Dynamic Genome Evolution and Blueprint of Complex Virocell Metabolism in Globally-Distributed Giant Viruses

Mohammad Moniruzzaman, Carolina A. Martinez-Gutierrez, Alaina R. Weinheimer, Frank O. Aylward*

Department of Biological Sciences, Virginia Tech, Blacksburg VA

*Email for correspondence: faylward@vt.edu

Key words: Giant viruses, NCLDV, Megavirales, virocell, viral diversity, marine viruses

Abstract

The discovery of giant viruses with large genomes has transformed our understanding of the limits of viral complexity in the biosphere, and subsequent research in model virus-host systems has advanced our knowledge of intricate mechanisms used by these viruses to take over host cells during infection. The extent of the metabolic diversity encoded by these viruses in the environment is less well-understood, however, and their potential impact on global biogeochemical cycles remains unclear. To address this, we generated 501 metagenome-assembled genomes (MAGs) of NCLDVs from diverse environments around the globe and analyzed their encoded functional diversity and potential for reprogramming host physiology. We found that 476 (95%) of the MAGs belonged to the *Mimiviridae* and *Phycodnaviridae* families, and of these we recovered 96% from aquatic environments, highlighting the diversity of these viral families in global freshwater and marine systems. MAGs encoded diverse genes predicted to be involved in nutrient uptake and processing, light harvesting, central nitrogen metabolism, and the manipulation of cell death, underscoring the complex interplay between these viruses and their hosts. Surprisingly, numerous genomes encoded genes involved in glycolysis, gluconeogenesis, and the TCA cycle, including one genome with a 70%-complete glycolytic pathway, suggesting that many of these viruses can even reprogram fundamental aspects of their host's central carbon metabolism. Phylogenetic trees of NCLDV metabolic genes together with their cellular homologs revealed distinct clustering of viral sequences into divergent clades, indicating these metabolic genes are virus-specific and were acquired in the distant past. Our findings reveal that diverse NCLDV genomes encode complex, cell-like metabolic capabilities with evolutionary histories that are largely independent of cellular life, strongly implicating them as distinct drivers of biogeochemical cycles in their own right.

Abbreviations: MAGs: Metagenome assembled genomes, NCLDV: Nucleo-cytoplasmic large DNA viruses

Introduction

Nucleocytoplasmic large DNA viruses (NCLDV) are a diverse group of eukaryotic viruses that include several families of "giants" known for both their large virion size, reaching up to 1.5 µm, and genomes reaching ~2.5 million base-pairs in length (1). The discovery of the first giant virus, *Acanthaomeba polyphaga mimivirus*, led to a paradigm shift in the field of virology by showing that, contrary to the traditional view of viruses as "filterable infectious agents", viruses could be larger than even some cellular lineages both in terms of physical size and genomic contents (2, 3). Several subsequent studies have continued to expand our knowledge of NCLDV diversity through the discovery of *Pithoviridae*, *Marseilleviridae*, *Pandoraviruses*, and several other members of the *Mimiviridae* family, all encoding large genomes with diverse genomic repertoires (4), and evolutionary genomic analysis has revealed common ancestry of all these groups together with algal viruses of the family *Phycodnaviridae* and vertebrate viruses of the families *Iridoviridae*, *Poxviridae*, and *Asfarviridae* (5).

Despite the genomic novelty of many NCLDVs, their diversity in the environment, mode of genome evolution, and potential role in shaping ecological processes remains poorly understood. Many pioneering discoveries of NCLDV over the last decade have leveraged Amoebozoa as a model host for isolation (1), but it is likely that a variety of other unicellular eukaryotes in the environment are infected by these viruses. Some *Phycodnaviridae* members that infect algae have been studied for decades (6), and several algae-infecting Mimiviruses have recently been isolated from diverse aquatic systems (7, 8). Moreover, recent cultivation-independent analyses have provided tantalizing evidence suggesting that some NCLDV groups are broadly distributed in nature and are potentially playing critical roles in the ecological and evolutionary dynamics of unicellular eukaryotes, particularly in aquatic environments (9, 10).

Understanding the phylogenetic and genomic diversity of NCLDV in the environment is especially critical given recent findings on the biogeochemical significance of virus-mediated metabolic reprogramming of host cells into "virocells" (11–13). The large and dynamic genomes of NCLDVs have been shown to encode a variety of metabolic genes, including those involved in nitrogen metabolism (14), fermentation (8, 14), and sphingolipid biosynthesis (11), which likely contribute to shifts in host physiology during infection. Although these genes are thought to be acquired by NCLDV from diverse sources through Lateral Gene Transfer (LGT), the origin of many of these genes and the extent to which they are characteristic of NCLDV genomes more broadly remains obscure. Given the emerging scientific consensus on their ability to rewire the physiology of globally-abundant protists and impact marine biogeochemistry (15), it is imperative to obtain a comprehensive view of the genome diversity, evolutionary dynamics, and potential metabolic activities of these 'giants' of the virosphere.

Results and Discussion

To address critical questions regarding the genomic diversity, evolutionary relationships, and virocell metabolism of NCLDVs in the environment, we developed a workflow to generate metagenome-assembled genomes (MAGs) of NCLDVs from publicly-available metagenomic data (see Methods, Figure S1). We surveyed 1,545 metagenomes and generated 501 novel NCLDV MAGs that ranged in size from 100-1,400 Kbp. Our workflow included steps to remove potential contamination from cellular organisms and bacteriophage and minimize possible strain heterogeneity in each MAG (see Methods). To ensure our NCLDV MAGs represented nearly-complete genomes, we only retained MAGs that contained at least 4 of 5 key NCLDV marker genes that are known to be highly conserved in these viruses (5) and had a total length > 100 Kbp (see Methods for details and rationale). Most of the MAGs were generated from marine and freshwater environments (444 and 36, respectively), but we also found 21 in metagenomes from bioreactors, wastewater treatment plants, oil fields, and soil samples (labeled “other” in Figure 1, S2; details in Dataset S1).

We constructed a multi-locus phylogenetic tree of the NCLDV MAGs together with 121 reference genomes using five highly conserved genes that have been used previously for phylogenetic analysis of these viruses (5) (Figure 1). The majority of our MAGs placed within the *Mimiviridae* and *Phycodnaviridae* families (350 and 126, respectively), but we also identified new genomes in the *Iridoviridae* (16), *Asfarviridae* (7), *Marseillviridae* (1), and *Pithoviridae* (1). Our phylogeny revealed that the *Phycodnaviridae* are polyphyletic and consist of at least two distinct monophyletic groups, one of which is sister to the *Mimiviridae* (Late *Phycodnaviridae*, 108 MAGs), and one which is basal branching to the *Mimiviridae*-Late *Phycodnaviridae* clade (Early *Phycodnaviridae*, 18 MAGs). In addition to the phylogeny, we evaluated the pairwise Average Amino Acid Identity (AAI) between NCLDV genomes to assess genomic divergence. AAI values provided results that were largely consistent with our phylogenetic analysis, with intra-family AAI values ranging from 26-100% (Figure 1b), highlighting the vast sequence divergence between even NCLDV genomes within the same family.

Given the large diversity within each of the NCLDV families, we sought to identify major clades within these groups that could be used for finer-grained classification. Using the rooted NCLDV phylogeny we calculated optimal clades within each family using the Dunn index (see Methods; (16)), resulting in 54 total clades, including 18 from the *Mimiviridae*, 13 for the Early *Phycodnaviridae*, and 6 for the Late *Phycodnaviridae* (Figure 1, Dataset S1). No cultured representatives were present in 31 of the clades (57%), including 2 from the *Asfarviridae*, 9 from the Early *Phycodnaviridae*, 1 in the *Iridoviridae*, 3 in the Late *Phycodnaviridae*, 1 in the *Marseilleviridae*, 14 in the *Mimiviridae*, and 1 in the *Pithoviridae*. Compared to references available in GenBank, this increases the number of available genomes in the *Mimiviridae* over eightfold (from 47 to 397) and the Phycodnaviridae by over threefold (from 46 to 121), highlighting the vast diversity of environmental NCLDV that have not been sampled using culture-based methods.

Analysis of the genome size distribution across the NCLDV phylogeny provided results that are consistent with the current knowledge of these viruses. For example, Late *Phycodnaviridae* clade 1 contained sequenced representatives of the Prasinoviruses, including known *Ostreococcus* and *Micromonas* viruses, which encode the smallest *Phycodnaviridae* genomes known (17). Consistent with this finding, the MAGs belonging to this clade were also smaller and ranged in size from 100-225 Kbp, suggesting that small genome size is broadly characteristic of this group. By comparison, genomes in the Early *Phycodnaviridae* were larger and formed more divergent groups with long branches, suggesting a large amount of untapped diversity in this clade (Figure 1a). The Pandoraviruses and *Mollivirus sibericum*, notable for their particularly large genomes, formed a distinct clade in the early *Phycodnaviridae*. The Coccolithoviruses and Phaeoviruses (18) were also placed in the Early *Phycodnaviridae*, and we identified 7 and 2 new members of these groups, respectively. Compared to the Late *Phycodnaviridae*, genome sizes of our MAGs were also notably higher in the *Mimiviridae*, which are known to encode among the largest viral genomes. In Mimivirus Clade 16, which includes *Acanthaeomeba polyphaga mimivirus*, we identified 19 new MAGs, 13 of which have genomes > 500 Kbp. Taken together, these results are consistent with the larger genomes that have been observed in the *Mimiviridae* compared to the Late *Phycodnaviridae* (5). Although our NCLDV MAGs contain most marker genes we would expect to find in these genomes, it is likely that many are not complete, and these genome size estimates are therefore best interpreted as underestimates.

To assess the diversity of protein families across the NCLDV families, we calculated orthologous groups (OGs) between our MAGs and 126 reference genomes, resulting in 81,411 OGs (Dataset S2). Of these, only 21,927 (27%) shared homology to known protein families, highlighting the high level of novel genes in NCLDV that has been observed in other studies. Moreover, 55,692 (68%) of the OGs were present in only one NCLDV genome (singleton OGs), and overall the degree distribution of protein family membership revealed only a small number of widely-shared protein families (Figure 2a,b), consistent with what has been shown for dsDNA viruses in general (19). To visualize patterns of gene sharing across the NCLDV we constructed a bipartite network in which both genomes and OGs can be represented (Figure 2c). Analysis of this network revealed primarily family-level clustering, with the *Mimiviridae* and early and late *Phycodnaviridae* clustering near each other, and the *Pithoviridae, Marseiviridae*, and *Poxviridae* clustering separately. Interestingly, although

Pandoraviruses are members of the Early *Phycodnaviridae* clade, they clustered independently in a small sub-network, indicating that the particularly large genomes and novel genomic repertoires in this group are distinct from all other NCLDVs. These patterns suggest that genomic content in the NCLDV is shaped in part by evolutionary history, but that large-scale gains or losses of genomic content can occur over short evolutionary timescales, as has occurred in the *Pandoraviruses*. Overall this is consistent with the “accordion-like” genome evolution that has been postulated for NCLDV (20), whereby lineages evolve through a balanced process of gene gain and loss over long evolutionary timescales. In many respects this mode of genome evolution is not unlike that of *Bacteria* and *Archaea*, where genomic repertoires are shaped by a mixture of vertical inheritance and HGT (21).

To further elucidate the evolutionary history of the large number of genes in NCLDVs, we investigated clade-specific patterns in gene sharing. We found distinct clustering of NCLDV OGs based on their presence in NCLDV clades, indicating that the majority of the OGs are unevenly distributed across clades (Figure 3a). This was confirmed by an enrichment analysis, where we identified sets of enriched OGs in each of the major NCLDV clades (Mann-Whitney U test, corrected p-value < 0.01). The most common functional categories among the clade-specific OGs are predicted to be involved in DNA replication, translation, and transcription. Translational machinery was particularly enriched in Mimivirus clade 16, which contains many cultivated representatives known to have the highest proportion of translation-associated genes of any virus (22, 23). The clade-specific genomic repertoires of NCLDV suggest that this is an appropriate phylogenetic scale for examining functional diversity across the NCLDV, and we anticipate these clades will be useful groupings that can be used in future studies examining spatiotemporal trends in viral diversity in the environment.

Relatively recent studies on model NCLDV-host systems have pointed out the presence of genes involved in rewiring key aspects of cell physiology during infection, such as apoptosis, nutrient processing and acquisition, and oxidative stress regulation (14, 24–26). We found a number of genes involved in such processes to be broadly encoded across NCLDVs, particularly in the *Mimiviridae* and *Phycodnaviridae* families (Figure 3b). Superoxide dismutase (SOD) and Glutathione peroxidase (GPx), key players in regulating cellular oxidative stress, are prevalent in phylogenetically divergent NCLDVs. Giant virus replication possibly occurs under high oxidative stress inside the host cells (26) and thus, the presence of enzymes with antioxidant activity might be crucial in preventing damage to the viral machineries. SOD was biochemically characterized in *Megavirus chilensis*, and was suggested to reduce the oxidative stress induced early in the infection (25). In addition, GPx was found to be upregulated during infection by algal giant viruses (25, 26). Genes putatively involved in the regulation of cellular apoptosis are also widespread in giant viruses, including

C14-family caspase-like proteins and several classes of apoptosis inhibitors, such as Bax1 (27). C14-family metacaspases were reported in a giant virus obtained through single virus genomics approach, while viral activation and recruitment of cellular metacaspase was found during *Emiliania huxleyi virus* (*EhV*) replication (24, 25). In *Chlorella* viruses, a K+ channel (KcV) protein mediates host cell membrane depolarization, facilitating genome delivery within the host (28). We identified KcV in genomes from all the major clades of late *Phycodnaviridae* and *Mimiviridae*, suggesting that host membrane depolarization is a widely-adopted aspect of NCLDV infection strategy. Lastly, in almost all the major *Mimiviridae* and *Phycodnaviridae* clades we detected genes involved in DNA repair and processing, such as photolyases, mismatch repair (*mutS)*, histones, and histone acetyl transferases, of which the latter two have previously been reported in a number of giant virus families, with a possible role of viral histones in packaging of DNA within the capsid (7, 29–31). All together, these results demonstrate that many important aspects of viral reproduction and infection found in cultivated NCLDV are widespread in nature and a common feature of virocell metabolism during giant virus infection.

Viruses are thought to restructure host metabolism during infection to align with virion production rather than cell growth, leading to altered nutrient demands inside the cell (12, 32). We found that genes involved in nutrient acquisition and light-driven energy generation are widespread in several NCLDV clades, including rhodopsins, chlorophyll a/b binding proteins, ferritin, central nitrogen metabolism, and diverse nutrient transporters (Figure 3b). Recent studies on the structure and mechanism of rhodopsin present in two giant viruses have revealed that these are light-driven proton pumps, with potential to reshape energy transfer within the infected host (33, 34). Similarly, widely-distributed chlorophyll a/b binding proteins in giant viruses might increase photosynthetic light-harvesting capacity of infected cells, since protists and plants are known to suppress their photosynthetic machineries, including the chlorophyll binding antenna proteins (25) in response to virus infection (25, 35, 36). Additionally, the presence of the key eukaryotic iron storage protein ferritin (37) and transporters predicted to target ammonium (AmT), phosphorus (Phosphate permease and Phosphate:Na+ symporters), sulfur (TauE/SafE family), and iron (Fe2+/Mn2+ transporters) highlights the shifting nutrient demands of virocells compared to their uninfected counterparts. Most of the MAGs were found in aquatic environments where nutrient availability may be limiting for cellular growth, and alteration of nutrient acquisition strategies during infection may be a key mechanism for increasing viral production. For example, although iron is crucial for photosynthesis and myriad other cellular processes (38), it is often present in low concentrations in marine environments (39, 40), and the production of viral ferritin may aid in regulating the availability of this key micronutrient during virion production. Moreover, nitrogen and phosphorus are limiting for microbial growth in many marine ecosystems, and given the N:C and N:P ratios of viral biomass are relatively higher than that of cellular material (41), it is likely crucial for viruses to boost acquisition of these nutrients with their own transporters. Indeed, a recent study has revealed that an NCLDV-encoded ammonium transporter (AmT) can influence the nutrient flux in host cells by altering the dynamics of ammonium uptake (14).

Strikingly, many NCLDV genomes encode genes involved in central carbon metabolism, including most of the enzymes for glycolysis, gluconeogenesis, the TCA cycle, and the glyoxylate shunt (Figures 3b, 4a, S5). Central carbon metabolism is generally regarded as a fundamental feature of cellular life, and so it is remarkable to consider that giant viruses cumulatively encode nearly every step of these pathways. These genes were particularly enriched in Mimivirus clades 1, 9, and 16, but a few of them were also present in several *Phycodnaviridae* members (Figure 3b,4a). The glycolytic enzymes glyceraldehyde-3-phosphate dehydrogenase (G3P), phosphoglycerate mutase (PGM), and phosphoglycerate kinase (PGK) as well as the TCA cycle enzymes aconitase and succinate dehydrogenase (SDH) were particularly prevalent. Additionally, we identified a fused gene in 16 MAGs that encode the glycolytic enzymes G3P and PGK, which carry out adjacent steps in glycolysis (Figure 4a, b), representing a unique domain architecture that has not been reported in cellular lineages before. Interestingly, in many MAGs, TCA cycle genes were co-localized on viral contigs, suggesting possible co-regulation of these genes during infection (Figure 4c). Remarkably, one NCLDV MAG (ERX552257.96) encoded enzymes for 7 out of 10 steps of glycolysis (Figure 4d), highlighting the high degree of metabolic independence that some giant viruses can achieve from their hosts. The fact that viruses encode these diverse central metabolic pathways underscores their ability to fundamentally reprogram virocell metabolism through manipulation of intracellular carbon fluxes.

Phylogenies of a number of viral metabolic genes identified here together with their cellular homologues revealed that NCLDV sequences tended to group together in deep-branching clades, except for a few cases were multiple acquisitions from cellular sources was evident (Figure 5, Figure S6). For example, aconitase, succinate dehydrogenase subunits B and C, PhoH, glyceraldehyde-3-phosphate dehydrogenase, and superoxide dismutase all showed distinct deep-branching viral clusters and were present in members of multiple NCLDV families, suggesting they diverged from their cellular homologs in the distant past (Figure 5, S6). This pattern was also observed for rhodopsin, similar to previous reports that NCLDV rhodopsins represent a virus-specific clade (33), although our study suggests that at least some NCLDVs independently acquired a bacterial rhodopsin. Phosphoglycerate kinase, chlorophyll a-b binding proteins, and ammonium transporter (AmT) also appear to have been acquired multiple times, but nonetheless show several deep-branching viral clades. These results demonstrate that while NCLDVs have acquired numerous central metabolic genes from cellular hosts, many of these metabolic genes have subsequently diversified into virus-specific lineages. Indeed, detailed functional characterization of viral rhodopsin and Cu-Zn superoxide dismutases has revealed that they have different structural and mechanistic properties compared to the cellular homologs (33, 42), indicating that many metabolic genes in giant viruses evolved to have specific functions in the context of host-virus interactions. Our finding of the fused G3P-PGK glycolytic enzyme in many *Mimivirus* MAGs further reinforces this view and demonstrates that NCLDV are unique drivers of evolutionary innovation in metabolic genes. These results run contrary to a canonical view of viral evolution in which viruses are seen as “pickpockets” that sporadically acquire genes from their cellular hosts rather than encoding their own virus-specific metabolic machinery (43). Although these metabolic enzymes were likely acquired from cellular lineages at some point, their distinct evolutionary trajectory differentiates them from their cellular counterparts and demonstrates that NCLDV are themselves a driver of evolutionary innovation in core metabolic pathways.

Viruses have historically been viewed as “accessories” to cellular life, and as such their influence on biogeochemical cycles has largely been viewed through the lens of their impact on host mortality, rather than any direct metabolic activities of their own. The large number of cellular metabolic genes encoded in NCLDV genomes that we reveal in this study brings to light an alternative view in which virus-specific enzymes have a direct role in shaping virocell physiology. Scaled across viral infections in global aquatic environments, this raises the possibility that viral enzymes can substantially alter global biogeochemical fluxes in their own right. Moreover, the distinct evolutionary lineages of viral metabolic genes implicate NCLDV as unique drivers of metabolic innovation, in stark contrast to the traditional view in which they are merely occasional “pickpockets” of cellular genes rather than *de facto* evolutionary innovators. Taken together, these findings argue that just as microbes are considered the “engines that shape global biogeochemical cycles” (44), viruses must be considered alongside their cellular counterparts as agents of metabolic fluxes with their own encoded physiology.

Figure Legends

Figure 1. A) Phylogeny of the 501 NCLDV MAGs presented in this study together with 121 reference genomes. The phylogeny was constructed from a concatenated alignment of 5 highly conserved marker genes that are present throughout the NCLDV families using the VT+F+I+G4 model in IQ-TREE. The tree is rooted at *Poxviridae/Asfarviridae* branch, consistent with previous studies (5). The inner strip is colored according to the phylogeny of the MAGs, while the outer strip is colored according to the habitat in which they were found.

The bar chart represents genome size, which ranges from 100 - 2,474 Kbp, and the dotted line denotes the 500 Kbp mark. Clades with >5 genomes are indicated with two letter abbreviations and clade numbers. Abbreviations: MM: *Mimiviridae*, EP: Early *Phycodnaviridae*, LP: Late *Phycodnaviridae*, IR: *Iridoviridae*, MR: *Marseilleviridae*, PT: *Pithoviridae*. For the list of all the clades, see Dataset 1. B) Average amino acid identity (AAI) heatmap of the MAGs and reference genomes, with rows and columns clustered according to the phylogeny.

Figure 2. A) The distribution of the orthologous groups (OGs) in the NCLDV MAGs and reference genomes. The barplot on the left shows the proportion of OGs in each frequency category that could be assigned an annotation, while the barplot on the right shows the total number of OGs in each frequency category (log scale). B) The degree distribution of the OG occurrence in the genomes analyzed. The best fit to a power law distribution is also shown. C) A bipartite network of the OGs, with large nodes corresponding to genomes and small nodes corresponding to OGs. The size of the genome nodes is proportional to their genome size, and they are colored according to their family-level classification.

Figure 3. A) The barplot shows the number of enriched OGs in each of the major NCLDV clades analyzed in this study. Only a subset of total functional categories are shown here; a full table can be found in Dataset S2. B) A heatmap showing the occurrence of OGs with >5 total members across the major NCLDV clades, with shading corresponding to the percent of MAGs in that clade that encode a given OG. C) A bubble plot of select metabolic genes detected in the NCLDV clades, with bubble size proportional to the percent of genomes in a clade that encode that protein. Abbreviations: G3P: glycerol-3-phosphate; LCM: Large conductance mechanosensitive; SCM; small conductance mechanosensitive.

Figure 4. A) Presence of central carbon metabolism enzymes in the NCLDVs. The number of genomes harboring a particular enzyme is provided beside its abbreviated name. Enzymes that were not detected in any of the studied NCLDVs are in grey. B) Representative CDS from genome ERX552243.92 illustrating the domain organization (PFAM and Interpro) of the fused-domain gene (G3P + PGK) involved in glycolysis, that was detected in 16 of the NCLDV MAGs. C) Example of co-localization of genes involved in TCA cycle on genomic contigs from five representative NCLDV MAGs. Location of a number of other genes commonly present in NCLDVs are also shown. D) Presence/absence of genes involved in central-carbon metabolism in NCLDV genomes assembled in this study. Only the genomes harboring 3 or more enzymes are shown. G3P+PGK indicates the fused-domain gene illustrated in panel B. Blue arrow indicates the genome that harbors 7 out of 10 enzymes involved in glycolysis. Abbreviations: **HK**: hexokinase, **PGI**: Phosphoglucoisomerase, **PFK**: Phosphofructokinase, **ALD**: aldolase, **TPI:** Triose-phosphate isomerase, **G3P**: Glyceraldehyde 3-phosphate

dehydrogenase, **PGK**: Phosphoglycerate kinase, **PGM**: Phosphoglycerate mutase, **ENO**: Enolase, **PYK**: Pyruvate kinase, **PEPCK**: PEP carboxykinase, **FBP**: Fructose 1,6-bisphosphatase, **G6P**: Glucose 6-phosphatase, **PDH**: Pyruvate dehydrogenase, **PC**: Pyruvate carboxylase, **CS**: Citrate synthase, **ACON**: Aconitase, **ICL**: Isocitrate lyase, **ICD**: Isocitrate dehydrogenase, **αKDH**: α-ketoglutarate dehydrogenase, **SCS**: Succinyl-CoA synthetase, **SD**: Succinate dehydrogenase (subunits A, B and C), **FH**: Fumarate hydratase, **MS**: Malate synthase, **MDH**: Malate dehydrogenase.

Figure 5. Phylogenetic reconstruction of a number of representative NCLDV genes likely involved in carbon and nutrient metabolism and light harvesting. NCLDV-specific clusters are encircled with dashed ovals in each of the trees, while number of genes from different NCLDV-clades contributing to these monophyletic groups are also provided (MM: *Mimiviridae*, EP: Early *Phycodnaviridae*, LP: Late *Phycodnaviridae*) Colors of the clade names correspond to those in Figure 1. Although node support values are not provided for better visual clarity, all the NCLDV-specific nodes are supported by >90% ultrafast bootstrap values (see Methods and Data availability statement for details).

* - Bacteriophage sequences are only present in the PhoH tree.

** - Unclassified sequences (environmental) are only present in the Rhodopsin tree.

Methods

Data Availability

Nucleotide sequences of NCLDV MAGs, predicted proteins, alignments used for phylogenies, raw orthologous groups membership files, and other major data products are available on the Aylward Lab Figshare account: https://figshare.com/authors/Frank_Aylward/5008391 in the project titled “NCLDV”.

Assembling NCLDV Genomes from Metagenomes. Although phylogenetic binning of metagenomic contigs belonging to Archaea and Bacteria is now commonplace (45), this approach is rarely used for viruses, and it was therefore necessary for us to develop a novel workflow to recover high-confidence NCLDV genomes from metagenomic data. Moreover, methods for assessing the completeness and potential contamination of prokaryotic bins, such as employed by the popular tool CheckM, rely on knowledge of shared single-copy protein families in different lineages, but this information is not applicable to viral bins given their fundamentally distinct genomic repertoires. We therefore also developed a workflow for quality-checking NCLDV bins. The overall process can be divided into three main stages: 1) initial binning of contigs, 2) identification of bins corresponding to NCLDV, and 3) quality-checking bins to ensure contamination is not present. An overview of this workflow can be found in Figure S1.

1) **Initial binning of contigs.** We obtained assembled contigs (> 10 Kbp) and coverage files for 1545 metagenomes that had been previously assembled in a large-scale study that examined bacterial and archaeal diversity (46). We chose to use the program Metabat2 (47) for binning because this program bins contigs based on sequence coverage and tetranucleotide frequencies, which are metrics that would be expected to be consistent in viral genomes and therefore useful for binning. Moreover, although some binning tools rely on marker gene sets that are specific to cellular lineages, MetaBat2 does not use marker genes for binning and is therefore more appropriate for binning viral sequences. We used the parameters -s 100000, -m 10000, --minS 75, --maxEdges 75, which are more stringent than the default parameters and would be expected to yield more conservative, high-confidence binning results.

2) **Screening Bins.** After binning contigs it is necessary to screen the bins to identify ones correspond to putative NCLDV genomes. It has previously been shown that 5 highly conserved and NCLDV-specific protein families are present in almost all known NCLDV genomes, and we therefore used these for screening. The protein families correspond to the Major Capsid Protein (MCP), Superfamily II helicase (SFII), Virus-like transcription factor (VLTF3), DNA Polymerase B (PolB), and packaging ATPase (A32). We predicted protein sequences from all bins using Prodigal (48) (default parameters), and matched proteins to the 5 NCLDV marker genes using HMMER3 (49) with custom Hidden Markov Models (HMMS, see section “protein families used for screening” below). We only considered bins that had >= 4 of the markers for further analysis in order to exclude NCLDV genomes that were mostly incomplete. Moreover, we only considered bins > 100 Kbp on the grounds that the smallest NCLDV genome is 103 Kbp (50), and that a higher cutoff may bias recovery against NCLDV clades with smaller genome sizes. After implementing these screens we recovered 517 candidate bins.

3) **Quality-checking Bins.** To ensure that the bins were indeed viral and did not include contamination for cellular sources, we screened all contigs using ViralRecall (https://github.com/faylward/viralrecall). This tool compares encoded proteins to virus-specific (from the VOG database) and cellular-specific HMMs to assess their provenance, and is therefore useful in determining contigs that may represent contamination from a cellular organism. ViralRecall uses custom subsets of the Viral Orthologous Groups (VOG: vogdb.org) and Pfam databases (51) for the virus-specific

and cellular-specific HMMs, respectively, and generates a score for each contig (negative scores indicating more hits to cellular HMMs, positive scores indicating more hits to viral HMMs). In addition to ViralRecall we also used LAST (parameter -m 500; (51, 52)) to compare all of the encoded proteins in each contig to RefSeq 92 (53), and recorded the top 5 hits for each protein. To remove contigs that derived from cellular organisms, we removed contigs that had a ViralRecall score < 0 (indicting a net cellular signal), contained < 3 encoded proteins with hits total to HMMs in the VOG database, and had no LAST hits to known NCLDV proteins. Additionally, to exclude possible bacteriophage sequences, we removed contigs for which the encoded proteins had at least one LAST hit to a bacteriophage and zero hits to a known NCLDV genome. A summary of all of the contigs removed in this way can be found in Dataset S1.

To uncover strain heterogeneity, we identified cases where marker genes were found in multiple copies. We used SFII, VLTF3, A32, and PolB, for this, excluding MCP because multiple copies of this gene is commonplace in NCLDV genomes. We identified 16 bins where more than one marker gene was found to be present in multiple copies, and excluded these bins from further analysis. Of the remaining 501 bins, 62 contained one marker gene that was not single-copy, but these were retained because some complete NCLDV genomes contain multiple copies, and overly strict thresholds would exclude potentially novel NCLDV lineages with genomic repertoires distinct from what has been observed. Overall 501 bins passed all screening procedures and are subsequently referred to as NCLDV Metagenome Assembled Genomes (MAGs). Overall by using strict binning parameters and excluding possible cellular or bacteriophage contigs, we recovered high quality MAGs that consisted of relatively few contigs (422 bins with < 20 contigs, including 2 bins with only a single contig; mean N50 contig size of 37.4 Kbp across MAGs).

Protein families used for screening. To screen preliminary bins and identify metagenome-assembled NCLDV genomes, we used a custom set of HMMs created for 5 NCLDV-specific protein families: The Major Capsid Protein (MCP), Superfamily II helicase (SFII), Virus-like transcription factor (VLTF3), DNA Polymerase B (PolB), and packaging ATPase (A32). These 5 protein families have previously been used for phylogenetic analysis of NCLDV and are typically not found in cellular organisms (5). To generate these models, we manually annotated proteins from 126 complete NCLDV genomes available in NCBI that span the 7 major families. We then generated model-specific HMMER3 score cutoffs based on the scores recovered from matching known protein family members to these HMMs (Figure S4). These scores were used in determining the presence/absence of these protein families in the metaBAT2 bins.

Phylogenetic Reconstruction of NCLDV MAGs. To assess the phylogeny of the NCLDV MAGs we generated a concatenated tree of all 501 MAGs together with 121 reference NCLDV genomes using the marker genes PolB, VLTF3, MCP, A32, and SFII. These proteins have previously been shown to be useful for phylogenetic analysis of NCLDV. In some cases, NCLDV are known to have introns or split genes, and we generated a Python script to identify these cases, check to ensure the proteins hit to the same HMM and had no sequence overlap, and subsequently concatenate the proteins. Alignments were created using ClustalOmega, and trimAl was used for trimming (parameter -gt 0.1). We ran IQ-TREE (54) with the "-m TEST" ModelFinder option (55), which identified VT+F+I+G4 as the optimal model. We then ran IQ-TREE on the alignment with 1000 ultrafast bootstraps to assess confidence (56).

Clade Delineation. Given the large phylogenetic diversity of NCLDV examined in this study, we sought to identify clades of closely related viruses within each of the major families. To this end we used the Dunn index to identify optimal clade-level delineations in our multi-locus phylogenetic tree of the NCLDV. We first generated a rooted ultrametric phylogenetic tree in R using the "ape" package and generated clades at different tree cut heights. For each cut height we calculated the Dunn index (16) using the "cluster.stats" package in R. We found that a height of 2.45 had the lowest Dunn index and therefore provided the best clustering (Figure S3). All clusters were then manually inspected and edited to ensure that they represented monophyletic groups, and these final clusters were used as clades.

Generation of Orthologous Groups and Annotation. We used ProteinOrtho v6.06 (57) to calculate the orthologous groups shared between the 501 NCLDV MAGs and 127 reference genomes. Protein files were generated using Prodigal with default parameters. Because of the accelerated evolutionary rate of viruses, we used the relaxed parameters "-e=1e-3 --identity=15 -p=blastp+ --selfblast --cov=40" for proteinortho. This resulted in 81,412 orthologous groups (OGs). For each OG, we randomly selected a representative for annotation. Representatives were compared to the EggNOG 4.5, TIGRFam, Pfam, VOG, and COG databases (e-value 1e-5), and best his were recorded.

Bipartite Network Analysis. Bipartite networks of NCLDV genomes and their protein families were created in igraph. For visualization purposes, only protein families present in > 5 genomes were analyzed. In the bipartite graph two node types were present: Genome nodes and Protein Family nodes. Each protein family node was connected to a genome node if it was encoded in that genome. The spring-directed layout was generated using the layout.fruchterman.reingold() command with 10,000 iterations.

Average Amino Acid Identity Calculation. AAI between the NCLDV genomes was calculated using a custom Python script available on GitHub (https://github.com/faylward/pangenomics/blob/master/lastp_aai.py). The script employs pairwise LASTP searches (52) (parameter -m 500) and calculates the AAI and alignment fraction (AF) between all genome pairs. For visualization purposes, genome pairs in which one member had an AF < 10 were considered to have an AAI of 0.

Clade-specific OG enrichment. To evaluate if specific NCLDV clades were enriched in particular OGs, we performed an enrichment analysis on all clades with >5 members. For each OG, a Mann-Whitney U test was performed on overall OG membership in that clade compared to all other NCLDV (including reference genomes). Only OGs with a known annotation that were present in >6 genomes were used. P-values were corrected using the Benjamini-Hochberg procedure in R (58), and values < 0.01 were considered significant.

Phylogenetic reconstruction of NCLDV metabolic genes: We collected the reference sequences for most of gene trees from the EggNOG database (59) with the following exceptions: for phosphate permease and chlorophyll a/b binding proteins we used sequences from the Pfam database (51), while rhodopsin sequences were collected from the MicRhoDE, a dedicated server for rhodopsins from different domains of life and environments (60). In case of the superoxide dismutase (SOD) and phoH genes, we curated additional sequences from viruses other than NCLDVs from the NCBI Refseq database. For each gene, a diagnostic tree was built using FastTree (61) implemented in the ETE3 package. The diagnostic trees were used to select a smaller set of reference sequences and remove short or redundant sequences. For construction of the final trees, we aligned the sequences using ClustalOmega and trimmed using trimAl (parameter -gt 0.1). IQ-TREE (54) was used to build maximum likelihood phylogenetic trees with the model 'LG + I +G4' and 1000 ultrafast bootstrap replicates (56).

References

1. N. Brandes, M. Linial, Giant Viruses – Big Surprises https:/doi.org/10.20944/preprints201904.0172.v1.

2. D. Raoult, P. Forterre, Redefining viruses: lessons from Mimivirus. *Nat. Rev. Microbiol.* **6**, 315–319 (2008).

3. B. La Scola, *et al.*, A giant virus in amoebae. *Science* **299**, 2033 (2003).

4. S. Aherfi, P. Colson, B. La Scola, D. Raoult, Giant Viruses of Amoebas: An Update. *Front. Microbiol.* **7**, 349 (2016).

5. E. V. Koonin, N. Yutin, Multiple evolutionary origins of giant viruses. *F1000Res.* **7** (2018).

6. A. Jeanniard, *et al.*, Towards defining the chloroviruses: a genomic journey through a

genus of large DNA viruses. *BMC Genomics* **14**, 158 (2013).

7. M. Moniruzzaman, *et al.*, Genome of brown tide virus (AaV), the little giant of the Megaviridae, elucidates NCLDV genome expansion and host-virus coevolution. *Virology* **466-467**, 60–70 (2014).

8. C. R. Schvarcz, G. F. Steward, A giant virus infecting green algae encodes key fermentation genes. *Virology* **518**, 423–433 (2018).

9. Y. Li, *et al.*, Degenerate PCR Primers to Reveal the Diversity of Giant Viruses in Coastal Waters. *Viruses* **10** (2018).

10. M. Moniruzzaman, *et al.*, Diversity and dynamics of algal Megaviridae members during a harmful brown tide caused by the pelagophyte, Aureococcus anophagefferens. *FEMS Microbiol. Ecol.* **92**, fiw058 (2016).

11. A. Vardi, *et al.*, Host-virus dynamics and subcellular controls of cell fate in a natural coccolithophore population. *Proceedings of the National Academy of Sciences* **109**, 19327–19332 (2012).

12. P. Forterre, The virocell concept and environmental microbiology. *ISME J.* **7**, 233–236 (2013).

13. B. M. Schieler, *et al.*, Nitric oxide production and antioxidant function during viral infection of the coccolithophore Emiliania huxleyi. *ISME J.* **13**, 1019–1031 (2019).

14. A. Monier, *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).

15. A. E. Zimmerman, *et al.*, Metabolic and biogeochemical consequences of viral infection in aquatic ecosystems. *Nature Reviews Microbiology* (2019) https:/doi.org/10.1038/s41579-019-0270-x.

16. J. C. Dunn†, Well-Separated Clusters and Optimal Fuzzy Partitions. *Journal of Cybernetics* **4**, 95–104 (1974).

17. K. D. Weynberg, M. J. Allen, W. H. Wilson, Marine Prasinoviruses and Their Tiny Plankton Hosts: A Review. *Viruses* **9** (2017).

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

19. J. Iranzo, M. Krupovic, E. V. Koonin, The Double-Stranded DNA Virosphere as a Modular Hierarchical Network of Gene Sharing. *MBio* **7** (2016).

20. J. Filée, Route of NCLDV evolution: the genomic accordion. *Curr. Opin. Virol.* **3**, 595–599 (2013).

21. V. Kunin, C. A. Ouzounis, The balance of driving forces during genome evolution in prokaryotes. *Genome Res.* **13**, 1589–1594 (2003).

22. F. Schulz, *et al.*, Giant viruses with an expanded complement of translation system components. *Science* **356**, 82–85 (2017).

23. J. S. Abrahão, R. Araújo, P. Colson, B. La Scola, The analysis of translation-related gene set boosts debates around origin and evolution of mimiviruses. *PLoS Genet.* **13**, e1006532 (2017).

24. K. D. Bidle, L. Haramaty, J. Barcelos E Ramos, P. Falkowski, Viral activation and recruitment of metacaspases in the unicellular coccolithophore, Emiliania huxleyi. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 6049–6054 (2007).

25. M. Moniruzzaman, E. R. Gann, S. W. Wilhelm, Infection by a Giant Virus Induces Widespread Physiological Reprogramming in Aureococcus Anophagefferens – A Harmful Bloom Algae https:/doi.org/10.1101/256149.

26. U. Sheyn, S. Rosenwasser, S. Ben-Dor, Z. Porat, A. Vardi, Modulation of host ROS metabolism is essential for viral infection of a bloom-forming coccolithophore in the ocean. *ISME J.* **10**, 1742–1754 (2016).

27. K. S. Robinson, A. Clements, A. C. Williams, C. N. Berger, G. Frankel, Bax Inhibitor 1 in apoptosis and disease. *Oncogene* **30**, 2391–2400 (2011).

28. M. Neupärtl, *et al.*, Chlorella viruses evoke a rapid release of K+ from host cells during the early phase of infection. *Virology* **372**, 340–348 (2008).

29. M. G. Fischer, M. J. Allen, W. H. Wilson, C. A. Suttle, Giant virus with a remarkable complement of genes infects marine zooplankton. *Proceedings of the National Academy of Sciences* **107**, 19508–19513 (2010).

30. V. Thomas, *et al.*, Lausannevirus, a giant amoebal virus encoding histone doublets. *Environ. Microbiol.* **13**, 1454–1466 (2011).

31. K. Okamoto, *et al.*, Cryo-EM structure of a Marseilleviridae virus particle reveals a large internal microassembly. *Virology* **516**, 239–245 (2018).

32. S. Rosenwasser, C. Ziv, S. G. van Creveld, A. Vardi, Virocell Metabolism: Metabolic Innovations During Host–Virus Interactions in the Ocean. *Trends in Microbiology* **24**, 821–832 (2016).

33. D. M. Needham, *et al.*, A distinct lineage of giant viruses brings a rhodopsin photosystem to unicellular marine predators. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 20574–20583 (2019).

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

35. P. Juneau, J. E. Lawrence, C. A. Suttle, P. J. Harrison, Effects of viral infection on photosynthetic processes in the bloom-forming alga Heterosigma akashiwo. *Aquatic Microbial Ecology* **31**, 9–17 (2003).

36. J. Rahoutei, I. Garcia-Luque, M. Baron, Inhibition of photosynthesis by viral infection: Effect on PSII structure and function. *Physiologia Plantarum* **110**, 286–292 (2000).

37. H. Botebol, *et al.*, Central role for ferritin in the day/night regulation of iron homeostasis in marine phytoplankton. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 14652–14657 (2015).

38. J. Morrissey, C. Bowler, Iron utilization in marine cyanobacteria and eukaryotic algae. *Front. Microbiol.* **3**, 43 (2012).

39. S. L. Hogle, *et al.*, Pervasive iron limitation at subsurface chlorophyll maxima of the California Current. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 13300–13305 (2018).

40. M. J. Behrenfeld, Z. S. Kolber, Widespread iron limitation of phytoplankton in the south pacific ocean. *Science* **283**, 840–843 (1999).

41. L. F. Jover, T. C. Effler, A. Buchan, S. W. Wilhelm, J. S. Weitz, The elemental composition of virus particles: implications for marine biogeochemical cycles. *Nat. Rev. Microbiol.* **12**, 519–528 (2014).

42. A. Lartigue, *et al.*, The megavirus chilensis Cu,Zn-superoxide dismutase: the first viral structure of a typical cellular copper chaperone-independent hyperstable dimeric enzyme. *J. Virol.* **89**, 824–832 (2015).

43. D. Moreira, P. López-García, Ten reasons to exclude viruses from the tree of life. *Nat. Rev. Microbiol.* **7**, 306–311 (2009).

44. P. G. Falkowski, T. Fenchel, E. F. Delong, The microbial engines that drive Earth's biogeochemical cycles. *Science* **320**, 1034–1039 (2008).

45. N. Sangwan, F. Xia, J. A. Gilbert, Recovering complete and draft population genomes from metagenome datasets. *Microbiome* **4**, 8 (2016).

46. D. H. Parks, *et al.*, Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol* **2**, 1533–1542 (2017).

47. D. D. Kang, *et al.*, MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).

48. D. Hyatt, *et al.*, Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).

49. S. R. Eddy, Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).

50. C. A. Tidona, G. Darai, The complete DNA sequence of lymphocystis disease virus. *Virology* **230**, 207–216 (1997).

51. S. El-Gebali, *et al.*, The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).

52. S. M. Kiełbasa, R. Wan, K. Sato, P. Horton, M. C. Frith, Adaptive seeds tame genomic sequence comparison. *Genome Res.* **21**, 487–493 (2011).

53. N. A. O'Leary, *et al.*, Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–45 (2016).

54. L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, B. Q. Minh, IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).

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

56. D. T. Hoang, O. Chernomor, A. von Haeseler, B. Q. Minh, L. S. Vinh, UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).

57. M. Lechner, *et al.*, Proteinortho: detection of (co-)orthologs in large-scale analysis. *BMC Bioinformatics* **12**, 124 (2011).

58. Y. Benjamini, Y. Hochberg, Controlling the False Discovery Rate: A Practical and

Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289–300 (1995).

59. J. Huerta-Cepas, *et al.*, eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* **44**, D286–93 (2016).

60. D. Boeuf, S. Audic, L. Brillet-Guéguen, C. Caron, C. Jeanthon, MicRhoDE: a curated database for the analysis of microbial rhodopsin diversity and evolution. *Database* **2015** (2015).

61. M. N. Price, P. S. Dehal, A. P. Arkin, FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).

Figure 1.

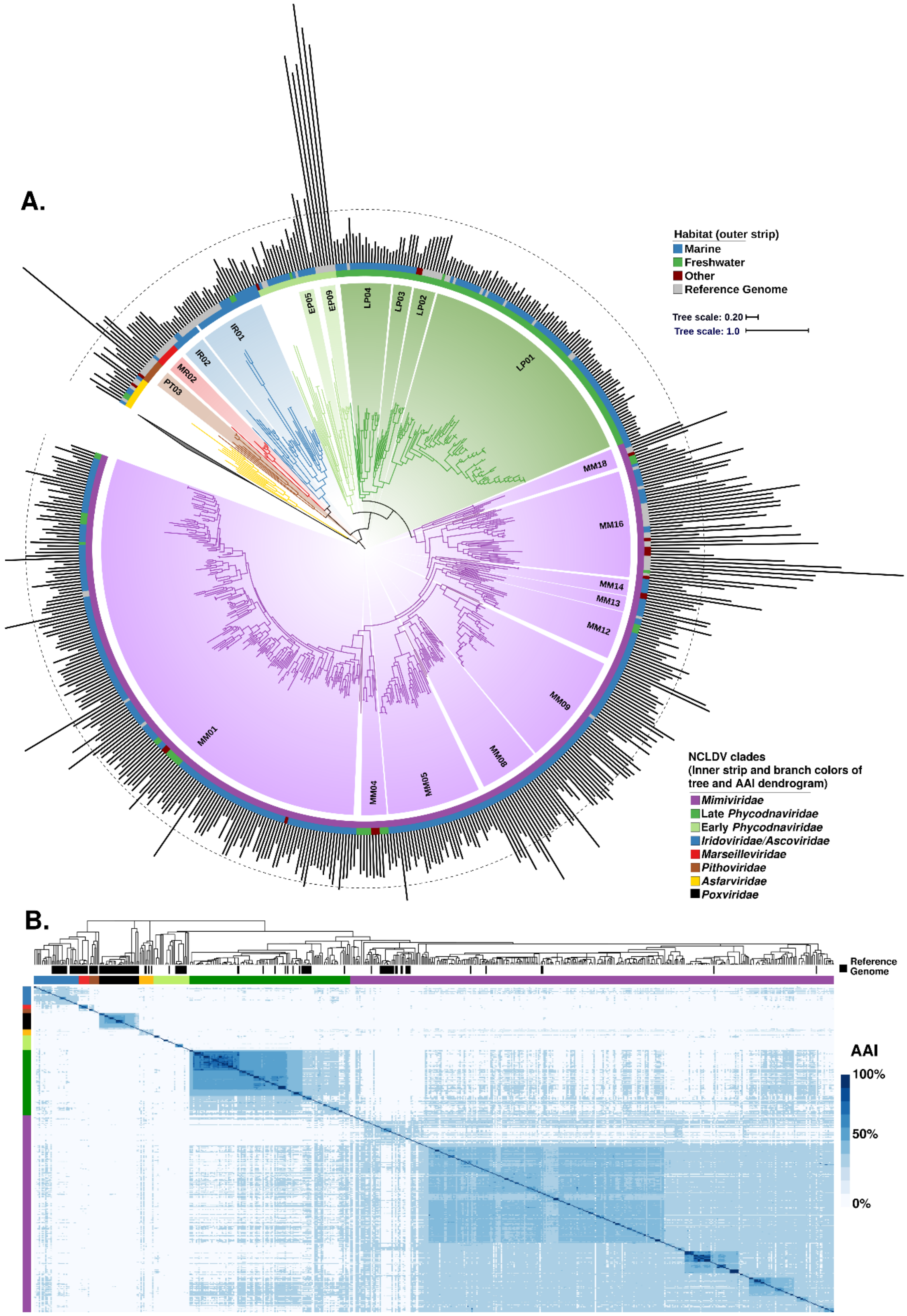

A.
Habitat (outer strip)
Marine
Freshwater
Other
Reference Genome
Tree scale: 0.20
Tree scale: 1.0
EP05
EP09
LP04
LP03
LP02
LP01
IR01
IR02
MR02
PT03
MM18
MM16
MM14
MM13
MM12
MM09
MM08
MM05
MM04
MM01
NCLDV clades
(Inner strip and branch colors of tree and AAI dendrogram)
Mimiviridae
Late Phycodnaviridae
Early Phycodnaviridae
Iridoviridae/Ascoviridae
Marseilleviridae
Pithoviridae
Asfarviridae
Poxviridae
B.
Reference Genome
AAI
100%
50%
0%

Figure 2.

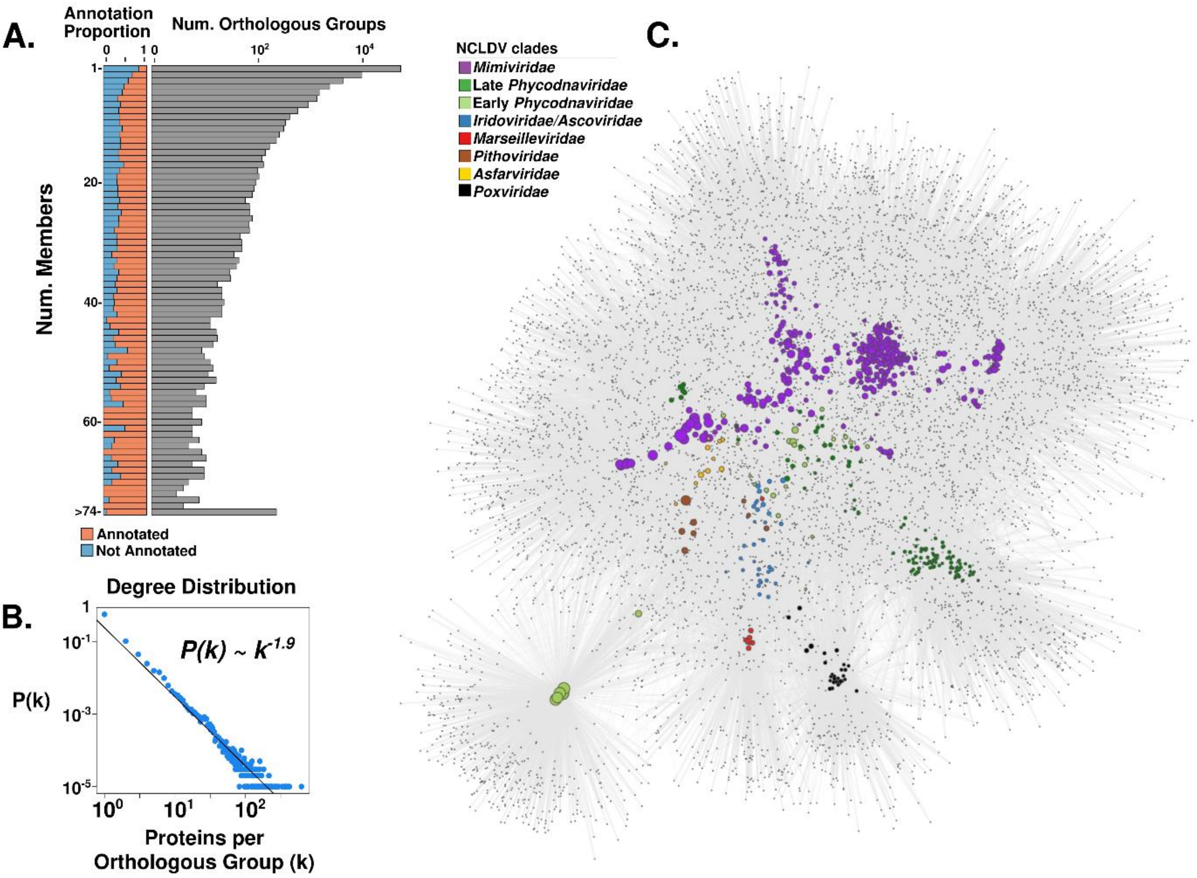

A.
Annotation Proportion
0
1
Num. Orthologous Groups
0
10²
10⁴
Num. Members
1
20
40
60
>74
Annotated
Not Annotated
B.
Degree Distribution
P(k) ~ k^-1.9
P(k)
1
10⁻¹
10⁻³
10⁻⁵
10⁰
10¹
10²
Proteins per Orthologous Group (k)
C.
NCLDV clades
Mimiviridae
Late Phycodnaviridae
Early Phycodnaviridae
Iridoviridae/Ascoviridae
Marseilleviridae
Pithoviridae
Asfarviridae
Poxviridae

Figure 3.

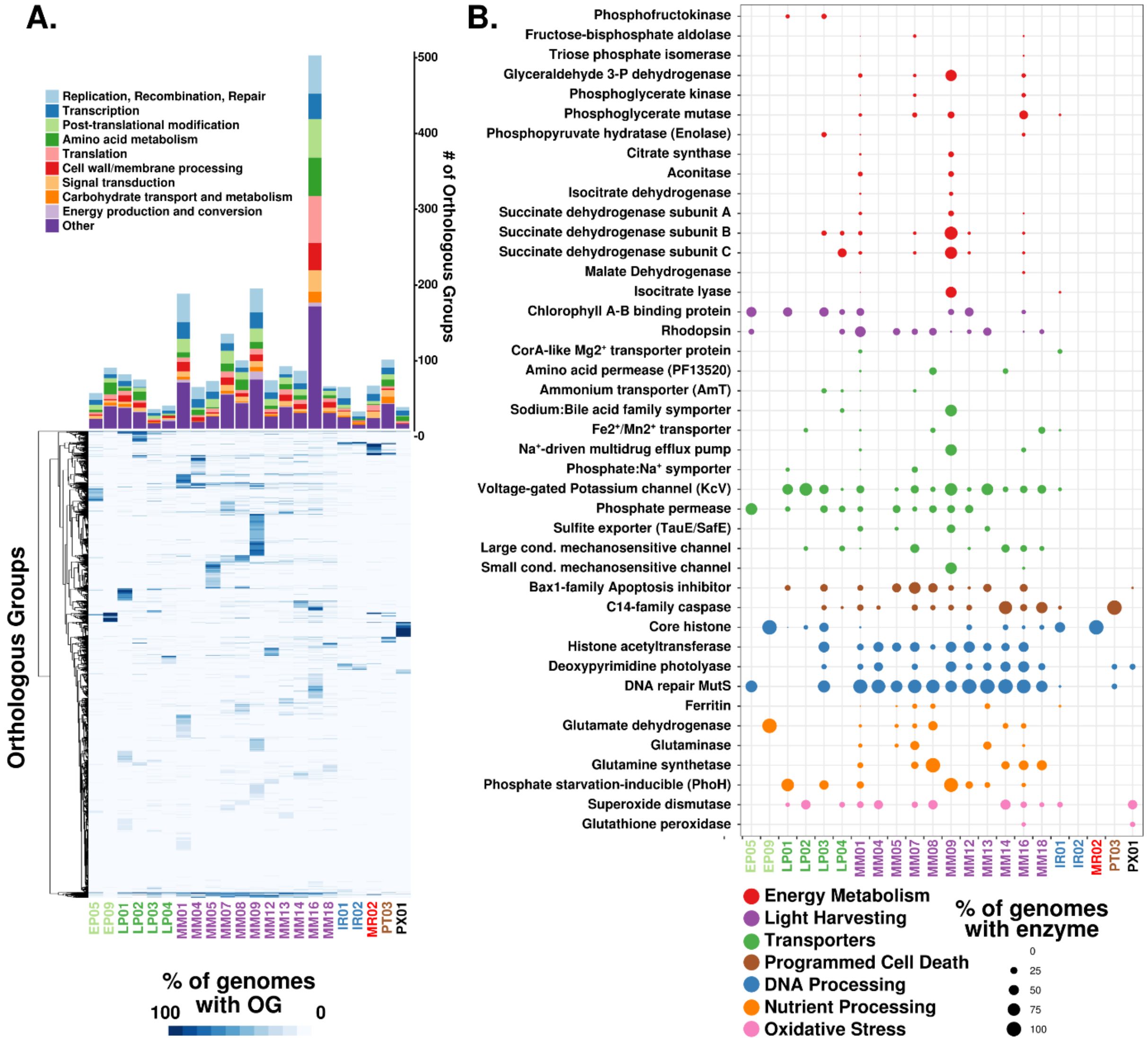

A.
Replication, Recombination, Repair
Transcription
Post-translational modification
Amino acid metabolism
Translation
Cell wall/membrane processing
Signal transduction
Carbohydrate transport and metabolism
Energy production and conversion
Other
# of Orthologous Groups
500
400
300
200
100
0
Orthologous Groups
EP05 EP09 LP01 LP02 LP03 LP04 MM01 MM04 MM05 MM07 MM08 MM09 MM12 MM13 MM14 MM16 MM18 IR01 IR02 MR02 PT03 PX01
% of genomes with OG
100
0
B.
Phosphofructokinase
Fructose-bisphosphate aldolase
Triose phosphate isomerase
Glyceraldehyde 3-P dehydrogenase
Phosphoglycerate kinase
Phosphoglycerate mutase
Phosphopyruvate hydratase (Enolase)
Citrate synthase
Aconitase
Isocitrate dehydrogenase
Succinate dehydrogenase subunit A
Succinate dehydrogenase subunit B
Succinate dehydrogenase subunit C
Malate Dehydrogenase
Isocitrate lyase
Chlorophyll A-B binding protein
Rhodopsin
CorA-like Mg2+ transporter protein
Amino acid permease (PF13520)
Ammonium transporter (AmT)
Sodium:Bile acid family symporter
Fe2+/Mn2+ transporter
Na+-driven multidrug efflux pump
Phosphate:Na+ symporter
Voltage-gated Potassium channel (KcV)
Phosphate permease
Sulfite exporter (TauE/SafE)
Large cond. mechanosensitive channel
Small cond. mechanosensitive channel
Bax1-family Apoptosis inhibitor
C14-family caspase
Core histone
Histone acetyltransferase
Deoxypyrimidine photolyase
DNA repair MutS
Ferritin
Glutamate dehydrogenase
Glutaminase
Glutamine synthetase
Phosphate starvation-inducible (PhoH)
Superoxide dismutase
Glutathione peroxidase
EP05 EP09 LP01 LP02 LP03 LP04 MM01 MM04 MM05 MM07 MM08 MM09 MM12 MM13 MM14 MM16 MM18 IR01 IR02 MR02 PT03 PX01
● Energy Metabolism
● Light Harvesting
● Transporters
● Programmed Cell Death
● DNA Processing
● Nutrient Processing
● Oxidative Stress
% of genomes with enzyme
0
25
50
75
100

Figure 4.

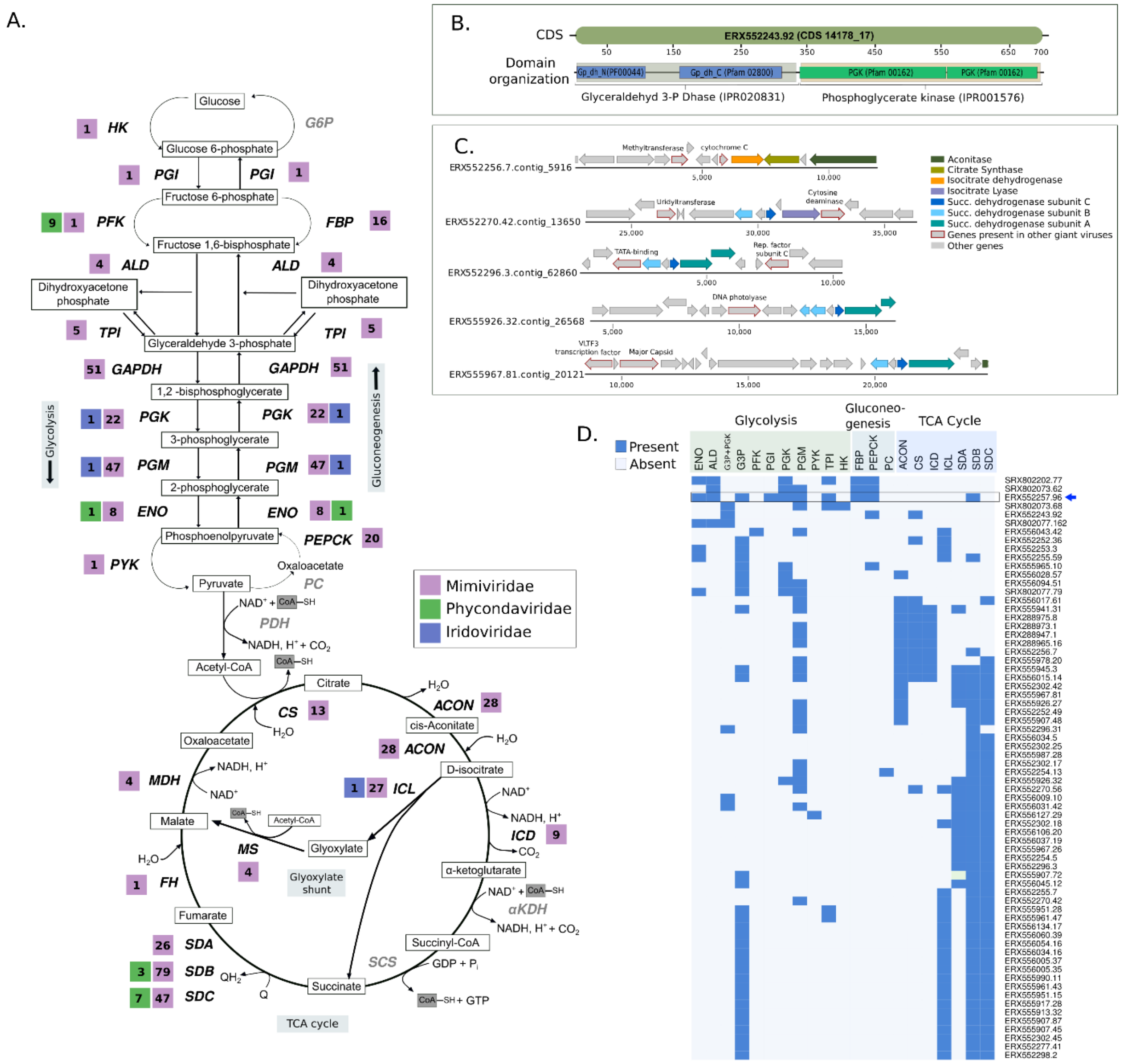
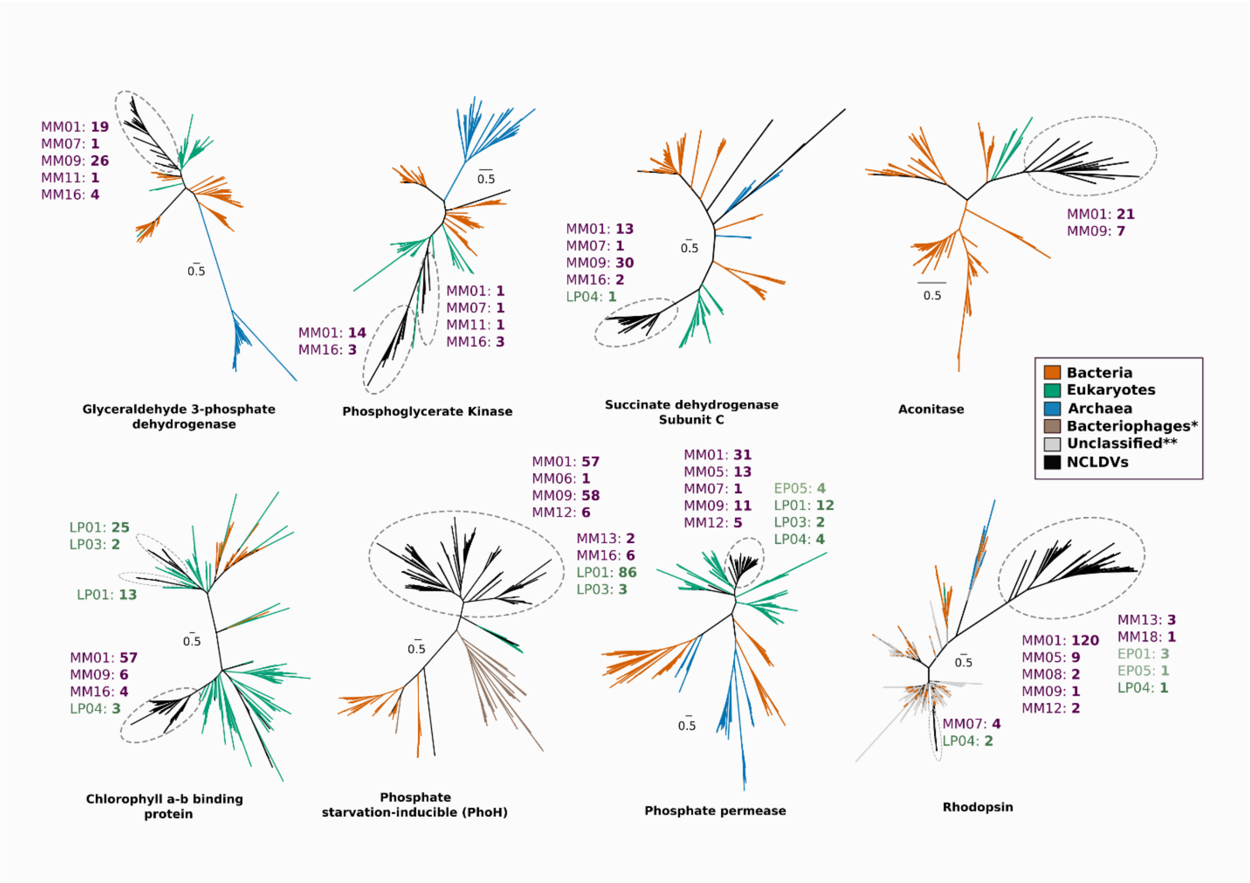

A.
Glucose
G6P
1 HK
Glucose 6-phosphate
1 PGI
PGI 1
Fructose 6-phosphate
9 1 PFK
FBP 16
Fructose 1,6-bisphosphate
4 ALD
ALD 4
Dihydroxyacetone phosphate
Dihydroxyacetone phosphate
5 TPI
TPI 5
Glyceraldehyde 3-phosphate
51 GAPDH
GAPDH 51
1,2 -bisphosphoglycerate
1 22 PGK
PGK 22 1
3-phosphoglycerate
1 47 PGM
PGM 47 1
2-phosphoglycerate
1 8 ENO
ENO 8 1
Phosphoenolpyruvate
PEPCK 20
1 PYK
Oxaloacetate
Pyruvate
PC
Glycolysis
Gluconeogenesis
NAD⁺ + CoA–SH
PDH
NADH, H⁺ + CO₂
Acetyl-CoA
CoA–SH
Citrate
H₂O
CS 13
ACON 28
H₂O
cis-Aconitate
Oxaloacetate
28 ACON
H₂O
NADH, H⁺
D-isocitrate
4 MDH
NAD⁺
1 27 ICL
NAD⁺
CoA–SH
Acetyl-CoA
NADH, H⁺
Malate
ICD 9
CO₂
Glyoxylate
MS
4
H₂O
1 FH
Glyoxylate shunt
α-ketoglutarate
NAD⁺ + CoA–SH
αKDH
NADH, H⁺ + CO₂
Fumarate
26 SDA
Succinyl-CoA
3 79 SDB
QH₂
SCS
GDP + Pᵢ
7 47 SDC
Q
Succinate
CoA–SH + GTP
TCA cycle
Mimiviridae
Phycondaviridae
Iridoviridae
B.
CDS
ERX552243.92 (CDS 14178_17)
50
150
250
350
450
550
650
700
Domain organization
Gp_dh_N(PF00044)
Gp_dh_C (Pfam 02800)
PGK (Pfam 00162)
PGK (Pfam 00162)
Glyceraldehyd 3-P Dhase (IPR020831)
Phosphoglycerate kinase (IPR001576)
C.
Methyltransferase
cytochrome C
ERX552256.7.contig_5916
5,000
10,000
Uridyltransferase
Cytosine deaminase
ERX552270.42.contig_13650
25,000
30,000
35,000
TATA-binding
Rep. factor subunit C
ERX552296.3.contig_62860
5,000
10,000
DNA photolyase
ERX555926.32.contig_26568
5,000
10,000
15,000
VLTF3 transcription factor
Major Capsid
ERX555967.81.contig_20121
10,000
15,000
20,000
Aconitase
Citrate Synthase
Isocitrate dehydrogenase
Isocitrate Lyase
Succ. dehydrogenase subunit C
Succ. dehydrogenase subunit B
Succ. dehydrogenase subunit A
Genes present in other giant viruses
Other genes
D.
Present
Absent
Glycolysis
Gluconeo-genesis
TCA Cycle
ENO
ALD
G3P+PGK
G3P
PFK
PGI
PGK
PGM
PYK
TPI
HK
FBP
PEPCK
PC
ACON
CS
ICD
ICL
SDA
SDB
SDC
SRX802202.77
SRX802073.62
ERX552257.96
SRX802073.68
ERX552243.92
SRX802077.162
ERX556043.42
ERX552252.36
ERX552253.3
ERX552255.59
ERX555965.10
ERX556028.57
ERX556094.51
SRX802077.79
ERX556017.61
ERX555941.31
ERX288975.8
ERX288973.1
ERX288947.1
ERX288965.16
ERX552256.7
ERX555978.20
ERX555945.3
ERX556015.14
ERX552302.42
ERX555967.81
ERX555926.27
ERX552252.49
ERX555907.48
ERX552296.31
ERX556034.5
ERX552302.25
ERX555987.28
ERX552302.17
ERX552254.13
ERX555926.32
ERX552270.56
ERX556009.10
ERX556031.42
ERX556127.29
ERX552302.18
ERX556106.20
ERX556037.19
ERX555967.26
ERX552254.5
ERX552296.3
ERX555907.72
ERX556045.12
ERX552255.7
ERX552270.42
ERX555951.28
ERX555961.47
ERX556134.17
ERX556060.39
ERX556054.16
ERX556034.16
ERX556005.37
ERX556005.35
ERX555990.11
ERX555961.43
ERX555951.15
ERX555917.28
ERX555913.32
ERX555907.87
ERX555907.45
ERX552302.45
ERX552277.41
ERX552298.2

Figure 5.


MM01: 19
MM07: 1
MM09: 26
MM11: 1
MM16: 4
0.5
Glyceraldehyde 3-phosphate dehydrogenase
0.5
MM01: 1
MM07: 1
MM11: 1
MM16: 3
MM01: 14
MM16: 3
Phosphoglycerate Kinase
MM01: 13
MM07: 1
MM09: 30
MM16: 2
LP04: 1
0.5
Succinate dehydrogenase Subunit C
MM01: 21
MM09: 7
0.5
Aconitase
Bacteria
Eukaryotes
Archaea
Bacteriophages*
Unclassified**
NCLDVs
LP01: 25
LP03: 2
LP01: 13
0.5
MM01: 57
MM09: 6
MM16: 4
LP04: 3
Chlorophyll a-b binding protein
MM01: 57
MM06: 1
MM09: 58
MM12: 6
MM13: 2
MM16: 6
LP01: 86
LP03: 3
0.5
Phosphate starvation-inducible (PhoH)
MM01: 31
MM05: 13
MM07: 1
MM09: 11
MM12: 5
EP05: 4
LP01: 12
LP03: 2
LP04: 4
0.5
Phosphate permease
MM13: 3
MM18: 1
EP01: 3
EP05: 1
LP04: 1
MM01: 120
MM05: 9
MM08: 2
MM09: 1
MM12: 2
0.5
MM07: 4
LP04: 2
Rhodopsin