Identification of branching order within the kingdom *Bamfordvirae*

Armen Kotsinyan[1], Harutyun Sahakyan[2,3], and Hovakim Zakaryan[1,4]

[1]Laboratory of Antiviral Drugs Discovery, Institute of Molecular Biology of NAS, Yerevan, Armenia

[2]Laboratory of Computational Modelling of Biological Processes, Institute of Molecular Biology of NAS, Yerevan, Armenia

[3] National Center for Biotechnology Information, National Library of Medicine, National Institutes for Health, Bethesda MD, USA

[4]Advanced Solutions Center, Foundation for Armenian Science and Technology, 0033, Hakobyan 3, Yerevan, Armenia

Corresponding author: Armen Kotsinyan, armenrkotsinyan@gmail.com

Key words: *Varidnaviria*; *Bamfordvirae*; *Nucleocytoviricota*; NCLDV; *Preplasmiviricota*, viral taxonomy

Abstract

The kingdom *Bamfordvirae* comprises the majority of the realm *Varidnaviria* and, according to the 2021 release of Virus Taxonomy by the International Committee on Taxonomy of Viruses, consists of the phyla *Nucleocytoviricota* and *Preplasmiviricota*. There are a number of fundamental unresolved issues related to the evolution of *Bamfordvirae*. These are questions concerning *Bamfordvirae* taxonomy include the branching order of *Nucleocytoviricota* and the question of the monophyly of *Preplasmiviricota*. Here, based on the analysis of individual core proteins phylogenies, supertree,

concatenated trees, dendrograms, as well as superdendrogram, we refine the branching order of major groups within phylum *Nucleocytoviricota* using the rooting of the entire phylum on the cellular outgroups. This effort results in several major changes in *Bamfordvirae* phylogeny. In particular, we show that *Nucleocytoviricota* consists of two sister clades, consisting of *Phycodnaviridae* sensu lato on the one hand and *Mimiviridae* sensu lato, *Iridoviridae*/*Ascoviridae*, *Marseilleviridae*, pithoviruses including *Cedratvirus*, *Solumvirus*, *Solivirus*, and *Orpheovirus*, mininucleoviruses, *Asfarviridae* sensu lato, and *Poxviridae* on the other hand. According to our data, *Asfarviridae* sensu lato and *Poxviridae* have likely originated from within the class *Megaviricetes*. We give evidence for polyphyly of the phylum *Preplasmiviricota* and argue for a transfer of the families *Lavidaviridae*, *Adintoviridae* and *Adenoviridae* from the phylum *Preplasmiviricota* in to the phylum *Nucleocytoviricota*. We also argue for the origin of the *Nucleocytoviricota* from small prokaryotic viruses and give arguments against the origin of *Nucleocytoviricota* from the *Adintoviridae*/Polinton-like viruses. Thus, our work, by clarifying the phylogenetic relationships within the *Bamfordvirae* kingdom, provides a solid basis for studying the evolutionary relationships of this kingdom with a vast diversity of supposedly related viruses outside of *Bamfordvirae*, including *Helvetiavirae* (*Varidnaviria*).

Introduction

The phylum *Nucleocytoviricota*, also known as nucleocytoplasmic large DNA viruses (NCLDV) comprises a group of large dsDNA viruses with presumably monophyletic origin [1-2]. Despite the official acceptance of this phylum in the ICTV Taxonomy, there are a number of unresolved problems associated with the origin and evolution of *Nucleocytoviricota*.

Two competing hypotheses have been proposed for the origin of NCLDV. According to first NCLDV originated from a specific group of bacteriophages [2-3]. Alternatively, NCLDV originated from

Polintons-like viruses, which, in turn, evolved from a tectivirus-like ancestor [4- 5]. In both cases, the NCLDV ancestor is believed to carried only a few dozen genes, and then the ancestral genome have rapidly expanded and diversified through gene duplications and horizontal gene transfer (HGT). Genomes of *Bamfordvirae* encode the FtsK-HerA superfamily ATPases, as well as double jelly roll major capsid proteins (DJR MCP) [5-6]. Moreover, the genomes of *Nucleocytoviricota* and the genomes of the part of *Preplasmiviricota*, in addition to these two genes, also encode the Family B DNA polymerase (PolB) and D5-like helicase-primase genes [5]. Excluding the DJR MCP genes, the rest of the mentioned genes have cellular homologues. Therefore, the construction of phylogenetic trees of proteins shared by *Bamfordvirae* using cellular homologues could tip the scales in favor of one of the hypotheses.

NCLDV share 20 conserved (core) orthologous genes [7]. Some of core genes also have the cellular homologues, but they have been used only in a handful of studies [7-9]. When cellular homologues are included, trees for different genes show conflicting topologies of branching order of NCLDV [8]. Perhaps this is a consequence of not identified and not eliminated homoplasies, leading to topological artifacts. The “rooted” trees sometimes presented in the literature tend to be rooted conditionally, on the basis of assumptions about the branching order of clades under study, but not with help of cellular outgroups from which core genes were transferred [9-11]. As far as we know there are no studies, investigating the influence of taxonomic composition of outgroups on topology of NCLDV core proteins. In other words, the branching order of *Nucleocytoviricota* is not determined and remains a subject of discussion.

Among the NCLDV core genes, there are genes encoding PolB, DJR MCP and FtsK-HerA superfamily ATPases. Genomes of several recently discovered viruses, such as *Pleurochrysis sp. endemic virus unk*, *Bovine rumen MELD virus*, *Sheep rumen MELD virus*, *Pleurochrysis sp. endemic virus 1a*, *Pleurochrysis sp. endemic virus 1b*, *Pleurochrysis sp. endemic virus 2* and highly divergent

Yaravirus brasiliensis appear to encode some or all these three genes, and this has been seen as a compelling reason to include these viruses in *Bamfordvirae* [12-13].

Another group of replicators apparently related to *Bamfordvirae* are mobile genetic elements (MGE) consisting of Polintons (also called Mavericks), cytoplasmic linear plasmids (CLP), and mitochondrial linear plasmids (MLP) [14-17]. The common genes composition, phylogenies of a number of proteins of NCLDV, *Lavidaviridae*, Polintons, CLPs, and MLPs have been examined in order to identify the phylogenetic relationships between these viruses and MGE [4-5, 18-19]. The phylogeny of *Bamfordvirae* and MGEs, however, has been studied in isolation from the problem of the branching order of the *Nucleocytoviricota*, without full use of cellular orthologues, as well as without investigation of impact of homoplasies and long branch attraction (LBA) effects on the resulting trees topology.

Phylogenetic relationships between individual homologous genes of NCLDV and their homologues from dsDNA viruses outside *Varidnaviria* also remain unclear. For instance, the possible relationships of DNA polymerases of NCLDV and *Herpesviridae* has been shown [20-22]. However, the cited works contain a number of unresolved fragments of phylogenetic relationships between PolB of *Eukaryota*, NCLDV and a number of dsDNA viruses.

According to the International Committee on Taxonomy of Viruses (ICTV) release from 2021, poxviruses and asfarviruses are combined into the class *Pokkesviricetes*, and the rest of *Nucleocytoviricota* form the class *Megaviricetes* (https://ictv.global/taxonomy/). Giving the equal taxonomic rank of *Pokkesviricetes* and *Megaviricetes* implies that the first bifurcation of the last common ancestor (LCA) of *Nucleocytoviricota* led to the emergence of two lineages leading to the LCA of *Pokkesviricetes* and to the LCA of *Megaviricetes*. The logic of dividing *Nucleocytoviricota* into the classes *Megaviricetes* and *Pokkesviricetes* is based on the greater sequence similarity within

Megaviricetes in comparison with *Pokkesviricetes*, as well as on the topologies of previously constructed phylogenetic trees [8, 10]. However, the lengths of the *Pokkesviricetes* branches of some core proteins may be associated with the accelerated evolution of this branch, and the topologies of the published trees, in our opinion, do not imply the division of *Nucleocytoviricota* into the classes *Megaviricetes* and *Pokkesviricetes* [8]. Similarly, the phylogenetic relationship between the phyla *Nucleocytoviricota* and *Preplasmiviricota* remains controversial [11]. Thus, the position of the root of the *Nucleocytoviricota* is logically justified, but it is not confidently determined and requires additional research.

Here we investigate all the above-mentioned problems related to the evolution of *Bamfordvirae*, and propose their solutions. Our main novel result is that based on the analysis of 15 individual core proteins phylogenies, supertree, 7 concatenated trees, 13 dendrograms, as well as superdendrogram, we propose a branching order of major groups within phylum *Nucleocytoviricota* based on the use of cellular, as well as viral (the NCLDV D6/D11 helicases) homologues as outgroups. Based on proposed branching order of *Nucleocytoviricota*, we propose for inclusion in the class *Megaviricetes* of the families *Asfarviridae* s.l. and *Poxviridae*. We give evidence for polyphyly of the phylum *Preplasmiviricota* and argue for a transfer of the families *Lavidaviridae*, *Adintoviridae* and *Adenoviridae* in to the phylum *Nucleocytoviricota*.

Results

Branching order of NCLDV. Phylogenies of 13 proteins support the primary split between phycodnaviruses s.l. and the rest of NCLDV

Replication module, transcription and RNA processing proteins

Eukaryotic genomes encode four homologues of NCLDV PolB: polymerases α (main component of the primase complex), δ (leading and lagging strand synthesis), ε (main replicative polymerase), and ζ (involved in translesion synthesis) [23-27]. Eukaryotic nuclear genomes also encode three, and plant genomes encode five homologues of each of the two largest subunits of NCLDV RNAP [28-34]. PolB and two RNAP subunits are nearly-universal markers of NCLDV, which can be used for establishing the phylogenetic relationships within the group, including the branching order between major clades of NCLDV. For the latter purpose, we need a rooted tree, preferably the one that is rooted on outgroup. PolB and the largest two subunits of RNAP are universally present in *Bacteria*, *Archaea* and *Eukarya* and, as a whole, are very likely to constitute outgroup to the NCLDV homologues [2, 8]. It has been shown, however, that the outgroup composition may affect the topological relationships between ingroup clades [35-38]. In order to account for the possibility of topological violations between NCLDV and outgroups, we first examined the phylogenetic relationships of the PolB and RNAP cellular orthologues.

Hereinafter, all PolB alignments refer to a fragment including the Exonuclease, N-terminal, and DNA polymerase domains, starting from the active site motif Exo I and including the MotifC. The topology of the cellular PolB tree confirmed the close relationships of the eukaryotic δ, ζ and archaeal PolB3 polymerases, as well as the sister relationship between the eukaryotic pol α clade and the joint δ + ζ + PolB3 clade (S1 Fig, S1 Align) [22, 39-40]. The second PolB clade includes archaeal PolB1, PolB2 and eukaryotic pol ε sequences and corresponds to the Pol-B1 + Pol-B2 + pol ε + *Candidatus Heimdallarchaeota* archaeon LC_3 (HeimC3_00430) clade [40].

On the tree of the RNAP α subunit, we observe sister relationships between the RPA1 and RPB1 clades (S2 Fig, S2 Align). In contrast, the RNAP β subunit tree shows the sister relationships between the RPA2 and RPC2 subunits (S3 Fig, S3 Align), which is consistent with previously published results [41].

The inclusion of the PolB, RNAP α and β subunits of NCLDV in the analysis showed the sister relationships of two clades within NCLDV. The first clade consists of phycodnaviruses s.l., and the second clade consists of all the remaining NCLDVs (Fig 1, S4 and 5 Figs, S4 and 6 Aligns). In order to clarify the topological position of PolBs of asfarviruses and poxviruses (the AP clade), we aligned the cellular and viral polymerases then added the PolB sequences of the AP clade to the alignment. We proceeded from the assumption that if PolBs of the AP clade are of non-NCLDV origin, they will be localized outside of NCLDV. However, the result obtained showed their origin from within the NCLDV, in contrast to the previously published data [22].

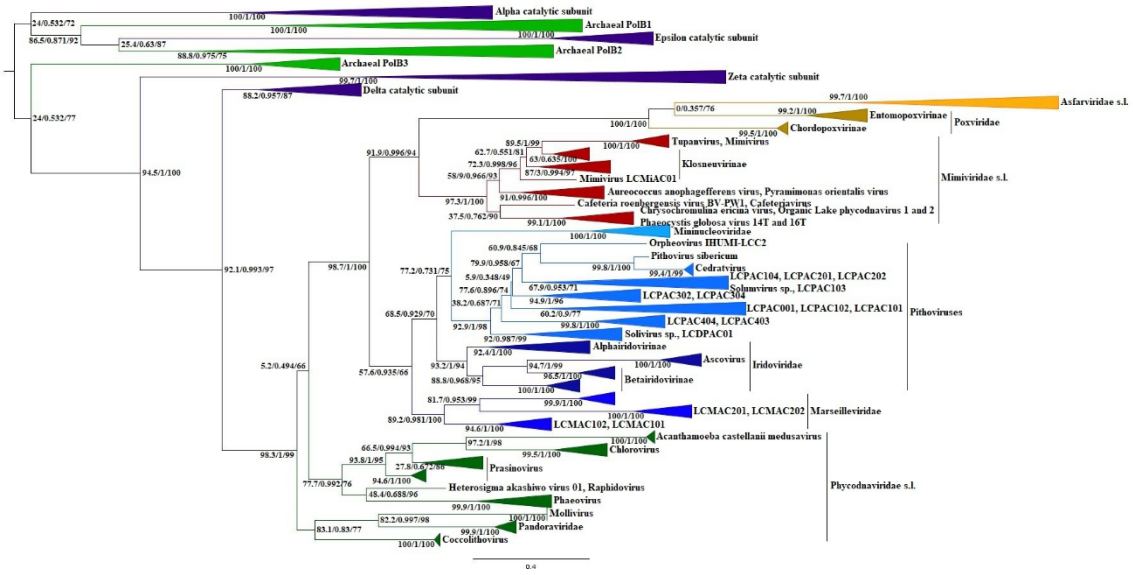


Fig 1. Maximum likelihood phylogenetic tree of archaeal (PolB1, PolB2 and PolB3), eukaryotic (α, δ, ε and ζ catalytic subunits), and NCLDV Family B DNA polymerases. The tree was constructed using the IQ TREE program (version 1.6.12) under the LG+F+R7 model and rooted according to the topology of the S1 Fig, as well as in published data [22, 40]. The corresponding alignment was built using the MUSCLE algorithm. Statistical support of the partitions, measured by SH-aLRT branch tests (1,000 replicates), approximate Bayes and ultrafast bootstrap (1,000

replicates), and is shown at the nodes. Scale bars represent the number of amino acid substitutions per site.

The branches corresponding to PolB, RNAP α and RNAP β of NCLDV are long and potentially may contain many character states uniting the NCLDV clades with outgroup clades (homoplasies), especially with the sister clades to NCLDV clades. This can lead to a false topological position of NCLDV among the cellular clades, as well as to false topological relationships of viral clades within NCLDV. In order to minimize the possibility of the influence of homoplasies, we removed the δ catalytic subunit, RPB1, and RPB2 clades from the corresponding trees and reconstructed the trees (S6-8 Figs, S7-9 Aligns). Reconstruction showed the invariability of topological relationships among the NCLDV clades, as well as among NCLDV and the cellular clades, which suggests that homoplasies were unlikely.

Changing the composition of outgroup is a reliable way to check the influence of the outgroup composition on the topology of ingroup [42]. We built the alignments of the PolB δ catalytic subunit, as well as RNAP α and β subunits based on the MUSCLE and MAFFT algorithms, in which we used only the sequences of NCLDV and eukaryotic sister clades (S9-14 Figs, S10-15 Aligns).

The complete removal of outgroups for the PolB, RNAP α and β subunits did not lead to a change in the topological relationships of the clades within NCLDV (S15-17 Figs, S16-18 Aligns).

The invariance of the topological positions of the PolB, RNAP α and β subunits of NCLDV, as well as the relationships of the clades within NCLDV, irrespective of changes in the composition of the outgroup, suggests that regardless of both possible potentialities i.e. i) capture of the specialized genes from the ancestral lineage leading to the modern eukaryotes, ii) capture of the specialized genes from now extinct protoeukaryotic lineages the identified branching order of NCLDV remains unchanged

and, therefore, may reflect the true evolutionary trajectory of NCLDV. Additional confidence in the results obtained is provided by a comparison of earlier results with ours, which confirm the close relationship between the eukaryotic δ catalytic subunits and NCLDV polymerases [20-22].

In contrast to other NCLDVs, in phycodnaviruses s.l., helicase-primase (SFIII helicase - D5-like primase) is encoded by two different ORFs. For instance, in *Paramecium bursaria Chlorella virus NY2A* (NC_009898), the SFIII helicase is located in the region 260150…262114 nucleotides, while the D5-like primase is located in the region 265781...267112 nucleotides. This situation does not exclude the possibility of different evolutionary trajectories of primase and helicase of phycodnaviruses s.l., of which only one can coincide with the helicase or primase trajectory of the rest of NCLDVs. In order to analyze two potentially different trajectories of D5-like primases and SFIII helicases of phycodnaviruses s.l., we examined these proteins as two separate proteins. Indeed, the topological relationships of the D5-like clades of NCLDV, obtained with using of outgroups, as well as without, repeat the trajectories obtained for polymerases, that is, phycodnaviruses form a sister clade to the rest of NCLDVs (S18-20 Figs, S19-21 Aligns). At the same time, the topology of SFIII helicases obtained with using of bacterial (phage/plasmid primase, P4 family) outgroup show affinity between the phage/plasmid primase sequences of *Gammaproteobacteria* and homologues of *Phycodnaviridae* (S21 Fig, S22 Align). While, the phylogenetic trajectories of the D5-like primases and SFIII helicases of the *Mimiviridae* s.l. and MAPIM + *Asfarviridae*/*Poxviridae* clades demonstrate the invariance of sister relationships (S18-22 Figs, S23 Aligns).

Based on preliminary constructed trees belonging to SFII helicases, we identified the sequences most closely related to poxvirus D6/D11 helicases and constructed a separate D6/D11 phylogenetic tree. The resulting tree was rooted between branches containing the homologues of the ASFV D1133L and ASFV Q706L proteins, since in this case, the branch containing the ASFV D1133L homologues allowed us to show the repeatability of the branching order of NCLDV, demonstrated above using the

PolB, RNAP α and β, D5-primase, and SFIII helicase topologies. That is, sister relationships were identified between the joint MAPIM + AP and *Mimiviridae* s.l. clades, as well as between the AP and MAPIM clades (Fig 2, S24 Align). However, in contrast to the topologies mentioned above, in the case of the D6/D11 helicases, the branching order repeatability was provided not by cellular outgroups, which can potentially be problematic (see Materials and Methods, as well as Discussion), but by extremely closely related homologues of NCLDV (branch containing the ASFV Q706L homologues).

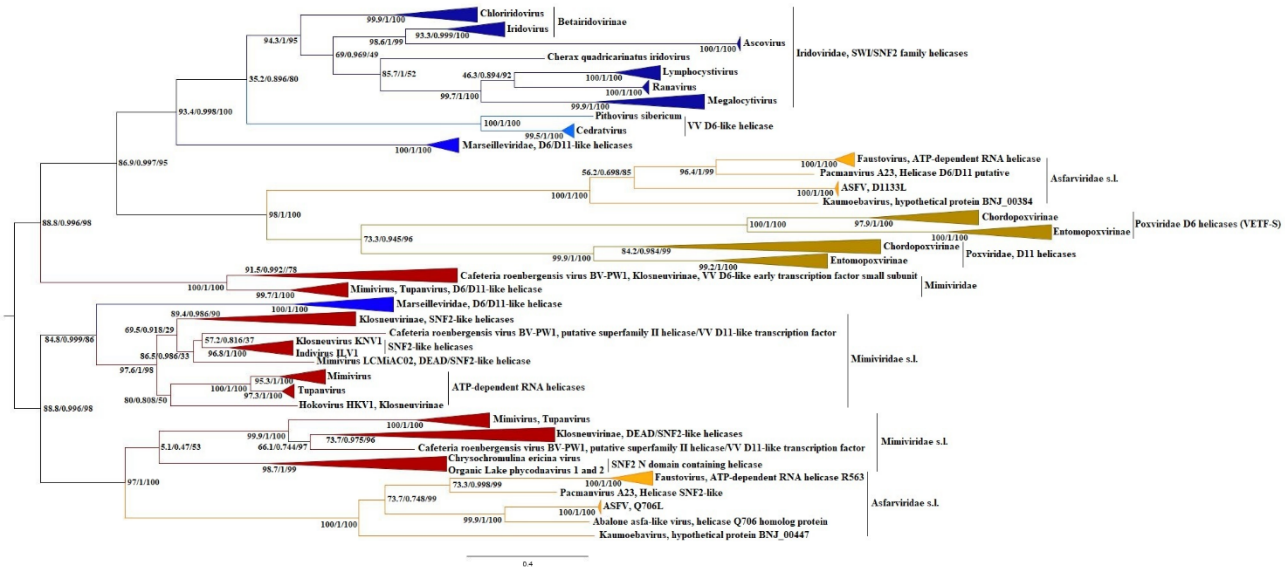


Fig 2. Maximum likelihood phylogenetic tree of superfamily II D6/D11-like helicases. The tree was constructed using the IQ TREE program (version 1.6.12) under the LG+F+R6 model and rooted between two clades containing the homologous sequences of *Mimiviridae* s.l., *Marseilleviridae*, and *Asfarviridae* s.l. (containing the ASFV D1133L and Q706L clades). The corresponding alignment was built using the MUSCLE algorithm. Statistical support of the partitions, measured by SH-aLRT branch tests (1,000 replicates), approximate Bayes and ultrafast bootstrap (1,000 replicates), and is shown at the nodes. Scale bars represent the number of amino acid substitutions per site.

Here we also present the phylogenies of the remaining proteins of the replication module, as well as proteins belonging to the functional category of RNA transcription and processing: AP (apurinic) endonuclease family 2 (S23 and 24 Figs, S25 and 26 Aligns), Metallopeptidase WLM (S25 and 26 Figs, S27 and 28 Aligns), PCNA (S27 and 28 Figs, S29 and 30 Aligns), RuvC (S29 and 30 Figs, S31 and 32 Aligns), Uracil DNA glycosylase (S31 and 32 Figs, S33 and 34 Aligns), Topoisomerase II (S33 and 34 Figs, S35 and 36 Aligns), mRNA capping enzyme (S35 and 36 Figs, S37 and 38 Aligns), and TFIIB (S37-39 Figs, S39-41 Aligns). The presented phylogenies were obtained both with the use of cellular outgroups and without them. Here we focus on the most important aspects of the obtained phylogenies summarized in (Table 1). The phylogenies of all individual proteins encoded by the genomes of *Phycodnaviridae* s.l. (with the exception of SFIII helicases, where the sequences of phycodnaviruses are localized among *Gammaproteobacteria* and *Betaproteobacteria* (S21 Fig)) confidently demonstrate primary split leading to the emergence of *Phycodnaviridae* s.l. and the clade consisting of the rest of NCLDVs. Phylogenies of eight proteins localize the AP clade sisterly to MAPIM, and only phylogenies of PolB (the MUSCLE algorithm derivatives) and Metallopeptidase WLM localize the AP clade sisterly to *Mimiviridae* s.l. Similar, phylogenies of eight proteins localize the joint AP + MAPIM clade sisterly to *Mimiviridae* s.l.

Table 1. Order of bifurcations of deep nodes of NCLDV.

	PolB	RNAP α	RNAP β	D5-like primase	D6/D11 helicase (SFII helicase)	SFIII helicase	AP (apurinic) endonuclease family 2	Metallopeptidase WLM	PCNA	RuvC	Uracil-DNA glycosylase	DNA topoisomerase II	mRNA capping enzyme (large subunit)	TFIIB	HerA-FtsK superfamily ATPases	DJR MCP	VLTF2	VETF
Ph/R	+	+	+	+	Ab	Ab	Ab	+	+	+	Ab	+	+	+	+	+	+	Ab
M/MAPIM+AP	-	-	+	+	+	+	+	-	-	+/-	+	+	+	-	+	+	+	+
M+AP/MAPIM	MU	-	-	-	-	-	-	MU	-	-	-	-	-	-	-	-	-	-
AP/MAPIM	-	-	+	+	+	+	+	-	-	-	+	+	+	-	+	+	+	+

Abbreviations: "/" – A sign indicating the presence of a node for clades, written to the right and to the left of it, "+" – Node exists, "-" – Node is missing, Ab – Lack of node due to lack of the *Phycodnaviridae* s.l. clade, Ph – the *Phycodnaviridae* s.l. clade, R – the NCLDV clades without the *Phycodnaviridae* s.l. clade, M – the *Mimiviridae* s.l. clade, MAPIM – the *Marseilleviridae* + Pithoviruses + *Iridoviridae*/*Ascoviridae* + *Mininucleoviridae* clade or some members of the MAPIM clade, AP – the *Asfarviridae* s.l. + *Poxviridae* clade or one of the two clades. $+^{A}$ – the sequences of the AP clade were not used, i.e. the corresponding phylogenetic trees (S29 and 30 Figs) show sister relationships between the *Mimiviridae* s.l. and MAPIM clades. MU - Present according to the MUSCLE algorithm.

The tertiary structure of proteins is largely but not completely determined by the primary structure. The prediction of the tertiary structures based on their primary structures is still an unsolved problem [43]. Moreover, unlike the computationally expensive but attractive physical interaction program, the evolutionary program imposes structural constraints based on bioinformatics analysis of the evolutionary history of proteins [43-45]. Thus, dendrograms calculated based on comparison of

crystallographic structures and predictions are not completely phylogeny-independent lines of evidence for the evolutionary trajectory of the clade under study. However, if they repeat the results of phylogenetic analysis, this speaks in favor of the reliability of the obtained phylogenetic trees. Logically more reliable should be dendrograms obtained based on crystallographic structures and physical interaction programs, since in both cases, tertiary structures and, as a result, dendrograms are independent of bioinformatics analysis, although they are more labor-intensive and resource-intensive. Based on the predictions obtained by ColabFold, we calculated dendrograms for the PolB (Fig 3, S40 Fig), RNAP α (S41 Fig), RNAP β (S42 Fig), PCNA (S43 Fig), TFIIB (S44 Fig), Topoisomerase II (S45 Fig), and Uracil DNA glycosylases (S46 Fig) proteins belonging to the replication module and functional category of RNA transcription and processing. The results of the PDB25 searches for the predicted structures and the Z-scores of the calculated dendrograms are shown in S1 and 2 Tables. The individual protein dendrograms shown here and below have been rooted according to the topologies of the corresponding individual protein phylogenies. The most important result, demonstrated by all dendrograms, is the repetition of the branching order of NCLDV obtained using the phylogenies of individual proteins. Phycodnaviruses either form a sister clade to all other NCLDVs (PolB (91-97 aa prediction, S40 Fig)), RNAP β (S42 Fig)), TFIIB (S44 Fig)) or are localized basal to crown NCLDVs (PolB (400-421 aa prediction, Fig 1), RNAP α (S41 Fig), PCNA (S43 Fig), Topoisomerase II (S45 Fig)). However, in contrast to individual phylogenies, the position of the AP clade is more variable. On the PolB 400-421 aa prediction dendrogram (Fig 1), the AP clade located sisterly to the joint *Mimiviridae* + MAPIM clade, whereas on the PolB 91-97 aa prediction dendrogram (S40 Fig), the AP clade located sisterly to the MAPIM clade, which is consistent with results for the vast majority of individual phylogenies. In turn, the topologies of the remaining dendrograms place the AP clade within the MAPIM clade.

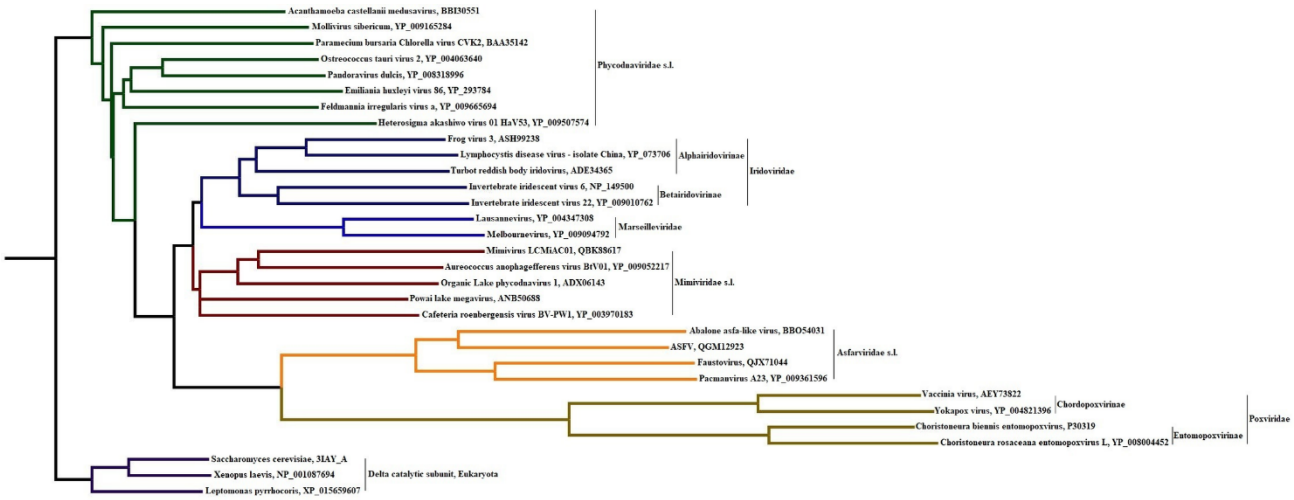


Fig 3. Eukaryotic and NCLDV PolB δ catalytic subunits dendrogram. The dendrogram based on the pairwise Z score comparisons and calculated using DALI. GenBank accessions are specified for each predicted protein structure. Protein structure prediction was done with ColabFold using MMseqs2 [44-45]. The length of the predicted structures ranged from 400 to 421 amino acids. Only unrelaxed rank 1 predicted protein structures were used to calculate the dendrogram. The dendrogram is rooted between the viral and cellular clades according to the topology of the Fig 1, as well as in published data [22, 40].

The morphogenetic module and synapomorphic proteins

Phylogenetic relationships of the FtsK-HerA superfamily ATPases are complex and incompletely resolved. Analysis of the presence and absence of characteristic secondary structure elements beyond the conserved core, as well as synapomorphic amino acid substitutions made it possible to delineate a clade consisting of the HerA, VirB3, VirB4, VirD4, FtsK and A32-like ATPases [46, 47]. We used the anchor amino acid substitutions identified by Iyer and colleagues [46] to create an alignment of the highly conserved Helix-1 - Strand 4 (sGsGKo – Q) fragment. The phylogeny of the clades of the

FtsK-HerA superfamily that we obtained coincide with Iyer and colleagues results [46] shown in Figure 4 (S47 Fig, S42 Align). This allowed us to reconstruct a tree consisting of the cellular homologues, as well as viral A32-like ATPase sequences (Fig 4, S43 Align). The topology obtained by us shows that bacterial FtsK and viral A32 form a clade with high statistical support. As in all other cases (Table 1) *Phycodnaviridae* s.l. and the rest of the NCLDV families form the two sister clades on the A32-like ATPase tree. We also constructed a tree without cellular orthologues (S48 Fig, S44 Align). The resulting topologies of both trees show that the MAPIM and AP clades are sisters, MAPIM + AP forms a sister clade to *Mimiviridae* s.l., which corresponds to the topologies of the eight proteins described above (Table 1). A similar direction of evolution demonstrated the FtsK-HerA superfamily ATPases dendrogram (S49 Fig). However, there are two differences between the topologies of the dendrogram and the phylogenic trees. First, on the dendrogram the *Emiliania huxleyi virus 86* prediction localized basally to the all crown NCLDVs, and the *Pandoraviridae* + *Prasinovirus* clade forms a sister clade to the all other NCLDV families. Second, the AP clade forms a sister group to the *Mimiviridae* s.l. clade, while on the phylogenetic tree, the AP clade is a sister clade to the MAPIM clade.

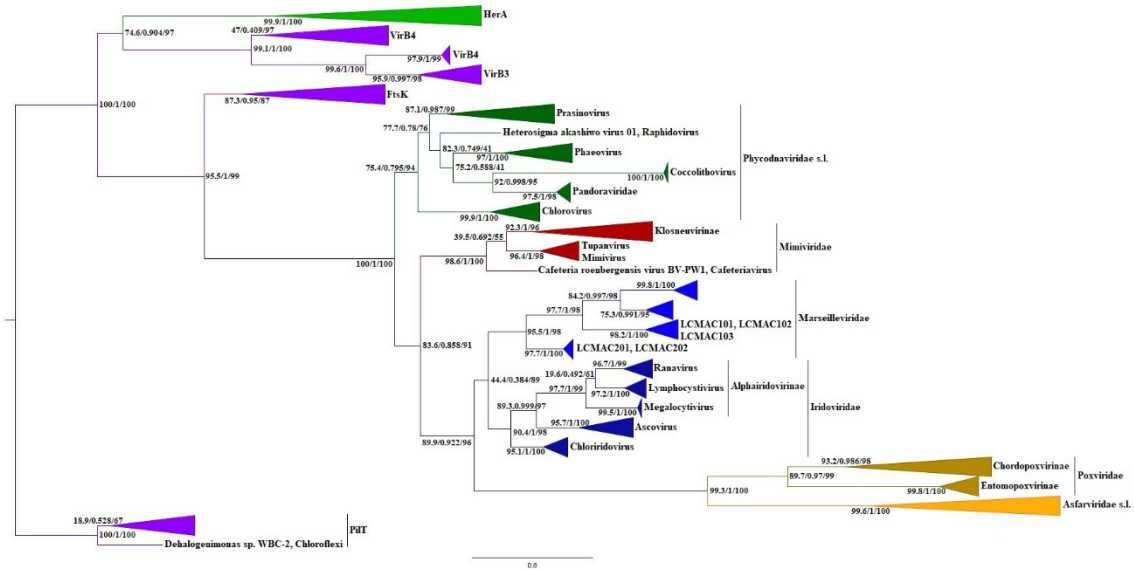


Fig 4. Maximum likelihood phylogenetic tree of the PilT, FtsK, HerA, VirB3, VirB4, and A32-like ATPases. The tree was constructed using the IQ TREE program (version 1.6.12) under the LG+R3 model and rooted according to the S47 Fig topology, as well as in published data [46]. The corresponding alignment of the Helix-1 – Strand 4 (sGsGKo – Q) fragment was built using the MUSCLE algorithm. Statistical support of the partitions, measured by SH-aLRT branch tests (1,000 replicates), approximate Bayes and ultrafast bootstrap (1,000 replicates), and is shown at the nodes. Scale bars represent the number of amino acid substitutions per site.

Next, we show the phylogenies of the DJR MCP (S50 Fig, S45 Align) and VLTF-2 (S51 Fig, S46 Align) proteins. DJR MCP belongs to the morphogenetic module, while VLTF-2 is involved in the transcription and RNA processing. Both phylogenies were constructed without using of outgroup and for this reason, both phylogenies were rooted according to the direction of evolution of individual proteins presented above. Notably, the topologies of both synapomorphic proteins showed sister relationships between the AP and MAPIM clades, as well as between the joint AP + MAPIM and *Mimiviridae* clades. However, it should be noted that the topologies of both proteins can be rerooted

in such a way that the direction of evolution of NCLDV will be consistent with the direction proposed by Koonin and Yutin [10].

Finally, we show the N and C terminal jelly roll barrel (tjrb) dendrograms (S52 and 53 Fig), and the dendrogram of full-size DJR MCPs (S54 Fig) whose tertiary structures were predicted using ColabFold, as well as the dendrogram of full-size DJR MCPs (Fig 5) predicted using I-TASSER and combined with crystallographic structures deposited in the PDB. All predictions were made without removing any amino acid residues within the sequences. All MCP dendrograms were rooted according to the phylogenies of individual protein phylogenies showing the direction of evolution of NCLDV presented above and obtained using outgroups. Unlike most of the dendrograms presented above, and, conversely, according to most individual phylogenies, the topologies of the N and C tjrbs show sister localization of the MAPIM and AP clades. However, in contrast to the both N and C tjrb dendrograms, the topology of the full-length DJR MCPs dendrogram calculated based on the ColabFold predictions placed the AP clade sisterly to the joint MAPIM + *Mimiviridae* s.l. clade. As an alternative, we calculated a dendrogram composed of the PDB crystallographic structures and I-TASSER predictions, excluding ColabFold predictions from it, since among the latter predictions of mimiviruses and the MAPIM clade demonstrated attraction to each other, separating from the rest of members of their clades, which are either crystallographic structures deposited in the PDB or I-TASSER predictions (S1 Data). The dendrogram calculated based on comparisons of Z-scores of the crystallographic structures and I-TASSER predictions repeated the topologies of the N and C tjrb of the DJR MCP dendrograms.

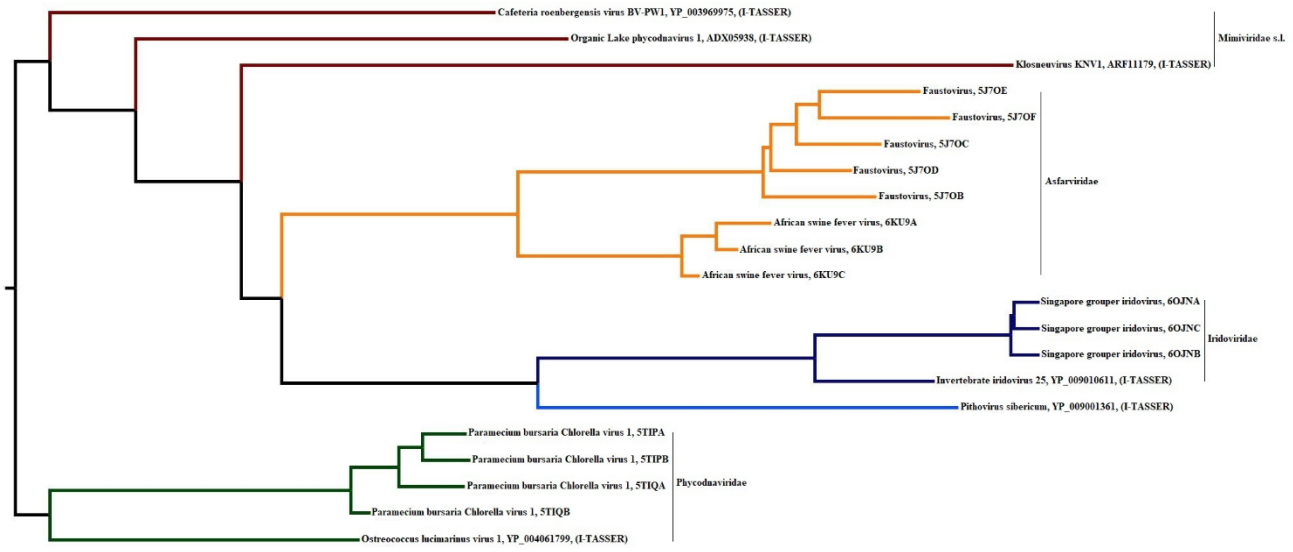


Fig 5. Dendrogram of the NCLDV DJR MCPs. The dendrogram based on the pairwise Z score comparisons calculated using DALI. Protein structure prediction was done with I-TASSER [48-50]. The dendrogram was rooted between *Phycodnaviridae* and *Mimiviridae* s.l. + MAPIM + AP clades according to the topologies of all phylogenetic trees with outgroups presented in this work. GenBank accessions are given for the sequences used for the I-TASSER predictions, as well as for the crystallographic structures deposited in the PDB. The I-TASSER prediction was carried out without removing of any amino acid residue within the sequences

The main conclusions from the analysis of 18 individual proteins are that: i) in all 11 cases (D5-like primase, DNA topoisomerase II, PolB, A32-like ATPase, Metallopeptidase WLM, mRNA capping enzyme, PCNA, RNAP α, RNAP β, RuvC, and TFIIB) where outgroup was used, the gene product was encoded by the genomes of phycodnaviruses and was not secondary transferred, there was a well-supported sister relationship between the *Phycodnaviridae* s.l. clade and the clade consisting of all other NCLDVs (Table 1); ii) sister relationships between the AP and MAPIM clades, as well as between the AP + MAPIM and *Mimiviridae* s.l. clades were observed in 12 cases for both clades

(Table 1). Dendrograms calculated using outgroups showed an evolutionary trajectory similar to individual phylogenies. Phycodnaviruses either form a sister clade to all other NCLDVs or are located basal to crown NCLDVs. The topological position of the AP clade on dendrograms shows somewhat greater variability compared to individual phylogenies. However, the AP clade is always located within the AP + MAPIM + *Mimiviridae* s.l. clade (see above).

Branching order of NCLDV. The SPR superdendrogram and SPR supertree phylogenies support the sister relationship between *Phycodnaviridae* s.l. and all other NCLDVs, as well as clarifying the position of the *Asfarviridae*/*Poxviridae* clade

One of approaches to resolve direction of evolution of some clade and topological disagreements resulting from phylogenetic analysis of various genes and proteins is to construct supertrees. Similarly, a superdendrogram can be constructed based on dendrograms of various proteins. In our case, the topological conflict refers to the position of the AP clade.

In order to construct the SPR superdendrogram, we generated 12 ColabFold derived dendrograms of the 9 core proteins (S2-13 Data). The topology of the superdendrogram shows sister relationships between the clades AP and MAPIM, as well as between the joint MAPIM + AP and *Mimiviridae* s.l. clades (Fig 6). It should be noted that the monophyletic *Phycodnaviridae* s.l. clade split into two clades. The first clade consists of the *Phaeovirus*, *Coccolithovirus* and *Pandoraviridae* predictions. This clade localized sisterly to the joint *Mimiviridae* s.l. + MAPIM + AP clade. The second clade consists of the *Raphidovirus*, *Acanthamoeba castellanii medusavirus*, *Chlorovirus*, *Prasinovirus*, and *Mollivirus sibericum* predictions. The second clade localized basal to the crown NCLDVs.

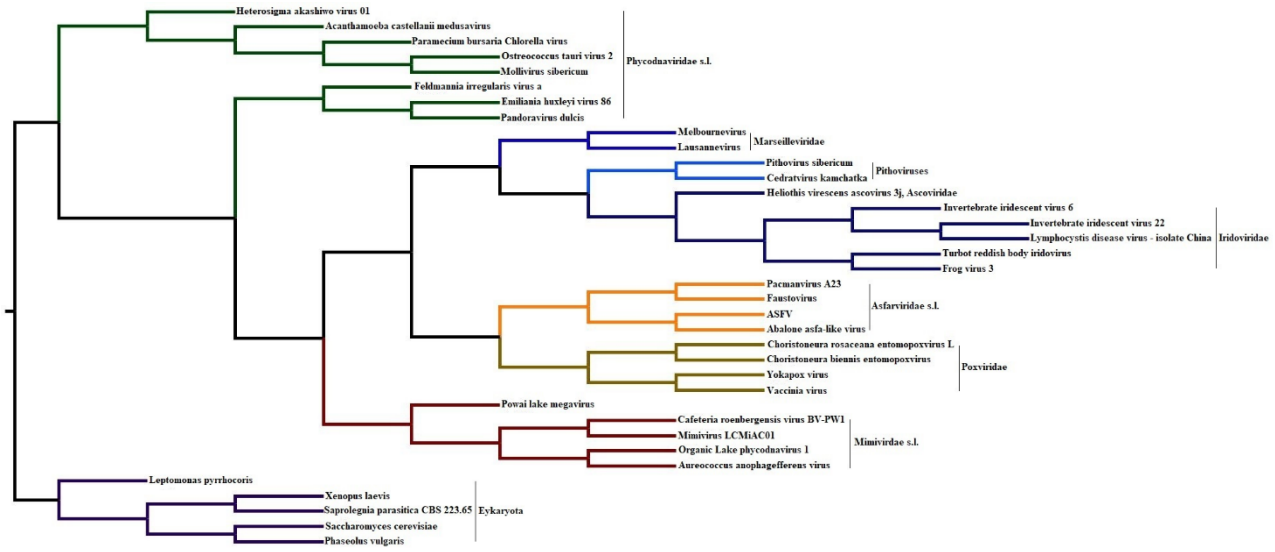


Fig 6. SPR superdendrogram. The dendrogram based on 12 ColabFold derived of the 9 core proteins (PolB, RNAP α, RNAP β, PCNA, Topoisomerase II, TFIIB, A32-like ATPase, Uracil DNA glycosylase, as well as N and C terminal jelly roll barrels of DJR MCP). The superdendrogram rooted between *Eukaryota* and NCLDV according to the topologies of all individual phylogenies containing the cellular and NCLDV clades presented in this work. The input dendrograms of the PolB, RNAP α, RNAP β, PCNA, Topoisomerase II, and TFIIB predictions contained the corresponding eukaryotic outgroups. The remaining input dendrograms were calculated without using outgroups.

In order to construct the SPR supertree, we generated the 23 ML trees of the 13 core proteins with a limited (1-4 sequences from each genus or subfamily) taxonomic composition, using the standard bootstrap method for statistical support of nodes (S14-59 Data). In contrast to the superdendrogram, the topology of the supertree has preserved monophyly of *Phycodnaviridae* s.l., which is localized sisterly to the joint *Mimiviridae* s.l. + MAPIM + AP clade (Fig 7). At the same time, the topological relationships between the AP and MAPIM and joint AP + MAPIM and *Mimiviridae* s.l. clades

repeated the topology of the superdendrogram as well as the vast majority of individual phylogenies (see above).

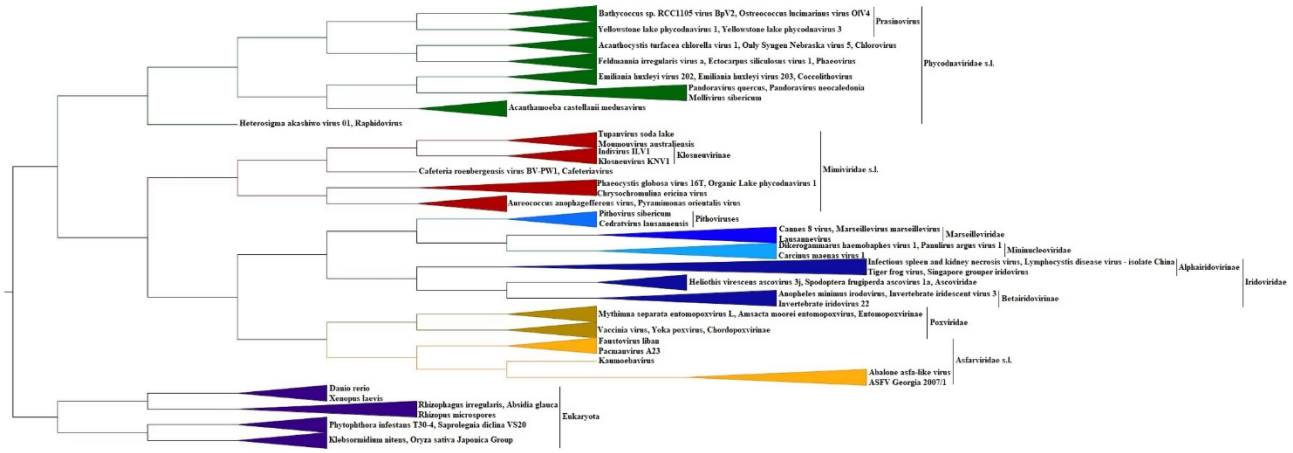


Fig 7. **SPR supertree.** The supertree constructed based on the 23 maximum likelihood (ML) trees of the 13 core proteins (A32-like ATPase, AP (apurinic) endonuclease family 2, D5-like primase, mRNA capping enzyme, PolB, RNAP α, RNAP β, DJR MCP, SFIII helicase, TFIIB, Uracil DNA glycosylase, VETF, and VLTF-2). The supertree was constructed with a bootstrap threshold equal to 80 and rooted between *Eukaryota* and NCLDV according to the topologies of all individual phylogenies containing the cellular and NCLDV clades presented in this work. The input trees of the D5-like primase, mRNA capping enzyme, PolB, RNAP α, RNAP β, and TFIIB were constructed with using of the corresponding eukaryotic outgroups. The remaining input trees were constructed without using outgroups.

Branching order of NCLDV. Supermatrix phylogenies support the sister relationship between *Phycodnaviridae* s.l. and all other NCLDVs, as well as clarifying the position of the *Asfarviridae*/*Poxviridae* clade

Concatenation is an alternative way of solving the phylogenetic disagreements among individual protein phylogenies. This refers to conflicting topologies of individual proteins with respect to the position of the AP clade. Alignments of the i) 16 individual proteins of eukaryotic/protoeukaryotic, bacterial, and unknown origin with and without using of eukaryotic outgroup (S3 Table, Supermatrixes 1 and 2), ii) 6 individual proteins of eukaryotic/protoeukaryotic origin with and without using of eukaryotic outgroup (S3 Table, Supermatrixes 3 and 4), iii) 4 individual proteins of bacterial origin with and without using of bacterial outgroup (S3 Table, Supermatrixes 5 and 6), and iiii) 3 individual proteins of unknown origin without outgroup (S3 Table, Supermatrix 7) were used for concatenation. Thus, based on the 7 concatenated alignments, we constructed 7 different concatenated phylogenies (Fig 8, S55-60 Figs; S47-53 Aligns).

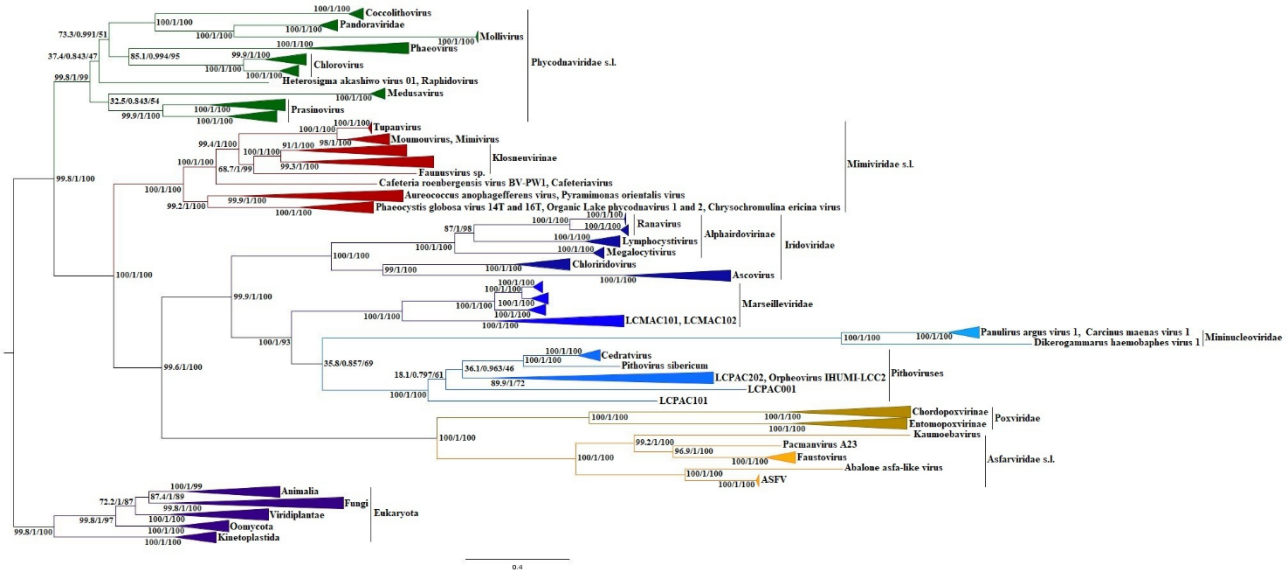


Fig 8. **Maximum likelihood phylogenetic tree of the NCLDVs.** The tree was constructed via concatenation of the 16 core protein (D5-like primase, PolB, mRNA capping enzyme (large subunit), PCNA, RNAP α, RNAP β, TFIIB, Topoisomerase II, FtsK-HerA superfamily ATPase, AP (apurinic) endonuclease family 2, Holliday junction resolvase, Uracil-DNA glycosylase, SFIII helicase, DJR MCP, VETF, and VLTF-2) alignments consisting of NCLDV and eukaryotic homologues using the IQ TREE program (version 1.6.12) under the EX_EHO+F+I+R4 model. The tree was rooted between *Eukaryota* and NCLDV according to the topologies of all individual phylogenies with outgroups presented in this work. The supermatrix was built by head-to-tail linking of already aligned (MUSCLE) individual protein alignments. Statistical support of the partitions, measured by SH-aLRT branch tests (1,000 replicates), approximate Bayes and ultrafast bootstrap (1,000 replicates), and is shown at the nodes. Scale bars represent the number of amino acid substitutions per site.

The cornerstone around which all the evolutionary trajectories of our work are built is the direction of evolution of NCLDV. The use of the eukaryotic and bacterial outgroups showed that the direction of evolution of the mixed, eukaryotic, bacterial, and "synapomorphic" concatenated trees, as expected, repeats the branching order of NCLDV obtained for the phylogenies and dendrograms of individual proteins, as well as for the supertree and superdendrogram. That is, the *Phycodnaviridae* s.l. and joint *Mimiviridae* s.l. + MAPIM + AP clades showed sister relationships. Moreover, all seven concatenated phylogenies repeated the topological relationships between the *Mimiviridae* s.l., MAPIM, and AP clades, i.e. the AP clade with high statistical support turned out to be sister to the MAPIM clade, and the joint MAPIM + AP clade, in turn, with high statistical support turned out to be sister to mimiviruses.

Phylogenetic relationships among *Nucleocytoviricota, Preplasmiviricota,* a number of classified and unclassified viruses and mobile genetic elements

Replication module

The results of our study show that the PolBs clades CLP + MLP + *Preplasmiviricota* (without *Turriviridae*, *Tectiviridae*, *Autolykiviridae*, and *Corticoviridae*) and *Herpesviridae* are a sister groups to eukaryotic δ catalytic subunits (S61 and 62 Figs, S54 and 55 Aligns). At the same time, the *Hytrosaviridae* PolBs were found to be localized within the δ clade, suggesting their origination from within the eukaryotic δ catalytic subunits clade (S63 Fig, S56 Align). Our results on the topological position of the *Herpesviridae* and *Hytrosaviridae* polymerases are consistent with previously published data [21-22].

We removed the δ catalytic subunit clades from the S54 and S55 Alignments, assuming that if the topological positions of viral clades are not the results of homoplasies, then the viral clades will retain their previous topological positions. Indeed, the *Herpesviridae* and *Adenoviridae*/Polintons clades retained their topological positions, thus confirming their relationship with eukaryotic δ subunits (S64 and 65 Figs; S57 and 58 Aligns).

The sister localization of the *Herpesviridae*, *Adenoviridae*/Polintons, and NCLDV clades to the eukaryotic δ catalytic subunits (for NCLDV see above) actualized the need to check the phylogenetic relationships among the *Herpesviridae*, *Adenoviridae*/Polintons, and NCLDV PolB clades. Reconstruction of the phylogenetic tree, including the eukaryotic δ catalytic subunits, NCLDV and *Herpesviridae* polymerases, showed that *Herpesviridae* form a sister clade to the joint *Mimiviridae* s.l. + MAPIM clade with high statistical support (S66 Fig; S59 Align). The topological localization of

herpesviruses, as well as high statistical support of the node that unites herpesviruses, mimiviruses and the MAPIM clade suggests that the ancestral *Herpesviridae* PolB gene has been acquired from the lineage leading to the last common ancestral gene of the *Mimiviridae* s.l. and MAIPIM clades.

Our results show that Family B DNA polymerases of Polintons, *Adintoviridae*, *Adenoviridae*, CLPs, MLPs, *Bovine rumen MELD virus*, *Sheep rumen MELD virus*, *Bidnaviridae*, and *Parvoviridae* form a clade with the highest statistical support whose last common ancestral gene originated from within NCLDV. Both trees based on the MUSCLE (Fig 9; S60 Align) and MAFFT (S67 Fig; S61 Align) algorithms showed the sister localization of the *Adenoviridae*/Polintons clade to the *Mimiviridae* s.l. clade. However, on both trees, the statistical support of the joint *Mimiviridae* s.l. + *Adenoviridae*/Polintons clade is low. We built de novo a new alignment using the MUSCLE algorithm adding to it the PolB sequences of the *Monoraphidium MELD virus SAG 48.87*, *Rhizophagus glomeromycete fungus MELD virus 6100*, *Blastocystis MELD virus 3996*, and *MELD virus sp. ctWXX4* viruses. During the building of this alignment, we have sought to maximize the preservation of conservative vertical blocks of the DALI structure-based sequence alignment of PolBs, highlighted in color in Supplementary Figure S1 [22]. We aligned the interval including the Exo I (DIE) and MotifC (YGDTDS) motifs and reconstructed the tree (S68 Fig, S62 Align). Next, we presented the topology of the dendrogram based on the ColabFold predictions which shows the sister localization of the *Adenoviridae*/Polintons clade to the joint *Mimiviridae* s.l. + MAPIM clade (Fig 10). Thus, we do not have reliable statistical arguments for suggesting the place of the *Adenoviridae*/Polintons clade among the *Mimiviridae* s.l. and MAPIM clades, but on all trees we obtained, the joint *Mimiviridae* s.l. + MAPIM + *Adenoviridae*/Polintons clade had high statistical support, which allowed us to confidently suggest the origin of the *Adenoviridae*/Polintons clade ancestral PolB gene or at least the ancestral Exo I – MotifC fragment of PolB from within the joint *Mimiviridae* s.l. + MAPIM clade.

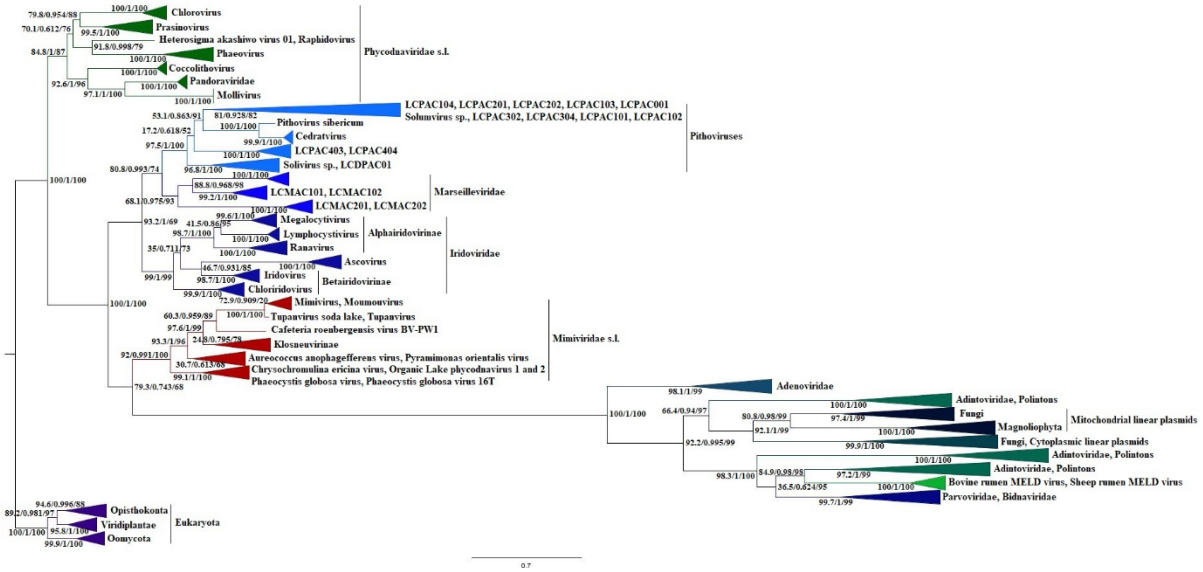


Fig 9. Maximum likelihood phylogenetic tree of Family B DNA polymerases. The tree consist of the eukaryotic δ catalytic subunits, the NCLDV, *Adenoviridae*, *Adintoviridae*, *Bidnaviridae*, *Parvoviridae*, *Bovine rumen MELD virus*, *Sheep rumen MELD virus*, Polintons, Mitochondrial linear plasmids, and Cytoplasmic linear plasmids Family B DNA polymerases. The tree was constructed using the IQ TREE program (version 1.6.12) under the LG+F+I+R4 model and rooted according to the Fig 1 and S9-10 Fig topologies. The corresponding alignment was built using the MUSCLE algorithm. Statistical support of the partitions, measured by SH-aLRT branch tests (1,000 replicates), approximate Bayes and ultrafast bootstrap (1,000 replicates), and is shown at the nodes. Scale bars represent the number of amino acid substitutions per site.

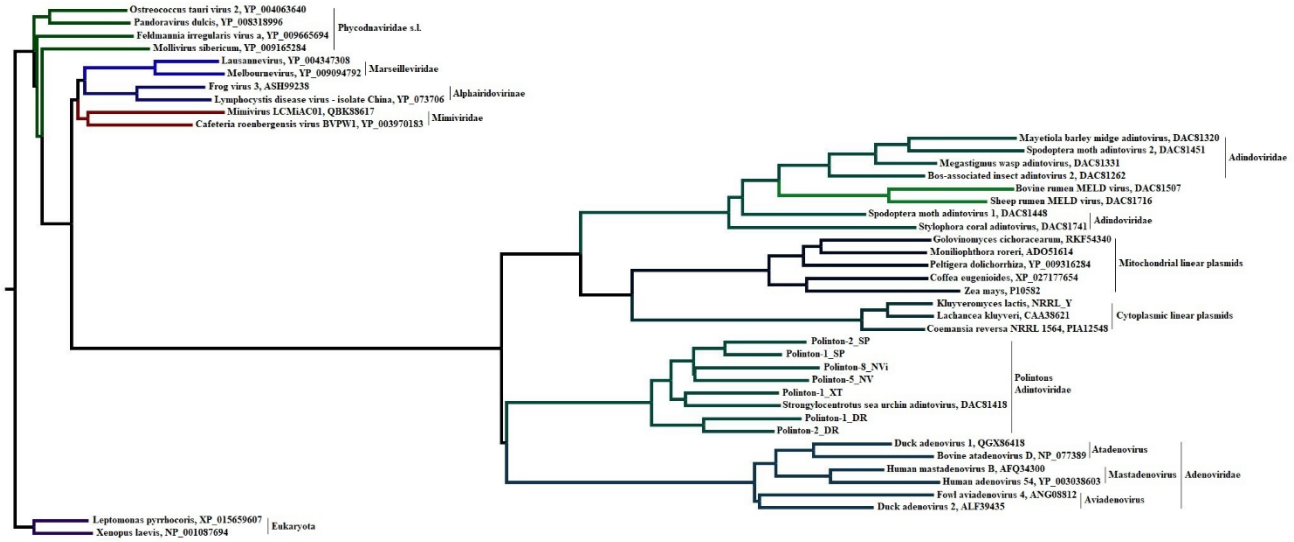


Fig 10. Dendrogram of Family B DNA polymerases. The dendrogram consist of the eukaryotic δ catalytic subunit, NCLDV, *Adenoviridae*, *Adintoviridae*, *Bidnaviridae*, *Parvoviridae*, *Bovine rumen MELD virus*, *Sheep rumen MELD virus*, Polintons, Mitochondrial linear plasmids, and Cytoplasmic linear plasmids Family B DNA polymerase predictions and based on the pairwise Z score comparisons calculated using DALI. GenBank accessions are specified for each predicted protein structure. Protein structure prediction was done with ColabFold using MMseqs2 [44, 45]. The length of the predicted structures ranged from 400 to 615 amino acids. Only unrelaxed rank 1 predicted protein structures were used to calculate the dendrogram. The dendrogram was rooted between the viral and cellular clades according to the Fig 1, Fig 9, and S9 and10 Figs topologies.

Finally, we show that the Exo I – MotifC fragment of *Tectiviridae*, *Autolykiviridae* (*Tectiliviricetes*, *Preplasmiviricota*), *Acidianus bottle-shaped virus* (*Ampullaviridae*), *Nitrosopumilus spindle-shaped virus* (*Fuselloviridae*), *His2 virus* (*Pleolipoviridae*), and *His 1 virus* (*Salterprovirus*) along with the homologous of *Euryarchaeota*, *Thaumarchaeota*, as well as *Firmicutes* sequences form a clade with

high statistical support, which, in turn, with high statistic support, are localized sisterly to the clade of eukaryotic ε catalytic subunits (S69 Fig, S63 Align).

Thus, we assume that at least the Exo I – MotifC fragment c of adenoviruses and adintoviruses on the one hand, tectiviruses and autolykiviruses on the other hand originate from different phylogenetic lineages (NCLDV and ε catalytic subunit), in other words, according to the Exo I – MotifC fragment phylogenies, *Preplasmiviricota* are polyphyletic.

The D5-like primase proteins of *Sputnikvirus* (*Lavidaviridae*, *Preplasmiviricota*) form a sister clade to the *Iridoviridae*/*Ascovirus* clade (S70 Fig, S64 Align) which is consistent with the data on PolB for *Adenoviridae*/Polintons line. Removal of the *Iridoviridae*/*Ascovirus* sequences from the alignment did not change the topological localization of sputnikviruses (S71 Fig, S65 Align). Thus, we have a strong claim for suspecting the D5-like primases of sputnikviruses originated from within the joint *Mimiviridae* s.l. + MAPIM clade.

Tree, consisting of the NCLDV D6/D11 (SFII) helicases, the DEAD-like helicases of *Fungi* (CLPs) and *Physcomitrella patens* (*Bryophyta*, chromosomes 3 and 13) showed that the fungal and *P. patens* sequences are localized within the MAPIM clade (Fig 11, S66 Align). However, the length of the branches of both sister clades (*Iridoviridae* + Pithoviruses and CLP + *P. patens*) is quite significant. The removal of the *Iridoviridae* + Pithoviruses sequences from the S66 Alignment, taking into account the topology of Fig 2, that is, in fact, the reciprocal removal of sister clades, showed that the CLP + *P. patens* clade does not change topological position, which speaks in favor of the absence of homoplasies and the LBA effect (S72 Fig, S67 Align). Thus, we suggest that the DEAD-like helicases of CLP and *P. patens* originated from within NCLDV. Above (see Fig 2), we spoke about the extreme importance of the D6/D11 helicase phylogeny, arguing that, in contrast to the phylogenies of proteins obtained using eukaryotic or bacterial outgroups, in the case of the D6/D11 helicases closely related

viral homologues were used as an outgroup. This allowed us to show that the phylogenetic relationships between the MAPIM and AP clades, as well as between the joint MAPIM + AP and *Mimiviridae* s.l. clades, are the same as in the cases of proteins whose phylogenies were obtained using cellular outgroups. The addition of the *Chlorovirus*, CLP, and *P. patens* homologues to the alignment did not change the topological relationships of the *Mimiviridae* s.l., MAPIM, and AP clades. Thus, we have strengthened the statistical arguments in favor of our proposed direction of evolution of NCLDV, justified by the use of cellular outgroups.

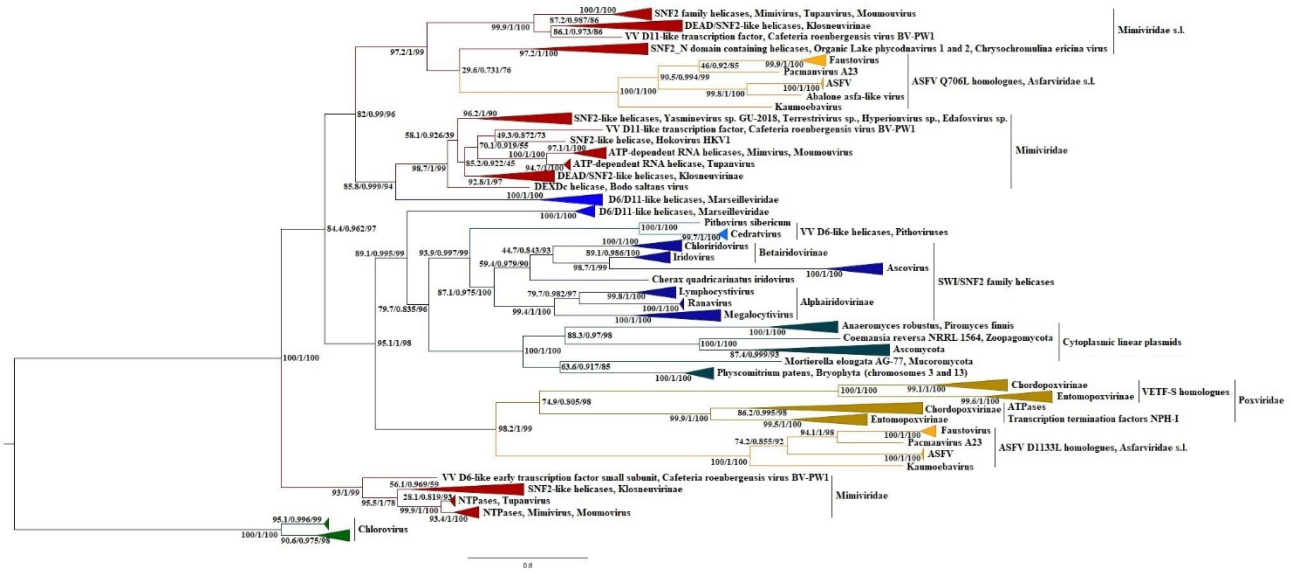


Fig 11. Maximum likelihood phylogenetic tree of the D6/D11 and DEAD-like helicases. The tree consist of the *Chlorovirus*, *Mimiviridae* s.l. *Marseilleviridae*, *Cedratvirus*, *Pithovirus sibericum*, *Iridoviridae*/*Ascoviridae*, *Asfarviridae* s.l., *Poxviridae*, fungal Cytoplasmic linear plasmid, and *Physcomitrella patens* (chromosomes 3 and 13) protein sequences. The tree was constructed using the IQ TREE program (version 1.6.12) under the LG+C20+F+I+R4 model and rooted between chloroviruses and the clade consisting of the rest of D6/D11 + CLP + *P. patens* homologues. The corresponding alignment was built using the MUSCLE algorithm. Statistical support of the partitions,

measured by SH-aLRT branch tests (1,000 replicates), approximate Bayes and ultrafast bootstrap (1,000 replicates), and is shown at the nodes. Scale bars represent the number of amino acid substitutions per site.

The phylogenetic tree consisting of the ASFV F1055L (SFII helicases) homologues of NCLDV (*Mimiviridae*, *Marseilleviridae*, and *Asfarviridae*) and *Herpesvirales* places the monophyletic *Herpesviridae* + *Malacoherpesviridae* clade in a sister position to *Asfarviridae* (S73 Fig, S68 Align). However, the statistical support of the node combining *Asfarviridae* and *Herpesvirales* is low. Moreover, the *Herpesvirales* branch is long and may contain homoplasies with the asfarvirus branch, which may cause the artificial topology. We removed the sequences of asfarviruses from alignment and reconstructed the tree (S74 Fig, S69 Align). As can be seen from S74 Fig, the topological position of *Herpesvirales* has not changed. Furthermore, the statistical supports for nodes joint asfarviruses, marseilleviruses, and herpesviruses are high in both trees, which suggests that the herpesvirus homologues originated from within NCLDV, similarly to the PolB orthologues of *Herpesviridae*.

In the next step, we included the *Herpesviridae* and *Lavidaviridae* homologues of the SFIII helicases (D5-like primase-helicase) into our analysis. Our results suggest that the LCA genes of unclassified virophages, *Sputnikvirus* (*Lavidaviridae*), as well as *Herpesviridae* originated from within NCLDV after the divergence of mimiviruses (S75 Fig, S70 Align). In fact, the removal of the joint MAPIM + AP clade (the sister clade to the *Lavidaviridae* and *Herpesviridae* clades) is reciprocal, since *Lavidaviridae* and *Herpesviridae* were added to the S22 Alignment corresponding to S21 Fig (S76 Fig, S71 Align). The resulting topology show that the position of herpesviruses and virophages remained unchanged, which implies the absence of the homoplasies and LBA effects. Thus, we suggest that the SFIII helicases of virophages and herpesviruses originated from within NCLDV.

Transcription and RNA processing

Reconstruction of the mRNA capping enzyme tree, consisting of homologous sequences of eukaryotes, NCLDV and CLPs, showed the origin of the plasmid sequences from within NCLDV, which is fully consistent with the data on PolB and D6/D11 helicases (S77 Fig, S72 Align). The removal of the sequences of marseilleviruses, pithoviruses, and cedratviruses from the alignment did not affect the topological position of the CLP clade (S78 Fig, S73 Align). This suggests that, like the PolB and D6/D11 genes, the ancestral mRNA capping enzyme gene of CLPs could also had originated from within NCLDV.

In contrast to the PolB, D6/D11 helicase and mRNA capping enzyme phylogenies we got a different result for the RNAP second largest (β) subunit. The CLP clade was found localized sisterly to the RPA2 clade (S79 Fig, S74 Align). Removal of the RPA2 sequences from the S74 Alignment did not change the topological position of the CLP clade (S80 Fig, S75 Align). Considering the sister relationship of the NCLDV and RPB2 clades, as well as the sister relationship of the CLP and RPA2 clades, we reject the possibility of the origin of the ancestral second largest (β) subunit gene of CLP from within NCLDV.

The interim conclusions from our data on the genome replication and expression machinery suggest that a number of genes of herpesviruses, *Lavidaviridae*, *Adenoviridae*, *Adintoviridae*, a number of unclassified viruses, CLPs (without RNAP β), and MLPs originated from within the NCLDV clade. The obtained data suggest that after bifurcation leading to the *Phycodnaviridae* s.l. and *Mimiviridae* s.l. + MAPIM + AP branches, the second branch became a kind of genetic "supernova", which took over the propagation function of the NCLDV genes not only vertically but also horizontally.

Morphogenetic module

There are conflicting data in the literature on the belonging of adenoviral IVA2 to different branches of ASCE ATPases [46-47]. It has been shown that the adenoviral packaging ATPases and terminases have an additional strand to the "right" of the core P-loop domain, which suggests their evolutionary relationship with the FtsK-HerA and PilT/VirB11 superfamilies [46, 51]. At the same time, it has been proposed that adenoviral IVA2 ATPases may have originated from the ABC superfamily of transport ATPases [47]. IVA2 and ABC transporters share two polar residues at the end of the sensor-1, one of which is highly conserved histidine [47]. In addition, prediction of the secondary structure of IVA2 suggests that the adenovirus ATPases contain an insert with β-strands after Helix-1. However, ABC transporters after β-strands contain one more strand what makes them different from IVA2 [46-47].

We aligned the Helix-1 – Strand 4 (sGsGKo – Q) fragment of the IVA2, HerA, and FtsK sequences with the ABC bacterial and eukaryotic homologues and reconstructed a tree. We proceeded from the assumption that if the adenoviral Helix-1 – Strand 4 fragment originates from ABC transporters, then it must be localized outside of the FtsK-HerA clade. However, the resulting topologies nested the *Adenoviridae* Helix-1 – Strand 4 fragment within the HerA-FtsK clade (S81 Fig, S76 Align). Localization of the IVA2 clade actualized the construction of a tree consisting of the Polintons, *Adintoviridae*, *Adenoviridae*, *Lavidaviridae* ATPases, as well as their cellular homologues (S82 Fig, S77 Align). At this step, we also used the anchor amino acid substitutions belonging to the Helix-1 – Strand 4 (sGsGKo – Q) fragment highlighted in Figure 2 [46]. The resulting topology shows that the *Adintoviridae* + Polintons + *Lavidaviridae* + *Adenoviridae* clade is a sister clade to the FtsK clade. The addition of the A32-like ATPase (NCLDV), *Pleurochrysis sp. endemic virus 2*, *Pleurochrysis sp. endemic virus 1a*, *Bovine rumen MELD virus*, and *Sheep rumen MELD virus* homologues to the analysis showed that the *Adenoviridae*/Polintons clade nested within NCLDV (Fig 12, S78 Align). Such a topological localization of the *Adenoviridae*/Polintons clade as a whole or its constituent parts is consistent with the PolB, D5-like primase, D6/D11 helicase, SFIII helicase and mRNA capping

enzyme data presented above. In order to test the potential influence of homoplasies on the topological position of the *Adenoviridae*/Polintons clade, we i) removed the sister mimiviruses from the alignment, ii) removed the entire outgroup of the joint NCLDV + *Adenoviridae*/Polintons clade and reconstructed trees (S83 and 84 Figs, S79 and 80 Align). The results obtained confirmed the origin of the FtsK-HerA superfamily ATPases of the *Adenoviridae*/Polintons clade from within the joint *Mimiviridae* + MAPIM clade. However, the low statistical support of the *Mimiviridae* + *Adenoviridae*/Polintons node (S84 Fig) did not allow pinpointing the topological position of the *Adenoviridae*/Polintons clade within the joint *Mimiviridae* + MAPIM clade.

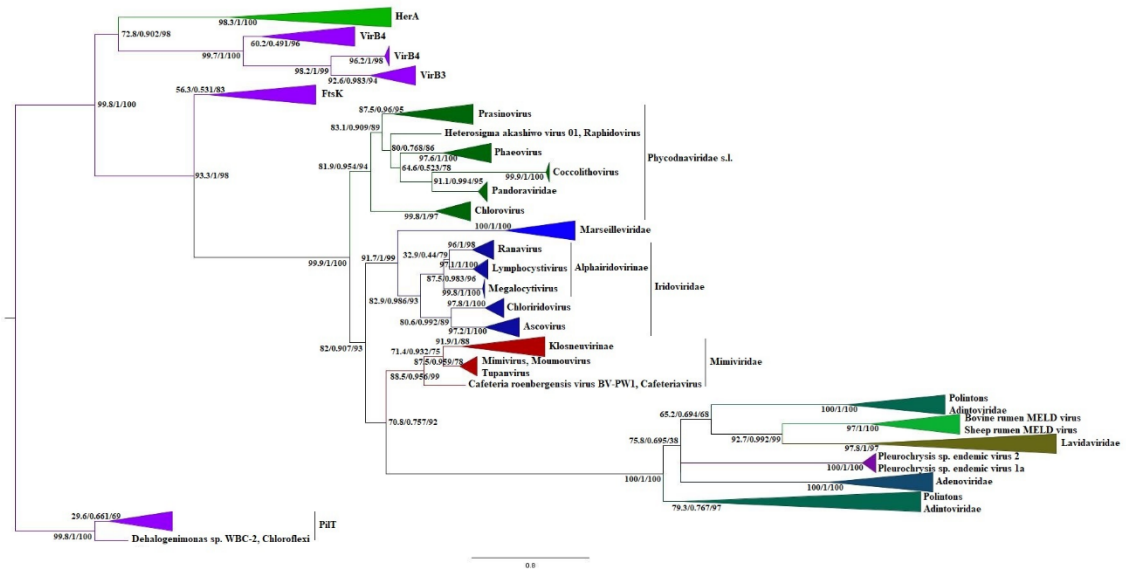


Fig 12. Maximum likelihood phylogenetic tree of the FtsK-HerA superfamily ATPases. The tree consist of the archaeal HerA, bacterial FtsK, ViB3, VirB4, A32-like (NCLDV), IVA2 (*Adenoviridae*), Polintons, *Adintoviridae*, *Lavidaviridae*, *Pleurochrysis sp. endemic virus 2*, *Pleurochrysis sp. endemic virus 1a*, *Bovine rumen MELD virus*, and *Sheep rumen MELD virus* homologues. The tree was constructed using the IQ TREE program (version 1.6.12) under the EHO+I+R4 model and rooted according to the topology of Fig 4, as well as in published data [46]. The corresponding alignment

was built using the MUSCLE algorithm. Statistical support of the partitions, measured by SH-aLRT branch tests (1,000 replicates), approximate Bayes and ultrafast bootstrap (1,000 replicates), and is shown at the nodes. Scale bars represent the number of amino acid substitutions per site.

In order to identify phylogenetic relationships of the DJR MCP proteins of the unclassified DJR MCP viruses, NCLDV, and *Preplasmiviricota* (without *Turriviridae*, *Tectiviridae*, *Autolykiviridae*, and *Corticoviridae*), we built two alternative alignments, consisting of the NCLDV, *Adenoviridae*, *Lavidaviridae*, *Adintoviridae*, *Yaravirus brasiliensis*, *Pleurochrysis sp. endemic virus 1a*, *Pleurochrysis sp. endemic virus 1b*, *Pleurochrysis sp. endemic virus 2*, *Bovine rumen MELD virus*, *Sheep rumen MELD virus* and Polintons sequences. In the first alignment, the sequences of *Preplasmiviricota*, Polintons, and unclassified DJR MCP viruses were aligned first, and then that multiple alignment was aligned with the prealigned NCLDV sequences (S85 Fig, S81 Align). For the second alignment, the sequences of *Preplasmiviricota*, Polintons, and unclassified DJR MCP viruses were aligned within family frameworks, and then we aligned each of these separately prealigned blocks with the prealigned NCLDV sequences (S86 Fig, S82 Align). In both cases, the unclassified viruses and *Preplasmiviricota* formed a clade, which speaks in favor of monophyletic origin of their DJR MCPs. However, in the case where all candidates were aligned initially, the clade of candidates was localized sisterly to mimiviruses (S85 Fig). Meanwhile, topology based on alternative alignment localized the clade of candidates sisterly to iridoviruses (S86 Fig). We removed the corresponding sister clades from alignments and reconstructed trees (without *Mimiviridae* – S87 Fig, S83 Align; without *Iridoviridae* – S88 Fig, S84 Align). In the case when the *Adenoviridae*/Polintons clade was sister to *Mimiviridae* (S85 Fig), the removal of mimiviruses resulted in a change in the expected position of the *Adenoviridae*/Polintons clade (S87 Fig), i.e. they became localized sisterly to iridoviruses. In contrast, in the case when the *Adenoviridae*/Polintons clade was sister to *Iridoviridae*

(S86 Fig), the removal of iridoviruses did not lead to a change in the expected position of the *Adenoviridae*/Polintons clade (S88 Fig). Thus, we tend to trust the S82 Alignment and its derived topology, since the S81 Alignment exhibits some level of homoplasies. In the next step, based on the pairwise Z score comparisons, we calculated the dendrogram of the full-length DJR MCP proteins of NCLDV, *Lavidaviridae*, and *Adenoviridae* using the I-TASSER predictions, as well as the crystallographic structures deposited in the PDB (Fig 13). According to the DALI dendrogram topology, *Lavidaviridae*, *Adenoviridae*, and Polintons form a monophyletic clade, which is consistent with the presented above DJR MCP phylogenies, as well as with topologies of other proteins. Moreover, the *Adenoviridae*/Polintons clade, consisting of the PDB crystallographic structures and our I-TASSER predictions, forms a sister clade to the MAPIM clade which speaks in favor of the topology shown in S86 Fig. We also note that dendrograms calculated only based on the I-TASSER predictions or only based on the PDB crystallographic structures gave the same direction of evolution as the combined dendrogram shown in Fig 13 (S60-61 Data).

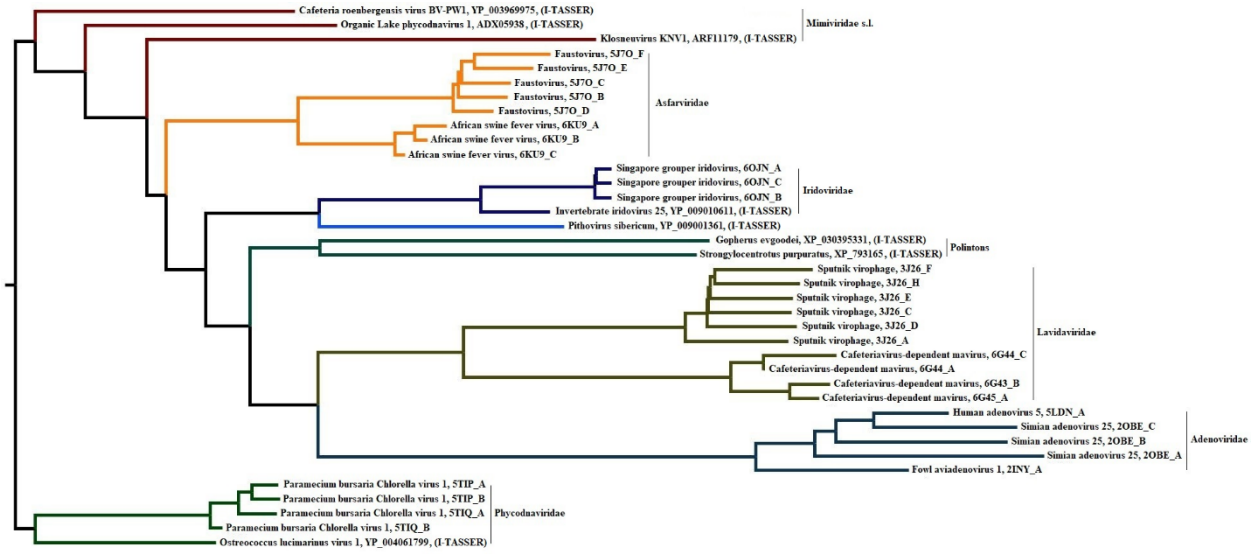


Fig 13. The NCLDV, *Lavidaviridae*, and *Adenoviridae* DJR MCPs dendrogram. The dendrogram based on the pairwise Z score comparisons calculated using DALI. Protein structure prediction was done with I-TASSER [48-50]. The dendrogram was rooted between *Phycodnaviridae* and the rest of clades according to the topologies of all individual phylogenies containing the cellular and NCLDV clades presented in this work. GenBank accessions are given for the sequences used for the I-TASSER predictions, as well as for the crystallographic structures deposited in the PDB. The I-TASSER prediction was carried out without removing of any amino acid residue within the sequences.

Discussion

To provide a more consistent phylogenetic basis to taxonomy it is necessary to identify the monophyletic clades and their hierarchical relationships. In particular, taxonomy of *Bamfordvirae* should be based on the identified branching order of monophyletic clades that form this kingdom. In the context of this idea, NCLDV is a good entry point into unraveling the branching order of *Bamfordvirae* since genomes of NCLDV contain a more than a dozen of nearly-omnipresent core genes that have homologues in eukaryotes and bacteria which can be used as outgroups [2, 7-8].

The core genes could have been transferred from the lineage leading to modern eukaryotes or/and from the different protoeukaryotic lineages that do not exist now, as well as from bacterial lineages. These events may have occurred over multiple times, potentially leading to the formation of the heterogenic plesiomorphic set of signals within the genome of the NCLDV last common ancestor. Possibly heterogeneity of NCLDV plesiomorphic signals increased the likelihood of homoplasies between the clades of modern NCLDV and the eukaryotic or bacterial outgroups we used, especially given the more rapid accumulation of substitutions in viral lines (long branches). There are two possibilities here. Potentially, we could get conflicting evolutionary trajectories of NCLDV for the

phylogenies and dendrograms of various individual proteins, as well as a unified one. However, we confidently obtained the realization of the second probability, that is, all phylogenies and dendrograms of individual core proteins of NCLDV gave the same direction of evolution. This means that in the conflict between synapomorphies and homoplasies, the contribution of synapomorphic signals of NCLDV for the reconstruction of the evolutionary trajectory of this group of giant viruses proved to be critically decisive.

The consistency and repeatability of our results was due, among other things, to the taxonomically widest use of outgroups [38, 52]. In order to identify potential homoplasies affecting the topological relationships among the NCLDV clades we applied a number of approaches identifying the effects of LBA as well as approaches that eliminates this effect described in the literature [42]. Firstly, we built supermatrixes consisting of core NCLDV proteins of protoeukaryotic and bacterial origin phylogenies of which showed the same direction of evolution of NCLDV. A comparative analysis of the seven supertree topologies did not reveal the influence of outgroups on the branching order within NCLDV. The same direction of evolution of NCLDV and relationships among clades within NCLDV were obtained using the supertree and superdendrogram. Moreover, the same direction of evolution of NCLDV and the same phylogenetic relationships within NCLDV were obtained by comparing the topologies of individual protein phylogenies of both protoeukaryotic and bacterial origin. Secondly, the same phylogenetic relationships between the AP and MAPIM clades, as well as between the AP + MAPIM and *Mimiviridae* s.l. clades were obtained using extremely closely related viral homologues of the D6/D11 helicases as an outgroup. Thirdly, the topologies of outgroups were examined separately before being combined with NCLDV sequences. They were checked for the identification of horizontal transfers and homoplasies, that is, for the preservation of monophyly of clades constituting outgroup, as well as for compliance with the phylogenies given in the literature. Fourth, in order to test the stability of the topological relationships among the NCLDV clades of individual

protein trees, we changed the composition of outgroups and constructed the same trees without outgroup [36, 42]. Fifth, dendrograms of individual core proteins where outgroups were used, as well as superdendrogram showed the same direction of evolution of NCLDV for major clades as individual phylogenies, concatenated trees, and supertree (first bifurcation between the *Phycodnaviridae* s.l. clade and the joint *Mimiviridae* s.l. + MAPIM + AP clade, as well as sister relationships between the *Mimiviridae* s.l. clade and the joint MAPIM + AP clade). All the approaches described above ultimately showed invariance the position of the root for individual trees, dendrograms, supertree, superdendrogram, and concatenated trees, as well as invariance of phylogenetic relationships of clades within NCLDV.

Our study revealed that in the vast majority of the individual phylogenies, as well as on concatenated trees and supertree phycodnaviruses s.l. resolved as a sister to the clade consisting of the *Mimiviridae* s.l. + MAPIM +AP clades (hereinafter the MAPIMMAP clade). The exceptions were the phylogenies of SFIII helicases and Uracil DNA glycosylases, where proteins of phycodnaviruses s.l. were separated from NCLDV by the bacterial or prophage homologues. The scarcity of alternative scenarios suggests that the first bifurcation of LCA of NCLDV must have led to the emergence of two branches. The first branch led to *Phycodnaviridae* s.l. including pandoraviruses, molliviruses and medusaviruses. The second branch led to the LCA of the MAPIMMAP clade, which, in turn, is branched into the MAPIM and AP clades (Figs 6-8). This branching order contradicts with the accepted ICTV taxonomy based on the data presented in the literature [2, 7, 9-10]. The taxonomy of ICTV is fully consistent with the branching order of NCLDV presented in 2019 [10]. Koonin and Yutin give arguments justifying the branching order of the NCLDV, according to which NCLDV consists of three clades, i.e., *Mimiviridae* + *Phycodnaviridae* (branch 1), MAPIM (branch 2), and AP (branch 3) [10]. Their concatenated tree was constructed without using of cellular orthologues and technically unrooted. However, the choice of branching order is based on two arguments - the greater

similarity of the core protein sequences of branches 1 and 2 (combined into the class *Megaviricetes*) in comparison with branch 3 (the class *Pokkesviricetes*) and on the topologies of the phylogenetic trees of individual proteins, in particular, helicase-primase, as well as DNA and RNA polymerases, even though individual protein trees gave conflicting topologies [8]. Phylogenies presented by other authors also included only viral sequences representing unrooted trees [9, 11, 53]. In fact, our phylogenetic data obtained using cellular outgroups refutes the most important argument about the greater similarity of the proteins of branches 1 and 2 in comparison with homologues of branch 3, on which the taxonomy of ICTV is based, as well as consistently demonstrate the direction of evolution of NCLDV, in contrast to the previously presented data [8]. Moreover, despite using of bioinformatics data on the evolutionary history of proteins, our dendrograms and superdendrogram based on ColabFold predictions also confirmed the similarity of the tertiary structures of the AP (branch 3 according to Koonin and Yutin 2019) and MAPIM (branch 2 according to Koonin and Yutin 2019) clades.

The main new set of observations from our studies is that, first, phycodnaviruses s.l. consistently resolve as the basal clade in relation to the all crown NCLDV clades and, second, the AP clade was always localized within NCLDV. Moreover, according to our results, with the exception of only two phylogenies (the PolB alignments made using the MUSCLE algorithm and Metallopeptidase WLM) the AP clade is consistently resolved as a sister to the MAPIM clade. While the MAFFT algorithm localized PolBs of the AP clade within the MAPIM clade. Of particular importance is the result obtained through the D6/D11 helicases, since the same relationships among the *Mimiviridae* s.l., MAPIM, and AP clades were obtained using as outgroup the extremely closely related viral D6/D11 homologues. Our results suggest that, in order to reconcile phylogeny and taxonomy, the class *Pokkesviricetes* should be abolished, and the families *Asfarviridae* s.l. and *Poxviridae* should be included in the class *Megaviricetes*.

The monophyletic origin of the *Adenoviridae*/*Lavidaviridae*/*Adintoviridae*/Polintons/CLP/MLP lineage (abbreviated as *Adenoviridae* lineage) supported by the phylogenies of the replication module proteins, as well as the mRNA capping enzyme and morphogenetic module proteins. Moreover, in all these phylogenies and dendrograms, the *Adenoviridae* lineage is found as a clade within NCLDV. The formation of the FtsK-HerA superfamily ATPase, DJR MCP, D5-like primase, D6/D11 helicase, SFIII helicase, PolB and mRNA capping enzyme clades by the *Adenoviridae* lineage and localization of these clades within NCLDV casts doubt on the monophyly of *Preplasmiviricota* and suggests that the viruses of the *Adenoviridae* lineage should be placed within the phylum *Nucleocytoviricota*. An apparent exception is the RNAP β subunit of CLPs, localizing outside of NCLDV and vice versa, the PolB, D6/D11 helicase and mRNA capping enzyme proteins of CLPs are localized within NCLDV. The CLP and MLP clades localization within NCLDV suggest that unlike the strategies of vertical inheritance of genetic information (viruses) and integration into the host genome (the *Adintoviridae*/Polintons group) the ancestor of the CLPs and MLPs could be a virus related to NCLDV, which chose the third, intermediate strategy - life in the cytoplasm and mitochondrial matrix.

The PolB, SFIII helicase, and ASFV F1055L homologues of herpesviruses tend to be nested within NCLDV. A possible explanation for these topologies could involve simultaneous infection of eukaryotic cells by NCLDV and ancient herpesviruses, which could have led to horizontal transfer of genes of the replication module from NCLDV to herpesviruses. However, the some differences in the topological localization of these proteins within NCLDV likely indicate that these hypothetical horizontal transfers have occurred more than once, and gene donors belonged to different clades of NCLDV.

There have been four competing scenarios for the origin of viruses. The first hypothesis postulates the origin of viruses from precellular genetic elements. According to the second hypothesis (‘regression’ or ‘reduction’ hypothesis), viruses evolved from cellular life forms through the reduction of genomes.

The third hypothesis involves the escape of genes from cellular hosts and the acquisition of partial replicative autonomy. The fourth "chimeric" hypothesis has been recently proposed and suggests that the replicative modules of different viral lineages may be of primordial, precellular origin, while capsid proteins have been captured from cellular organisms in different evolutionary periods multiple times and independently [6, 54]. There also may be combinations of some or all of those four scenarios. Our results do not refute the "chimeric" hypothesis about the primordial origin of the replication module, in particular, for monophyletic *Riboviria* [55]. It is also possible that the substantial portions of replicative modules of *Bamfordvirae* and *Herpesviridae* were of primordial origin. However, it appears that some of them were replaced with the corresponding host genes, as in the case of *Bamfordvirae*, where we observe the presence of eukaryotic xenologues encoded by genomes of eukaryotic viruses. At the same time, our results show that three genes of the replicative module in *Herpesviridae* originate from within NCLDV. Likewise, one of the key genes of the morphogenetic module of *Bamfordvirae* the FtsK-HerA superfamily ATPases had been clearly recruited from bacterial homologues.

Finally, the phylogenies and dendrograms of the *Adenoviridae* lineage presented and discussed here disprove the possibility of NCLDV origin from the *Adintoviridae*/Polinton-like viruses and vice versa, our data supports the hypothesis of the origin of NCLDV from prokaryotic viruses with small genomes proposed by Yutin and colleagues [3].

Materials and Methods

General information

In this study, we used 18 core proteins of NCLDV, as defined by Yutin and colleagues and Aylward and colleagues [7, 53]. Correspondences of the names of the used proteins, NCVOGs and GVOGs are

given in the S4 Table. These proteins, together with their homologues from the *Preplasmiviricota* phylum, were collected from approximately 650 genomes belonging (or tentatively assigned) to realm *Varidnaviria* and related mobile elements, as well as from 109 genomes of *Herpesvirales*, *Nudiviridae*, *Baculoviridae*, *Nimaviridae*, and *Hytrosaviridae*. The overview of this data set is given in the S5 Table.

Sources of amino acid sequences and gene prediction

Amino acid sequences were downloaded from the NCBI non-redundant protein sequences (nr) database [56]. The genomes of Polintons were downloaded from the RepBase database [57-58]. The selection of Polintons genomes was made from the 20181026 release. The prediction of the Polintons genes has been carried out using the Heuristic Approach for Gene Prediction web server http://exon.gatech.edu/GeneMark/heuristic_gmhmmp.cgi [59-60]. For further work, we selected Polintons, whose genomes encode the PolB, FtsK-HerA superfamily ATPase, and DJR MCP genes or at least PolB and one of the latter two genes (S6 Table).

Sequence selection

In most families of NCLDV the majority of the core NCVOG (GVOG) representatives are single-copy (see S5 Table and [61]). In addition, within families (clades) with multiple copies of a gene, there may be genera whose genomes encode only one homologue. We used single-copy homologues of NCLDVs as queries and performed the blastp (protein-protein BLAST) and PSI-BLAST (Position Specific Iterated BLAST) searches against the nr database [62-63]. For each search, we entered taxon ID and downloaded only subjects with the highest max score for each species (E value < 0.005). Then, the same procedure was carried out with the received objects, thus maximally covering the entire possible range of subjects. Next, we constructed preliminary LG+F+R4 trees, which were examined for detecting violations of monophyly of established viral clades (nesting of a clade or viral sequence

with known belonging to certain clade within a phylogenetically distinct clade) and removed all sequences violating monophyly of any viral clade. Thus, we removed the results of potential horizontal transfers or homoplasies from further analysis. If a taxon had two or more homologues that did not violate the monophyly of other viral clades, we examined alignments to detect plesiomorphies and apomorphies among the viral homologues and outgroup under study, and retained the viral homologue in which plesiomorphies were not substituted, and apomorphies of the family (clade) were best represented. In some cases, however, the topological relationships of homologues were of interest, and we retained both or more homologues in alignment (for instance, the D6/D11 helicases). Additional information about distantly related homologues was obtained by searching the protein domain databases, such as PFAM, using relevant keywords. The list of GenBank accessions used in this study is presented in the S7 Table.

The initial pool of cellular homologues was selected by PSI-BLAST searches against the nr database where queries were the NCLDV sequences. We selected the cellular subjects with the highest max scores and used them as queries. After the accumulation of a redundant pool of cellular homologues, we treated it in the same way as described for NCLDV above, starting from the stage of construction of preliminary trees. In the cases of DNA and RNA polymerases, as well as FtsK-HerA superfamily ATPases, as queries were used not only the sequences identified as a result of seeding of the NCLDV sequences, but also sequences presented in the literature.

Multiple sequence alignments

Multiple sequence alignments were generated using the MEGA7 software package [64] by the MUSCLE [65] algorithms. The MUSCLE alignments errors were minimized by using anchor substitutions reported in the literature and by identifying conserved sequence blocks [22, 46]. In order

to verify topologies of some key trees obtained using MUSCLE, we used the MAFFT online 7.47 version available in https://www.ebi.ac.uk/Tools/msa/mafft/ [66-67].

We employed multi-step alignment strategy. A) At first, viral genera, families or genera/family-level clades were aligned using the gap open 2 and gap extend 0 or 1 options (MUSCLE). At the next step, we removed all positions in which more than 10% of the sequences had gaps. Further, the alignment process was repeated using the gap open 1 and gap extend 0 options (MUSCLE). Then, alignment positions with more than 10% gaps were removed. B) To create NCLDV alignments, ready aligned genera/family blocks were combined and the above described alignment procedure was repeated. C) To create alignment blocks of cellular organisms, the same approach was used, except that instead of genera and families the archaeal and bacterial phyla or the unranked *Animalia* (*ParaHoxozoa*, *Bilateria*), *Fungi*, *Oomycota*, *Archaeplastida*, and *Kinetoplastida* eukaryotic clades were initially aligned. Then prokaryotic phyla or eukaryotic clades were combined and aligned. D) Finally, depending on the goals of the both final MUSCLE and MAFFT alignments, the pre-aligned viral and cellular blocks were combined and aligned. For the final MUSCLE alignments building we used the gap open 2 or 3 and gap extend 0 or 1 options. For the MAFFT final alignments building we used matrix none or BLOSUM30, gap open 1 and gap extension 0.0001penalties. The choice of the MAFFT matrix was carried out depending on the preservation of anchor amino acid substitutions of alignments, published in the literature. For alignments of both algorithms, positions with more than 40% gaps were removed.

The supermatrixes were built by head-to-tail linking of already aligned (MUSCLE) individual protein alignments. In concatenated alignments, in contrast to individual protein alignments, due to the absence of some proteins in some taxa, gaps were not removed.

Phylogenetic analysis

Best-fit general amino acid replacement matrix models were selected from the common model list using the IQ-TREE web server (http://iqtree.cibiv.univie.ac.at/), as well as the IQ TREE program (version 1.6.12) [68-69]. The Bayesian selection criterion (BIC) was used by default. Single-protein and concatenated protein phylogenetic trees were constructed within the maximum likelihood (ML) framework using the IQ TREE program (version 1.6.12) [68]. Sets of trees based on all single-protein and concatenated protein alignments were constructed under the LG, LG4X, LG4M, LG+C20, EX2, EX3, EHO, EX_EHO, UL2, and UL3 (+ accessories) models. For statistical support of the partitions on the trees obtained by IQ TREE, we used four tests, i.e., the SH-aLRT branch, approximate Bayes (aBayes), ultrafast bootstrap (Ultrafast), and standard bootstrap [70-74]. The default parameters were applied in all three tests, and the number of replicates for the standard bootstrap test was set to 100.

Sets of trees obtained from all individual and concatenated protein alignments were subjected to topology testing using the TreeFinder program under the LG+G4 model [75]. We applied the following tests: ELW, BP, SH, WSH, AU and likelihood [70, 76-78]. The results of topological tests of individual and concatenated proteins are shown in S8-27 Tables. The best AU or AU/likelihood trees were used for visualization and for supertree construction.

Supermatrix analysis

We obtained seven supermatrixes by concatenating the alignments, the protein compositions of which are shown in S3 Tables. Before concatenation, preliminary LG+F+R4 trees of individual core proteins with and without outgroups were constructed for each viral protein. All phylogenetic trees of individual proteins were inspected for monophyly of clades. Any clade or sequence violating monophyly of another clade (nesting of a sequence or clade within a phylogenetically distinct clade) was removed from the further analysis. Information about constructing of concatenated trees, as well as testing topologies, is given above.

Supertree construction

To construct supertrees, we constructed the set of input bootstrapped trees for 13 protein families (total 23 trees) substitution matrices of which corresponded to the best AU values (S25 Table). In order to avoid obvious errors caused by HGT and homoplasies into supertrees, for construction of input bootstrapped trees we used the same condition as in case of construction of concatenated supermatrix, i.e., preservation of monophyly of well-established clades. For construction of bootstrapped input trees we limited the taxonomic composition of the sequences used to 61 species (52 viral and 9 eukaryotic). To construct supertrees, we used the subtree prune-and-regraft distance (SPR supertrees) method. We used SPR Supertree 1.2.1 downloaded version [79]. We constructed a supertree with bootstrap thresholds equal to 80 and the -allow_multi and -i 1000 options. The rest of the options were used by default.

Determination of the root position

The root position of the individual protein trees and dendrograms, concatenated trees, supertree, and superdendrogram could be pinpointed due to the presence of eukaryotic homologues, as well as their viral homologues (PolB, DNA-directed RNA polymerase subunit α (RNAP α), DNA-directed RNA polymerase subunit β (RNAP β), mRNA capping enzyme, Topoisomerase II, D5-like primase, PCNA, TFIIB, and Metallopeptidase WLM). For many of these viral genes have been shown to be of eukaryotic (protoeukaryotic) or eukaryotic (protoeukaryotic)/archaeal origin [2]. Similar, the root position of the NCLDV clade could be pinpointed for proteins (genes) of viral (for instance between the D6/D11 helicase orthologues), bacterial or bacteriophage/plasmid origin (FtsK-HerA superfamily ATPase, AP (apurinic) endonuclease family 2, Holliday junction resolvase (RuvC), Uracil-DNA glycosylase, and Superfamily III (SFIII) helicase) [2, 8, 46]. Of the 14 proteins (genes) of cellular origin (S5 Table), 9 are of eukaryotic/protoeukaryotic origin, and 5 are of bacterial origin. We rooted

these trees in the following way. Trees containing the multifamily and multiple type proteins (PolB, RNAP α, RNAP β, HerA-FtsK ATPase) were rooted according to the branching order of the cellular clades (*Archaea*, *Bacteria*, and *Eukaryota*) as in published data [22, 39-40, 46-47, 80]. The remaining trees with non-multifamily and non-multiple type proteins rooted according to the branching order of cellular clades presented in the literature if clades of more than one cellular domain were present. Alternatively, trees, dendrograms, concatenated trees, supertree, and superdendrogram (topologies with only one clade as outgroup) were rooted between the viral and cellular clades. Topologies where the outgroup was absent were rooted according to the topologies of the corresponding proteins where the outgroup was present. Topologies of synapomorphic proteins were rooted according to the branching orders of NCLDV identified for trees with outgroups.

Identification and elimination of potential homoplasies are described in the Results section for each individual case. The relevance of using of outgroups is discussed in the Discussion section.

Prediction of three-dimensional structures of proteins using ColabFold

With the exception of the DJR MCP sequences, the sequences of interest were aligned using the MUSCLE5 software using the mpc algorithm and the permutation none, perturbation 0 options [81]. In order to exclude from further analysis spatial fluctuations that are not typical for most of proteins of given alignment and that could potentially affect the results of the analysis we removed from the alignment fragments between constant blocks, formed by single or/and unaligned sequences, as well as aligned sequences characterizing a genus or subfamily if other members of alignment also had a gap in this area. Thus, we obtained presumably functionally active standardized fragments of the minimum size with a length of ≈ 80 - 600 amino acids.

To predict the three-dimensional structures of protein fragments, we used the online version of the ColabFold software, which speed-ups the three-dimensional prediction of protein structures by

replacing the AlphaFold2's input feature generation stage with a fast MMseqs2 (https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb) [43-45, 82]. The options were used by default. Initially we selected from the redundant set of predictions a subset of predictions consisting of individual predictions of the unrelaxed rank 1 with the pLDDT > 80 for an entire prediction. The length and quality of predicted fragments was controlled using the output predicted aligned error and predicted LDDT per position plots. In order to exclude from the analysis amino acid regions that give an unreliable prediction (pLDDT < 70) we established the following criterion. In the predicted structures, the total length of regions with pLDDT < 70 embedded in the regions with the pLDDT > 70 had not to exceed 20 % of the total length of prediction, otherwise the alignment was considered unsuitable for prediction. This criterion was applied to individual sets of predictions of the *Eukaryota*, *Bacteria*, and NCLDV clades, as well as to the *Preplasmiviricota* families, but not to individual predictions, since this criterion refers to the applicability of the alignment. If amino acid region with the pLDDT < 70 (for more than half of family/clade members), with a length exceeding 20% of the total length of prediction, flanked the prediction, then such amino acid region was removed from the alignment. If such a region fell inside the alignment, then we built an alignment of another fragment of the same protein. Then, the predictions were subjected to the exhaustive PDB25 search (S1 Table). All predictions intended for dendrograms calculation had Z-scores > 2 [83-84]. Used in this work the unrelaxed rank 1 predictions (.PDB files), corresponding GenBank accessions, corresponding RepBase codes of Polintons as well as alignments (.fas) of predicted fragments are given in S62-84 Data.

Prediction of full-size three-dimensional structures of DJR MCPs using I-TASSER

We used a threading method implemented in the I-TASSER (Iterative Threading ASSEmbly Refinement) program for full-size DJR MCP structures prediction (https://zhanggroup.org/I-TASSER/about.html) [48-50]. Predicted structures, as well as corresponding GenBank accessions are given in S85-92 Data.

Dendrograms and superdendrogram calculation

Structural comparisons to the Protein Data Bank were performed using the PDB25 algorithm through the DALI server [83-84]. Structural similarities were evaluated based on the DALI Z scores [85], and dendrograms were calculated using the DALI server. Z scores of dendrograms calculated using DALI shown in S2 Table. Based on 12 ColabFold derived dendrograms of 9 proteins we constructed a SPR superdendrogram using the DALI server (see below). For calculation of a superdendrogram we used the –reroot and -i 10000 options. The rest of the options were used by default.

Visualization

The phylogenetic trees, dendrograms, concatenated trees, supertree, and superdendrogram were rooted and visualized with FigTree v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/). Phylogenetic trees and concatenated trees (Figures and S Figures with extension .treefile), as well as dendrograms, superdendrogram, and supertree (with extension .nwk) are presented in S93-202 Data.

Acknowledgments

We are grateful to Dr. Arcady Mushegian for many helpful discussions in the course of this project.

Author Contributions

AK – conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, software, supervision, validation, visualization, original draft preparation, review and editing. HS – formal analysis, resources, software, validation, original draft preparation, review and editing. HZ – conceptualization, project administration, supervision, original draft preparation, review and editing.

Funding information

No funding

Supporting Information

Supporting information can be found at: https://zenodo.org/record/6579602#.Yo30fqhByUl

DOI: 10.5281/zenodo.6579602

References

1. Iyer LM, Aravind L, Koonin EV. Common Origin of Four Diverse Families of Large Eukaryotic DNA Viruses. Journal of Virology. 2001;75:11720–11734. PMID: 11689653. doi: 10.1128/JVI.75.23.11720-11734.2001
2. Koonin EV, Yutin N. Origin and evolution of eukaryotic large nucleo-cytoplasmic DNA viruses. Intervirology. 2010;53:284–292. PMID: 20551680. doi: 10.1159/000312913
3. Yutin N, Wolf YI, Koonin EV. Origin of giant viruses from smaller DNA viruses not from a fourth domain of cellular life. Virology. 2014;466-467:38–52. PMID: 25042053. doi: 10.1016/j.virol.2014.06.032

4. Koonin EV, Krupovic M, Yutin N. Evolution of double-stranded DNA viruses of eukaryotes: from bacteriophages to transposons to giant viruses. Annals of the New York Academy of Sciences. 2015;1341:10–24. PMID: 25727355. doi: 10.1111/nyas.12728
5. Krupovic M, Koonin EV. Polintons: a hotbed of eukaryotic virus, transposon and plasmid evolution. Nature reviews microbiology. 2015;13:105–115. PMID: 25534808. doi: 10.1038/nrmicro3389
6. Krupovic M, Yutin N, Koonin EV. Evolution of a major virion protein of the giant pandoraviruses from an inactivated bacterial glycoside hydrolase. Virus Evolution. 2020;6:veaa059. PMID: 33686356. doi: org/10.1093/ve/veaa059
7. Yutin N, Wolf YI, Raoult D, Koonin EV. Eukaryotic large nucleo-cytoplasmic DNA viruses: clusters of orthologous genes and reconstruction of viral genome evolution. Virology Journal. 2009;6:223. PMID: 20017929. doi: 10.1186/1743-422X-6-223
8. Yutin N, Koonin EV. Hidden evolutionary complexity of nucleo-cytoplasmic large DNA viruses of eukaryotes. Virology Journal. 2012;9:161. PMID: 22891861. doi: 10.1186/1743-422X-9-161
9. Guglielmini J, Woo AC, Krupovic M, Forterre P, Gaia M. Diversification of giant and large eukaryotic dsDNA viruses predated the origin of modern eukaryotes. Proceedings of the National Academy of Sciences of the United States of America. 2019;116:19585–19592. PMID: 31506349. doi: 10.1073/pnas.1912006116
10. Koonin EV, Yutin N. Evolution of the large nucleocytoplasmic DNA viruses of eukaryotes and convergent origins of viral gigantism. Advances in Virus Research. 2019;103:167–202. PMID: 30635076. doi: 10.1016/bs.aivir.2018.09.002
11. Woo AC, Gaia M, Guglielmini J, Da Cunha V, Forterre P. Phylogeny of the Varidnaviria morphogenesis module: congruence and incongruence with the tree of life and viral taxonomy.

Frontiers in Microbiology. 2021;12:704052. PMID: 34349745. doi: org/10.3389/fmicb.2021.704052

12. Boratto PVM, Oliveira GP, Machado TB, Andrade AC SP, Baudoin J-P, Klose T, et al. Yaravirus: A novel 80-nm virus infecting Acanthamoeba castellanii. Proceedings of the National Academy of Sciences of the United States of America. 2020;117:16579–16586. PMID: 32601223. doi: 10.1073/pnas.2001637117
13. Starrett GJ, Tisza MJ, Welch NL, Belford AK, Peretti A, Pastrana DV, et al. Adintoviruses: a proposed animal-tropic family of midsize eukaryotic linear dsDNA (MELD) viruses. Virus Evolution. 2021;7:veaa055. PMID: 34646575. doi: 10.1093/ve/veaa055
14. Griffiths AJ. Natural plasmids of filamentous fungi. Microbiological Reviews. 1995;59:673–685. PMID: 8531891. doi: 10.1128/mr.59.4.673-685.1995
15. Kapitonov VV, Jurka J. Self-synthesizing DNA transposons in eukaryotes. Proceedings of the National Academy of Sciences of the United States of America. 2006;103:4540–4545. PMID: 16537396. doi: org/10.1073/pnas.0600833103
16. Pritham EJ, Putliwala T, Feschotte C. Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. Gene. 2007;390:3–17. PMID: 17034960. doi: 10.1016/j.gene.2006.08.008
17. Handa H. Linear plasmids in plant mitochondria: peaceful coexistences or malicious invasions? Mitochondrion. 2008;8:15–25. PMID: 18326073. doi: 10.1016/j.mito.2007.10.002
18. Yutin N, Raoult D, Koonin EV. Virophages, polintons, and transpovirons: a complex evolutionary network of diverse selfish genetic elements with different reproduction strategies. Virology Journal. 2013;10:158. PMID: 23701946. doi: 10.1186/1743-422X-10-158

19. Koonin EV, Krupovic M. Polintons, virophages and transpovirons: a tangled web linking viruses, transposons and immunity. Current Opinion in Virology. 2017;25:7–15. PMID: 28672161. doi: 10.1016/j.coviro.2017.06.008
20. Boyer M, Madoui M-A, Gimenez G, La Scola B, Raoult D. Phylogenetic and phyletic studies of informational genes in genomes highlight existence of a 4 domain of life including giant viruses. PLoS One. 2010;5:e15530. PMID: 21151962. doi: 10.1371/journal.pone.0015530
21. Mushegian A, Karin EL, Pupko T. Sequence analysis of malacoherpesvirus proteins: Pan-herpesvirus capsid module and replication enzymes with an ancient connection to "Megavirales". Virology. 2018;513:114–128. PMID: 29065352. doi: 10.1016/j.virol.2017.10.009
22. Kazlauskas D, Krupovic M, Guglielmini J, Forterre P, Venclovas Č. Diversity and evolution of B-family DNA polymerases. Nucleic Acids Research. 2020;48:10142–10156. PMID: 32976577. doi: 10.1093/nar/gkaa760
23. Kunkel TA, Burgers PM. Dividing the Workload at a Eukaryotic Replication Fork. Trends in Cell Biology. 2008;18:521–527. PMID: 18824354. doi: 10.1016/j.tcb.2008.08.005
24. Pursell ZF, Isoz I, Lundström E-B, Johansson E, Kunkel TA. Yeast DNA polymerase epsilon participates in leading-strand DNA replication. Science. 2007;317:127–130. PMID: 17615360. doi: 10.1126/science.1144067
25. Gan GN, Wittschieben JP, Wittschieben B Ø, Wood RD. DNA Polymerase Zeta (Pol Zeta) in Higher Eukaryotes. Cell Research. 2008;18:174–183. PMID: 18157155. doi: 10.1038/cr.2007.117
26. Pavlov YI, Shcherbakova PV. DNA Polymerases at the Eukaryotic Fork-20 years Later. Mutation Research. 2010;685:45–53. PMID: 19682465. doi: org/10.1016/j.mrfmmm.2009.08.002

27. Lujan SA, Williams JS, Kunkel TA. DNA Polymerases Divide the Labor of Genome Replication. Trends in Cell Biology. 2016;26:640–654. PMID: 27262731. doi: 10.1016/j.tcb.2016.04.012
28. Louis BG, Fitt PS. Nucleic Acid Enzymology of Extremely Halophilic Bacteria. Halobacterium cutirubrum Deoxyribonucleic Acid-Dependent Ribonucleic Acid Polymerase. The Biochemical Journal. 1971;121:621–627. PMID: 4940048. doi: 10.1042/bj1210621
29. Willis IM. RNA Polymerase III. Genes, Factors and Transcriptional Specificity. European Journal of Biochemistry. 1993;212:1–11. PMID: 25840332. doi: 10.1111/j.1432-1033.1993.tb17626.x
30. Grummt I. Regulation of mammalian ribosomal gene transcription by RNA polymerase I. Progress in Nucleic Acid Research and Molecular Biology. 1999;62:109–154. PMID: 9932453. doi: 10.1016/s0079-6603(08)60506-1
31. Ebright RH. RNA Polymerase: Structural Similarities Between Bacterial RNA Polymerase and Eukaryotic RNA Polymerase II. Journal of Molecular Biology. 2000;304:687–698. PMID: 11124018. doi: 10.1006/jmbi.2000.4309
32. Lee Y, Kim M, Han J, Yeom K-H, Lee S, Baek SH, et al. MicroRNA Genes are Transcribed by RNA Polymerase II. The EMBO Journal. 2004;23:4051–4060. PMID: 15372072. doi: 10.1038/sj.emboj.7600385
33. Herr AJ, Jensen MB, Dalmay T, Baulcombe DC. RNA Polymerase IV Directs Silencing of Endogenous DNA. Science. 2005;308:118–120. PMID: 15692015. doi: 10.1126/science.1106910
34. Wierzbicki AT, Ream TS, Haag JR, Pikaard CS. RNA Polymerase V Transcription Guides ARGONAUTE4 to Chromatin. Nature Genetics. 2009;41:630–640. PMID: 19377477. doi: 10.1038/ng.365

35. Giribet G, Ribera C. The Position of Arthropods in the Animal Kingdom: A Search for a Reliable Outgroup for Internal Arthropod Phylogeny. Molecular Phylogenetics and Evolution. 1997:9:481–488. PMID: 9667996. doi: 10.1006/mpev.1998.0494
36. Barriel V, Tassy P. Rooting with Multiple Outgroups: Consensus Versus Parsimony. Cladistics. 1998;14:193–200. PMID: 34902930. doi: 10.1111/j.1096-0031.1998.tb00332.x
37. De la Torre-Bárcena JE, Kolokotronis S-O, Lee EK, Stevenson DW, Brenner ED, Katari MS, et al. The Impact of Outgroup Choice and Missing Data on Major Seed Plant Phylogenetics Using Genome-Wide EST Data. PLoS One. 2009;4:e5764. PMID: 19503618. doi: org/10.1371/journal.pone.0005764
38. Wilberg EW. What's in an Outgroup? The Impact of Outgroup Choice on the Phylogenetic Position of Thalattosuchia (Crocodylomorpha) and the Origin of Crocodyliformes. Systematic Biology. 2015;64:621–637. PMID: 25840332. doi: org/10.1093/sysbio/syv020
39. Makarova KS, Krupovic M, Koonin EV. Evolution of Replicative DNA Polymerases in Archaea and Their Contributions to the Eukaryotic Replication Machinery. Frontiers in Microbiology. 2014;5:354. PMID: 25101062. doi: org/10.3389/fmicb.2014.00354
40. Zaremba-Niedzwiedzka K, Caceres EF, Saw JH, Bäckström D, Juzokaite L, Vancaester E, et al. Asgard archaea Illuminate the Origin of Eukaryotic Cellular Complexity. Nature. 2017;541:353–358. PMID: 28077874. doi: 10.1038/nature21031
41. Brown JR, Doolittle WF. Archaea and the Prokaryote-to-Eukaryote Transition. Microbiology and Molecular Biology Reviews. 1997;61:456–502. PMID: 9409149. doi: 10.1128/mmbr.61.4.456-502.1997
42. Bergsten JA. Review of Long-Branch Attraction. Cladistics. 2005;21:163–193. PMID: 34892859. doi: org/10.1111/j.1096-0031.2005.00059.x

43. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596:583–589. PMID: 34265844. doi: org/10.1038/s41586-021-03819-2

44. Mirdita M, Steinegger M, Söding J. MMseqs2 desktop and local web server app for fast, interactive sequence searches. Bioinformatics. 2019;35:2856–2858. PMID: 30615063. doi: org/10.1093/bioinformatics/bty1057

45. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold - Making protein folding accessible to all. BioRxiv [Preprint]. [posted 2022 February 08; cited 2022 May 07]: [8 p.]. Available from: https://www.biorxiv.org/content/10.1101/2021.08.15.456425v3 doi.org/10.1101/2021.08.15.456425

46. Iyer LM, Makarova KS, Koonin EV, Aravind L. Comparative Genomics of the FtsK–HerA Superfamily of Pumping ATPases: Implications for the Origins of Chromosome Segregation, Cell Division and Viral Capsid Packaging. Nucleic Acids Research. 2004;32:5260–5279. PMID: 15466593. doi: 10.1093/nar/gkh828

47. Burroughs AM, Iyer LM, Aravind L. Comparative Genomics and Evolutionary Trajectories of Viral ATP Dependent DNA-Packaging Systems. Genome Dynamics. 2007;3:48–65. PMID: 18753784. doi: 10.1159/000107603

48. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. Nature Protocols. 2010;5:725–738. PMID: 20360767. doi: 10.1038/nprot.2010.5

49. Yang J, Zhang Y. I-TASSER server: new development for protein structure and function predictions. Nucleic Acids Research. 2015;43:W174–W181. PMID: 25883148. doi: 10.1093/nar/gkv342

50. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER Suite: Protein structure and function prediction. Nature Methods. 2015;12:7–8. PMID: 25549265. doi: 10.1038/nmeth.3213.
51. Mitchell MS, Matsuzaki S, Imai S, Rao VB. Sequence Analysis of Bacteriophage T4 DNA Packaging/Terminase Genes 16 and 17 Reveals a Common ATPase Center in the Large Subunit of Viral Terminases. Nucleic Acids Research. 2002;30:4009–4021. PMID: 12235385. doi: 10.1093/nar/gkf524
52. Nixon KC, Carpenter JM. On outgroups. Cladistics. 1994;9:413–426. PMID: 34929985. doi: 10.1111/j.1096-0031.1993.tb00234.x
53. Aylward FO, Moniruzzaman M, Ha AD, Koonin EV. A phylogenomic framework for charting the diversity and evolution of giant viruses. PLOS Biology. 2021;19:e3001430. PMID: 34705818. doi: 10.1371/journal.pbio.3001430
54. Krupovic M, Dolja VV, Koonin EV. Origin of Viruses: Primordial Replicators Recruiting Capsids from Hosts. Nature Reviews Microbiology. 2019;17:449–458. PMID: 31142823. doi: 10.1038/s41579-019-0205-6
55. Wolf YI, Kazlauskas D, Iranzo J, Lucía-Sanz A, Kuhn JH, Krupovic M, et al. Origins and Evolution of the Global RNA Virome. mBio. 2018;9:e02329-18. PMID: 30482837. doi: 10.1128/mBio.02329-18
56. Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Research. 2016;44:D67–D72. PMID: 26590407. doi: org/10.1093/nar/gkv1276
57. Jurka J. Repbase Update: a Database and an Electronic Journal of Repetitive 312 Elements. Trends in Genetics. 2000;16:418–420. PMID: 10973072. doi: 10.1016/s0168-9525(00)02093-x

58. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase Update, a Database of Eukaryotic Repetitive Elements. Cytogenetic and Genome Research. 2005;110:462–467. PMID: 16093699. doi: 10.1159/000084979

59. Besemer J, Borodovsky M. Heuristic approach to deriving models for gene finding. Nucleic Acids Research. 1999;27:3911–3920. PMID: 10481031. doi: org/10.1093/nar/27.19.3911

60. Zhu W, Lomsadze A, Borodovsky M. Ab initio gene identification in metagenomic sequences. Nucleic Acids Research. 2010;38:e132. PMID: 20403810. doi: org/10.1093/nar/gkq275

61. 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:e02938–19. PMID: 31937645. doi: org/10.1128/mBio.02938-19

62. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic Local Alignment Search Tool. Journal of Molecular Biology. 1990;215:403–410. PMID: 2231712. doi: 10.1016/S0022-2836(05)80360-2

63. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a New Generation of Protein Database Search Programs. Nucleic Acids Research. 1997;25:3389–3402. PMID: 9254694. doi: 10.1093/nar/25.17.3389

64. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Molecular Biology and Evolution. 2016;33:1870–1874. PMID: 27004904. doi: 10.1093/molbev/msw054.

65. Edgar RC. MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput. Nucleic Acids Research. 2004;32:1792–1797. PMID: 15034147. doi: org/10.1093/nar/gkh340

66. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform. Nucleic Acids Research. 2002;30:3059–3066. PMID: 12136088. doi: 10.1093/nar/gkf436
67. Madeira F, Park Y, Lee J, Buso N, Gur T, Madhusoodanan N, et al. The EMBL-EBI Search and Sequence Analysis Tools APIs in 2019. Nucleic Acids Research. 2019;47:W636–W641. PMID: 30976793. doi: org/10.1093/nar/gkz268
68. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Molecular Biology and Evolution. 2015;32:268–274. PMID: 25371430. doi: 10.1093/molbev/msu300
69. Trifinopoulos J, Nguyen L-T, von Haeseler A, Minh BQ. W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. Nucleic Acids Research. 2016;44:W232–W235. PMID: 27084950. doi: org/10.1093/nar/gkw256
70. Felsenstein J. Confidence Limits on Phylogenies: an Approach Using the Bootstrap. Evolution. 1985;39:783–791. PMID: 28561359. doi: 10.1111/j.1558-5646.1985.tb00420.x
71. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Systematic Biology. 2010;59:307–321. PMID: 20525638. doi: 10.1093/sysbio/syq010
72. Anisimova M, Gil M, Dufayard J-F, Dessimoz C, Gascuel, O. Survey of Branch Support Methods Demonstrates Accuracy, Power, and Robustness of Fast Likelihood-based Approximation Schemes. Systematic Biology. 2011;60:685–699. PMID: 21540409. doi: 10.1093/sysbio/syr041

73. Minh BQ, Nguyen MA, von Haeseler A. Ultrafast Approximation for Phylogenetic Bootstrap. Molecular Biology and Evolution. 2013;30:1188–1195. PMID: 23418397. doi: 10.1093/molbev/mst024

74. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Molecular Biology and Evolution. 2017;35:518–522. PMID: 29077904. doi: 10.1093/molbev/msx281

75. Jobb G, von Haeseler A, Strimmer K. TREEFINDER: a Powerful Graphical Analysis Environment for Molecular Phylogenetics. BMC Evolutionary Biology. 2004;28:4–18. PMID: 15222900. doi: 10.1186/1471-2148-4-18

76. Strimmer K, von Haeseler A. Quartet Puzzling: a Quartet Maximum Likelihood Method for Reconstructing Tree Topologies. Molecular Biology and Evolution. 1996;13:964–969. doi: org/10.1093/oxfordjournals.molbev.a025664

77. Shimodaira H, Hasegawa M. Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. Molecular Biology and Evolution. 1999;16:1114–1116. doi: org/10.1093/oxfordjournals.molbev.a026201

78. Shimodaira H. An Approximately Unbiased Test of Phylogenetic Tree Selection. Systematic Biology. 2002;51:492–508. PMID: 12079646. doi: 10.1080/10635150290069913

79. Whidden C, Zeh N, Beiko RG. Supertrees Based on the Subtree Prune-and-Regraft Distance. Systematic Biology. 2014;63:566–581. PMID: 24695589. doi: 10.1093/sysbio/syu023

80. Petitjean C, Deschamps P, López-García P, Moreira D. Rooting the Domain Archaea by Phylogenomic Analysis Supports the Foundation of the New Kingdom Proteoarchaeota. Genome Biology and Evolution. 2014;7:191–204. PMID: 25527841. doi: 10.1093/gbe/evu274

81. Edgar, R. C. MUSCLE v5 enables improved estimates of phylogenetic tree confidence by ensemble bootstrapping. bioRxiv [Preprint]. [Posted 2021 June 21; cited 2022 May 07]: [27

p.]. Available from: https://www.biorxiv.org/content/10.1101/2021.06.20.449169v1.full doi: org/10.1101/2021.06.20.449169

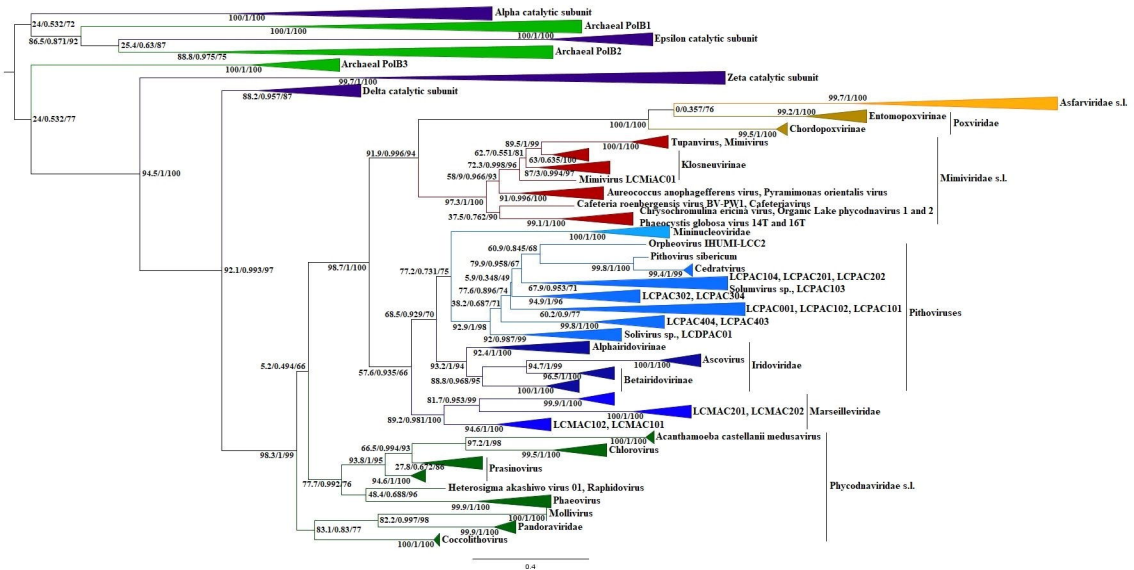
82. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nature Biotechnology. 2017;35:1026-1028. PMID: 29035372. doi: 10.1038/nbt.3988

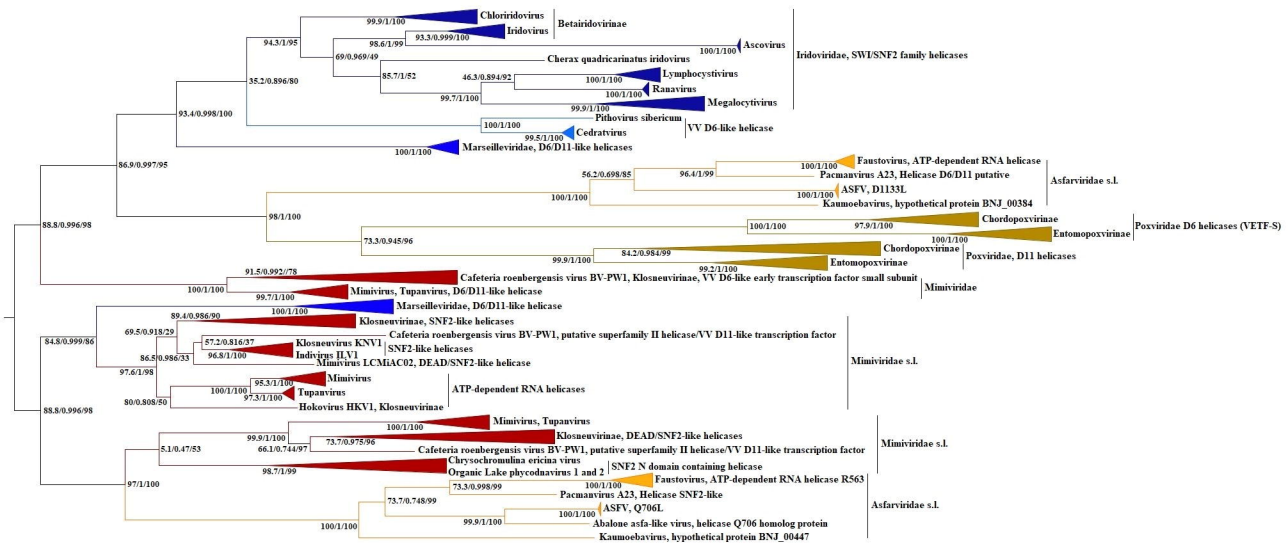
83. Holm L. DALI and the Persistence of Protein Shape. Protein Science. 2020;29:128–140. PMID: 31606894. doi: org/10.1002/pro.3749

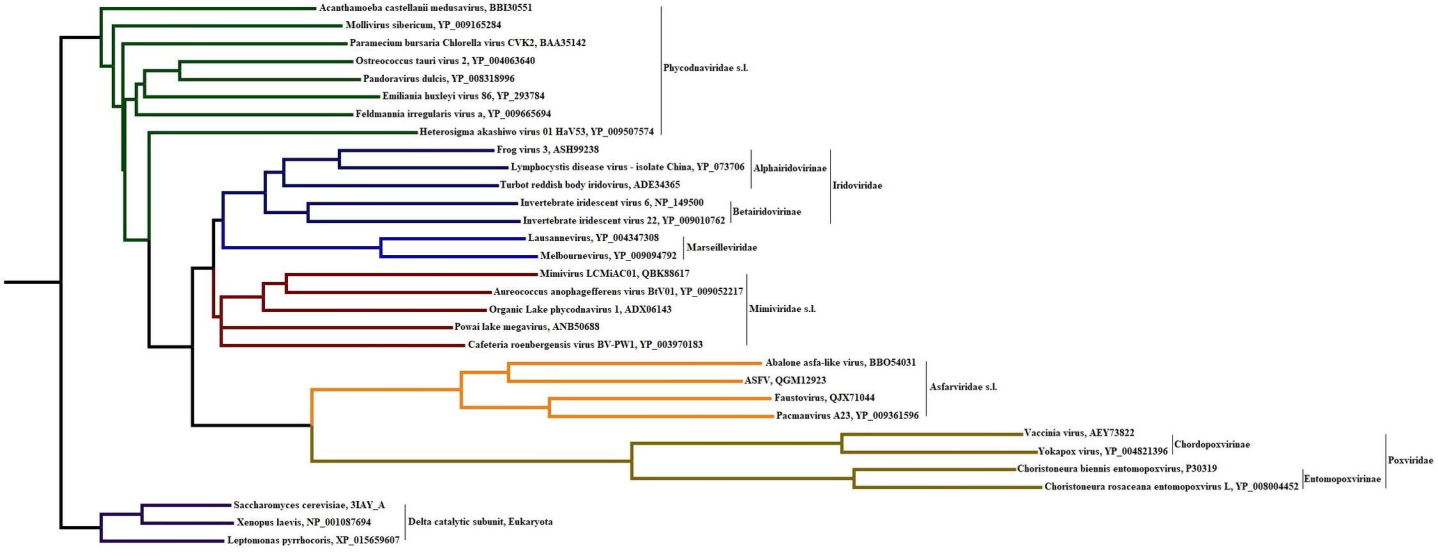
84. Holm L. Using Dali for Protein Structure Comparison. Comparative Study. 2020;2112:29–42. PMID: 32006276. doi: 10.1007/978-1-0716-0270-6_3

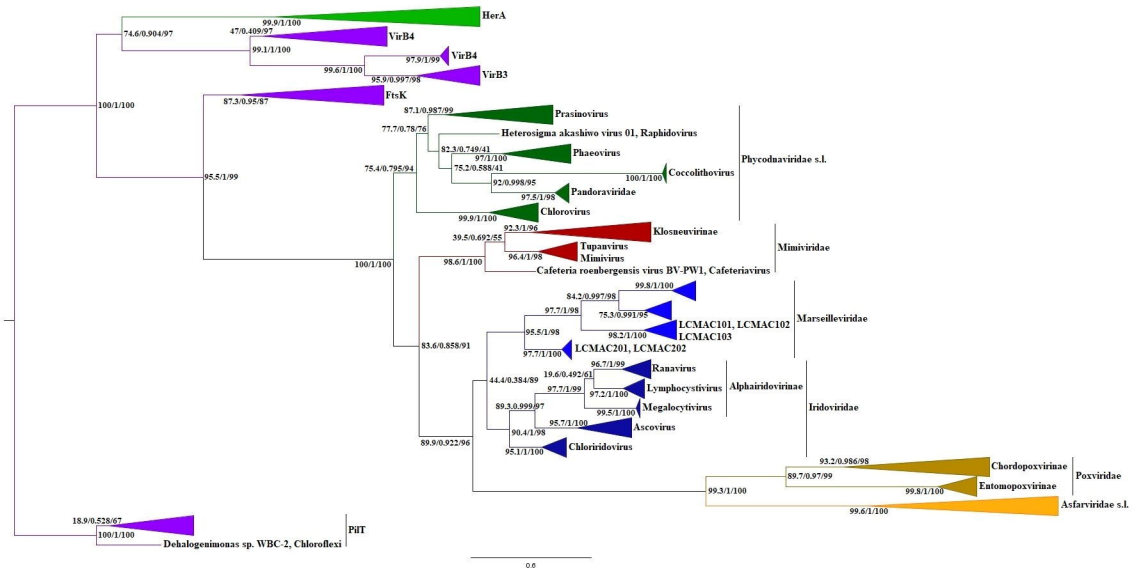
85. Holm L, Sander C. Dali: A Network Tool for Protein Structure Comparison. Trends in Biochemical Sciences. 1995;20:478–480. PMID: 8578593. doi: 10.1016/s0968-0004(00)89105-7

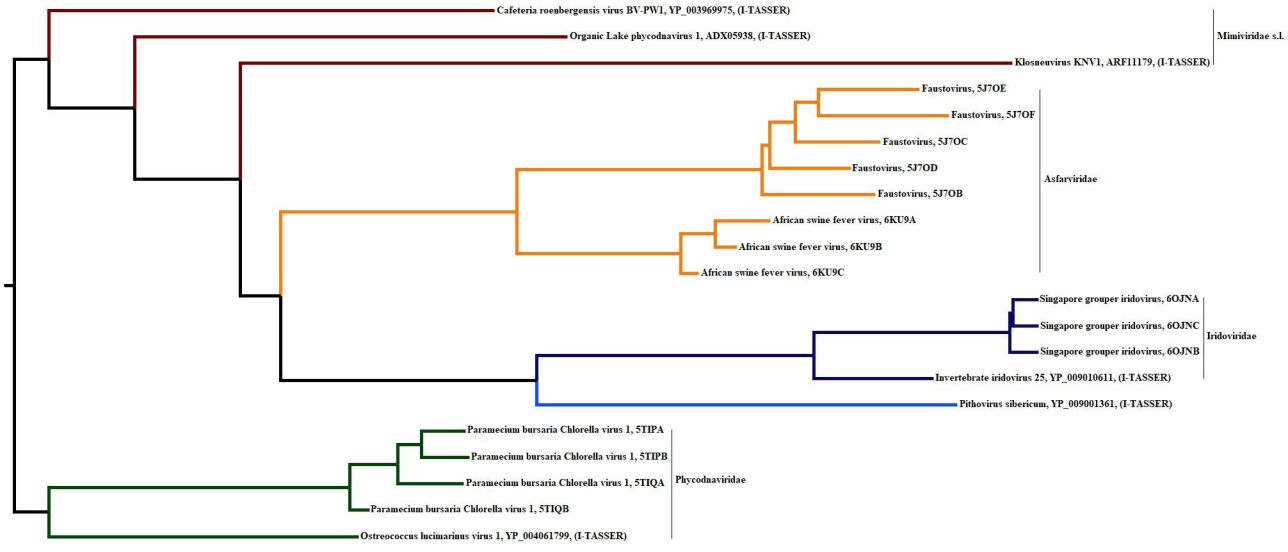
86. Battistuzzi FU, Hedges SB. A Major Clade of Prokaryotes with Ancient Adaptations to Life on Land. Molecular Biology and Evolution. 2009;26:335–343. PMID: 18988685. doi: 10.1093/molbev/msn247

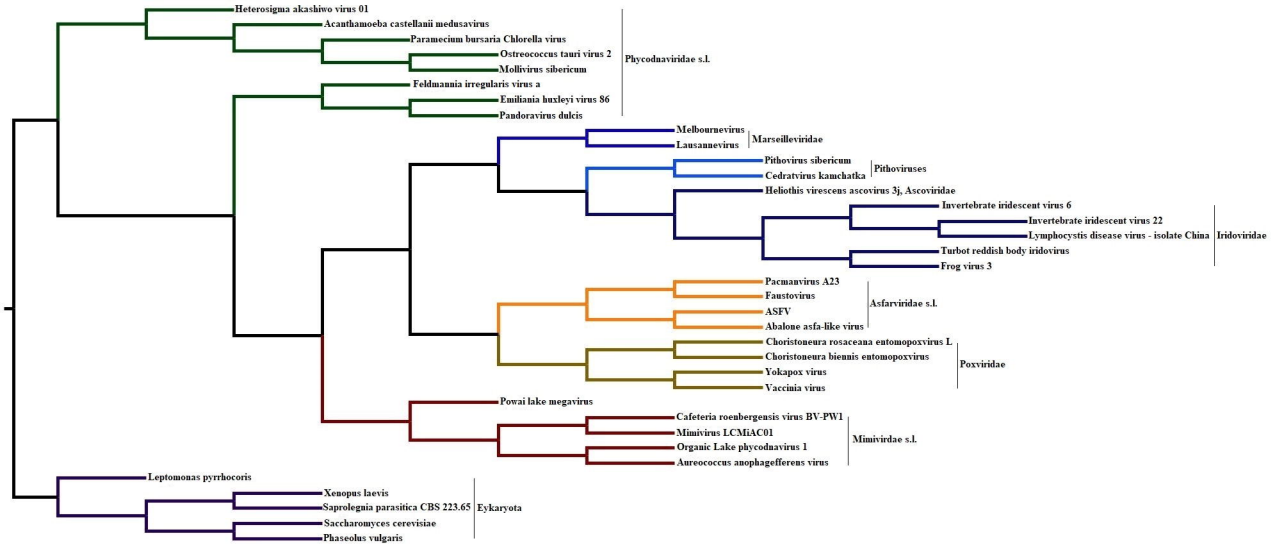

Alpha catalytic subunit
Archaeal PolB1
Epsilon catalytic subunit
Archaeal PolB2
Archaeal PolB3
Zeta catalytic subunit
Delta catalytic subunit
Asfarviridae s.l.
Entomopoxvirinae
Chordopoxvirinae
Poxviridae
Tupanvirus, Minivirus
Klosneuvirinae
Mimivirus LCMiAC01
Mimiviridae s.l.
Aureococcus anophagefferens virus, Pyramimonas orientalis virus
Cafeteria roenbergensis virus BV-PW1, Cafeteriavirus
Chrysochromulina ericina virus, Organic Lake phycodnavirus 1 and 2
Phaeocystis globosa virus 14T and 16T
Minunucleoviridae
Orpheovirus IHUMI-LCC2
Pithovirus sibericum
Cedratvirus
LCPAC104, LCPAC201, LCPAC202
Solumvirus sp., LCPAC103
LCPAC302, LCPAC304
LCPAC001, LCPAC102, LCPAC101
LCPAC404, LCPAC403
Solivirus sp., LCDPAC01
Pithoviruses
Alphairidovirinae
Ascovirus
Iridoviridae
Betairidovirinae
LCMAC201, LCMAC202
Marseilleviridae
LCMAC102, LCMAC101
Acanthamoeba castellanii medusavirus
Chlorovirus
Prasinovirus
Heterosigma akashiwo virus 01, Raphidovirus
Phaeovirus
Mollivirus
Pandoraviridae
Coccolithovirus
Phycodnaviridae s.l.
24/0.532/72
100/1/100
86.5/0.871/92
100/1/100
25.4/0.63/87
100/1/100
88.8/0.975/75
100/1/100
99.7/1/100
88.2/0.957/87
24/0.532/77
94.5/1/100
0/0.357/76
99.2/1/100
99.7/1/100
100/1/100
99.5/1/100
89.5/1/99
100/1/100
62.7/0.551/81
63/0.635/100
72.3/0.998/96
87/3/0.994/97
58/9/0.966/93
91.9/0.996/94
97.3/1/100
91/0.996/100
37.5/0.762/90
99.1/1/100
100/1/100
60.9/0.845/68
79.9/0.958/67
99.8/1/100
99.4/1/99
5.9/0.348/49
77.6/0.896/74
67.9/0.953/71
38.2/0.687/71
94.9/1/96
60.2/0.9/77
99.8/1/100
92.9/1/98
92/0.987/99
92.4/1/100
93.2/1/94
94.7/1/99
100/1/100
88.8/0.968/95
96.5/1/100
100/1/100
81.7/0.953/99
99.9/1/100
100/1/100
94.6/1/100
97.2/1/98
100/1/100
99.5/1/100
66.5/0.994/93
93.8/1/95
27.8/0.672/86
94.6/1/100
48.4/0.688/96
99.9/1/100
82.2/0.997/98
100/1/100
99.9/1/100
100/1/100
92.1/0.993/97
98.7/1/100
77.2/0.731/75
68.5/0.929/70
57.6/0.935/66
89.2/0.981/100
5.2/0.494/66
98.3/1/99
77.7/0.992/76
83.1/0.83/77
0.4

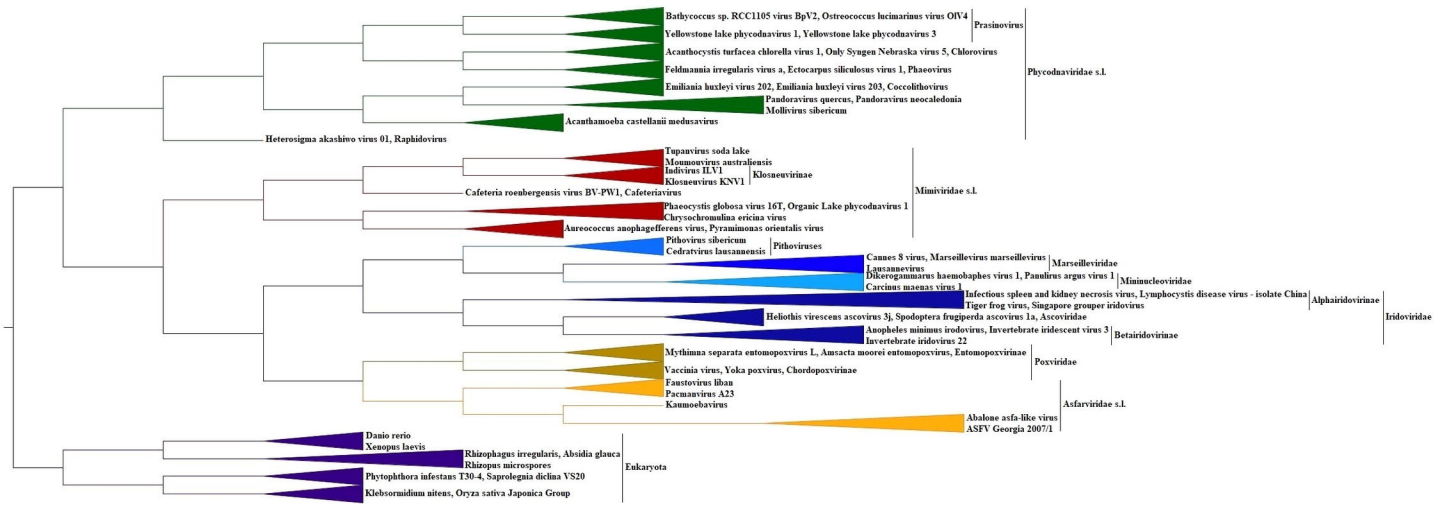

Chloriridovirus
Iridovirus
Betairidovirinae
Ascovirus
Cherax quadricarinatus iridovirus
Lymphocystivirus
Ranavirus
Megalocytivirus
Iridoviridae, SWI/SNF2 family helicases
Pithovirus sibericum
Cedratvirus
VV D6-like helicase
Marseilleviridae, D6/D11-like helicases
Faustovirus, ATP-dependent RNA helicase
Pacmanvirus A23, Helicase D6/D11 putative
ASFV, D1133L
Kaumoebavirus, hypothetical protein BNJ_00384
Asfarviridae s.l.
Chordopoxvirinae
Entomopoxvirinae
Poxviridae D6 helicases (VETF-S)
Chordopoxvirinae
Entomopoxvirinae
Poxviridae, D11 helicases
Cafeteria roenbergensis virus BV-PW1, Klosneuvirinae, VV D6-like early transcription factor small subunit
Mimivirus, Tupanvirus, D6/D11-like helicase
Mimiviridae
Marseilleviridae, D6/D11-like helicase
Klosneuvirinae, SNF2-like helicases
Cafeteria roenbergensis virus BV-PW1, putative superfamily II helicase/VV D11-like transcription factor
Klosneuvirus KNV1
Indivirus ILV1
SNF2-like helicases
Mimivirus LCMiAC02, DEAD/SNF2-like helicase
Mimivirus
Tupanvirus
Hokovirus HKV1, Klosneuvirinae
ATP-dependent RNA helicases
Mimiviridae s.l.
Mimivirus, Tupanvirus
Klosneuvirinae, DEAD/SNF2-like helicases
Cafeteria roenbergensis virus BV-PW1, putative superfamily II helicase/VV D11-like transcription factor
Chrysochromulina ericina virus
Organic Lake phycodnavirus 1 and 2
SNF2 N domain containing helicase
Mimiviridae s.l.
Faustovirus, ATP-dependent RNA helicase R563
Pacmanvirus A23, Helicase SNF2-like
ASFV, Q706L
Abalone asfa-like virus, helicase Q706 homolog protein
Kaumoebavirus, hypothetical protein BNJ_00447
Asfarviridae s.l.
0.4

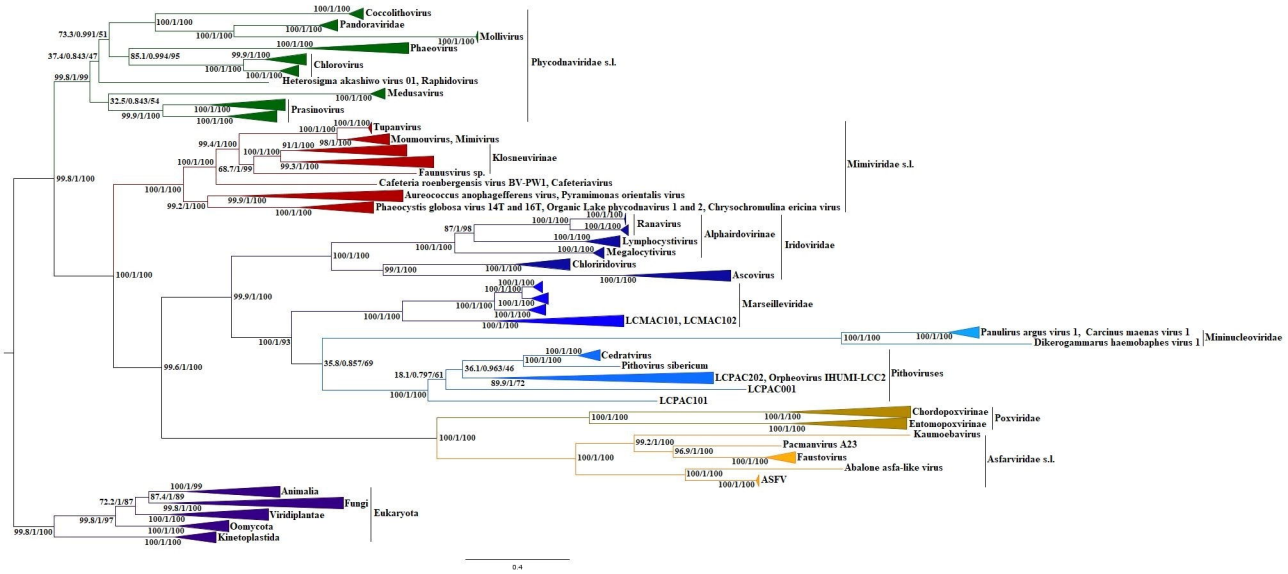

Acanthamoeba castellanii medusavirus, BBI30551
Mollivirus sibericum, YP_009165284
Paramecium bursaria Chlorella virus CVK2, BAA35142
Ostreococcus tauri virus 2, YP_004063640
Pandoravirus dulcis, YP_008318996
Emiliania huxleyi virus 86, YP_293784
Feldmannia irregularis virus a, YP_009665694
Heterosigma akashiwo virus 01 HaV53, YP_009507574
Phycodnaviridae s.l.
Frog virus 3, ASH99238
Lymphocystis disease virus - isolate China, YP_073706
Turbot reddish body iridovirus, ADE34365
Alphairidovirinae
Invertebrate iridescent virus 6, NP_149500
Invertebrate iridescent virus 22, YP_009010762
Betairidovirinae
Iridoviridae
Lausannevirus, YP_004347308
Melbournevirus, YP_009094792
Marseilleviridae
Mimivirus LCMiAC01, QBK88617
Aureococcus anophagefferens virus BtV01, YP_009052217
Organic Lake phycodnavirus 1, ADX06143
Powai lake megavirus, ANB50688
Cafeteria roenbergensis virus BV-PW1, YP_003970183
Mimiviridae s.l.
Abalone asfa-like virus, BBO54031
ASFV, QGM12923
Faustovirus, QJX71044
Pacmanvirus A23, YP_009361596
Asfarviridae s.l.
Vaccinia virus, AEY73822
Yokapox virus, YP_004821396
Chordopoxvirinae
Choristoneura biennis entomopoxvirus, P30319
Choristoneura rosaceana entomopoxvirus L, YP_008004452
Entomopoxvirinae
Poxviridae
Saccharomyces cerevisiae, 3IAV_A
Xenopus laevis, NP_001087694
Leptomonas pyrrhocoris, XP_015659607
Delta catalytic subunit, Eukaryota

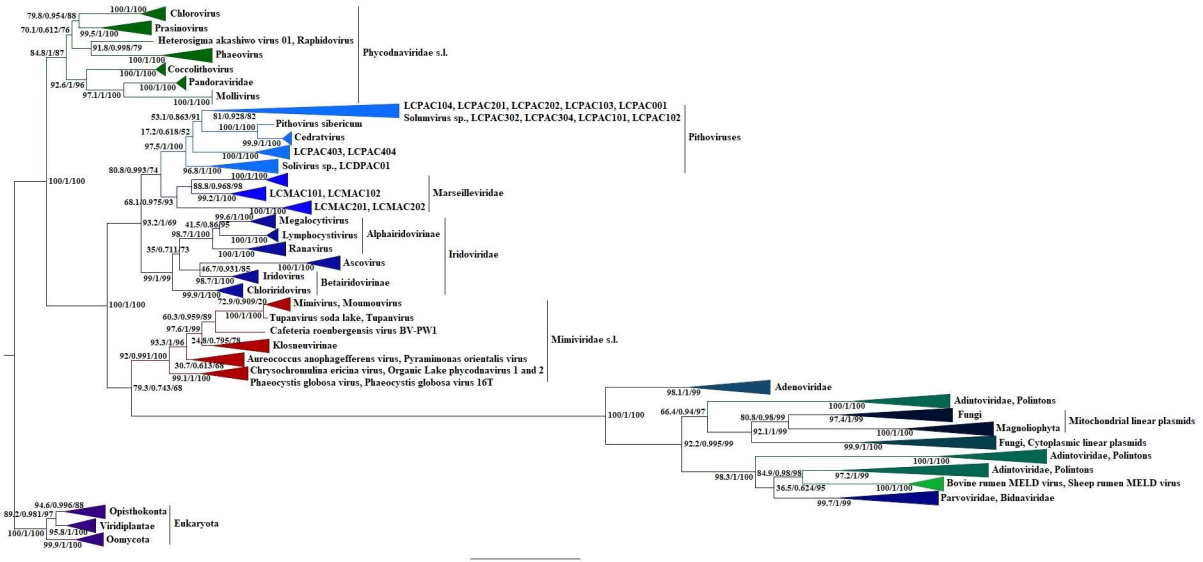

HerA
VirB4
VirB4
VirB3
FtsK
Prasinovirus
Heterosigma akashiwo virus 01, Raphidovirus
Phaeovirus
Coccolithovirus
Phycodnaviridae s.l.
Pandoraviridae
Chlorovirus
Klosneuvirinae
Tupanvirus
Mimivirus
Mimiviridae
Cafeteria roenbergensis virus BV-PW1, Cafeteriavirus
LCMAC101, LCMAC102
LCMAC103
LCMAC201, LCMAC202
Marseilleviridae
Ranavirus
Lymphocystivirus
Megalocytivirus
Alphairidovirinae
Ascovirus
Chloriridovirus
Iridoviridae
Chordopoxvirinae
Entomopoxvirinae
Poxviridae
Asfarviridae s.l.
PilT
Dehalogenimonas sp. WBC-2, Chloroflexi
74.6/0.904/97
99.9/1/100
47/0.409/97
99.1/1/100
97.9/1/99
99.6/1/100
95.9/0.997/98
100/1/100
87.3/0.95/87
87.1/0.987/99
77.7/0.78/76
82.3/0.749/41
97/1/100
75.2/0.588/41
100/1/100
92/0.998/95
97.5/1/98
99.9/1/100
95.5/1/99
75.4/0.795/94
92.3/1/96
39.5/0.692/55
96.4/1/98
100/1/100
98.6/1/100
99.8/1/100
84.2/0.997/98
97.7/1/98
75.3/0.991/95
98.2/1/100
95.5/1/98
97.7/1/100
83.6/0.858/91
96.7/1/99
19.6/0.492/61
97.7/1/99
97.2/1/100
44.4/0.384/89
89.3/0.999/97
99.5/1/100
95.7/1/100
90.4/1/98
95.1/1/100
89.9/0.922/96
93.2/0.986/98
89.7/0.97/99
99.8/1/100
99.3/1/100
99.6/1/100
18.9/0.528/67
100/1/100
0.6

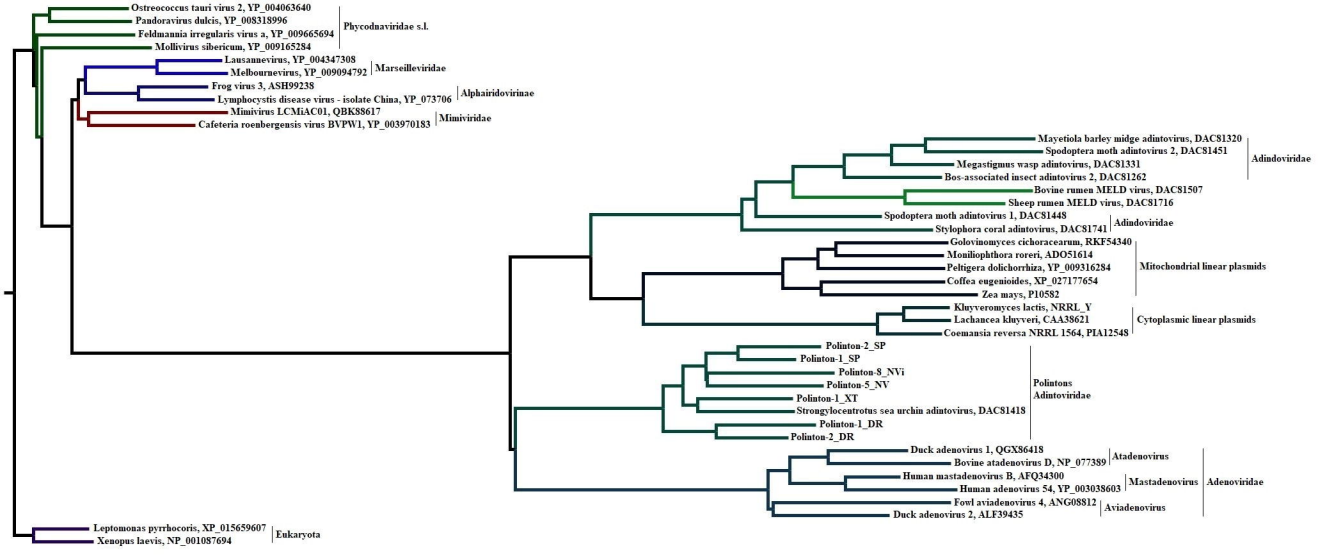

Cafeteria roenbergensis virus BV-PW1, YP_003969975, (I-TASSER)
Organic Lake phycodnavirus 1, ADX05938, (I-TASSER)
Klosneuvirus KNV1, ARF11179, (I-TASSER)
Minivirdae s.l.
Faustovirus, 5J7OE
Faustovirus, 5J7OF
Faustovirus, 5J7OC
Faustovirus, 5J7OD
Faustovirus, 5J7OB
African swine fever virus, 6KU9A
African swine fever virus, 6KU9B
African swine fever virus, 6KU9C
Asfarviridae
Singapore grouper iridovirus, 6OJNA
Singapore grouper iridovirus, 6OJNC
Singapore grouper iridovirus, 6OJNB
Invertebrate iridovirus 25, YP_009010611, (I-TASSER)
Iridoviridae
Pithovirus sibericum, YP_009001361, (I-TASSER)
Paramecium bursaria Chlorella virus 1, 5TIPA
Paramecium bursaria Chlorella virus 1, 5TIPB
Paramecium bursaria Chlorella virus 1, 5TIQA
Paramecium bursaria Chlorella virus 1, 5TIQB
Phycodnaviridae
Ostreococcus lucimarinus virus 1, YP_004061799, (I-TASSER)

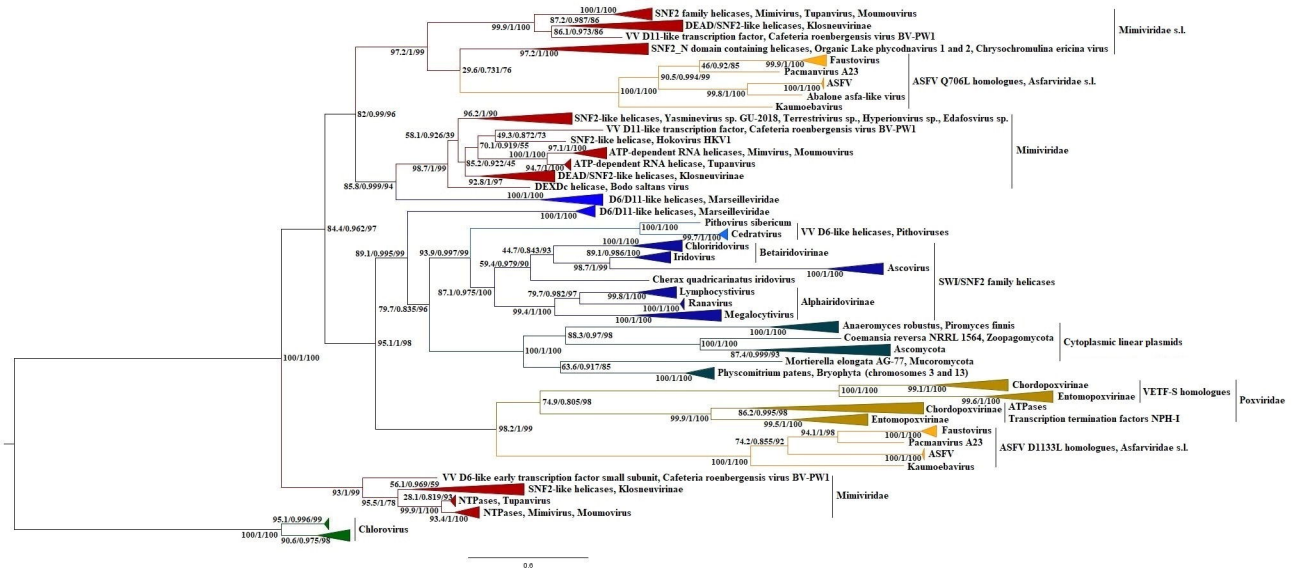

Heterosigma akashiwo virus 01
Acanthamoeba castellanii medusavirus
Paramecium bursaria Chlorella virus
Ostreococcus tauri virus 2
Mollivirus sibericum
Feldmannia irregularis virus a
Emiliania huxleyi virus 86
Pandoravirus dulcis
Phycodnaviridae s.l.
Melbournevirus
Lausannevirus
Marseilleviridae
Pithovirus sibericum
Cedratvirus kamchatka
Pithoviruses
Heliothis virescens ascovirus 3j, Ascoviridae
Invertebrate iridescent virus 6
Invertebrate iridescent virus 22
Lymphocystis disease virus - isolate China
Turbot reddish body iridovirus
Frog virus 3
Iridoviridae
Pacmanvirus A23
Faustovirus
ASFV
Abalone asfa-like virus
Asfarviridae s.l.
Choristoneura rosaceana entomopoxvirus L
Choristoneura biennis entomopoxvirus
Yokapox virus
Vaccinia virus
Poxviridae
Powai lake megavirus
Cafeteria roenbergensis virus BV-PW1
Mimivirus LCMiAC01
Organic Lake phycodnavirus 1
Aureococcus anophagefferens virus
Mimivirdae s.l.
Leptomonas pyrrhocoris
Xenopus laevis
Saprolegnia parasitica CBS 223.65
Saccharomyces cerevisiae
Phaseolus vulgaris
Eykaryota

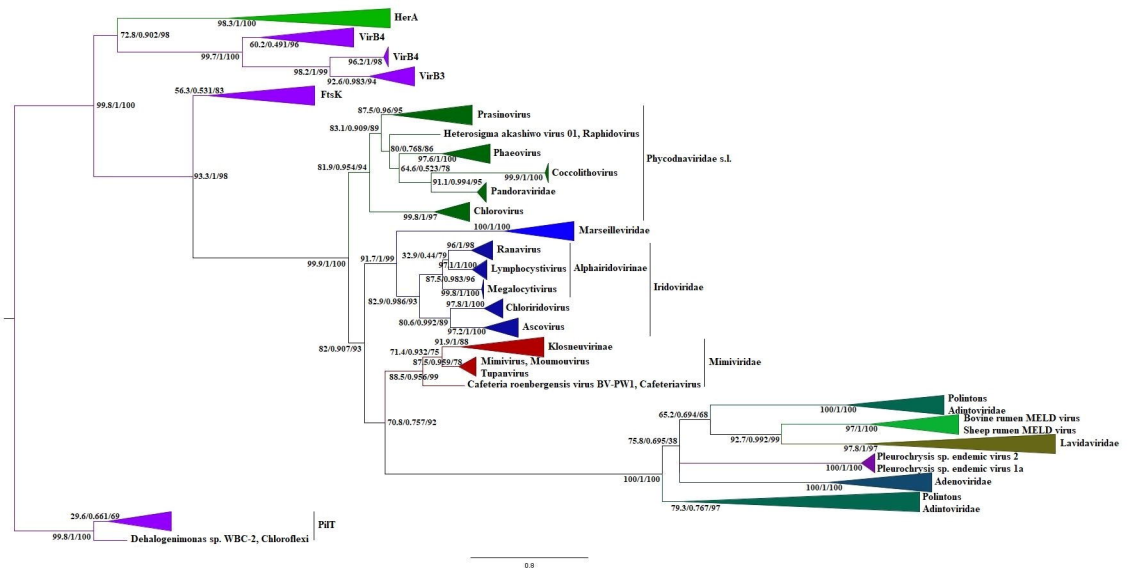

Bathycoccus sp. RCC1105 virus BpV2, Ostreococcus lucimarinus virus OlV4
Yellowstone lake phycodnavirus 1, Yellowstone lake phycodnavirus 3
Prasinovirus
Acanthocystis turfacea chlorella virus 1, Only Syngen Nebraska virus 5, Chlorovirus
Feldmannia irregularis virus a, Ectocarpus siliculosus virus 1, Phaeovirus
Emiliania huxleyi virus 202, Emiliania huxleyi virus 203, Coccolithovirus
Pandoravirus quercus, Pandoravirus neocaledonia
Mollivirus sibericum
Phycodnaviridae s.l.
Acanthamoeba castellanii medusavirus
Heterosigma akashiwo virus 01, Raphidovirus
Tupanvirus soda lake
Mounouvirus australiensis
Indivirus ILV1
Klosneuvirus KNV1
Klosneuvirinae
Cafeteria roenbergensis virus BV-PW1, Cafeteriavirus
Mimiviridae s.l.
Phaeocystis globosa virus 16T, Organic Lake phycodnavirus 1
Chrysochromulina ericina virus
Aureococcus anophagefferens virus, Pyramimonas orientalis virus
Pithovirus sibericum
Cedratvirus lausannensis
Pithoviruses
Cannes 8 virus, Marseillevirus marseillevirus
Lausannevirus
Marseilleviridae
Dikerogammarus haemobaphes virus 1, Paulirus argus virus 1
Carcinus maenas virus 1
Minincleoviridae
Infectious spleen and kidney necrosis virus, Lymphocystis disease virus - isolate China
Alphairidovirinae
Tiger frog virus, Singapore grouper iridovirus
Heliothis virescens ascovirus 3j, Spodoptera frugiperda ascovirus 1a, Ascoviridae
Iridoviridae
Anopheles minimus irodovirus, Invertebrate iridescent virus 3
Invertebrate iridovirus 22
Betairidovirinae
Mythimna separata entomopoxvirus L, Amsacta moorei entomopoxvirus, Entomopoxvirinae
Vaccinia virus, Yoka poxvirus, Chordopoxvirinae
Poxviridae
Faustovirus liban
Pacmanvirus A23
Kaumoebavirus
Asfarviridae s.l.
Abalone asfa-like virus
ASFV Georgia 2007/1
Danio rerio
Xenopus laevis
Rhizophagus irregularis, Absidia glauca
Rhizopus microspores
Phytophthora infestans T30-4, Saprolegnia diclina VS20
Klebsormidium nitens, Oryza sativa Japonica Group
Eukaryota

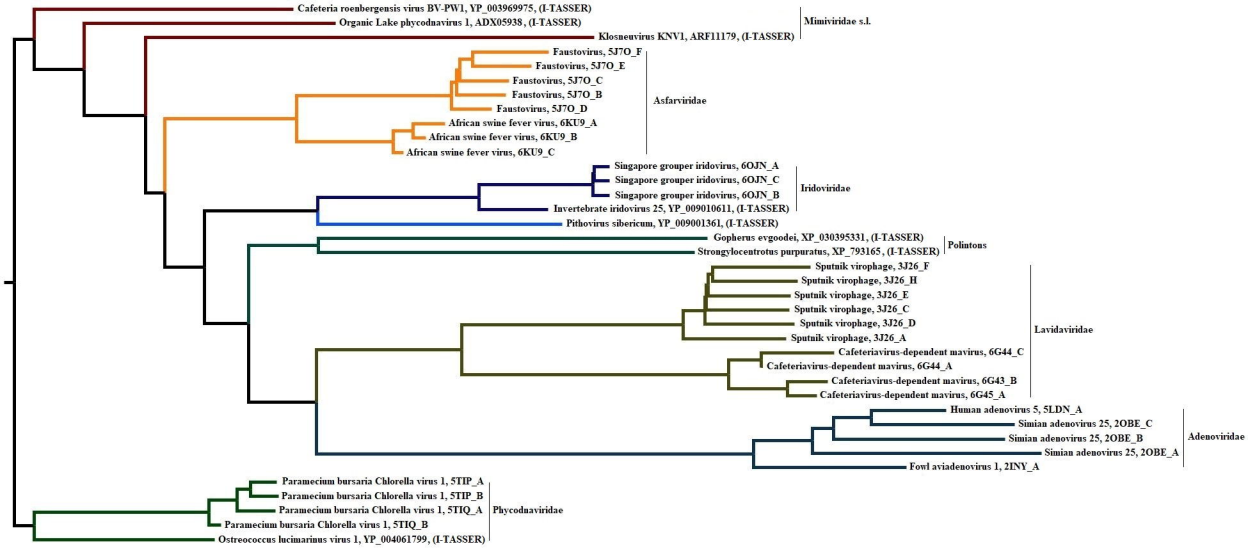

Coccolithovirus
Pandoraviridae
Mollivirus
Phaeovirus
Chlorovirus
Heterosigma akashiwo virus 01, Raphidovirus
Medusavirus
Phycodnaviridae s.l.
Prasinovirus
Tupanvirus
Mommouvirus, Mimivirus
Klosneuvirinae
Faunusvirus sp.
Cafeteria roenbergensis virus BV-PW1, Cafeteriavirus
Aureococcus anophagefferens virus, Pyraminonas orientalis virus
Phaeocystis globosa virus 14T and 16T, Organic Lake phycodnavirus 1 and 2, Chrysochromulina ericina virus
Mimiviridae s.l.
Ranavirus
Lymphocystivirus
Megalocytivirus
Alphairdovirinae
Chloridovirus
Ascovirus
Iridoviridae
LCMAC101, LCMAC102
Marseilleviridae
Panulirus argus virus 1, Carcinus maenas virus 1
Dikerogammarus haemobaphes virus 1
Mininucleoviridae
Cedratvirus
Pithovirus sibericum
LCPAC202, Orpheovirus IHUMI-LCC2
LCPAC001
LCPAC101
Pithoviruses
Chordopoxvirinae
Entomopoxvirinae
Kaumoebavirus
Poxviridae
Pacmanvirus A23
Faustovirus
Abalone asfa-like virus
ASFV
Asfarviridae s.l.
Animalia
Fungi
Viridiplantae
Eukaryota
Oomycota
Kinetoplastida
0.4


Chlorovirus
Prasinovirus
Heterosigma akashiwo virus 01, Raphidovirus
Phaeovirus
Coccolithovirus
Pandoraviridae
Mollivirus
Plycodnaviridae s.l.
LCPAC104, LCPAC201, LCPAC202, LCPAC103, LCPAC001
Solumvirus sp., LCPAC302, LCPAC304, LCPAC101, LCPAC102
Pithovirus sibericum
Cedratvirus
LCPAC403, LCPAC404
Solivirus sp., LCDPAC01
Pithoviruses
LCMAC101, LCMAC102
LCMAC201, LCMAC202
Marseilleviridae
Megalocytivirus
Lymphocystivirus
Alphairidovirinae
Ranavirus
Iridoviridae
Ascovirus
Iridovirus
Betairidovirinae
Chloriridovirus
Mimivirus, Moumouvirus
Tupanvirus soda lake, Tupanvirus
Cafeteria roenbergensis virus BV-PW1
Klosneuvirinae
Aureococcus anophagefferens virus, Pyramimonas orientalis virus
Chrysochromulina ericina virus, Organic Lake phycodnavirus 1 and 2
Phaeocystis globosa virus, Phaeocystis globosa virus 16T
Mimiviridae s.l.
Adenoviridae
Adintoviridae, Polintons
Fungi
Magnoliophyta
Mitochondrial linear plasmids
Fungi, Cytoplasmic linear plasmids
Adintoviridae, Polintons
Adintoviridae, Polintons
Bovine rumen MELD virus, Sheep rumen MELD virus
Parvoviridae, Bidnaviridae
Opisthokonta
Viridiplantae
Oomycota
Eukaryota
0.7


Ostreococcus tauri virus 2, YP_004063640
Pandoravirus dulcis, YP_008318996
Feldmannia irregularis virus a, YP_009665694
Mollivirus sibericum, YP_009165284
Phycodnaviridae s.l.
Lausannevirus, YP_004347308
Melbournevirus, YP_009094792
Marseilleviridae
Frog virus 3, ASH99238
Lymphocystis disease virus - isolate China, YP_073706
Alphairidovirinae
Mimivirus LCMiAC01, QBK88617
Cafeteria roenbergensis virus BVPW1, YP_003970183
Mimiviridae
Mayetiola barley midge adintovirus, DAC81320
Spodoptera moth adintovirus 2, DAC81451
Megastigmus wasp adintovirus, DAC81331
Bos-associated insect adintovirus 2, DAC81262
Adindoviridae
Bovine rumen MELD virus, DAC81507
Sheep rumen MELD virus, DAC81716
Spodoptera moth adintovirus 1, DAC81448
Stylophora coral adintovirus, DAC81741
Adindoviridae
Golovinomyces cichoracearum, RKF54340
Moniliophthora roreri, ADO51614
Peltigera dolichorrhiza, YP_009316284
Coffea eugenioides, XP_027177654
Zea mays, P10582
Mitochondrial linear plasmids
Kluyveromyces lactis, NRRL_Y
Lachancea kluyveri, CAA38621
Coemansia reversa NRRL 1564, PIA12548
Cytoplasmic linear plasmids
Polinton-2_SP
Polinton-1_SP
Polinton-8_NVi
Polinton-5_NV
Polinton-1_XT
Strongylocentrotus sea urchin adintovirus, DAC81418
Polinton-1_DR
Polinton-2_DR
Polintons
Adintoviridae
Duck adenovirus 1, QGX86418
Bovine atadenovirus D, NP_077389
Atadenovirus
Human mastadenovirus B, AFQ34300
Human adenovirus 54, YP_003038603
Mastadenovirus
Adenoviridae
Fowl aviadenovirus 4, ANG08812
Duck adenovirus 2, ALF39435
Aviadenovirus
Leptomonas pyrrhocoris, XP_015659607
Xenopus laevis, NP_001087694
Eukaryota


SNF2 family helicases, Mimivirus, Tupanvirus, Moumouvirus
DEAD/SNF2-like helicases, Klosneuvirinae
VV D11-like transcription factor, Cafeteria roenbergensis virus BV-PW1
SNF2_N domain containing helicases, Organic Lake phycodnavirus 1 and 2, Chrysochromulina ericina virus
Faustovirus
Pacmanvirus A23
ASFV
Abalone asfa-like virus
Kaumoebavirus
ASFV Q706L homologues, Asfarviridae s.l.
Mimiviridae s.l.
SNF2-like helicases, Yasminevirus sp. GU-2018, Terrestrivirus sp., Hyperionvirus sp., Edafosvirus sp.
VV D11-like transcription factor, Cafeteria roenbergensis virus BV-PW1
SNF2-like helicase, Hokovirus HKV1
ATP-dependent RNA helicases, Mimivirus, Moumouvirus
ATP-dependent RNA helicase, Tupanvirus
DEAD/SNF2-like helicases, Klosneuvirinae
DEXDc helicase, Bodo saltans virus
Mimiviridae
D6/D11-like helicases, Marseilleviridae
D6/D11-like helicases, Marseilleviridae
Pithovirus sibericum
VV D6-like helicases, Pithoviruses
Cedratvirus
Chloriridovirus
Iridovirus
Betairidovirinae
Ascovirus
Cherax quadricarinatus iridovirus
Lymphocystivirus
Ranavirus
Alphairidovirinae
Megalocytivirus
SWI/SNF2 family helicases
Anaeromyces robustus, Piromyces finnis
Coemansia reversa NRRL 1564, Zoopagomycota
Ascomycota
Mortierella elongata AG-77, Mucoromycota
Physcomitrium patens, Bryophyta (chromosomes 3 and 13)
Cytoplasmic linear plasmids
Chordopoxvirinae
Entomopoxvirinae
VETF-S homologues
Poxviridae
Chordopoxvirinae
Entomopoxvirinae
ATPases
Transcription termination factors NPH-I
Faustovirus
Pacmanvirus A23
ASFV
Kaumoebavirus
ASFV D1133L homologues, Asfarviridae s.l.
VV D6-like early transcription factor small subunit, Cafeteria roenbergensis virus BV-PW1
SNF2-like helicases, Klosneuvirinae
NTPases, Tupanvirus
NTPases, Mimivirus, Moumouvirus
Mimiviridae
Chlorovirus
0.6


HerA
VirB4
VirB4
VirB3
FtsK
Prasinovirus
Heterosigma akashiwo virus 01, Raphidovirus
Phaeovirus
Coccolithovirus
Pandoraviridae
Phycodnaviridae s.l.
Chlorovirus
Marseilleviridae
Ranavirus
Lymphocystivirus
Alphairidovirinae
Megalocytivirus
Iridoviridae
Chloriridovirus
Ascovirus
Klosneuvirinae
Mimivirus, Moumouvirus
Tupanvirus
Cafeteria roenbergensis virus BV-PW1, Cafeteriavirus
Mimiviridae
Polintons
Adintoviridae
Bovine rumen MELD virus
Sheep rumen MELD virus
Lavidaviridae
Pleurochrysis sp. endemic virus 2
Pleurochrysis sp. endemic virus 1a
Adenoviridae
Polintons
Adintoviridae
PilT
Dehalogenimonas sp. WBC-2, Chloroflexi
98.3/1/100
72.8/0.902/98
60.2/0.491/96
99.7/1/100
96.2/1/98
98.2/1/99
92.6/0.983/94
56.3/0.531/83
99.8/1/100
87.5/0.96/95
83.1/0.909/89
80/0.768/86
97.6/1/100
64.6/0.523/78
99.9/1/100
91.1/0.994/95
81.9/0.954/94
93.3/1/98
99.8/1/97
100/1/100
96/1/98
32.9/0.44/79
91.7/1/99
97/1/100
87.5/0.983/96
99.3/1/100
99.9/1/100
82.9/0.986/93
97.8/1/100
80.6/0.992/89
97.2/1/100
91.9/1/88
82/0.907/93
71.4/0.932/75
87/0.989/78
88.5/0.956/99
70.8/0.757/92
100/1/100
65.2/0.694/68
97/1/100
92.7/0.992/99
75.8/0.695/38
97.8/1/97
100/1/100
100/1/100
100/1/100
79.3/0.767/97
29.6/0.661/69
99.8/1/100
0.8


Cafeteria roenbergensis virus BV-PW1, YP_003969975, (I-TASSER)
Organic Lake phycodnavirus 1, ADX05938 , (I-TASSER)
Klosneuvirus KNV1, ARF11179, (I-TASSER)
Mimiviridae s.l.
Faustovirus, 5J7O_F
Faustovirus, 5J7O_E
Faustovirus, 5J7O_C
Faustovirus, 5J7O_B
Faustovirus, 5J7O_D
African swine fever virus, 6KU9_A
African swine fever virus, 6KU9_B
African swine fever virus, 6KU9_C
Asfarviridae
Singapore grouper iridovirus, 6OJN_A
Singapore grouper iridovirus, 6OJN_C
Singapore grouper iridovirus, 6OJN_B
Invertebrate iridovirus 25, YP_009010611, (I-TASSER)
Iridoviridae
Pithovirus sibericum, YP_009001361, (I-TASSER)
Gopherus evgoodei, XP_030395331, (I-TASSER)
Strongylocentrotus purpuratus, XP_793165 , (I-TASSER)
Polintons
Sputnik virophage, 3J26_F
Sputnik virophage, 3J26_H
Sputnik virophage, 3J26_E
Sputnik virophage, 3J26_C
Sputnik virophage, 3J26_D
Sputnik virophage, 3J26_A
Cafeteriavirus-dependent mavirus, 6G44_C
Cafeteriavirus-dependent mavirus, 6G44_A
Cafeteriavirus-dependent mavirus, 6G43_B
Cafeteriavirus-dependent mavirus, 6G45_A
Lavidaviridae
Human adenovirus 5, 5LDN_A
Simian adenovirus 25, 2OBE_C
Simian adenovirus 25, 2OBE_B
Simian adenovirus 25, 2OBE_A
Fowl aviadenovirus 1, 2INY_A
Adenoviridae
Paramecium bursaria Chlorella virus 1, 5TIP_A
Paramecium bursaria Chlorella virus 1, 5TIP_B
Paramecium bursaria Chlorella virus 1, 5TIQ_A
Paramecium bursaria Chlorella virus 1, 5TIQ_B
Ostreococcus lucimarinus virus 1, YP_004061799, (I-TASSER)
Phycodnaviridae

	PolB	RNAP α	RNAP β	D5-like primase	D6/D11 helicase (SFII helicase)	SFIII helicase	AP (apurinic) endonuclease family 2	Metallopeptidase WLM	PCNA	RuvC	Uracil-DNA glycosylase	DNA topoisomerase II	mRNA capping enzyme (large subunit)	TFIIB	HerA-FtsK superfamily ATPases	DJR MCP	VLTF2	VETF
Ph/R	+	+	+	+	Ab	Ab	Ab	+	+	+	Ab	+	+	+	+	+	+	Ab
M/MAPIM+AP	-	-	+	+	+	+	+	-	-	+/-	+	+	+	-	+	+	+	+
M+AP/MAPIM	MU	-	-	-	-	-	-	MU	-	-	-	-	-	-	-	-	-	-
AP/MAPIM	-	-	+	+	+	+	+	-	-	-	+	+	+	-	+	+	+	+

Abbreviations: "/" – A sign indicating the presence of a node for clades, written to the right and to the left of it, "+" – Node exists, "-" – Node is missing, Ab – Lack of node due to lack of the *Phycodnaviridae* s.l. clade, Ph – the *Phycodnaviridae* s.l. clade, R – the NCLDV clades without the *Phycodnaviridae* s.l. clade, M – the *Mimiviridae* s.l. clade, MAPIM – the *Marseilleviridae* + Pithoviruses + *Iridoviridae*/*Ascoviridae* + *Mininucleoviridae* clade or some members of the MAPIM clade, AP – the *Asfarviridae* s.l. + *Poxviridae* clade or one of the two clades. $+^{A}$ – the sequences of the AP clade were not used, i.e. the corresponding phylogenetic trees (S29 and 30 Figs) show sister relationships between the *Mimiviridae* s.l. and MAPIM clades. MU - Present according to the MUSCLE algorithm.