PLOS BIOLOGY

METHODS AND RESOURCES

A phylogenomic framework for charting the diversity and evolution of giant viruses

Frank O. Aylward[1,2]*, **Mohammad Moniruzzaman**[1], **Anh D. Ha**[1], **Eugene V. Koonin**[3]

1 Department of Biological Sciences, Virginia Tech, Blacksburg, Virginia, United States of America, **2** Center for Emerging, Zoonotic, and Arthropod-borne Pathogens, Virginia Tech, Blacksburg, Virginia, United States of America, **3** National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, United States of America

* faylward@vt.edu

OPEN ACCESS

Citation: Aylward FO, Moniruzzaman M, Ha AD, Koonin EV (2021) A phylogenomic framework for charting the diversity and evolution of giant viruses. PLoS Biol 19(10): e3001430. https://doi.org/10.1371/journal.pbio.3001430

Academic Editor: Curtis Suttle, University of British Colombia, CANADA

Received: June 4, 2021

Accepted: September 29, 2021

Published: October 27, 2021

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pbio.3001430

Data Availability Statement: All data products described in this study are available on the Giant Virus Database: https://faylward.github.io/GVDB/. Reference trees of concatenated alignments can be

Abstract

Large DNA viruses of the phylum Nucleocytoviricota have recently emerged as important members of ecosystems around the globe that challenge traditional views of viral complexity. Numerous members of this phylum that cannot be classified within established families have recently been reported, and there is presently a strong need for a robust phylogenomic and taxonomic framework for these viruses. Here, we report a comprehensive phylogenomic analysis of the Nucleocytoviricota, present a set of giant virus orthologous groups (GVOGs) together with a benchmarked reference phylogeny, and delineate a hierarchical taxonomy within this phylum. We show that the majority of Nucleocytoviricota diversity can be partitioned into 6 orders, 32 families, and 344 genera, substantially expanding the number of currently recognized taxonomic ranks for these viruses. We integrate our results within a taxonomy that has been adopted for all viruses to establish a unifying framework for the study of Nucleocytoviricota diversity, evolution, and environmental distribution.

Main text

Large double-stranded DNA viruses of the phylum Nucleocytoviricota are a diverse group of viruses with virion sizes reaching up to 1.5 μm and genome sizes up to 2.5 Mb, comparable to many bacteria and archaea as well as picoeukaryotes [1–5]. The recognized taxonomic ranks in this phylum currently include 2 classes, 5 orders, 7 families, and 41 genera. The viruses in the families Asfarviridae, Ascoviridae, Iridoviridae, and Poxviridae infect metazoans, whereas those in the families Marseilleviridae, Mimiviridae, and Phycodnaviridae primarily infect algae or heterotrophic unicellular eukaryotes [6–8]. Members of the Nucleocytoviricota span an exceptionally broad range of genome sizes, from below 100 kbp to more than 2.5 Mbp. Several comparative genomic analyses have documented the highly complex, chimeric nature of their genomes in which numerous genes appear to have been acquired from diverse cellular lineages and other viruses [9–13]. These multiple, dynamic gene exchanges between viruses and their hosts [14–17] as well as the large phylogenetic breadth of this viral group [12,18,19] make the investigation of the evolution and taxonomic classification of the Nucleocytoviricota a challenging task. Despite these difficulties, early comparative genomic analyses studies succeeded in identifying a small set of core genes that could be reliably used to produce phylogenies that

found on the interactive Tree of Life: https://itol.embl.de/shared/faylward.

Funding: This work was supported by a Simons Early Career Award in Marine Microbial Ecology and Evolution to F.O.A, an the NSF (IIBR-1918271) award to F.O.A., and the Intramural Research Program of the National Institutes of Health (National Library of Medicine) for E.V.K. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist

Abbreviations: AaV, Aureococcus anophagefferens virus; ANI, average nucleotide identity; ASFV, African swine fever virus; A32, A32-like packaging ATPase; GVOG, giant virus orthologous group; HMM, Hidden Markov Model; IC, Internode Certainty; ICTV, International Committee on the Taxonomy of Viruses; MAG, metagenome-assembled genome; MCP, major capsid protein; OG, orthologous group; PolB, family B DNA Polymerase; RED, relative evolutionary distance; RNAPL, large RNA polymerase subunit; RNAPS, small RNA polymerase subunit; SFII, superfamily II helicase; TC, Tree Certainty; TetV, Tetraselmis virus; TFIIB, TFIIB transcriptional factor; TopoII, Topoisomerase family II; VLTF3, virus late transcription factor 3.

encompass the entire diversity of Nucleocytoviricota, leading to the conclusion that all these viruses share common evolutionary origins [18,20].

Recent studies have reported numerous new Nucleocytoviricota genomes, many of which seem to represent novel lineages with only distant phylogenetic affinity for previously identified taxa [10,16,21]. For example, many viruses that infect a variety of protist genera have been discovered that are related to Mimiviridae but do not fall within the same clade as the canonical *Acanthamoeba polyphaga mimivirus* [9,22,23]. Moreover, numerous metagenome-assembled genomes (MAGs) have been reported that also appear to form novel sister clades to the Mimiviridae, Asfarviridae, and other families [10,16,21]. Uncertainty in the phylogenetic relationships within the Nucleocytoviricota is a major impediment to the ongoing efforts that seek to characterize the diversity of these viruses in the environment, as well as studies aiming to better understand the evolutionary origins of unique traits within this viral phylum. As more studies begin to chart the environmental diversity of Nucleocytoviricota, defining taxonomic groupings that encompass equivalent phylogenetic breadths will be critical for the exploration of the geographic and temporal variability in viral diversity and for comparing results from different studies. Moreover, the evolutionary origins of large genomes, virion sizes, and complex metabolic repertoires in many Nucleocytoviricota are of great interest, and ancestral state reconstructions and the tracking of horizontal gene transfers fully depend on a robust phylogenetic framework.

Here, we present a phylogenomic framework for charting the diversity and evolution of Nucleocytoviricota. We first assess the strength of the phylogenetic signals from different marker genes that are found in a broad array of distantly related viruses and arrive at a set of 7 genes that performs well in our benchmarking of concatenated protein alignments. Using this hallmark gene set, we then perform a large-scale phylogenetic analysis and clade delineation of the Nucleocytoviricota to produce a hierarchical taxonomy. Our taxonomy includes the established families Poxviridae, Asfarviridae, Iridoviridae, Phycodnaviridae, Marseilleviridae, and Mimiviridae as well as 26 proposed new family-level clades and 1 proposed new order. Sixteen of the families are represented only by genomes derived from cultivation-independent approaches, underscoring the enormous diversity of these viruses in the environment that have not yet been isolated. We integrate these family-level classifications into the broader hierarchical taxonomy of all viruses that has recently been adopted (i.e., a "megataxonomy" [3]) to arrive at a unified and hierarchical classification scheme for the entire phylum Nucleocytoviricota.

Results

Phylogenetic benchmarking of marker genes

We first generated a dataset of protein families to identify phylogenetic marker genes that are broadly represented across Nucleocytoviricota. To this end, we selected a set of 1,380 quality-checked Nucleocytoviricota genomes that encompassed all established families (S1 Data; see Methods). By clustering the protein sequences encoded in these genomes, we then generated a set of 8,863 protein families, which we refer to as giant virus orthologous groups (GVOGs). We examined 25 GVOGs that were represented in >70% of all genomes and ultimately arrived at a set of 9 GVOGs that were potentially useful for phylogenetic analysis, which is largely consistent with the previous studies that have identified phylogenetic marker genes in Nucleocytoviricota [19,20,24] (Table 1, Figs A-Y in S1 Text, see Methods for details; descriptions of the 25 GVOGs provided in S2 Data). These GVOGs included 5 genes that we have previously used for phylogenetic analysis of Nucleocytoviricota: the family B DNA Polymerase (PolB), A32-like packaging ATPase (A32), virus late transcription factor 3 (VLTF3), superfamily II

Table 1. Broadly represented GVOGs used for phylogenetic benchmarking.

GVOG ID	Name	Annotation
GVOGm0003	MCP	NCLDV major capsid protein
GVOGm0013	SFII	DEAD/SNF2-like helicase
GVOGm0022	RNAPS	DNA-directed RNA polymerase beta subunit
GVOGm0023	RNAPL	DNA-directed RNA polymerase alpha subunit
GVOGm0054	PolB	DNA polymerase family B
GVOGm0172	TFIIB	Transcription initiation factor IIB
GVOGm0461	TopoII	DNA topoisomerase II
GVOGm0760	A32	Packaging ATPase
GVOGm0890	VLTF3	Poxvirus Late Transcription Factor VLTF3

https://doi.org/10.1371/journal.pbio.3001430.t001

helicase (SFII), and major capsid protein (MCP) [10]. In addition, this set included the large and small RNA polymerase subunits (RNAPL and RNAPS, respectively), the TFIIB transcriptional factor (TFIIB), and the Topoisomerase family II (TopoII).

We evaluated individual marker genes and concatenated marker sets using the Internode Certainty and Tree Certainty metrics (IC and TC, respectively), which provide a measure of the phylogenetic strength of each individual marker gene [25,26]. The TC values were highest for the RNAP subunits, PolB, and TopoII (Fig 1A), consistent with the view that, in most cases, longer genes carry a stronger phylogenetic signal, likely due to the larger number of phylogenetically informative characters. A similar observation has also been made for phylogenetic marker genes of bacteria and archaea [27]. The MCP marker had markedly lower TC values than PolB, TopoII, or either of the RNAP subunits; this is potentially because Nucleocytoviricota genomes often encode multiple copies of MCP, which complicates efforts to distinguish orthologs from paralogs (Fig 1A). This is especially true when using metagenome-derived genomes that are incomplete, because orthologous MCP copies may be missing even while paralogs are present. When this occurs, a paralogous MCP will have the best match to this protein family and will be included even if it has experienced distinct evolutionary pressures compared to the orthologous copy. SFII, TFIIB, A32, and VLTF3 showed lower TC values than the other 5 markers, but these were also the shortest marker genes and would not be expected to yield high quality phylogenies when used individually.

Next, we sought to identify which marker genes provide for the best phylogenetic inference when used together in a concatenated alignment. If markers produce incongruent phylogenetic signals, they will yield trees with low TC values when concatenated, even if the individual phylogenetic strength of the markers is high [26]. We evaluated 8 marker gene sets in total. We began by assessing the TC of the 5-gene set that we have previously used [10]. Surprisingly, the TC of this set was lower than that of some individual markers (TC of 0.865; Fig 1B), suggesting that some of the markers provide incongruent signals. We surmised that this was most likely due to the MCP, given that the presence of multiple copies of this protein in some Nucleocytoviricota may complicate efforts to identify the appropriate ortholog to use for tree construction (Fig 1A). As we suspected, removal of MCP increased the TC of the concatenated tree (from 0.865 to 0.875) (Fig 1B). The addition of the RNAPS, RNAPL, TFIIB, or TopoII markers to the 4-gene set increased the TC (Fig 1B), although a 7-gene marker set that excluded RNAPS performed best overall (TC of 0.898). The existence of RNAPS paralogs has been observed before [23], and it is likely that this is the cause of the lower TC value when using this marker. Overall, the 7-gene marker set represents an improvement over the initial 5-gene set, and we therefore used these genes for subsequent phylogenetic analysis and clade demarcation. Importantly, the benchmarking results we present here are specific to the genome set that we analyzed, and the

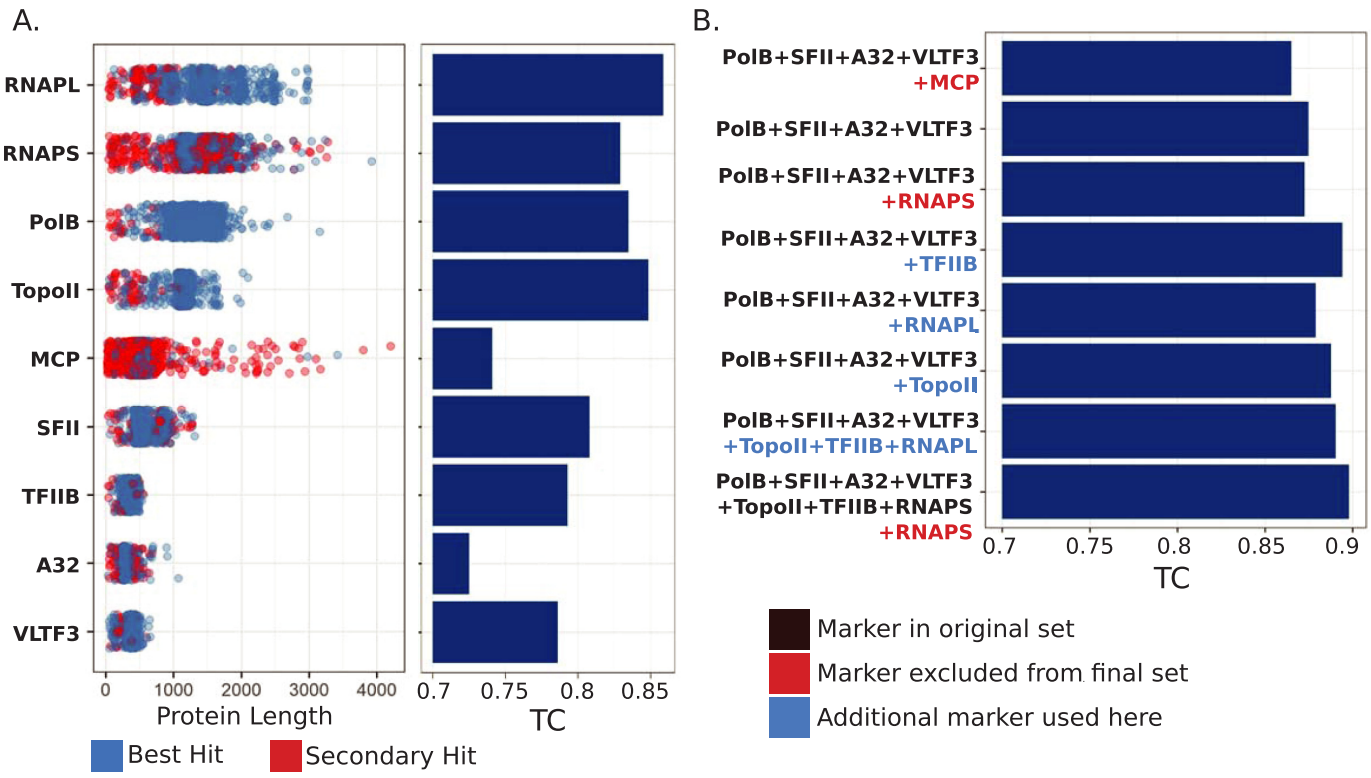

Fig 1. Benchmarking of phylogenetic marker genes for Nucleocytoviricota. (A) Dotplot of protein lengths for each of the 9 marker genes examined in detail. Blue dots represent proteins that were the best hit against marker gene HMMs and likely represent true orthologs, while red dots represent multiple copies of marker genes present in a genome. The TC scores of the markers are presented on the barplot on the right. (B) TC values for phylogenies made from concatenated alignments of different marker sets. Black text denotes markers we have used previously, red text denotes markers that we did not include in the final set, and blue text denotes additional markers used here compared to our original 5-gene set. Note that MCP was used in our original marker set but is excluded from the final 7-gene set. Protein lengths and TC values are provided in S2 Data. A32, A32-like packaging ATPase; HMM, Hidden Markov Model; MCP, major capsid protein; PolB, family B DNA Polymerase; RNAPL, large RNA polymerase subunit; RNAPS, small RNA polymerase subunit; SFII, superfamily II helicase; TC, Tree Certainty; TFIIB, TFIIB transcriptional factor; TopoII, Topoisomerase family II; VLTF3, virus late transcription factor 3.

https://doi.org/10.1371/journal.pbio.3001430.g001

use of MCP and RNAPS as phylogenetic markers may still be useful in other contexts. For example, when analyzing only complete genomes the presence of multiple paralogous copies of these genes may be less problematic.

A hierarchical taxonomy for Nucleocytoviricota

The best-quality phylogenetic tree produced with the 7-gene marker set could be broadly divided into 2 class-level and 6 order-level clades, 5 of which were consistent with the orders in the recently adopted megataxonomy of viruses (Fig 2) [3]. The Chitovirales and Asfuvirales orders, which respectively contain the Poxviridae and Asfarviridae, formed a distinct group with a long stem branch (class Pokkesviricetes) that we used to root the tree, consistent with previous studies [20,28]. The Pimascovirales, which includes Pithoviruses, Marseilleviruses, and Iridoviridae/Ascoviridae, also formed a highly supported monophyletic group. The current order Algavirales, which includes the Phycodnaviridae, Chloroviruses, Pandoraviruses, Molliviruses, Prasinoviruses, and Coccolithoviruses, was paraphyletic, and we split this order into 2 groups based on their placement in the phylogeny. In the proposed taxonomy, we retain the existing Algavirales name for the clade that contains the Chloroviruses and Prasinoviruses and additionally propose the order pandoravirales for the group that includes the

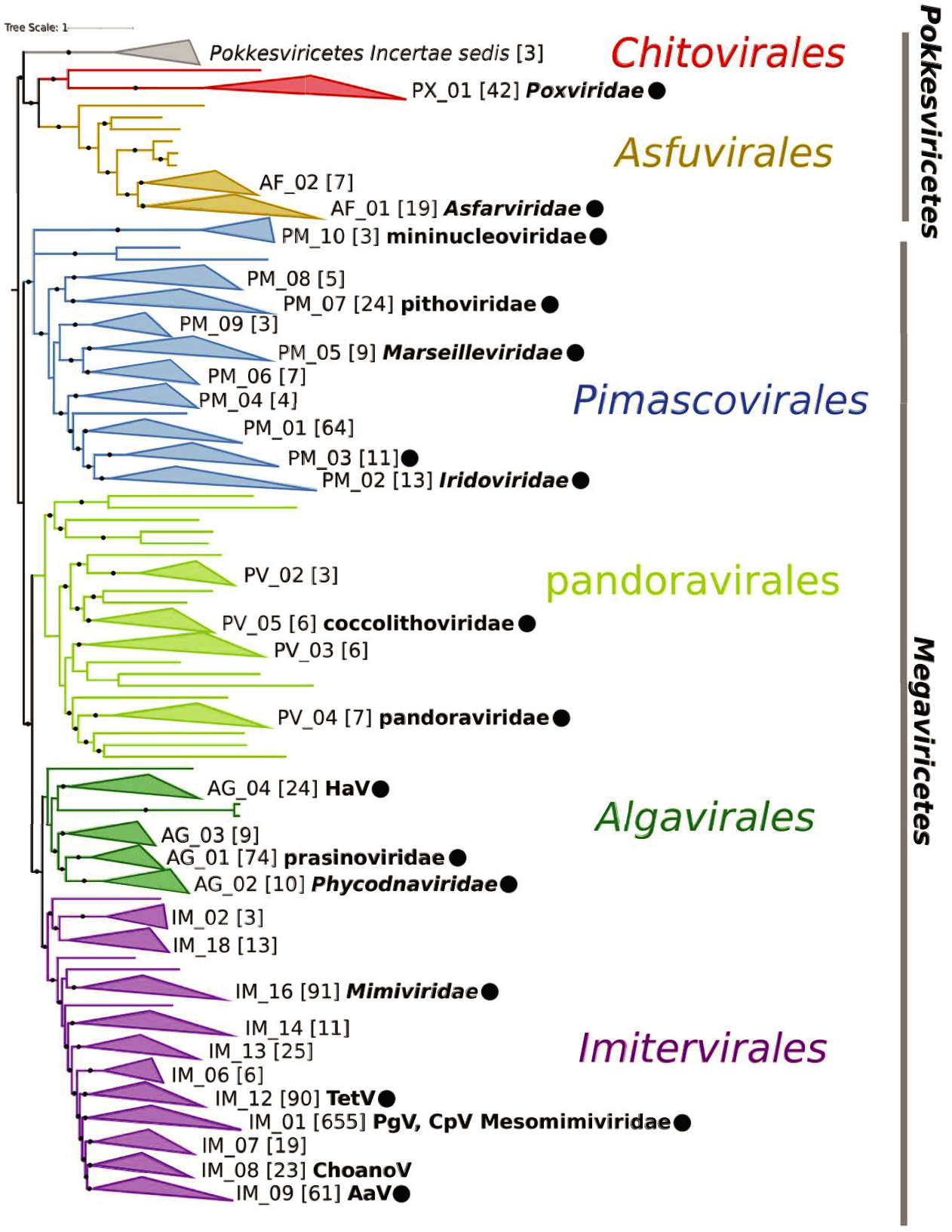
Tree Scale: 1
Pokkesviricetes Incertae sedis [3]
Chitovirales
PX_01 [42] Poxviridae
Asfuvirales
AF_02 [7]
AF_01 [19] Asfarviridae
PM_10 [3] mininucleoviridae
PM_08 [5]
PM_07 [24] pithoviridae
PM_09 [3]
PM_05 [9] Marseilleviridae
PM_06 [7]
PM_04 [4]
Pimascovirales
PM_01 [64]
PM_03 [11]
PM_02 [13] Iridoviridae
PV_02 [3]
pandoravirales
PV_05 [6] coccolithoviridae
PV_03 [6]
PV_04 [7] pandoraviridae
AG_04 [24] HaV
Algavirales
AG_03 [9]
AG_01 [74] prasinoviridae
AG_02 [10] Phycodnaviridae
IM_02 [3]
IM_18 [13]
IM_16 [91] Mimiviridae
IM_14 [11]
IM_13 [25]
Imitervirales
IM_06 [6]
IM_12 [90] TetV
IM_01 [655] PgV, CpV Mesomimiviridae
IM_07 [19]
IM_08 [23] ChoanoV
IM_09 [61] AaV
Pokkesviricetes
Megaviricetes

Fig 2. Phylogeny of Nucleocytoviricota based on the 7-gene marker gene set that had the highest TC value of those tested. The phylogeny was inferred using the LG+I+F+G4 model in IQ-TREE. Solid circles denote IC values >0.5. Families are denoted by collapsed clades, with their nonredundant identifier provided at their right. The number of genomes in each clade is provided in brackets. Established family names are provided in bold italics, and proposed names are provided in lowercase. The presence of notable cultivated viruses is provided in bold next to some clades. Aav, *Aureococcus anophagefferens* virus; ChoanoV1, Choanoflagellate virus; CpV, *Chrysochromulina parva* virus; HaV, *Heterosigma akashiwo* virus; IC, Internode Certainty; PgV, *Phaeocystis globosa* virus; TC, Tree Certainty; TetV, *Tetraselmis* virus.

https://doi.org/10.1371/journal.pbio.3001430.g002

Pandoraviruses and Coccolithoviruses. The Imitervirales, which contain the Mimiviridae, formed a sister group to the Algavirales.

From our reference tree, we delineated taxonomic levels using the relative evolutionary distance (RED) of each clade as a guide, using an approach similar to the one recently employed for bacteria and archaea [29]. RED values vary between 0 and 1, with lower values denoting phylogenetically broad groups that branch closer to the root and higher values denoting phylogenetically shallow groups that branch closer to the leaves. The RED of the Nucleocytoviricota classes ranges from 0.017 to 0.032, whereas the values for the orders range from 0.158 to 0.240 (Fig 3A and S3 Data). We delineated family- and genus-level clades so that they had nonoverlapping RED values that were higher than their next-highest taxonomic rank (Fig 3A). This approach yielded clades that were consistent with families and genera currently recognized by the International Committee on Taxonomy of Viruses (ICTV; [30]), such as the *Chlorovirus*, *Prasinovirus*, and Mimiviridae (see below; full classification information in S1 Data). To ensure that putative families were not defined by spurious placement of individual genomes, we accepted only groups with $\geq$3 members and left other genomes in the tree as singletons with incertae sedis as the family identifier. This approach yielded a total of 32 families, not including 22 singleton genomes that potentially represent additional families and are listed as incertae sedis here. We provided tentative genus-level identifiers for all genomes, leading to 344 total genera (Figs 2 and 3). Of these, 213 genera contain only a single representative, and additional merging or splitting of these groups may be necessary as more genomes become available and fine-scale phylogenetic patterns are clarified.

Of the 32 families, 6 correspond to the families currently recognized by the ICTV, for which we retained the existing nomenclature (Asfarviridae, Poxviridae, Marseilleviridae, Iridoviridae, Phycodnaviridae, and Mimiviridae). The Ascoviridae are included within the Iridoviridae, and so we use the latter family name here. In addition, we propose 6 family names here: "prasinoviridae," which include the prasinoviruses, "pandoraviridae," which include the Pandoraviruses and *Mollivirus sibericum*, "coccolithoviridae," which include the coccolithoviruses, "pithoviridae," which include Pithoviruses, Cedratviruses, and Orpheoviruses, "mesomimiviridae," which includes several haptophyte viruses previously defined as "extended Mimiviridae," and "mininucleoviridae," which has previously been described and includes several viruses of Crustacea [31]. Some of these family names have been used previously, such as pandoraviridae and mininucleoviridae, but so far have not been formally recognized by the ICTV. For other proposed families, we provide nonredundant identifiers corresponding to their order, and we anticipate that future studies will provide information for selecting appropriate family names once more is learned on the host ranges and molecular traits of these viruses. Two of the families contained only a single cultivated representative (AG_04 and IM_09), whereas 16 families included none.

Notably, the Imitervirales contain 11 families, as well as 4 singleton viruses that potentially represent additional family-level clades. This underscores the vast diversity of the large viruses in this group, which is consistent with the results of several studies reporting an enormous diversity of Mimiviridae-like viruses in the biosphere, in particular in aquatic environments [10,32–34]. Other studies have suggested additional nomenclature to refer to these

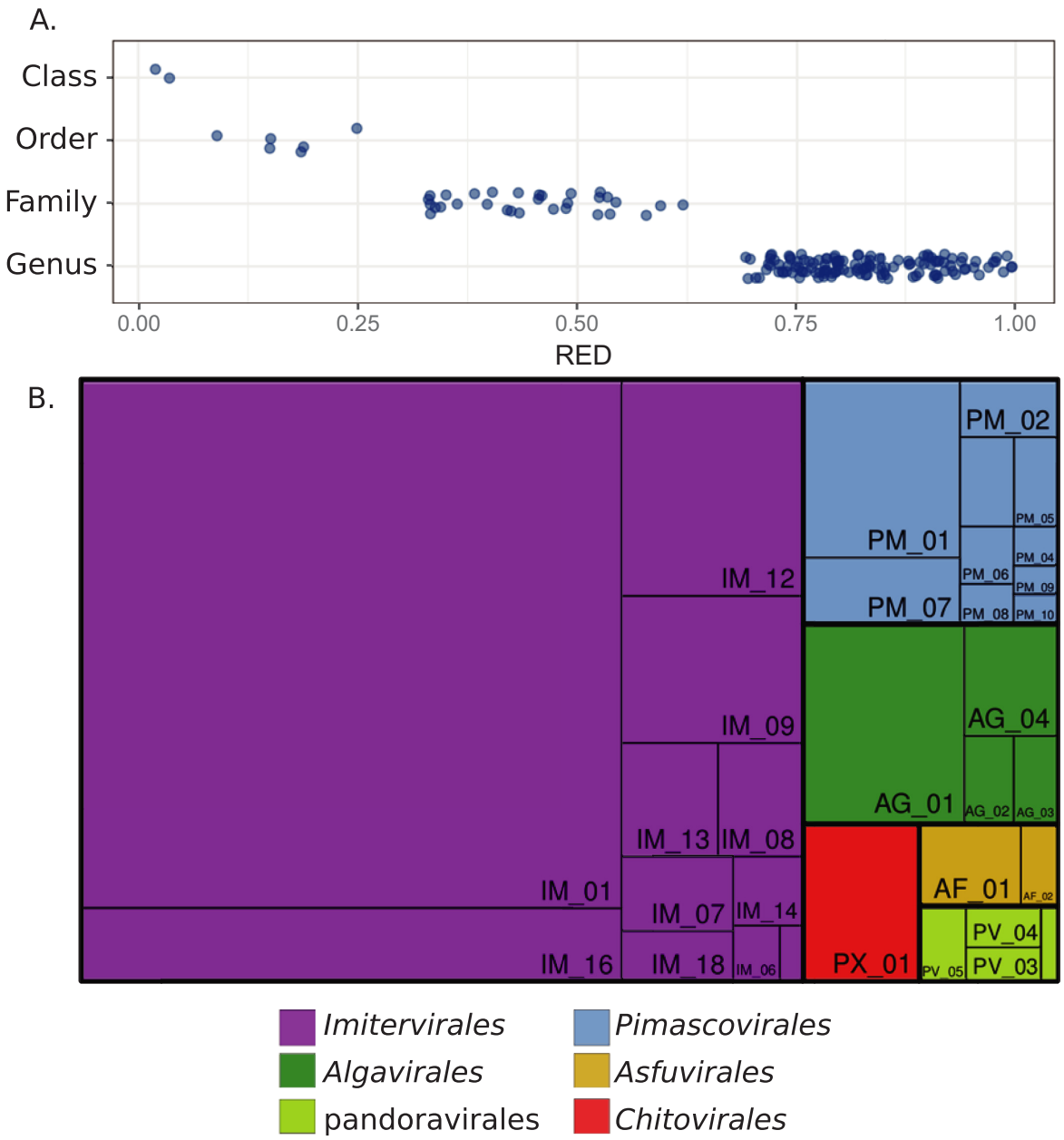

Fig 3. Summary of the Nucleocytoviricota taxonomy. (A) RED values for Nucleocytoviricota classes, orders, and families, and genera. (B) Treemap diagram of the Nucleocytoviricota in which orders and families are shown. The area of each rectangle is proportional to the number of genomes in the respective taxon. RED values can be found in S3 Data. RED, relative evolutionary divergence.

https://doi.org/10.1371/journal.pbio.3001430.g003

Mimiviridae-like viruses, such as the "extended Mimiviridae" and the subfamilies Mesomimivirinae, or Megamimivirinae, but our results suggest that an extensive array of new families is warranted within Imitervirales, given the broad genomic and phylogenetic diversity within this group. Several of the proposed new families contain representatives that have recently been described; IM_12 contains the *Tetraselmis* virus (TetV), which encodes several fermentation genes [11], IM_09 contains *Aureococcus anophagefferens* virus (AaV), which is thought to play an important role in brown tide termination [35], and IM_08 contains a virus of

Choanoflagellates [36] (Fig 2). Family IM_01 contains cultivated viruses that infect haptophytes of the genera *Chrysochromulina* and *Phaeocystis*, which were previously proposed to be classified in the subfamily mesomimivirinae [23]. We propose the name mesomimiviridae to denote the family-level status of this lineage, while still retaining reference to this original name. Notably, the Mesomimiviridae includes by far the largest total number of genomic representatives in our analysis ($n = 655$, including 652 MAGs; Figs 2 and 3B), the vast majority of which are derived from aquatic environments (Fig Z in S1 Text), suggesting that members of this family are important components of global freshwater and marine ecosystems. Within the Mimiviridae, we recovered 3 clades that correspond to previously proposed subfamilies. One of these clades contains Klosneuviruses and corresponds to the proposed subfamily Klosneuvirinae [37]; this subfamily also includes *Bodo saltans* virus as well as several genomes recovered from forest soils [38,39]. The second clade corresponds to the subfamily Megamimivirinae and includes *A. polyphaga mimivirus*, Tupanviruses, and *Megavirus chilensis*, among others [40–42]. Lastly, we recovered a clade that includes *Cafeteria roenbergensis* virus [9], several "PacV" viruses obtained from flow sorting and sequencing of marine samples [43], and a variety of MAGs.

All families within the Imitervirales except one included members with genome sizes >500 kbp, highlighting the "giant" genomes that are characteristic of this lineage (Fig 4A). Genes involved in translation, including tRNA synthetases and translation initiation factors, were consistently highly represented in the Imitervirales, showing that the rich complement of these genes that has been described for the Mimiviridae is broadly characteristic of other families in this order (Fig 4B) [40,42]. Throughout the Imitervirales genes involved in glycolysis and the TCA cycle, cytoskeleton components such as viral-encoded actin, myosin, and kinesin proteins, and nutrient transporters including those that target ammonia and phosphate were also common (Fig 4B) [10,44–46], underscoring the complex functional repertoires of this virus order.

The Algavirales is a sister lineage to the Imitervirales that contains 4 families encompassing several well-studied algal viruses. The Prasinoviridae (AG_01) is a family that includes viruses known to infect the prasinophyte genera *Bathycoccus*, *Micromonas*, and *Ostreococcus* [8], and cultivation-independent surveys have provided evidence that the MAGs in this clade are also associated with prasinophytes [46]. Similarly, our approach yielded a well-defined Phycodnaviridae family (AG_02) composed mostly of chloroviruses, consistent with the similar host range of these viruses [47]. All 4 families of the Algavirales have smaller genome sizes compared to the Imitervirales (Fig 4A), but there were still several similarities in their encoded functional repertoires. As noted previously [10,17,36], genes involved in light sensing, including rhodopsins and chlorophyll-binding proteins, were common across the Imitervirales and Algavirales, perhaps because many of the viruses are found in sunlit aquatic environments where manipulation of host light sensing during infection is advantageous. Moreover, genes involved in nutrient transport, translation, and even some components of glycolysis and the TCA cycle were found in the Algavirales, consistent with the complex repertoires of metabolic genes that have been reported for some of these viruses despite their relatively small genome sizes [48,49].

The pandoravirales, a new order we propose here, consists of 4 families, including the pandoraviridae and the coccolithoviridae. The pandoraviridae (PV_04) include *Mollivirus sibericum* as well as the pandoraviruses, which possess the largest viral genomes known [50]. Grouping of these viruses together in the same family is consistent with previous studies that have shown that *M. sibericum* and the Pandoraviruses have shared ancestry [51,52], and comparative genomic analysis that have shown that they all encode a unique duplication in the glycosyl hydrolase that has been co-opted as a major virion protein in the Pandoraviruses [53].

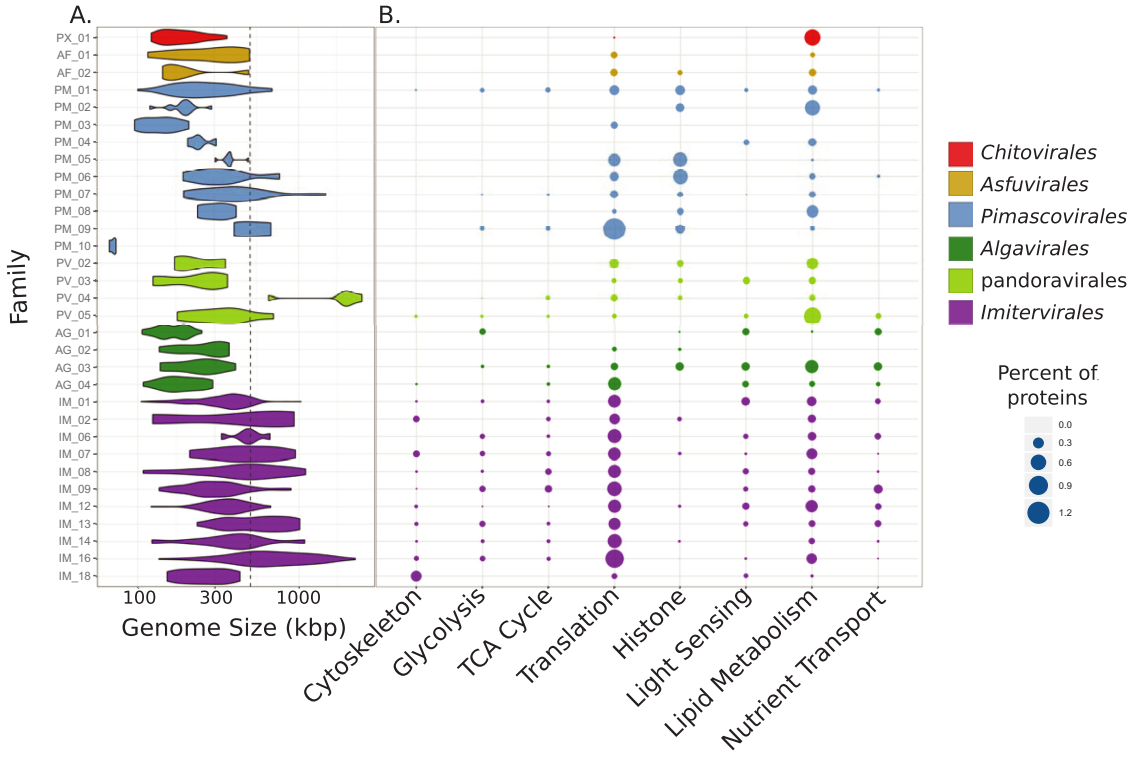

Fig 4. Genomic characteristics of the Nucleocytoviricota. (A) Violin plot showing the genome size distribution across the Nucleocytoviricota families. The dashed gray line denotes 500 kbp. (B) Bubble plot showing the percent of total proteins in each family that could be assigned to GVOGs that belonged to particular functional categories (details in S2 Data). GVOG, giant virus orthologous group.

https://doi.org/10.1371/journal.pbio.3001430.g004

The coccolithoviridae (PV_05) is mostly comprised of viruses that infect the marine coccolithophore *Emiliania huxleyi*; although much smaller than the genomes of the Pandoraviruses, genomes of cultivated representatives of this family exceed 400 kbp and encode diverse functional repertoires including sphingolipid biosynthesis genes [54].

Although most orders contained primarily genomes that could be readily grouped into families, the pandoravirales also included 15 singleton genomes out of the 37 total. This is potentially due to the lack of adequate genome sampling in this group, which would result in many distinct lineages represented by only individual genomes. If this is the case, more well-defined families will become evident as additional genomes are sequenced. Alternatively, the lack of clearly defined families could result from longer branches in this group that obfuscate the clustering of well-defined groups. The Medusavirus, which is included in this order, encodes a divergent PolB marker gene that is likely the result of gene transfer with a eukaryotic homolog [55]. Frequent gene transfers among phylogenetic marker genes might be another explanation for the presence of many long branches in the pandoravirales clade.

The Pimascovirales encompass 10 families including the Iridoviridae (PM_02), Marseilleviridae (PM_05), and Pithoviridae (PM_07) and notably includes both *Pithovirus sibericum*, which has the largest viral capsid currently known (1.5 μm [56]), as well as crustacean viruses in the family Mininucleoviridae (PM_10), which possess the smallest genomes recorded for any Nucleocytoviricota (67 to 71 kbp [31]). The Mininucleoviridae have highly degraded genomes that lack several phylogenetic marker genes. Although they can be classified within

the Pimascovirales with high confidence, their relationship to other families is uncertain, and we therefore placed them in a polytomous node at the base of this order (Fig 2). The uncharacterized family PM_01 contains the largest number of genomes ($n = 64$) within this order, all of which are MAGs. The majority of these MAGs were derived from aquatic metagenomes, and some have been recovered in marine metatranscriptomes [46], suggesting that they play an important but currently unknown role in marine systems. Overall, the repertoires of encoded proteins in the Pimascovirales were notably different from the Imitervirales, pandoravirales, and Algavirales; while cytoskeleton components, nutrient transporters, light sensing genes, and central carbon metabolism components were prevalent in the latter 3 families, they were largely absent in the Pimascovirales (Fig 4B). Conversely, histone components appeared to be more prevalent in the latter order; indeed, the histones encoded in marseilleviruses have recently become a model for understanding their structure and interactions with viral DNA [57,58]. Genes involved in translation and lipid metabolism were present in the Pimascovirales in addition to most other orders.

In addition to the families that fall within the established orders and families, we also identified several lineages or individual genomes that may represent novel taxonomic ranks (Fig 2). One of these groups consists of 3 genomes that is basal-branching to the Pokkesviricetes class, which we refer to as Pokkesviricetes incertae sedis (Fig 2). The basal-branching placement of this group suggests that it might comprise a new class that is a sister group to the Pokkesviricetes. The placement of this lineage remains tentative, however, and to clarify evolutionary relationships within the Nucleocytoviricota further phylogenetic work with additional genomes will be necessary both for this lineage as well as other putative novel taxa that are represented by individual genomes.

Discussion

Although only 6 families of Nucleocytoviricota have been established to date, recent cultivation-independent studies have revealed a vast diversity of these viruses in the environment, and their classification, together with cultivated representatives, has remained challenging. Here, we present a unified taxonomic framework based on a benchmarked set of phylogenetic marker genes that establishes a hierarchical taxonomy of Nucleocytoviricota. This taxonomy encompasses 6 orders and 32 families, including 1 order and 26 families we propose here. Remarkably, the Imitervirales contain 11 families, including the Mimiviridae, underscoring the vast diversity of large viruses within this order. This framework substantially increases the total number of Nucleocytoviricota families, and we expect that the number will continue to increase as new genomes are incorporated. In particular, we identified 22 singleton genomes that likely represent additional families, the status of which will be clarified as more genomes become available.

We anticipate that the phylogenetic and taxonomic framework we develop here will be a useful community resource for several future lines of inquiry into the biology of Nucleocytoviricota. Firstly, the GVOGs are a large set of viral protein families constructed using many recently produced Nucleocytoviricota MAGs, and they will likely be useful for the genome annotation and the examination of trends in gene content across viral groups. Secondly, the reference phylogeny we present will facilitate work that delves into ancestral Nucleocytoviricota lineages, examines the timing and nature of gene acquisitions, and classifies newly discovered viruses. For example, giant viral genomes (>500 kbp) evolved independently in multiple orders, and future studies that examine the similarities and differences in these genome expansion events will be important for pinpointing the driving forces of viral gigantism. Lastly, analysis of the environmental distribution of different taxonomic ranks of Nucleocytoviricota

across Earth's biomes will be an important direction for future work that reveals prominent biogeographic patterns and helps to clarify the ecological impact of these viruses.

Methods

Nucleocytoviricota genome set

We compiled a set of Nucleocytoviricota genomes that included MAGs as well as genomes of cultured isolates. For this, we first downloaded all MAGs available from several recent studies [10,16,21]. We also included all Nucleocytoviricota genomes available in NCBI RefSeq as of June 1, 2020. Lastly, we also included several Nucleocytoviricota genomes from select publications that were not yet available in NCBI, such as the cPacV, ChoanoV, *Pyramimonas orientalis* virus O1B (MT663543), and AbALV viruses that have recently been described [15,36,43,59]. After compiling this set, we dereplicated the genomes, since the presence of highly similar or identical genomes is not necessary for broad-scale phylogenetic inference. For dereplication, we compared all genomes against each other using MASH v. 2.0 [60] ("mash dist" parameters -k 16 and -s 300), and clustered genomes together using a single-linkage clustering, with all genomes with a MASH distance of $\leq$0.05 linked together. The MASH distance of 0.05 was chosen since it has been roughly found to correspond to an average nucleotide identity (ANI) of 95% [60]; although gene flow can occur over a broad range of genome identity values [61], this is still a useful threshold for genome dereplication. From each cluster, we chose the genome with the highest N50 contig length as the representative. We then decontaminated the genomes through analysis with ViralRecall v.2.0 [62] (-c parameter), with all contigs with negative scores removed on the grounds that they represent non-Nucleocytoviricota contamination or highly unusual gene composition that cannot be validated by our present knowledge of Nucleocytoviricota genomic content. We only considered contigs >10 kbp, given the inherent difficulty in eliminating contamination derived from short contigs. To ensure that we only used genomes that could be placed in a phylogeny, we then screened the genome set and retained only those with a PolB marker and 3 of the 4 markers A32, SFII, VLTF3, and MCP, consistent with our previous methodology [10]. After this, we arrived at a set of 1,380 genomes, including 1,253 MAGs and 127 complete genomes of cultivated viruses.

GVOG construction

To construct GVOGs, we first predicted proteins from all genomes using Prodigal v. 2.6.2. Proteins that did not have a recognizable start or stop codon at the ends of contigs were removed on the grounds that they may represent fragmented genes and obfuscate orthologous group (OG) predictions. We then calculated OGs using Proteinortho v. 6.06 [63] (parameters -e = 1e-5—identity = 25 -p = blastp+—selfblast—cov = 50 -sim = 0.80). We constructed Hidden Markov Models (HMMs) from proteins by aligning them with Clustal Omega v1.2.3 [64] (default parameters), trimming the alignment with trimAl v1.4.rev15 [65] (parameters -gt 0.1), and generating the HMM from the trimmed alignment with hmmbuild in HMMER v3.3 [66]. The goal of this analysis was to identify broad-level protein families, and we therefore sought to merge HMMs that bore similarity to each other and therefore derived from related protein families. For this, we then compared the proteins in each OG to the HMM of every other OG (hmmsearch -E 1e-20—domtblout option, hits retained only if 30% of the query protein aligned to the HMM). In cases where >50% of the proteins in one OG also had hits to the HMM of another OG, and vice versa, we then merged all of the proteins together and constructed a new merged HMM from the full set of proteins. The final set contained 8,863 HMMs, and we refer to these as the GVOGs. To provide annotations for GVOGs, we compared all of the proteins in each GVOG to the EggNOG 5.0 [67], Pfam [68], and NCVOG

databases [69] (hmmsearch, -E 1e-3). For NCVOGs, we obtained protein sequences from the original NCVOG study and generated HMMs using the same methods we used for GVOGs. Annotations were assigned to a GVOG if >50% of the proteins used to make a GVOG had hits to the same HMM in one of these databases. Details regarding all GVOGs and their annotations can be found in S2 Data.

Benchmarking phylogenetic marker genes for Nucleocytoviricota

To identify phylogenetic markers for Nucleocytoviricota, we cataloged GVOGs that were broadly represented in the 1,380 viral genomes that we used for benchmarking. We searched all proteins encoded in the genomes against the GVOG HMMs using hmmsearch (e-value cutoff 1e-10) and identified a set of 25 GVOGs that were found in >70% of the genomes in our set (hmmsearch, -E 1e-5). We constructed individual phylogenetic trees of these protein families to assess their individual evolutionary histories. For individual phylogenetic trees, we calibrated bit score cutoffs so that poorly matching proteins would not be included. These cutoffs were generally equivalent to the fifth percentile score of all of the best protein matches for each genome. We then examined several features of these trees. Firstly, we only considered GVOGs present in all established families that would therefore be useful as universal or nearly universal phylogenetic markers. Secondly, we examined each tree individually to assess the degree to which taxa from different orders clustered together in distinct monophyletic groups, which was taken as a signature of HGT. High levels of gene transfer would produce topologies incongruent with other marker genes and therefore compromise the reliability of a given marker when used on a concatenated alignment. For individual marker gene trees, we aligned proteins from each GVOG using Clustal Omega, trimmed the alignment using trimAl (-gt 0.1 option), and constructed the phylogeny using IQ-TREE with ultrafast bootstraps calculated (-m TEST, -bb 1000, -wbt options).

We arrived at a set of 9 GVOGs that met the criteria described above and could potentially serve as robust phylogenetic markers (Table 1). We evaluated the phylogenetic strength of these markers individually using the recently developed TC and IC metrics. These metrics are an alternative to the traditional bootstrap because they take into account the frequency of contrasting bipartitions and can therefore be viewed as a measure of the phylogenetic strength of a gene [25,26]. We generated alignments using Clustal Omega, trimmed with TrimAl, and generated trees with IQ-TREE v1.6.9 [70] with ultrafast bootstraps [71] (parameters -wbt -bb 1000 -m LG+I+G4). We calculated TC and IC values in RaxML v8.2.12 (-f i option, ultrafast bootstraps used with the -z flag) [72]. We also evaluated the TC and IC values of trees generated from concatenated alignments. To construct concatenated alignments, we used the python program "ncldv_markersearch.py" that we developed for this purpose: https://github.com/faylward/ncldv_markersearch.

For the final tree used for clade demarcation, we ran IQ-TREE 5 times using the parameters "-m LG+F+I+G4 -bb 1000 -wbt," and we chose the resulting tree with the highest TC value for subsequent clade demarcation and RED calculation. Three genomes in the Mininucleoviridae family were included in the final tree but were not used for the benchmarking analysis because they have been shown to have highly degraded genomes that are not necessarily representative of Nucleocytoviricota more broadly [31]. Moreover, the MAG ERX555967.47 was found to have highly variable placement in different orders in different trees we analyzed, and we therefore did not include this genome in the final tree on the grounds that it represented a rogue taxa that may reduce overall tree quality [73]. We rooted the final tree between the Pokkesviricetes and Megaviricetes, consistent with previous studies [6,28]. We placed the 3 genomes of Pokkesviricetes incertae sedis adjacent to the Pokkesvirictes clade due to the clustering of

several GVOGs of this group with members of the Pokkesvirictes (SFII: Fig C in S1 Text, PolB: Fig I in S1 Text).

Family delineation and nomenclature

We calculated RED values in R using the get_reds function in the package "castor" [74]. As input, we used a rooted tree derived from the 7-gene marker set described above. For the Poxviridae, Asfarviridae, Iridoviridae, Phycodnaviridae, Marseilleviridae, mininucleoviridae, and Mimiviridae, we retained existing nomenclature, and clades assigned these names based on the initially characterized viruses that were assigned to these families. For example, the Phycodnaviridae was assigned to AG_02 because the chloroviruses within this clade were the first-described members of this family, while the prasinoviruses were assigned to a new family, although they are commonly referred to as Phycodnaviridae. Similarly, Mimiviridae was assigned based on the placement of *A. polyphaga mimivirus*, Iridoviridae was assigned based on the placement of *Invertebrate iridescent virus 6*, Asfarviridae was assigned to the clade containing African swine fever virus (ASFV), and Marseilleviridae was assigned to the clade containing the marseilleviruses. The treemap visualization was generated using the R package "treemap."

Supporting information

S1 Text. Supporting figures. Fig A. Major Capsid Protein GVOGm0003 phylogeny. Fig B. Disulfide (thiol) oxidoreductase GVOGm0004 phylogeny. Fig C. Superfamily II helicase GVOGm0013 phylogeny. Fig D. Patatin phospholipase GVOGm0018 phylogeny. Fig E. DEAD/SNF2-like helicase GVOGm0020 phylogeny. Fig F. DNA-directed RNA polymerase subunit beta (RNAPS) GVOGm0022 phylogeny. Fig G. DNA-directed RNA polymerase subunit alpha (RNAPL) GVOGm0023 phylogeny. Fig H. mRNA capping enzyme GVOGm0036 phylogeny. Fig I. DNA polymerase family B GVOGm0054 phylogeny. Fig J. TATA box binding protein (TBP) GVOGm0056 phylogeny. Fig K. Ribonucleoside diphosphate reductase, alpha subunit GVOGm0088 phylogeny. Fig L. D5-like helicase-primase GVOGm0095 phylogeny. Fig M. Uncharacterized, C-terminal domain GVOGm0115 phylogeny. Fig N. Uncharacterized protein GVOGm0152 phylogeny. Fig O. Transcription initiation factor IIB GVOGm0172 phylogeny. Fig P. RuvC, Holliday junction resolvases (HJRs) GVOGm0189 phylogeny. Fig Q. Ubiquitin carboxyl-terminal hydrolase GVOGm0214 phylogeny. Fig R. Proliferating cell nuclear antigen GVOGm0239 phylogeny. Fig S. DNA topoisomerase II GVOGm0461 phylogeny. Fig T. Divergent DNA-directed RNA polymerase subunit 5 GVOGm0694 phylogeny. Fig U. Packaging ATPase GVOGm0760 phylogeny. Fig V. Metallopeptidase WLM GVOGm0787 phylogeny. Fig W. Ribonuclease III GVOGm0798 phylogeny. Fig X. Virus Late Transcription Factor 3 VLTF3 GVOGm0890 phylogeny. Fig Y. Ribonucleotide reductase small subunit GVOGm1574 phylogeny. Fig Z. Barchart of source habitats for the Nucleocytoviricota families. Full information is provided in S1 Data.
(PDF)

S1 Data. Taxonomy, genome statistics, and other metadata for the Nucleocytoviricota genomes analyzed in this study.
(XLSX)

S2 Data. Statistics and descriptions of the 25 GVOGs present in 70% of the genomes analyzed. Full annotations of all GVOGs are also provided, and TC values for the trees of this study.
(XLSX)

S3 Data. RED values for taxonomic ranks presented in this study.
(XLSX)

Acknowledgments

We acknowledge the use of the Virginia Tech Advanced Research Computing Center for bioinformatic analyses performed in this study.

Author Contributions

Conceptualization: Frank O. Aylward.

Data curation: Frank O. Aylward, Anh D. Ha.

Formal analysis: Frank O. Aylward, Mohammad Moniruzzaman, Anh D. Ha.

Funding acquisition: Frank O. Aylward.

Investigation: Frank O. Aylward, Mohammad Moniruzzaman.

Methodology: Frank O. Aylward, Mohammad Moniruzzaman, Anh D. Ha.

Project administration: Frank O. Aylward.

Resources: Frank O. Aylward, Eugene V. Koonin.

Software: Frank O. Aylward.

Supervision: Frank O. Aylward, Eugene V. Koonin.

Validation: Frank O. Aylward.

Visualization: Frank O. Aylward.

Writing – original draft: Frank O. Aylward, Eugene V. Koonin.

Writing – review & editing: Frank O. Aylward, Eugene V. Koonin.

References

1. Fischer MG. Giant viruses come of age. Curr Opin Microbiol. 2016; 31:50–7. https://doi.org/10.1016/j.mib.2016.03.001 PMID: 26999382
2. Koonin EV, Yutin N. Origin and evolution of eukaryotic large nucleo-cytoplasmic DNA viruses. Intervirology. 2010; 53:284–92. https://doi.org/10.1159/000312913 PMID: 20551680
3. Koonin EV, Dolja VV, Krupovic M, Varsani A, Wolf YI, Yutin N, et al. Global Organization and Proposed Megataxonomy of the Virus World. Microbiol Mol Biol Rev. 2020; 84. https://doi.org/10.1128/MMBR.00061-19 PMID: 32132243
4. Raoult D, Forterre P. Redefining viruses: lessons from Mimivirus. Nat Rev Microbiol. 2008; 6:315–9. https://doi.org/10.1038/nrmicro1858 PMID: 18311164
5. Wilhelm S, Bird J, Bonifer K, Calfee B, Chen T, Coy S, et al. A Student's Guide to Giant Viruses Infecting Small Eukaryotes: From Acanthamoeba to Zooxanthellae. Viruses. 2017:46. https://doi.org/10.3390/v9030046 PMID: 28304329
6. Koonin EV, Yutin N. Evolution of the Large Nucleocytoplasmic DNA Viruses of Eukaryotes and Convergent Origins of Viral Gigantism. Adv Virus Res. 2019; 103:167–202. https://doi.org/10.1016/bs.aivir.2018.09.002 PMID: 30635076
7. Karki S, Moniruzzaman M, Aylward FO. Comparative Genomics and Environmental Distribution of Large dsDNA Viruses in the Family Asfarviridae. Front Microbiol. 2021; 12. https://doi.org/10.3389/fmicb.2021.657471 PMID: 33790885
8. Weynberg KD, Allen MJ, Wilson WH. Marine Prasinoviruses and Their Tiny Plankton Hosts: A Review. Viruses. 2017;9. https://doi.org/10.3390/v9030043 PMID: 28294997

9. Fischer MG, Allen MJ, Wilson WH, Suttle CA. Giant virus with a remarkable complement of genes infects marine zooplankton. Proc Natl Acad Sci U S A. 2010; 107:19508–13. https://doi.org/10.1073/pnas.1007615107 PMID: 20974979

10. Moniruzzaman M, Martinez-Gutierrez CA, Weinheimer AR, Aylward FO. Dynamic genome evolution and complex virocell metabolism of globally-distributed giant viruses. Nat Commun. 2020; 11:1710. https://doi.org/10.1038/s41467-020-15507-2 PMID: 32249765

11. Schvarcz CR, Steward GF. A giant virus infecting green algae encodes key fermentation genes. Virology. 2018; 518:423–33. https://doi.org/10.1016/j.virol.2018.03.010 PMID: 29649682

12. Boyer M, Yutin N, Pagnier I, Barrassi L, Fournous G, Espinosa L, et al. Giant Marseillevirus highlights the role of amoebae as a melting pot in emergence of chimeric microorganisms. Proc Natl Acad Sci U S A. 2009; 106:21848–53. https://doi.org/10.1073/pnas.0911354106 PMID: 20007369

13. Monier A, Pagarete A, de Vargas C, Allen MJ, Read B, Claverie J-M, et al. Horizontal gene transfer of an entire metabolic pathway between a eukaryotic alga and its DNA virus. Genome Res. 2009; 19:1441–9. https://doi.org/10.1101/gr.091686.109 PMID: 19451591

14. Moniruzzaman M, Weinheimer AR, Martinez-Gutierrez CA, Aylward FO. Widespread endogenization of giant viruses shapes genomes of green algae. Nature. 2020; 588:141–5. https://doi.org/10.1038/s41586-020-2924-2 PMID: 33208937

15. Rozenberg A, Oppermann J, Wietek J, Fernandez Lahore RG, Sandaa R-A, Bratbak G, et al. Lateral Gene Transfer of Anion-Conducting Channelrhodopsins between Green Algae and Giant Viruses. Curr Biol. 2020; 30:4910–4920.e5. https://doi.org/10.1016/j.cub.2020.09.056 PMID: 33065010

16. Schulz F, Roux S, Paez-Espino D, Jungbluth S, Walsh DA, Denef VJ, et al. Giant virus diversity and host interactions through global metagenomics. Nature. 2020; 578:432–6. https://doi.org/10.1038/s41586-020-1957-x PMID: 31968354

17. Yutin N, Koonin EV. Proteorhodopsin genes in giant viruses. Biol Direct. 2012; 7:34. https://doi.org/10.1186/1745-6150-7-34 PMID: 23036091

18. Iyer LM, Aravind L, Koonin EV. Common origin of four diverse families of large eukaryotic DNA viruses. J Virol. 2001; 75:11720–34. https://doi.org/10.1128/JVI.75.23.11720-11734.2001 PMID: 11689653

19. Yutin N, Koonin EV. Hidden evolutionary complexity of Nucleo-Cytoplasmic Large DNA viruses of eukaryotes. Virol J. 2012; 9:161. https://doi.org/10.1186/1743-422X-9-161 PMID: 22891861

20. Iyer LM, Balaji S, Koonin EV, Aravind L. Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. Virus Res. 2006; 117:156–84. https://doi.org/10.1016/j.virusres.2006.01.009 PMID: 16494962

21. Bäckström D, Yutin N, Jørgensen SL, Dharamshi J, Homa F, Zaremba-Niedwiedzka K, et al. Virus Genomes from Deep Sea Sediments Expand the Ocean Megavirome and Support Independent Origins of Viral Gigantism. mBio. 2019;10. https://doi.org/10.1128/mBio.02497-18 PMID: 30837339

22. Santini S, Jeudy S, Bartoli J, Poirot O, Lescot M, Abergel C, et al. Genome of Phaeocystis globosa virus PgV-16T highlights the common ancestry of the largest known DNA viruses infecting eukaryotes. Proc Natl Acad Sci U S A. 2013; 110:10800–5. https://doi.org/10.1073/pnas.1303251110 PMID: 23754393

23. 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. 2017;91. https://doi.org/10.1128/JVI.00230-17 PMID: 28446675

24. Guglielmini J, Woo AC, 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; 116:19585–92. https://doi.org/10.1073/pnas.1912006116 PMID: 31506349

25. Salichos L, Stamatakis A, Rokas A. Novel information theory-based measures for quantifying incongruence among phylogenetic trees. Mol Biol Evol. 2014; 31:1261–71. https://doi.org/10.1093/molbev/msu061 PMID: 24509691

26. Salichos L, Rokas A. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature. 2013; 497:327–31. https://doi.org/10.1038/nature12130 PMID: 23657258

27. Martinez-Gutierrez CA, Aylward FO. Phylogenetic Signal, Congruence, and Uncertainty across Bacteria and Archaea. Mol Biol Evol. 2021. https://doi.org/10.1093/molbev/msab254 PMID: 34436605

28. Koonin EV, Yutin N. Multiple evolutionary origins of giant viruses. F1000Res. 2018:7. https://doi.org/10.12688/f1000research.13350.2 PMID: 29527296

29. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol. 2018; 36:996–1004. https://doi.org/10.1038/nbt.4229 PMID: 30148503

30. Lefkowitz EJ, Dempsey DM, Hendrickson RC, Orton RJ, Siddell SG, Smith DB. Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV). Nucleic Acids Res. 2018: D708–17. https://doi.org/10.1093/nar/gkx932 PMID: 29040670

31. Subramaniam K, Behringer DC, Bojko J, Yutin N, Clark AS, Bateman KS, et al. A New Family of DNA Viruses Causing Disease in Crustaceans from Diverse Aquatic Biomes. mBio. 2020; 11. https://doi.org/10.1128/mBio.02938-19 PMID: 31937645

32. Mihara T, Koyano H, Hingamp P, Grimsley N, Goto S, Ogata H. Taxon Richness of "Megaviridae" Exceeds those of Bacteria and Archaea in the Ocean. Microbes Environ. 2018; 33:162–71. https://doi.org/10.1264/jsme2.ME17203 PMID: 29806626

33. Monier A, Larsen JB, Sandaa R-A, Bratbak G, Claverie J-M, Ogata H. Marine mimivirus relatives are probably large algal viruses. Virol J. 2008; 5:12. https://doi.org/10.1186/1743-422X-5-12 PMID: 18215256

34. Ghedin E, Claverie J-M. Mimivirus relatives in the Sargasso sea. Virol J. 2005; 2:62. https://doi.org/10.1186/1743-422X-2-62 PMID: 16105173

35. Moniruzzaman M, LeCleir GR, Brown CM, Gobler CJ, Bidle KD, Wilson WH, et al. Genome of brown tide virus (AaV), the little giant of the Megaviridae, elucidates NCLDV genome expansion and host-virus coevolution. Virology. 2014; 466–467:60–70. https://doi.org/10.1016/j.virol.2014.06.031 PMID: 25035289

36. Needham DM, Yoshizawa S, Hosaka T, Poirier C, Choi CJ, Hehenberger E, et al. A distinct lineage of giant viruses brings a rhodopsin photosystem to unicellular marine predators. Proc Natl Acad Sci. 2019:20574–83. https://doi.org/10.1073/pnas.1907517116 PMID: 31548428

37. Schulz F, Yutin N, Ivanova NN, Ortega DR, Lee TK, Vierheilig J, et al. Giant viruses with an expanded complement of translation system components. Science. 2017; 356:82–5. https://doi.org/10.1126/science.aal4657 PMID: 28386012

38. Deeg CM, Chow C-ET, Suttle CA. The kinetoplastid-infecting Bodo saltans virus (BsV), a window into the most abundant giant viruses in the sea. Elife. 2018;7. https://doi.org/10.7554/eLife.33014 PMID: 29582753

39. Schulz F, Alteio L, Goudeau D, Ryan EM, Yu FB, Malmstrom RR, et al. Hidden diversity of soil giant viruses. Nat Commun. 2018; 9:4881. https://doi.org/10.1038/s41467-018-07335-2 PMID: 30451857

40. Abrahão J, Silva L, Silva LS, Khalil JYB, Rodrigues R, Arantes T, et al. Tailed giant Tupanvirus possesses the most complete translational apparatus of the known virosphere. Nat Commun. 2018; 9:749. https://doi.org/10.1038/s41467-018-03168-1 PMID: 29487281

41. 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. 2011; 108:17486–91. https://doi.org/10.1073/pnas.1110889108 PMID: 21987820

42. Raoult D, Audic S, Robert C, Abergel C, Renesto P, Ogata H, et al. The 1.2-megabase genome sequence of Mimivirus. Science. 2004; 306:1344–50. https://doi.org/10.1126/science.1101485 PMID: 15486256

43. Needham DM, Poirier C, Hehenberger E, Jiménez V, Swalwell JE, Santoro AE, 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. 2019; 374:20190086. https://doi.org/10.1098/rstb.2019.0086 PMID: 31587639

44. Kijima S, Delmont TO, Miyazaki U, Gaia M, Endo H, Ogata H. Discovery of Viral Myosin Genes With Complex Evolutionary History Within Plankton. Front Microbiol. 2021; 12:683294. https://doi.org/10.3389/fmicb.2021.683294 PMID: 34163457

45. Cunha VD, Da Cunha V, Gaia M, Ogata H, Jaillon O, Delmont TO, et al. Giant viruses encode novel types of actins possibly related to the origin of eukaryotic actin: the viractins. bioRxiv. https://doi.org/10.1101/2020.06.16.150565

46. Ha AD, Moniruzzaman M, Aylward FO. High Transcriptional Activity and Diverse Functional Repertoires of Hundreds of Giant Viruses in a Coastal Marine System. mSystems. 2021; 6:e0029321.

47. Van Etten JL, Agarkova IV, Dunigan DD. Chloroviruses. Viruses. 2019; 12:20.

48. Moreau H, Piganeau G, Desdevises Y, Cooke R, Derelle E, Grimsley N. Marine prasinovirus genomes show low evolutionary divergence and acquisition of protein metabolism genes by horizontal gene transfer. J Virol. 2010; 84:12555–63. https://doi.org/10.1128/JVI.01123-10 PMID: 20861243

49. Weynberg KD, Allen MJ, Gilg IC, Scanlan DJ, Wilson WH. Genome sequence of Ostreococcus tauri virus OtV-2 throws light on the role of picoeukaryote niche separation in the ocean. J Virol. 2011; 85:4520–9. https://doi.org/10.1128/JVI.02131-10 PMID: 21289127

50. Philippe N, Legendre M, Doutre G, Couté Y, Poirot O, Lescot M, et al. Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. Science. 2013; 341:281–6. https://doi.org/10.1126/science.1239181 PMID: 23869018

51. Yutin N, Koonin EV. Pandoraviruses are highly derived phycodnaviruses. Biol Direct. 2013; 8:25. https://doi.org/10.1186/1745-6150-8-25 PMID: 24148757

52. Legendre M, Lartigue A, Bertaux L, Jeudy S, Bartoli J, Lescot 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. 2015; 112: E5327–35. https://doi.org/10.1073/pnas.1510795112 PMID: 26351664

53. Krupovic M, Yutin N, Koonin E. Evolution of a major virion protein of the giant pandoraviruses from an inactivated bacterial glycoside hydrolase. Virus Evol. 2020; 6. https://doi.org/10.1093/ve/veaa059 PMID: 33686356

54. Wilson WH, Schroeder DC, Allen MJ, Holden MTG, Parkhill J, Barrell BG, et al. Complete genome sequence and lytic phase transcription profile of a Coccolithovirus. Science. 2005; 309:1090–2. https://doi.org/10.1126/science.1113109 PMID: 16099989

55. Yoshikawa G, Blanc-Mathieu R, Song C, Kayama Y, Mochizuki T, Murata K, et al. Medusavirus, a Novel Large DNA Virus Discovered from Hot Spring Water. J Virol. 2019;93. https://doi.org/10.1128/JVI.02130-18 PMID: 30728258

56. Legendre M, Bartoli J, Shmakova L, Jeudy S, Labadie K, Adrait A, et al. Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses with a pandoravirus morphology. Proc Natl Acad Sci U S A. 2014; 111:4274–9. https://doi.org/10.1073/pnas.1320670111 PMID: 24591590

57. Liu Y, Bisio H, Toner CM, Jeudy S, Philippe N, Zhou K, et al. Virus-encoded histone doublets are essential and form nucleosome-like structures. Cell. 2021; 184:4237–4250.e19. https://doi.org/10.1016/j.cell.2021.06.032 PMID: 34297924

58. Valencia-Sánchez MI, Abini-Agbomson S, Wang M, Lee R, Vasilyev N, Zhang J, et al. The structure of a virus-encoded nucleosome. Nat Struct Mol Biol. 2021; 28:413–7. https://doi.org/10.1038/s41594-021-00585-7 PMID: 33927388

59. Matsuyama T, Takano T, Nishiki I, Fujiwara A, Kiryu I, Inada M, et al. A novel Asfarvirus-like virus identified as a potential cause of mass mortality of abalone. Sci Rep. 2020; 10:4620. https://doi.org/10.1038/s41598-020-61492-3 PMID: 32165658

60. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 2016; 17:132. https://doi.org/10.1186/s13059-016-0997-x PMID: 27323842

61. Bobay L-M, Ochman H. Biological species in the viral world. Proc Natl Acad Sci. 2018:6040–5. https://doi.org/10.1073/pnas.1717593115 PMID: 29784828

62. Aylward FO, Moniruzzaman M. ViralRecall-A Flexible Command-Line Tool for the Detection of Giant Virus Signatures in 'Omic Data. Viruses. 2021; 13. https://doi.org/10.3390/v13020150 PMID: 33498458

63. Lechner M, Findeiss S, Steiner L, Marz M, Stadler PF, Prohaska SJ. Proteinortho: detection of (co-) orthologs in large-scale analysis. BMC Bioinformatics. 2011; 12:124. https://doi.org/10.1186/1471-2105-12-124 PMID: 21526987

64. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011; 7:539. https://doi.org/10.1038/msb.2011.75 PMID: 21988835

65. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009; 25:1972–3. https://doi.org/10.1093/bioinformatics/btp348 PMID: 19505945

66. Eddy SR. Accelerated Profile HMM Searches. PLoS Comput Biol. 2011; 7:e1002195. https://doi.org/10.1371/journal.pcbi.1002195 PMID: 22039361

67. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. 2019; 47:D309–14. https://doi.org/10.1093/nar/gky1085 PMID: 30418610

68. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: The protein families database in 2021. Nucleic Acids Res. 2021; 49:D412–9. https://doi.org/10.1093/nar/gkaa913 PMID: 33125078

69. Yutin N, Wolf YI, Raoult D, Koonin EV. Eukaryotic large nucleo-cytoplasmic DNA viruses: clusters of orthologous genes and reconstruction of viral genome evolution. Virol J. 2009; 6:223. https://doi.org/10.1186/1743-422X-6-223 PMID: 20017929

70. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015; 32:268–74. https://doi.org/10.1093/molbev/msu300 PMID: 25371430

71. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol Biol Evol. 2018; 35:518–22. https://doi.org/10.1093/molbev/msx281 PMID: 29077904

72. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014; 30:1312–3. https://doi.org/10.1093/bioinformatics/btu033 PMID: 24451623

73. Aberer AJ, Krompass D, Stamatakis A. Pruning Rogue Taxa Improves Phylogenetic Accuracy: An Efficient Algorithm and Webservice. Syst Biol. 2013:162–6. https://doi.org/10.1093/sysbio/sys078 PMID: 22962004

74. Louca S, Doebeli M. Efficient comparative phylogenetics on large trees. Bioinformatics. 2018; 34:1053–5. https://doi.org/10.1093/bioinformatics/btx701 PMID: 29091997