happened relatively recently, following the formation of viral genera, although some were estimated to have occurred during early evolutionary stages, corresponding to order-level divergence.

4.4.3 Polar-specific viral functions

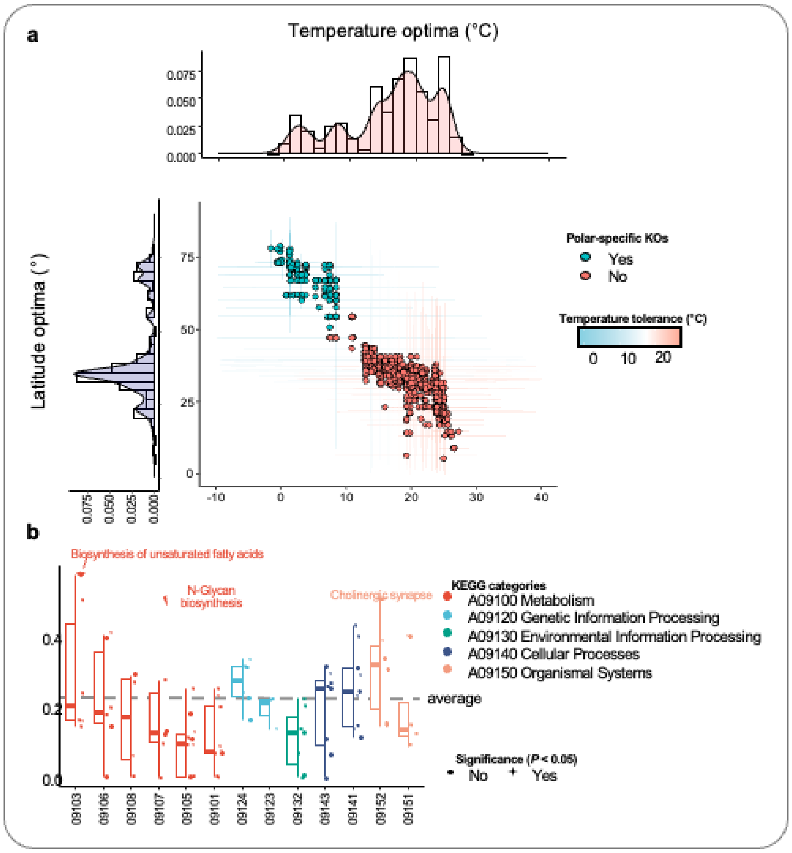


Fig. 4-3 Ecological niche of KEGG Orthologs (KOs) and polar-enriched pathways. a, Distribution of the temperature optima and latitude optima for KEGG Orthologs (KOs) found in viral genomes. Colours of dots represent the Polar or Nonpolar niche for each KO. Bars indicate the tolerance ranges of temperature (horizontal) and latitude (vertical). Histograms show the distributions of temperature and latitude optima. b, A boxplot with jitter of ratio of Polar KOs in each pathway. The x axis shows the second-level categories of KEGG pathways. This figure has been published and modified under the CC-BY license from the paper by L. Meng et al., 2023.

To investigate genomic adaptations to polar regions, gene functions encoded in viral genomes were annotated using KEGG Orthologs. Temperature and latitude optima for genes observed in multiple genomes were calculated. This analysis revealed two major categories of genes: one primarily distributed in high-latitude/low-temperature areas and another in lower-latitude/higher-temperature regions (Fig. 4-3). Polar-specific genes had temperature optima below 10°C and latitude optima above 50°. These genes showed a relatively narrow phylogenetic distribution compared to other genes, indicating that factors beyond speciation history, such as environmental conditions or host distributions, impact their distribution.

Further analysis showed a significantly higher proportion of Polar-specific genes in genomes classified as Polar compared to those classified as Nonpolar or Unknown. Specific functions like ceramide glucosyltransferase and dihydrofolate reductase were found exclusively in polar genomes, implying unique adaptive strategies. At the pathway level, unsaturated fatty acid biosynthesis, N-glycan biosynthesis, and cholinergic synapse pathways were significantly enriched with Polar-specific genes, suggesting multiple avenues through which polar viruses could be adapting to their environment.

4.4.4 Other potential polar adapted functions

The study conducted further enrichment analyses to examine gene functions in Polar and Nonpolar viral genomes across different evolutionary scales. The analyses highlighted 265 functions that were significantly enriched in Polar genomes. A finer examination of one Mesomimiviridae clade revealed four functions that were more prevalent in Polar genomes, including a near-complete CMP-KDO biosynthesis module, which is involved in lipopolysaccharide biosynthesis (Fig. 4-4). This suggests that Polar viral genomes may utilize glycoconjugates to enhance virion-host interactions or virion stability.

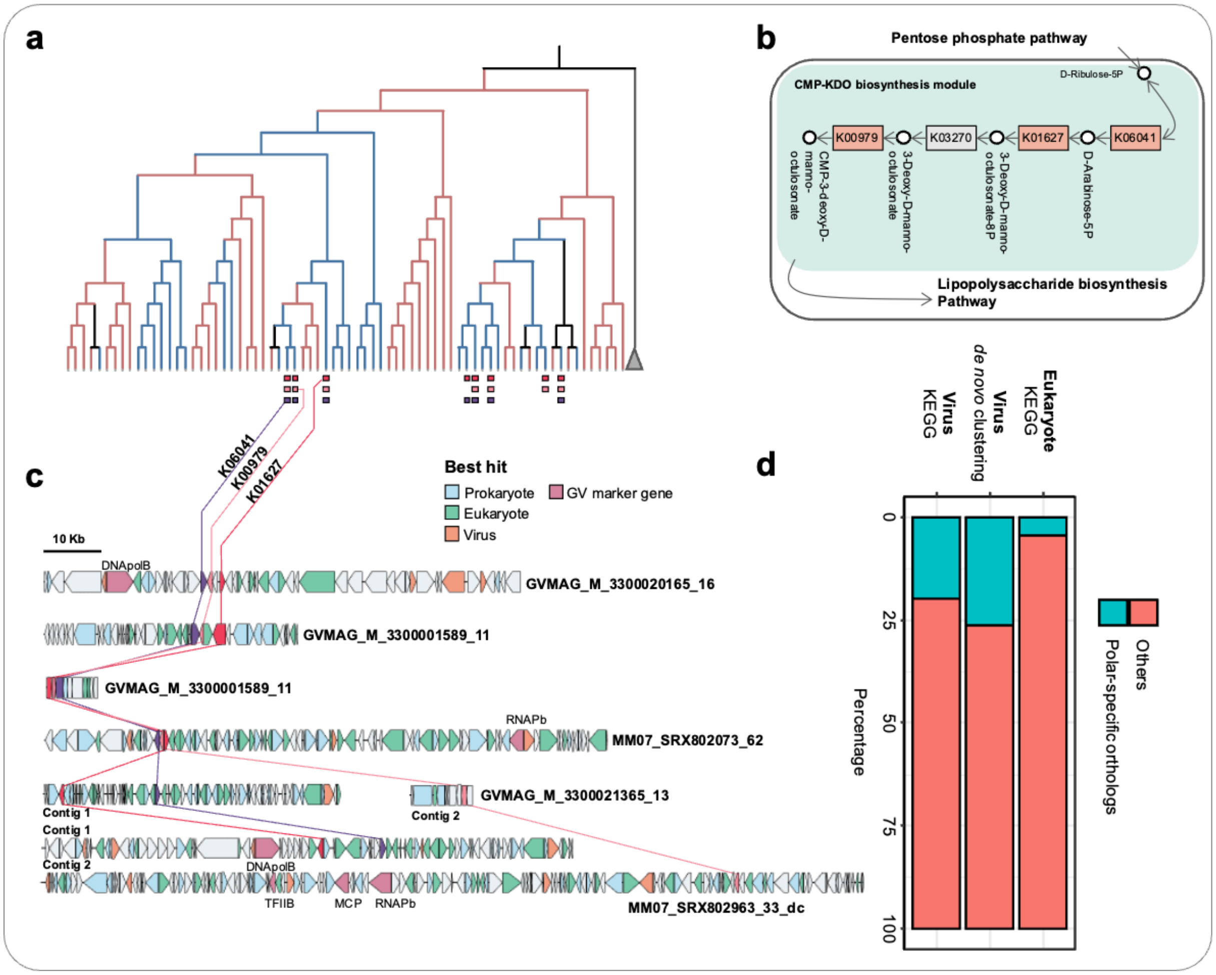


Fig. 4-4 Independent genomic adaptation of giant viruses.

244 functions (KOs) were enriched at individual lineages. One example was given in a, Three KOs that were present exclusively in more than five Polar genomes in a selected Mesomimiviridae clade. Three of them (K01627, K00979, K06041) were encoded in the same genomes and formed a near-complete CMP−KDO biosynthesis module shown in b, Schematic of the three Polar enzymatic steps in the CMP–KDO biosynthesis module. c, Genome maps of MAGs encoding three CMP-KDO KOs. Best matched taxonomies of genes are shown using the same colours, with the key provided at the top right. Coloured lines connect detected CMP-KDO KOs between every two contigs. “contig1” and “contig2” indicate two contigs come from the same MAG. d, Proportion of Polar and Nonpolar specific functions (KOs and GCCs) in viruses and eukaryotes. This figure has been published and modified under the CC-BY license from the paper by L. Meng et al., 2023.

In addition, the study analyzed gene cluster communities (de novo clustering)[14], revealing a higher proportion of Polar-specific gene clusters than identified through KO annotations, suggesting the presence of unknown functions with Polar-specific distributions. The Polar genomes also had a higher proportion of Alanine-rich low-complexity regions. These low-complexity sequences potentially have an anti-freeze function, as alanine-rich helical structure is one of the significant characteristics of type I antifreeze proteins for ice growth inhibition. Although not statistically significant, a larger percentage of Polar genomes encoded antifreeze protein homologs compared to other genomes, hinting at additional adaptation mechanisms for extreme conditions.

Finally, to examine whether genomic adaptation of eukaryotic plankton is related to the one in viruses, the temperature and latitude optima were calculated for KOs in plankton (n = 11,988). A similar pattern of Polar and Nonpolar KO groups was identified, although the proportion of the Polar KO group (n = 523, 4.36%) was much smaller than that for viruses (19.74%) (Fig. 4-4). Interestingly, of the 523 KOs in the eukaryotic Polar group, only four were found in the viral Polar group. These were PPM family protein phosphatase, L-galactose dehydrogenase, transcription factor S, and ATP-dependent DNA helicase DinG. This result indicates that most Polar viral functions do not exhibit the same temperature/latitude optima as their homologs in eukaryotic genomes, suggesting that genomic adaptations are uncoupled between viruses and eukaryotes.

4.5 Discussion

A prerequisite for a virus to adapt to a new environment is that its host has already adapted to that environment. This host adaptation would give rise to additional environmental (or micro-

[14] AGNOSTOS, high-quality remote gene clusters approach

environmental) changes for the virus. Such micro-environmental changes include alterations of cell surface structures as well as intracellular metabolic states. Virus-host interactions involve different processes such as adhesion to the cell, metabolic remodeling, viral genome replication, genome packaging and egress from the cell. These processes are likely affected not only by ambient physio-chemical conditions (such as temperature) but also, and more profoundly, by the biochemical and physiological conditions of the host cell that adapts to the target environment. Therefore, for a virus to adapt to a new environment, it needs to cope with both environmental changes and environment-induced host cell alterations. Our results suggest that adaptation of large and giant DNA viruses to polar environments involves the alteration or innovation of viral metabolic strategies, which is manifested in viral genomic changes. In this adaptive process, viruses appear to take their own strategies that are distinct from the host strategies for their adaption in the same habitat.

The adaptation of cellular organisms to their environments could be largely manifested in their functional repertoire. In viruses, previous discoveries of presence of ecologically significant genes (such as lipid metabolism and rhodopsin) (Needham et al., 2019; Rosenwasser et al., 2014) indicated that functional repertoire could also be important for adaptive evolution of viruses. However, the functional adaptation of viruses at a wide geographic scale has not been investigated as deeply as for cellular organisms. Thanks to the recent progress in metagenomics, the links between the biogeography, host types, and gene repertoire of large and giant DNA viruses infecting marine eukaryotes could be investigated. The existence of a strong polar/nonpolar barrier for these viruses and revealed size fraction-dependent Arctic diversity hotspots for some virus groups was confirmed, which may reflect a high diversity of their hosts in cold environments. The phylogenomic tree and ancestral state reconstruction revealed back-and-forth adaptations between lower- and higher-temperature niches that occurred recurrently throughout the long evolutionary course of these viruses. Numerous functions, especially ones related to host interactions, were found to be specific to viral polar adaptation, but most of them were not identified as polar-specific functions in eukaryotes, suggesting a decupling of viral and host polar adaptations. Furthermore, the gene repertoire of these large DNA viral genomes appears more evolutionarily flexible and responsive to temperature change than that of eukaryotic genomes. Together, this Chapter provides new insights, suggesting that the evolution of viruses could be influenced by their surrounding physical environments.

Chapter 5 Discovery of a novel phylum '*Mirusviricota*'

5.1 Abstract

DNA viruses, including giant viruses and phages, play a significant role in the ecology and evolution of cellular organisms, yet their diversity and evolutionary trajectories remain poorly understood. Starting with a comprehensive database of giant virus MAGs and phylogenetic analyses, a novel phylum '*Mirusviricota*', consisting of plankton-infecting relatives of herpesviruses, was discovered and identified in sunlit oceanic metagenomes. The virion morphogenesis module of '*Mirusviricota*' shares critical characteristic features with the realm *Duplodnaviria*. Specifically, the major capsid proteins of '*Mirusviricota*' represent an intermediary evolutionary stage between Caudoviruses and Herpesviruses. In contrast, a substantial portion of the genes in the informational module of '*Mirusviricota*' are absent in herpesviruses but show homology with the giant viruses of *Varidnaviria*. Revealing a fascinating chimeric nature, '*Mirusviricota*' serves as an evolutionary bridge between the two viral realms of *Duplodnaviria* and *Varidnaviria*. Moreover, the genomes of mirusviruses encode several functional genes crucial for infecting plankton. Being both prevalent and transcriptionally active in the ocean, mirusviruses may also have a significant impact on marine ecological dynamics.

5.2 Introduction

According to hybrid hypotheses on viral evolution (See Chapter 1) and the taxonomic system of ICTV, double-stranded DNA viruses are classified into two major realms: *Duplodnaviria* and *Varidnaviria* (Krupovic et al., 2019; Siddell et al., 2023). *Duplodnaviria* comprises tailed bacteriophages and related archaeal viruses of the class *Caudoviricetes* and eukaryotic viruses of the order *Herpesvirales*. *Varidnaviria*, as described above, includes giant viruses from the phylum *Nucleocytoviricota* as well as smaller viruses with tailless icosahedral capsid (Koonin et al., 2020b). The two realms were established based on the non-homologous sets of virion morphogenesis genes (virion module), including the structurally unrelated MCPs, namely the 'double jelly-roll' (DJR) and HK97 MCP folds in *Varidnaviria* and *Duplodnaviria*, respectively (Krupovic et al., 2019). Members of both realms infect organisms across all domains of life, with the respective ancestors thought to date back to the last universal cellular ancestor (Krupovic, Dolja, et al., 2020).

Within *Duplodnaviria*, caudoviruses infect bacterial and archaeal and display a continuous range of genome sizes, whereas herpesviruses, exclusively infect animal hosts, have genomes in the range of 100-300 kb. Evidence suggests that herpesviruses likely evolved from bacteriophages. However, the absence of viruses that infect eukaryotes other than animals raises questions about their actual evolutionary trajectory within *Duplodnaviria* (Koonin et al., 2015). On the other hand, members of the *Varidnaviria* display a wide range of genome sizes, from ~10 Kbp to >2 Mbp, but there is a discontinuity in the complexity between giant viruses of the and the rest of varidnaviruses with genomes smaller than 50 Kbp. It has been suggested that *Nucleocytoviricota* have evolved from a smaller varidnavirus ancestor (Guglielmini et al., 2019; Krupovic & Koonin, 2015; Woo et al., 2021), but the acquisition of multiple informational genes and the gigantism remains to be fully understood.

Viruses in the *Caudoviricetes* and *Nucleocytoviricota* groups are abundant in the sunlit ocean, where they play a crucial role in regulating both the composition and blooming activity of plankton communities (Kaneko et al., 2021; Mann, 2003; Schulz et al., 2020). Therefore, a comprehensive metagenomic survey of marine samples could significantly advance our understanding of the diversity of dsDNA viruses and their ecological impact. The *Tara* Oceans expedition is a global-scale survey on marine ecosystems that expands our knowledge of microbial diversity, organismal interactions, and ecological drivers of community structure. Epipelagic zone samples (including surface and Deep Chlorophyll Maximum layers) in the *Tara* Oceans project provide nearly 300 billion metagenomic reads (Sunagawa et al., 2020). A comprehensive database enriched in large and giant marine eukaryotic dsDNA viruses (thereafter called Global Ocean Eukaryotic Viral 'GOEV' database) was constructed using the surface metagenomes. The survey using the database led to the discovery of plankton-infecting relatives of herpesviruses that form a putative new phylum dubbed '*Mirusviricota*'.

5.3 Methods

5.3.1 Identification of '*Mirusviricota*'

As described in the published method (Gaïa et al., 2023), the identification of '*Mirusviricota'* involved a multifaceted approach to search MCPs and morphogenetic module proteins, with AGNOSTOS (Vanni et al., 2022) aiding in determining potential candidates. These candidates underwent protein structural modeling using advanced tools, namely AlphaFold2 (Jumper et al., 2021) and RoseTTAFold (Baek et al., 2021). Generated models were then critically compared to known capsid protein structures. The functionality of proteins in the '*Mirusviricota*' core gene clusters, especially those lacking sequence-based functional annotation, was inferred through structural modeling with AlphaFold2 and functionality prediction via the DALI server. To further assign the viral realm of Mirus genes, two custom Hidden Markov Model (HMM) databases were created through meticulous curation and amalgamation of various coding sequence datasets.

5.3.2 Functional annotation

Orthologous groups (OGs) in Mirus MAGs (N = 111), a Mirus near-complete contiguous genome, and reference genomes from the Virus-Host Database (Mihara et al., 2016) (including 1,754

Duplodnaviria, 184 *Varidnaviria*, and 11 unclassified genomes) were generated, resulting in 26,045 OGs. A subset of these (9,631 OGs) with at least five genome observations was utilized for further genome clustering. AGNOSTOS was implemented to categorize protein-coding genes from the GOEV database, generating low functional entropy gene groups and facilitating functional annotation via remote homology methods. *Nucleocytoviricota* genomes were functionally inferred by comparing genes against multiple databases like Virus-Host DB, RefSeq, and others, using tools like Diamond (Buchfink et al., 2021) and Hmmsearch with specific E-value cut-offs. Further annotations and functional categorizations were achieved with additional resources like the GVOG and eggnogmapper (Cantalapiedra et al., 2021). tRNAs were predicted by tRNAscan-SE (Lowe & Eddy, 1997).

5.4 Results

5.4.1 Overview of marine eukaryotic DNA viral genomes

From 798 metagenomes from the *Tara* Oceans expeditions, over 2,500 non-redundant environmental RNAPb protein sequences were identified through a broad-spectrum profile hidden Markov model. Based on the RNAPb-guided search, 698 viral MAGs of *Nucleocytoviricota* were generated. Phylogenetic signal for those genomes in the database not only enlarges a notable diversity of marine *Nucleocytoviricota* but also unveiled novel, deep-branching lineages, dubbed ‘Mirus’, with apparent phylogenetic independence from the three recognized domains of life (Fig. 5-1). Among these, 587 MAGs belonging to *Nucleocytoviricota* have genome sizes reaching up to 1.45 Mbp and an average length of approximately 270 Kbp. Additionally, 111 non-redundant Mirus MAGs were found, with the largest genome size up to 438 Kbp and an average length of around 200 Kbp. After integrating MAGs from two previous surveys and public reference *Nucleocytoviricota* genomes.

A total of 1,817 MAGs were retained in the final GOEV database, possessing approximately 0.6 million genes.

5.4.2 Discovery of Mirusviruses

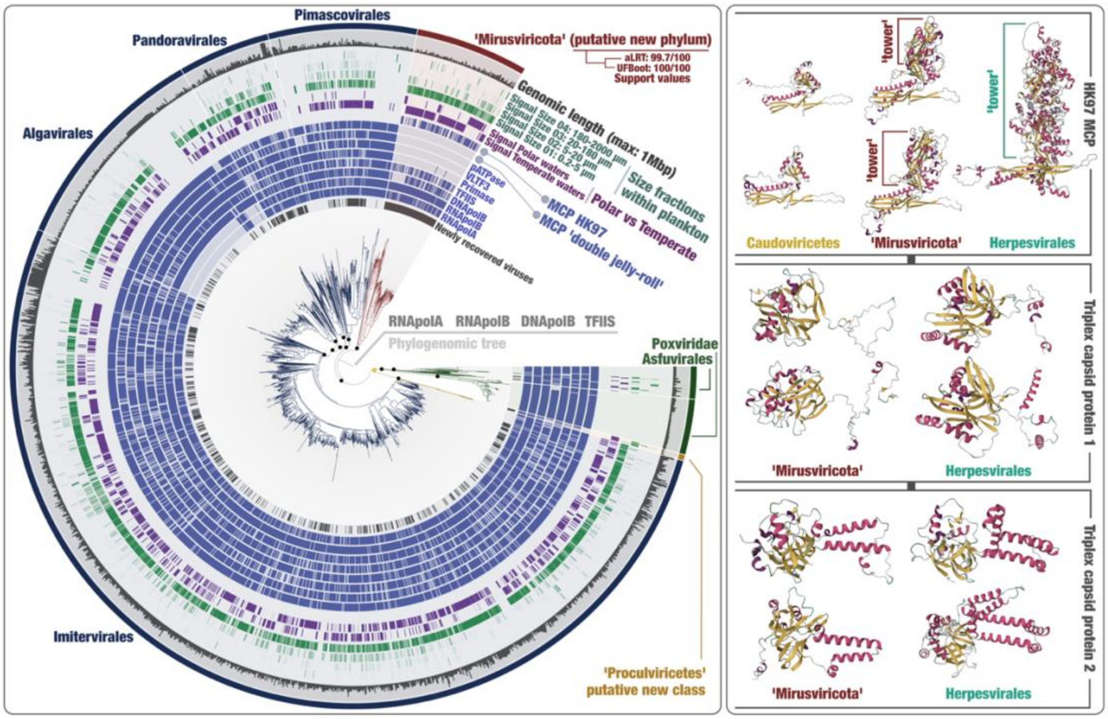


Fig. 5-1 Evolutionary relationships between *Nucleocytoviricota*, *Herpesvirales* and mirusviruses.

The left section of the figure illustrates a phylogenetic tree for the GOEV database. This tree is based on a concatenation of RNApolA, RNApolB, DNApolB, and TFIIS genes On the right, the panel presents the 3D structures of the MCP from three distinct sources: Escherichia phage HK97 representing *Caudoviricetes*, a representative genome for mirusviruses predicted using Alphafold, and the human cytomegalovirus from the *Herpesvirales*.This figure has been published and modified under the CC-BY license from the paper by M. Gaia#, L. Meng# et al., 2023.

A deep dive into the *Nucleocytoviricota* MAGs unveiled most of the signature genes pertinent to the virion and informational modules typical to this viral phylum. In addition to mirusviruses, a potential new class-level group, '*Proculviricetes*' (Fig. 5-1), also broadened our understanding of the diversity within *Nucleocytoviricota*. Meanwhile, several key genes in the *Nucleocytoviricota* informational module, such as RNApolA and RNApolB, were encoded by mirusviruses, revealing that mirusviruses are evolutionarily linked to the *Nucleocytoviricota*. However, the phylogenetic tree showed mirusviruses are monophyletic and distinct from all known *Nucleocytoviricota* classes (Fig. 5-1). Mirus MAGs, organized into seven subclades M1 to M7 (Fig. 5-2), exhibited notable hallmark genes related. One notable characteristic is that mirusviruses lack identifiable homologs of DJR-fold MCP, which is representative in the virion module of *Nucleocytoviricota*. Instead, protein structure prediction helped identify a new clade of HK97-fold MCP in most mirusviruses, assigning them to the realm *Duplodnaviria* according to the ICTV classification system (See Chapter 1). These mirusvirus MCPs, intriguingly, had similarities with both *Caudoviricetes* (no 'tower' in MCP) and *Herpesvirales* (larger MCP 'tower'), suggesting an intermediate evolutionary state (Fig. 5-1). The detailed analysis of these viruses revealed an evident coevolution of two functional (i.e., virion and informational) modules (Fig. 5-2). Although definitive phylogenetic comparisons are challenging due to extensive protein sequence divergences, the findings strongly supported the classification of mirusviruses as a new viral phylum[15], termed *'Mirusviricota'*. This new phylum served as a third clade within the *Duplodnaviria* realm.

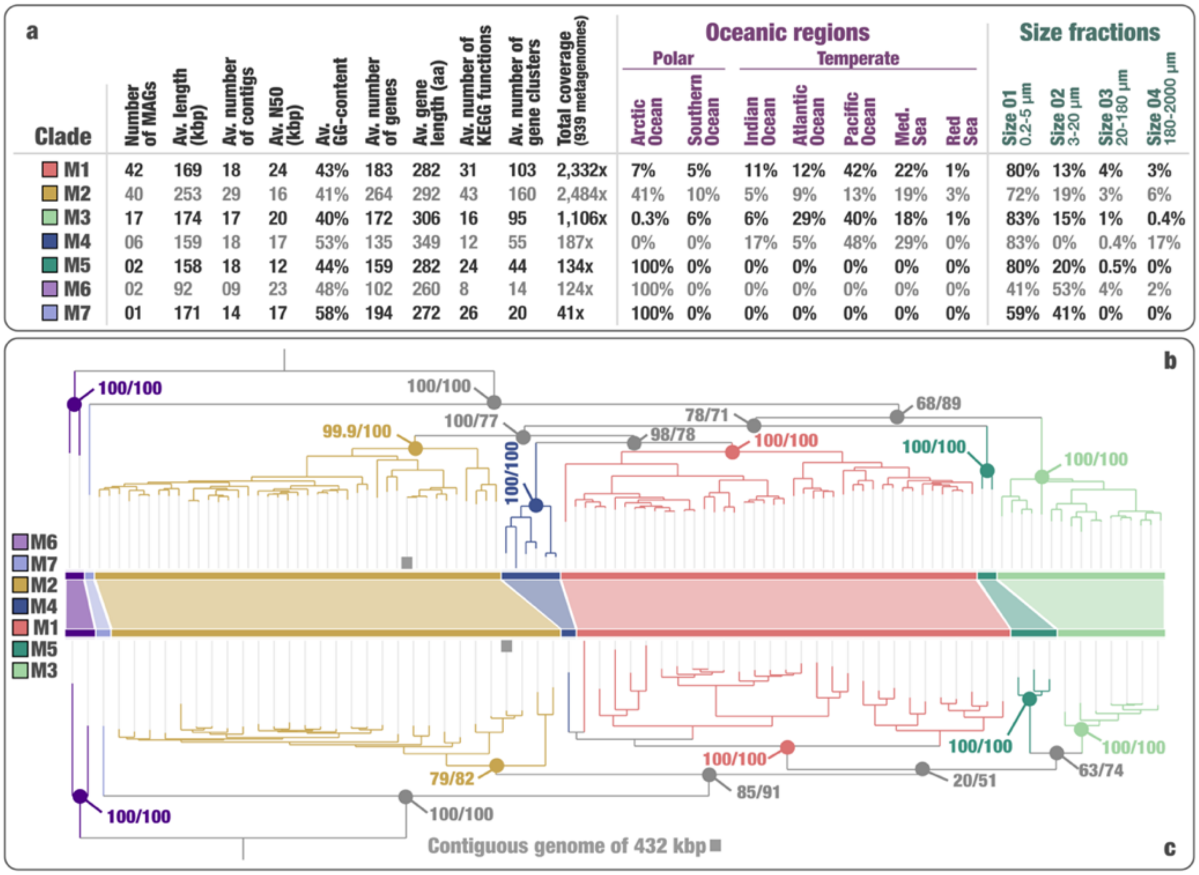
5.4.3 Features of Mirusvirus Functions

The exploration of 111 ‘*Mirusviricota*’ MAGs, encompassing a total of 22,242 genes, reveals a complexity in the functional capabilities of mirusviruses, highlighting the mechanisms they are involved signal transduction, degrade proteins, manipulate critical cellular mechanisms and replicate within the host. These MAGs contain 35 core gene clusters (defined as present in at least half of

[15] The common attributes of mirusviruses are distinct from other two phyla, ‘*Uroviricota*’ (phage) and ‘*Peploviricota*’ (herpesvirus). Meanwhile, the 7 clades within mirusviruses are highly divergent to each other.

mirusviruses), 1,825 non-core gene clusters, and 9,018 singletons, covering various aspects of the life cycle, regulation, and functioning of cells and viruses. Intriguingly, nine core gene clusters, critical to ‘Mirusviricota’ and coding for proteins with confidently predicted structures, have not yet been functionally annotated.

Despite the high virion module similarity between mirusviruses and herpesviruses, there was a pronounced functional affiliation between mirusviruses and the *Nucleocytoviricota*. A notable overlap exists between the '*Mirusviricota*' and *Nucleocytoviricota* genomes in terms of gene clusters related to DNA replication. These shared clusters prominently feature genes involved in key replication processes, including those coding for glutaredoxin/ribonucleotide reductase, Holliday junction resolvase, proliferating cell nuclear antigen, dUTPase, and DNA topoisomerase II. This suggests that the functional connectivity between these two phyla extends well beyond the realm of basic information processing, reaching into critical aspects of genome replication.



a

Clade	Number of MAGs	Av. length (kbp)	Av. number of contigs	Av. N50 (kbp)	Av. GG-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							Size fractions			
											Polar		Temperate								
											Arctic Ocean	Southern Ocean	Indian Ocean	Atlantic Ocean	Pacific Ocean	Med. Sea	Red Sea	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%	12%	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%	6%	6%	29%	40%	18%	1%	83%	15%	1%	0.4%
🟦 M4	06	159	18	17	53%	135	349	12	55	187x	0%	0%	17%	5%	48%	29%	0%	83%	0%	0.4%	17%
🟩 M5	02	158	18	12	44%	159	282	24	44	134x	100%	0%	0%	0%	0%	0%	0%	80%	20%	0.5%	0%
🟪 M6	02	92	09	23	48%	102	260	8	14	124x	100%	0%	0%	0%	0%	0%	0%	41%	53%	4%	2%
🟦 M7	01	171	14	17	58%	194	272	26	20	41x	100%	0%	0%	0%	0%	0%	0%	59%	41%	0%	0%

Fig.5-2 Analysis of '*Mirusviricota*' genomics and evolution

Panel a showcases genomic and environmental statistics across seven Mirusviricota subclades, highlighting average statistics, amino acid (aa) counts, and relevant KEGG data. Panel b illustrates a maximum-likelihood phylogenomic tree of 'Mirusviricota' MAGs. This tree is constructed based on concatenated sequences of four key informational genes (RNApolA, RNApolB, DNApolB, and TFIIS). Panel c presents another maximum-likelihood phylogenetic tree, focusing on the MCP. This figure has been published and modified under the CC-BY license from the paper by M. Gaia#, L. Meng# et al., 2023.

However, distinct lifestyles for the two clades were represented by functional genes were implied by significant enrichment of such genes in either mirusviruses or *Nucleocytoviricota*. Certain core gene clusters in the mirusviruses, including trypsin, M16-family peptidase, TATA-binding protein, heliorhodopsin, and histone, were significantly less represented among *Nucleocytoviricota* genomes. Moreover, phylogenetic evaluations of histones and rhodopsins reveal several mirusviruses-specific monophyletic clades, suggesting that these critical functions were acquired at ancient time and are specific to mirusviruses.

5.4.4 The Chimeric Nature of mirusviruses

Subsequently, '*Mirusviricota*' MCPs were screened in a database containing hundreds of metagenomic assemblies from the 0.2-3 μm size fraction of the surface oceans. A nearly complete contiguous '*Mirusviricota*' genome was found in the Mediterranean Sea with a length of 431.5 kb, just 6 kb shorter than the longest MAG. This near-complete contiguous genome contains all marker genes of '*Mirusviricot*a' as well as a non-core gene that might represent longest viral gene (over 11.5 Kaa). This genome could be assigned to the clade M2 based on both the information module and MCP phylogenies.

The 355 genes found in the near-complete contiguous genome were compared to two comprehensive genomic databases corresponding to the realms *Duplodnaviria* and *Varidnaviria*, and 86 significant hits were found (Fig. 5-3). Only six of them had better matches within the *Duplodnaviria* database and included the terminase protein. The remaining 80 genes had a better match within the *Varidnaviria* and occurred relatively homogeneously across the genome. These included the RNApolA, RNApolB, DNApolB, DNA 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 supported the strong functional connectivity between mirusviruses and the large and giant DNA viruses within the realm *Varidnaviria*.

On one hand, mirusviruses belong to the realm *Duplodnaviria* based on their virion module. On the other hand, hallmark informational markers and other relevant functions missing in all the known *Herpesvirales* lineages display surprisingly high sequence similarities to the corresponding proteins encoded by members of the phylum *Nucleocytoviricota*. Notably, the near-complete contiguous '*Mirusviricota*' genome perfectly recapitulated all chimeric attributes initially observed based on 111 manually curated MAGs. Thus, this putative new phylum is not only an integral component of the ecology of eukaryotic plankton but also fills critical evolutionary gaps between *Duplodnaviria* and *Varidnaviria*, the two major realms of double-stranded DNA viruses.

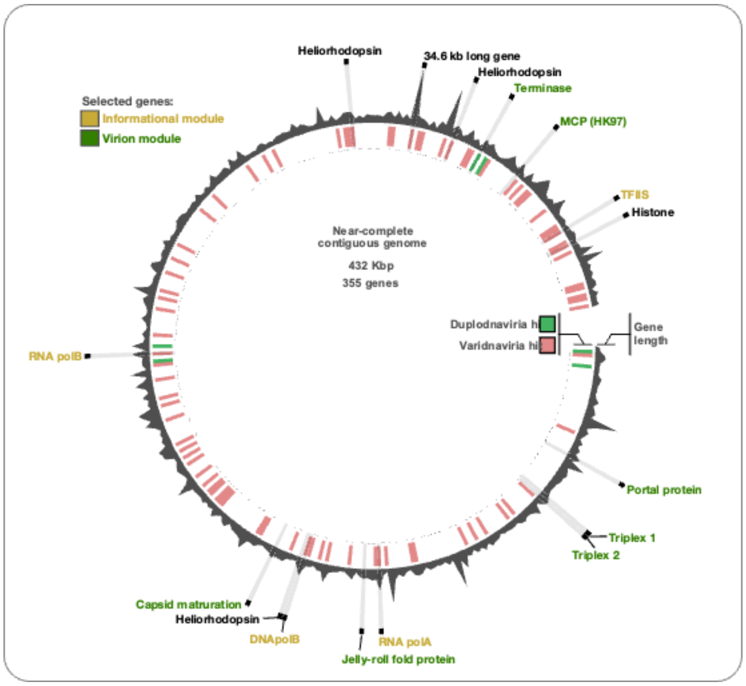


Fig. 5-3 A near-complete genome for '*Mirusviricota*'.

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. This figure has been published and modified under the CC-BY license from the paper by M. Gaia#, L. Meng# et al., 2023.

5.5 Discussion

'*Mirusviricota*', a potentially new phylum within *Duplodnaviria* discovered in the global ocean, represents an impactful clade of large eukaryotic DNA viruses. Mirusviruses exhibit distinct genomic features and substantial functional overlaps with significant eukaryotic varidnaviruses, presenting a complex lifestyle and impactful influence on the ecology of pivotal marine eukaryotes. Furthermore, their gene repertoires, imply a consequential role in marine eukaryotic plankton ecology and hint at a potentially underappreciated lifestyle and evolutionary influence through gene flow within marine biomes. Functions that are more prevalent in mirusviruses compared to *Nucleocytoviricota* feature unique phylogenetic branches of H3 histones, which play a role in eukaryotic chromatin formation, as well as heliorhodopsins, light-sensitive receptor proteins used by giant viruses as proton channels during infection. Intriguingly, the core heliorhodopsin in micromonas may have originated from a mirusvirus, suggesting that mirusviruses are significant contributors to the evolutionary development of micromonas through gene exchange. Additionally, the close sequence resemblance between heliorhodopsins in mirusviruses and micromonas points to the possibility that green algae may have been hosts for certain '*Mirusviricota*' lineages. Taken together, the evidence from biogeographic distribution, functional gene profiles, and metatranscriptomic data suggests that mirusviruses may play an important role in shaping the ecology of key marine eukaryotic organism

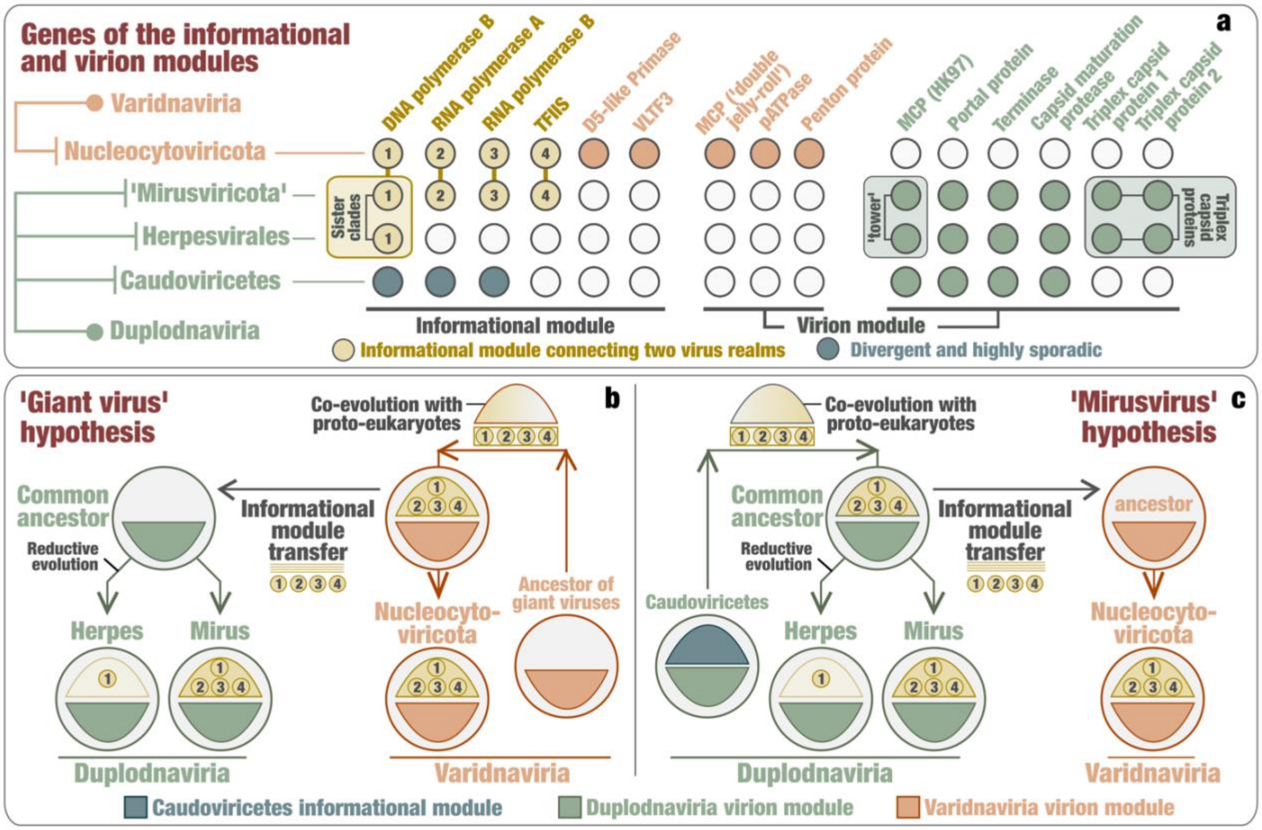


Fig. 5-4 Hypothesis of dsDNA virus evolution based on the discovery of '*Mirusviricota*'

Panel a presents an overview of the distribution of key genes within the informational and virion modules across different viral groups: *Nucleocytoviricota*, mirusviruses, herpesviruses, and *Caudoviricetes*. Panels b and c delve into two distinct evolutionary hypotheses explaining the origins of the informational module in eukaryote-infecting viruses within the realms of *Duplodnaviria* and *Varidnaviria*. Panel b describes the 'Giant Virus Hypothesis,' which posits that the informational module first emerged in the ancestor of *Nucleocytoviricota*. Conversely, Panel c outlines the 'Mirusvirus Hypothesis,' suggesting that this module initially appeared in the ancestor of mirusviruses. Both scenarios provide different perspectives on the evolutionary pathways of these significant viral realms. This figure has been published and modified under the CC-BY license from the paper by M. Gaia#, L. Meng# et al., 2023.

Some small varidnaviruses have been speculated that might represent evolutionary intermediates between phages and giant viruses. Owing to the chimeric attributes, '*Mirusviricota*' sheds light on these evolutionary mysteries. Two possible scenarios were proposed based on all the findings (Fig. 5-4). The first one is informational module may have been transferred from *Nucleocytoviricota* to the ancestor of mirusviruses and herpesviruses, contributing to the complexity of eukaryotic duplodnaviruses. In this scenario, a *Nucleocytoviricota* virus might have exchanged its virion module with that of an unidentified duplodnavirus co-infecting the same host, while retaining its advanced informational module. The second scenario is that an ancient transfer of the informational module occurred from the ancestorsof mirusvirus to small and simple ancestors of *Nucleocytoviricota*, as suggested by the mirusvirus origin hypothesis. This could elucidate the significant evolutionary jump from 'small' varidnaviruses to the complex *Nucleocytoviricota*. Regardless of which hypothesis is considered, mirusviruses shed light on the evolutionary path of eukaryotic double-stranded DNA viruses from both realms.

Chapter 6: Conclusion, Limitations and Perspective

6.1 Importance and achievements

This dissertation has significantly contributed to the understanding of the ecology and evolution of giant viruses in marine environments. Here I list key findings and implications from each chapter. In Chapter 2, the biogeographical analysis of giant viruses underscores the critical influence of temperature on their distribution. This insight is vital for understanding how environmental factors shape the presence and proliferation of marine viruses. In Chapter 3, The use of co-occurrence networks for predicting virus-host interactions, while promising, faces limitations due to the complexity of microbial interactions and time-lagged abundance correlations. This chapter marked a significant step forward in quantitatively assessing and improving host predictions based on these networks, enhancing our understanding of virus-host dynamics. In Chapter 4, based on the biogeography results, investigating the functional gene repertoires of marine eukaryote-infecting large and giant DNA viruses revealed a clear divide between polar and nonpolar viral communities, indicating adaptations driven by genomic alterations. These adaptations, particularly in polar environments, are crucial in the context of global climate change and its impact on microbial ecosystems. Last but not the least, in Chapter 5, the discovery of '*Mirusviricota*' highlights the evolutionary connection between two viral realms and suggests ancient gene exchanges. This group's prevalence in the global ocean underscores its potential role in infecting plankton, marking an important evolutionary narrative.

6.2 Limitations

The study, while extensive in its metagenomic approach, is constrained by several limitations that warrant attention for a more comprehensive understanding of marine giant viruses. The primary limitation lies in the exclusive reliance on metagenomic data, which, although informative, lacks the concreteness of experimental validation. For example, this gap particularly affects our grasp of the life cycle and physiological characteristics of the newly identified '*Mirusviricota*'. The inability to isolate these novel viruses and their hosts leaves a significant void in understanding their ecological roles and interactions in marine environments. Moreover, the adaptation scenario and polar specific functions also call for the further experimental evolutions.

Additionally, the interpretations and classifications within this study are contingent upon the existing ICTV evolutionary model. The prevailing scenario, which emphasizes distinct origins for the informational and virion modules, is widely accepted due to its logical robustness, alignment with virological traits, and extensive support within the scientific community. However, it's notable to acknowledge that there is still some debate regarding the higher ranks of the taxonomic system based on the model (Caetano-Anollés et al., 2023). While future alterations in evolution models and taxonomy systems will not significantly change the results and findings of this dissertation, it is anticipated that the current model need and will undergo further revisions, potentially influencing the interpretation of findings written in this dissertation. Conversely, future theoretical frameworks, necessitating continuous reassessment of our understanding, also need to be constructed based on a growing body of findings, such as those presented in this dissertation.

6.3 Conclusion and future plans

Looking ahead, the dissertation sets a groundwork for future explorations in the realm of marine virology. The next phase of research should ideally focus on experimental validation, particularly aiming to isolate and study '*Mirusviricota*' viruses and their hosts. This will not only validate the findings derived from metagenomic data but also provide deeper insights into the physiological and ecological aspects of these viruses. Furthermore, the study highlights the imperative to delve deeper into the environmental factors influencing the distribution and evolution of marine viruses. Such research is critical in the era of global climate change, where understanding the resilience and adaptability of marine life (including viruses) to changing environmental conditions becomes increasingly important. Modelling work on climate change should consider involving viruses. The development and refinement of predictive models for virus-host interactions is another crucial area for future research. The current study's application of co-occurrence networks opens new avenues in this direction, but also underscores the complexity of marine microbial interactions and the challenges in accurately predicting these dynamics. Some high-throughput experimental approaches, such as single-cell sequencing, struggle with distinguishing between infection and predation, or are limited by their low sensitivity on environmental viruses. Recently, a improved single-cell metatranscriptomics for identifying the native hosts of giant viruses has been established (Fromm et al., 2023). Such methods could be a viable solution for validating virus-host pairs in future research.

In the long term, this dissertation could enhance our understanding of giant virus ecology in the ocean. It sets the stage for comprehensive studies that could unravel the mechanisms governing the interactions between marine viruses and their hosts, their evolutionary paths, and their roles in marine ecosystems. This knowledge is fundamental for virology and marine biology, as well as for broader ecological and environmental sciences. It contributes significantly to the understanding of life in one of the planet's most expansive and diverse habitats.

Data Availability

The metagenome data from Tara Oceans is available at the ENA under accession PRJEB402 (https://www.ebi.ac.uk/ena/browser/view/PRJEB402), the metadata of metagenomes used in this study was summarized in the Supplementary Data 1. FASTA files for the 1,380 giant virus genomes from the Global Ocean Eukaryotic Viral (GOEV) database can be accessed via https://doi.org/10.6084/m9.figshare.20284713. Additionally, the accession numbers of 1,593 non-redundant marine *Nucleocytoviricota* and mirusvirus MAGs and 224 reference genomes in the GOEV database is provided in the Supplementary Data 2. There are other data used in this study: Giant Virus Orthologous Groups (GVOGs) database (https://faylward.github.io/GVDB/); Virus-Host Database (https://www.genome.jp/virushostdb); Tara Oceans Eukaryotic Genomes Database (https://www.genoscope.cns.fr/tara); NCBI database (https://www.ncbi.nlm.nih.gov/genome) . The data utilized in this study can be accessed from GenomeNet at https://www.genome.jp/ftp/db/community/tara/PolarAdaptaiton/data/; https://www.genome.jp/ftp/db/community/tara/Cooccurrence. Source data are provided with this paper. The script used to calculate robust ecological optima is available at https://github.com/LingjieEcoEvo/PolarAdaptaiton/tree/main/optimum

Acknowledgements

I would like to express my sincere thanks and gratitude to my Ph.D. supervisor, Professor Hiroyuki Ogata. The decision to leave China and embark on this educational journey in Japan was a substantial transition in my life. However, at the beginning, the life was tough, and I was not meeting the expected standards, largely because of gaps in my academic background in both viruses and bioinformatics. It's thanks to Prof. Ogata's trust and mentorship that I have been able to grow academically and reach this milestone of graduation. I deeply appreciate the confidence he placed in me and the invaluable training I've received under his supervision. I want to thank every professor, Hisashi Endo, Romain Blanc-Mathieu, Yusuke Okazaki, Hiroyuki Hikida in Ogata Lab. Each one of them is an exceptional researcher who has set a high standard for academic excellence. I also want to thank everyone in Ogata Lab. A special thank you goes to Russell Young Neches, Junyi Wu, Rodrigo Hernández-Velázquez, and Hiroto Kaneko for their valuable contributions to my projects.

I would also express my gratitude to Dr. Tom O. Delmont for his indispensable contributions. His work on the GOEV database has not only provided a foundational framework that has significantly aided my own research, but has also enriched the broader field of virus research. His effective organization of teamwork has been both instructive and inspirational and his encouragement has been a great support. I want to thank Dr. Morgan Gaïa for his invaluable discussions and insights on the phylogeny of giant viruses. I'm grateful for the collaborative spirit that enabled us to complete a significant piece of work as co-authors. I want to thank Dr. Mart Krupovic. His intellectual contributions have elevated the importance of '*Mirusviricota*' to a new level in the viral evolution. I also want to thank Prof. Samuel Chaffron, Prof. Patrick Forterre and all my collogues in the projects.

Take the opportunity of this Ph.D. dissertation, I would like to extend special thanks to Prof. Minoru Kanehisa, renowned for developing the KEGG bioinformatics database, where I am currently employed. I also want to thank Bioinformatics Center and the supercomputer system in Institute for Chemical Research.

I would like to offer thanks to all my friends for their warm accompany.

Lastly, I owe a lot of gratitude to my parents for their unconditional support throughout my Ph.D. journey and for standing by all the choices I've made.

Reference

Abrahão, J., Silva, L., Silva, L. S., Khalil, J. Y. B., Rodrigues, R., Arantes, T., Assis, F., Boratto, P., Andrade, M., & Kroon, E. G. (2018). Tailed giant Tupanvirus possesses the most complete translational apparatus of the known virosphere. *Nature Communications*, *9* (1), 749.

Alfredo, J., Castruita, S., & Westbury, M. V. (2020). Analyses of key genes involved in Arctic adaptation in polar bears suggest selection on both standing variation and de novo mutations played an important role. *BMC Genomics*, *0*, 1–8.

Alneberg, J., Bjarnason, B. S., De Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., Lahti, L., Loman, N. J., Andersson, A. F., & Quince, C. (2014). Binning metagenomic contigs by coverage and composition. *Nature Methods*, *11* (11), 1144–1146.

Andreani, J., Khalil, J. Y. B., Baptiste, E., Hasni, I., Michelle, C., Raoult, D., Levasseur, A., & La Scola, B. (2018). Orpheovirus IHUMI-LCC2: a new virus among the giant viruses. *Frontiers in Microbiology*, *8*, 2643.

Arslan, D., Legendre, M., Seltzer, V., Abergel, C., & Claverie, J.-M. (2011). Distant Mimivirus relative with a larger genome highlights the fundamental features of Megaviridae. *Proceedings of the National Academy of Sciences of the United States of America*, *108* (42), 17486–17491. https://doi.org/10.1073/pnas.1110889108

Aylward, F. O., Abrahão, J. S., Brussaard, C. P. D., Fischer, M. G., Moniruzzaman, M., Ogata, H., & Suttle, C. A. (2023). Taxonomic update for giant viruses in the order Imitervirales (phylum Nucleocytoviricota). *Archives of Virology*, *168* (11), 283.

Aylward, F. O., Moniruzzaman, M., Ha, A. D., & Koonin, E. V. (2021). A phylogenomic framework for charting the diversity and evolution of giant viruses. *PLoS Biology*, *19* (10 October), 1–18. https://doi.org/10.1371/JOURNAL.PBIO.3001430

Bäckström, D., Yutin, N., Jørgensen, S. L., Dharamshi, J., Homa, F., Zaremba-Niedwiedzka, K., Spang, A., Wolf, Y. I., Koonin, E. V., & Ettema, T. J. G. G. (2019). Virus genomes from deep sea sediments expand the ocean megavirome and support independent origins of viral gigantism. *MBio*, *10* (2), 1–23. https://doi.org/10.1128/mBio.02497-18

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., & Schaeffer, R. D. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, *373* (6557), 871–876.

Bell, P. J. L. (2022). Eukaryogenesis: the rise of an emergent superorganism. *Frontiers in Microbiology*, *13*, 858064.

Bivand, R., Keitt, T., Rowlingson, B., Pebesma, E., Sumner, M., & Hijmans, R. (2018). rgdal: bindings for the geospatial data abstraction library 2017. *URL Https://CRAN. R-Project. Org/Package= Rgdal. R Package Version*, 1.

Blanc-Mathieu, R., Dahle, H., Hofgaard, A., Brandt, D., Ban, H., Kalinowski, J., Ogata, H., & Sandaa, R.-A. (2021). A persistent giant algal virus, with a unique morphology, encodes an unprecedented number of genes involved in energy metabolism. *Journal of Virology*, *95* (8), 10–1128.

Boyer, M., Yutin, N., Pagnier, I., Barrassi, L., Fournous, G., Espinosa, L., Robert, C., Azza, S., Sun, S., Rossmann, M. G., Suzan-Monti, M., La Scola, B., Koonin, E. V., & Raoult, D. (2009). Giant Marseillevirus highlights the role of amoebae as a melting pot in emergence of chimeric microorganisms. *Proceedings of the National Academy of Sciences of the United States of America*, *106* (51), 21848–21853. https://doi.org/10.1073/pnas.0911354106

Breitbart, M. Y. A., Thompson, L. R., Suttle, C. A., & Sullivan, M. B. (2007). Exploring the vast diversity of marine viruses. *Oceanography*, *20* (2), 135–139.

Buchfink, B., Reuter, K., & Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods*, *18* (4), 366–368.

Caetano-Anollés, G., Claverie, J.-M., & Nasir, A. (2023). A critical analysis of the current state of virus taxonomy. *Frontiers in Microbiology*, *14*.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*, 1–9. https://doi.org/10.1186/1471-2105-10-421

Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., & Huerta-Cepas, J. (2021). eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Molecular Biology and Evolution*, *38* (12), 5825–5829.

Carradec, Q., Pelletier, E., Da Silva, C., Alberti, A., Seeleuthner, Y., Blanc-Mathieu, R., Lima-Mendez, G., Rocha, F., Tirichine, L., Labadie, K., Kirilovsky, A., Bertrand, A., Engelen, S., Madoui, M. A., Méheust, R., Poulain, J., Romac, S., Richter, D. J., Yoshikawa, G., … Wincker, P. (2018). A global ocean atlas of eukaryotic genes. *Nature Communications*, *9* (1), 373. https://doi.org/10.1038/s41467-017-02342-1

Chow, C. E. T., & Suttle, C. A. (2015). Biogeography of Viruses in the Sea. *Annual Review of Virology*, *2*, 41–66. https://doi.org/10.1146/annurev-virology-031413-085540

Coenen, A. R., & Weitz, J. S. (2018). Limitations of Correlation-Based Inference in Complex Virus-Microbe Communities. *MSystems*, *3* (4), 7–9. https://doi.org/10.1128/msystems.00084-18

Colson, P., Aherfi, S., & La Scola, B. (2017). Evidence of giant viruses of amoebae in the human gut. *Human Microbiome Journal*, *5*, 14–19.

Colson, P., Gimenez, G., Boyer, M., Fournous, G., & Raoult, D. (2011). The giant Cafeteria roenbergensis virus that infects a widespread marine phagocytic protist is a new member of the fourth domain of life. *PLoS ONE*, *6* (4), 13–17. https://doi.org/10.1371/journal.pone.0018935

Da Cunha, V., Gaia, M., Ogata, H., Jaillon, O., Delmont, T. O., & Forterre, P. (2022). Giant viruses encode actin-related proteins. *Molecular Biology and Evolution*, *39* (2), msac022.

Deeg, C. M., Chow, C.-E. T., & Suttle, C. A. (2018). The kinetoplastid-infecting Bodo saltans virus (BsV), a window into the most abundant giant viruses in the sea. *Elife*, *7*, e33014.

Demory, D., Arsenieff, L., Simon, N., Six, C., Rigaut-Jalabert, F., Marie, D., Ge, P., Bigeard, E., Jacquet, S., Sciandra, A., Bernard, O., Rabouille, S., & Baudoux, A. C. (2017). Temperature is a key factor in Micromonas-virus interactions. *ISME Journal*, *11* (3), 601–612. https://doi.org/10.1038/ismej.2016.160

Demory, D., Weitz, J. S., Baudoux, A. C., Touzeau, S., Simon, N., Rabouille, S., Sciandra, A., & Bernard, O. (2021). A thermal trade-off between viral production and degradation drives virus-phytoplankton population dynamics. *Ecology Letters*, *24* (6), 1133–1144. https://doi.org/10.1111/ele.13722

DeVries, A. L., & Cheng, C. H. C. (2005). Antifreeze Proteins and Organismal Freezing Avoidance in Polar Fishes. *Fish Physiology*, *22* (C), 155–201. https://doi.org/10.1016/S1546-5098 (04)22004-0

Dominguez-Huerta, G., Zayed, A. A., Wainaina, J. M., Guo, J., Tian, F., Pratama, A. A., Bolduc, B., Mohssen, M., Zablocki, O., Pelletier, E., Delage, E., Alberti, A., Aury, J. M., Carradec, Q., da Silva, C., Labadie, K., Poulain, J., Bowler, C., Eveillard, D., … Sullivan, M. B. (2022). Diversity and ecological footprint of Global Ocean RNA viruses. *Science*, *376* (6598), 1202–1208. https://doi.org/10.1126/science.abn6358

Duffy, S., Shackelton, L. A., & Holmes, E. C. (2008). Rates of evolutionary change in viruses: patterns and determinants. *Nature Reviews Genetics*, *9* (4), 267–276.

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Computational Biology*, *7* (10), e1002195.

Endo, H., Blanc-Mathieu, R., Li, Y., Salazar, G., Henry, N., Labadie, K., de Vargas, C., Sullivan, M. B., Bowler, C., Wincker, P., Karp-Boss, L., Sunagawa, S., & Ogata, H. (2020). Biogeography of marine giant viruses reveals their interplay with eukaryotes and ecological functions. *Nature Ecology and Evolution*, *4* (12), 1639–1649. https://doi.org/10.1038/s41559-020-01288-w

Eren, A. M., Esen, O. C., Quince, C., Vineis, J. H., Morrison, H. G., Sogin, M. L., & Delmont, T. O. (2015). Anvi'o: An advanced analysis and visualization platformfor 'omics data. *PeerJ*, *2015* (10), 1–29. https://doi.org/10.7717/peerj.1319

Finke, J. F., Winget, D. M., Chan, A. M., & Suttle, C. A. (2017). Variation in the genetic repertoire of viruses infecting Micromonas pusilla reflects horizontal gene transfer and links to their environmental distribution. *Viruses*, *9* (5), 1–18. https://doi.org/10.3390/v9050116

Fischer, M. G., Allen, M. J., Wilson, W. H., & Suttle, C. A. (2010). Giant virus with a remarkable complement of genes infects marine zooplankton. *Proceedings of the National Academy of Sciences of the United States of America*, *107* (45), 19508–19513. https://doi.org/10.1073/pnas.1007615107

Fromm, A., Hevroni, G., Vincent, F., Schatz, D., Martinez-Gutierrez, C. A., Aylward, F. O., & Vardi, A. (2023). Homing in on the rare virosphere reveals the native host of giant viruses. *BioRxiv*.

Gaïa, M., Meng, L., Pelletier, E., Forterre, P., Vanni, C., Fernandez-Guerra, A., Jaillon, O., Wincker, P., Ogata, H., & Krupovic, M. (2023). Mirusviruses link herpesviruses to giant viruses. *Nature*, 1–7.

Gallot-Lavallée, L., & Blanc, G. (2017). A glimpse of nucleo-cytoplasmic large DNA virus biodiversity through the eukaryotic genomicswindow. *Viruses*, *9* (1), 17. https://doi.org/10.3390/v9010017

Gastrich, M. D., Leigh-Bell, J. A., Gobler, C. J., Anderson, O. R., Wilhelm, S. W., & Bryan, M. (2004). Viruses as potential regulators of regional brown tide blooms caused by the alga, Aureococcus anophagefferens. *Estuaries*, *27* (1), 112–119. https://doi.org/10.1007/BF02803565

Gilbert, N. E., LeCleir, G. R., Pound, H. L., Strzepek, R. F., Ellwood, M. J., Twining, B. S., Roux, S., Boyd, P. W., & Wilhelm, S. W. (2023). Giant Virus Infection Signatures Are Modulated by Euphotic Zone Depth Strata and Iron Regimes of the Subantarctic Southern Ocean. *MSystems*, *8* (2). https://doi.org/10.1128/msystems.01260-22

Gorbalenya, A. E., Krupovic, M., Mushegian, A., Kropinski, A. M., Siddell, S. G., Varsani, A., Adams, M. J., Davison, A. J., Dutilh, B. E., Harrach, B., Harrison, R. L., Junglen, S., King, A. M. Q., Knowles, N. J., Lefkowitz, E. J., Nibert, M. L., Rubino, L., Sabanadzovic, S., Sanfaçon, H., … Kuhn, J. H. (2020). The new scope of virus taxonomy: partitioning the virosphere into 15 hierarchical ranks. *Nature Microbiology*, *5* (5), 668–674. https://doi.org/10.1038/s41564-020-0709-x

Gregory, A. C., Zayed, A. A., Conceição-Neto, N., Temperton, B., Bolduc, B., Alberti, A., Ardyna, M., Arkhipova, K., Carmichael, M., Cruaud, C., Dimier, C., Domínguez-Huerta, G., Ferland, J., Kandels, S., Liu, Y., Marec, C., Pesant, S., Picheral, M., Pisarev, S., … Roux, S. (2019). Marine DNA Viral Macro- and Microdiversity from Pole to Pole. *Cell*, *177* (5), 1109-1123.e14. https://doi.org/10.1016/j.cell.2019.03.040

Guglielmini, J., Woo, A. C., Krupovic, M., Forterre, P., & Gaia, M. (2019). Diversification of giant and large eukaryotic dsDNA viruses predated the origin of modern eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America*, *116* (39), 19585–19592. https://doi.org/10.1073/pnas.1912006116

Hillebrand, H. (2004). On the generality of the latitudinal diversity gradient. *American Naturalist*, *163* (2), 192–211. https://doi.org/10.1086/381004

Hingamp, P., Grimsley, N., Acinas, S. G., Clerissi, C., Subirana, L., Poulain, J., Ferrera, I., Sarmento, H., Villar, E., Lima-Mendez, G., Faust, K., Sunagawa, S., Claverie, J. M., Moreau, H., Desdevises, Y., Bork, P., Raes, J., De Vargas, C., Karsenti, E., … Ogata, H. (2013). Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans microbial metagenomes. *ISME Journal*, *7* (9), 1678–1695. https://doi.org/10.1038/ismej.2013.59

Hopkins, D. R. (1983). *The Greatest Killer: Smallpox in History, with a New Introduction*. University of Chicago Press.

Ibarbalz, F. M., Henry, N., Brandão, M. C., Martini, S., Busseni, G., Byrne, H., Coelho, L. P., Endo, H., Gasol, J. M., Gregory, A. C., Mahé, F., Rigonato, J., Royo-Llonch, M., Salazar, G., Sanz-Sáez, I., Scalco, E., Soviadan, D., Zayed, A. A., Zingone, A., … Zinger, L. (2019a). Global Trends in Marine Plankton Diversity across Kingdoms of Life. *Cell*, *179* (5), 1084-1097.e21. https://doi.org/10.1016/j.cell.2019.10.008

Ibarbalz, F. M., Henry, N., Brandão, M. C., Martini, S., Busseni, G., Byrne, H., Coelho, L. P., Endo, H., Gasol, J. M., Gregory, A. C., Mahé, F., Rigonato, J., Royo-Llonch, M., Salazar, G., Sanz-Sáez, I., Scalco, E., Soviadan, D., Zayed, A. A., Zingone, A., … Zinger, L. (2019b). Global Trends in Marine Plankton Diversity across Kingdoms of Life. *Cell*, *179* (5), 1084-1097.e21. https://doi.org/10.1016/j.cell.2019.10.008

Iranzo, J., Krupovic, M., & Koonin, E. V. (2016). The double-stranded DNA virosphere as a modular hierarchical network of gene sharing. *MBio*, *7* (4), 1–21. https://doi.org/10.1128/mBio.00978-16

Irwin, N. A. T., Pittis, A. A., Richards, T. A., & Keeling, P. J. (2022). Systematic evaluation of horizontal gene transfer between eukaryotes and viruses. *Nature Microbiology*, *7* (2), 327–336. https://doi.org/10.1038/s41564-021-01026-3

Iyer, L. M., Balaji, S., Koonin, E. V, & Aravind, L. (2006). Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. *Virus Research*, *117* (1), 156–184.

Johannessen, T. V., Bratbak, G., Larsen, A., Ogata, H., Egge, E. S., Edvardsen, B., Eikrem, W., & Sandaa, R.-A. (2015). Characterisation of three novel giant viruses reveals huge diversity among viruses infecting Prymnesiales (Haptophyta). *Virology*, *476*, 180–188.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596* (7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kaneko, H., Blanc-Mathieu, R., Endo, H., Chaffron, S., Delmont, T. O., Gaia, M., Henry, N., Hernández-Velázquez, R., Nguyen, C. H., Mamitsuka, H., Forterre, P., Jaillon, O., De Vargas, C., Sullivan, M. B., Suttle, C. A., Guidi, L., & Ogata, H. (2021). Eukaryotic virus composition can predict the efficiency of carbon export in the global ocean. *IScience*, *24* (24), 102002. https://doi.org/doi.org/10.1101/710228

Kelly, L., Ding, H., Huang, K. H., Osburne, M. S., & Chisholm, S. W. (2013). Genetic diversity in cultured and wild marine cyanomyoviruses reveals phosphorus stress as a strong selective agent. *The ISME Journal*, *7* (9), 1827–1841.

Kijima, S., Delmont, T. O., Miyazaki, U., Gaia, M., Endo, H., & Ogata, H. (2021). Discovery of viral myosin genes with complex evolutionary history within plankton. *Frontiers in Microbiology*, *12*, 683294.

Koonin, E. V, Dolja, V. V, & Krupovic, M. (2015). Origins and evolution of viruses of eukaryotes: The ultimate modularity. *Virology*, *479*, 2–25.

Koonin, E. V, Dolja, V. V, Krupovic, M., Varsani, A., Wolf, Y. I., Yutin, N., Zerbini, F. M., & Kuhn, J. H. (2020a). crossm Global Organization and Proposed Megataxonomy of the. *Microbiology and Molecular Biology Reviews*, *84* (2), e00061-19.

Koonin, E. V., Dolja, V. V., Krupovic, M., Varsani, A., Wolf, Y. I., Yutin, N., Zerbini, F. M., & Kuhn, J. H. (2020b). Global Organization and Proposed Megataxonomy of the Virus World. *Microbiology and Molecular Biology Reviews*, *84* (2), e00061-19. https://doi.org/10.1128/mmbr.00061-19

Koonin, E. V., & Yutin, N. (2019). Evolution of the Large Nucleocytoplasmic DNA Viruses of Eukaryotes and Convergent Origins of Viral Gigantism. In *Advances in Virus Research* (1st ed., Vol. 103). Elsevier Inc. https://doi.org/10.1016/bs.aivir.2018.09.002

Krupovic, M., Dolja, V. V, & Koonin, E. V. (2019). Origin of viruses: primordial replicators recruiting capsids from hosts. *Nature Reviews Microbiology*, *17* (7), 449–458.

Krupovic, M., Dolja, V. V., & Koonin, E. V. (2020). The LUCA and its complex virome. *Nature Reviews Microbiology*, *18* (11), 661–670. https://doi.org/10.1038/s41579-020-0408-x

Krupovic, M., & Koonin, E. V. (2015). Polintons: a hotbed of eukaryotic virus, transposon and plasmid evolution. *Nature Reviews Microbiology*, *13* (2), 105–115.

Krupovic, M., Yutin, N., & Koonin, E. (2020). Evolution of a major virion protein of the giant pandoraviruses from an inactivated bacterial glycoside hydrolase. *Virus Evolution*, *6* (2), 1–8. https://doi.org/10.1093/ve/veaa059

La Scola, B., Audic, S., Robert, C., Jungang, L., De Lamballerie, X., Drancourt, M., Birtles, R., Claverie, J. M., & Raoult, D. (2003). A giant virus in amoebae. *Science*, *299* (5615), 2033. https://doi.org/10.1126/science.1081867

Legendre, M., Bartoli, J., Shmakova, L., Jeudy, S., Labadie, K., Adrait, A., Lescot, M., Poirot, O., Bertaux, L., Bruley, C., Couté, Y., Rivkina, E., Abergel, C., & Claverie, J.-M. M. (2014). Thirtythousand-year-old distant relative of giant icosahedral DNA viruses with a pandoravirus morphology. *Proceedings of the National Academy of Sciences of the United States of America*, *111* (11), 4274–4279. https://doi.org/10.1073/pnas.1320670111

Li, D., Liu, C.-M., Luo, R., Sadakane, K., & Lam, T.-W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, *31* (10), 1674–1676.

Li, W., & Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22* (13), 1658–1659.

Li, Y., Hingamp, P., Watai, H., Endo, H., Yoshida, T., & Ogata, H. (2018). Degenerate PCR primers to reveal the diversity of giant viruses in coastal waters. *Viruses*, *10* (9), 496. https://doi.org/10.3390/v10090496

Louca, S., & Doebeli, M. (2018). Efficient comparative phylogenetics on large trees. *Bioinformatics*, *34* (6), 1053–1055.

Lowe, T. M., & Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*, *25* (5), 955–964.

M Boratto, P. V, Oliveira, G. P., Machado, T. B., Cláudia P Andrade, A. S., Baudoin, J.-P. P., Klose, T., Schulz, F., Azza, S., Decloquement, P., Chabrière, E., Colson, P., Levasseur, A., La Scola, B., Abrahão, J. S., Boratto, P. V. M. M., Oliveira, G. P., Machado, T. B., Andrade, A. C. S. P. P., Baudoin, J.-P. P., … Abrahão, J. S. (2020). Yaravirus: A novel 80-nm virus infecting Acanthamoeba castellanii. *Proceedings of the National Academy of Sciences of the United States of America*, *117* (28), 16579–16586. https://doi.org/10.1073/pnas.2001637117

Machado, T. B., Picorelli, A. C. R., de Azevedo, B. L., de Aquino, I. L. M., Queiroz, V. F., Rodrigues, R. A. L., Araújo Jr, J. P., Ullmann, L. S., dos Santos, T. M., & Marques, R. E. (2023).

Gene duplication as a major force driving the genome expansion in some giant viruses. *Journal of Virology*, e01309-23.

Mann, N. H. (2003). Phages of the marine cyanobacterial picophytoplankton. *FEMS Microbiology Reviews*, *27* (1), 17–34.

Martínez Martínez, J., Schroeder, D. C., Larsen, A., Bratbak, G., & Wilson, W. H. (2007). Molecular dynamics of Emiliania huxleyi and cooccurring viruses during two separate mesocosm studies. *Applied and Environmental Microbiology*, *73* (2), 554–562. https://doi.org/10.1128/AEM.00864-06

Matsuyama, T., Takano, T., Nishiki, I., Fujiwara, A., Kiryu, I., Inada, M., Sakai, T., Terashima, S., Matsuura, Y., & Isowa, K. (2020). A novel Asfarvirus-like virus identified as a potential cause of mass mortality of abalone. *Scientific Reports*, *10* (1), 4620.

Meints, R. H., Van Etten, J. L., Kuczmarski, D., Lee, K., & Ang, B. (1981). Viral infection of the symbiotic chlorella-like alga present in Hydra viridis. *Virology*, *113* (2), 698–703.

Meng, L., Delmont, T. O., Gaïa, M., Pelletier, E., Fernàndez-Guerra, A., Chaffron, S., Neches, R. Y., Wu, J., Kaneko, H., Endo, H., & Ogata, H. (2023). Genomic adaptation of giant viruses in polar oceans. *Nature Communications*, *14* (1), 6233. https://doi.org/10.1038/s41467-023-41910-6

Meng, L., Endo, H., Blanc-Mathieu, R., Chaffron, S., Hernández-Velázquez, R., Kaneko, H., & Ogata, H. (2021). Quantitative Assessment of Nucleocytoplasmic Large DNA Virus and Host Interactions Predicted by Co-occurrence Analyses. *MSphere*, *6* (2). https://doi.org/10.1128/msphere.01298-20

Methé, B. A., Nelson, K. E., Deming, J. W., Momen, B., Melamud, E., Zhang, X., Moult, J., Madupu, R., Nelson, W. C., Dodson, R. J., Brinkac, L. M., Daugherty, S. C., Durkin, A. S., DeBoy, R. T., Kolonay, J. F., Sullivan, S. A., Zhou, L., Davidsen, T. M., Wu, M., … Fraser, C. M. (2005). The psychrophilic lifestyle as revealed by the genome sequence of Colwellia psychrerythraea 34H through genomic and proteomic analyses. *Proceedings of the National*

Academy of Sciences of the United States of America, *102* (31), 10913–10918. https://doi.org/10.1073/pnas.0504766102

Mihara, T., Koyano, H., Hingamp, P., Grimsley, N., Goto, S., & Ogata, H. (2018). Taxon richness of "Megaviridae" exceeds those of bacteria and archaea in the ocean. *Microbes and Environments*, *33* (2), 162–171. https://doi.org/10.1264/jsme2.ME17203

Mihara, T., Nishimura, Y., Shimizu, Y., Nishiyama, H., Yoshikawa, G., Uehara, H., Hingamp, P., Goto, S., & Ogata, H. (2016). Linking virus genomes with host taxonomy. *Viruses*, *8* (3), 10–15. https://doi.org/10.3390/v8030066

Monier, A., Claverie, J. M., & Ogata, H. (2008). Taxonomic distribution of large DNA viruses in the sea. *Genome Biology*, *9* (7), R106. https://doi.org/10.1186/gb-2008-9-7-r106

Moniruzzaman, M., Erazo-Garcia, M. P., & Aylward, F. O. (2022). Endogenous giant viruses contribute to intraspecies genomic variability in the model green alga Chlamydomonas reinhardtii. *Virus Evolution*, *8* (2), veac102.

Moniruzzaman, M., Martinez-Gutierrez, C. A., Weinheimer, A. R., & Aylward, F. O. (2020). Dynamic genome evolution and complex virocell metabolism of globally-distributed giant viruses. *Nature Communications*, *11* (1), 1–11. https://doi.org/10.1038/s41467-020-15507-2

Moniruzzaman, M., Weinheimer, A. R., Martinez-Gutierrez, C. A., & Aylward, F. O. (2020). Widespread endogenization of giant viruses shapes genomes of green algae. *Nature*, *588* (7836), 141–145. https://doi.org/10.1038/s41586-020-2924-2

Moniruzzaman, M., Wurch, L. L., Alexander, H., Dyhrman, S. T., Gobler, C. J., & Wilhelm, S. W. (2017). Virus-host relationships of marine single-celled eukaryotes resolved from metatranscriptomics. *Nature Communications*, *8*, 16054. https://doi.org/10.1038/ncomms16054

Nagasaki, K., & Yamaguchi, M. (1997). Isolation of a virus infectious to the harmful bloom causing microalga Heterosigma akashiwo (Raphidophyceae). *Aquatic Microbial Ecology*, *13* (2), 135–140. https://doi.org/10.3354/ame013135

Needham, D. M., Yoshizawa, S., Hosaka, T., Poirier, C., Choi, C. J., Hehenberger, E., Irwin, N. A. T., Wilken, S., Yung, C. M., Bachy, C., Kurihara, R., Nakajima, Y., Kojima, K., Kimura-Someya, T., Leonard, G., Malmstrom, R. R., Mende, D. R., Olson, D. K., Sudo, Y., … Worden, A. Z. (2019). A distinct lineage of giant viruses brings a rhodopsin photosystem to unicellular marine predators. *Proceedings of the National Academy of Sciences of the United States of America*, *116* (41), 20574–20583. https://doi.org/10.1073/pnas.1907517116

Ogata, H., Toyoda, K., Tomaru, Y., Nakayama, N., Shirai, Y., Claverie, J. M., & Nagasaki, K. (2009). Remarkable sequence similarity between the dinoflagellate-infecting marine girus and the terrestrial pathogen African swine fever virus. *Virology Journal*, *6*, 1–8. https://doi.org/10.1186/1743-422X-6-178

Ogunbunmi, E. T., Love, S. D., Rhodes, K. A., Morales, A., Wilch, M. H., Jonas, J., & Fane, B. A. (2022). Low-Temperature Adaptation Targets Genome Packing Reactions in an Icosahedral Single-Stranded DNA Virus. *Journal of Virology*, *96* (7). https://doi.org/10.1128/jvi.01970-21

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, H. H., Szoecs, E., & Wagner, H. (2018). vegan: Community Ecology Package. R package. version 2.5-3. *Https://CRAN.R-Project.Org/Package=vegan*.

Paradis, E., & Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, *35* (3), 526–528.

Paul Shannon, Andrew Markiel, Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: A Software Environment for Integrated Models. *Genome Research*, *13* (22), 426. https://doi.org/10.1101/gr.1239303.metabolite

Pennell, M. W., Eastman, J. M., Slater, G. J., Brown, J. W., Uyeda, J. C., FitzJohn, R. G., Alfaro, M. E., & Harmon, L. J. (2014). geiger v2. 0: an expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. *Bioinformatics*, *30* (15), 2216–2218.

Philippe, N., Legendre, M., Doutre, G., Couté, Y., Poirot, O., Lescot, M., Arslan, D., Seltzer, V., Bertaux, L., Bruley, C., Garin, J., Claverie, J. M., & Abergel, C. (2013). Pandoraviruses: Amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. *Science*, *341* (6143), 281–286. https://doi.org/10.1126/science.1239181

Raoult, D., Audic, S., Robert, C., Abergel, C., Renesto, P., Ogata, H., La Scola, B., Suzan, M., & Claverie, J.-M. (2004). The 1.2-megabase genome sequence of Mimivirus. *Science*, *306* (5700), 1344–1350.

Rosenwasser, S., Mausz, M. A., Schatz, D., Sheyn, U., Malitsky, S., Aharoni, A., Weinstock, E., Tzfadia, O., Ben-Dor, S., Feldmesser, E., Pohnert, G., & Vardi, A. (2014). Rewiring host lipid metabolism by large viruses determines the fate of Emiliania huxleyi, a bloom-forming alga in the ocean. *Plant Cell*, *26* (6), 2689–2707. https://doi.org/10.1105/tpc.114.125641

Roux, S., Adriaenssens, E. M., Dutilh, B. E., Koonin, E. V., Kropinski, A. M., Krupovic, M., Kuhn, J. H., Lavigne, R., Brister, J. R., Varsani, A., Amid, C., Aziz, R. K., Bordenstein, S. R., Bork, P., Breitbart, M., Cochrane, G. R., Daly, R. A., Desnues, C., Duhaime, M. B., … Eloe-Fadrosh, E. A. (2019). Minimum information about an uncultivated virus genome (MIUVIG). *Nature Biotechnology*, *37* (1), 29–37. https://doi.org/10.1038/nbt.4306

Roux, S., Chan, L. K., Egan, R., Malmstrom, R. R., McMahon, K. D., & Sullivan, M. B. (2017). Ecogenomics of virophages and their giant virus hosts assessed through time series metagenomics. *Nature Communications*, *8* (1), 858. https://doi.org/10.1038/s41467-017-01086-2

Roux, S., Páez-Espino, D., Chen, I. M. A., Palaniappan, K., Ratner, A., Chu, K., Reddy, T., Nayfach, S., Schulz, F., Call, L., Neches, R. Y., Woyke, T., Ivanova, N. N., Eloe-Fadrosh, E. A., & Kyrpides, N. C. (2021). IMG/VR v3: An integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses. *Nucleic Acids Research*, *49* (D1), D764–D775. https://doi.org/10.1093/nar/gkaa946

Schulz, F., Abergel, C., & Woyke, T. (2022). Giant virus biology and diversity in the era of genome-resolved metagenomics. *Nature Reviews Microbiology*, *20* (12), 721–736. https://doi.org/10.1038/s41579-022-00754-5

Schulz, F., Alteio, L., Goudeau, D., Ryan, E. M., Yu, F. B., Malmstrom, R. R., Blanchard, J., & Woyke, T. (2018). Hidden diversity of soil giant viruses. *Nature Communications*, *9* (1), 1–9. https://doi.org/10.1038/s41467-018-07335-2

Schulz, F., Roux, S., Paez-Espino, D., Jungbluth, S., Walsh, D. A., Denef, V. J., McMahon, K. D., Konstantinidis, K. T., Eloe-Fadrosh, E. A., Kyrpides, N. C., & Woyke, T. (2020). Giant virus diversity and host interactions through global metagenomics. *Nature*, *578* (7795), 432–436. https://doi.org/10.1038/s41586-020-1957-x

Schulz, F., Yutin, N., Ivanova, N. N., Ortega, D. R., Lee, T. K., Vierheilig, J., Daims, H., Horn, M., Wagner, M., Jensen, G. J., Kyrpides, N. C., Koonin, E. V, & Woyke, T. (2017). *Giant viruses with an expanded complement of translation system components*. *85* (April), 82–85. http://science.sciencemag.org/

Siddell, S. G., Smith, D. B., Adriaenssens, E., Alfenas-Zerbini, P., Dutilh, B. E., Garcia, M. L., Junglen, S., Krupovic, M., Kuhn, J. H., & Lambert, A. J. (2023). Virus taxonomy and the role of the International Committee on Taxonomy of Viruses (ICTV). *Journal of General Virology*, *104* (5), 1840.

Subramaniam, K., Behringer, D. C., Bojko, J., Yutin, N., Clark, A. S., Bateman, K. S., van Aerle, R., Bass, D., Kerr, R. C., Koonin, E. V., Stentiford, G. D., & Waltzek, T. B. (2020). A new family of DNA viruses causing disease in crustaceans from diverse aquatic biomes. *MBio*, *11* (1), 1–14. https://doi.org/10.1128/mBio.02938-19

Sullivan, M. B. (2015). Viromes, Not Gene Markers, for Studying Double-Stranded DNA Virus Communities. *Journal of Virology*, *89* (5), 2459–2461. https://doi.org/10.1128/jvi.03289-14

Sun, T.-W. W., Yang, C.-L. L., Kao, T.-T. T., Wang, T.-H. H., Lai, M.-W. W., & Ku, C. (2020). Host Range and Coding Potential of Eukaryotic Giant Viruses. *Viruses*, *12* (11), 1337. https://doi.org/10.3390/v12111337

Sunagawa, S., Acinas, S. G., Bork, P., Bowler, C., Acinas, S. G., Babin, M., Bork, P., Boss, E., Bowler, C., Cochrane, G., de Vargas, C., Follows, M., Gorsky, G., Grimsley, N., Guidi, L., Hingamp, P., Iudicone, D., Jaillon, O., Kandels, S., … de Vargas, C. (2020). Tara Oceans: towards global ocean ecosystems biology. *Nature Reviews Microbiology*, *18* (8), 428–445. https://doi.org/10.1038/s41579-020-0364-5

Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D. R., Alberti, A., Cornejo-Castillo, F. M., Costea, P. I., Cruaud, C., D'Ovidio, F., Engelen, S., Ferrera, I., Gasol, J. M., Guidi, L., Hildebrand, F., … Bork, P. (2015). Ocean plankton. Structure and function of the global ocean microbiome. *Science*, *348* (6237), 1261359. https://doi.org/10.1126/science.1261359

Suttle, C. A. (2007). Marine viruses - Major players in the global ecosystem. *Nature Reviews Microbiology*, *5* (10), 801–812. https://doi.org/10.1038/nrmicro1750

Tackmann, J., Matias Rodrigues, J. F., & von Mering, C. (2019). Rapid Inference of Direct Interactions in Large-Scale Ecological Networks from Heterogeneous Microbial Sequencing Data. *Cell Systems*, *9* (3), 286-296.e8. https://doi.org/10.1016/j.cels.2019.08.002

Tang, K. F. J., Redman, R. M., Pantoja, C. R., Le Groumellec, M., Duraisamy, P., & Lightner, D. V. (2007). Identification of an iridovirus in Acetes erythraeus (Sergestidae) and the development of in situ hybridization and PCR method for its detection. *Journal of Invertebrate Pathology*, *96* (3), 255–260.

Tarutani, K., Nagasaki, K., & Yamaguchi, M. (2000). Viral impacts on total abundance and clonal composition of the harmful bloom-forming phytoplankton: Heterosigma akashiwo. *Applied and Environmental Microbiology*, *66* (11), 4916–4920. https://doi.org/10.1128/AEM.66.11.4916-4920.2000

Tomaru, Y., Katanozaka, N., Nishida, K., Shirai, Y., Tarutani, K., Yamaguchi, M., & Nagasaki, K. (2004). Isolation and characterization of two distinct types of HcRNAV, a single-stranded RNA virus infecting the bivalve-killing microalga Heterocapsa circularisquama. *Aquatic Microbial Ecology*, *34* (3), 207–218. https://doi.org/10.3354/ame034207

Vanni, C., Schechter, M. S., Acinas, S. G., Barberán, A., Buttigieg, P. L., Casamayor, E. O., Delmont, T. O., Duarte, C. M., Eren, A. M., Finn, R. D., Kottmann, R., Mitchell, A., Sanchez, P., Siren, K., Steinegger, M., Glöckner, F. O., & Fernandez-Guerra, A. (2022). Unifying the known and unknown microbial coding sequence space. *ELife*, *11*, 1–60. https://doi.org/10.7554/eLife.67667

Whittington, R. J., Becker, J. A., & Dennis, M. M. (2010). Iridovirus infections in finfish–critical review with emphasis on ranaviruses. *Journal of Fish Diseases*, *33* (2), 95–122.

Wilson, W. H., Schroeder, D. C., Allen, M. J., Holden, M. T. G., Parkhill, J., Barrell, B. G., Churcher, C., Hamlin, N., Mungall, K., Norbertczak, H., Quail, M. A., Price, C., Rabbinowitsch, E., Walker, D., Craigon, M., Roy, D., & Ghazal, P. (2005). Complete genome sequence and lytic phase transcription profile of a Coccolithovirus. *Science*, *309* (5737), 1090–1092. https://doi.org/10.1126/science.1113109

Woo, A. C., Gaia, M., Guglielmini, J., Da Cunha, V., & Forterre, P. (2021). Phylogeny of the Varidnaviria morphogenesis module: congruence and incongruence with the tree of life and viral taxonomy. *Frontiers in Microbiology*, *12*, 704052.

Wu, J., Meng, L., Gaia, M., Hikida, H., Okazaki, Y., Endo, H., & Ogata, H. (2023). Gene transfer among viruses substantially contributes to gene gain of giant viruses. *BioRxiv*, 2009–2023.

Xia, J., Kameyama, S., Prodinger, F., Yoshida, T., Cho, K. H., Jung, J., Kang, S. H., Yang, E. J., Ogata, H., & Endo, H. (2022). Tight association between microbial eukaryote and giant virus communities in the Arctic Ocean. *Limnology and Oceanography*, *67* (6), 1343–1356. https://doi.org/10.1002/lno.12086

Yau, S., Krasovec, M., Felipe Benites, L., Rombauts, S., Groussin, M., Vancaester, E., Aury, J. M., Derelle, E., Desdevises, Y., Escande, M. L., Grimsley, N., Guy, J., Moreau, H., Sanchez-Brosseau, S., van de Peer, Y., Vandepoele, K., Gourbiere, S., & Piganeau, G. (2020). Virus-host coexistence in phytoplankton through the genomic lens. *Science Advances*, *6* (14), eaay2587. https://doi.org/10.1126/sciadv.aay2587

Yau, S., Lauro, F. M., DeMaere, M. Z., Brown, M. V., Thomas, T., Raftery, M. J., Andrews-Pfannkoch, C., Lewis, M., Hoffman, J. M., Gibson, J. A., & Cavicchioli, R. (2011). Virophage control of antarctic algal host-virus dynamics. *Proceedings of the National Academy of Sciences of the United States of America*, *108* (15), 6163–6168. https://doi.org/10.1073/pnas.1018221108

Yoshikawa, G., Blanc-Mathieu, R., Song, C., Kayama, Y., Mochizuki, T., Murata, K., Ogata, H., & Takemura, M. (2019). Medusavirus, a Novel Large DNA Virus Discovered from Hot Spring Water. *Journal of Virology*, *93* (8), 1–3. https://doi.org/10.1128/jvi.02130-18

Zhang, R., Takemura, M., Murata, K., & Ogata, H. (2023). "Mamonoviridae", a proposed new family of the phylum Nucleocytoviricota. *Archives of Virology*, *168* (3), 80.

Zhao, H., Zhang, R., Wu, J., Meng, L., Okazaki, Y., Hikida, H., & Ogata, H. (2023). A 1.5 Mb continuous endogenous viral region in the arbuscular mycorrhizal fungus Rhizophagus irregularis. *BioRxiv*, 2004–2023.