Giant virus biology and diversity in the era of genome-resolved metagenomics

Frederik Schulz, Chantal Abergel, Tanja Woyke

Giant viruses in an era of genome-resolved metagenomics

Authors
Frederik Schulz*1, Chantal Abergel2, Tanja Woyke*1

*corresponding authors: fschulz@lbl.gov, twoyke@lbl.gov

Author affiliations
1 DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, California, USA
2 Aix Marseille University, CNRS-AMU, Marseille, France

Abstract
The discovery of giant viruses, with capsids as large as some bacteria, megabase-range genomes, and a wide variety of traits typically found only in cellular organisms, was one of the most spectacular breakthroughs in biology. Until recently, most of our knowledge about these giant viruses came from ~100 species-level isolates for which genome sequences were available. However, these isolates were primarily derived from laboratory-based co-cultivation with few cultured protists and algae and thus did not reflect the true diversity of giant viruses in the environment. While virus co-cultures enabled important detailed insights into giant virus biology, many questions regarding the origin, evolution and ecological importance of giant viruses remain unanswered. With recent advances in sequencing technologies and bioinformatics, our understanding of giant viruses has drastically expanded, as genome-resolved metagenomics and environmental surveys of viral hallmark genes are beginning to shed light on the taxonomic breadth of this group of viruses, their biogeography and ecology, and their evolutionary trajectories. In this review, we summarize our current understanding of giant virus diversity and biology based on viral isolates, as laboratory cultivation has facilitated extensive insights into viral morphological features and infection strategies. We then expand on how more recent cultivation-independent approaches led to an unparalleled increase of the coding potential and diversity of the Nucleocytoviricota. We further elaborate on how metagenomics has revolutionized our perspective of giant viruses by revealing their distribution across our planet's biomes where they impact the biology and ecology of a wide range of eukaryotic hosts and ultimately affect global nutrient cycles.

Introduction

Large and giant viruses are part of a group of double stranded DNA viruses, the Nucleocytoplasmic Large DNA Viruses (NCLDV)[1,2], which constitutes the viral phylum Nucleocytoviricota[3]. Viruses of this phylum infect a wide range of eukaryotic hosts, from the tiniest known unicellular choanoflagellates to multicellular animals[4]. NCLDV typically replicate in so-called viral factories built in the host cytoplasm or use the host nucleus to replicate and sometimes assemble their progeny[5,6]. Hallmark features of these viruses are large genomes, greater than 100 kb and up to 2.5 Mb, and virions that can reach more than 2 microns in length[7]. The term giant virus was coined in the early 2000s, when the first virus with a genome in the megabase range was discovered; initial light microscopy observations led to the

assumption that its particles corresponded to a gram positive bacterial pathogen of amoebae[8,9]. More detailed ultrastructural analysis revealed the typical icosahedral shape of viruses and genome sequencing yielded a 1.2 Mb viral genome[10]. This virus was named “Mimivirus”, short for microbe mimicking virus, and represented an unexpected novelty in the viral world, not only due to its exceptional particle and genome sizes but also its coding potential that included several genes with possible roles in protein biosynthesis[10]. Since this discovery of giant viruses their coding potential has been full of surprises and the presence of hallmark genes of cellular life led to the hypothesis that these viruses might represent an enigmatic fourth domain of life[10–12]. Equally intriguing, much smaller viruses, so-called “virophages", were found to infect some NCLDV that have exclusively cytoplasmic infectious cycles. Virophages parasitize and sometimes eradicate their viral hosts[13]. Also discovered was a third partner coined transpoviron, which corresponds to a 7 kb dsDNA episome, able to propagate using both the giant virus and the virophage particles as vehicles[14,15].

For well over a decade, giant viruses had chiefly been discovered through cultivation-based approaches until very recently when virology followed the footsteps of microbial genomics by applying cultivation-independent metagenomics to investigate the biology of these viruses at an unparalleled pace. In the following, we build on a wealth of experimental data that has revealed many insights into giant virus biology, in particular their virion structure and unique infection strategies. We expand this framework by integrating the latest sequence-based studies that expanded NCLDV diversity, biogeography, coding potential and putative host range. Further, there is now compelling evidence that the presence of a variety of cellular hallmark genes in giant virus genomes enable the virus to reprogram host metabolism, and that direct integration of giant virus genetic material into host genomes may impact the biology and evolution of the eukaryotic cell.

Giant virus discovery through isolation

The earliest discovered NCLDVs were the Poxviruses, the causative agents of smallpox, which were the first viral particles seen under a microscope more than 130 years ago[16]. The genome of Vaccinia virus was sequenced in the early 1990s[17] and shortly thereafter additional genomes of Poxviruses were sequenced (Figure 1) with sizes ranging from 120 to 360 kb[18]. Subsequently, other viruses that infect animals, including Ascoviruses, Iridoviruses and Asfarviruses were found and their genomes sequenced[19–21]. Genomes of viruses in these groups were comparably small with up to 220 kb, and even smaller in the recently discovered shrimp-associated Mininucleoviridae with only 70-80 kb[22]. In addition to animal-infecting NCLDV, a wide range of NCLDV were detected in diverse eukaryotic algae, including chlorophytes, haptophytes, pelagophytes, brown algae and dinoflagellates in the early 2000s[23]. These algae-associated NCLDV were affiliated with the Phycodnaviruses[23] and Mesomimivirinae[24], and although most of their genomes are around 200-500 kb[23,25], the genomes of Tetraselmis Virus and *Prymnesium kappa* virus RF01 are 668 kb[26] and 1.4 Mb[27], respectively.

After the discovery of Mimivirus in 2003[9] other NCLDV with larger virions and genomes above 500 kb have been found to infect heterotrophic protists[28], mainly members of the Amoebozoa.

For more than a decade *Acanthamoeba* strains had chiefly been used as bait for the co-cultivation of new viruses, leading to the frequent isolation of closely related giant viruses able to infect this unicellular host[29]. *Acanthamoeba* has proven as a particularly suitable host for many Megamimivirinae and Marseilleviridae[29]. Consequently, viruses from these taxonomic groups are currently among the most thoroughly cultivated NCLDV with more than 30 genome sequences readily available in public databases, including the novel Megamimivirinae lineages Tupanvirus[7] and Cotonvirus[30]. The co-cultivation approach has been widely successful and also led to the recovery of isolates from divergent NCLDV clades, facilitating the organization and naming of Pithoviruses, Pandoraviruses, Molliviruses and, most recently, Medusavirus[31]. More recently the use of alternative hosts, such as *Vermamoeba*, has led to the co-cultivation of several new Faustovirus isolates[32] and some Orpheoviruses[33], Pacmanvirus[34] and Kaumoebaviruses[35], all distant relatives of Pitho-and Asfarviruses. A newly developed high-throughput co-cultivation-based approach using high content screening has proven a powerful tool for giant virus discovery and isolation[36]. Yet, co-cultivation is limited by host specificity of giant viruses[4]; some NCLDV lineages are able to infect only specific hosts, such as certain species of *Acanthamoeba*[37], while others may be more versatile, exhibiting a broader host range[7]. Considering the immense diversity of eukaryotes[38], and in particular of microeukaryotes, it is likely that giant viruses that have been recovered through isolation reflect only a minute fraction of NCLDV lineages extant in the wild.

Virion structures and infection strategies in the Nucleocytoviricota

One of the best studied and most well understood infectious cycles is that of the amoeba-infecting Mimivirus[8]. Similar to other known members of the Mimiviridae it replicates in its host's cytoplasm[39–41] (Figure 2A). The ~700 nm icosahedral virion of Mimivirus (Figure 2B) contains RNA polymerase and transcript maturation machinery together with the genome. Mimivirus enters its host by triggering phagocytosis upon adhering to the host cell membrane with its glycosylated fibrils. Bacterial-type sugars are uniquely synthesized by the virally encoded glycosylation machinery and are the building blocks of the recently determined complex 70 kDa and 25 kDa polysaccharide structures that decorate the Mimivirus fibrils[42]. The structure of the Mimivirus capsid has also been determined, but due to its large size, the triangulation (T)-number remains unknown[43]. Detailed Atomic Force Microscopy provided additional insights into virion composition[44], further underlining the complexity of the capsid and an enclosed compartment, the so-called nucleoid, that contains hundreds of proteins in addition to the structural ones. It has been proposed that the non-structural proteins are required to initiate the viral infectious cycle, protect the virion from oxidative stress and perform early transcription[5,40]. Once in the host vacuole, the Mimivirus virion features a specific structure, the stargate, at one vertex of the icosahedron. Upon opening, the internal membrane is pulled outside the capsid to fuse with the host vacuole's membrane[45]. Next, the nucleoid is transferred into the host cytoplasm[46,47] where early transcription begins using the virally encoded transcription machinery, which, at first, remains confined in the nucleoid[48]. The accumulation of nucleic acids due to active transcription and replication leads to an increase in size of the virion factory and newly synthesized virions start budding at its periphery, recycling host cell membranes derived from the endoplasmic reticulum[46,48] or golgi[30]. The last step of virion maturation, after genome

loading into the nucleoid, is the addition of the fibril layer to the capsids[49], and then hundreds of newly synthesized virions are released after cell lysis.

Several viruses related to Mimivirus have similar infectious cycles but smaller virions. Among them is *Cafeteria roenbergensis* virus, which presents an icosahedral capsid of 300 nm in diameter (Figure 2B) with a lipidic membrane underneath the capsid shell. Its mode of infection is not fully understood but, similar to Mimivirus, a nucleoid structure has been observed in the cytoplasm as well as extracellular empty capsids, supporting an external opening of the capsids followed by the fusion of the internal membrane with that of the cell, thus allowing the transfer of the nucleoid into the host cytoplasm. Virions contain about 150 proteins, which either make up the icosahedral capsid or are necessary to initiate the infectious cycle[46]. Neo-synthesized virions assemble during the late stage of infection and are released through cell lysis. The structure of the complex capsid, determined by cryo-EM, corresponds to a T-number of 499 and provided a new model for capsid assembly[50]. Another member of the Mimiviridae, with a similar icosahedral capsid of 300 nm in diameter, is *Bodo saltans* virus. Its capsid appears to be made by two proteinaceous layers surrounded by 40nm long fibrils. A possible stargate-like structure is present at one vertex of the capsid and there are two membranes, one lining the external protein shell and one internal to the nucleoid compartment containing the genome. The infectious cycle is similar to the one of Mimivirus except that the host's nuclear genome appears to be degraded. The viral factory develops at the posterior pole of the cell to fill two thirds of the cell space, pushing aside the nucleus and organelles. Lipid vesicles are recruited for virion assembly that takes place at one side of the viral factory and mature virions detach after genome loading to migrate to the posterior pole of the cell. Virions are released by budding in vesicles from the host membrane and ultimately after cell lysis[41].

Virion morphology and infection mechanisms differ vastly between viruses of the Nucleocytoviricota. In particular, viruses that infect algae have been found to form small virions. Among the smallest members of the Nucleocytoviricota are Prasinoviruses with virion diameters of around 120 nm. This feature is important for infecting and replicating within *Ostreococcus,* which is one of the smallest free-living eukaryotes with only 0.8 µm cells[51]. Following viral infection, genome replication begins almost immediately and within hours new virions assemble in the cytoplasm; in less than 24 hours post-infection, cellular lysis of the host occurs. The host cell nucleus, mitochondria and chloroplast remain intact throughout this period. Slightly larger are Chloroviruses with 190 nm in diameter, an icosahedral shape (T-number 169) and an internal lipid membrane. A prototype for this group of viruses is *Paramecium bursaria* chlorella virus[52](Figure 2B). It has a spike-like structure at one vertex and a few external fibers that extend from some of the capsomers[53]. The outer capsid layer covers a single lipid bi-layered membrane, which is essential for infection. Its major capsid protein (Vp54) is a glycoprotein, and three Vp54s form a trimeric capsomere, which has pseudo-six-fold symmetry[54]. Some of the largest viruses that infect algae are affiliated with the Mimiviridae. All of them have icosahedral capsids with sizes ranging from 150 nm in *Aureococcus* virus (Figure 2B) to 370 nm in the recently described *Prymnesium kappa* virus[27]. These viruses also build a viral factory in the host cytoplasm but it is still unknown if the transcription machinery is loaded in the capsids allowing an entirely cytoplasmic infectious cycle.

Outside of the Mimiviridae there are smaller amoeba-infecting viruses such as the Melbournevirus, which has an icosahedral virion of ~250 nm diameter (Figure 2B). The Cryo-EM structure of the capsid was obtained for two members of the family at various resolutions revealing a T-number of 309 and a complex capsid structure[55–57] with many minor capsid proteins. Melbournevirus and other members of the Marseilleviruses are taken up by phagocytosis and then lose their icosahedral appearance to become spherical after the disappearance of the vacuole membrane. In contrast to Mimiviruses their RNA polymerase is not loaded in the virion. Instead, the cell nucleus starts changing appearance and becomes leaky through a still unknown mechanism triggered by the viral infection[58]. The nuclear proteins are recruited to the early viral factory, including the host RNA polymerase that will perform early transcription. Shortly after, the nucleus integrity is restored and the virally encoded RNA polymerase performs intermediate and late transcription[59] and icosahedral particles assemble inside the viral factory (Figure 2A). Mature particles can gather in large vesicles[60] and cell lysis leads to the release of both individual virions and filled vacuoles.

Larger amoeba-infecting viruses are Pandoraviruses, with amphora-shaped virions that can be up to 1 µm in length and 500 nm in diameter (Figure 2C). There is at least one lipidic membrane lining a thick tegument made of three layers including one made of cellulose[61]. The particles are taken up through phagocytosis followed by the opening of the virion ostiole-like structure to allow the fusion of the internal membrane with the phagosome and, consequently, delivery of the genome and necessary proteins into the cytoplasm. In contrast to Mimiviruses the Pandoraviruses build their viral factory in the host cell nucleus (Figure 2A). Although they encode RNA polymerase, there is none present in the capsids and thus they rely on the host cell for early transcription of viral genes. New virions start to assemble from the apex and the neo-synthesized virions are released either by cell lysis or by exocytosis through membrane fusion with the plasma membrane when they are in vacuoles[62,63].

Some of the largest virions found in the Nucleocytoviricota are those of the Pitho- and Cedratviruses (Figure 2C), which have very large amphora shaped capsids that can be up to 2 µm long and 600 nm wide. These capsids are closed by corks, one cork for Pithovirus[64,65] (Figure 2C) and two for Cedratvirus[66], that are made by proteins organized in a honeycomb array. Despite a virion morphology that closely resembles that of Pandoraviruses, the external tegument is different, appears to be made of parallel strips and does not present cellulose. The particle appears to be coated with short sparse fibrils[33,64]. The infectious cycle proceeds as for other amoeba-infecting viruses, by phagocytosis followed by capsid opening and membrane fusion with the phagosome[5]. For the Pithoviruses and Cedratviruses, the RNA polymerase loaded in the virion starts early transcription in the cytoplasm and the host nucleus remains intact during the entire infectious cycle. The virions are assembled from the cork and will lead to rectangular uncoated virions, the tegument being built by patches from reservoirs in the cytoplasm, and the morphology changes to amphora-shaped. The neo-synthesized virions can exit by exocytosis and upon cell lysis[64,66].

A different replication strategy has been described for Medusaviruses[31]; after uptake into the host cytoplasm, its DNA is replicated in the host nucleus while virions assemble in the cytoplasm. Medusaviruses have icosahedral virions that are 260 nm in diameter, covered by spherical-headed spikes extending from each capsomer, and a lipidic membrane surrounds the capsid interior. A low-resolution structure was determined by Cryo-EM, which returned a T-number of 277[67]. The mechanism of entry and egress of the Medusavirus virion from its host is yet to be determined.

As these examples illustrate, there is no shared blueprint for giant virus structure and infection mechanism; these characteristics vary between giant virus lineages and are likely shaped by the host organisms. The host range of the currently experimentally characterized giant viruses is limited to a few amoeba and algae lineages representing only a tiny fraction of the eukaryotic diversity. Thus, we expect that many more unusual virions and infection strategies will be revealed in the near future when new viruses will be captured together with their native hosts.

Cultivation-independent genomic approaches to studying giant viruses

Many important discoveries on giant virus biology and diversity have been made through giant virus isolation and cultivation. However, such approaches are constrained by the need to satisfy optimal growth requirements in a laboratory setting and are often restricted to lytic viruses. Cultivation-independent methods have been proven to be an indispensable tool to discover the genetic make-up of giant viruses from environmental samples.

In the earlier days of metagenomics, single-marker gene based surveys (Box 1) revealed that several marine NCLDV affiliated with *Phycodnaviridiae* and *Mimiviridae* were present in a wide range of marine metagenomes collected during the Tara Oceans and the Sargasso Sea expeditions[68,69] and that these viruses were more abundant in the photic layer than eukaryotes[69]. Interestingly, in a follow-up study, data from these surveys gave rise to the hypothesis that giant viruses are more diverse in the oceans than any cellular organism[70]. Subsequently, a large-scale analysis of the NCLDV major capsid protein (MCP), in which more than 50k of these proteins were found across Earth's biomes, highlighted the global dispersal of giant viruses, including in terrestrial ecosystems[71].

Other approaches that facilitated the discovery of novel NCLDV are single cell genomics and mini-metagenomics (Box 1). First, sorting single cells from marine samples enabled detection of NCLDV particles affiliated with viruses that had previously been found in the algae *Ostreococcus* sp. and *Phaeocystis globosa*[72]. This led to the sequencing of several so-called gvSAGs, of which the largest was a 813 kb genome affiliated with Mimiviridae that encoded a metacaspase, which potentially facilitates autocatalytic cell death of the host cell[73]. Single-cell methods were also used to identify and genome sequence 5 giant viruses associated with marine choanoflagellates[74,75]; comparative genomics together with all other NCLDV genomes revealed that viruses that infect hosts with similar trophic modes, including host habitat and lifestyles, express distinct genetic features[75,76]. Further, mini-metagenomics (Box 1) on a single forest soil sample led to the discovery of 15 diverse giant virus MAGs, including several members of the Klosneuvirinae, revealing that giant viruses are commonly found in soil[77].

The most successful approach for obtaining NCLDV genomes from environmental sequence data is genome-resolved metagenomics (Box 1). Since the early 2000s this approach has become common practice for recovering genomes of bacteria and archaea from complex environmental samples[78], yet it took nearly another decade before the first giant virus metagenome assembled genomes (GVMAGs) appeared in public databases (Figure 1). Yau and colleagues reconstructed the first GVMAGs as a byproduct of their work on virophages in metagenomes from the Organic Lake in Antarctica[79]. Several years later four additional potentially algae-associated GVMAGs were retrieved from environmental sequence data from Yellowstone lake and found to be related to phycodna- and mimiviruses, which shared some genes with virophages that co-occurred in the same sample[80]. Cultivation-independent approaches for discovery of giant virus genome-centric sequence information gained traction when members of a Mimiviridae-affiliated subfamily, the proposed Klosneuvirinae, were recovered from metagenomic data[81]. The fact that these were found in freshwater and sewage metagenomes from samples originating from four different continents suggested this novel group of giant viruses is cosmopolitan[81]. More than 20 giant virus MAGs from deep sea metagenomes were subsequently discovered, including 15 affiliated with Pithoviruses, indicating a surprisingly high prevalence of pithovirus-like viruses in the ocean[82] and additional likely algae-associated freshwater giant viruses from Dishui Lake[83,84]. The unique strength of cultivation-independent approaches to viral genomics became even more evident when over 2000 GVMAGs were extracted from metagenome datasets generated from thousands of samples collected from diverse biomes[71] plus an additional 500 GVMAGs collected shortly after from marine systems[85]. Viruses discovered through cultivation-independent approaches enabled a comprehensive update of the taxonomic framework of the Nucleocytoviricota[86] (Figure 3). The addition of the GVMAGs to the Nucleocytoviricota species tree led the phylogenetic diversity to increase by more than 10-fold and expanded their pangenome to more than 900,000 proteins[71]. This translated to an extensively expanded repertoire of functional genes, providing not only many novel insights into how giant viruses may interact with their hosts and the environment, but also generating novel hypotheses about their evolutionary roles[71,85,87,88].

Exploring the host ranges of giant viruses

Genome-resolved metagenomics facilitated the discovery of thousands of viral genomes, of which many represented lineages divergent from viruses recovered by isolation or co-cultivation[71,85] (Figure 3). However, NCLDV discovered through metagenomics typically lack direct connections to host organisms[89]. Approaches that have been used to overcome this limitation are the detection of viruses and potential eukaryotic hosts co-occurring in the same sample. Further, as horizontal swapping of genetic material between viruses and their hosts is a common phenomenon, the analysis of viral genes that may have been acquired through recent horizontal gene transfer (HGT) can potentially identify host organisms. In the early days of giant virus metagenomics, read-mapping based co-occurrence analysis (Box 1) revealed that the presence of giant virus sequences in some marine samples was positively correlated with those of Oomycetes[69], which is a eukaryote that has not yet been found associated with NCLDV. More recently, co-occurrences of sequences of NCLDV and microalgae, including haptophytes,

chlorophytes, dinophytes, pelagophytes and raphidophytes, were detected across a wide range of marine samples collected during the Tara Oceans expedition[90]. The strongest sequence co-occurrences were found between a virus affiliated with the *Mimiviridae* and Chrysophytes as its potential host[90]. Subsequent detection of putative HGT events between giant virus metagenome-assembled genomes and Chrysophyte genomes and transcriptomes provided further support for this host-virus connection[90]. A systematic analysis of HGT candidates present in more than 2000 NCLDV genomes, of which most were MAGs from diverse global sampling sites, revealed thousands of genes likely derived from host genomes through recent HGT[71]. Based on these results it was possible to propose connections between NCLDV and members of all major eukaryotic phyla[71]. Though most of these predicted hosts have not yet been found to be infected by giant viruses, more than 20 distinct connections could be confirmed through existing isolates[71] (Figure 4A).

While sequence-based computational host predictions provide a means to expand the range of putative NCLDV hosts, the approaches have some potential challenges and biases. For example, co-occurrence analysis is dependent on sufficient host genome coverage for detection in metagenome data and HGT analysis likewise requires the availability of the host genomic sequences. Further, it is difficult to detect ancient HGT from previous hosts, and the integration of NCLDV genes into host genomes whose sequences were deposited in public databases can potentially yield misleading results. The quality of databases used for the analyses is another important factor and the presence of NCLDV genomes misannotated as bacteria, archaea, or eukarya is a major weakness when employing automated tools for HGT detection[71,91]. Despite some of these limitations, expanding the putative host range of metagenome-derived NCLDV provides a basis for targeted sampling of putative hosts, for studying virus-host co-evolution, and for identifying viral-encoded functions for targeted modulation of host metabolism. Sequence-based inferences of viruses and their hosts may then be extrapolated to assess the impact of such interactions on the environment to enable predictive ecosystem modeling.

Not only is horizontal transfer of single genes a common phenomenon but some giant viruses can even integrate their entire genomes into the host's chromosome (Figure 4). This so-called endogenization is a term initially coined for a mechanism discovered in retroviruses[92,93]. Arrays of NCLDV genes have occasionally been found in genomes of eukaryotes, in particular in algae, plants[94–96] and amoebae[97–99]. A recent survey of published eukaryotic genomes and transcriptomes revealed the presence of giant virus genes in 66 different eukaryotes including several *Acanthamoeba* species, flagellates, ciliates, stramenopiles, oomycetes, fungi, arthropods and diverse unicellular and multicellular algae[100] (Figure 4). Yet, for many of these eukaryotes giant virus infections have not been observed. The integration of NCLDV genes often appears to be highly host specific, with viral genes detected in one eukaryotic species being unrelated to viral genes found in its sister species[100]. Among the integrated genes were NCLDV hallmark genes that were, in some instances, dispersed over the host chromosome whereas in other cases genes were co-localized in islands composed of more than 100 genes[100]. The integration of complete viral genomes has been described for Phaeoviruses; for example, *Ectocarpus siliculosus* virus integrated into its brown algal host more than 20 years ago[96] likely through use of integrases[101]. Phaeoviruses are lysogenic viruses leading to

persistent infections[102,103], which is in stark contrast to many other known NCLDV lineages that were successfully isolated based on the fact that they lyse their amoeba host[5]. In a recent study, the analysis of existing algal genomes and transcriptome data revealed other examples of whole giant virus genomes integrated into eukaryotic host genomes[104]. Some regions encoded more than 1,500 viral genes, up to 10% of the genes of the green algal host[104]. Several of the detected viral genes were annotated as enzymes with roles in carbohydrate metabolism, chromatin remodeling, signal transduction, energy production and translation[104].

It remains unknown if integrated giant viruses are dormant with no or minimal benefit to the host, or if the host cell benefits from some of the viral genes that may provide or finetune metabolic capabilities. Another unanswered question is whether there are mechanisms encoded in the integrated viral genome that may reinstate an active infection after transcribing and translating some of the integrated viral genes. This would then be followed by the release of the giant virus genetic material during host replication and effective dispersal to new hosts. If there is no reactivation of viral infection, it is conceivable that some giant virus genes would decay over time leading to pseudogenization. Genomics, metagenomics and metatranscriptomics enabled the discovery of NCLDV integration and revealed potential model systems, such as the green algae *Tetrabaena socialis* and *Chlamydomonas eustigma*[104] for experimental work. Future investigation of integration of giant virus genes is expected to provide some answers for how endogenization has shaped and continues to shape the evolution and ecology of eukaryotic organisms.

Reprogramming of the host and its impact on host populations

Reprogramming of the host can occur when the virus transforms its host into a so-called virocell[105] in which particular viral auxiliary metabolic genes augment host metabolism to further promote viral replication[106,107]. Analogous to bacteriophages, giant viruses seem to contribute genes to their hosts to augment and/or modulate metabolic capabilities of the host cell (Figure 5). One of the first described examples was a host-derived nitrogen transporter in *Ostreococcus taurus* virus that had been found expressed during the infection of its green algal host[108]. Experimental characterization provided evidence that this transporter may enhance the uptake of nitrogen by the host cell[108]. Other studies revealed the presence of fermentation genes in the Tetraselmis virus genome with possible implications for host metabolism in nutrient-limited marine systems[26]. A survey of giant virus isolates and metagenome-assembled genomes revealed the widespread presence of genes for cytochrome P450 monooxygenases, potentially enabling or modulating complex metabolic processes such as synthesis of sterols and other fatty acids[88]. Distant homologs of actins and also myosin have been found in NCDLV genomes[87,109] indicating a viral impact on cell structure, motility and intracellular transport processes, but further functional validation is needed. Furthermore, a giant virus affiliated with Mesomimivirinae that infects heterotrophic Choanoflagellates was found to encode type-1 rhodopsins together with the pathway for synthesis of the required pigment, β-carotene[74]. Experimental characterization of the NCLDV rhodopsin showed that the putative rhodopsin likely functions as a proton pump - generating energy from light[74]. A phylogenetically distinct NCLDV rhodopsin was found in the Organic Lake virus metagenome-assembled genome and experimental characterization of this protein revealed that it may function as a light-gated

pentameric ion channel, potentially impacting ion homeostasis and phototaxis of the host cell[110]. Further, through global metagenomics it was predicted that genes encoding various substrate transport processes, energy generation, i.e. photosynthesis, carbon fixation and also glycolysis are commonly found in giant virus metagenome-assembled genomes affiliated with diverse lineages of the Nucleocytoviricota[71,85] (Figure 5). More detailed phylogenetic analysis revealed that some auxiliary metabolic genes encoding transporters for iron, phosphate, magnesium and ammonium originated in eukaryotic hosts and were likely recently acquired by giant viruses through horizontal gene transfer[71,74,85]. However, other genes encoding several rhodopsins, succinate hydrogenase, aconitase, and glyceraldehyde 3-phosphate dehydrogenase showed a pattern that suggested a viral origin or a common evolutionary origin in one of the ancestral hosts[71,74,85]. Taken together, the widespread presence of metabolic genes in diverse NCLDV lineages implies that augmenting host metabolic capacities is likely a strategy more commonly used by members of the NCLDV than initially assumed. However, the current lack of experimental evidence on the functions and activities of most of these genes and pathways, and their effects on the host cell, demand further experimental investigation.

Metabolic reprogramming has direct consequences on host population structure and dynamics. One striking example is the cosmopolitan marine coccolithophore *Emiliana huxleyi* which forms massive blooms that play important roles in global carbon and sulfur cycles[111]. *E. huxleyi* populations are subject to persistent but ultimately lytic infections by the *E. huxleyi* virus[23](Figure 4B). Once lysis is induced this leads to the termination of the algal bloom coupled to the deposition of massive amounts of calcite and nutrients into the ocean which increases the marine pool of dissolved organic matter[112–114]. Importantly, viral infections do not only lead to host lysis but also enhance viral replication by rewiring host physiology, in particular the turnover of sugars and synthesis of fatty acids and lipids[115–117]. Comparably little is known about the impact of other large and giant viruses on host populations but it is conceivable that similar principles are omnipresent and important in shaping our planet's biomes and biogeochemical cycles.

Towards independence from the host

Among the most intriguing features found in giant virus genomes are signatures of cellular life, such as genes involved in protein biosynthesis. This phenomenon was first described upon sequencing the Mimivirus genome[8]. Subsequent analyses revealed the phylogenetic placement of virus-encoded cellular genes between bacteria and eukaryotes suggesting an ancient origin[10]. Other cellular hallmark genes with similarly deep branching patterns were found in other giant virus genomes and led to the hypotheses that giant viruses may either represent a fourth domain of life[12] or are remnants of a highly degraded eukaryotic cell derived by reductive evolution[11]. However, the use of more complex phylogenetic models revealed that many of these genes had most likely been acquired from different eukaryotic hosts[118,119]. Some of these genes might represent ancient transfers from undiscovered eukaryotic hosts. This finding provided evidence for the hypothesis that giant viruses may have evolved from smaller viruses[119]. Yet, other studies have reported alternative topologies for some housekeeping and other metabolic genes of cellular organisms, including rhodopsins[71,74,85] and cytochrome P450[88]. It cannot be ruled out that these genes were transferred from ancestral giant viruses to past

eukaryotic hosts, or even a proto-eukaryote, highlighting a potentially integral role of giant viruses in the evolution of the eukaryotic cell[120,121]. Further, it is possible that some genes that may function as part of the eukaryotic core metabolism were introduced upon integration of giant virus genetic material into the genome of an ancient eukaryotic cell - further shaping eukaryotic evolution[120,122].The presence of genes for aminoacyl tRNA synthetases (aaRS) and eukaryotic translation factors has been recorded multiple times in newly recovered giant virus genomes. Indeed, a nearly complete set of 20 aaRS has been reported in Klosneuvirus from metagenomic data[81]. Shortly after two Tupanviruses were isolated with genomes that contained a full set of aaRS[7] and subsequently the first isolates affiliated with the Klosneuvirinae were described, which also contained a complete set of aaRS[123]. Especially in Klosneuviruses, the presence of aaRS with lineage-specific evolutionary histories provided additional support that these genes derived from different eukaryotic hosts[81]. The presence of genes for extended sets of aaRS is currently limited to members of the Mimiviridae and information on the role of giant virus aaRS in host interaction is limited, however, some have been experimentally studied and were indeed functional[124]. Nevertheless, there is some experimental evidence for potential roles of these genes in making giant viruses less dependent on host machinery, e.g., during shutdown of host translation in response to viral infection or other adverse conditions[125]. On the other hand, a suspected role in enhancing viral translation by providing additional copies of aaRS to support host translation has not yet been confirmed. Another path towards host independence is implied by encoding translation system components, in addition to recently discovered giant virus genes predicted to be involved in energy generation[26,85]. A recent study reported on an active membrane potential in *Pandoravirus massiliensis* virions together with the expression of several remote homologs of TCA genes[126]. Despite encoding expanded genetic complements for functions that were recently thought to be exclusively present in cellular organisms, there is currently no evidence that giant viruses may perform protein translation without host-derived ribosomes or host-independent energy generation.

Conclusions

Nearly 20 years of giant virus isolation has yielded viral isolates representing highly diverse lineages. Complementary detailed research on the biology of these viruses has revealed many important details of virion structures and infection strategies. It has become obvious that there are stark differences in virion size and structure and, while there are some similarities in how these viruses enter and exit the host cell, most giant viruses employ contrasting strategies for replicating within and exploiting their host cells. Sequencing of viral isolates has led to the largest and smallest known genomes of viruses of the Nucleocytoviricota.

Cultivation-independent approaches have accelerated the discovery of genome sequences of new giant viruses and other large viruses in the NCLDV, providing novel insights into their phylogenetic diversity and functional potential. These approaches also revealed that these viruses can be found anywhere on Earth, are affiliated with diverse eukaryotes and are likely manipulating host physiology through metabolic reprogramming - ultimately altering structure and function of host communities in the environment. At the same time, estimates based on NCLDV hallmark genes in metagenomic datasets indicated that only a small fraction of giant virus genomes have been discovered so far[71], and that the diversity of giant viruses may be far

greater than that of bacteria, at least in the oceans[70]. A controlled metagenomic binning experiment where giant viruses were spiked into an environmental sample showed that genome fragments of many giant viruses that are present in a given sample likely remain below the detection limit, highlighting the need for ultra-deep metagenome sequencing[127] or targeted isolation efforts[128]. Furthermore, there is a strong bias towards detecting giant viruses that are similar to those already known, as tools used to identify viruses from metagenomes rely heavily on features observed in sequenced NCLDV genomes, such as large sets of conserved genes[71,82,85,129,130]. However, giant virus genomes exhibit extensive plasticity, such that viruses within the same clade quickly diverge and share very few genes[28]. A recent stunning example of NCLDV diversity is Yaravirus, which was isolated with its native amoeba host[131], yet no close relative of Yaravirus was detectable in public metagenomic datasets. Its placement within NCLDV was challenging due to greater than 90 percent of its genes lacking similarity to those in public databases and the paucity of most viral hallmark genes[131]. Such excessive gene novelty of viruses in the NCLDV, observed through both cultivation and cultivation-independent methods, further underlines that many giant viruses are likely to be hiding in plain sight, awaiting discovery.

Author contributions

FS analyzed and visualized the data. Figures were prepared by FS and CA (Figure 2). FS, CA and TW wrote the manuscript.

Acknowledgements

This work was conducted by the US Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, under Contract No. DE-AC02–05CH11231 and made use of resources of the National Energy Research Scientific Computing Center, also supported by the DOE Office of Science under Contract No. DE-AC02–05CH11231. We thank Dr. Xiao (River) Chuan from Department of Chemistry and Biochemistry, University of Texas El Paso for providing cryo electron micrographs for AaV, Mimivirus and CroV. Further we acknowledge R. Watanabe and K. Murata, ExCELLS, NINS, Japan who provided cryo electron micrographs for Medusavirus, and Dr. R.N. Burton-Smith and Prof. K. Murata, ExCELLS, NINS, Japan for cryo electron micrographs of Melbournevirus.

Competing interests

The authors declare no competing interests.

References

1. Fischer, M. G. Giant viruses come of age. *Curr. Opin. Microbiol.* **31**, 50–57 (2016).

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

3. Koonin, E. V. *et al.* Global Organization and Proposed Megataxonomy of the Virus World.

Microbiol. Mol. Biol. Rev. **84**, (2020).

4. Sun, T.-W. *et al.* Host Range and Coding Potential of Eukaryotic Giant Viruses. *Viruses* **12**, (2020).
5. Abergel, C., Legendre, M. & Claverie, J.-M. The rapidly expanding universe of giant viruses: Mimivirus, Pandoravirus, Pithovirus and Mollivirus. *FEMS Microbiol. Rev.* **39**, 779–796 (2015).
6. Iyer, L. M., Aravind, L. & Koonin, E. V. Common origin of four diverse families of large eukaryotic DNA viruses. *J. Virol.* **75**, 11720–11734 (2001).
7. Abrahão, J. *et al.* Tailed giant Tupanvirus possesses the most complete translational apparatus of the known virosphere. *Nat. Commun.* **9**, 749 (2018).
8. Colson, P., La Scola, B., Levasseur, A., Caetano-Anollés, G. & Raoult, D. Mimivirus: leading the way in the discovery of giant viruses of amoebae. *Nat. Rev. Microbiol.* **15**, 243–254 (2017).
9. La Scola, B. *et al.* A giant virus in amoebae. *Science* **299**, 2033 (2003).
10. Raoult, D. *et al.* The 1.2-megabase genome sequence of Mimivirus. *Science* **306**, 1344–1350 (2004).
11. Legendre, M., Arslan, D., Abergel, C. & Claverie, J.-M. Genomics of Megavirus and the elusive fourth domain of Life. *Commun. Integr. Biol.* **5**, 102–106 (2012).
12. Colson, P., de Lamballerie, X., Fournous, G. & Raoult, D. Reclassification of giant viruses composing a fourth domain of life in the new order Megavirales. *Intervirology* **55**, 321–332 (2012).
13. La Scola, B. *et al.* The virophage as a unique parasite of the giant mimivirus. *Nature* **455**, 100–104 (2008).
14. Jeudy, S. *et al.* Exploration of the propagation of transpovirons within Mimiviridae reveals a unique example of commensalism in the viral world. *ISME J.* **14**, 727–739 (2020).
15. Desnues, C. *et al.* Provirophages and transpovirons as the diverse mobilome of giant

viruses. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 18078–18083 (2012).

16. Fenner, F. Adventures with poxviruses of vertebrates. *FEMS Microbiol. Rev.* **24**, 123–133 (2000).

17. Goebel, S. J. *et al.* The complete DNA sequence of vaccinia virus. *Virology* **179**, 247–66, 517–63 (1990).

18. Oliveira, G. P., Rodrigues, R. A. L., Lima, M. T., Drumond, B. P. & Abrahão, J. S. Poxvirus Host Range Genes and Virus–Host Spectrum: A Critical Review. *Viruses* **9**, 331 (2017).

19. İnce, İ. A., Özcan, O., Ilter-Akulke, A. Z., Scully, E. D. & Özgen, A. Invertebrate Iridoviruses: A Glance over the Last Decade. *Viruses* **10**, (2018).

20. Piégu, B., Asgari, S., Bideshi, D., Federici, B. A. & Bigot, Y. Evolutionary relationships of iridoviruses and divergence of ascoviruses from invertebrate iridoviruses in the superfamily Megavirales. *Mol. Phylogenet. Evol.* **84**, 44–52 (2015).

21. Dixon, L. K., Chapman, D. A. G., Netherton, C. L. & Upton, C. African swine fever virus replication and genomics. *Virus Res.* **173**, 3–14 (2013).

22. Subramaniam, K. *et al.* A New Family of DNA Viruses Causing Disease in Crustaceans from Diverse Aquatic Biomes. *MBio* **11**, (2020).

23. Wilson, W. H., Van Etten, J. L. & Allen, M. J. The Phycodnaviridae: the story of how tiny giants rule the world. *Curr. Top. Microbiol. Immunol.* **328**, 1–42 (2009).

24. Gallot-Lavallée, L., Blanc, G. & Claverie, J.-M. Comparative Genomics of Chrysochromulina Ericina Virus and Other Microalga-Infecting Large DNA Viruses Highlights Their Intricate Evolutionary Relationship with the Established Mimiviridae Family. *J. Virol.* **91**, (2017).

25. Claverie, J.-M. & Abergel, C. Mimiviridae: An Expanding Family of Highly Diverse Large dsDNA Viruses Infecting a Wide Phylogenetic Range of Aquatic Eukaryotes. *Viruses* **10**, (2018).

26. Schvarcz, C. R. & Steward, G. F. A giant virus infecting green algae encodes key

fermentation genes. *Virology* **518**, 423–433 (2018).

27. Blanc-Mathieu, R. *et al.* A persistent giant algal virus, with a unique morphology, encodes an unprecedented number of genes involved in energy metabolism. *J. Virol.* (2021) doi:10.1128/JVI.02446-20.

28. Koonin, E. V. & Yutin, N. Evolution of the Large Nucleocytoplasmic DNA Viruses of Eukaryotes and Convergent Origins of Viral Gigantism. *Adv. Virus Res.* **103**, 167–202 (2019).

29. Pagnier, I. *et al.* A decade of improvements in Mimiviridae and Marseilleviridae isolation from amoeba. *Intervirology* **56**, 354–363 (2013).

30. Takahashi, H., Fukaya, S., Song, C., Murata, K. & Takemura, M. Morphological and Taxonomic Properties of the Newly Isolated Cotonvirus japonicus , a New Lineage of the Subfamily Megavirinae. *Journal of Virology* vol. 95 (2021).

31. Yoshikawa, G. *et al.* Medusavirus, a Novel Large DNA Virus Discovered from Hot Spring Water. *J. Virol.* **93**, (2019).

32. Reteno, D. G. *et al.* Faustovirus, an asfarvirus-related new lineage of giant viruses infecting amoebae. *J. Virol.* **89**, 6585–6594 (2015).

33. Andreani, J. *et al.* Orpheovirus IHUMI-LCC2: A New Virus among the Giant Viruses. *Front. Microbiol.* **8**, 2643 (2017).

34. Andreani, J. *et al.* Pacmanvirus, a New Giant Icosahedral Virus at the Crossroads between Asfarviridae and Faustoviruses. *J. Virol.* **91**, (2017).

35. Bajrai, L. H. *et al.* Kaumoebavirus, a new virus that clusters with faustoviruses and Asfarviridae. Viruses 8: 278. (2016).

36. Francis, R., Ominami, Y., Bou Khalil, J. Y. & La Scola, B. High-throughput isolation of giant viruses using high-content screening. *Commun Biol* **2**, 216 (2019).

37. Dornas, F. P. *et al.* Isolation of new Brazilian giant viruses from environmental samples using a panel of protozoa. *Front. Microbiol.* **6**, 1086 (2015).

38. Burki, F., Roger, A. J., Brown, M. W. & Simpson, A. G. B. The New Tree of Eukaryotes. *Trends Ecol. Evol.* **35**, 43–55 (2020).
39. Renesto, P. *et al.* Mimivirus giant particles incorporate a large fraction of anonymous and unique gene products. *J. Virol.* **80**, 11678–11685 (2006).
40. Fischer, M. G., Allen, M. J., Wilson, W. H. & Suttle, C. A. Giant virus with a remarkable complement of genes infects marine zooplankton. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 19508–19513 (2010).
41. Deeg, C. M., Chow, C.-E. T. & Suttle, C. A. The kinetoplastid-infecting Bodo saltans virus (BsV), a window into the most abundant giant viruses in the sea. *Elife* **7**, (2018).
42. Notaro, A. *et al.* Expanding the Occurrence of Polysaccharides to the Viral World: The Case of Mimivirus. *Angew. Chem. Int. Ed Engl.* **60**, 19897–19904 (2021).
43. Klose, T. *et al.* The three-dimensional structure of Mimivirus. *Intervirology* **53**, 268–273 (2010).
44. Kuznetsov, Y. G. *et al.* Atomic force microscopy investigation of the giant mimivirus. *Virology* **404**, 127–137 (2010).
45. Zauberman, N. *et al.* Distinct DNA exit and packaging portals in the virus Acanthamoeba polyphaga mimivirus. *PLoS Biol.* **6**, e114 (2008).
46. Fischer, M. G., Kelly, I., Foster, L. J. & Suttle, C. A. The virion of Cafeteria roenbergensis virus (CroV) contains a complex suite of proteins for transcription and DNA repair. *Virology* **466-467**, 82–94 (2014).
47. Arslan, D., Legendre, M., Seltzer, V., Abergel, C. & Claverie, J.-M. Distant Mimivirus relative with a larger genome highlights the fundamental features of Megaviridae. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 17486–17491 (2011).
48. Mutsafi, Y., Zauberman, N., Sabanay, I. & Minsky, A. Vaccinia-like cytoplasmic replication of the giant Mimivirus. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 5978–5982 (2010).
49. Kuznetsov, Y. G., Klose, T., Rossmann, M. & McPherson, A. Morphogenesis of Mimivirus

and Its Viral Factories: an Atomic Force Microscopy Study of Infected Cells. *Journal of Virology* vol. 88 3055–3055 (2014).

50. Fischer, M. G. & Hackl, T. Host genome integration and giant virus-induced reactivation of the virophage mavirus. *Nature* **540**, 288–291 (2016).
51. Bellec, L., Grimsley, N., Moreau, H. & Desdevises, Y. Phylogenetic analysis of new Prasinoviruses (Phycodnaviridae) that infect the green unicellular algae Ostreococcus, Bathycoccus and Micromonas. *Environ. Microbiol. Rep.* **1**, 114–123 (2009).
52. Van Etten, J. L., Agarkova, I. V. & Dunigan, D. D. Chloroviruses. *Viruses* **12**, (2019).
53. Cherrier, M. V. *et al.* An icosahedral algal virus has a complex unique vertex decorated by a spike. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 11085–11089 (2009).
54. De Castro, C. *et al.* Structure of the chlorovirus PBCV-1 major capsid glycoprotein determined by combining crystallographic and carbohydrate molecular modeling approaches. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E44–E52 (2018).
55. Burton-Smith, R. N. *et al.* The 4.4 Å structure of the giant Melbournevirus virion belonging to the Marseilleviridae family. *bioRxiv* (2021) doi:10.1101/2021.07.14.452405.
56. Chihara, A. *et al.* A novel capsid protein network allows the characteristic inner membrane structure of Marseilleviridae giant viruses. *bioRxiv* 2021.02.03.428533 (2021) doi:10.1101/2021.02.03.428533.
57. Fang, Q. *et al.* Near-atomic structure of a giant virus. *Nat. Commun.* **10**, 388 (2019).
58. Rodrigues, R. A. L. *et al.* Analysis of a Marseillevirus Transcriptome Reveals Temporal Gene Expression Profile and Host Transcriptional Shift. *Frontiers in Microbiology* vol. 11 (2020).
59. Oliveira, G. P. *et al.* The investigation of promoter sequences of marseilleviruses highlights a remarkable abundance of the AAATATTT motif in intergenic regions. *J. Virol.* **91**, (2017).
60. Okamoto, K. *et al.* Cryo-EM structure of a Marseilleviridae virus particle reveals a large internal microassembly. *Virology* **516**, 239–245 (2018).

61. Legendre, M. *et al.* In-depth study of Mollivirus sibericum, a new 30,000-y-old giant virus infecting Acanthamoeba. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E5327–35 (2015).
62. Scheid, P., Balczun, C. & Schaub, G. A. Some secrets are revealed: parasitic keratitis amoebae as vectors of the scarcely described pandoraviruses to humans. *Parasitol. Res.* **113**, 3759–3764 (2014).
63. Akashi, M. & Takemura, M. Co-isolation and characterization of two pandoraviruses and a Mimivirus from a riverbank in Japan. *Viruses* **11**, 1123 (2019).
64. Levasseur, A. *et al.* Comparison of a Modern and Fossil Pithovirus Reveals Its Genetic Conservation and Evolution. *Genome Biol. Evol.* **8**, 2333–2339 (2016).
65. Andreani, J. *et al.* Cedratvirus, a Double-Cork Structured Giant Virus, is a Distant Relative of Pithoviruses. *Viruses* **8**, (2016).
66. Bertelli, C. *et al.* Cedratvirus lausannensis - digging into Pithoviridae diversity. *Environ. Microbiol.* **19**, 4022–4034 (2017).
67. Yoshida, K. *et al.* Draft Genome Sequence of Medusavirus Stheno, Isolated from the Tatakai River of Uji, Japan. *Microbiol Resour Announc* **10**, (2021).
68. Ghedin, E. & Claverie, J.-M. Mimivirus relatives in the Sargasso sea. *Virol. J.* **2**, 62 (2005).
69. Hingamp, P. *et al.* Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans microbial metagenomes. *ISME J.* **7**, 1678–1695 (2013).
70. Mihara, T. *et al.* Taxon Richness of 'Megaviridae' Exceeds those of Bacteria and Archaea in the Ocean. *Microbes Environ.* **33**, 162–171 (2018).
71. Schulz, F. *et al.* Giant virus diversity and host interactions through global metagenomics. *Nature* **578**, 432–436 (2020).
72. Martínez Martínez, J., Swan, B. K. & Wilson, W. H. Marine viruses, a genetic reservoir revealed by targeted viromics. *ISME J.* **8**, 1079–1088 (2014).
73. Wilson, W. H. *et al.* Genomic exploration of individual giant ocean viruses. *ISME J.* **11**, 1736–1745 (2017).

74. Needham, D. M. *et al.* A distinct lineage of giant viruses brings a rhodopsin photosystem to unicellular marine predators. *Proc. Natl. Acad. Sci. U. S. A.* 201907517 (2019).
75. Needham, D. M. *et al.* Targeted metagenomic recovery of four divergent viruses reveals shared and distinctive characteristics of giant viruses of marine eukaryotes. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **374**, 20190086 (2019).
76. Sun, T.-W. & Ku, C. Unraveling gene content variation across eukaryotic giant viruses based on network analyses and host associations. *Virus Evol* **7**, veab081 (2021).
77. Schulz, F. *et al.* Hidden diversity of soil giant viruses. *Nat. Commun.* **9**, 4881 (2018).
78. Allen, E. E. & Banfield, J. F. Community genomics in microbial ecology and evolution. *Nat. Rev. Microbiol.* **3**, 489–498 (2005).
79. Yau, S. *et al.* Virophage control of antarctic algal host–virus dynamics. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 6163–6168 (2011).
80. Zhang, W. *et al.* Four novel algal virus genomes discovered from Yellowstone Lake metagenomes. *Sci. Rep.* **5**, 15131 (2015).
81. Schulz, F. *et al.* Giant viruses with an expanded complement of translation system components. *Science* **356**, 82–85 (2017).
82. Bäckström, D. *et al.* Virus Genomes from Deep Sea Sediments Expand the Ocean Megavirome and Support Independent Origins of Viral Gigantism. *MBio* **10**, (2019).
83. Chen, H. *et al.* The genome of a prasinoviruses-related freshwater virus reveals unusual diversity of phycodnaviruses. *BMC Genomics* **19**, 49 (2018).
84. Xu, S. *et al.* Novel Cell-Virus-Virophage Tripartite Infection Systems Discovered in the Freshwater Lake Dishui Lake in Shanghai, China. *J. Virol.* **94**, (2020).
85. Moniruzzaman, M., Martinez-Gutierrez, C. A., Weinheimer, A. R. & Aylward, F. O. Dynamic genome evolution and complex virocell metabolism of globally-distributed giant viruses. *Nat. Commun.* **11**, 1710 (2020).
86. Aylward, F. O., Moniruzzaman, M., Ha, A. D. & Koonin, E. V. A phylogenomic framework

for charting the diversity and evolution of giant viruses. *PLoS Biol.* **19**, e3001430 (2021).

87. Da Cunha, V., Gaia, M., Ogata, H., Jaillon, O. & Delmont, T. O. Giant viruses encode novel types of actins possibly related to the origin of eukaryotic actin: the viractins. *bioRxiv* (2020).

88. Lamb, D. C. *et al.* On the occurrence of cytochrome P450 in viruses. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 12343–12352 (2019).

89. Mihara, T. *et al.* Linking Virus Genomes with Host Taxonomy. *Viruses* **8**, 66 (2016).

90. Endo, H. *et al.* Biogeography of marine giant viruses reveals their interplay with eukaryotes and ecological functions. *Nat Ecol Evol* **4**, 1639–1649 (2020).

91. Andreani, J., Verneau, J., Raoult, D., Levasseur, A. & La Scola, B. Deciphering viral presences: two novel partial giant viruses detected in marine metagenome and in a mine drainage metagenome. *Virol. J.* **15**, 66 (2018).

92. Feschotte, C. & Gilbert, C. Endogenous viruses: insights into viral evolution and impact on host biology. *Nat. Rev. Genet.* **13**, 283–296 (2012).

93. Chiba, S. *et al.* Widespread endogenization of genome sequences of non-retroviral RNA viruses into plant genomes. *PLoS Pathog.* **7**, e1002146 (2011).

94. Maumus, F., Epert, A., Nogué, F. & Blanc, G. Plant genomes enclose footprints of past infections by giant virus relatives. *Nature Communications* vol. 5 (2014).

95. Wang, L. *et al.* Endogenous viral elements in algal genomes. *Acta Oceanologica Sinica* vol. 33 102–107 (2014).

96. Delaroque, N., Maier, I., Knippers, R. & M ºller, D. G. Persistent virus integration into the genome of its algal host, Ectocarpus siliculosus (Phaeophyceae). *J. Gen. Virol.* **80 (Pt 6)**, 1367–1370 (1999).

97. Maumus, F. & Blanc, G. Study of Gene Trafficking between Acanthamoeba and Giant Viruses Suggests an Undiscovered Family of Amoeba-Infecting Viruses. *Genome Biol. Evol.* **8**, 3351–3363 (2016).

98. Clarke, M. *et al.* Genome of Acanthamoeba castellanii highlights extensive lateral gene transfer and early evolution of tyrosine kinase signaling. *Genome Biol.* **14**, R11 (2013).
99. Chelkha, N. *et al.* Vermamoeba vermiformis CDC-19 draft genome sequence reveals considerable gene trafficking including with candidate phyla radiation and giant viruses. *Scientific Reports* vol. 10 (2020).
100. Gallot-Lavallée, L. & Blanc, G. A Glimpse of Nucleo-Cytoplasmic Large DNA Virus Biodiversity through the Eukaryotic Genomics Window. *Viruses* **9**, (2017).
101. Delaroque, N. & Boland, W. The genome of the brown alga Ectocarpus siliculosus contains a series of viral DNA pieces, suggesting an ancient association with large dsDNA viruses. *BMC Evol. Biol.* **8**, 110 (2008).
102. Stevens, K. *et al.* A novel evolutionary strategy revealed in the phaeoviruses. *PLoS One* **9**, e86040 (2014).
103. Cock, J. M. *et al.* The Ectocarpus genome and the independent evolution of multicellularity in brown algae. *Nature* **465**, 617–621 (2010).
104. Moniruzzaman, M., Weinheimer, A. R., Martinez-Gutierrez, C. A. & Aylward, F. O. Widespread endogenization of giant viruses shapes genomes of green algae. *Nature* **588**, 141–145 (2020).
105. Forterre, P. The virocell concept and environmental microbiology. *ISME J.* **7**, 233–236 (2013).
106. Howard-Varona, C. *et al.* Phage-specific metabolic reprogramming of virocells. *The ISME Journal* vol. 14 881–895 (2020).
107. Hurwitz, B. L., Hallam, S. J. & Sullivan, M. B. Metabolic reprogramming by viruses in the sunlit and dark ocean. *Genome Biol.* **14**, R123 (2013).
108. Monier, A. *et al.* Host-derived viral transporter protein for nitrogen uptake in infected marine phytoplankton. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E7489–E7498 (2017).
109. Kijima, S. *et al.* Discovery of Viral Myosin Genes With Complex Evolutionary History Within

Plankton. *Front. Microbiol.* **12**, 683294 (2021).

110. Bratanov, D. *et al.* Unique structure and function of viral rhodopsins. *Nat. Commun.* **10**, 4939 (2019).

111. Paasche, E. A review of the coccolithophorid Emiliania huxleyi (Prymnesiophyceae), with particular reference to growth, coccolith formation, and calcification-photosynthesis interactions. *Phycologia* **40**, 503–529 (2001).

112. Kuhlisch, C. *et al.* Viral infection of algal blooms leaves a unique metabolic footprint on the dissolved organic matter in the ocean. *Sci Adv* **7**, (2021).

113. Breitbart, M. Marine viruses: truth or dare. *Ann. Rev. Mar. Sci.* **4**, 425–448 (2012).

114. Wilhelm, S. W. & Suttle, C. A. Viruses and Nutrient Cycles in the Sea: Viruses play critical roles in the structure and function of aquatic food webs. *Bioscience* **49**, 781–788 (1999).

115. Malitsky, S. *et al.* Viral infection of the marine alga Emiliania huxleyi triggers lipidome remodeling and induces the production of highly saturated triacylglycerol. *New Phytol.* **210**, 88–96 (2016).

116. Schleyer, G. *et al.* In plaque-mass spectrometry imaging of a bloom-forming alga during viral infection reveals a metabolic shift towards odd-chain fatty acid lipids. *Nature Microbiology* **4**, 527–538 (2019).

117. Rosenwasser, S. *et al.* Rewiring Host Lipid Metabolism by Large Viruses Determines the Fate of Emiliania huxleyi, a Bloom-Forming Alga in the Ocean. *Plant Cell* **26**, 2689–2707 (2014).

118. Williams, T. A., Embley, T. M. & Heinz, E. Informational gene phylogenies do not support a fourth domain of life for nucleocytoplasmic large DNA viruses. *PLoS One* **6**, e21080 (2011).

119. Yutin, N., Wolf, Y. I. & Koonin, E. V. Origin of giant viruses from smaller DNA viruses not from a fourth domain of cellular life. *Virology* **466-467**, 38–52 (2014).

120. Guglielmini, J., Woo, A. C., Krupovic, M., Forterre, P. & Gaia, M. Diversification of giant and large eukaryotic dsDNA viruses predated the origin of modern eukaryotes. *Proc. Natl.*

Acad. Sci. U. S. A. (2019) doi:10.1073/pnas.1912006116.

121. Bell, P. J. Viral eukaryogenesis: was the ancestor of the nucleus a complex DNA virus? *Journal of molecular evolution* vol. 53 251–256 (2001).

122. Cheng, S., Wong, G. K.-S. & Melkonian, M. Giant DNA viruses make big strides in eukaryote evolution. *Cell host & microbe* vol. 29 152–154 (2021).

123. Hussein Bajrai, L. *et al.* Isolation of Yasminevirus, the first member of Klosneuvirinae isolated in coculture with Vermamoeba vermiformis, demonstrates an extended arsenal of translational apparatus components. *J. Virol.* (2019) doi:10.1128/JVI.01534-19.

124. Abergel, C., Rudinger-Thirion, J., Giegé, R. & Claverie, J.-M. Virus-encoded aminoacyl-tRNA synthetases: structural and functional characterization of mimivirus TyrRS and MetRS. *J. Virol.* **81**, 12406–12417 (2007).

125. Silva, L. C. F. *et al.* Modulation of the expression of mimivirus-encoded translation-related genes in response to nutrient availability during Acanthamoeba castellanii infection. *Front. Microbiol.* **6**, 539 (2015).

126. Aherfi, S. *et al.* Incomplete tricarboxylic acid cycle and proton gradient in Pandoravirus massiliensis: is it still a virus? *The ISME Journal* (2021) doi:10.1038/s41396-021-01117-3.

127. Schulz, F. *et al.* Advantages and Limits of Metagenomic Assembly and Binning of a Giant Virus. *mSystems* **5**, (2020).

128. Christo-Foroux, E. *et al.* Characterization of Mollivirus kamchatka, the First Modern Representative of the Proposed Molliviridae Family of Giant Viruses. *J. Virol.* **94**, (2020).

129. Aylward, F. O. & Moniruzzaman, M. ViralRecall: A Flexible Command-Line Tool for the Detection of Giant Virus Signatures in Omic Data. doi:10.1101/2020.12.15.422924.

130. Nayfach, S. *et al.* CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* (2020) doi:10.1038/s41587-020-00774-7.

131. Boratto, P. V. M. *et al.* Yaravirus: A novel 80-nm virus infecting Acanthamoeba castellanii. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 16579–16586 (2020).

132. Philippe, N. *et al.* Pandoraviruses: Amoeba Viruses with Genomes Up to 2.5 Mb Reaching That of Parasitic Eukaryotes. *Science* (2013) doi:10.1126/science.1239181.
133. Legendre, M. *et al.* Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses with a pandoravirus morphology. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 4274–4279 (2014).
134. Gann, E. R. *et al.* Structural and Proteomic Studies of the Aureococcus anophagefferens Virus Demonstrate a Global Distribution of Virus-Encoded Carbohydrate Processing. *Front. Microbiol.* **11**, 2047 (2020).
135. Xiao, C. *et al.* Structural studies of the giant mimivirus. *PLoS Biol.* **7**, e92 (2009).
136. Xiao, C. *et al.* Cryo-EM reconstruction of the Cafeteria roenbergensis virus capsid suggests novel assembly pathway for giant viruses. *Sci. Rep.* **7**, 5484 (2017).
137. Kerepesi, C. & Grolmusz, V. The 'Giant Virus Finder' discovers an abundance of giant viruses in the Antarctic dry valleys. *Arch. Virol.* **162**, 1671–1676 (2017).
138. Chatterjee, A. & Kondabagil, K. Giant viral genomic signatures in the previously reported gut metagenomes of pre-school children in rural India. *Arch. Virol.* **164**, 2819–2822 (2019).
139. Pires de Souza, G. A., Rolland, C., Nafeh, B., La Scola, B. & Colson, P. Giant virus-related sequences in the 5300-year-old Ötzi mummy metagenome. *Virus Genes* **57**, 222–227 (2021).
140. Verneau, J., Levasseur, A., Raoult, D., La Scola, B. & Colson, P. MG-Digger: An Automated Pipeline to Search for Giant Virus-Related Sequences in Metagenomes. *Front. Microbiol.* **7**, 428 (2016).
141. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
142. Bowers, R. M. *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**,

725–731 (2017).

143. Woyke, T., Doud, D. F. R. & Schulz, F. The trajectory of microbial single-cell sequencing. *Nat. Methods* **14**, 1045–1054 (2017).

144. Martínez, J. M., Martinez-Hernandez, F. & Martinez-Garcia, M. Single-virus genomics and beyond. *Nat. Rev. Microbiol.* **18**, 705–716 (2020).

145. Khalil, J. Y. B. *et al.* High-Throughput Isolation of Giant Viruses in Liquid Medium Using Automated Flow Cytometry and Fluorescence Staining. *Front. Microbiol.* **7**, 26 (2016).

146. Yu, F. B. *et al.* Microfluidic-based mini-metagenomics enables discovery of novel microbial lineages from complex environmental samples. *Elife* **6**, (2017).

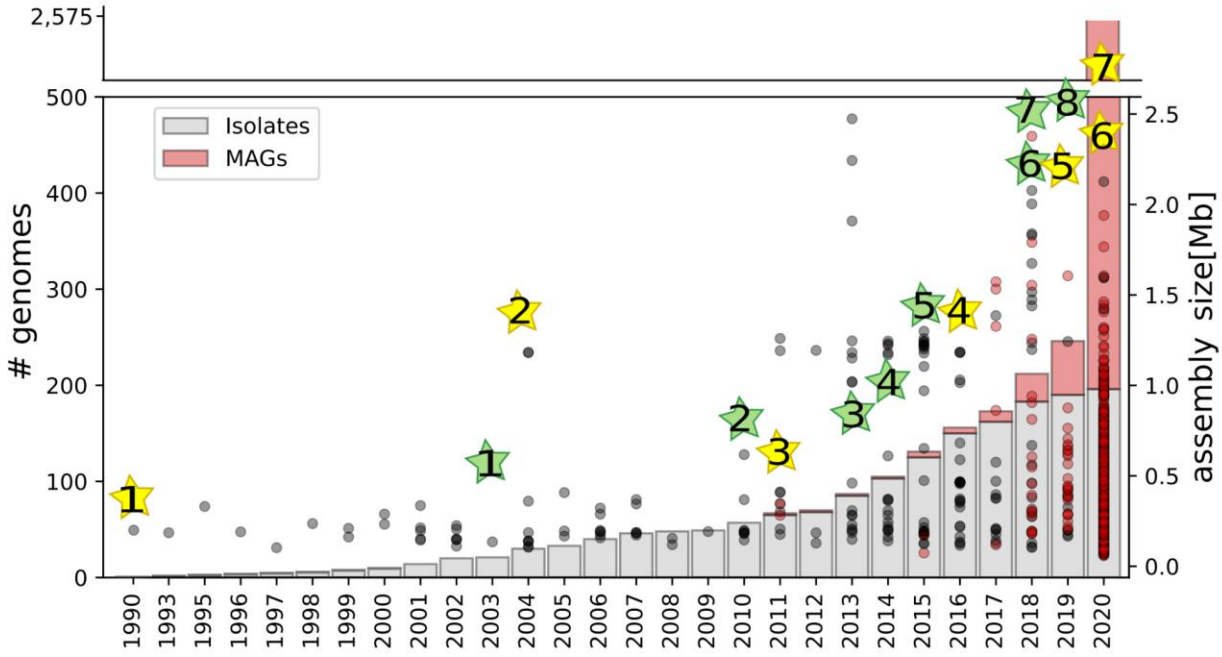

Figure 1. Timeline of genomic and metagenomic discoveries in the Nucleocytoviricota. Stacked bars indicate the number of virus isolate genomes (grey, left y-axis) and giant virus metagenome-assembled genomes (GVMAGs; red, left y-axis) of members of the Nucleocytoviricota that have been published and/or became available at the NCBI Genbank database for each year on the x-axis. Filled circles indicate genome size (right y-axis) of virus isolates (grey) and GVMAGs (red). Important genomic and metagenomic events are highlighted by yellow stars: 1. First genome of a member of the NCLDV became available[17]; 2. Sequencing of the first giant virus with a genome size above 1 Mb, *Acanthamoeba polyphaga* Mimivirus[10]; 3. First-time recovery of GVMAGs (from Organic Lake, Antarctica)[79]; 4. Viral subfamily "Klosneuvirinae" proposed based on GVMAGs recovered from environmental sequence data[81]; 5. Single-cell genomics enabled discovery of Choanovirus from marine choanoflagellates[74]; 6. First large-scale global metagenomics study leading to the recovery of over 2000 GVMAGs facilitating an 11-fold increase in phylogenetic diversity and a 10-fold expansion in functional diversity[71], 7. Detection of whole giant virus genomes integrated in host chromosomes[104]. Important cultivation based events are highlighted by green stars: 1. Isolation of Mimivirus in amoeba co-cultivation [9] 2. Isolation of a giant virus together with its *Cafeteria roenbergensis* host[40] 3. Isolation of amphora-shaped Pandoravirus with 2.5Mb genome[132] 4. Recovery of Pithovirus, from a 30,000 year old ice core, through co-cultivation with an amoeba[133] 5. Isolation of Faustovirus in co-cultivation with *Vermamoeba*[32] 6. Isolation of *Bodo saltans* virus the first isolated member of the Klosneuvirinae with its native Kinetoplastid host [41] 7. Isolation of Tupanvirus in amoeba - currently the largest giant virus based on its capsid diameter and length[7] 8. Isolation of Medusavirus in amoeba represents a divergent new lineage in the Nucleocytoviricota[31].

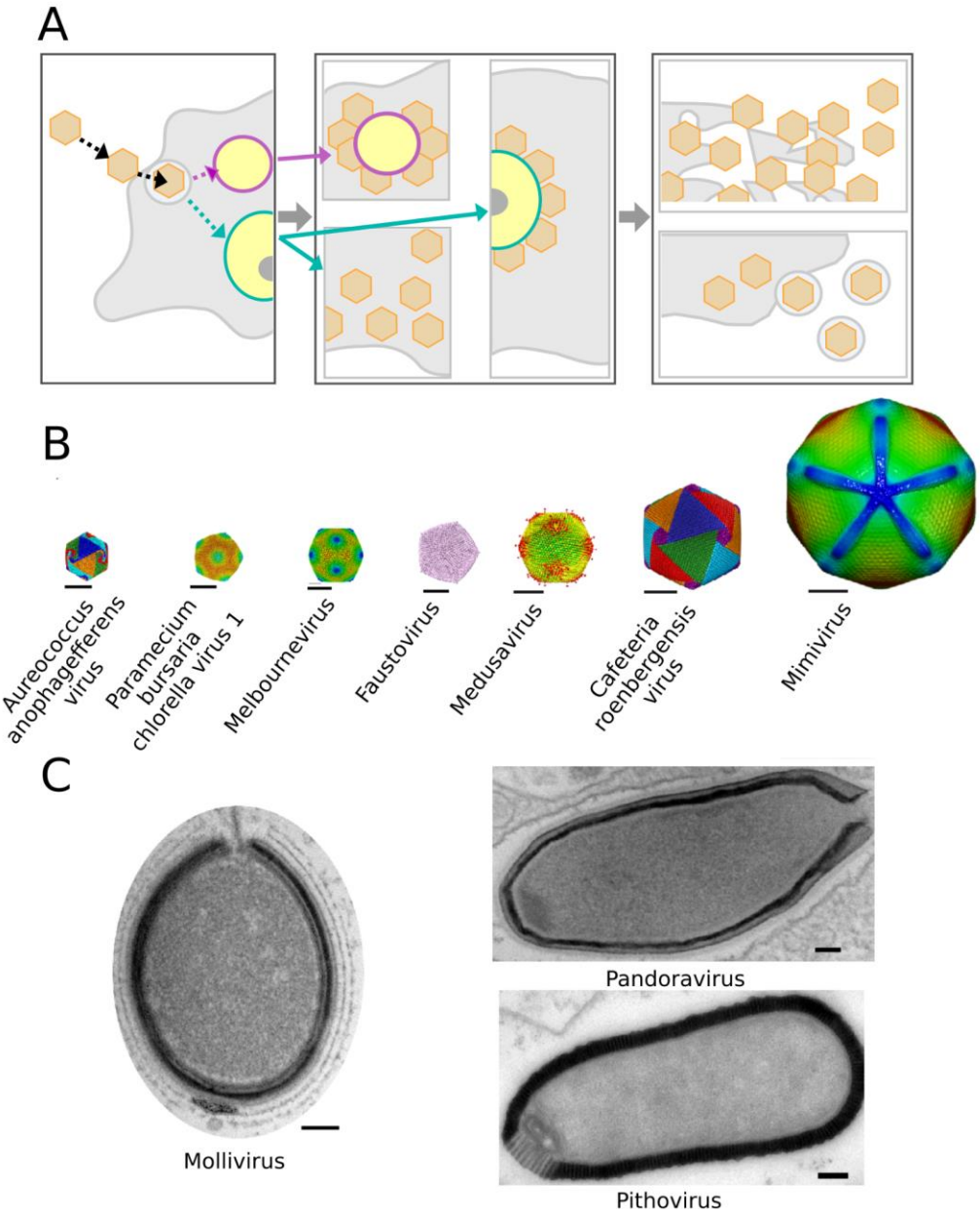

Figure 2. Giant virus infection mechanisms and virion structures. A Attachment to the host cell envelope followed by uptake by endocytosis and activation of giant virus transcription in the nucleus (turquoise arrow) or the cytoplasm / viral factory (purple arrow). This is followed by giant replication and assembly of new virions in the periphery of the cytoplasmic viral factory, in the host nucleus or scattered in the virocell. Virions are released after host cell lysis or exocytosis of membrane-bound virions. **B** Structures of isolated giant viruses resolved by cryo-electron microscopy[55,134–136]. Note the blue-colored stargate structure on Mimivirus. **C** Transmission electron micrographs of ultrathin sections of non icosahedral viruses embedded in resin. The scale bars in B and C are 100 nm.

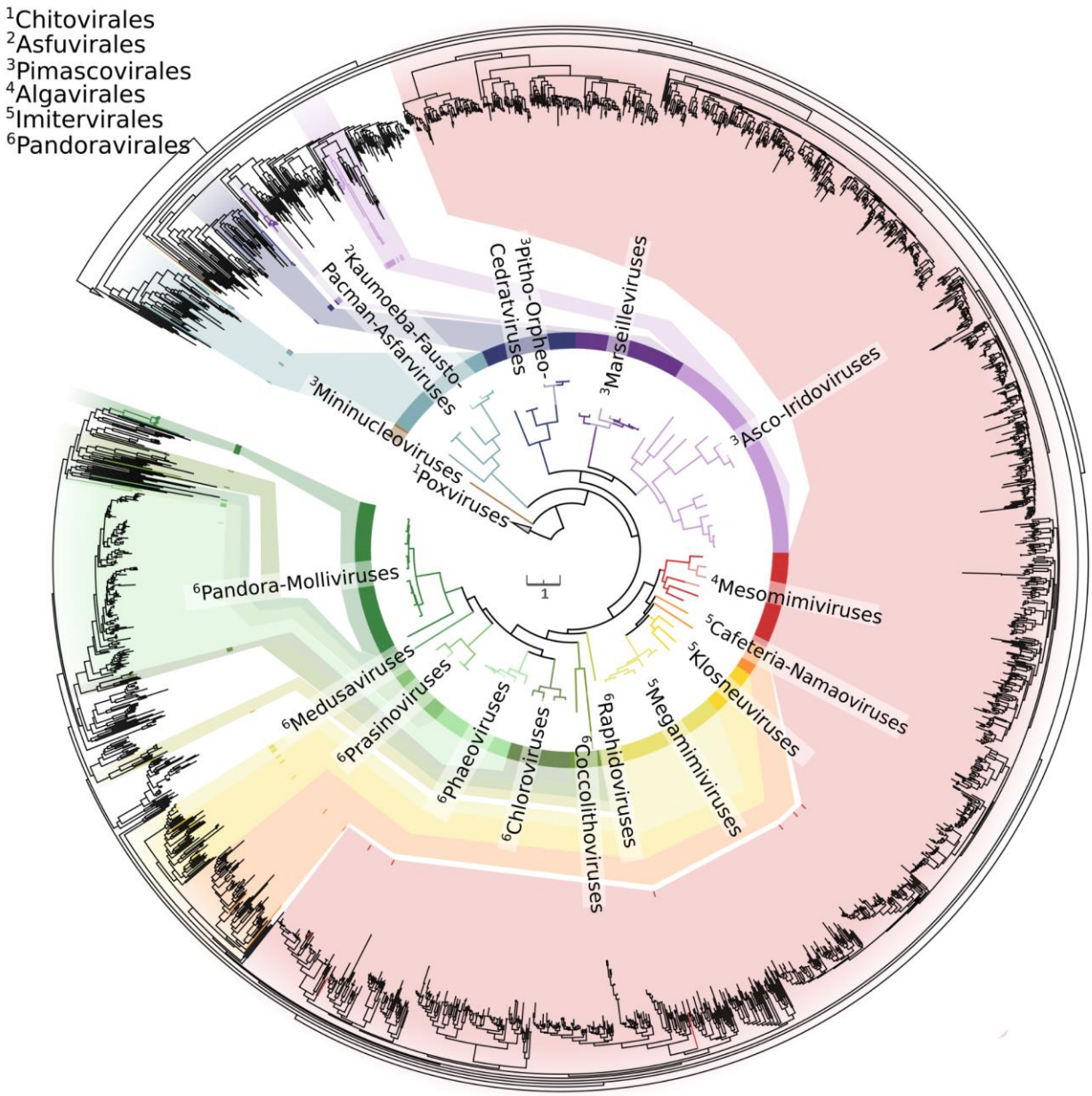

Figure 3. Expansion of the Nucleocytoviricota phylogenetic diversity through metagenomics. Shown are two species trees of the Nucleocytoviricota: the inner tree depicts the diversity of NCLDV based on the genomes of viral isolates and the outer inverted tree highlights the expansion of species diversity through genomes derived from cultivation-independent sequencing approaches. Branches are colored based on affiliation to taxonomic groups consisting of isolates and extrapolated to the expanded diversity of the outer tree. Superscripts map to the proposed updated taxonomic framework for Nucleocytoviricota lineages[86].

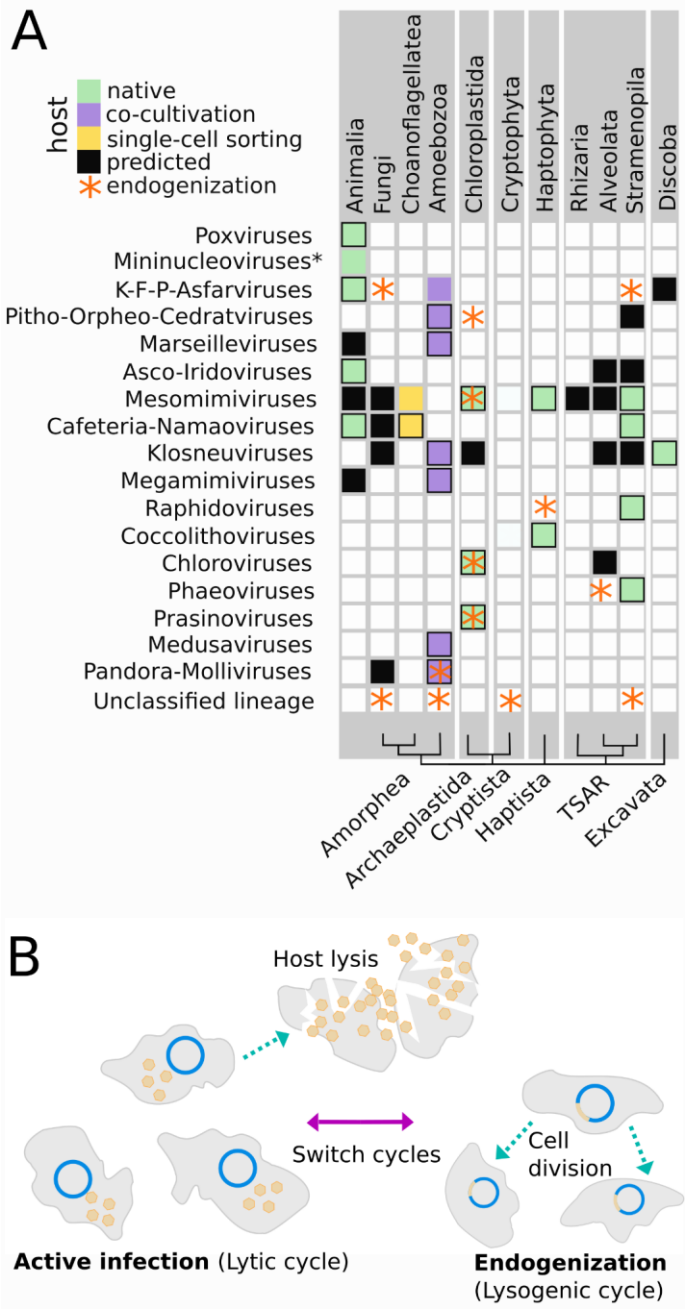

Figure 4. Experimentally verified and computationally predicted host ranges of the Nucleocytoviricota. A Shown are host lineages identified through isolation with the native host, co-cultivation, single-cell sorting, and *in silico* HGT-based predictions. Black outline of colored boxes indicates that experimentally verified interaction has been also predicted computationally. K-F-P-Asfarviruses: Kaumoeba-Fausto-Pacman-Asfarviruses; *genomes of this group were not included in the HGT analysis. Topology of the eukaryotic species tree adapted from[38]**. B** Lysis of host cells and entire host populations are the typical outcomes of a giant virus infection. There is growing evidence that long term persistent infections may be prevalent in algae populations, including the integration of giant virus genomes, so-called endogenization, into the host chromosome.

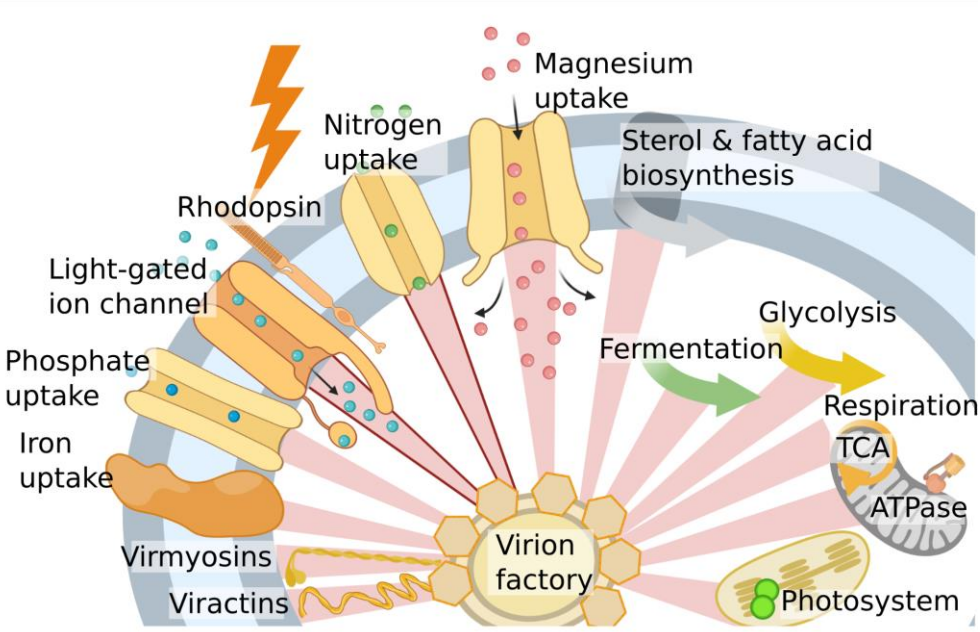

Figure 5: Predicted metabolic reprogramming of a giant virus-derived virocell and consequences of giant virus infection for host populations. Shown is a hypothetical virocell with a combination of metabolic roles that different giant viruses are predicted to have during host infection, based on the presence of auxiliary metabolic genes in giant virus genomes. Dark red lines denote metabolic roles that are supported by some functional data obtained through experiments.

Box 1:

Toolkits for giant virus discovery: cultivation independent genomic approaches

Read mapping-based approaches

Mapping metagenomic reads to giant virus reference genomes has been successfully applied to detect giant viruses and estimate their abundances in the environment[41,69,137–139] and several tools have been published[137,140]. Read mapping-based approaches are advantageous because they are sensitive enough to detect lowly abundant giant viruses[139]; however, they typically do not lead to the recovery of viral genomes and thus cannot provide information on genome features and coding potential. Moreover, mapping approaches are highly dependent on the quality of the reference genome database, and if low mapping stringency is used, false positive hits may occur. Detection of giant viruses may also be hindered if taxonomic classification of metagenome-assembled genomes was performed using automated tools; this has resulted in several NCLDV MAGs that have been misclassified as being of bacterial, archaeal or eukaryotic origin[71,91]. In addition, genes that have recently been integrated into NCLDV genomes after being horizontally acquired from bacteria or eukaryotes may produce sequence reads that map to cellular genomes, resulting in false positive hits.

Marker gene surveys

Detection and phylogenetic analysis of signature genes in complex environmental datasets is a commonly used approach to assess viral diversity in metagenome data. For Nucelocytoviricota genes that encode major capsid protein, DNA Polymerase B, or viral packaging ATPase have been used as marker genes. The approach is less error-prone than read mapping as it can be coupled with phylogenetic analysis to confirm the monophyly of the respective gene homologs found in known NCLDV genomes. This approach has been successfully applied in several studies[69–71,90] and, while being less sensitive than read mapping, it can detect viruses that were not abundant enough in a metagenome to be successfully assembled and binned[127].

Genome-resolved metagenomics

The reconstruction of metagenome-assembled genomes through metagenomic binning is an established approach to recover microbial genomes. Due to their virion sizes giant viruses are often present in environmental samples that have been selectively filtered to target microbes. In contrast to smaller viruses such as most bacteriophages, the large genomes of most members of the Nucleocytoviricota typically require metagenomic binning to increase genome completeness[127]. However, in most microbe-centric metagenome projects giant virus genome bins were frequently ignored, as tools that estimate genome quality[141] predict viral genomes to be of low completeness based on their lack of cellular marker genes[127] which then leads to their exclusion from downstream analyses[142]. Several recent studies employed custom workflows to identify giant virus MAGs and to estimate completeness and contamination by, for example, identifying copy numbers of conserved NCLDV genes[74,75,77,81,85,130] or inferring deviations from lineage specific copy numbers of low-copy orthologs[71]. It is important to note that genome sizes of GVMAGs typically do not represent the size of the complete viral genome.

Single cell genomic approaches

Flow-sorting and sequencing of single viruses can be used to detect viruses in environmental samples[143,144], yet only a few such studies have discovered novel giant viruses[72–75,77]. Due to large virion sizes and a bright signal using DNA stains[73,145] giant viruses are a promising target for sorting. A drawback of this approach is that the subsequent whole genome amplification, if

performed on a single virus, may lead to low genome recovery[143]. An alternative approach to direct sorting of giant viruses from an environmental sample is targeted sorting of host cells[74,75]. Viruses actively replicating inside a host cell can produce hundreds to thousands of virions with clonal copies of viral genomes, which would greatly improve whole genome amplification[143]. Further, if successful, this approach enables identification of the virus and its native host. Similarly, mini metagenomics uses fluorescence activated single cell sorting or microfluidics to collect tens to hundreds of cell-sized particles[77,146]. The presence of many identical viral particles, either through repeated sorting of clonal single viruses, an infected host cell, or the sorting of vacuoles filled with giant viruses, would then increase genome recovery.