Supplementary Information for

Diversification of giant and large eukaryotic dsDNA viruses predated the origin of modern eukaryotes

Julien Guglielmini, Anthony C Woo, Mart Krupovic, Patrick Forterre, Morgan Gaia

Corresponding authors:
Patrick Forterre
patrick.forterre@pasteur.fr
Morgan Gaia
morgan.gaia@pasteur.fr

This PDF file includes:

Other supplementary materials for this manuscript include the following:

Additional data (https://doi.org/10.5281/zenodo.3368642)

SI Discussions

Conservation of genes among NCLDVs

In order to determine a set of core genes that would both be informative and provide reliable markers for our analysis, we designed a Best Bidirectional BLAST-Hit (BBH)-based pipeline with manual curation and built a first dataset based on 96 NCLDV genomes (see Methods and SI Appendix, Table S1), representing every family and including some of the most recently identified viruses. Highly redundant genomes were removed to limit imbalance between families, reducing the dataset to 73 genomes. This step was necessary to avoid some genes to be considered as highly conserved by the algorithm when they are actually missing in several under-represented families but present in over-represented families (many highly-related strains). This analysis determined that 3 protein-coding genes were strictly conserved among the 73 selected NCLDV genomes. This finding was rather remarkable considering the number of viral genomes analysed, the observed divergences between NCLDVs and the common assumption that horizontal gene transfer plays a central role in the evolution of dsDNA viruses. Indeed, one usually expects to find fewer core genes when more genomes are included in phylogenomic analyses. However, if we compare our results to previous NCLDV genomic analyses, it is noteworthy that despite substantially increasing the number of genomes (especially from novel or under-represented families), the number of core genes did not really decrease. Indeed, in 2012, Yutin and Koonin (1) studied 45 genomes and defined 5 genes strictly conserved in every NCLDV family: the DNA pol B, the primase, the VLTF3-like, the MCP, and the packaging ATPase. Only the two latter were not found in our strict core genes set, but this was only due to the inclusion in our dataset of genomes known to lack the MCP or ATPase (2–5). These two genes are otherwise present in every other NCLDV. Considering for both studies a relaxed definition of the

core genomes (presence in 92% of genomes), our results are compatible: the most conserved genes then include the RNAP-a and -b, TFIIS, the late transcription factor VLTF-2, and the disulfide (thiol) oxidoreductase of the Erv1/Alr family. Given that the core genome did not significantly change when using two different methodologies and after adding significantly more genomes, we postulate that these 10 genes represent the actual NCLDV core genome. Two genes of the relaxed core gene set (VLTF-2 and Erv1/Alr) were not included in our in-depth phylogenetic analyses because they were not matching other criteria fulfilled by the other relaxed core genes (slightly higher conservation for TFIIS and long proteins for the RNA polymerases). The final list of markers we selected for our study thus comprises 8 proteins: 6 are related to informational processes – genome's expression and replication (DNA pol B, primase, VLTF3-like, TFIIS, RNAP-a, and RNAP-b) – and 2 to virion structure and morphogenesis (pATPase and MCP). While this list is rather short in comparison with the total content of NCLDV genomes, these markers were selected for investigating the backbone of their evolution. The many other genes might also contain valuable sources of information on that matter, and remain to be further investigated.

Interestingly, our analysis of separate core genes in each family reveals different levels of diversifications. For instance, the *Phycodnaviridae* (excluding Pandoraviruses and Mollivirus), with genomes encoding between 150 and 860 genes, only possess 10 strict core genes. Similarly, the *Poxviridae* and one subset of the *Iridoviridae* (the *Alphairidovirinae*) each have 29 core genes, for genomes encoding 130 to 334 and 95 to 239 genes, respectively. By contrast, *Marseilleviridae* share 289 core genes (genomes encoding 403 to 484 genes), and Pandoraviruses/*Mollivirus* – 112 (523 genes for *Mollivirus*; 1,497 and 2,541 genes for Pandoraviruses). *Mimiviridae* share 59 core genes (for 544 to 1,545 genes encoded in their genomes), while the *Mimiviridae*-related viruses

have 46 (for 326 to 512 genes in their genomes); together within the putative order "Megavirales" they would share 25 core genes. These discrepancies could reflect comparable biological constraints for viruses belonging to the same clade, or the level of represented diversity from the isolates/reconstructed genomes.

Viral phylogenies

Because of the divergence generally observed between homologous proteins from different viral families, building a viral phylogeny is not a trivial task. This holds true in the case of the NCLDVs, despite the presence of 3 strictly conserved protein-coding genes and 5 highly conserved genes. Notably, among these 8 core proteins, the two transcription factors (TFIIS and VLTF3-like) produced the least supported trees (trees in Additional data at https://doi.org/10.5281/zenodo.3368642); this was however not unexpected as they represent the shortest markers. In these two trees, several families were not monophyletic, with one or two taxa branching outside of well recognized families. In the trees obtained from the larger markers, some incongruences were also observed. In the primase phylogenetic tree, Cedratvirus A11 is notably branching next to the *Asfarviridae*, while *Heterosigma akashiwo* virus is branching outside – but next to – the *Phycodnaviridae*. The *Ascoviridae* are paraphyletic in the MCP tree, just as the *Phycodnaviridae*. The latter are also paraphyletic in the pATPase tree. In all our trees, the *Poxviridae* had long branches and were ambiguously located with varying positions. Not only their position seems difficult to confidently determine, but also their mere presence in datasets is a potential source of bias. Notably, their inclusion in the analyses often reduces the branch supports at most nodes. Their frequent grouping with the *Asfarviridae* (3/8 of the single protein trees) could hence result from an attraction with the clade displaying the second longest branch (trees in Additional data at

https://doi.org/10.5281/zenodo.3368642). Interestingly, when *Poxviridae* were included in the phylogenetic analysis based on 8 concatenated markers in the ML framework, they branched between the MAPI and PAM putative superclades (SI Appendix, Fig. S15), a position reminiscent of that of Polintoviruses with the concatenated structural proteins (SI Appendix, Fig. S5). This suggests that *Poxviridae* diverged from NCLDVs before the separation between these two superclades. However, this position can also be due to their long branches and one cannot exclude that *Poxviridae* also belong to one of them. One possibility is that *Poxviridae* have undergone an evolutionary history more complex than other NCLDV families, at least concerning the core proteins investigated herein. It is possible that they acquired their genes through several horizontal transfers from other viral families, just like they do with genes of eukaryotic origin (6–8). This could explain why *Poxviridae* have very long branches in all trees. The evolution of *Poxviridae*, as well as their position among NCLDVs, thus remains to be elucidated.

A similar situation was observed with *Aureococcus anophagefferens* virus. This virus is known to be hard to position with confidence. Moniruzzaman and colleagues suggested that this virus could be related to "Megavirales", based on the core gene phylogenies and comparative genomic analyses (9). While indeed located close to *Mimiviridae* in 5 out of 8 of our individual protein trees, it was branching at various other positions for the 3 other markers (trees in Additional data at https://doi.org/10.5281/zenodo.3368642). Considering its long branch, we thus removed *Aureococcus anophagefferens* virus from the dataset. Removing both the *Poxviridae* and *Aureococcus anophagefferens* virus improved greatly the resolution of the single-protein trees, which were much better supported and more congruent, especially in terms of relationships between the viral families (SI Appendix, Fig. S1). This is

particularly noticeable when the trees are rooted between the MAPI and the PAM superclades. Despite the paraphyly of *Phydodnaviridae* in the TFIIS and MCP trees, comparative phylogenetic tests, based on all possible combinations of 6 out of 8 markers, did not detect any major incongruence between the different combinations of core proteins (SI Appendix, Table S2). These results warranted concatenation of the 8 marker genes to determine the global NCLDV phylogeny. We thus performed Bayesian inferences with the CAT-GTR model, designed to deal with site and sequence heterogeneities, and obtained chains that reached a stable and good convergence according to the software's manual (maxdiff <0.1). The very robust resulting tree had all nodes but two minor ones at maximum support (PP=1), and thus appears much more reliable than a tree we obtained with the same dataset but using the ML framework (SI Appendix, Fig. S16).

In parallel, we constructed a supertree based on the subtree prune-and-regraft (SPR) distances (see Methods; SI Appendix, Fig. S2). This method has been designed to help to recover the species tree despite the presence of transfers and is entirely independent of any concatenation since the reconstructed tree is directly based on the single-protein phylogenetic trees. Considering the transfers that occurred for the RNA polymerases, and the still possible presence of hidden conflicting signals, such an approach could indeed be useful. Both approaches, the Bayesian inferences and the SPR Supertree, produced strikingly identical phylogenetic trees, adding strong confidence in the obtained topology and again supporting the absence of conflicting signals within the core genes. This implies that these trees likely represent the vertical evolution of NCLDVs' core genes and that the informational proteins within it co-evolved with the markers involved in virion formation. We hence separately concatenated these two sets of proteins: the DNA polB, RNAP-a and -b, and the primase on one side (considering the previous results obtained in single-protein trees, we did not include the short and

126 possibly confusing VLTF3-like and TFIIS markers), and the MCP along with the pATPase
127 on the other hand. In both trees (SI Appendix, Fig. S3 and S4), all NCLDV families were
128 monophyletic, except for the *Iridoviridae* which were split by the *Ascoviridae* in the tree
129 constructed from the concatenation of informational proteins (SI Appendix, Fig. S3). The
130 two phylogenies had similar topologies, with the same clusters of NCLDV families as
131 observed in the trees obtained from Bayesian inferences and SPR Supertree
132 reconstruction (Fig. 1; SI Appendix, Fig. S2). Some positions within these clusters might
133 be affected by differences between the two datasets: 2 of the 4 informational proteins are
134 absent in all but one *Phycodnaviridae* genus, while the Pitho-like viruses lack the pATPase
135 gene. The congruence between the two trees still supports the co-evolution of the
136 informational markers with those involved in virion formation.
137
138 The robust tree we obtained (Fig. 1) calls for a reconsideration of taxonomy and
139 nomenclature among the NCLDVs. This is particularly true for the *Asfarviridae*, initially
140 comprising the *African swine fever* virus only but now including amoeba-infecting viruses.
141 Similarly, the *Phycodnaviridae* clade groups very diverse marine viruses, infecting not
142 only algae but also protists with pandoraviruses and mollivirus, raising questions about
143 their taxonomic-level and their actual monophyly. One of the most robust clusters, but
144 also one of the most confusing ones with regard to its nomenclature, corresponds to the
145 *Mimiviridae* with a clade of related viruses infecting algae and referred to as the "extended
146 Mimiviridae" (10) or "Mesomimivirinae" (11). We proposed herein to name this cluster
147 the "Megavirales" order, since the vast majority of this cluster is currently represented by
148 giant viruses. The term "Megavirales" has already been proposed with different
149 definitions. For instance, Arlsan and colleagues (12) proposed a name "Megaviridae" to
150 refer to the giant DNA viruses with genome sizes larger than 1 Mb. However, the latter

virus group corresponds to the previously created and officially recognized *Mimiviridae* family, and is thus unjustified (but still used in literature, albeit rarely). One year later, Coslon and colleagues (13) proposed to unify the families included in the NCLDV assemblage into the "Megavirales" order, on the basis of phylogenetic reconstructions and conserved features. This name has not been officially adopted though, and one could argue that most families among the NCLDVs do not encompass any truly giant viruses. The definition we propose herein somewhat matches the one previously described by Santini, Moniruzzaman, and their respective colleagues with the "Megaviridae" family (9, 14), except that we raised it to the taxonomic rank of order, so as to remain consistent with the current ICTV classification comprising the *Mimiviridae* family.

The DNA-dependent RNA polymerase

The two largest subunits of DNA-dependent RNA polymerase (RNAP) are the largest universal markers and are present in all three cellular domains. As such, they are good candidates to study deep phylogenies such as the relationships between cells and NCLDVs. However, unlike Bacteria and Archaea that have a single polymerase processing every type of RNAs, all eukaryotes have three different RNA polymerases: one responsible for the synthesis of ribosomal RNA (except 5S rRNA) (RPA), another responsible for the synthesis of mRNA (RPB), and a third responsible for the synthesis of transfer RNA and small rRNA (RPC). To avoid confusion with the alphabetical names of the subunits, we used only the Roman numbers in this manuscript: RNAP-I, RNAP-II, and RNAP-III, respectively. The nomenclature of the RNAP subunits is especially confusing with the two largest subunits being respectively named β' and β in Bacteria, A and B in Archaea, 1 and 2 in Eukaryotes, and alpha and beta in NCLDVs. For clarity, we decided to name all of them *a* and *b* here.

The second largest subunit, RNAP-b, has already been used in different controversial studies discussing whether NCLDVs correspond to a fourth domain of life (15–17). The first study, performed by Boyer and colleagues, displayed a RNAP-b phylogenetic tree in which the NCLDVs form a separate monophyletic clade close to Eukaryotes, prompting them to claim that NCLDVs should be considered as a fourth domain of life (based on other protein trees as well) (15). Their analyses of the RNAP-b comprised 272 aligned positions for 80 taxa. In these trees, Archaea were, however, paraphyletic (and underrepresented, with only 2 members of the phylum Euryarchaeota), many nodes were unsupported, and some phyla (especially in Bacteria and some NCLDVs) presented very long branches. In particular, *Candidatus* Korarchaeum cryptofilum was branching with Bacteria, suggesting the presence of a long branch attraction artefact (LBA). This study was criticized by Williams and colleagues, who suggested that the monophyly of NCLDV in the tree of Boyer and colleagues was probably due to the use of inappropriate models of protein evolution (JTT+CAT in maximum-likelihood, and WAG in Bayesian inferences) (16). From the same dataset (80 taxa and 272 positions), Williams and colleagues performed a Bayesian inference with a model better suited to deal with heterogeneity (CAT60) and obtained a tree in which NCLDVs were no longer monophyletic. While one group was still branching between Archaea and Eukaryotes, the others were branching among Eukaryotes. Their tree nonetheless still displayed the paraphyly of underrepresented Archaea and low supports. *Ca*. Korarchaeum cryptofilum was this time branching next to Eukaryotes/NCLDVs, still suggesting an LBA. Furthermore, the tree contained many polytomies, and *Poxviridae* still presented a significantly longer branch. A few years later, Sharma and colleagues obtained again RNAP-b phylogenies similar to those obtained by Boyer and colleagues (with NCLDV monophyletic) using the same dataset enriched with new NCLDV sequences

(15, 18–20). However, they only performed maximum-likelihood analyses with the WAG model.

At the same time, Moreira and Lopez-Garcia proposed a re-analysis of the RNAP-b, and suggested that the previous studies were affected by poor taxon sampling (17). As a consequence, they added several new taxa, mostly eukaryotes. In parallel, they removed Bacteria and used Archaea as the outgroup. This allowed them to increase the number of aligned positions to 427 positions for 127 taxa. Their tree, performed in Bayesian framework with the CAT model, displays the Archaea as monophyletic, and the NCLDVs branching at various positions among the Eukaryotes. The authors concluded that the RNAP-b was acquired several times independently by NCLDVs after the emergence of modern eukaryotes, in agreement with their views that large DNA viruses are mainly pick-pockets of cellular genes that were rather recently acquired in the history of life (21). However, their tree is poorly supported (with many nodes having posterior probabilities values below 0.9). Furthermore, the resolution of the intra-domain phylogenies was not recovered, with for instance, Thaumarchaea and Euryarchaea branching within Crenarchaeota in Archaea. The eukaryotic part of the tree was not resolved, with many very short branches, possibly because it was strongly enriched in fast-evolving species (such as Cryptomonads). Several consensus NCLDVs families, such as the *Iridoviridae*, were not monophyletic. Finally, the viruses were never branching close to their known or supposed host, in contradiction with the “pick-pocket hypothesis”.

A common feature for these analyses was the very limited number of positions for the RNAP-b. This protein is usually between 1,000 and 1,500 amino-acid long, yet the alignments were 272 positions-long for 80 sequences in the two first studies (15, 16) and up to 427 positions for 127 taxa in the third (17). The analysis of Sharma and colleagues in 2014 similarly included 420 positions for 99 sequences (including Bacteria) (18). This

indicates very stringent conditions for trimming the aligned sequences, an approach known for drastically reducing the signal carried by the protein, potentially up to the point where it cannot be differentiated from mere noise (22).

Notably, all these analyses included only one eukaryotic RNAP (mostly RNAP-II). In 2010, Lane and Darst included all of them with viral sequences in their analyses, yet their work was specifically oriented on the conservation of domains within the RNAP genes with a special focus on Bacteria (23). The only study on the NCLDV evolution that included the three eukaryotic RNAP was published in 2012 by Yutin and Koonin (1). They obtained phylogenetic trees very similar to our single subunit trees (the number of positions for each subunit was, however, not mentioned). They concluded that the ancestral NCLDV RNAP-a possibly derived from the eukaryotic RNAP-Ia before being replaced in *Mimiviridae* and *Asfarviridae* by eukaryotic RNAP-IIa and Ia, respectively. The second largest subunit, according to their results, could either display the NCLDVs as polyphyletic or monophyletic, with a more recent transfer of RNAP-IIb to the *Mimiviridae*. Their analyses, published in 2012, were however lacking some representatives that were isolated or described more recently, and the analyses were performed in the ML framework with limited options concerning the models. In addition, the *Poxviridae* were still included, and the results were essentially interpreted as a modular evolution, in the sense that genes were systematically analysed separately, congruence between trees was not considered, nor concatenations performed.

Our RNAP analyses were performed with considerations for the above-mentioned issues. We also performed topology tests (Approximately Unbiased tests) against trees constrained for the monophyly of cellular sequences or NCLDVs sequences. These alternative topologies were rejected, reinforcing the confidence in our RNAP phylogeny

250 in which the NCLDV assemblage is not monophyletic but nested between the different
251 clades of eukaryotic RNAPs (Fig. 2).

252 Our results strongly suggest that the true eukaryotic ortholog of archaeal and
253 bacterial RNAP is actually the eukaryotic RNAP-III. This is in line with the presence in
254 Archaea of a homologue of the RNAP-III specific subunit, RPC34 (24, 25). Genes encoding
255 these archaeal proteins (dubbed TFE-β) (25) were initially reported in Crenarchaeota,
256 Thaumarchaeota and some Euryarchaeota (24) and later on in Asgard archaea (26).
257 Interestingly, we failed to detect homologues of these proteins in NCLDVs. This suggests
258 that this subunit was lost during the recruitment of the proto-eukaryotic RNAP by the
259 ancestor of NCLDVs.

260 Our global RNAPs tree displays three clades of NCLDVs, corresponding to i) the
261 monophyletic MAPI superclade, which is a sister group to the *Phycodnaviridae*, ii) the
262 "Megavirales", and iii) the *Asfarviridae*. Notably, the RNAP tree does not recover the
263 monophyly of the PAM supergroup and the rooting between the PAM and MAPI obtained
264 in the MCP-pATPase tree using Polintonviruses as an outgroup. Instead, while the relative
265 positions of the NCLDV families are still matching the topology obtained in the absence of
266 cellular sequences (SI Appendix, Fig. S11 and S12), the RNAP phylogeny suggests rooting
267 the NCLDV tree between the *Asfarviridae* and all other NCLDVs, using eukaryotic RNAP-
268 III/Archaea as outgroups. This suggests that the rooting of the NCLDV tree remains an
269 open question. However, we noticed that the RNAP-based rooting suffers two
270 weaknesses: i) one cannot exclude an attraction of the long branches of the
271 *Asfarviridae*/RNAP-I assemblage by outgroup sequences (Archaea, RNAP-III), and ii) the
272 absence of RNAP genes in most *Phycodnaviridae* could have influenced the position of the
273 root. Thus, in our evolutionary scenario (Fig. 3), we used the rooting between the MAPI

and PAM supergroups, but further investigations will be required to confirm or disprove this particular rooting.

Considering the paraphyly of the PAM superclade in the viral/cellular RNAP tree, the position of the *Phycodnaviridae* as a sister group to the MAPI superclade could be due to insufficient signal due to their low representation for these specific markers, but could also suggest an early replacement of their RNAP by the ancestral MAPI variant. We hence performed a ML phylogenetic reconstruction of the concatenation of the two RNAP subunits from the NCLDVs and used the eukaryotic RNAP-III as an outgroup (SI Appendix, Fig. S17). In this tree, the *Phycodnaviridae* are branching before the MAPI and the "Megavirales"/*Asfarviridae* bipartitions. This branching pattern is not consistent with the transfer of RNAP from the MAPI superclade to the *Phycodnaviridae*; nonetheless, the *Phycodnaviridae* are not branching with the other PAM families either. It is thus possible that this virus family indeed acquired a NCLDV-like RNAP complex from a different currently unknown source more closely related to the MAPI superclade. However, the most parsimonious scenario fits with our hypothesis depicted in Fig. 3, which posits the emergence of the *Phycodnaviridae* shortly after the separation between the MAPI and the PAM superclades. The RNAP of the "Megavirales"/*Asfarviridae* common ancestor has followed a specific evolutionary trajectory, whereas the *Phycodnaviridae* retained a RNAP complex more similar to the NCLDV and MAPI ancestral variants. It should be noted that, at the moment, alternative scenarios for the origin of the *Phycodnaviridae* RNAP cannot be ruled out with confidence. Furthermore, the absence of this complex in all genera but the *Coccolithovirus* genus could suggest a specific evolutionary pathway. Altogether, their low representation in the RNAP phylogeny calls for caution when interpreting their position, and further data would be needed to resolve this uncertainty.

298 The concatenated RNAP-subunits tree, along with the trees obtained through
299 consensus bootstrap and ancestral sequence reconstruction (SI Appendix, Fig. S13),
300 strongly support the relationships between the eukaryotic RNAP-I and -II with the
301 *Asfarviridae* and the "Megavirales", respectively. If the bipartition corresponding to the
302 MAPI superclade is still strongly supported in both the two single-subunit phylogenetic
303 trees, these latter offered more contrasted information regarding the relationships
304 between the cellular and viral RNAP-subunits. Indeed, the RNAP-I and -II are sister clades
305 to *Asfarviridae* and "Megavirales", respectively, in the *a*-subunit tree (SI Appendix, Fig.
306 S8), whereas the RNAP-II alone is a sister group to a clade encompassing both *Asfarviridae*
307 and "Megavirales" (the former being nested the latter) in the *b*-subunit tree. In this tree,
308 the *b*-subunit of the eukaryotic RNAP-I is branching with the RNAP-III.
309 Our results strongly suggest that horizontal transfers occurred for the largest
310 RNAP subunit (RNAP-a) between (i) "Megavirales" and eukaryotic RNAP-II, and (ii)
311 *Asfarviridae* and eukaryotic RNAP-I. The second largest subunit, RNAP-b, was also
312 horizontally transferred between eukaryotic RNAP-II and a clade including both
313 "Megavirales" and *Asfarviridae.* It is possible that the two subunits were simultaneously
314 transferred between the proto-eukaryotes and the common ancestor of "Megavirales"
315 and *Asfarviridae* before the largest subunit was later again transferred between
316 *Asfarviridae* and cells. Alternatively, it is possible that the RNAP-a and -b were transferred
317 separately from the beginning, but this seems less likely considering the multimeric
318 nature of RNAPs. Interestingly, the RNAP trees are fully compatible with the concatenated
319 markers trees. The transfer of RNAP-b between proto-eukaryotes and a clade grouping
320 *Asfarviridae* and "Megavirales", but not with *Phycodnaviridae*, is coherent with the
321 Bayesian inference (CAT-GTR model) (Fig. 1) and the SPR supertree obtained with the

concatenated markers and showing the sisterhood of “Megavirales” and *Asfarviridae* (SI Appendix, Fig. S2).

Importantly, the comparative phylogenetic test we performed for the markers suggested a strong congruence between the NCLDV tree topologies of every possible combination of 6 markers out of 8, hence including a concatenation lacking the two RNAP subunits (that otherwise correspond to 47% of the positions in the total alignment). This shows that the signal corresponding to the global concatenation is not only carried by the two RNAP subunits (that would have oriented the final topology toward their own.) but also by the other markers that were not subject to the transfers. This strongly suggests that the core genes were vertically inherited in all modern NCLDV families. In other words, the obvious important horizontal exchanges that occurred for RNAP-a and -b apparently did not perturb the signal likely to represent the NCLDV vertical evolution, and the RNAP trees were still congruent with the other concatenations. Notably, a similar topology is obtained with all the markers, with and without the RNAP genes (Fig. 1 and SI Appendix, Fig. S10, respectively), and with the viral RNAP genes only (SI Appendix, Fig. S11). Despite these transfer events involving two major clades of NCLDVs, the topology of the concatenated RNAP-subunits tree still matches the topology of NCLDVs from most trees in our study, as shown in SI Appendix, Fig. S12. Considering the proportion of positions corresponding to the RNAP genes in the concatenation, major cell-to-virus transfers in these two markers would have likely impacted the topology of NCLDVs. The absence of substantial impact on the NCLDV tree topology, even from the position of *Asfarviridae*, seems unlikely in the events of transfers from cells to viruses as proposed by Yutin and Koonin (1). On the contrary, this strongly suggests that the transfers of RNAPs between cells and viruses were oriented from the latter to the former. This would also

346 explain why the RPC34 subunit, lost in NCLDVs, is not associated with eukaryotic RNAP-I
347 and II.

348 In addition, considering the two main alternative scenarios involving transfers of
349 the eukaryotic RNAP-I and -II to the *Asfarviridae* and the "Megavirales", replacing their
350 ancestral NCLDV RNAP more alike modern eukaryotic RNAP-III, would have likely led to
351 different topologies for the RNAP phylogenetic trees (SI Appendix, Fig. S14). If the
352 eukaryotic RNAP-I and -II emerged by duplication events before the first transfer of RNAP
353 to the ancestor of NCLDVs, one could expect the two large subunits to carry a congruent
354 signal for a clade grouping the eukaryotic RNAP-I and -II with the *Asfarviridae* and the
355 "Megavirales", and for another clade with the eukaryotic RNAP-III and the
356 *Phycodnaviridae* and the MAPI putative superclade. On the opposite, a first transfer to the
357 ancestor of NCLDVs occurring before the emergence by duplication of the eukaryotic
358 RNAP-I and -II would have likely induce a congruent signal in the two subunits for a clade
359 encompassing the three eukaryotic RNAPs with the *Asfarviridae* and the "Megavirales",
360 and another clade containing the *Phycodnaviridae* and the MAPI putative superclade.
361 None of these clades were observed in our RNAP phylogenies, adding more credit to our
362 hypothetical scenario for the transfers of RNAPs.

363

364 **Evolution of NCLDVs**

365 Our results, displaying a robust phylogeny of NCLDVs, highlight particular points
366 about their evolution that had been debated. Notably, with Pandoraviruses related to
367 *Phycodnaviridae* and giant *Mimiviridae* encompassed within the "Megavirales" order with
368 smaller related viruses, it appears that gigantism in viral genomes was not a unique event,
369 but occurred at least twice independently within the PAM superclade. In addition,
370 *Orpheovirus*, a member of the Pitho-like group in the MAPI superclade, also exhibits a

giant genome at odds compared to related viruses such as *Cedratvirus* and *Pithovirus*, which still produce giant particles but encapsidate smaller genomes. Even though more genomes/viruses belonging to this family are necessary to understand the directionality of evolution and extent of its actual diversity, the giant genome of Orpheovirus suggests that the switch toward the accumulation of genes also occurred independently in the Pitho-like virus lineage. This is in contradiction with the hypotheses advocating a giant cellular or viral ancestor of NCLDVs that evolved through parasitic reduction (27, 28). This scenario would indeed involve the parallel reduction in many different viral families and sub-families from a giant NCLDV ancestor, or potentially a giant PRD1-Adenovirus lineage ancestor, and would thus be less parsimonious given that many viruses of this lineage infect bacteria and archaea with comparatively small genomes and cell sizes. In contrast, our results favour models in which NCLDV genomes evolved from a smaller ancestor by successive steps of genome reduction and expansion (29, 30). Genome expansion in giant viruses could be related to host-virus interactions in the context of hosts evolving themselves toward gigantism, a situation favouring exchanges of genetic material, gene family expansion and *de novo* emergences of viral genes, as hypothesized for some years (31, 32) and demonstrated in Pandoraviruses more recently (33). Up to now, all giant viruses have been isolated in amoeba (but not all viruses isolated in amoebas are giant), and even if this corresponds to a methodological bias and primary hosts are still essentially unknown, it is reasonable to consider that these viruses naturally infect phagotrophic organisms where similar genetic dynamics are possible. Additional representatives from different NCLDV families and studies on virus-host interactions are necessary to unveil the prerequisite conditions for a virus to become giant.

Altogether, our different results prompted us to elaborate a putative scenario of the NCLDVs evolution compatible with our observations (Fig. 3). We hypothesize that the smaller ancestor of NCLDVs acquired an RNA polymerase complex from a proto-eukaryotic host soon after the divergence of the latter from Archaea. This ancestral eukaryotic polymerase corresponds to modern RNAP-III, the actual ortholog of archaeal and bacterial RNAPs, and was able to switch its transcription toward coding or non-coding RNAs. Later on, this lineage of ancestral NCLDV viruses split into two groups, the MAPI and the PAM superclades. From the MAPI superclade then emerged different modern families, the *Marseilleviridae*, Pitho-like viruses, *Iridoviridae*, and later *Ascoviridae*, without any major transfers involving the core genes analysed in our study. On the other side, the PAM superclade first divided into proto-*Phycodnaviridae* and the common ancestor of the "Megavirales" and *Asfarviridae*. Proto-eukaryotes acquired from this latter group a new RNAP complex (at least the two largest subunits) that was already or subsequently became specialized towards the transcription of mRNA (RNAP-II). After the emergence of the specific *Asfarviridae* and "Megavirales" clades, the largest subunit of the new proto-eukaryotic RNAP (RNAP-I) that potentially originated by a duplication event from RNAP-III, was transferred between the *Asfarviridae* and the proto-eukaryotes.

Regardless of the hypothetical scenario considered for the orientation of the transfers, they occurred between NCLDVs and proto-eukaryotes, and the diversification of NCLDVs predated that of modern eukaryotes.

Supplementary Legends

Fig. S1. Maximum likelihood (ML) single-protein trees of the 8 core genes from the NCLDVs after removal of the *Poxviridae* and of *Aureococcus anophagefferens* virus.
Fig. S2. Supertree of the 8 core proteins from the NCLDVs.
Fig. S3. Maximum likelihood (ML) phylogenetic tree of the concatenated informational proteins from NCLDVs.
Fig. S4. Maximum likelihood (ML) phylogenetic tree of the concatenated structural proteins from NCLDVs.
Fig. S5. Relationships between Polintoviruses and NCLDVs.
Fig. S6. Maximum likelihood (ML) phylogenetic tree of the concatenated two largest RNA polymerase subunits from Bacteria, Archaea, and Eukaryotes, including the 3 eukaryotic polymerases.
Fig. S7. Maximum likelihood (ML) phylogenetic tree of the concatenated two largest RNA polymerase subunits from Bacteria, Archaea, Eukaryotes, and NCLDVs.
Fig. S8. Maximum likelihood (ML) single-protein trees of the two largest RNA polymerase subunits from Archaea, Eukaryotes, and NCLDVs.
Fig. S9. Maximum likelihood (ML) phylogenetic tree of the concatenated two largest RNA polymerase subunits from Archaea, Eukaryotes, and NCLDVs.
Fig. S10. Maximum likelihood (ML) phylogenetic tree of the concatenation of all core proteins but the two RNAP subunits from NCLDVs.
Fig. S11. Maximum likelihood (ML) phylogenetic tree of the concatenated two largest RNA polymerase subunits from NCLDVs.
Fig. S12. Schematic representation of the congruence in NCLDV topologies obtained before and after the inclusion of cellular sequences in the concatenated RNAP-subunits tree.
Fig. S13. Phylogenetic trees of the concatenated two largest RNA polymerase subunits from Archaea, Eukaryotes, and NCLDVs, obtained through consensus bootstrap reconstruction (left) and maximum likelihood (ML) with ancestral sequences reconstructed (right).
Fig. S14. Schematic representations of two alternative scenarios for the transfers of RNAPs from cells to viruses with the congruent signals expected from the two subunits.
Fig. S15. Maximum likelihood (ML) phylogenetic tree of the concatenated 8 core genes from the NCLDVs, including *Poxviridae*.
Fig. S16. Maximum likelihood (ML) phylogenetic tree of the concatenated 8 core genes from the NCLDVs.
Fig. S17. Maximum likelihood (ML) phylogenetic tree of the concatenated two largest RNA polymerase subunits from the NCLDVs and the eukaryotic RNAP-III.

Table S1. List and access numbers of NCLDV genomes included in this study.
Table S2. Results of the comparative phylogenetic analyses (congruence test), based on the presence/absence of reference features in the ML phylogenetic trees of every possible concatenations of 6 out of 8 markers (systematically referred by the two missing genes).
Table S3. List and taxon IDs of the cellular taxa used in this study.

Not included in SI Appendix:
Additional data. Sequence and tree files, and table listing the core genes and their access numbers among the NCLDV families (accessible https://doi.org/10.5281/zenodo.3368642).

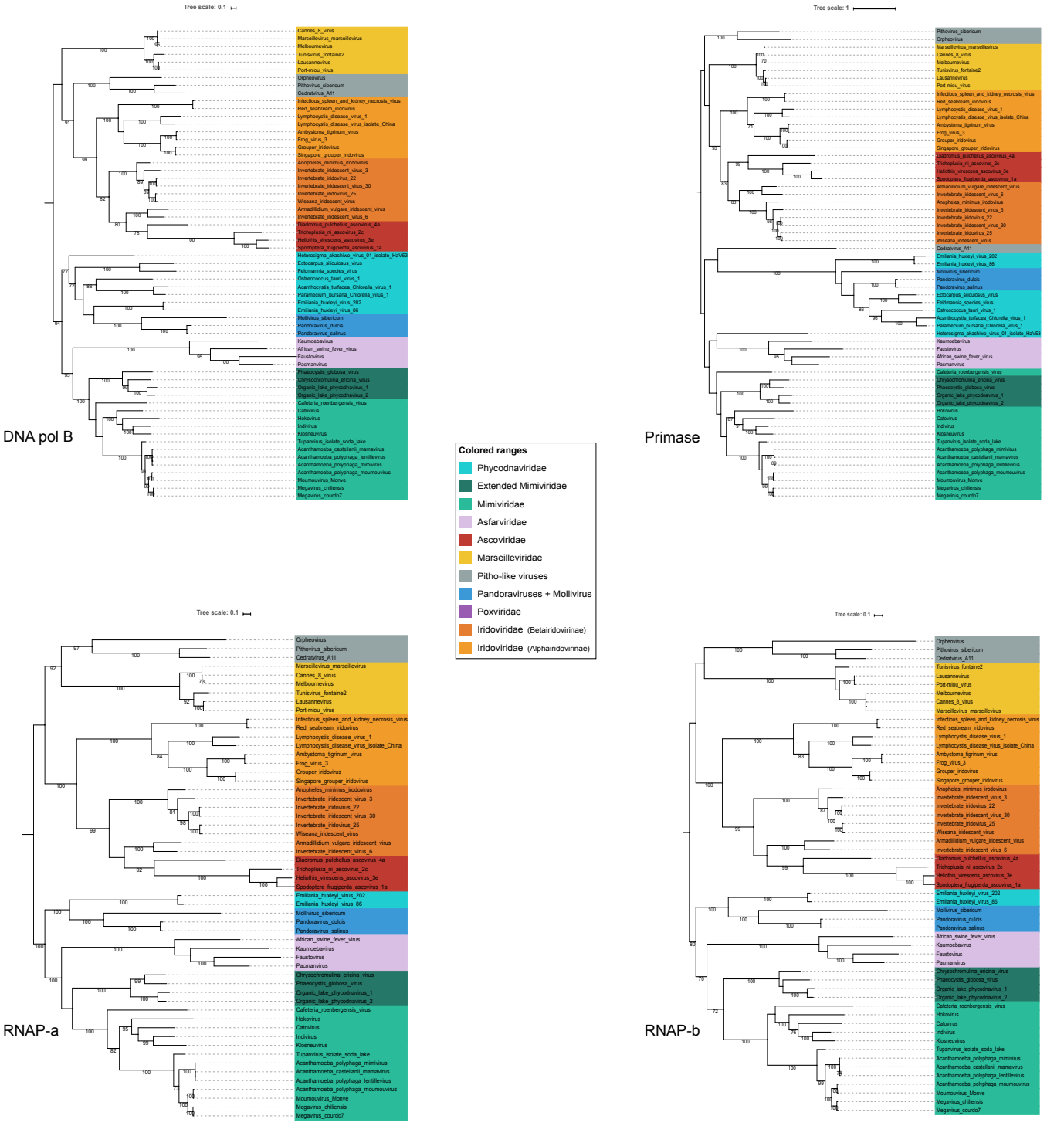
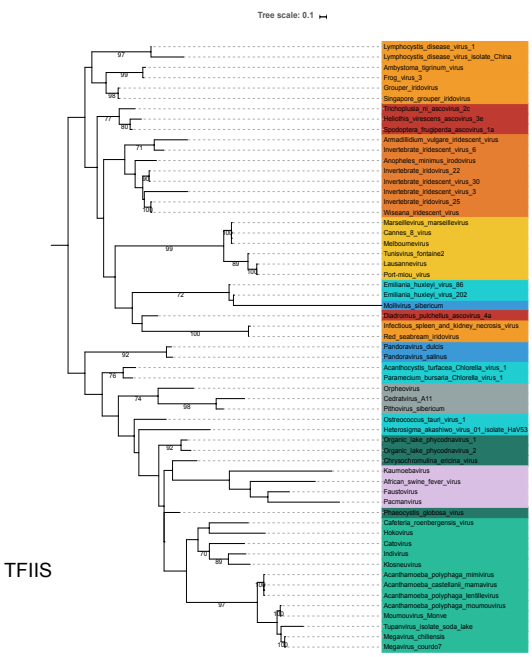
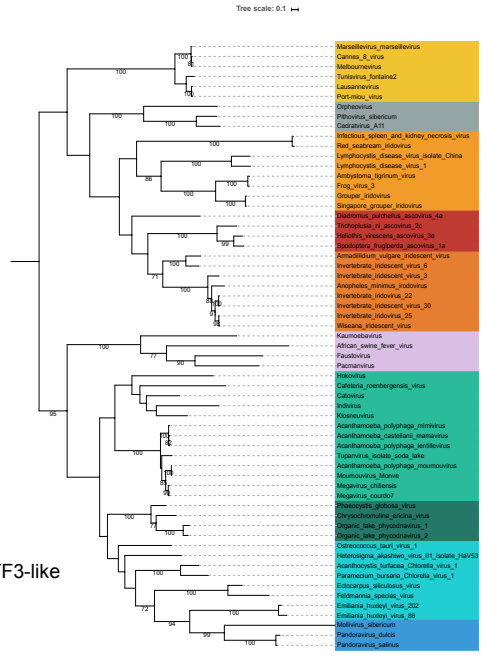


Fig. S1. Maximum likelihood (ML) single-protein trees of the 8 core genes from the NCLDVs after removal of the *Poxviridae* and of *Aureococcus anophagefferens* virus. The scale-bars indicate the average number of substitutions per site. Values on branches represent support calculcated by nonparametric bootstrap; only supports superior to 70% are shown. The trees are rooted between the PAM and the MAPI putative superclades.



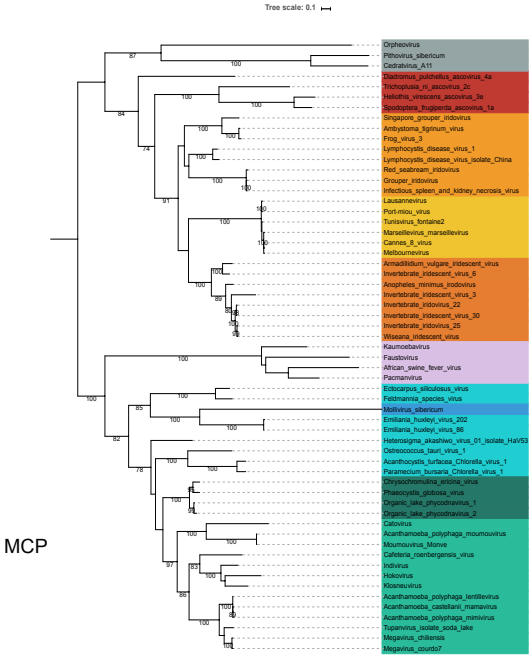
TFIIS



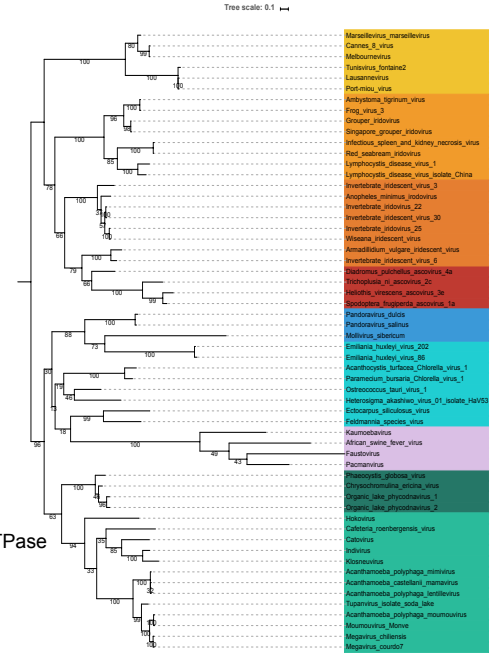
VLTF3-like

Colored ranges

- Phycodnaviridae
- Extended Mimiviridae
- Mimiviridae
- Asfarviridae
- Ascoviridae
- Marseilleviridae
- Pitho-like viruses
- Pandoraviruses + Mollivirus
- Poxviridae
- Iridoviridae (Betairidovirinae)
- Iridoviridae (Alphairidovirinae)



MCP



pATPase

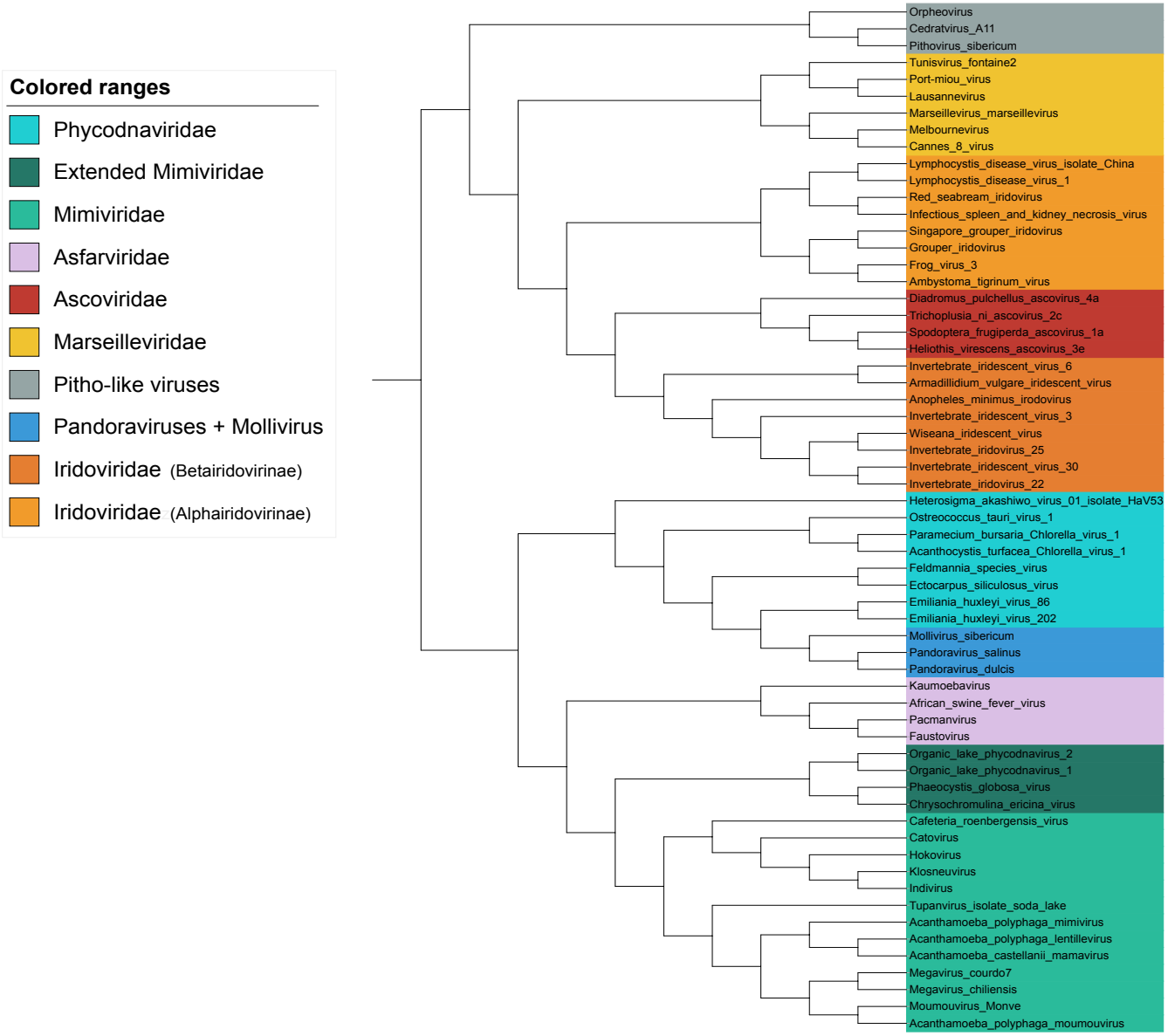


Fig. S2. SPR Supertree of the 8 core proteins from the NCLDVs. Supertree based on the subtree prune-and-regraft (SPR) distance from the DNA pol B, Primase, RNAP-a, RNAP-b, MCP, pATPase, TFIIS, and VLTF3-like sequences from NCLDVs after removal of *Poxviridae* and *Aureococcus anophagefferens* virus.

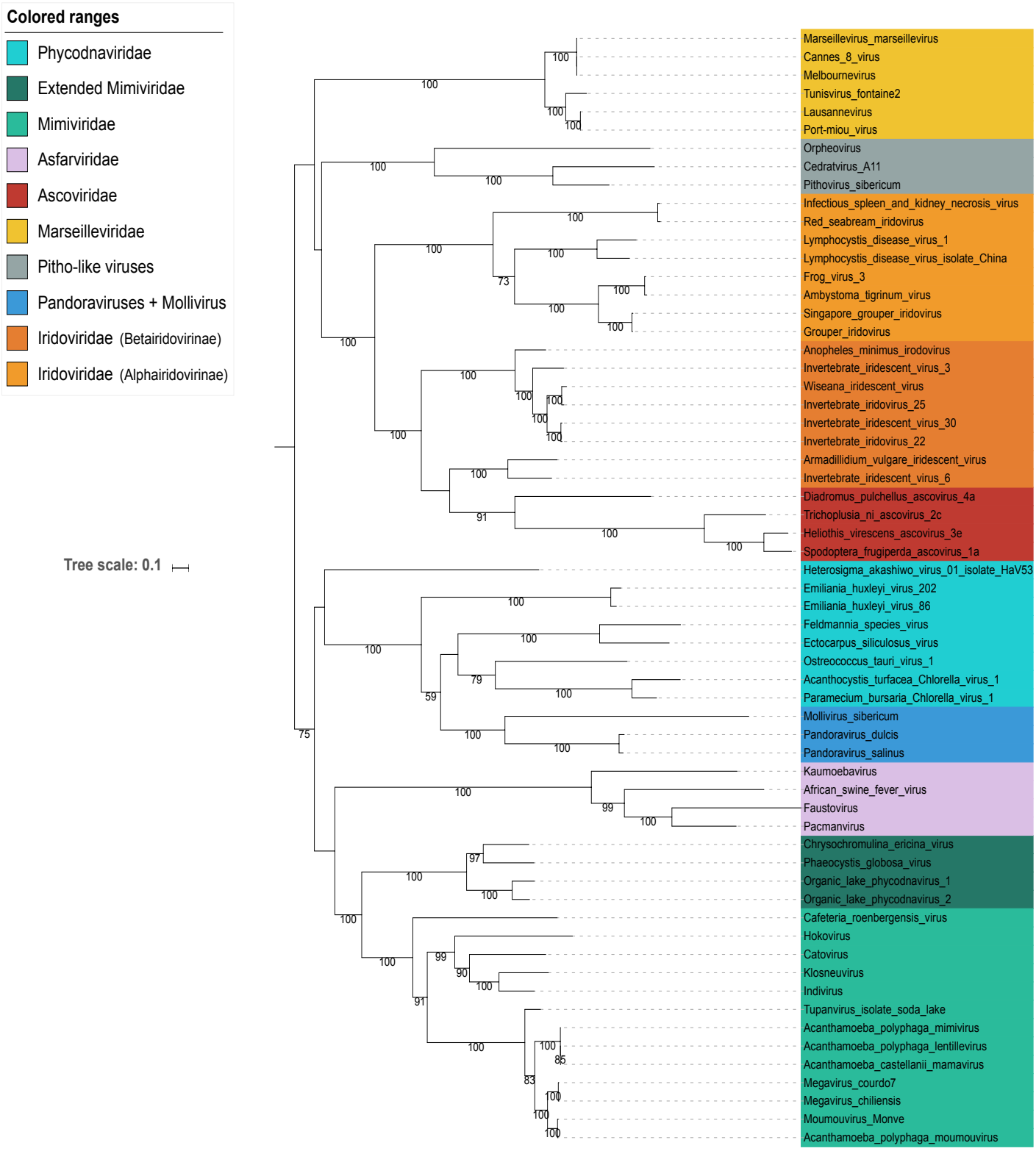


Fig. S3. Maximum likelihood (ML) phylogenetic tree of the concatenated informational proteins from NCLDVs. ML phylogenetic tree of the concatenation of the DNA pol B, Primase, RNAP-a, and RNAP-b sequences from NCLDVs after removal of *Poxviridae* and *Aureococcus anophagefferens* virus. The scale-bar indicates the average number of substitutions per site. Values on branches represent support calculated by nonparametric bootstrap; only supports superior to 70% are shown.

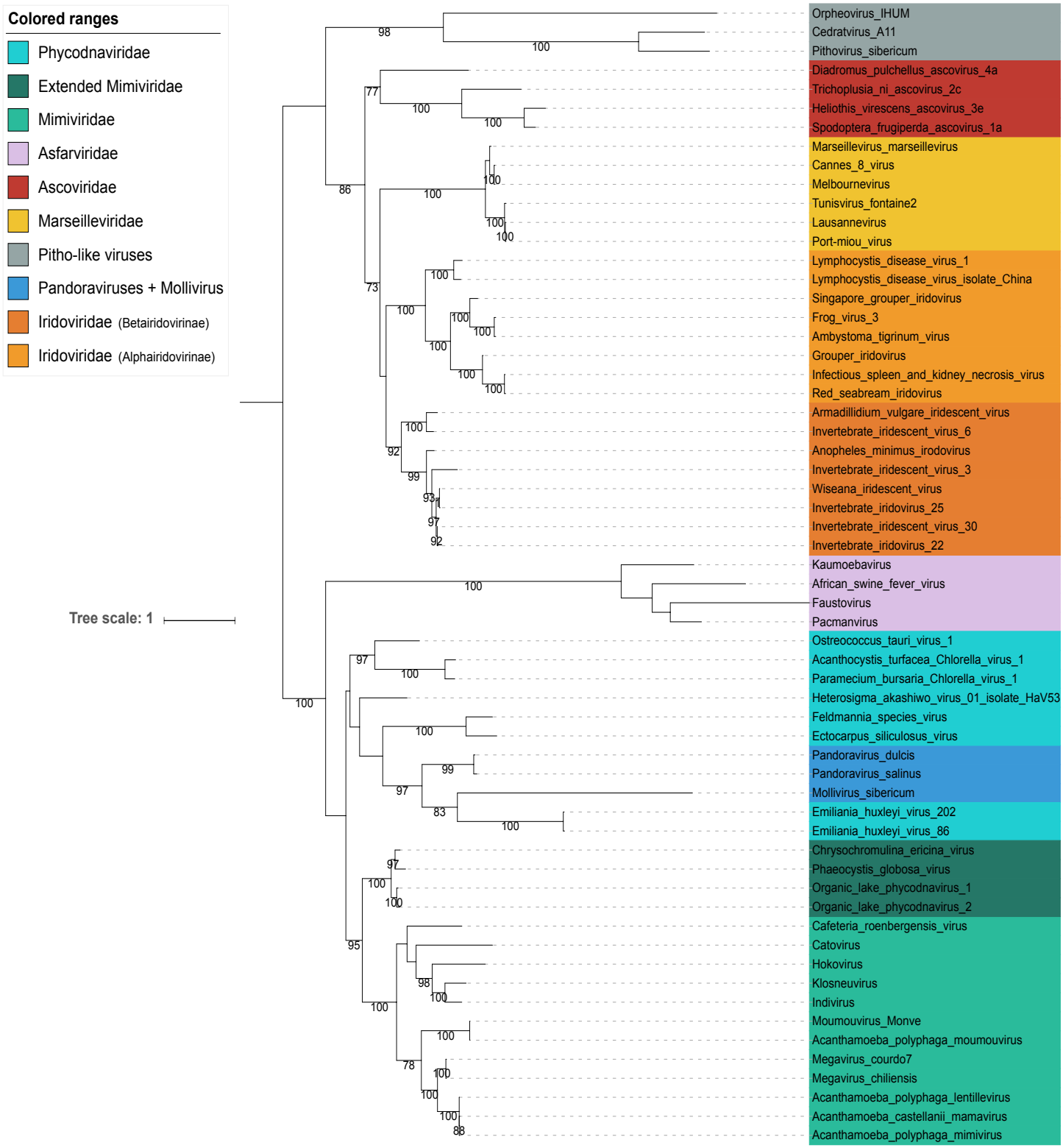


Fig. S4. Maximum likelihood (ML) phylogenetic tree of the concatenated structural proteins from NCLDVs. ML phylogenetic tree of the concatenation of the MCP and pATPase from NCLDVs after removal of *Poxviridae* and *Aureococcus anophagefferens* virus. The scale-bar indicates the average number of substitutions per site. Values on branches represent support calculated by nonparametric bootstrap; only supports superior to 70% are shown.

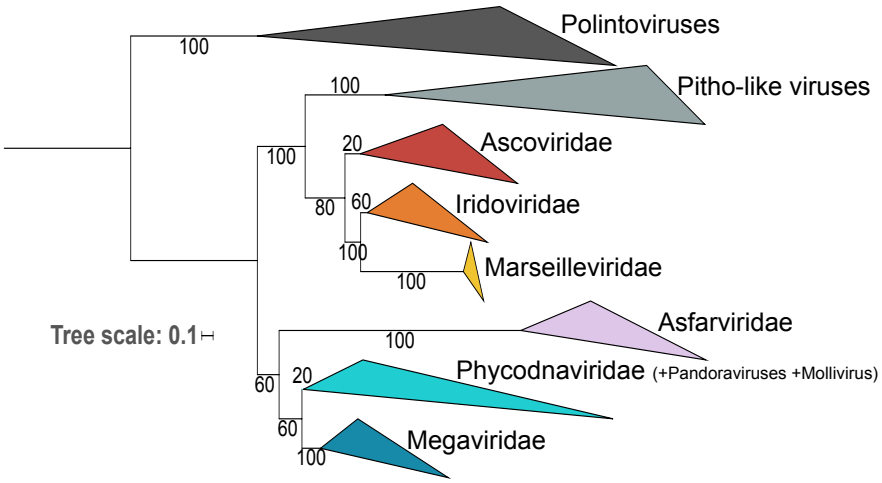


Fig. S5. Relationships between Polintoviruses and NCLDVs. Maximum likelihood (ML) phylogenetic tree of the concatenated structural proteins from Polintoviruses and NCLDVs after removal of *Poxviridae* and *Aureococcus anophagefferens* virus. The scale-bar indicates the average number of substitutions per site. The values at branches represent support calculated by nonparametric bootstrap.

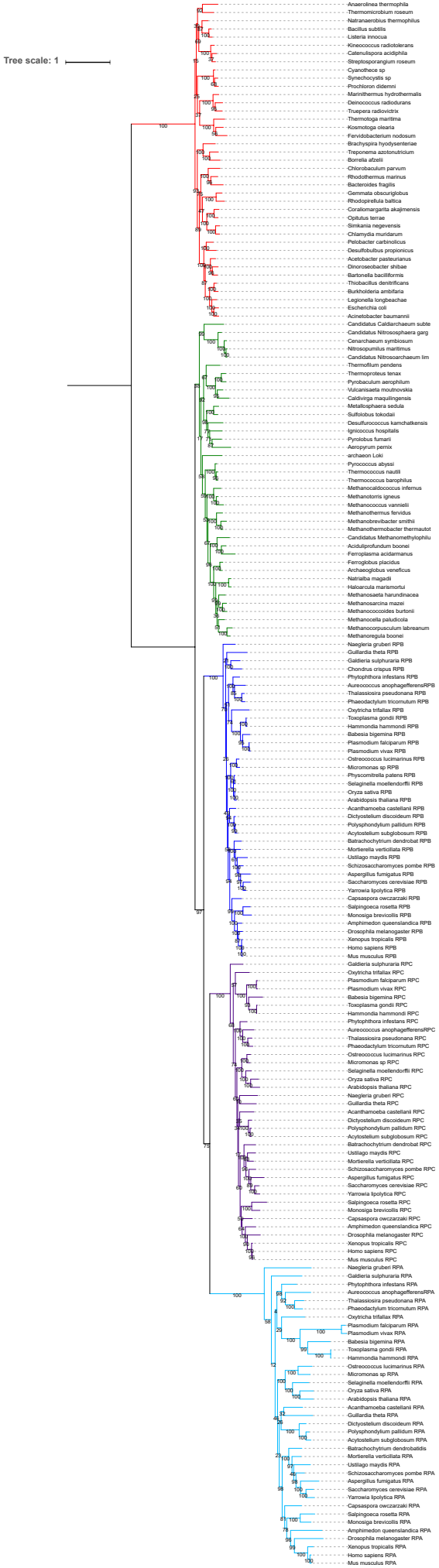


Fig. S6. Maximum-likelihood (ML) phylogenetic tree of the concatenated two largest RNA polymerase subunits from Bacteria, Archaea, and Eukaryotes, including the 3 eukaryotic polymerases. Bacteria have been used as the outgroup. Bacteria, Archaea, eukaryotic RNAP-I, -II, and -III are indicated in red, green, light blue, dark blue, and purple, respectively. The scale-bar indicates the average number of substitutions per site. Values on branches represent support calculated by nonparametric bootstrap (100 replicates).

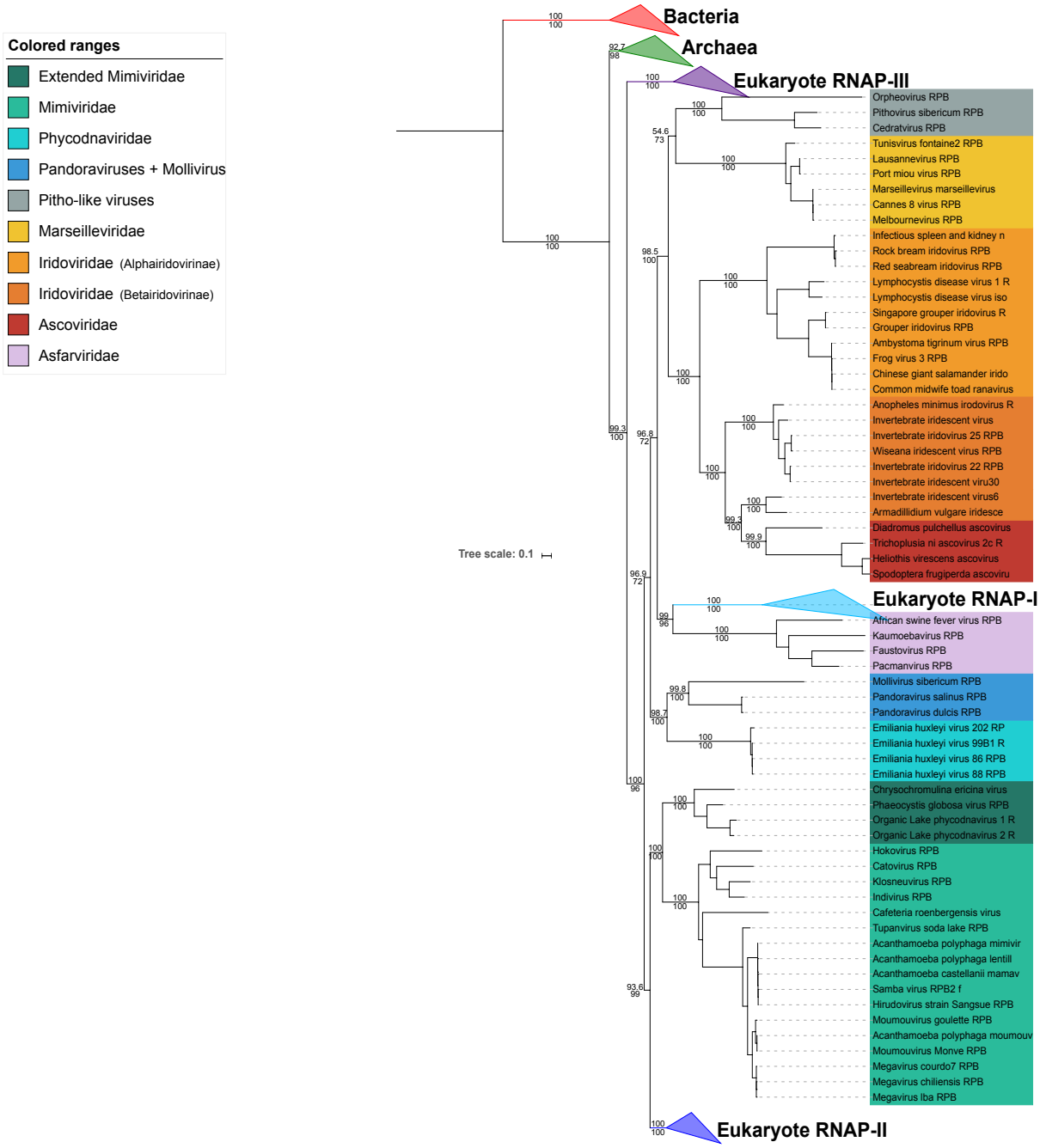


Fig. S7. Maximum likelihood (ML) phylogenetic tree of the concatenated two largest RNA polymerase subunits from Bacteria, Archaea, Eukaryotes, and NCLDVs. ML phylogenetic tree of the concatenation of RNAP-a and RNAP-b, with Bacteria used as the outgroup. The scale-bar indicates the average number of substitutions per site. Values on top and below branches represent supports calculated by SH-like approximate likelihood ratio test (aLRT; 1000 replicates) and ultrafast bootstrap approximation (UFBoot; 1000 replicates), respectively. Only values for major branches (main clades and their relationships) are shown.

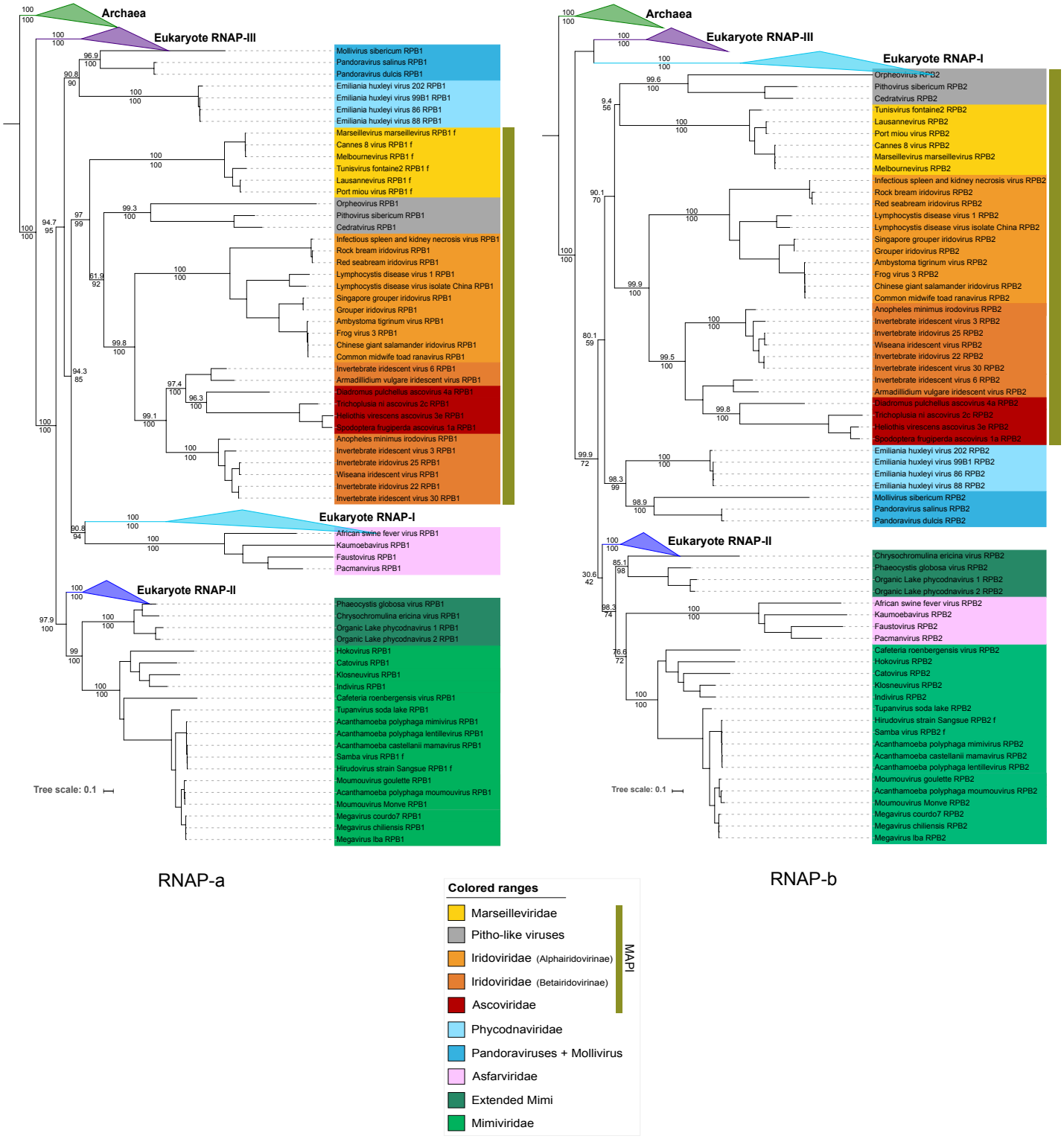


Fig. S8. Maximum likelihood (ML) single-protein trees of the two largest RNA polymerase subunits from Archaea, Eukaryotes, and NCLDVs. ML phylogenetic trees of the RNAP-a (left) and RNAP-b (right), with Archaea used as the outgroup. The scale-bars indicate the average number of substitutions per site. Values on top and below branches represent supports calculated by SH-like approximate likelihood ratio test (aLRT; 1000 replicates) and ultrafast bootstrap approximation (UFBoot; 1000 replicates), respectively. Only values for major branches (main clades and their relationships) are shown.

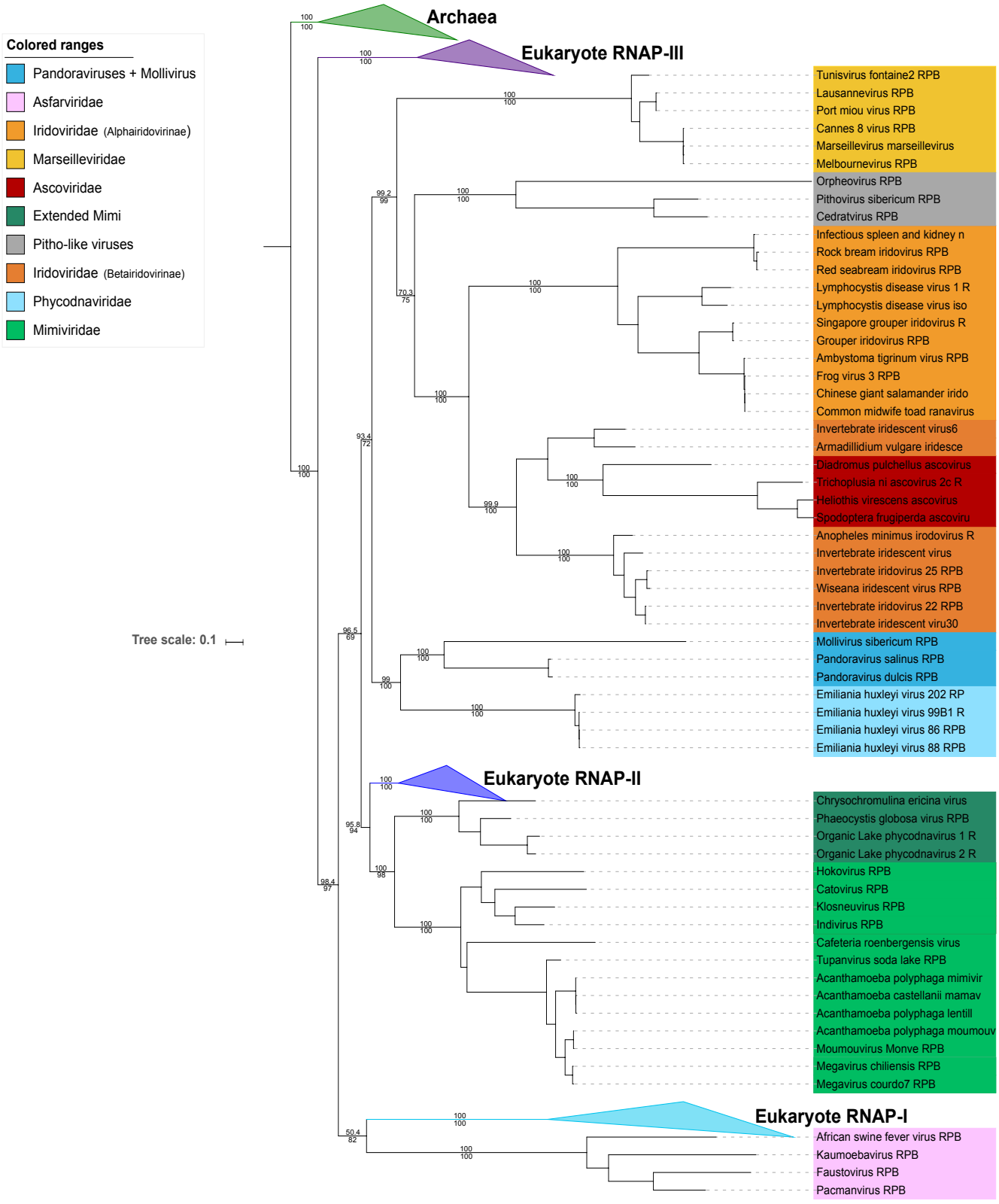


Fig. S9. Maximum likelihood (ML) phylogenetic tree of the concatenated two largest RNA polymerase subunits from Archaea, Eukaryotes, and NCLDVs. ML phylogenetic tree of the concatenation of RNAP-a and RNAP-b, with Archaea used as the outgroup. The scale-bar indicates the average number of substitutions per site. Values on top and below branches represent supports calculated by SH-like approximate likelihood ratio test (aLRT; 1000 replicates) and ultrafast bootstrap approximation (UFBoot; 1000 replicates), respectively. Only values for major branches (main clades and their relationships) are shown.

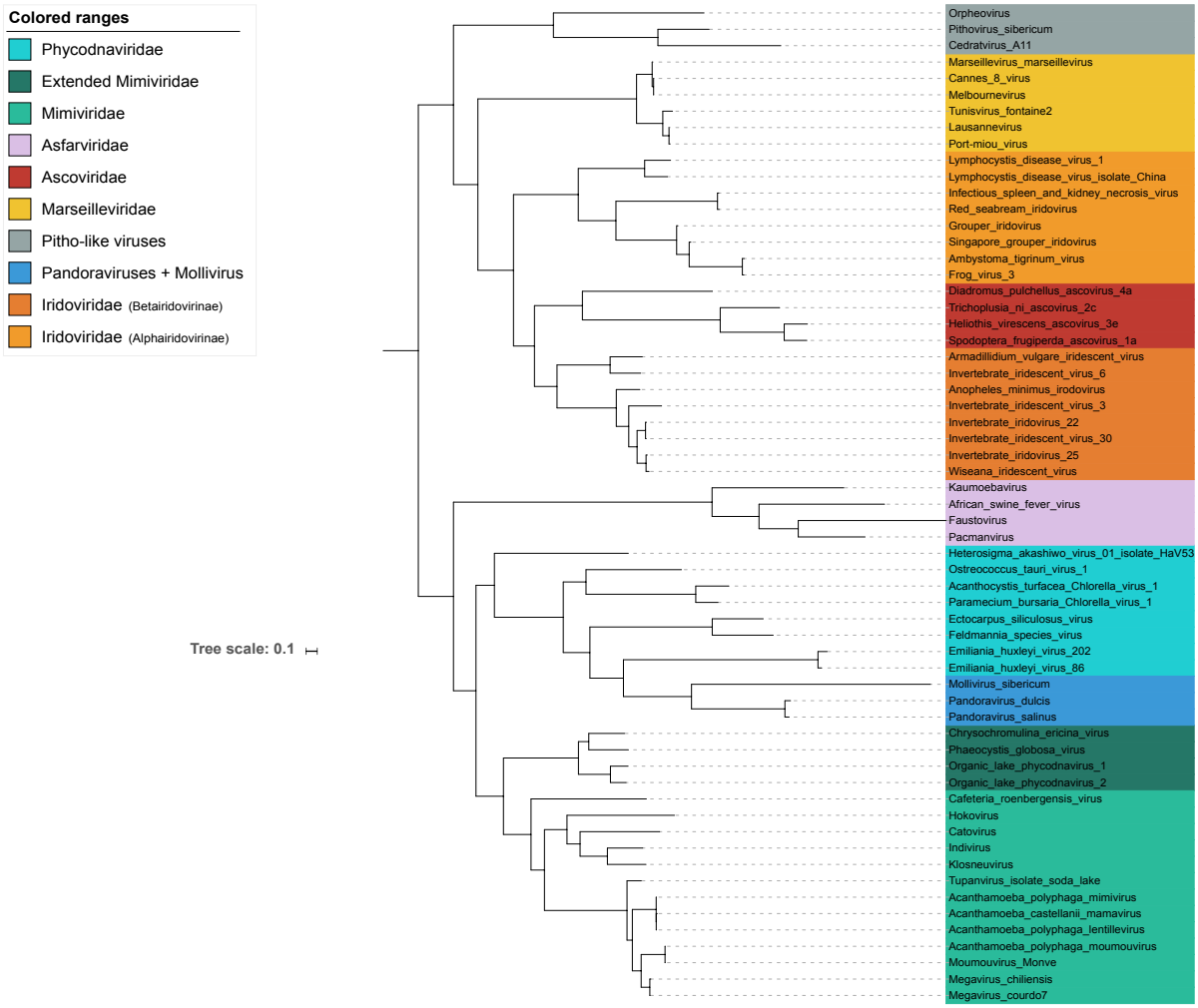


Fig. S10. Maximum likelihood (ML) phylogenetic tree of the concatenation of all core proteins but the two RNAP subunits from NCLDVs. ML tree of the concatenation of the DNA pol B, Primase, MCP, pATPase, TFIIS, and VLTF3-like sequences from NCLDVs obtained during the comparative phylogenetics test (see Methods and Table S3). The scale-bar indicates the average number of substitutions per site.

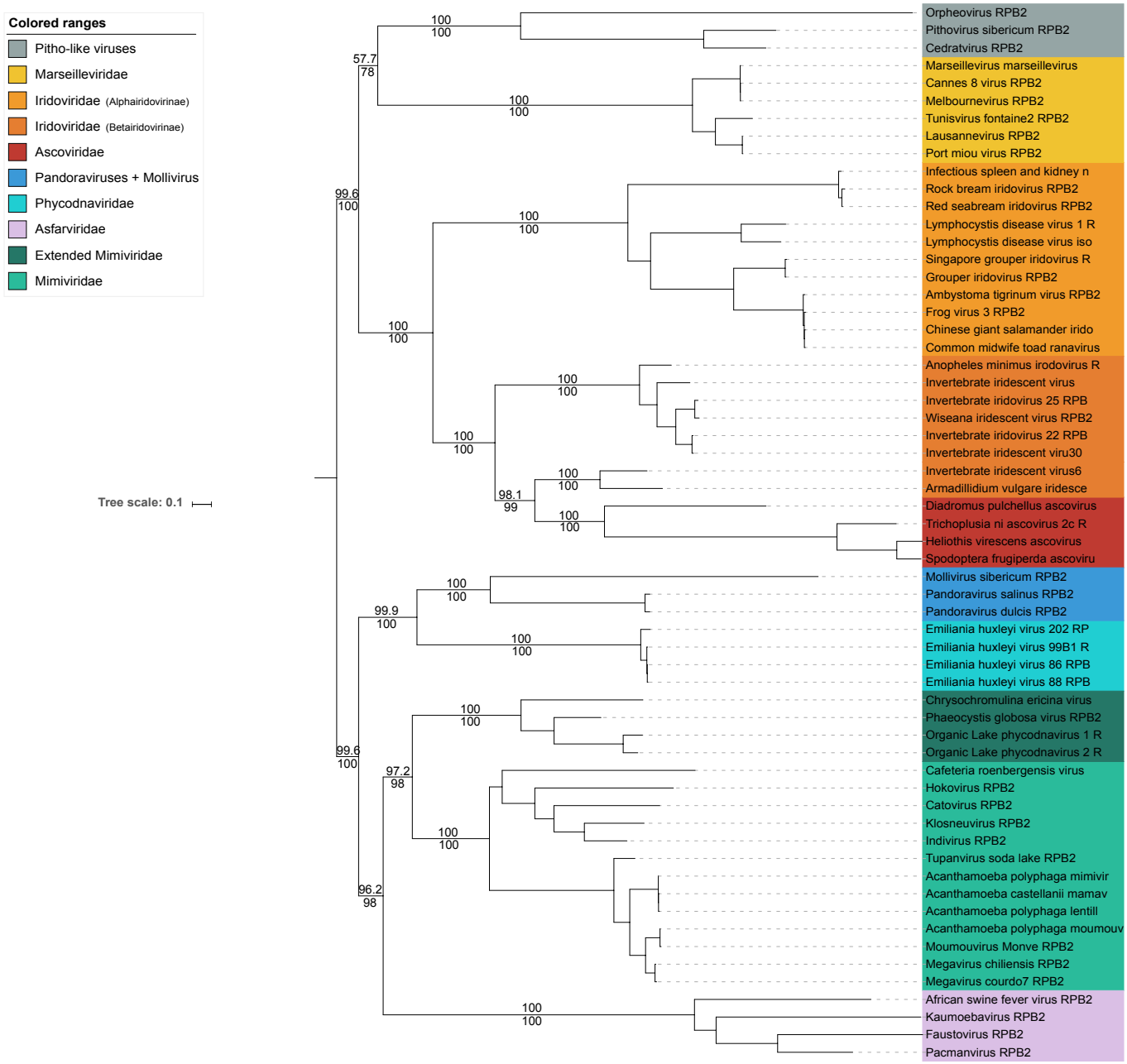


Fig. S11. Maximum likelihood (ML) phylogenetic tree of the concatenated two largest RNA polymerase subunits from the NCLDVs. The scale-bars indicates the average number of substitutions per site. Values on top and below branches represent supports calculated by SH-like approximate likelihood ratio test (aLRT; 1000 replicates) and ultrafast bootstrap approximation (UFBoot; 1000 replicates), respectively. Only values for major branches (main clades and their relationships) are shown.

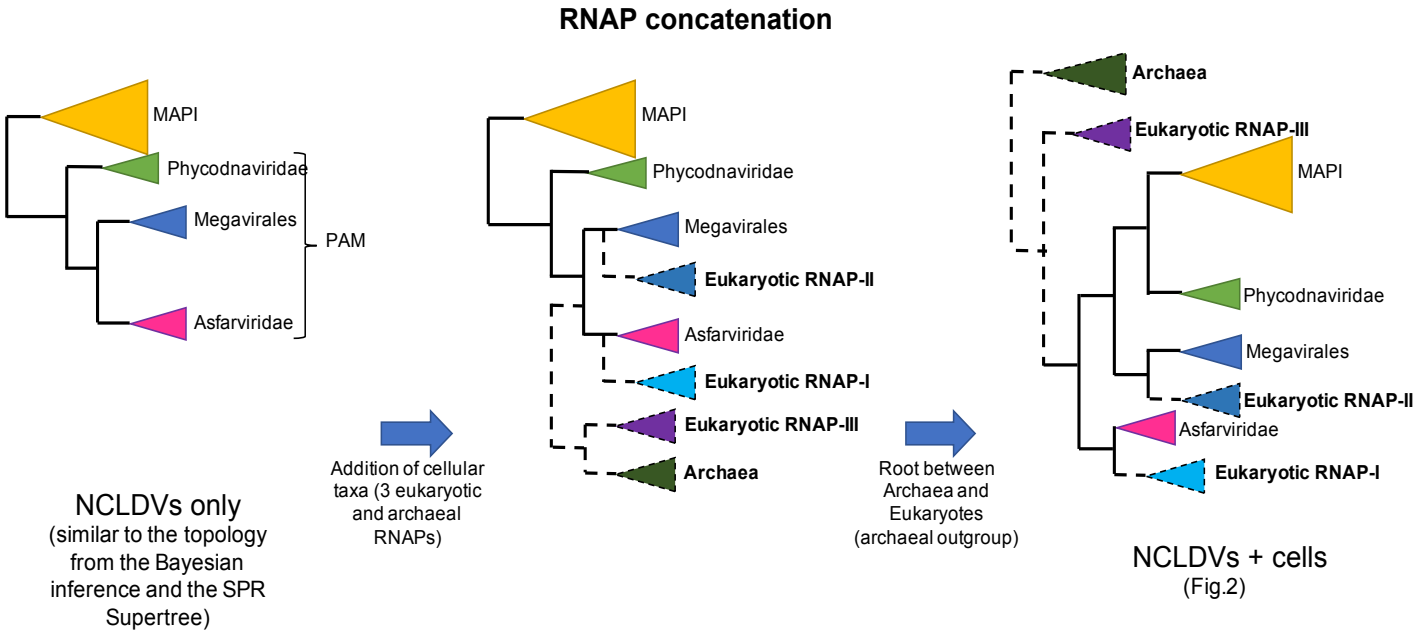


Fig. S12. Schematic representation of the congruence in NCLDV topologies obtained before and after the inclusion of cellular sequences in the concatenated RNAP-subunits tree.
Schematic representation of the evolution of the RNAP concatenation topologies after the inclusion of sequences from eukaryotes and archaea, and then after the rooting between the Archaea and the rest of the tree.

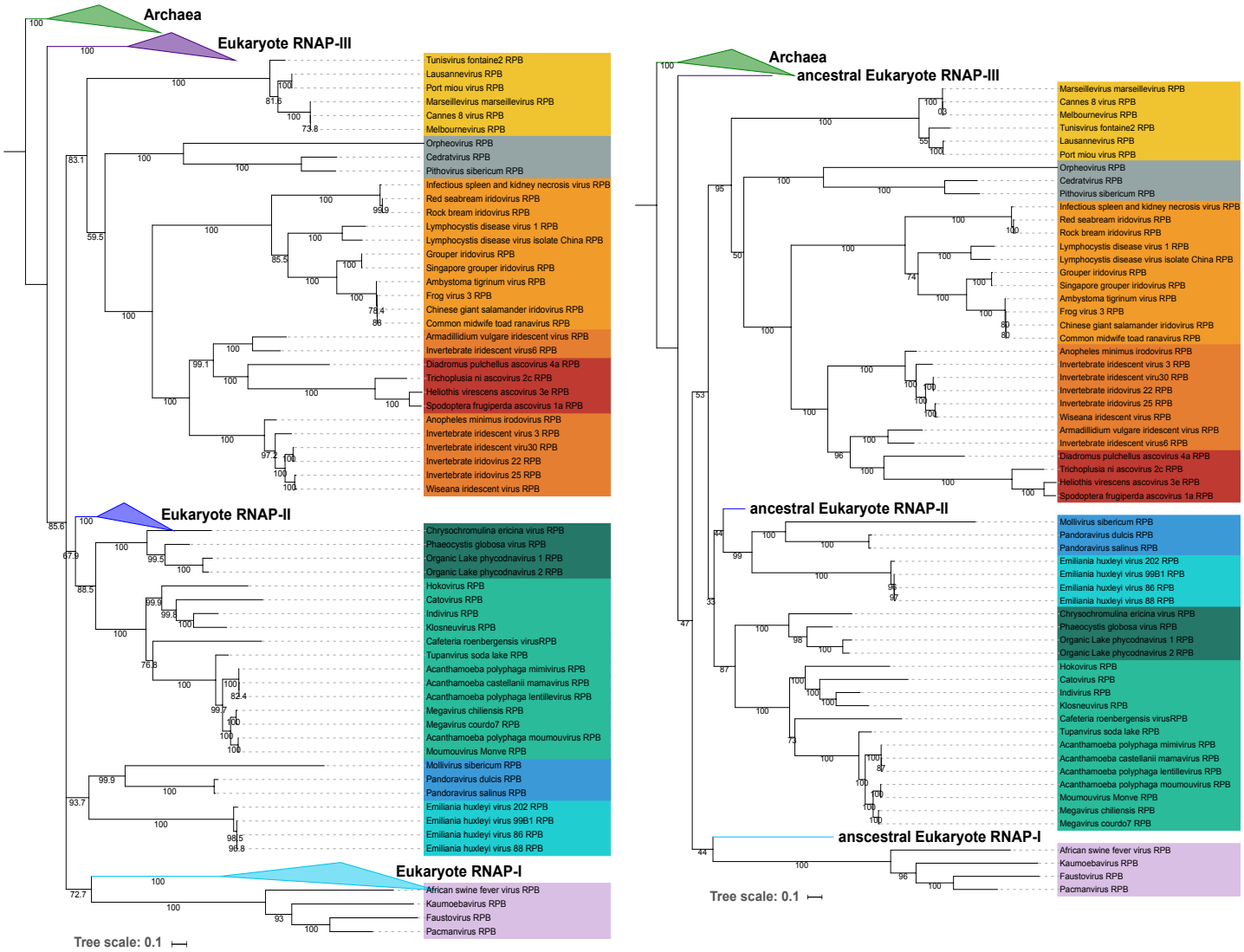


Fig. S13. Phylogenetic trees of the concatenated two largest RNA polymerase subunits from Archaea, Eukaryotes, and NCLDVs, obtained through consensus bootstrap reconstruction (left) and maximum likelihood (ML) with ancestral sequences reconstructed (right). Consensus bootstrap tree (left) obtained from the concatenation of RNAP-a and RNAP-b, with Archaea used as the outgroup. ML phylogenetic tree (right) of the concatenation of RNAP-a and RNAP-b, with Archaea used as the outgroup and the eukaryotic polymerases replaced by their reconstructed ancestral sequences. The scale-bars indicate the average number of substitutions per site. Supports were calculated by nonparametric bootstrap.

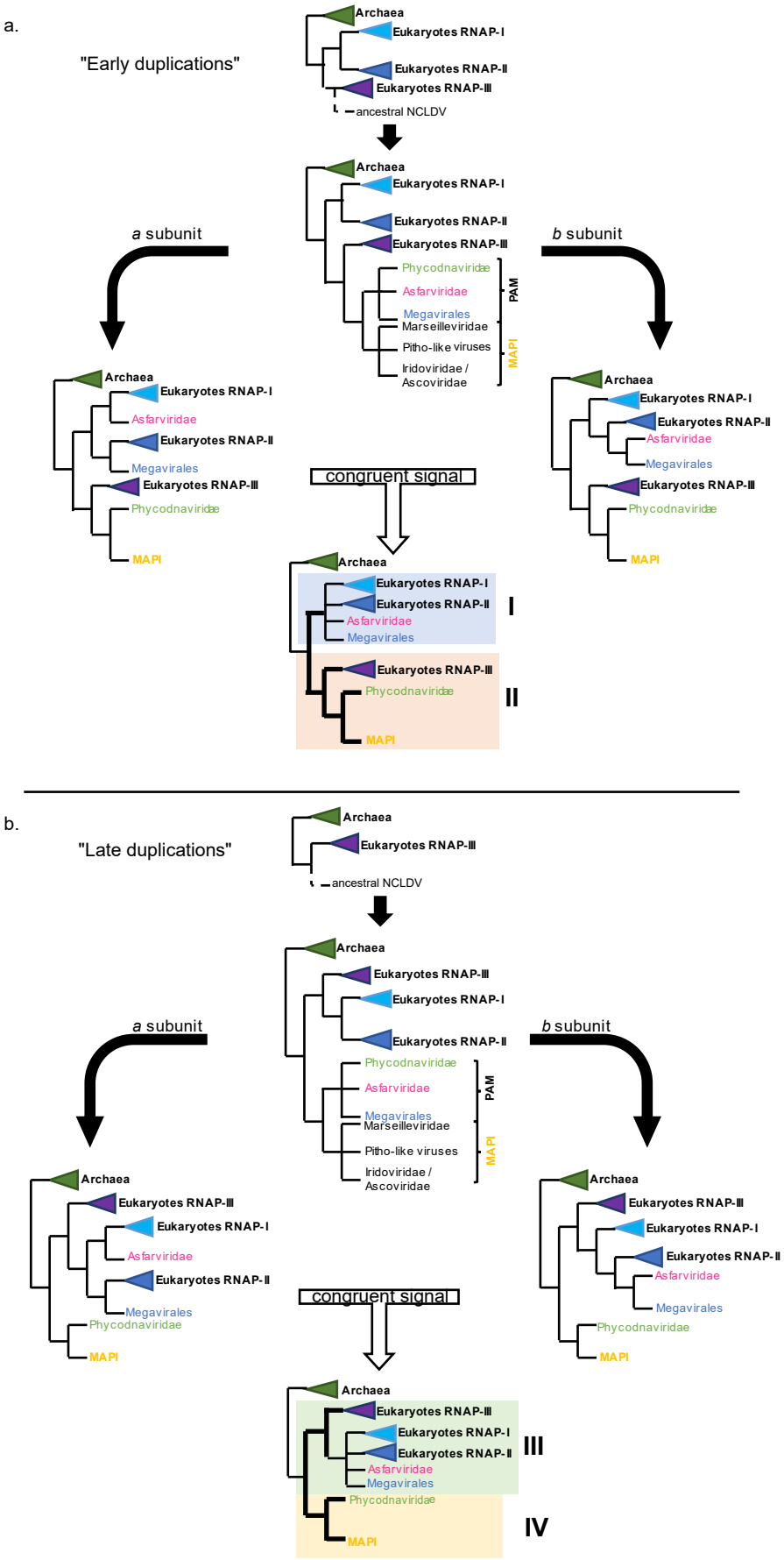


Fig. S14. Schematic representations of two alternative scenarios for the transfers of RNAPs from cells to viruses with the congruent signals expected from the two subunits. The eukaryotic RNAP-I and -II originated from duplication events, either before the transfer of the ancestral eukaryotic RNAP (more alike RNAP-III) to the ancestor of NCLDVs ("Early duplications"), or after the transfer ("Late duplications"). In the first scenario (**a.**), the two subunits should contain a congruent signal for a clade containing the eukaryotic RNAP-I/-II together with the "Megavirales" and the *Asfarviridae* (I), and another containing the Eukaryotic RNAP-III with the MAPI and the *Phycodnaviridae* (II). In the other scenario (**b.**), a congruent signal should be expected for a clade grouping the MAPI superclade with the *Phycodnaviridae* (IV) branching separately from a clade comprising the *Asfarviridae*, the "Megavirales", and the three eukaryotic RNAPs (III). None of these clades are observed in the phylogenetic trees.

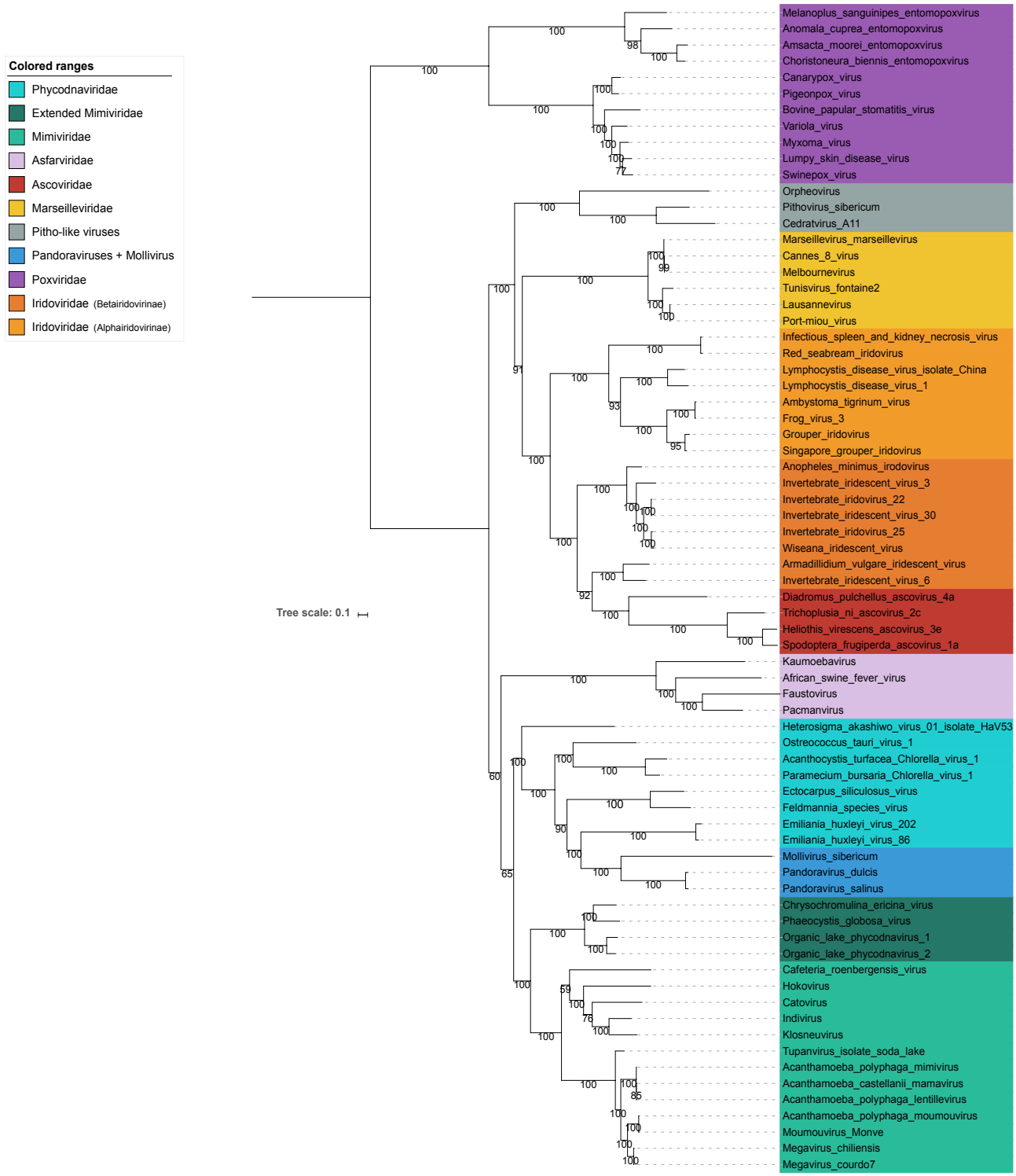


Fig. S15. Maximum likelihood (ML) phylogenetic tree of the concatenated 8 core genes from the NCLDVs, including *Poxviridae*. ML phylogenetic tree of the concatenation of the DNA pol B, Primase, RNAP-a, RNAP-b, MCP, pATPase, TFIIS, and VLTF3-like sequences from NCLDVs, with Poxviridae used as the outgroup. The scale-bar indicates the average number of substitutions per site. Values on branches represent support calculated by nonparametric bootstrap.

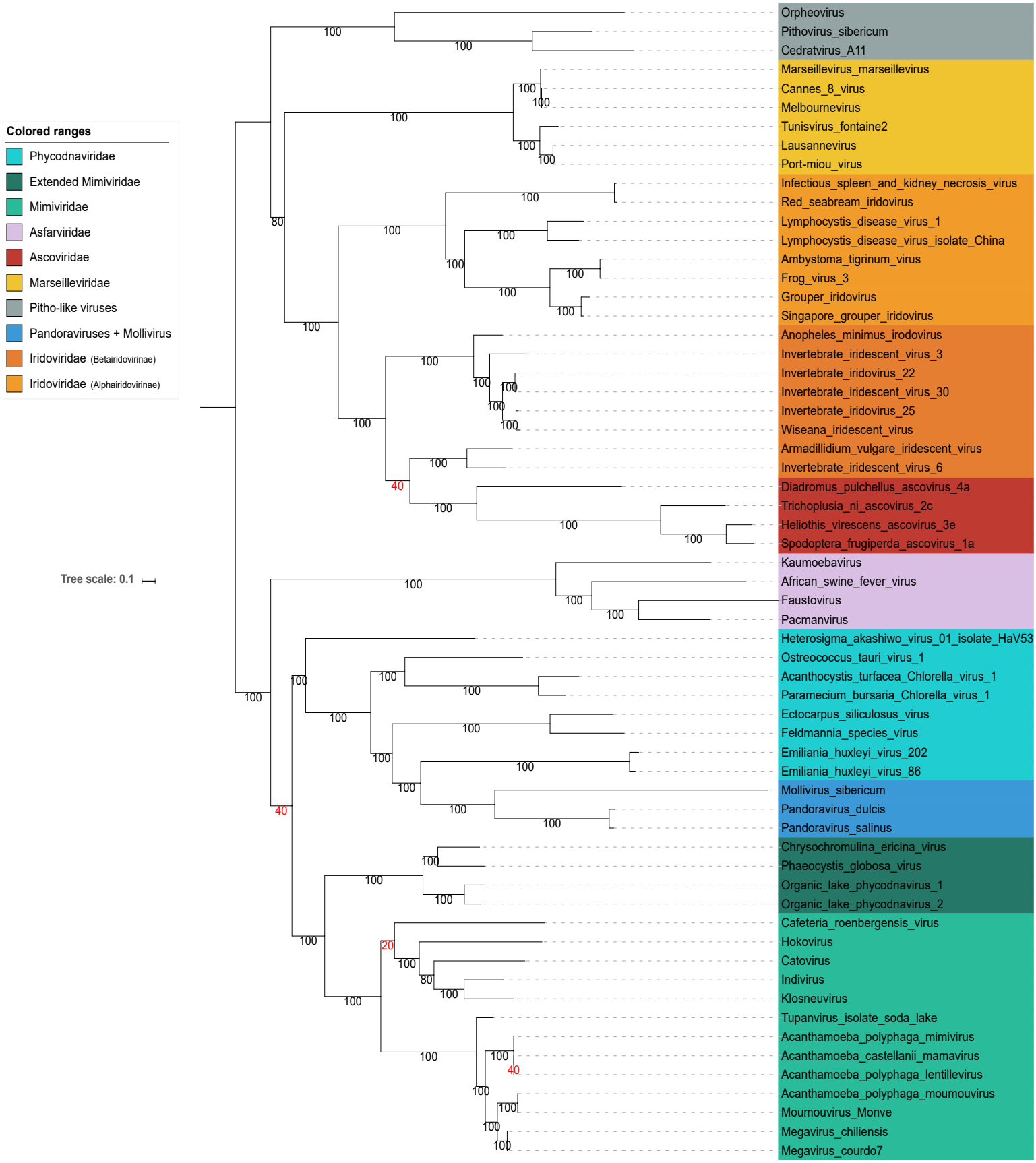


Fig. S16. Maximum likelihood (ML) phylogenetic tree of the concatenated 8 core genes from the NCLDVs. ML phylogenetic tree of the concatenation of the DNA pol B, Primase, RNAP-a, RNAP-b, MCP, pATPase, TFIIS, and VLTF3-like sequences from NCLDVs after removal of *Poxviridae* and *Aureococcus anophagefferens* virus. The scale-bar indicates the average number of substitutions per site. Values on branches represent support calculated by nonparametric bootstrap; supports inferior to 70% are shown in red.

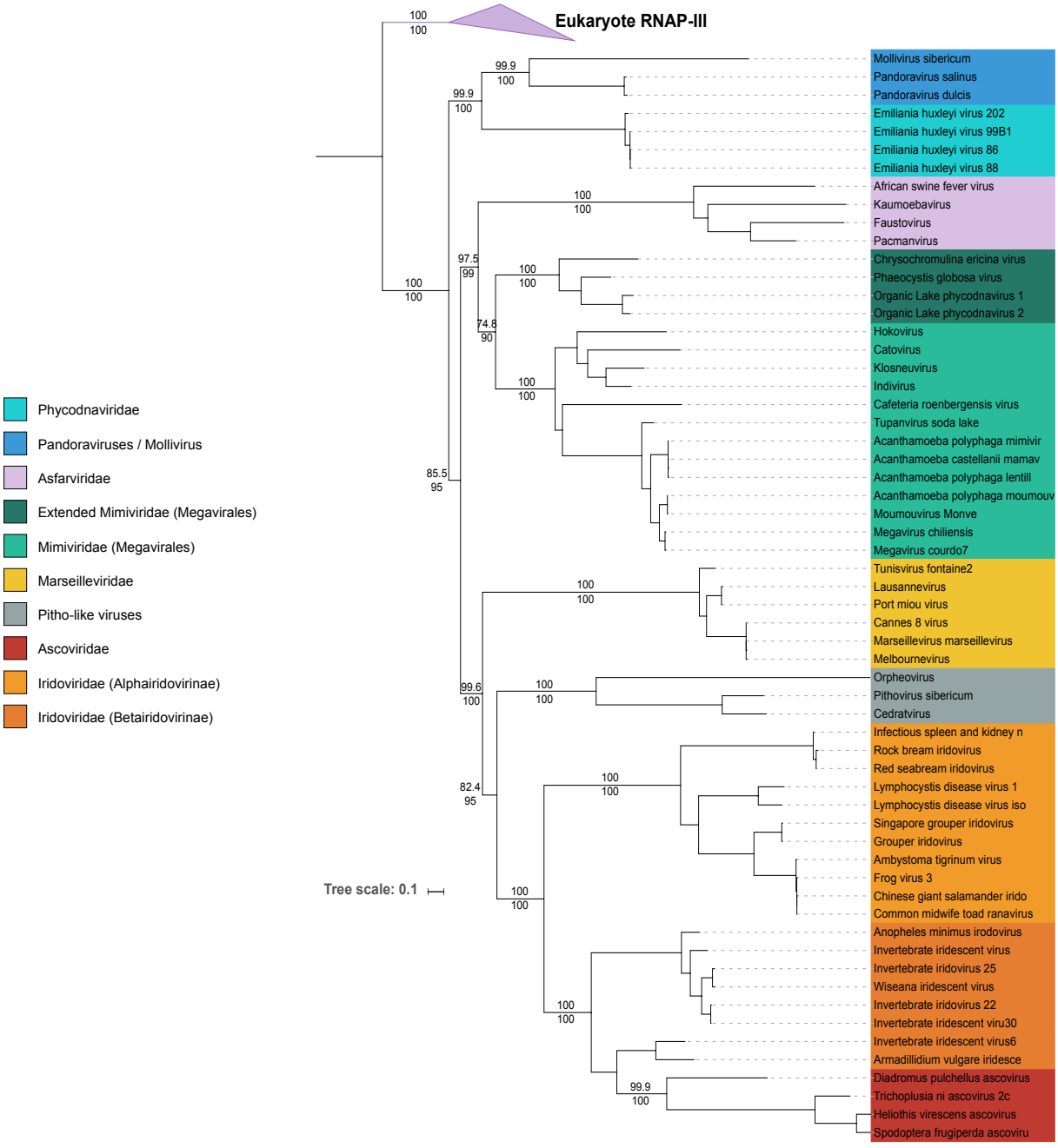


Fig. S17. Maximum likelihood (ML) phylogenetic tree of the concatenated two largest RNA polymerase subunits from the NCLDVs and the eukaryotic RNAP-III. The scale-bar indicates the average number of substitutions per site. Values on top and below branches represent supports calculated by SH-like approximate likelihood ratio test (aLRT; 1000 replicates) and ultrafast bootstrap approximation (UFBoot; 1000 replicates), respectively. Only values for major branches (main clades and their relationships) are shown. The eukaryotic RNAP-III sequences have been used as outgroup.

Table S1. List and access numbers of NCLDV genomes included in this study.

Name	Genome ID	Core genome	Phylogenetic analyses
Aureococcus anophagefferens virus	NC_024697.1	√	x
Diadromus pulchellus ascovirus 4a	NC_011335.1	√	√
Heliothis virescens ascovirus 3e	NC_009233.1	√	√
Spodoptera frugiperda ascovirus 1a	NC_008361.1	√	√
Trichoplusia ni ascovirus 2c	NC_008518.1	√	√
African swine fever virus	NC_001659.2	√	√
Faustovirus	KJ614390.1	√	√
Ambystoma tigrinum virus	NC_005832.1	√	√
Andrias davidianus ranavirus	KC865735.1	√	x
Anopheles minimus irodovirus	NC_023848.1	√	√
Armadillidium vulgare iridescent virus	NC_024451.1	√	√
Chinese giant salamander iridovirus	KF512820.1	√	x
Common midwife toad ranavirus	KP056312.1	√	x
Epizootic haematopoietic necrosis virus	NC_028461.1	√	x
European catfish virus	NC_017940.1	√	x
Frog virus 3	NC_005946.1	√	√
German gecko ranavirus	KP266742.1	√	x
Grouper iridovirus	AY666015.1	√	√
Invertebrate iridovirus 22	NC_021901.1	√	√
Invertebrate iridovirus 25	NC_023613.1	√	√
Infectious spleen and kidney necrosis virus	NC_003494.1	√	√
Invertebrate iridescent virus 3	NC_008187.1	√	√
Invertebrate iridescent virus 6	NC_003038.1	√	√
Invertebrate iridescent virus 30	NC_023611.1	√	√
Lymphocystis disease virus 1	NC_001824.1	√	√
Orange-spotted grouper iridovirus	AY894343.1	√	x
Rana grylio iridovirus	JQ654586.1	√	x
Red seabream iridovirus	AB104413.1	√	√
Rock bream iridovirus	KC244182.1	√	x
Singapore grouper iridovirus	NC_006549.1	√	√
Soft-shelled turtle iridovirus	EU627010.1	√	x
Testudo hermanni ranavirus	KP266741.1	√	x
Tiger frog virus	AF389451.1	√	x
Tortoise ranavirus	KP266743.1	√	x
Turbot reddish body iridovirus	GQ273492.1	√	x
Wiseana iridescent virus	NC_015780.1	√	√
Cannes 8 virus	KF261120.1	√	√
Lausannevirus	NC_015326.1	√	√
Marseillevirus marseillevirus	NC_013756.1	√	√
Melbournevirus	NC_025412.1	√	√
Port-miou virus	NC_028047.1	√	√
Tunisvirus fontaine2	KF483846.1	√	√
Acanthamoeba castellanii mamavirus	JF801956.1	√	√
Acanthamoeba polyphaga lentillevirus	AFYC01000001.1 AFYC01000002.1 AFYC01000003.1 AFYC01000004.1 AFYC01000005.1 AFYC01000006.1 AFYC01000007.1 AFYC01000008.1 AFYC01000009.1 AFYC01000010.1	√	√
Acanthamoeba polyphaga mimivirus	NC_014649.1	√	√
Acanthamoeba polyphaga moumouvirus	NC_020104.1	√	√
Cafeteria roenbergensis virus	NC_014637.1	√	√
Hirudovirus strain Sangsue	KF493731.1	√	x
Megavirus chiliensis	NC_016072.1	√	√
Megavirus courdo7	JN885990.1	√	√
Megavirus courdo7	JN885991.1 JN885991.1 JN885992.1 JN885993.1	√	√
Megavirus lba	NC_020232.1	√	x
Moumouvirus goulette	KC008572.1	√	x
Moumouvirus Monve	JN885994.1 JN885995.1 JN885996.1 JN885997.1 JN885998.1 JN885999.1 JN886000.1 JN886001.1	√	√

Name	Genome ID	Core genome	Phylogenetic analyses
Samba virus	KF959826.2	√	x
Pandoravirus dulcis	NC_021858.1	√	√
Pandoravirus salinus	NC_022098.1	√	√
Acanthocystis turfacea Chlorella virus 1	NC_008724.1	√	√
Chrysochromulina ericina virus	NC_028094.1	√	√
Ectocarpus siliculosus virus	NC_002687.1	√	√
Emiliania huxleyi virus 86	NC_007346.1	√	√
Feldmannia species virus	NC_011183.1	√	√
Organic lake phycodnavirus 1	HQ704802.1	√	√
Organic lake phycodnavirus 2	HQ704803.1	√	√
Ostreococcus tauri virus 1	NC_013288.1	√	√
Paramecium bursaria Chlorella virus 1	NC_000852.5	√	√
Phaeocystis globosa virus	NC_021312.1	√	√
Amsacta moorei entomopoxvirus	AF250284.1	√	x
Anomala cuprea entomopoxvirus	NC_023426.1	√	x
Bovine papular stomatitis virus	NC_005337.1	√	x
Canarypox virus	NC_005309.1	√	x
Choristoneura biennis entomopoxvirus	NC_021248.1	√	x
Cowpox virus	NC_003663.2	√	x
Lumpy skin disease virus	NC_003027.1	√	x
Melanoplus sanguinipes entomopoxvirus	NC_001993.1	√	x
Myxoma virus	NC_001132.2	√	x
Penguinpox virus	NC_024446.1	√	x
Pigeonpox virus	NC_024447.1	√	x
Swinepox virus	NC_003389.1	√	x
Vaccinia virus	NC_006998.1	√	x
Variola virus	NC_001611.1	√	x
Cedratvirus A11	NC_032108.1	x	√
Mollivirus sibericum	NC_027867.1	√	√
Pithovirus sibericum	NC_023423.1	√	√
Heterosigma akashiwo virus 01 isolate HaV53	KX008963.1	x	√
Kaumoebavirus	NC_034249.1	x	√
Pacmanvirus	NC_034383.1	x	√
Klosneuvirus	KY684123.1 KY684122.1 KY684121.1 KY684120.1 KY684119.1 KY684118.1 KY684117.1 KY684116.1 KY684115.1 KY684114.1 KY684113.1 KY684112.1 KY684111.1 KY684110.1 KY684109.1 KY684108.1	x	√
Indivirus	KY684102.1 KY684101.1 KY684100.1 KY684099.1 KY684098.1 KY684097.1 KY684096.1 KY684095.1 KY684094.1 KY684093.1 KY684092.1 KY684091.1 KY684090.1 KY684089.1 KY684088.1 KY684087.1 KY684086.1 KY684085.1	x	√
Catovirus	KY684083.1 KY684084.1	x	√
Hokovirus	KY684103.1 KY684104.1 KY684105.1 KY684106.1 KY684107.1	x	√
Tupanvirus isolate soda lake	KY523104.1	x	√
Orpheovirus	LT906555.1	x	√

Table S2. Results of the comparative phylogenetic analyses (congruence test), based on the presence/absence of reference features in the ML phylogenetic trees of every possible concatenations of 6 out of 8 markers (systematically referred by the two missing genes).

			Reference (Fig 1)	PrimaseB / PolB	PolB / MCP	PrimaseA / MCP	PolB / RNAP-a	PrimaseA / RNAP-a	MCP / RNAP-a	PolB / TF2S	PrimaseA / TF2S	TF2S / MCP	TF2S / RNAP-a	PolB / pATPase	PrimaseA / ATPase	MCP / pATPase	pATPase / RNAP-a	TF2S / pATPase	PolB / RNAP-b	PrimaseA / RNAP-b	MCP / RNAP-b	RNAP-a / RNAP-b	TF2S / RNAP-b	pATPase / RNAP-b	PolB / VLTF3	PrimaseA / VLTF3	VLTF3 / MCP	VLTF3 / RNAP-a	VLTF3 / TF2S	VLTF3 / pATPase	VLTF3 / RNAP-b	Features	Notes
MAPI			**1**	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	28	
		Pitho-like	**1**	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	28	
	MIA		**1**	1	0	1	1	1	0	1	1	0	1	1	1	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	21	
		Marseilleviridae	**1**	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	28	
		Alphaindovirinae	**1**	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	28	
	Ascoviridae + Iridoviridae		**1**	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	28	
	Ascoviridae + Betairidovirinae		**1**	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	28	Ascoviridae are almost systematically branching within Betairidovirinae
		Betairidovirinae	**0**	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	7	
		Ascoviridae	**1**	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	28	
PAM			**1**	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	28	
		Phyco + Pando + Molli	**1**	0	1	0	1	0	1	1	0	1	1	1	1	0	1	1	1	0	1	1	1	1	0	1	1	1	1	1	1	21	
		Pando + Molli	**1**	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	28	
	Megavirales + Asfarviridae		**1**	0	1	1	0	1	0	0	1	0	0	0	0	0	1	0	0	0	1	0	1	1	0	1	1	0	1	1	1	14	
		Asfarviridae	**1**	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	28	
	Megavirales		**1**	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	28	
		Mimiviridae	**1**	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	28	
		Extended Mimiviridae	**1**	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	28	
Score			**16**	15	15	14	16	14	15	15	14	15	15	16	14	15	16	15	16	15	15	16	16	17	15	15	15	15	16	16	16		

Table S3. List and taxon IDs of the cellular taxa used in this study.

	Phylum	Species	Taxon ID
Bacteria			
PVC Planctomycetes		Gemmata obscuriglobus UQM 2246	214688
		Rhodopirellula baltica strain SH 1	243090
Bacterioidetes		Rhodothermus marinus DSM 4252	518766
		Bacteriodes fragilis	862962
		Chlorobaculum parvum NCIB 8327	517417
Gammaproteobacteria		Escherichia coli str. K-12 substr. MG1655 (W3110)	511145
		Legionella longbeachae NSW150	661367
		Acinetobacter baumannii 1656-2	696749
Firmicutes		Bacillus subtilis subsp. Subtilis str. 168	224308
		Natranaerobius thermophilus JW/NM-WN-LF	457570
		Listeria innocua Clip11262	272626
Cyanobacteria		Synechocystis sp. PCC 6714	1147
		Prochloron didemni	1216
		Cyanothece sp. PCC 7424	65393
Deinococcus-thermus		Deinococcus radiodurans R1	243230
		Truepera radiovictrix DSM 17093	649638
		Marinithermus hydrothermalis DSM 14884	869210
Thermotogae		Kosmotoga olearia TBF 19.5.1	521045
		Fervidobacterium nodosum Rt17-B1	381764
		Thermotoga maritima MSB8	243274
Chloroflexi		Anaerolinea thermophila UNI-1	926569
		Thermomicrobium roseum DSM 5159	309801
Actinobacteria		Catenulispora acidiphila DSM 44928	479433
		Streptosporangium roseum DSM 43021	479432
		Kineococcus radiotolerans SRS30216	266940
Spirochaetes		Brachyspira hyodysenteriae WA1	565034
		Treponema azotonutricium ZAS-9	545695
		Borrelia afzelii Pko	390236
PVC Verrucomicrobia		Coraliomargarita akajimensis DSM 45221	583355
		Opitutus terrae PB90-1	452637
PVC Chlamydiae		Simkania negevensis Z	331113
		Chlamydia muridarum Nigg	1434773
Deltaproteobacteria		Pelobacter carbinolicus DSM 2380	338963
		Desulfobulbus propionicus DSM 2032	577650
Alphaproteobacteria		Acetobacter pasteurianus IFO 3283-01-42C	634458
		Dinoroseobacter shibae DFL 12	398580
		Bartonella bacilliformis KC583	360095
Betaproteobacteria		Thiobacillus denitrificans ATCC 25259	292415
		Burkholderia ambifaria AMMD	339670
Archaea			
Crenarchaeota	Desulfurococcales	Pyrolobus fumarii 1A	694429
		Aeropyrum pernix K1	272557
		Desulfurococcus kamchatkensis 1221n	490899
		Ignicoccus hospitalis KIN4_I	453591
	Sulfolobales	Metallosphaera sedula DSM 5348	399549
		Sulfolobus tokodaii str.7	273063
	Thermoproteales	Thermoproteus tenax Kra 1	768679
		Thermofilum pendens Hrk 5	368408
		Vulcanisaeta moutnovskia 768-28	985053
		Caldivirga maquilingensis IC-167	397948
		Pyrobaculum aerophilum str. IM2	178306
Thaumarchaeota		Nitrosopumilus maritimus SCM1	436308
		Cenarchaeum symbiosum A	414004
		Candidatus Nitrosoarchaeum limnia SFB1	886738
		Candidatus Nitrososphaera gargensis Ga9.2	1237085
	Aigarchaeota	Candidatus Caldiarchaeum subterraneum ASM27032	311458
Asgard		Lokiarchaeum sp. GC14_75	1538547
Euryarchaeota Cluster I	Thermococcales	Thermococcus nautili 30-1	195522
		Thermococcus barophilus MP	391623
		Pyrococcus abyssi GE5	272844
	Methanococcales	Methanotorris igneus Kol 5	880724
		Methanococcus vannielii SB	406327
		Methanocaldococcus infernus ME	573063
	Methanobacteriales	Methanothermus fervidus DSM 2088	523846
		Methanobrevibacter smithii ATCC 35061	420247
		Methanothermobacter thermautotrophicus str. Delta H	187420
Euryarchaeota Cluster II	Archaeoglobales	Ferroglobus placidus DSM 10642	589924
		Archaeoglobus veneficus	693661
	Thermoplasmatales	Ferroplasma acidarmanus fer1	333146
	Methanomassiliicoccales	Candidatus Methanomethylophilus alvus Mx1201	1236689
	DHEV2	Aciduliprofundum boonei T469	439481
	Methanosarcinales	Methanosarcina mazei Go1	192952
		Methanococcoides burtonii DSM 6242	259564
		Methanosaeta harundinacea 6Ac	1110509
	Methanomicrobiales	Methanocorpusculum labreanum Z	410358
		Methanoregula boonei 6A8	456442
	Halobacteriales	Natrialba magadii ATCC 43099	547559
		Haloarcula marismortui ATCC 43049	272569
	Methanocellales	Methanocella paludicola SANAE	304371
Eukaryotes			
Opisthokonta	insertae sedis	Capsaspora owczarzaki	595528
	Metazoa	Homo sapiens	9606
		Drosophila melanogaster	7227
		Xenopus (Silurana) tropicalis	8364
		Amphimedon queenslandica	400682
		mus musculus domesticus C57BL/6J	10092
	Choanoflagellida	Salpingoeca rosetta	946362

		Monosiga brevicollis	431895
	Fungi	Aspergillus fumigatus Af293	330879
		Schizosaccharomyces pombe 972h	284812
		Saccharamyces cerevisae	765312
		Batrachochytrium dendrobatidis	684364
		Yarrowia lipolytica	284591
		Ustilago maydis	237631
		Mortierella verticillata	1069443
Amoebozoa	Mycetozoa	Dictyostelium discoideum	352472
		Polysphondylum pallidum	670386
		Acytostelium subglobosum	1410327
	Discosea	Acanthamoeba castellanii	1257118
Discoba	Heterolobosea	Naegleria gruberi	744533
Viridiplantae	Sreptophyta	Physcomitrella patens ‡	3218
		Oryza sativa	39946
		Arabidopsis thaliana	3702
		Selaginella moellendorfii	88036
	Chlorophyta	Ostreococcus lucimarinus	436017
		Micromonas sp.	296587
	Rhodophyta	Galdieria sulphuraria	130081
		Chondrus crispus	2769
	Pyrenomonadales	Guillardia theta	905079
SAR	Stramenopiles	Phytophthora infestans	403677
		Thalassiosira pseudonana	296543
		Phaeodactylum tricornutum	556484
		Aureococcus anophagefferens	44056
	Alveolata	Oxytricha trifallax	1172189
		Toxoplasma gondii	508771
		Plasmodium falciparum	36329
		Plasmodium vivax	126793
		Babesia bigemina	5866
		Hammondia hammondi	99158

‡ only the RNAP-II sequence of *Physcomitrella patens* is included in our analyses, as the RNAP-I and -III sequences resulted in extremely long branches in preliminary phylogenetic analyses.

References

1. Yutin N, Koonin EV (2012) Hidden evolutionary complexity of Nucleo-Cytoplasmic Large DNA viruses of eukaryotes. *Virol J* 9(1):161.

2. Legendre M, et al. (2014) Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses with a pandoravirus morphology. *Proc Natl Acad Sci* 111(11):4274–4279.

3. Philippe N, et al. (2013) Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. *Science* 341(6143):281–286.

4. Andreani J, et al. (2016) Cedratvirus, a double-cork structured giant virus, is a distant relative of Pithoviruses. *Viruses* 8(11):300.

5. Andreani J, et al. (2018) Orpheovirus IHUMI-LCC2: a new virus among the giant viruses. *Front Microbiol* 8. doi:10.3389/fmicb.2017.02643.

6. Hughes AL, Friedman R (2005) Poxvirus genome evolution by gene gain and loss. *Mol Phylogenet Evol* 35(1):186–195.

7. Bratke KA, McLysaght A (2008) Identification of multiple independent horizontal gene transfers into poxviruses using a comparative genomics approach. *BMC Evol Biol* 8(1):67.

8. Hughes AL, Irausquin S, Friedman R (2010) The evolutionary biology of poxviruses. *Infect Genet Evol* 10(1):50–59.

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

10. Yutin N, Colson P, Raoult D, Koonin EV (2013) Mimiviridae: clusters of orthologous genes, reconstruction of gene repertoire evolution and proposed expansion of the giant virus family. *Virol J* 10(1):106.

11. Gallot-Lavallée L, Blanc G, Claverie J-M (2017) Comparative genomics of Chrysochromulina Ericina virus and other microalga-infecting large DNA viruses highlights their intricate evolutionary relationship with the established Mimiviridae family. *J Virol* 91(14). doi:10.1128/JVI.00230-17.

12. Arslan D, Legendre M, Seltzer V, Abergel C, Claverie J-M (2011) Distant Mimivirus relative with a larger genome highlights the fundamental features of Megaviridae. *Proc Natl Acad Sci* 108(42):17486–17491.

13. Colson P, de Lamballerie X, Fournous G, Raoult D (2012) Reclassification of giant viruses composing a fourth domain of Life in the new order Megavirales. *Intervirology* 55(5):321–332.

14. Santini S, et al. (2013) Genome of Phaeocystis globosa virus PgV-16T highlights the common ancestry of the largest known DNA viruses infecting eukaryotes. *Proc Natl Acad Sci* 110(26):10800–10805.

15. Boyer M, Madoui M-A, Gimenez G, La Scola B, Raoult D (2010) Phylogenetic and phyletic studies of informational genes in genomes highlight existence of a 4 domain of life including giant viruses. *PloS One* 5(12):e15530.

16. Williams TA, Embley TM, Heinz E (2011) Informational gene phylogenies do not support a fourth domain of Life for nucleocytoplasmic large DNA viruses. *PLoS ONE* 6(6):e21080.

17. Moreira D, López-García P (2015) Evolution of viruses and cells: do we need a fourth domain of life to explain the origin of eukaryotes? *Philos Trans R Soc B Biol Sci* 370(1678):20140327.

18. Sharma V, Colson P, Giorgi R, Pontarotti P, Raoult D (2014) DNA-Dependent RNA polymerase detects hidden giant viruses in published databanks. *Genome Biol Evol* 6(7):1603–1610.

19. Sharma V, et al. (2015) Welcome to pandoraviruses at the "Fourth TRUC" club. *Front Microbiol* 6. doi:10.3389/fmicb.2015.00423.

20. Sharma V, Colson P, Chabrol O, Pontarotti P, Raoult D (2015) Pithovirus sibericum, a new bona fide member of the “Fourth TRUC” club. *Front Microbiol* 6. doi:10.3389/fmicb.2015.00722.

21. Moreira D, López-García P (2009) Ten reasons to exclude viruses from the tree of life. *Nat Rev Microbiol* 7(4):306–311.

22. Tan G, et al. (2015) Current methods for automated filtering of multiple sequence alignments frequently worsen single-Gene phylogenetic inference. *Syst Biol* 64(5):778–791.

23. Lane WJ, Darst SA (2010) Molecular evolution of multisubunit RNA polymerases: sequence analysis. *J Mol Biol* 395(4):671–685.

24. Blombach F, et al. (2009) Identification of an ortholog of the eukaryotic RNA polymerase III subunit RPC34 in Crenarchaeota and Thaumarchaeota suggests specialization of RNA polymerases for coding and non-coding RNAs in Archaea. *Biol Direct* 4(1):39.

25. Blombach F, et al. (2015) Archaeal TFEα/β is a hybrid of TFIIE and the RNA polymerase III subcomplex hRPC62/39. *eLife* 4:e08378.

26. Eme L, Spang A, Lombard J, Stairs CW, Ettema TJG (2017) Archaea and the origin of eukaryotes. *Nat Rev Microbiol* 15(12):711–723.

27. Claverie J-M, Abergel C (2013) Open questions about giant viruses. *Advances in Virus Research* (Elsevier), pp 25–56.

28. Claverie J-M, Abergel C (2016) Giant viruses: The difficult breaking of multiple epistemological barriers. *Stud Hist Philos Sci Part C Stud Hist Philos Biol Biomed Sci* 59:89–99.

29. Filée J (2013) Route of NCLDV evolution: the genomic accordion. *Curr Opin Virol* 3(5):595–599.

30. Filée J (2018) Giant viruses and their mobile genetic elements: the molecular symbiosis hypothesis. *Curr Opin Virol* 33:81–88.

31. Forterre P (2006) The origin of viruses and their possible roles in major evolutionary transitions. *Virus Res* 117(1):5–16.

32. Forterre P (2012) Virocell concept, The. *ELS*, ed John Wiley & Sons, Ltd (John Wiley & Sons, Ltd, Chichester, UK). doi:10.1002/9780470015902.a0023264.

33. Legendre M, et al. (2018) Diversity and evolution of the emerging Pandoraviridae family. *Nat Commun* 9(1). doi:10.1038/s41467-018-04698-4.