

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

Julien Guglielmini<sup>1</sup>, Anthony Woo<sup>2</sup>, Mart Krupovic<sup>2</sup>, Patrick Forterre<sup>2\*</sup>, Morgan Gaia<sup>2\*</sup>

<sup>1</sup>HUB Bioinformatique et Biostatistique, C3BI, USR 3756 IP CNRS, Institut Pasteur, Paris, France

<sup>2</sup>Unité de Biologie Moléculaire du Gène chez les Extrêmophiles (BMGE), Département de Microbiologie, Institut Pasteur, Paris, France

\*corresponding authors:

Patrick Forterre: [patrick.forterre@pasteur.fr](mailto:patrick.forterre@pasteur.fr)

Morgan Gaia: [morgan.gaia@pasteur.fr](mailto:morgan.gaia@pasteur.fr)

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

3 Julien Guglielmini<sup>1</sup>, Anthony Woo<sup>2</sup>, Mart Krupovic<sup>2</sup>, Patrick Forterre<sup>2</sup>, Morgan Gaia<sup>2</sup>

4

5 <sup>1</sup>HUB Bioinformatique et Biostatistique, C3BI, USR 3756 IP CNRS, Institut Pasteur, Paris, France

6 <sup>2</sup>Unité de Biologie Moléculaire du Gène chez les Extrêmophiles (BMGE), Département de

7 Microbiologie, Institut Pasteur, Paris, France

8

9 **Abstract**

10 Giant and large eukaryotic double-stranded DNA viruses from the Nucleo-Cytoplasmic  
11 Large DNA Virus (NCLDV) assemblage represent a remarkably diverse and potentially  
12 ancient component of the eukaryotic virome. However, their origin(s), evolution and  
13 potential roles in the emergence of modern eukaryotes remain a subject of intense  
14 debate. Since the characterization of the mimivirus in 2003, many big and giant viruses  
15 have been discovered at a steady pace, offering a vast material for evolutionary  
16 investigations. In parallel, phylogenetic tools are constantly being improved, offering  
17 more rigorous approaches for reconstruction of deep evolutionary history of viruses  
18 and their hosts. Here we present robust phylogenetic trees of NCLDVs, based on the 8  
19 most conserved proteins responsible for virion morphogenesis and informational  
20 processes. Our results uncover the evolutionary relationships between different NCLDV  
21 families and support the existence of two superclades of NCLDVs, each encompassing  
22 several families. We present evidence strongly suggesting that the NCLDV core genes,  
23 which are involved in both informational processes and virion formation, were acquired  
24 vertically from a common ancestor. Among them, the largest subunits of the DNA-  
25 dependent RNA polymerase were seemingly transferred from two clades of NCLDVs to

26 proto-eukaryotes, giving rise to two of the three eukaryotic DNA-dependent RNA  
27 polymerases. Our results strongly suggest that these transfers and the diversification of  
28 NCLDV's predated the emergence of modern eukaryotes, emphasizing the major role of  
29 viruses in the evolution of cellular domains.

30

31

32 The discovery of giant viruses in the early 21<sup>st</sup> century has revived the debate on  
33 the nature of viruses and their role in evolution<sup>1-13</sup>. The 1 $\mu$ m-long particles of  
34 pithoviruses<sup>14</sup> can be seen under a light microscope and the 2.5Mb-long genomes of  
35 pandoraviruses, larger than those of many cellular organisms, encode for more than  
36 2,000 proteins, mostly ORFans<sup>15</sup>. However, these unexpected features notwithstanding,  
37 giant viruses are a *bona fide* part of the virosphere, relying on the infected cells for the  
38 production of energy and protein synthesis. Phylogenetic and comparative genomics  
39 analyses showed that giant viruses together with smaller eukaryotic dsDNA viruses  
40 form a supergroup, dubbed the Nucleo-Cytoplasmic Large DNA Viruses (NCLDV)<sup>16,17</sup>.  
41 This assemblage encompasses families of large and giant viruses, including *Poxviridae*,  
42 *Iridoviridae*, *Ascoviridae*, *Asfarviridae*, *Marseilleviridae*, *Mimiviridae*, and *Phycodnaviridae*  
43 as well as several lineages of as yet unclassified viruses, such as pithoviruses,  
44 pandoraviruses, molliviruses and faustoviruses<sup>18</sup>. Altogether, the NCLDV's are associated  
45 with diverse eukaryotic phyla, from phagotrophic protists to insects and mammals, and  
46 some cause devastating diseases, such as smallpox (*Poxviridae*) or swine fever  
47 (*Asfarviridae*), or play important ecological roles, such as termination of algal blooms  
48 (*Phycodnaviridae*<sup>19</sup>).

49

50           The origin and evolution of the NCLDV remain a subject of controversy. It is still  
51 unclear if these viruses form a monophyletic group, if proteins conserved in most  
52 NCLDVs had a congruent evolutionary history or if some of them were acquired several  
53 times independently from their hosts. Most phylogenetic analyses performed up to now  
54 were based on individual proteins or various subsets of conserved proteins<sup>20,21</sup>. These  
55 analyses usually recovered the monophyly of various NCLDV families, but often offered  
56 contradicting results and the relationships between the families remained debated. For  
57 instance, it has been proposed that the giant pandoraviruses are related to members of  
58 the *Phycodnaviridae*<sup>22</sup>, but this grouping was not recovered in a recent phylogeny based  
59 on their DNA polymerases<sup>23</sup>. According to some studies, the different families of the  
60 NCLDVs emerged during the diversification of modern eukaryotes<sup>24</sup>, whereas in other  
61 studies, NCLDVs form a monophyletic group branching between Archaea and  
62 Eukarya<sup>29</sup>/10/2018 13:51:00. Some authors have even suggested that several families  
63 of giant viruses could have originated independently from extinct cellular lineages,  
64 possibly even before the last universal common ancestor (LUCA) of Archaea, Bacteria,  
65 and Eukarya<sup>11,25</sup>.

66

67           With phylogenetic tools being constantly improved and new genomes of large  
68 and giant viruses steadily unearthed, we decided to perform an updated and in-depth  
69 phylogenetic analysis of the NCLDVs. We mined available genomes for homologous  
70 genes, built clusters of orthologous genes, and performed extensive phylogenetic  
71 analyses on the 8 most conserved ones, separately and in concatenations. In addition,  
72 we have investigated the relationships between NCLDVs and eukaryotes through the  
73 phylogeny of the DNA-dependent RNA polymerases (RNAP). Unlike in previous  
74 analyses, we included in our study the three eukaryotic RNAP (RNAP I, II, and III) and

75 concatenated their two largest subunits. The robust phylogenies we obtained show that  
76 core genes involved in virion morphogenesis as well as genome transcription and  
77 replication have co-evolved in the entire NCLDV lineage. Furthermore, our results  
78 revealed the existence of two superclades of NCLDVs that diverged after the separation  
79 of the archaeal and eukaryotic lineages, but before the emergence of the Last Eukaryotic  
80 Common Ancestor (LECA). Surprisingly, our data suggest that eukaryotic RNAP-III is the  
81 actual cellular ortholog of the archaeal and bacterial RNAP, while eukaryotic RNAP-II  
82 and possibly RNAP-I were transferred between two viral families and proto-eukaryotes.  
83 Overall, our results reveal that the diversification of NCLDVs predates the origin of  
84 modern eukaryotes: the ancestors of contemporary NCLDVs co-evolved with proto-  
85 eukaryotes and could have played an important role in the emergence and  
86 diversification of modern eukaryotes.

87

## 88 **Results**

### 89 **Identification of the core genes**

90 Many new NCLDV genomes have been published following the latest  
91 comprehensive comparative genomics analyses<sup>21,26</sup>, substantially increasing their  
92 known diversity and enriching families that were previously poorly represented. As a  
93 result, the list of the most conserved genes among the NCLDVs could have drastically  
94 changed since the last estimation, prompting us to re-analyse it. To identify NCLDV  
95 orthologs, we designed a pipeline based on Best Bidirectional BLAST Hit combined with  
96 manual curation in order to remain as exhaustive as possible while avoiding inclusion of  
97 paralogs (see details in Methods section). The sets of conserved proteins classified  
98 according to their conservation among NCLDVs are summarized in Supplementary Table  
99 1.

100 Our results show that only 3 proteins are strictly conserved among the 73  
101 selected NCLDV genomes: family B DNA polymerase (DNApol B), the D5-like primase-  
102 helicase (primase hereinafter) and homologs of the Poxvirus Late Transcription Factor  
103 VLTF3 (VLTF3-like) (list of genomes in Supplementary Table 2; selection criteria in  
104 Methods). Acknowledging various reasons which may preclude detection of homologous  
105 genes (e.g., due to high divergence or genuine loss in a taxon), we decided to lower our  
106 conservation threshold to include genes found in at least 95% of the genomes. This  
107 resulted in the increase of our set of core genes by three: the transcription elongation  
108 Factor II-S (TFIIS), the genome packaging ATPase (pATPase), and the major capsid  
109 protein (MCP). Notably, no homolog of the MCP has been found in pandoraviruses<sup>15</sup>,  
110 whereas pATPases are apparently lacking in Pithovirus<sup>14</sup>, Cedratvirus<sup>27</sup>, and  
111 Orpheovirus<sup>28</sup>. Conservation of the NCLDV genes is further discussed in the  
112 Supplementary Information.

113

114 To this set of six proteins (3 strictly conserved and 3 conserved in 95% of the  
115 genomes), we added the two largest RNAP subunits (RNAP-a and -b) despite their  
116 notable absence in all genera of the *Phycodnaviridae* family, except for the  
117 *Coccolithovirus* genus. Indeed, these two proteins are otherwise highly conserved among  
118 the NCLDVs (present in 92% of the genomes) and are the largest universal markers  
119 (found in all members of the three cellular domains), which makes them perfectly suited  
120 for reconstructing the evolutionary relationships between NCLDVs and cellular  
121 organisms. Thus, the set of 8 proteins contains 6 proteins related to informational  
122 processes – genomes expression and replication (DNApol B, primase, VLTF3-like, TFIIS,  
123 RNAP-a, and RNAP-b) – and 2 proteins involved in virion structure and morphogenesis  
124 (pATPase and MCP).

125

## 126 **The core markers share a similar phylogenetic signal**

127       Using a maximum-likelihood (ML) framework, the monophyly of all known  
128 NCLDV families, except the *Phycodnaviridae*, was obtained with high support in most of  
129 the 8 single-protein phylogenetic trees (Supplementary Figure 1). As often observed in  
130 published NCLDV phylogenies<sup>26</sup>, *Ascoviridae* were however nested within the  
131 *Iridoviridae* in most trees. The grouping of the *Mimiviridae* with related unclassified  
132 viruses with smaller genomes often referred to as the “extended Mimiviridae”<sup>21</sup> or more  
133 recently the “Mesomimivirinae”<sup>29</sup>, was obtained in five out of the 8 trees. We will refer  
134 to this grouping as the “Megavirales” putative order (see Supplementary Information).

135

136       The *Poxviridae* clade consistently formed a long branch and displayed the most  
137 unstable position, branching next to various families (see Supplementary Information).  
138 The same was true for *Aureococcus anophagefferens* virus. Thus, to avoid potential  
139 artefacts, we decided to remove these taxa from most of our subsequent analyses.  
140 Phylogenetic analyses of the resultant dataset resulted in globally congruent trees of  
141 individual core proteins (Supplementary Figure 2). Notably, the *Marseilleviridae*, the  
142 *Ascoviridae*, the *Iridoviridae*, and a clade grouping *Pithovirus sibericum* with *Cedratvirus*  
143 A11 and *Orpheovirus IHUM-LCC2* (thereafter referred as the Pitho-like viruses), group  
144 seemingly together, while the *Phycodnaviridae* (including *Pandoraviruses* and  
145 *Mollivirus*), *Asfarviridae*, and the “Megavirales” also form a cluster.

146

147       In order to verify if the NCLDV informational proteins have indeed co-evolved  
148 with proteins involved in virion formation, we first concatenated independently the 4  
149 largest informational proteins (i.e. the DNA and RNA polymerases, and the primase) and

150 next the 2 proteins involved in the formation of virions (the MCP and the pATPase). In  
151 both trees (Supplementary Figure 3 and 4), all NCLDV families were monophyletic,  
152 except for the *Iridoviridae* which again were split by the *Ascoviridae* in the tree  
153 constructed from the concatenation of informational proteins (Supplementary Figure 3).  
154 The two phylogenies had similar topologies, with the same clusters of NCLDV families as  
155 observed in single-protein trees. Some positions within these clusters might be affected  
156 by differences between the two datasets: 2 of the 4 informational proteins are absent in  
157 all but one *Phycodnaviridae* genera, while the Pitho-like viruses lack the pATPase gene.  
158 The congruence between the two trees still suggests that informational proteins of the  
159 NCLDVs have mostly co-evolved with proteins involved in the formation of virions. The  
160 8 core genes hence likely underwent through a similar evolutionary history.

161 To further confirm that the 8 core proteins have a similar evolutionary history  
162 and to detect potential incongruences within the selected proteins that could prevent  
163 their global concatenation, we performed a home-made congruence test based on  
164 comparative phylogenetic analyses of differential concatenations (see details in  
165 Methods; Supplementary Table 3). The topologies of the resulting trees were congruent,  
166 with most features systematically present, such as the two clusters of NCLDV families,  
167 the presence of groups regularly observed in the ML trees, and the monophyly of  
168 families. This test thus did not reveal any major incongruences between the different  
169 combinations of core proteins and consequently strongly supports the absence of  
170 conflicting signal embedded in a sequence or in a subset of proteins, confirming that the  
171 core proteins were likely presents in a common ancestor of NCLDVs and all evolved  
172 vertically along their co-evolution with their hosts.

173

174 **The evolution of NCLDVs**



175           We concatenated the 8 core proteins together to improve the robustness and  
176 resolution of the NCLDV phylogeny. We obtained a ML tree (Supplementary Figure 5) in  
177 which the NCLDV families are again clustering into two superclades: the *Marseilleviridae*  
178 with the *Ascoviridae*, the Pitho-like viruses' clade, and the *Iridoviridae* (thereinafter  
179 referred as the MAPI superclade), and the *Phycodnaviridae* with the *Asfarviridae* and the  
180 "Megavirales" (thereinafter referred as the PAM superclade). All positions in this tree  
181 are strongly supported except for the position of the *Asfarviridae* (see Supplementary  
182 Information). We further performed Bayesian inferences with the CAT-GTR model,  
183 designed to deal with sites and sequences heterogeneity, considering that this could  
184 allow a more trustful and accurate reconstruction provided that a satisfactory  
185 convergence could be obtained (see Methods). After reaching a good convergence  
186 (maxdiff <0.1), we obtained a phylogenetic tree with all nodes at maximum support  
187 (Posterior Probabilities = 1), except for two nodes corresponding to minor internal  
188 positions within the *Mimiviridae* family. The Bayesian tree was almost identical to the  
189 ML tree, except that *Phycodnaviridae* are now sister group to a clade clustering  
190 *Asfarviridae* and "Megavirales" (Fig 1). This topology was also confirmed using a  
191 supertree approach (Supplementary Figure 6; details in Methods and Supplementary  
192 Information).

193

194           This tree confidently positions recently identified viruses. The *Mimiviridae* hence  
195 include Klosneuvirus, Indivirus, Catovirus, Hokovirus<sup>30</sup>, and Tupanvirus<sup>31</sup>, and are  
196 associated with related viruses within the putative "Megavirales" order. The still  
197 unclassified Pitho-like viruses, which herein consists of Pithovirus sibericum,  
198 Cedratvirus A11, and Orpheovirus IHUM-LCC2, seem to represent a new separate family  
199 whose position within the putative MAPI superclade remains to be investigated to

200 further extent considering their still low representation. Faustovirus<sup>32,33</sup>, Pacmanvirus<sup>34</sup>,  
201 and Kaumoebavirus<sup>35</sup>, form a well-supported clade with the African swine fever virus  
202 (ASFV-1) of the *Asfarviridae*, as previously suggested<sup>36</sup>. The *Phycodnaviridae* encompass  
203 pandoraviruses and Mollivirus sibericum. The monophyly of this family however  
204 remains a matter of debate as it is not observed in half of the single-protein trees and  
205 has low support in the ML tree based on the concatenated structural proteins. This is  
206 possibly due to the very large diversity of the viruses within this family. Altogether, our  
207 in-depth phylogenetic analyses nonetheless strongly support the existence of the two  
208 major superclades, the MAPI and the PAM.

209

210 The evolution and origin of NCLDV is regularly debated, most notably in term of  
211 their connections to other viruses<sup>18</sup>. Interestingly, homologs of the MCP and pATPase  
212 can be found in viruses from various families belonging to the PRD1-Adenovirus lineage.  
213 This lineage was initially proposed based on the structural conservation of the major  
214 capsid proteins as well as shared principles of virion assembly and genome packaging<sup>37-</sup>  
215 <sup>39</sup>. The closest outgroup to NCLDVs in this lineage could be Polintoviruses<sup>40,41</sup>. When  
216 using Polintoviruses as an outgroup (see Methods), the ML tree of the MCP-pATPase  
217 concatenation is split between the MAPI and PAM putative superclades, suggesting that  
218 these two clusters indeed form monophyletic assemblages (Fig 2). Notably, the MCP-  
219 pATPase tree remains almost identical to the one obtained with the NCLDVs alone (the  
220 only difference being the position of the *Phycodnaviridae*), and the number of positions  
221 was not dramatically reduced (601 positions with Polintoviruses versus 625 positions  
222 without). This indicates that the split between the MAPI and PAM superclades was  
223 probably the earliest event in the evolution of known modern NCLDVs from their  
224 common ancestor.

225

## 226 **The relationship between NCLDV and the three cellular domains**

227         The RNA and DNA polymerases of NCLDV have homologues in the three domains  
228 of life (Archaea, Bacteria and Eukarya), making it *a priori* possible to investigate their  
229 evolutionary relationships with cellular organisms. However, the family B DNA  
230 polymerase, often used to tentatively affiliate new NCLDV genomes to known taxa<sup>42</sup>,  
231 cannot be used for this task since they are absent from most Bacteria and their  
232 phylogenetic analyses produce complex scenarios with the two major subgroups of  
233 archaeal DNA polymerases intermingled with the four types of eukaryotic family B DNA  
234 polymerases ( $\alpha$ ,  $\delta$ ,  $\epsilon$ ,  $\zeta$ )<sup>43</sup>. In contrast, phylogeny of the two largest RNAP subunits,  
235 which are also the largest universal markers, recovered the monophyly of the three  
236 cellular domains<sup>44</sup>. Thus, RNAPs are good candidates to study the relationships between  
237 the cellular domains and NCLDVs.

238

239         Most phylogenetic analyses of RNAPs performed until now included only the  
240 eukaryotic RNA polymerase II (RNAP-II), which is the most studied and usually  
241 considered as the most similar to the archaeal RNAPs<sup>45</sup>. Here, we decided to include all  
242 three eukaryotic RNAPs (RNAP-I, RNAP-II and RNAP-III) (we used a normalized  
243 nomenclature, see Supplementary Information). Importantly, these three multi-subunit  
244 RNAPs are present in all eukaryotes, indicating that they were already all present in the  
245 Last Eukaryotic Common Ancestor (LECA). Their inclusion in our dataset thus should  
246 both reduce the length of the eukaryotic branch and provide three universal eukaryotic  
247 phylogenies, thus three positions for LECA in the cellular/NCLDV RNAP tree.

248

249 We have previously obtained a robust phylogenetic RNAP tree with a  
250 concatenation of the two largest RNAP subunits (in ML and Bayesian frameworks), in  
251 which the three domains are monophyletic, with Eukaryotes and Archaea being sister  
252 groups (the so-called Woese's tree). We obtained this result using a balanced dataset  
253 (same number of species for each of the three domains) and avoiding known fast-  
254 evolving species to prevent long branch attraction artefacts<sup>44,46</sup>29/10/2018 13:51:00.  
255 Since our initial dataset included only RNAP-II as the eukaryotic representative, we  
256 added the eukaryotic RNAP-I and RNAP-III (list of selected taxa in Supplementary Table  
257 4). Interestingly, Archaea and Eukarya again form two monophyletic sister groups in our  
258 new concatenated RNAP subunits tree, despite the drastic reduction of the eukaryotic  
259 branch length (Supplementary Figure 7). Remarkably, RNAP-I was not attracted by  
260 Bacteria despite its very long branch. These observations suggest that the three-domain  
261 topology of the RNAP tree did not result from the attraction of eukaryotes by the long  
262 bacterial branch. Interestingly, the three eukaryotic RNAPs displayed globally congruent  
263 phylogenies, corroborating their presence in LECA.

264  
265 We included the sequences of NCLDV into this new dataset (except for  
266 *Poxviridae* and *Aureococcus anophagefferens* virus) in order to investigate the timeline of  
267 NCLDV diversification in the context of cellular evolution. The ML phylogenetic analysis  
268 of concatenated RNAP subunits yielded the three-domain topology (Supplementary  
269 Figure 8) in which NCLDV branch after the divergence of the archaeal and eukaryotic  
270 lineages. We then removed Bacteria from our subsequent analyses in order to increase  
271 the resolution (single-protein trees in Fig 3 and in Supplementary Figure 9;  
272 concatenation in Supplementary Figure 10). The trees were highly similar after selecting  
273 the Archaea as the outgroup, and supports for several nodes indeed became stronger.

274 Since each of the cellular clades (the Archaea and the three eukaryotic homologs) was  
275 well represented and systematically monophyletic, we decided to use the cellular  
276 sequences as constraints during the alignment process (each of the 4 clades of cellular  
277 sequences corresponding to an independent constraint; see details in Methods),  
278 allowing us to check if this could improve the resolution by limiting mis-alignments  
279 from small insertions or deletions in the viral sequences. The resulting concatenation of  
280 the two subunits switched from 1,683 positions to 1,595, and the highly supported  
281 reconstructed tree obtained in ML framework (LG+C60 model) (Fig 4) was strictly  
282 identical to the one without any constraint. The most significant feature of the  
283 viral/cellular RNAP tree is that LECA, despite being a single timepoint in the history of  
284 eukaryotes, is represented three times among the diversity of NCLDVs, indicating that  
285 NCLDVs predated LECA. This reveals that the diversification of NCLDVs itself predated  
286 that of modern eukaryotes, and consequently, different NCLDV families or superclades  
287 were already infecting proto-eukaryotes.

288

289 Surprisingly, in the tree based on concatenated RNAP subunits, the eukaryotic  
290 RNAP-III appears to be the closest to the archaeal outgroup after addition of viral  
291 sequences with strong supports, suggesting that it could be the actual ortholog of the  
292 archaeal enzyme (Fig 4). A major feature of this tree is that NCLDVs do not form a  
293 monophyletic group, but three monophyletic subgroups well separated from the three  
294 eukaryotic RNAPs, instead of emerging from within eukaryotic diversity. In order to test  
295 this result, we performed an Approximately Unbiased (AU) tree topology test and  
296 compare this tree to two others constraining either the monophyly of NCLDVs or  
297 cellular organisms (see Methods). The AU test rejected these two alternative trees with  
298 p-values  $<1e-3$ . Remarkably, the relative positions of the NCLDV families and

299 superclades in the RNAP tree are completely congruent with the NCLDV topology in the  
300 Bayesian tree previously obtained with the 8 core proteins (Fig 1) and highly similar to  
301 the tree obtained using the concatenation from which the two RNAP subunits were  
302 omitted during the congruence test (Supplementary Table 3; Supplementary Figure 11).  
303 In particular, we recovered the monophyly of the MAPI superclade, and its internal  
304 phylogeny is highly similar to that obtained previously (the positions of *Marseilleviridae*  
305 and Pitho-like viruses are flipped).

306

307         Four clades of the NCLDVs are distinguishable in this viral-cellular RNAP tree,  
308 corresponding to the monophyletic MAPI superclade, the *Phycodnaviridae*, the  
309 “Megavirales” and the *Asfarviridae*. The PAM superclade is indeed not monophyletic in  
310 the RNAP tree because eukaryotic RNAP-I and -II are branching within it. The relative  
311 positions of the three PAM families compared to each other are still matching the NCLDV  
312 tree topology obtained with the 8 core proteins in the Bayesian framework (Fig 1), but  
313 in the viral/cellular RNAP tree, the eukaryotic RNAP-II is sister group to the  
314 “Megavirales” whereas the eukaryotic RNAP-I is sister group to *Asfarviridae*. In order to  
315 assess the robustness of these groupings, and notably of the *Asfarviridae* and RNAP-I  
316 that both display long branches, we reconstructed a consensus bootstrap tree of the  
317 concatenated RNAP subunits. In parallel, we also performed a phylogenetic analysis  
318 based on reconstructed ancestral sequences to replace the three eukaryotic RNAP clades  
319 (see Methods). Both methods supported the relationships between the “Megavirales”  
320 and the eukaryotic RNAP-II as well as between the *Asfarviridae* and the eukaryotic  
321 RNAP-I, suggesting that they reflect a genuine evolutionary signal (Supplementary  
322 Figure 12). Worth-noting, the position of the *Asfarviridae* differs in the two single-  
323 protein subunit trees: they are sister group to the RNAP-I in the individual *a* subunit

324 tree (Fig 3a), as in the tree based on concatenated RNAP subunits (Fig 4), whereas they  
325 branch within the “Megavirales” in the *b* subunit tree (Fig 3b). This suggests that two  
326 transfers might have occurred between proto-eukaryotes and ancestors of the  
327 *Asfarviridae* and could explain the long branch of the *Asfarviridae* in the RNAP trees.

328

329         Considering the branching of NCLDVs after the eukaryotic RNAP-III, it seems that  
330 they have originally obtained their RNAP from proto-eukaryotes after their divergence  
331 from the archaeal lineage. The unexpected positions of RNAP-I and -II within NCLDVs  
332 could suggest that these two eukaryotic RNAPs were either recruited from NCLDVs or  
333 transferred to the ancestors of the *Asfarviridae* family and “Megavirales” order. The  
334 latter hypothesis seems unlikely because replacements of the two largest core genes of  
335 two major NCLDV families by their cellular counterparts would have likely resulted in  
336 substantial alterations in the NCLDV topologies obtained during the congruence test.  
337 This was not the case, and notably, the tree produced without RNAP genes during this  
338 test (Supplementary Figure 12) was highly similar with the 8-core-proteins tree (Fig 1),  
339 and with the trees from the concatenated RNAP genes only, with (Fig 4) or without cells  
340 (Supplementary Figure 13). The only difference is the position of *Phycodnaviridae*,  
341 which are sister group to “Megavirales” in the absence of RNAP genes. This is  
342 remarkable since the RNAP proteins represent nearly half of the total positions in the  
343 global concatenation. These data strongly suggest that the transfers of the RNAP-  
344 encoding genes were directed from viruses to cells, after the diversification of these  
345 RNAPs within NCLDVs. Based on this observation, we postulate a possible scenario  
346 depicted in Fig 5. In this hypothesis, the ancestral eukaryotic RNAP (at least the two  
347 largest subunits), more similar to RNAP-III, was first transferred to the ancestor of  
348 NCLDVs. After the divergence between the MAPI and the PAM superclades, this viral

349 RNAP diverged in the common ancestor of “Megavirales” and *Asfarviridae*, and was  
350 transferred to proto-eukaryotes, later to become the RNAP-II. Separately, a duplication  
351 of the ancestral RNAP-III in proto-eukaryotes occurred, before the largest subunit of this  
352 newly formed RNAP was replaced by that of *Asfarviridae*: this new complex, partly viral  
353 and partly cellular from duplication, resulted in the RNAP-I.

354

## 355 **Discussion**

356 From our investigation of the NCLDV genomes, including those of most recently  
357 identified giant and large dsDNA viruses, we could reconstruct a robust phylogenetic  
358 tree of this group likely to represent their vertical evolutionary history. Our results  
359 provide a solid framework for proposed and sometimes debated positions of different  
360 NCLDV families. Notably, Pithovirus and related viruses form a separate, yet to be  
361 named family most closely related to the *Marseilleviridae*. Pandoraviruses and Mollivirus  
362 branch within the *Phycodnaviridae*, as a sister group to *Coccolithovirus* genus,  
363 confirming the results of Yutin and Koonin<sup>22</sup>. Our results reveal two robust  
364 monophyletic superclades, the MAPI and the PAM, each of which includes several virus  
365 families and a number of unclassified viruses. These results call for reassessment of the  
366 taxonomy of large and giant dsDNA viruses included in the NCLDV assemblage. In  
367 particular, the expansion of the *Mimiviridae* family and discovery of associated but more  
368 distantly related viruses suggests that a family-level taxon might not be adequate to  
369 encompass this diversity. Consequently, the *Mimiviridae* and the related algal viruses as  
370 well as viruses discovered by metagenomics might have to be unified into a new order,  
371 the “Megavirales”. Furthermore, the *Asfarviridae* clade, in addition to ASFV-1, includes  
372 the Faustovirus<sup>32,33</sup>, Kaumoebavirus<sup>35</sup> and Pacmanvirus<sup>34</sup>, which have been suggested to  
373 represent separate families<sup>35</sup>. Thus, an order-level taxon would be needed for



374 classification of these viruses. Similarly, in the MAPI superclade, the placement of the  
375 pandoraviruses and the mollivirus within the *Phycodnaviridae* indicates that this family  
376 might not be monophyletic and should be revised. *Ascoviridae* regularly branch within  
377 *Iridoviridae*, advocating for a reconsideration of these two families. The elusive position  
378 of the *Poxviridae*, which were removed from most of our analyses, and their actual  
379 association to NCLDVs remain to be investigated.

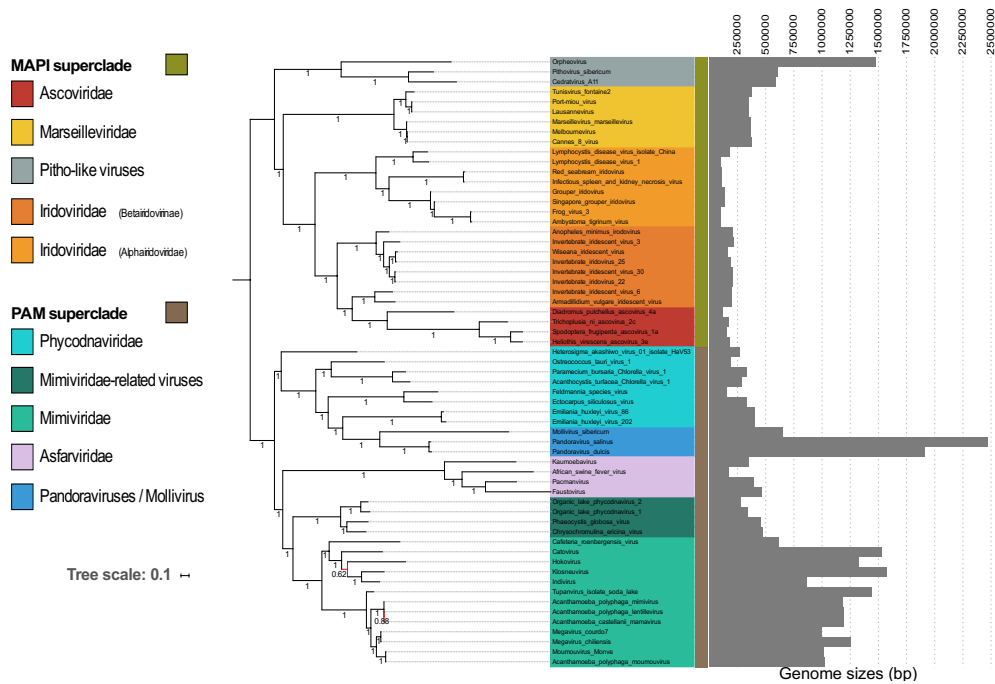
380

381         The monophyly of NCLDVs is not recovered in the cellular/NCLDV RNAP tree:  
382 NCLDVs do not form a fourth domain of life, as proposed by some<sup>20</sup>, nor nest among  
383 eukaryotes<sup>24</sup>. While some genes in the NCLDV genomes might have been recruited from  
384 different sources, notably their modern hosts and bacteria, we have shown that a  
385 congruent vertical evolutionary history of NCLDVs is traceable and sound. The 8  
386 selected core genes selected indeed shared a similar vertical evolution, and were  
387 inherited from a common ancestor, which was likely smaller, as hypothesized before<sup>47</sup>,  
388 and specifically related to polintoviruses<sup>12</sup>. Notably, these core genes are involved in  
389 both genome replication and virion formation, key features of viruses, supporting their  
390 evolution from a viral ancestor. The division into the two superclades that our results  
391 confidently describe seems to have been the most basal event in the evolutionary  
392 history from this ancestor toward modern NCLDVs. The MAPI superclade gave rise to  
393 *Marseilleviridae*, *Ascoviridae*, Pitho-like viruses, and *Iridoviridae*. The second superclade,  
394 PAM, comprises the *Phycodnaviridae*, the *Asfarviridae*, and the “Megavirales”.  
395 Interestingly, giant viruses do not cluster together in the NCLDV trees. Most of them are  
396 present in the PAM superclade, but in two separate families (*Mimiviridae* and  
397 *Phycodnaviridae*), whereas Orpheovirus is present in the MAPI superclade (Fig 1). The  
398 scattered distribution of giant viruses within the diversity of NCLDVs strongly opposes a

399 giant – viral or cellular – ancestor scenario as proposed previously<sup>11,25</sup>. By contrast, it  
400 suggests that along the evolution of NCLDVs massive increases in genome size have  
401 occurred several times independently in different virus groups, potentially through  
402 successive steps of reduction and expansion of their genomes<sup>48,49</sup>.

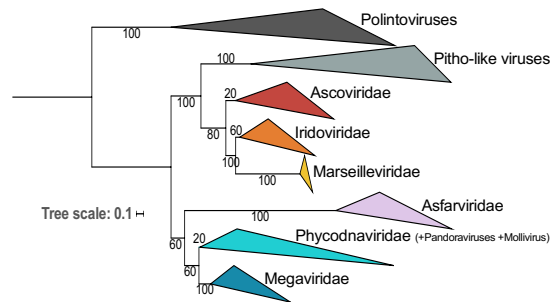
403

404 Our analyses of the two largest subunits of the RNAP, including the three  
405 eukaryotic polymerases, revealed that the genuine ortholog of the archaeal and bacterial  
406 RNAP might actually be the eukaryotic RNAP-III. In agreement with this unexpected  
407 result, homologs of the eukaryotic RNAP-III specific subunit RPC34 are present in most  
408 archaeal lineages<sup>50,51</sup>. Importantly, the inclusion in our analyses of the three eukaryotic  
409 polymerases, which emerged and were fixed in the LECA before the emergence of  
410 modern eukaryotes, provided a relative timeframe for the NCLDVs' origin and  
411 diversification. Our RNAP trees, by positioning the three monophyletic eukaryotic  
412 homologs, representing LECA, within the diversity of NCLDV families strongly imply that  
413 the evolution of NCLDVs toward the MAPI and PAM superclades and subsequent  
414 emergence of the constituent families predated the evolutionary bottleneck that marked  
415 the emergence of modern eukaryotes. Several authors have suggested that NCLDVs have  
416 played a central role in the origin of eukaryotes<sup>7,9,52-54</sup>. Our results indeed suggest that  
417 modern eukaryotes obtained two of their three RNAP, RNAP-I and RNAP-II from  
418 NCLDVs. Preliminary studies also suggested that eukaryotes obtained their major type II  
419 DNA topoisomerases from NCLDVs<sup>55</sup>. It will be interesting to test these enzymes as  
420 alternative outgroups to root the eukaryotic tree. Our results indicate that further  
421 digging into the diversity and molecular biology of NCLDV will probably have a major  
422 impact on our understanding of the origin and early evolution of eukaryotes.



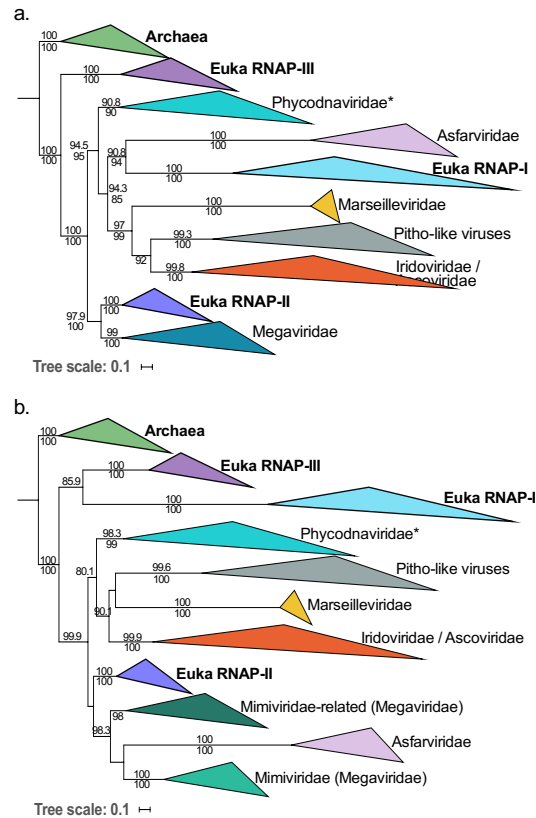
423  
 424 **Fig 1. Phylogenetic tree of the NCLDVs.** Bayesian inference (CAT-GTR model) of the  
 425 concatenated 8 core proteins from the NCLDVs after removal of *Poxviridae* and  
 426 *Aureococcus anophagefferens* virus. Genome sizes (in bp) are represented next to each  
 427 virus name. The scale-bar indicates the average number of substitutions per site. The  
 428 values at branches represent Bayesian posterior probabilities. Nodes without maximum  
 429 support are indicated in red.

430  
 431  
 432  
 433



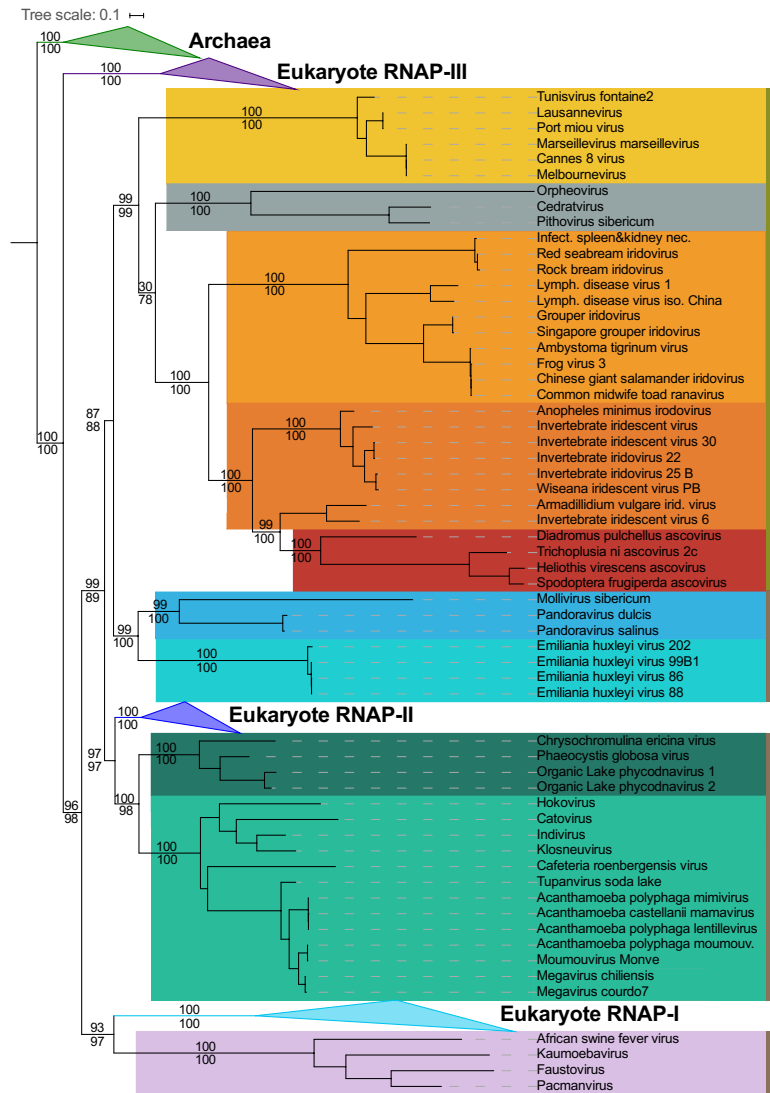
434  
 435 **Fig 2. Relationships between Polintoviruses and NCLDVs.** Maximum likelihood (ML)  
 436 phylogenetic tree of the concatenated structural proteins from Polintoviruses and  
 437 NCLDVs after removal of *Poxviridae* and *Aureococcus anophagefferens* virus. The scale-  
 438 bar indicates the average number of substitutions per site. The values at branches  
 439 represent support calculated by nonparametric bootstrap.

440  
 441  
 442

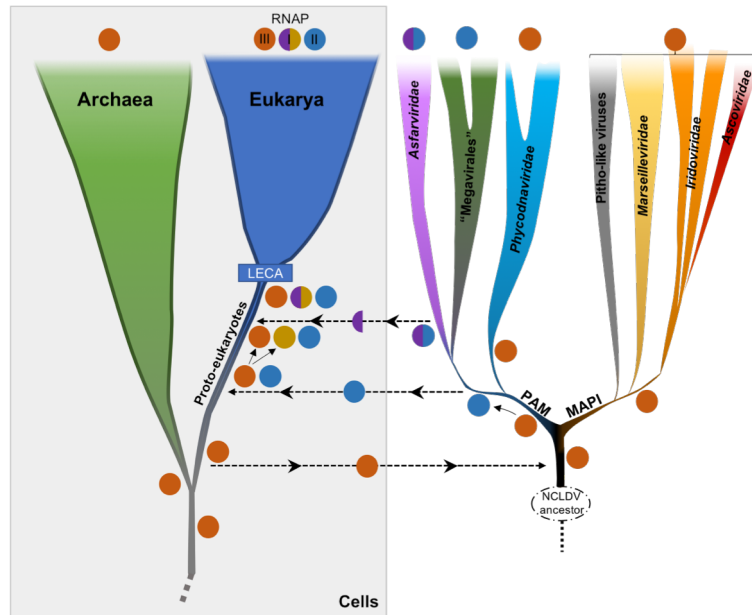


443  
444  
445  
446  
447  
448  
449  
450  
451  
452

**Fig 3. 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 (a) and RNAP-b (b) subunits, with Archaea used as the outgroup. The scale-bars indicate the average number of substitutions per site. Values on top and below branches represent support calculated by SH-like approximate likelihood ratio test (aLRT; 1,000 replicates) and ultrafast bootstrap approximation (UFBoot; 1,000 replicates), respectively. Only values superior to 80 are shown.



453  
 454 **Fig 4. Maximum likelihood (ML) phylogenetic tree of the concatenated two largest**  
 455 **RNAP subunits from Archaea, Eukaryotes, and NCLDVs.** ML phylogenetic tree of the  
 456 concatenated of the two largest RNAP subunits, with Archaea used as the outgroup.  
 457 Among the PAM superclade (light brown), “Megavirales”, *Asfarviridae*, and  
 458 *Phycodnaviridae* are indicated in light/dark green, pink, and light/dark blue,  
 459 respectively. Among the MAPI superclade (olive green), the *Marseilleviridae*, Pitho-like  
 460 viruses, *Iridoviridae*, and *Ascoviridae* are indicated in dark yellow, grey, light/dark  
 461 orange, and red, respectively. The scale-bar indicates the average number of  
 462 substitutions per site. Values on top and below branches represent support calculated  
 463 by SH-like approximate likelihood ratio test (aLRT; 1,000 replicates) and ultrafast  
 464 bootstrap approximation (UFBoot; 1,000 replicates), respectively.  
 465



466  
467  
468  
469  
470  
471  
472  
473  
474  
475

**Fig 5. Schematic representation of a putative scenario for the transfers of RNAP between cells and NCLDVs.** An ancestral RNAP that later gave rise to the eukaryotic RNAP-III, actual ortholog of the archaeal RNAP, was transferred (at least the two largest subunits) from proto-eukaryotes to the ancestor of modern NCLDVs. A significantly divergent RNAP was later on transferred from the common ancestor of *Asfarviridae* and "Megavirales" to proto-eukaryotes. A new eukaryotic RNAP also emerged from a duplication event from the RNAP-III, before its largest subunit was replaced by that of *Asfarviridae*. These events occurred before LECA, the Last Eukaryotic Common Ancestor, that marked the emergence of modern eukaryotes.

476 **Methods**

477 **Datasets**

478 We initially collected a total of 96 NCLDV genomes from public databases  
479 (Supplementary Table 2) that we used to build their core genome (see below). This  
480 dataset comprises 17 Mimiviridae, 6 Marseilleviruses, 30 Iridoviridae, 4 Ascoviridae, 14  
481 Poxviridae, 4 Asfarviridae, 15 Phycodnaviridae, 3 unclassified viruses (referred to as  
482 Pitho-like viruses), 2 Pandoraviruses, 1 Mollivirus.

483 Preliminary phylogenetic analyses showed high redundancy within some groups  
484 already comprising many members compared to others. We thus decided to remove  
485 some genomes in order to obtain a more balanced sampling (Supplementary Table 2):  
486 14 *Iridoviridae*, 2 *Phycodnaviridae* and 4 *Mimiviridae*. These analyses also revealed that  
487 the *Poxviridae* on the one hand, and a single virus (*Aureococcus anophagefferens virus*)  
488 on the other hand, always produce long branches and tend to change position in the tree  
489 depending on the considered proteins or concatenation of proteins. We thus decided to  
490 remove these viruses (14 *Poxviridae* and *Aureococcus anophagefferens virus*) from  
491 subsequent analyses, leading to the dataset of 61 genomes used in the phylogenetic  
492 analyses.

493 Ten polintoviruses sequences were collected from the Rebase collection<sup>56</sup>  
494 ([http://www.girinst.org/Rebase\\_Update.html](http://www.girinst.org/Rebase_Update.html)): Polinton-1\_HM, Polinton-3\_TC,  
495 Polinton-5\_NV, Polinton-2\_NV, Polinton-1\_DY, Polinton-1\_TC, Polinton-1\_SP, Polinton-  
496 2\_SP, Polinton-2\_DR, Polinton-1\_DR.

497 The cellular taxa included in some analyses were selected based on previous works  
498 performed by some of us<sup>44</sup>. The list of selected taxa is presented as Supplementary Table  
499 4.

500

## 501 **Core genome building**

502 Because of the high divergence level of NCLDV genomes, we were not able to directly  
503 identify genes shared among all of them. This is why we first started from two subsets of  
504 NCLDVs, both being coherent enough and comprising enough members. Those two  
505 subsets were the viruses annotated as *Mimiviridae* on the one hand and *Marseilleviridae*  
506 on the other hand.

507 For each subset of genomes, we proceeded as follow. We defined groups of orthologous  
508 genes by blasting one proteome against all the others. We only considered hits that had  
509 an E-value less than  $1e^{-10}$ . We then identified pairwise reciprocal best hits with at least  
510 20% similarity, and at least 40% of alignment coverage. We finally identified the union  
511 of all the sets of orthologs and retained those present in more than half of the members  
512 of the subset.

513 The result was two sets of orthologs, one for each subset of NCLDVs genomes. We  
514 compared these two sets by identifying the matching proteins using BLAST and HMM  
515 profiles and obtained orthologs found in both *Mimiviridae* and *Marseilleviridae*. Using  
516 the aforementioned BLAST criteria, we checked for the presence of these orthologs in  
517 other NCLDVs proteomes. When a protein was missing, we checked the presence of a  
518 corresponding gene using TBLASTN to account for incomplete annotations of the  
519 genomes, and also used HMM profiles to account for high sequence divergence. This  
520 whole process resulted in a set of putative orthologous proteins found in all NCLDV  
521 families.

522 In order to detect errors, typically different proteins assigned to the same group, we  
523 used HMMer<sup>57</sup> to find a matching HMM profile in the PFAM database  
524 (<http://pfam.xfam.org/>) for each group and discarded those significantly matching  
525 more than one PFAM profile (after checking that these profiles were not from the same



526 protein family). We finally aligned the remaining orthologs and visually inspected the  
527 alignments as a last control.

528 We obtained a list of orthologs that we ordered according to their presence in NCLDV  
529 genomes to define different categories of core proteins.

530

### 531 **Phylogenetic analyses**

#### 532 **Alignments**

533 All alignments were performed using MAFFT v7.397 and the E-INS-i algorithm<sup>58</sup>, which  
534 is designed to align sequences that are susceptible to contain large insertions. For one  
535 RNA polymerase analysis (see manuscript), constraints in the alignments were used  
536 with the seed option: independent alignments of each cellular clade (Archaea and the  
537 three eukaryotic RNA polymerases) performed separately were used as constraints for  
538 the global alignment. For the viral phylogenies, we trimmed each alignment of the  
539 positions containing more than 20% of gaps using our own scripts. For the RNA  
540 polymerase phylogenies with cellular sequences, the alignments were trimmed with  
541 BMGE (with the -m BLOSUM30 and -b 1 options)<sup>59</sup>.

542

#### 543 **Maximum likelihood phylogenies**

544 Single-protein and concatenated protein phylogenies were conducted within the  
545 Maximum Likelihood (ML) framework using IQ-TREE v1.6.3<sup>60</sup>. We first performed a  
546 model test with the Bayesian Information Criterion (BIC) by including protein mixture  
547 models<sup>61</sup>. For mixture model analyses, we used the PMSF models<sup>62</sup>. The support values  
548 were either computed from 100 bootstrap replicates in the case of nonparametric  
549 bootstrap, or from 1,000 replicates for SH-like approximation likelihood ratio test  
550 (aLRT)<sup>63</sup> and ultrafast bootstrap approximation (UFBoot)<sup>64</sup>.

551

## 552 **Congruence analysis**

553 To detect potential incongruences within the signal carried by core proteins (after  
554 removal of Poxviridae and Aureococcus anophagefferens virus) that could prevent their  
555 global concatenation, we performed comparative phylogenetic analyses of every  
556 possible combinations of 6 out of 8 core proteins through ML framework (see ML  
557 method aforementioned). The 36 ML trees generated were carefully analyzed for  
558 reference features estimated from the Bayesian phylogenetic tree (Fig 1), as well as from  
559 most phylogenetic trees obtained throughout this study. The presence or absence of  
560 these features were counted, and accordingly each feature was scored for its observed  
561 frequency among the trees, as well as each tree was scored according to the number of  
562 observed reference features (Supplementary Table 3).

563

## 564 **Supermatrix analysis**

565 We obtained a supermatrix by concatenating the 8 amino acid alignments of the core  
566 genes. Supermatrices containing more characters, we computed ML trees with the  
567 aforementioned method and performed Bayesian analyses using phyloBayes MPI  
568 v1.5a<sup>65</sup> and the CAT-GTR model<sup>66</sup>. Four independent chains were run until at least two  
569 reached convergence with a maximum difference value  $<0.1$ . The tree presented in Fig 1  
570 was obtained from the convergence (maxdiff value: 0.097) of two chains of 3,426 and  
571 3,276 generations. The first 25% of trees were removed as burn-in. The consensus tree  
572 was obtained by selecting one out of every two trees. In order to account for  
573 composition bias, we also applied two different character recodings, using 4 bins  
574 according to two different binnings: the adaptation of the 6 Dayhoff groups<sup>67</sup> to 4 bins

575 proposed by Lartillot in phyloBayes manual, and the one proposed by Susko and  
576 Rogers<sup>68</sup>. For these analyses, a GTR+ $\Gamma_4$ +I model was used.

577

### 578 **Supertree analysis**

579 Horizontal gene transfers can deeply impact tree reconstruction when using alignment-  
580 based methods. Supertree methods aim at reconciling sets of phylogenetic trees,  
581 typically gene/protein trees, into an organismal tree even when such evolutionary  
582 phenomena occur. Among the different proposed criteria for supertree methods, the  
583 subtree prune-and-regraft (SPR) distance has proven to lead to more accurate tree  
584 reconstructions<sup>69</sup>. We used the software SPR Supertree v1.2.1<sup>69</sup> from the 8 single  
585 protein phylogenies we previously inferred, after collapsing the clades for which the  
586 support was less than 95%.

587

### 588 **Ancestral sequence reconstruction**

589 In order to try to reduce the risk of long branch attraction, we replaced, in the RNAP  
590 tree, the eukaryotic clades by their ancestral sequences. These sequences were inferred  
591 using IQ-TREE. We selected sites with a posterior probability greater than 0.7 and  
592 replace the other sites by gaps.

593

### 594 **Topology test**

595 IQ-TREE v1.6.3 was used to perform Approximately Unbiased (AU) tree topology tests<sup>70</sup>  
596 for comparing the tree obtained with the concatenated RNAP genes (Fig 4) with two  
597 other ones we built using the same methodology but constraining i) the monophyly of  
598 the NCLDVs and ii) the monophyly of the cellular organisms. The AU tests rejected these  
599 two new trees with p-values <1e-3.

600

## 601 Visualization

602 The phylogenetic trees were visualized with FigTree v1.4.3

603 (<http://tree.bio.ed.ac.uk/software/figtree/>) and iTOL<sup>71</sup>.

604

605

## 606 References

- 607 1. La Scola, B. *et al.* A Giant Virus in Amoebae. *Science* **299**, 2033–2033 (2003).
- 608 2. Claverie, J.-M. Viruses take center stage in cellular evolution. *Genome Biol.* **5** (2006).
- 609 3. Raoult, D. & Forterre, P. Redefining viruses: lessons from Mimivirus. *Nat. Rev.*  
610 *Microbiol.* **6**, 315–319 (2008).
- 611 4. Moreira, D. & López-García, P. Ten reasons to exclude viruses from the tree of life.  
612 *Nat. Rev. Microbiol.* **7**, 306–311 (2009).
- 613 5. Filée, J. & Chandler, M. Gene Exchange and the Origin of Giant Viruses. *Intervirology*  
614 **53**, 354–361 (2010).
- 615 6. Forterre, P. Giant Viruses: Conflicts in Revisiting the Virus Concept. *Intervirology* **53**,  
616 362–378 (2010).
- 617 7. Nasir, A., Forterre, P., Kim, K. M. & Caetano-Anollés, G. The distribution and impact of  
618 viral lineages in domains of life. *Front. Microbiol.* **5**, 194 (2014).
- 619 8. Takemura, M., Yokobori, S. & Ogata, H. Evolution of Eukaryotic DNA Polymerases via  
620 Interaction Between Cells and Large DNA Viruses. *J. Mol. Evol.* **81**, 24–33 (2015).
- 621 9. Forterre, P. & Gaïa, M. Giant viruses and the origin of modern eukaryotes. *Curr. Opin.*  
622 *Microbiol.* **31**, 44–49 (2016).
- 623 10. Forterre, P. To be or not to be alive: How recent discoveries challenge the traditional  
624 definitions of viruses and life. *Stud. Hist. Philos. Biol. Biomed. Sci.* **59**, 100–108 (2016).
- 625 11. Claverie, J.-M. & Abergel, C. Giant viruses: The difficult breaking of multiple  
626 epistemological barriers. *Stud. Hist. Philos. Sci. Part C Stud. Hist. Philos. Biol. Biomed. Sci.* **59**,  
627 89–99 (2016).
- 628 12. Koonin, E. V. & Krupovic, M. Polintons, virophages and transpovirons: a tangled web  
629 linking viruses, transposons and immunity. *Curr. Opin. Virol.* **25**, 7–15 (2017).
- 630 13. Mihara, T. *et al.* Taxon Richness of ‘Megaviridae’ Exceeds those of Bacteria and  
631 Archaea in the Ocean. *Microbes Environ.* **33**, 162–171 (2018).
- 632 14. Legendre, M. *et al.* Thirty-thousand-year-old distant relative of giant icosahedral DNA  
633 viruses with a pandoravirus morphology. *Proc. Natl. Acad. Sci.* **111**, 4274–4279 (2014).
- 634 15. Philippe, N. *et al.* Pandoraviruses: Amoeba Viruses with Genomes Up to 2.5 Mb  
635 Reaching That of Parasitic Eukaryotes. *Science* **341**, 281–286 (2013).
- 636 16. Iyer, L. M., Aravind, L. & Koonin, E. V. Common Origin of Four Diverse Families of  
637 Large Eukaryotic DNA Viruses. *J. Virol.* **75**, 11720–11734 (2001).
- 638 17. Koonin, E. V. & Yutin, N. Nucleo-cytoplasmic Large DNA Viruses (NCLDV) of  
639 Eukaryotes. in *eLS* (ed. John Wiley & Sons, Ltd) (John Wiley & Sons, Ltd, 2012).

- 640 18. Koonin, E. V., Krupovic, M. & Yutin, N. Evolution of double-stranded DNA viruses of  
641 eukaryotes: from bacteriophages to transposons to giant viruses. *Ann. N. Y. Acad. Sci.* **1341**,  
642 10–24 (2015).
- 643 19. Brussaard, C., Kempers, R., Kop, A., Riegman, R. & Heldal, M. Virus-like particles in a  
644 summer bloom of *Emiliana huxleyi* in the North Sea. *Aquat. Microb. Ecol.* **10**, 105–113  
645 (1996).
- 646 20. Boyer, M., Madoui, M.-A., Gimenez, G., La Scola, B. & Raoult, D. Phylogenetic and  
647 phyletic studies of informational genes in genomes highlight existence of a 4 domain of life  
648 including giant viruses. *PLoS One* **5**, e15530 (2010).
- 649 21. Yutin, N., Colson, P., Raoult, D. & Koonin, E. V. Mimiviridae: clusters of orthologous  
650 genes, reconstruction of gene repertoire evolution and proposed expansion of the giant  
651 virus family. *Viol. J.* **10**, 106 (2013).
- 652 22. Yutin, N. & Koonin, E. V. Pandoraviruses are highly derived phycodnaviruses. *Biol.*  
653 *Direct* **8**, (2013).
- 654 23. Legendre, M. *et al.* Diversity and evolution of the emerging Pandoraviridae family.  
655 *Nat. Commun.* **9**, (2018).
- 656 24. Moreira, D. & López-García, P. Evolution of viruses and cells: do we need a fourth  
657 domain of life to explain the origin of eukaryotes? *Philos. Trans. R. Soc. B Biol. Sci.* **370**,  
658 20140327 (2015).
- 659 25. Claverie, J.-M. & Abergel, C. Open Questions About Giant Viruses. in *Advances in*  
660 *Virus Research* **85**, 25–56 (Elsevier, 2013).
- 661 26. Yutin, N. & Koonin, E. V. Hidden evolutionary complexity of Nucleo-Cytoplasmic Large  
662 DNA viruses of eukaryotes. *Viol. J.* **9**, 161 (2012).
- 663 27. Andreani, J. *et al.* Cedratvirus, a Double-Cork Structured Giant Virus, is a Distant  
664 Relative of Pithoviruses. *Viruses* **8**, 300 (2016).
- 665 28. Andreani, J. *et al.* Orpheovirus IHUMI-LCC2: A New Virus among the Giant Viruses.  
666 *Front. Microbiol.* **8**, (2018).
- 667 29. Gallot-Lavallée, L., Blanc, G. & Claverie, J.-M. Comparative Genomics of  
668 Chrysochromulina Ericina Virus and Other Microalga-Infecting Large DNA Viruses Highlights  
669 Their Intricate Evolutionary Relationship with the Established Mimiviridae Family. *J. Virol.* **91**,  
670 (2017).
- 671 30. Schulz, F. *et al.* Giant viruses with an expanded complement of translation system  
672 components. *Science* **356**, 82–85 (2017).
- 673 31. Abrahão, J. *et al.* Tailed giant Tupanvirus possesses the most complete translational  
674 apparatus of the known virosphere. *Nat. Commun.* **9**, (2018).
- 675 32. Reteno, D. G. *et al.* Faustovirus, an Asfarvirus-Related New Lineage of Giant Viruses  
676 Infecting Amoebae. *J. Virol.* **89**, 6585–6594 (2015).
- 677 33. Klose, T. *et al.* Structure of faustovirus, a large dsDNA virus. *Proc. Natl. Acad. Sci.* **113**,  
678 6206–6211 (2016).
- 679 34. Andreani, J. *et al.* Pacmanvirus, a New Giant Icosahedral Virus at the Crossroads  
680 between Asfarviridae and Faustoviruses. *J. Virol.* **91**, (2017).
- 681 35. Bajrai, L. *et al.* Kaumobavirus, a New Virus That Clusters with Faustoviruses and  
682 Asfarviridae. *Viruses* **8**, 278 (2016).
- 683 36. Oliveira, G. P., de Aquino, I. L. M., Luiz, A. P. M. F. & Abrahão, J. S. Putative Promoter  
684 Motif Analyses Reinforce the Evolutionary Relationships Among Faustoviruses,  
685 Kaumobavirus, and Asfarvirus. *Front. Microbiol.* **9**, (2018).

- 686 37. Bamford, D. H., Burnett, R. M. & Stuart, D. I. Evolution of Viral Structure. *Theor.*  
687 *Popul. Biol.* **61**, 461–470 (2002).
- 688 38. Krupovic, M. & Bamford, D. H. Virus evolution: how far does the double beta-barrel  
689 viral lineage extend? *Nat. Rev. Microbiol.* **6**, 941–948 (2008).
- 690 39. Abrescia, N. G. A., Bamford, D. H., Grimes, J. M. & Stuart, D. I. Structure Unifies the  
691 Viral Universe. *Annu. Rev. Biochem.* **81**, 795–822 (2012).
- 692 40. Krupovic, M., Bamford, D. H. & Koonin, E. V. Conservation of major and minor jelly-  
693 roll capsid proteins in Polinton (Maverick) transposons suggests that they are bona fide  
694 viruses. *Biol. Direct* **9**, 6 (2014).
- 695 41. Krupovic, M. & Koonin, E. V. Polintons: a hotbed of eukaryotic virus, transposon and  
696 plasmid evolution. *Nat. Rev. Microbiol.* **13**, 105–115 (2015).
- 697 42. Fischer, M. G. Giant viruses come of age. *Curr. Opin. Microbiol.* **31**, 50–57 (2016).
- 698 43. Filée, J., Forterre, P., Sen-Lin, T. & Laurent, J. Evolution of DNA Polymerase Families:  
699 Evidences for Multiple Gene Exchange Between Cellular and Viral Proteins. *J. Mol. Evol.* **54**,  
700 763–773 (2002).
- 701 44. Da Cunha, V., Gaia, M., Gadelle, D., Nasir, A. & Forterre, P. Lokiarchaea are close  
702 relatives of Euryarchaeota, not bridging the gap between prokaryotes and eukaryotes. *PLoS*  
703 *Genet.* **13**, e1006810 (2017).
- 704 45. Werner, F. & Grohmann, D. Evolution of multisubunit RNA polymerases in the three  
705 domains of life. *Nat. Rev. Microbiol.* **9**, 85–98 (2011).
- 706 46. Da Cunha, V., Gaia, M., Nasir, A. & Forterre, P. Asgard archaea do not close the  
707 debate about the universal tree of life topology. *PLOS Genet.* **14**, e1007215 (2018).
- 708 47. Yutin, N., Wolf, Y. I. & Koonin, E. V. Origin of giant viruses from smaller DNA viruses  
709 not from a fourth domain of cellular life. *Virology* **466–467**, 38–52 (2014).
- 710 48. Filée, J. Route of NCLDV evolution: the genomic accordion. *Curr. Opin. Virol.* **3**, 595–  
711 599 (2013).
- 712 49. Filée, J. Giant viruses and their mobile genetic elements: the molecular symbiosis  
713 hypothesis. *Curr. Opin. Virol.* **33**, 81–88 (2018).
- 714 50. Eme, L., Spang, A., Lombard, J., Stairs, C. W. & Ettema, T. J. G. Archaea and the origin  
715 of eukaryotes. *Nat. Rev. Microbiol.* **15**, 711–723 (2017).
- 716 51. Blombach, F. *et al.* Identification of an ortholog of the eukaryotic RNA polymerase III  
717 subunit RPC34 in Crenarchaeota and Thaumarchaeota suggests specialization of RNA  
718 polymerases for coding and non-coding RNAs in Archaea. *Biol. Direct* **4**, 39 (2009).
- 719 52. Takemura, M. Poxviruses and the Origin of the Eukaryotic Nucleus. *J. Mol. Evol.* **52**,  
720 419–425 (2001).
- 721 53. Bell, P. J. Viral Eukaryogenesis: Was the Ancestor of the Nucleus a Complex DNA  
722 Virus? *J. Mol. Evol.* **53**, 251–256 (2001).
- 723 54. Forterre, P. & Prangishvili, D. The Great Billion-year War between Ribosome- and  
724 Capsid-encoding Organisms (Cells and Viruses) as the Major Source of Evolutionary  
725 Novelties. *Ann. N. Y. Acad. Sci.* **1178**, 65–77 (2009).
- 726 55. Forterre, P., Gribaldo, S., Gadelle, D. & Serre, M.-C. Origin and evolution of DNA  
727 topoisomerases. *Biochimie* **89**, 427–446 (2007).
- 728 56. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements.  
729 *Cytogenet. Genome Res.* **110**, 462–467 (2005).
- 730 57. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195  
731 (2011).

- 732 58. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7:  
733 improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
- 734 59. Criscuolo, A. & Gribaldo, S. BMGE (Block Mapping and Gathering with Entropy): a  
735 new software for selection of phylogenetic informative regions from multiple sequence  
736 alignments. *BMC Evol. Biol.* **10**, 210 (2010).
- 737 60. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and  
738 effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol.*  
739 *Evol.* **32**, 268–274 (2015).
- 740 61. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S.  
741 ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**,  
742 587–589 (2017).
- 743 62. Wang, H.-C., Minh, B. Q., Susko, E. & Roger, A. J. Modeling Site Heterogeneity with  
744 Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation. *Syst.*  
745 *Biol.* **67**, 216–235 (2018).
- 746 63. Guindon, S. *et al.* New Algorithms and Methods to Estimate Maximum-Likelihood  
747 Phylogenies: Assessing the Performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
- 748 64. Minh, B. Q., Nguyen, M. A. T. & von Haeseler, A. Ultrafast Approximation for  
749 Phylogenetic Bootstrap. *Mol. Biol. Evol.* **30**, 1188–1195 (2013).
- 750 65. Lartillot, N., Lepage, T. & Blanquart, S. PhyloBayes 3: a Bayesian software package for  
751 phylogenetic reconstruction and molecular dating. *Bioinforma. Oxf. Engl.* **25**, 2286–2288  
752 (2009).
- 753 66. Lartillot, N. & Philippe, H. A Bayesian mixture model for across-site heterogeneities in  
754 the amino-acid replacement process. *Mol. Biol. Evol.* **21**, 1095–1109 (2004).
- 755 67. Embley, T. M., van der Giezen, M., Horner, D. S., Dyal, P. L. & Foster, P. Mitochondria  
756 and hydrogenosomes are two forms of the same fundamental organelle. *Philos. Trans. R.*  
757 *Soc. Lond. B. Biol. Sci.* **358**, 191–201–202 (2003).
- 758 68. Susko, E. & Roger, A. J. On reduced amino acid alphabets for phylogenetic inference.  
759 *Mol. Biol. Evol.* **24**, 2139–2150 (2007).
- 760 69. Whidden, C., Zeh, N. & Beiko, R. G. Supertrees Based on the Subtree Prune-and-  
761 Regraft Distance. *Syst. Biol.* **63**, 566–581 (2014).
- 762 70. Shimodaira, H. An Approximately Unbiased Test of Phylogenetic Tree Selection. *Syst.*  
763 *Biol.* **51**, 492–508 (2002).
- 764 71. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display  
765 and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **44**, W242–245 (2016).
- 766