



HAL
open science

Past and present giant viruses diversity explored through permafrost metagenomics

Sofia Rigou, Sébastien Santini, Chantal Abergel, Jean-Michel Claverie,
Matthieu Legendre

► To cite this version:

Sofia Rigou, Sébastien Santini, Chantal Abergel, Jean-Michel Claverie, Matthieu Legendre. Past and present giant viruses diversity explored through permafrost metagenomics. 2022. hal-03810478

HAL Id: hal-03810478

<https://hal.science/hal-03810478>

Preprint submitted on 11 Oct 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Past and present giant viruses diversity explored through permafrost metagenomics

Sofia Rigou

Aix-Marseille University <https://orcid.org/0000-0002-9107-3509>

Sébastien Santini

CNRS-AMU UMR7256

Chantal Abergel

Aix-Marseille University

Jean-Michel Claverie

University of Mediterranee School of Medicine

Matthieu Legendre (✉ legendre@igs.cnrs-mrs.fr)

Aix-Marseille University - CNRS <https://orcid.org/0000-0002-8413-2910>

Article

Keywords:

Posted Date: February 10th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1328080/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at Nature Communications on October 7th, 2022. See the published version at <https://doi.org/10.1038/s41467-022-33633-x>.

Past and present giant viruses diversity explored through permafrost metagenomics

Sofia Rigou¹, Sébastien Santini¹, Chantal Abergel¹, Jean-Michel Claverie¹, Matthieu Legendre^{1,*}

¹Aix–Marseille University, Centre National de la Recherche Scientifique, Information Génomique & Structurale, Unité Mixte de Recherche 7256 (Institut de Microbiologie de la Méditerranée, FR3479), 13288 Marseille Cedex 9, France

*Correspondence: legendre@igs.cnrs-mrs.fr

Abstract

Giant viruses are abundant in aquatic environments and ecologically important through the metabolic reprogramming of their hosts. Lesser is known about giant viruses from soil although two of them, belonging to different families, were reactivated from 30,000-y-old Siberian permafrost samples, suggesting an untapped diversity of Nucleocytoviricota in this environment. Through permafrost metagenomics, we reveal a high heterogeneity in the abundance of giant viruses representing up to 12% of the total community in one sample. *Pithoviridae* and *Orpheoviridae*-like viruses were the most important contributors, followed by *Mimiviridae*. A comparison to other terrestrial metagenomes confirmed that the diversity pattern in these samples is quite unique. In contrast, *Pandoraviridae* sequences remained scarce. Using a stringent methodology, we were able to assemble large genomes, including a complete circular 1.6 Mb

21 *Pithoviridae*-like from a 42,000-y-old sample. The uncovered *Pithoviridae* diversity
22 also provided insights into the family evolution. The phylogenetic reconstruction of
23 specific functions not only revealed gene transfers between cells and viruses, but also
24 between viruses from different families. Finally, the extensive annotation of the
25 permafrost viral sequences revealed a patchwork of predicted functions amidst an
26 even larger reservoir of anonymous genes of unknown functions.

27

28 **Introduction**

29 The permafrost, soil remaining continuously frozen for at least 2 years, covers 15% of
30 the Northern hemisphere (1) and gathers complex communities of living organisms
31 and variable soil types. The microbial community of the surface cryosol is in some
32 cases subject to freezing and thawing every year (2) whereas communities from
33 deeper layers are trapped in the syngenetic (as old as the sediment) or epigenetic
34 (more recent than the sediment) permafrost. Pleistocene permafrost has been showed
35 to harbor up to 5×10^7 cells per wet gram of soil of which about a fifth is alive (3). The
36 permafrost has thus the ability to preserve organisms for tens if not hundreds
37 thousands of years and acts as a huge reservoir of ancient microorganisms. It has
38 been shown for instance that numerous bacteria isolated from permafrost samples
39 remained viable (4, 5), even potentially up to 1.1 million years (6). Even in low
40 biomass-containing frozen environments such as glacier ice, metagenomics
41 approaches have recently revealed hundreds of distinct bacterial genera (7).
42 Unicellular (8–10) and even multicellular (11, 12) eukaryotes can also be preserved
43 for thousands of years and be revived from such frozen environments.

44 Besides cellular organisms, metagenomics studies have revealed bacteriophages
45 communities archived in surface (13) or deeper (7) glacier ice, the majority of which
46 being taxonomically unassigned. Following the high bacterial abundance (14),
47 bacteriophages are expected to be the most abundant viruses in the permafrost.
48 However, in the unfiltered size fraction, the eukaryotic viruses Nucleocytoviricota
49 (formerly known as Nucleocytoplasmic large DNA viruses or NCLDVs) are also highly
50 represented (14). This phylum gathers large double stranded DNA viruses such as
51 Pokkesviricetes (*Poxviridae* and *Asfarviridae*) as well as all the known giant viruses

52 (i.e. viruses visible by light microscopy): the Megaviricetes (*Phycodnaviridae*,
53 *Mimiviridae* and Pimascovirales). Likewise, a handful of scaffolds of potential
54 *Phycodnaviridae*, also belonging to this phylum, were identified in a metagenomic
55 study of glacial environments (13). More importantly, among Nucleocytoviricota, two
56 giant viruses, namely Pithovirus sibericum and Mollivirus sibericum, were reactivated
57 from a 30,000-y-old permafrost sample on *Acanthamoeba castellanii* (15, 16).
58 Together with the presence of numerous protists and in particular amoeba in
59 permafrost (9), this hints at the existence of many more giant viruses in such
60 environments.

61 Recently, several studies specifically targeting the viral dark matter from
62 environmental metagenomics data have started to grasp the diversity and gene-
63 content of the Nucleocytoviricota (17–19). It became clear that the genomes of these
64 viruses code for various auxiliary metabolic genes, making them capable of
65 reprogramming their host's metabolism and hence are potentially important drivers of
66 global biogeochemical cycles (17, 18, 20). They also seem to be widespread in aquatic
67 environments. More specifically, *Mimiviridae* (in particular the proposed
68 *Mesomimivirinae* sub-family (21)) and *Phycodnaviridae* are major contributors of the
69 marine viromes all over the world, as revealed by thousands of metagenome-
70 assembled viral genome (MAG) sequences (17–19). They also have been found
71 active at the surface layer of the ocean by metatranscriptomics (22). The
72 Nucleocytoviricota ecological functions and diversity in terrestrial samples on the other
73 hand is far less known, with the exception of *Klosneuvirinae* sequences recovered
74 from forest soil samples (23) and of *Pithoviridae* sequences assembled from the Loki's
75 castle deep sea sediments sequences (24). The overwhelming proportion of marine-

76 related as compared to terrestrial Nucleocytoviricota sequences from metagenomic
77 studies is most likely due to the difficulty at revealing their hidden diversity in these
78 environments (23). Soils host highly complex microbial communities making
79 metagenomic studies notoriously challenging as population heterogeneity with closely
80 related strains can hamper sequence assembly (25, 26).

81 Current giant viruses' metagenomic studies rely on the detection of Nucleocytoviricota
82 core genes (17, 18, 23, 24). However, among the very few shared genes some are
83 highly divergent or even completely absent from certain viral families. For instance, a
84 packaging ATPase, presumably encoded by a "core" gene in large DNA viruses, is
85 absent in *Pithoviridae* (27). Likewise, the Major Capsid Protein (MCP) often used as a
86 marker gene to detect Nucleocytoviricota within metagenomic assemblies (18) is only
87 present in a divergent form in *Pithoviridae* (15) and completely absent from
88 *Pandoraviridae* (27, 28). Thus, the probability to detect these types of non-icosahedral
89 giant viruses is drastically lowered.

90 Although two distinct non-icosahedral giant viruses were initially isolated from
91 permafrost samples (15, 16) little is known on the Nucleocytoviricota diversity in this
92 type of environment. Here we propose an analysis of these viruses from eleven
93 permafrost samples ranging from the active layer up to 49,000-y-old (14). We show
94 that the permafrost is a great source of viral diversity. Although the samples are very
95 heterogeneous in Nucleocytoviricota content, they reach up to 5% of the assembled
96 sequences and 12% of the total coverage in one deep permafrost sample. We found
97 here that *Pithoviridae* and *Orpheoviridae*-like families as well as *Mimiviridae* are the
98 main contributors of the giant virus diversity of the deep permafrost.

99 **Results**

100 **Cryosol metagenomes assemblies**

101 We gathered permafrost and surface cryosol raw metagenomic data produced by (14)
102 on the three surface samples from Kamchatka (C-D-E, Table S1) that are also the
103 samples from which Cedratvirus kamchatka (29) and Mollivirus kamchatka (30) were
104 isolated, and on eight deep samples from the Yukechi Alas area radio-carbon-dated
105 from 53 to over 49,000-y-old, seven of which are syngenetic (Table S1).

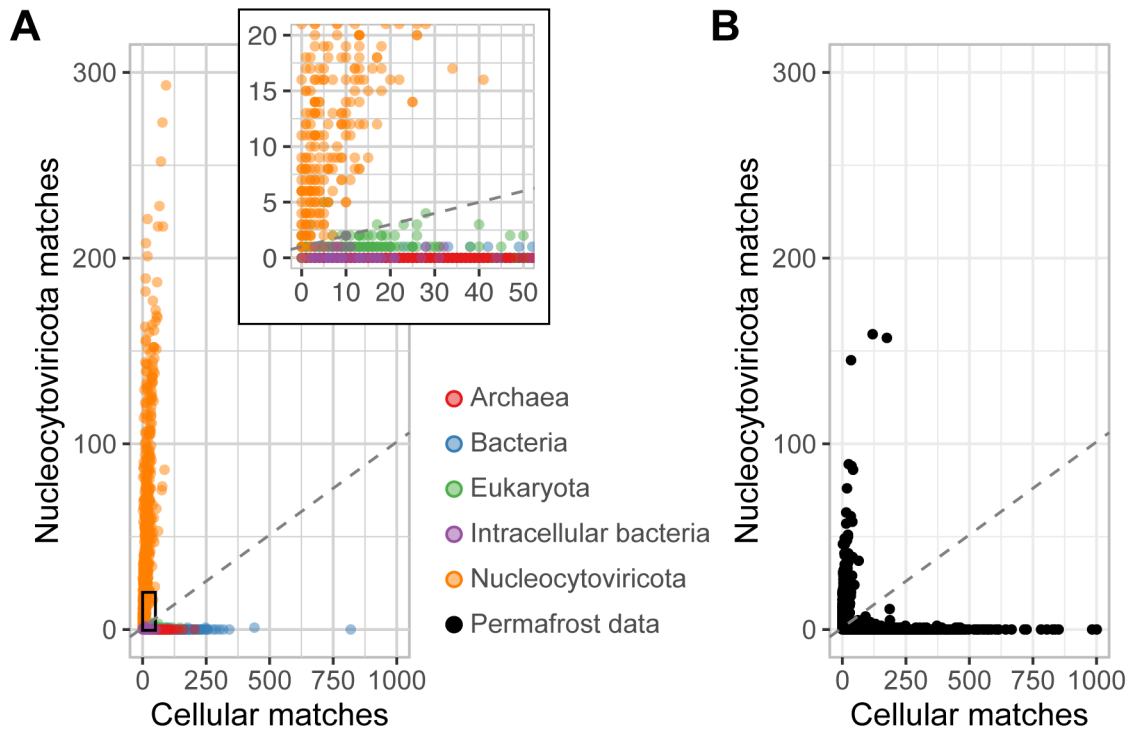
106 We first performed an assembly of the reads (Table S2) followed by binning and
107 assessed the quality of the bins, mostly composed of prokaryotic sequences (90%),
108 using Checkm (32). This revealed potential chimeras (Fig. S1A). We thus chose not
109 to consider them as unique organisms but instead used binning as a procedure to
110 decrease complexity in our datasets. The reads were first separated according to the
111 bin they belonged to and a second *de novo* assembly was made within each bin. This
112 resulted in significantly longer scaffolds and a larger total assembly (Table S2).
113 Applying Checkm to this final dataset identified nearly no chimera (Fig. S1B). Thus,
114 our method significantly gained in reliability by lowering the proportion of chimeras in
115 comparison to conventional binning.

116 **Discriminating Nucleocytoviricota in metagenomic samples**

117 From this dataset we then sought to extract Nucleocytoviricota sequences. Our
118 method is based on the detection of both Nucleocytoviricota genes (including the ones
119 specific to the non-icosahedral *Pithoviridae* and *Pandoraviridae*) and cellular ones.
120 Clearly the combination of the two showed a very distinct pattern for Nucleocytoviricota
121 compared to cellular genomic sequences (Fig. 1A), as revealed by a control
122 metagenomic mimicking database containing reference Nucleocytoviricota genomes

123 from (31), cellular genomes randomly sampled from Genbank in addition to amoeba
124 and algae genomes (known to be the hosts of Nucleocytoviricota) as well as amoeba-
125 hosted intracellular bacteria (*Babela massiliensis* and *Parachlamydia*
126 *acanthamoebae*). The control database was also used to find the optimal parameters
127 discriminating Nucleocytoviricota sequences (slope = 0.1, intercept = 1; Fig. 1),
128 yielding high classification performance (sensitivity = 98.16% and specificity \geq 99.53%;
129 Fig. S2). For comparison we also tested the Viralrecall tool (35) that confirmed 1848
130 out of the 1973 (94%) scaffolds we detected. Finally, further controls for contamination
131 in the Nucleocytoviricota dataset involved a search for ribosomal sequences, none of
132 which were found. Manual functional annotation also allowed the identification of 7
133 scaffolds potentially belonging to intracellular bacteria, a phage and a nudivirus that
134 were removed. At the end, our Nucleocytoviricota identification method on the
135 permafrost dataset resulted in 1966 scaffolds ranging from 10 kb up to 1.6 Mb,
136 corresponding to 1% of all scaffolds over 10 kb in size (Fig. 1B).

137



138

139 **Figure 1: Extraction method of viral scaffolds**

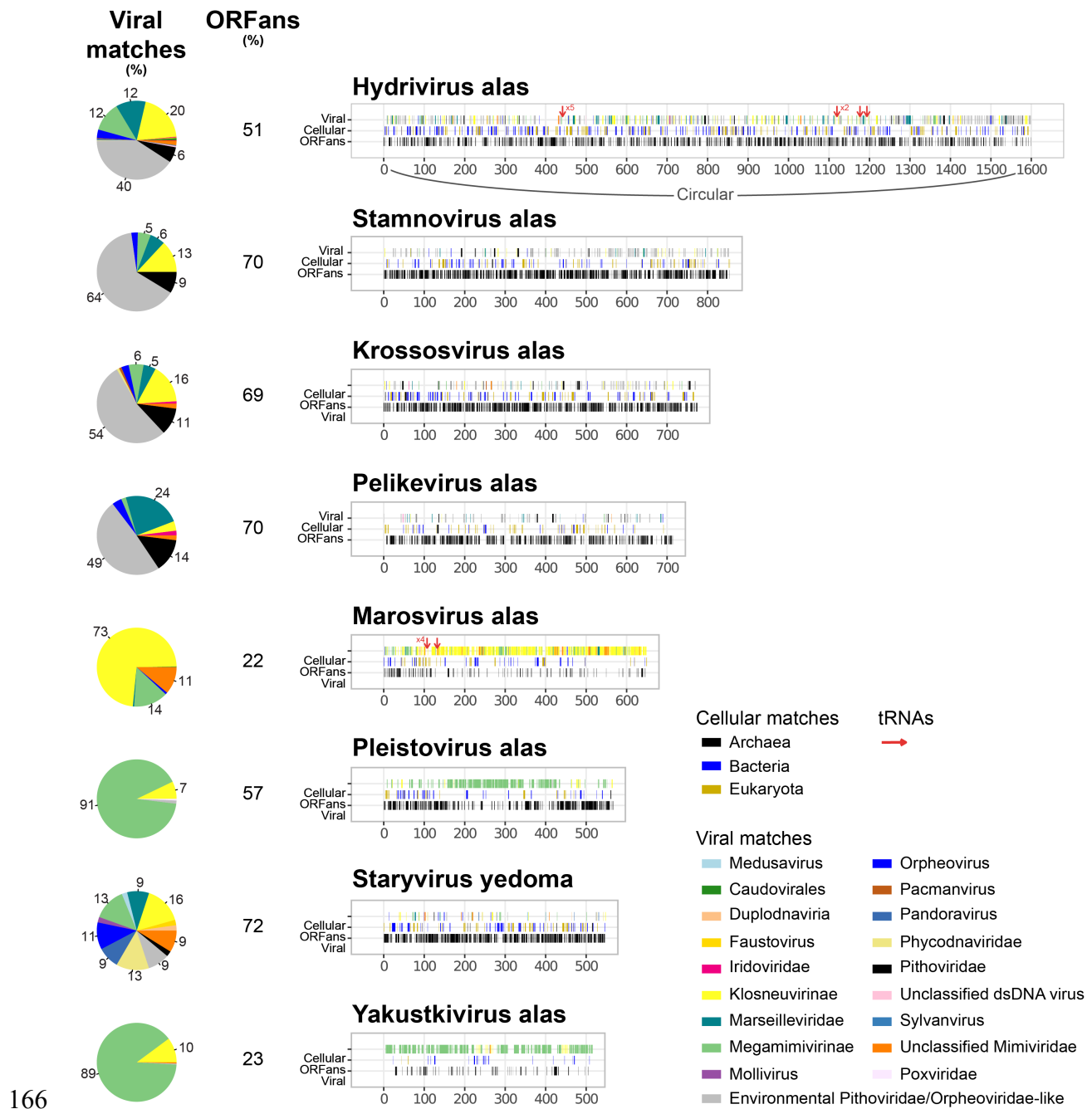
140 Each point corresponds to one scaffold. Viral matches (y-axis) were counted as the number of ORFs
 141 matching a Nucleocytoviricota-specific HMM. These HMMs come from a previous study (18) to which
 142 were added specific HMMs from the VOG database and HMMs constructed on *Pandoraviridae* and
 143 *Pithoviridae* genomes. Cellular matches (x-axis) are the number of Diamond blastP matches against
 144 the cellular Refseq database with a threshold of 35% of sequence identity. The dashed lines represent
 145 the chosen threshold excluding all point under or on the line. (A) Control dataset. The inset is a zoom
 146 of the bottom-left corner of the plot. For clarity, 1 bacterial point with over 1000 cellular matches and 1
 147 viral match are not shown. (B) Permafrost data. For clarity, 5 points with over 1000 cellular matches are
 148 not shown.

149 As said, Nucleocytoviricota metagenomic studies often rely on the MCP as a bait,
 150 making it hard, if not impossible, to catch some of the non-icosahedral viruses. By
 151 adding *Pithoviridae* and *Pandoraviridae* HMMs to the original profiles (18) and VOG's
 152 HMMs, we gained 5% (n = 110) more scaffolds that were mainly unclassified or from
 153 *Pithoviridae* and divergent *Pithoviridae* families (see further for phylogenies).

154 **Large viral genomes from deep permafrost**

155 Although our strategy to exclude conventional binning was primarily designed to
156 capture high confidence MAGs at the price of completeness, we were still able to
157 recover large Nucleocytoviricota genomes in single scaffolds with no apparent
158 chimera. Eight of them, assembled from 16m to 19m deep permafrost samples (R, N
159 and M, Table S1) dating from 42,000 to 49,000 years, reached over 500kb (Fig. 2).
160 The largest one of 1.6 Mb, referred to as Hydrivirus alas, is most likely complete as it
161 was successfully circularized. Although these large scaffolds are deeply sequenced
162 (with an average coverage in between 14 and 72), they are far from belonging to the
163 most abundant viruses in their samples (the highest coverages are of 53, 181 and
164 1572 in samples M, N, R respectively).

165



166

167 **Figure 2: Gene content of the large genomes recovered from ancient**
 168 **permafrost samples**

169 For each genome, the position of ORFans (ORFs with no match in the NR database), cellular and viral
 170 matches are recorded along the genome. The positions of tRNAs are also showed as red arrows. The
 171 pie charts present the proportion and taxonomy of viral matches with slices $\geq 5\%$ labeled. The
 172 environmental *Pithoviridae/Orpheoviridae*-like category contains metagenomic sequences from (23,
 173 24). The Hydrivirus alas genome was circularized.

174 These MAGs vary in divergence from known genomes, having from 22% up to 72%
175 of ORFans for *Staryvirus yedoma* (Fig. 2). As always for newly discovered giant
176 viruses, their genomes also match cellular genes from all domains of life (with very
177 few Archaea). When looking at the viral matches, two scaffolds seem close to
178 *Megamimivirinae* (*Pleistovirus alas* and *Yakustkivirus alas*), one to *Klosneuvirinae*
179 (*Marosvirus alas*) and four to *Pithoviridae/Orpheoviridae* (*Hydrivirus alas*, *Stamnovirus*
180 *alas*, *Krossosvirus alas* and *Pelikevirus alas*). The most divergent, *Staryvirus yedoma*,
181 shows an even distribution of viral best BlastP matches with no specific family standing
182 out (Fig 2). Together with its high ORFan content, this suggests that it belongs to a
183 Nucleocytoviricota viral family with no previous isolate so far.

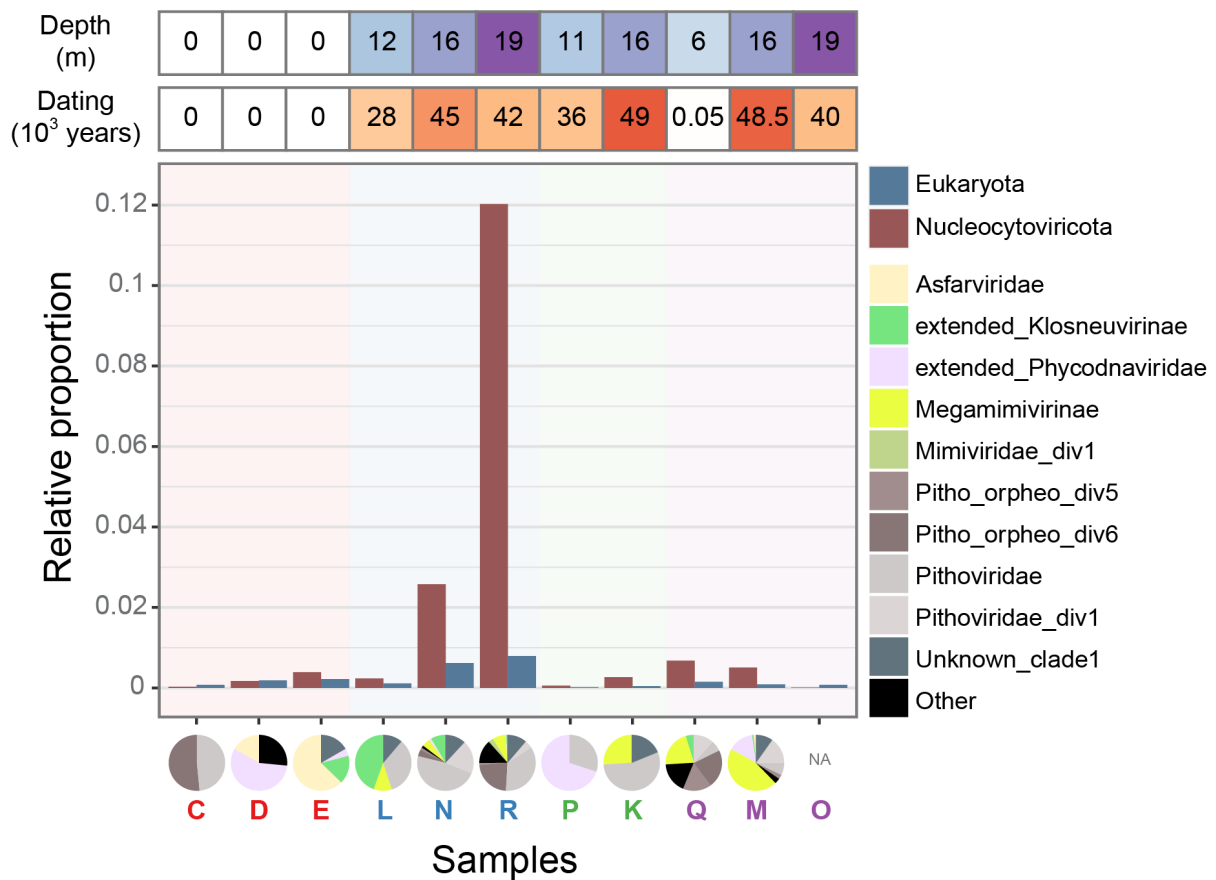
184 The complete 1.6 Mb *Hydrivirus alas* genome reaches a size similar to the isolated
185 *Orpheovirus* (32). The other 715 to 855 kb scaffolds (*Stamnovirus alas*, *Krossosvirus*
186 *alas* and *Pelikevirus alas*) are slightly larger than isolated *Pithoviridae* (ca. 600 kb) (15,
187 33, 34). However they were not circularized as expected for a *Pithoviridae* genome
188 structure (15) and are potentially even larger. Still, in the four of them, most of the core
189 genes are present (Table S3). Furthermore, except for *Pelikevirus alas*, all the
190 *Pithoviridae*-like large genomes and *Marosvirus alas* have a near complete base
191 excision repair system.

192 **Heterogeneous Nucleocytoviricota abundance of the Russian** 193 **cryosol**

194 The permafrost samples were very heterogeneous in Nucleocytoviricota relative
195 abundance (Fig. 3) and number of scaffolds, ranging from 2 found in the sample O
196 (core permafrost under a lake in Yedoma, frozen for 40,000 years) to 721 scaffolds
197 found in sample R (core permafrost under a drained thermokast lake, frozen for over
198 42,000 years). This sample was also the richest in eukaryotes with mostly

199 Streptophyta (35%), Dikarya (14%), Platyhelminthes (9%), Eumycetozoa (8%) and
 200 Longamoebia (7%). In other permafrost core samples, the most abundant eukaryote
 201 clades were Streptophyta (34%), Dikarya (18%), Chordata (7%), Arthropoda (5%) and
 202 Eumycetozoa (3%) (Fig. S3). Interestingly, amoebas (Longamoebia) are a lot more
 203 abundant in sample R than in other samples (Fig. S3).

204



205

206 **Figure 3: Relative abundance of Nucleocytoviricota and Eukaryota across**
 207 **samples**

208 The relative abundance is calculated as the sum of coverages belonging to the given group divided by
 209 the total sample coverage. Sample names in red are surface samples from Kamchatka while samples
 210 in blue, green and purple indicate that they come from three different forages in the Yukechi Alas area.
 211 The pie charts indicate the taxonomy of the Nucleocytoviricota in different samples (see further for
 212 phylogeny). Only classified scaffolds were considered.

213 The relative proportion of giant viruses abundance (Fig. 3) and the number of scaffolds
214 were correlated to the ones of Eukaryota, with Spearman correlation coefficients of
215 $\rho=0.72$ (p-value=0.017) and $\rho=0.83$ (p-value=0.003), respectively. Such correlation
216 could simply be explained by host-parasites dynamics. Alternatively, one could
217 hypothesize that Nucleocytoviricota scaffolds correspond to endogenized viruses in
218 eukaryotes (GEVE), as previously shown in green algae (35). The confusion is
219 possible as 57% (193 out of 338) of the GEVE pseudo-contigs (see Methods) were
220 captured by our Nucleocytoviricota detection method. To explore this possibility, we
221 thus checked for endogenization signs in the viral scaffolds using Viralrecall (36)
222 (example in Fig. S4) but none was found. In addition, Nucleocytoviricota largely
223 outnumber eukaryotes with a 4:1 Nucleocytoviricota/Eukaryota ratio in the sum of
224 coverages (mean=4.06, sd=4.22) and number of scaffolds (mean=4.40, sd=3.34).
225 Altogether, this suggests that most of the discovered permafrost Nucleocytoviricota
226 scaffolds correspond to *bona fide* free viruses.

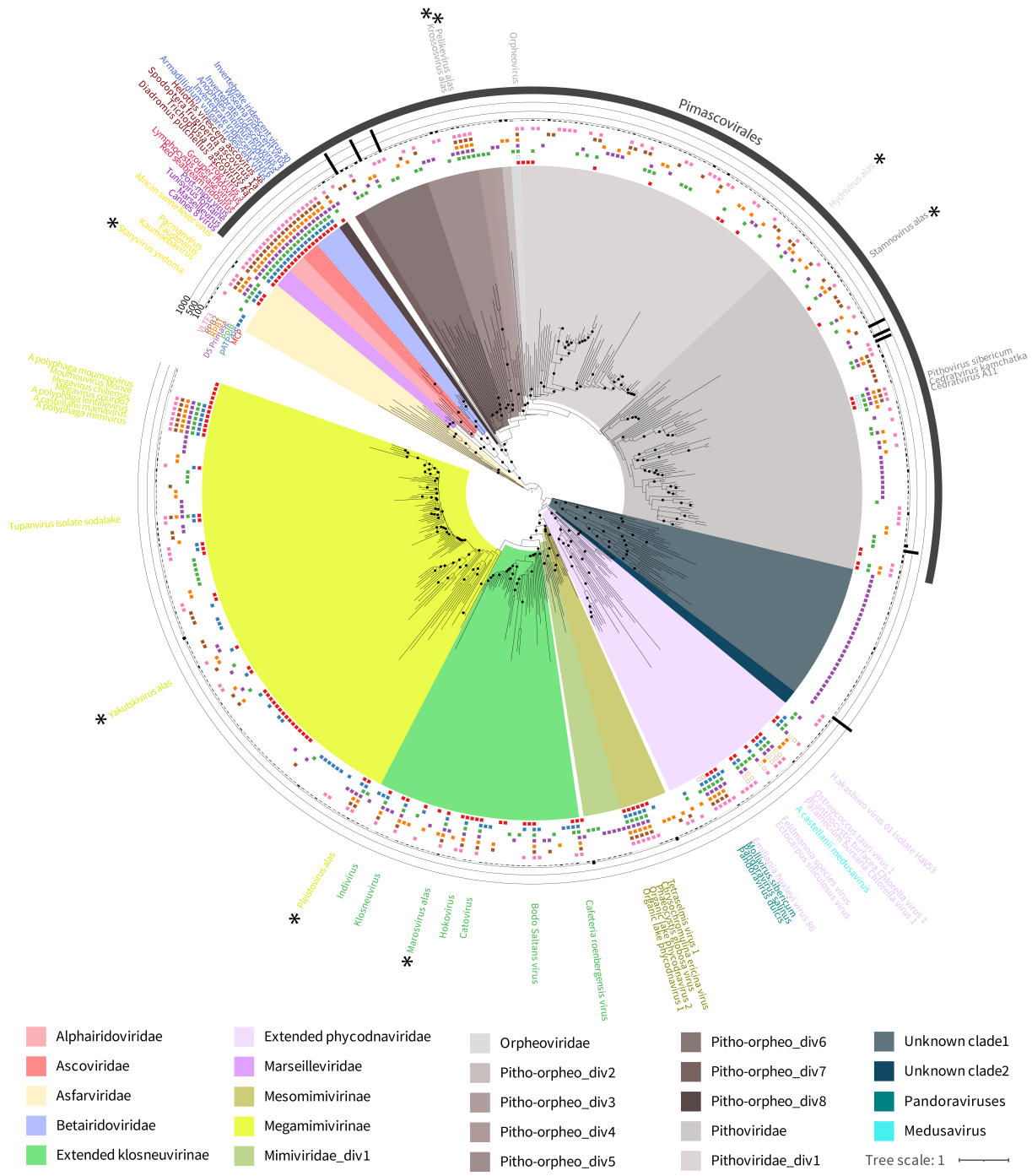
227 **Exploration of the sequence diversity**

228 To further investigate which viral families were present in the samples, we conducted
229 a phylogenetic analysis based on the 7 marker genes (Table S3) from (31) We
230 excluded the transcription elongation factor TFIIS from our analysis as its evolution
231 was unclear and not coherent with the known Nucleocytoviricota clades (Fig. S5). It
232 should also be noted that the primase D5 revealed an unexpected grouping of the
233 Cedratviruses with *Phycodnaviridae* instead of *Pithoviridae*, suggesting that this gene
234 was acquired from an unknown source in Cedratviruses (Fig. S5).

235 With this method, 369 Nucleocytoviricota scaffolds (19%) were taxonomically
236 classified (Fig. 4) corresponding to 40% of the 72 Mb of total Nucleocytoviricota

237 identified sequences. *Pithoviridae* and *Orpheoviridae*-like viral families were clearly
238 the most diverse followed by *Megamimivirinae*. In contrast, *Marseilleviridae*,
239 *Alphairidoviridae*, *Betairidoviridae* and *Ascoviridae* were completely absent from our
240 samples. *Poxviridae* were not included in the phylogeny as they were absent from our
241 samples and adding their marker genes lowered the tree bootstraps values. In addition
242 to our strategy to combine different marker genes, we also computed a phylogenetic
243 tree from a single conserved one, the DNA polymerase, confirming that *Pithoviridae*
244 and *Orpheoviridae*-like sequences were the most diverse families in our samples (Fig.
245 S6).

246



247

248 **Figure 4: Nucleocytoviricota phylogeny in all samples**

249 Consensus of 1000 bootstrapped trees calculated by a partitioned analysis on 7 marker genes. Black
 250 dots represent branch bootstrap support ≥ 90 . The colored labels indicate the reference genomes and
 251 the large viruses identified in this study (marked with a star). The colored clades were manually created
 252 to be monophyletic based on reference genomes. The marker genes used for this phylogeny are
 253 indicated as colored squares. Empty squares correspond to marker genes absent from the reference

254 genomes. Black bars show the normalized mean coverage of the scaffold. Pimascovirales are defined
255 as the clade composed of all the *Ascoviridae*, *Iridoviridae*, *Marseilleviridae*, but also *Pithoviridae*,
256 *Orpheoviridae* and the metagenomic intermediate clades. The Extended_phycodnaviridae group
257 includes *Pandoraviridae* and Mollivirus. The Extended_klosneuvirinae group includes the Cafeteria
258 roenbergensis virus.

259 The permafrost data appears to reveal a whole new clade branching before the
260 *Phycodnaviridae* and with no previously isolated representatives (Fig. 4). This is
261 probably an artefact due to the divergent Cedratviruses primase D5 gene closer to the
262 *Phycodnaviridae* primase. On the primase D5 tree (Fig. S7), the clade is split between
263 Cedratviruses and the other half (probably *Phycodnaviridae*) that remains in the same
264 position in the tree. A second unknown clade branching right before *Phycodnaviridae*
265 (Fig. 4) had four members and was mildly supported by the bootstrap analysis (71%).

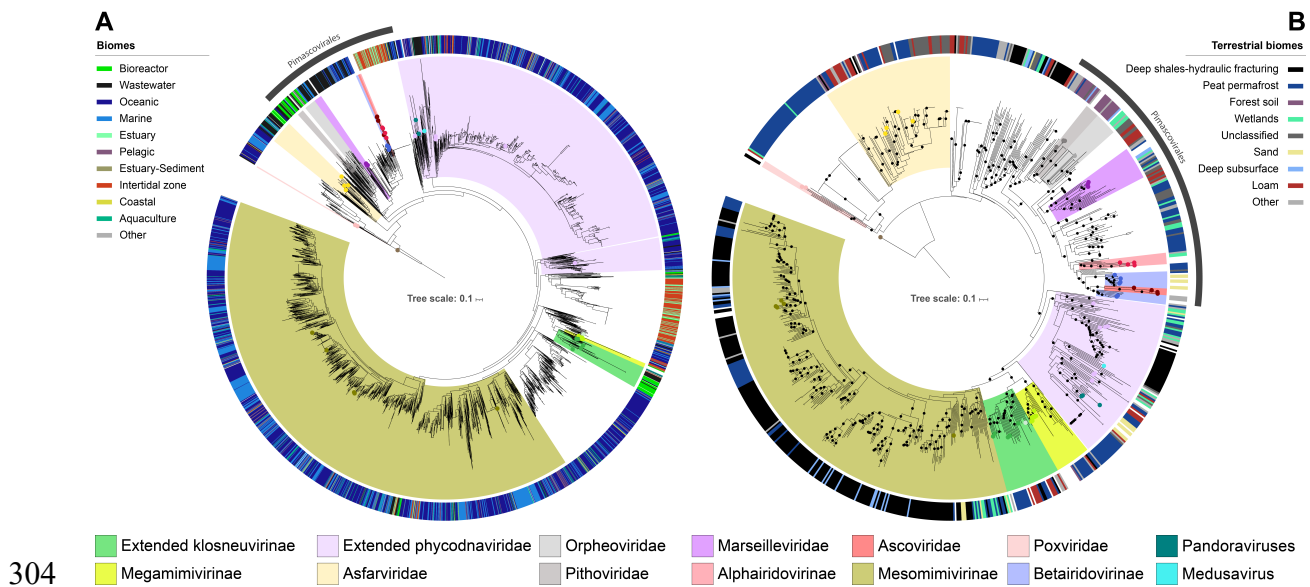
266 In order to explore the genome content diversity, we next analyzed the best BlastP
267 matches against the NR database. Sequences unclassified by our phylogenetic
268 approach were dominated (50.6%) by ORFans, in the same range than
269 phylogenetically classified permafrost scaffolds (from 25.2% to 73.9% with on average
270 54%, Fig. S8A). This suggests that these sequences are not more divergent to known
271 relatives than any other Nucleocytoviricota sequence. They remained unclassified
272 simply because they lack the marker genes. Secondly, even though viral Blast results
273 alone are only crude taxonomy indicators, they were nevertheless consistent with our
274 phylogenetic analyses, with unclassified scaffolds mainly composed of
275 *Megamimivirinae* (43.4%), *Klosneuvirinae* (26.2%) and *Pithoviridae* (22.6%) related
276 sequence (Fig. S8B).

277 Not only *Pithoviridae* were unexpectedly diverse (Fig. 4), they were also the most
278 abundant Nucleocytoviricota according to their normalized coverage (Fig. 3).

279 *Pithoviridae/Orpheoviridae*-like families appear in all samples and particularly in R and
280 N where they are very abundant (Fig. 3). The most covered sequences in five samples
281 (C, N, R, K and Q) come from these, and from extended_*Phycodnaviridae*,
282 *Megamimivirinae* and *Klosneuvirinae* in other samples. Along with the relative
283 abundance, the diversity of Nucleocytoviricota is quite heterogeneous with the
284 exception of samples N and R from the same borehole (16 and 19m respectively)
285 having a similar distribution (Fig. 3). Most viruses are specific to the sample they were
286 recovered from, in particular the ones from surface samples (Fig. S9). Surprisingly, we
287 also found viruses that were common to samples from close locations in Central
288 Yakutia but from different ages (samples K, L, M, N, P, Q and R; Table S1). This
289 indicates that part of the viral community was maintained over time.

290 **Worldwide Nucleocytoviricota distribution**

291 The *Pithoviridae* diversity and abundance observed in two samples from the Russian
292 permafrost highlight the richness of this viral family in this environment, or alternatively,
293 a Nucleocytoviricota detection method more adapted to non-icosahedral viruses. To
294 investigate the presence of Nucleocytoviricota in other environments we applied the
295 same methodology to the Mgnify database (37), resulting in 3564 classified contigs.
296 Since biomes are unevenly present in this database, with marine samples being
297 largely predominant, we found more Nucleocytoviricota in such samples (Fig. 5A). The
298 phylogenetic distribution of the scaffolds confirmed previous results highlighting the
299 high diversity of *Mesomimivirinae* and *Phycodnaviridae* in oceanic samples (17–19).
300 On the other hand, *Pithoviridae* and *Orpheoviridae*'s diversity was much lower, with
301 corresponding sequences mostly found in engineered samples (bioreactors and
302 wastewater).



304

305 **Figure 5: Worldwide Nucleocytoviricota phylogenetic distribution**

306 (A) 3664 contigs assembled from 427 datasets of the EBI Mgnify database and (B) 804 contigs
 307 assembled from 147 terrestrial datasets of the JGI IMG/M database. Viral contigs were detected using
 308 the previously described method and placed on tree using at least one of the seven marker genes. The
 309 tree was made using Cyprinid herpesvirus 2 as outgroup. Clades containing the reference sequences
 310 were manually drawn. Colored circles at tips represent reference genomes and the outer circle shows
 311 the corresponding biome.

312 Terrestrial biomes being completely absent from the Mgnify database, we completed
 313 this analysis by using 1835 terrestrial datasets collected from the JGI IMG/M database
 314 (38). The vast majority of the samples exhibited no Nucleocytoviricota at all and few
 315 contigs over 10 kb in general (Fig. S10), probably due to the difficulty at assembling
 316 sequence data from these complex environments. Our Russian samples, along with
 317 few outliers from this database, stood out for having a high number of viral and total
 318 contigs. *Mesomimivirinae* was the most represented sub-family in this terrestrial
 319 dataset (Fig. 5B), mainly due to its presence in two deep shales samples also rich in
 320 *Phycodnaviridae*. Noteworthy, Pandoravirus-like sequences were found in sand and a

321 900 kb contig grouping next to *Pandoraviridae* and *Molliviridae* in peat permafrost
322 samples. Pimascovirales were found in a variety of soil samples.

323 Overall, *Pithoviridae* and *Orpheoviridae* were more abundant in terrestrial samples
324 than in aquatic samples (Fig. 5A and Fig. 5B). Russian permafrost samples were
325 particularly and highly significantly enriched in these viruses, followed by forest soil,
326 bioreactors and wastewater samples (Fig. S11).

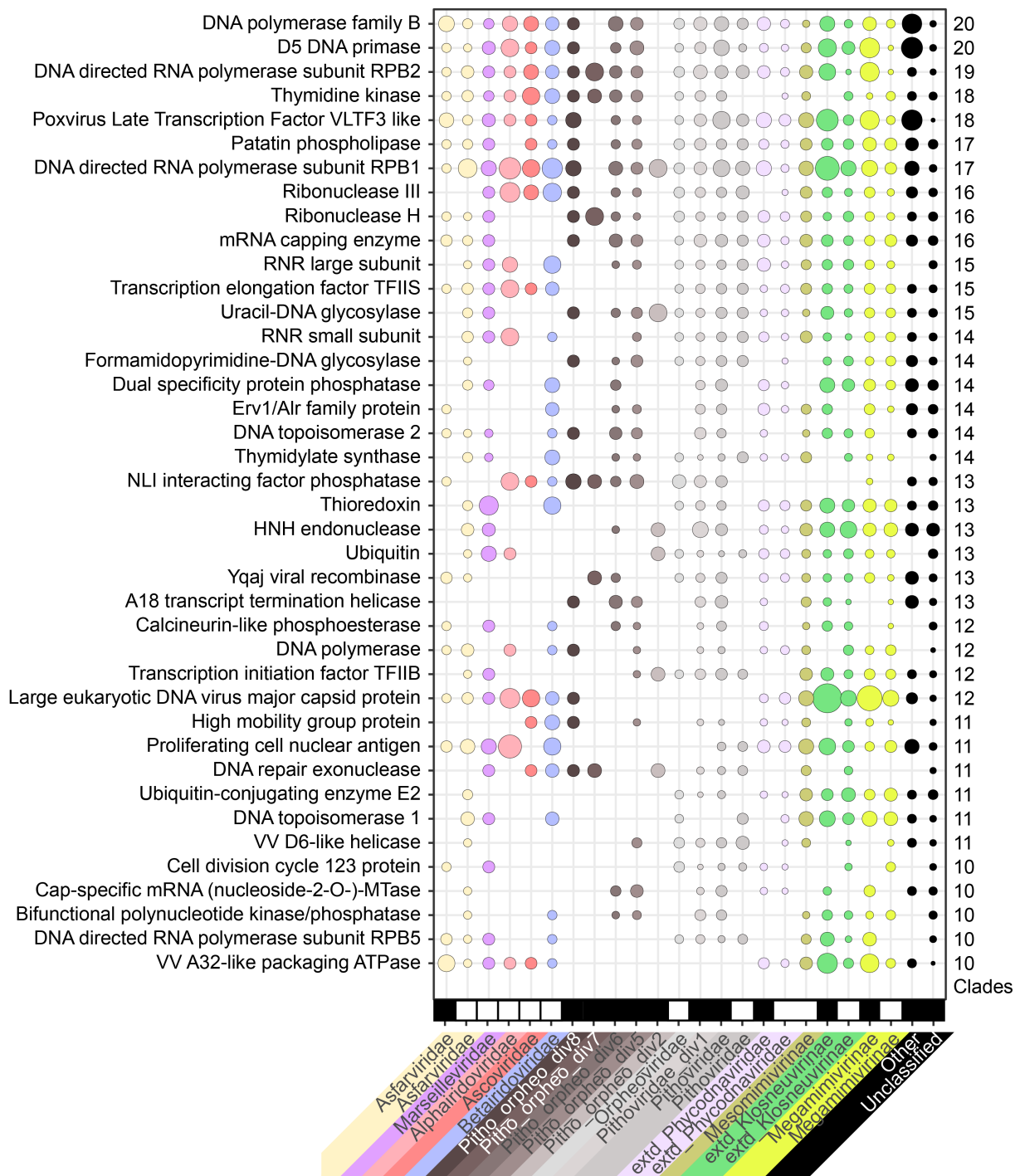
327 **Functions encoded in the permafrost Nucleocytoviricota** 328 **sequences**

329 A total of 64,648 viral ORFs over 50 amino acids were manually annotated and
330 assigned to functional categories. Most of the permafrost metagenomes predicted
331 proteins are of unknown function (81%), as expected from the high proportion of ORFs
332 of that category in reference genomes (64%, Fig. S12). With a stringent minimal ORF
333 size of 150 amino acids, the proportions are still of 76% and 56%, respectively.
334 Unspecific annotations such as Ankyrin repeat proteins, F-box proteins and FNIP
335 repeat proteins, represent 1.4, 0.2 and 0.5% of the permafrost viral proteins while they
336 represent 4.6, 1.8 and 0.6% in the reference ones. Most genes with a known function
337 are involved in DNA replication, recombination and/or repair. There are also auxiliary
338 metabolic genes that are scattered within the different viral families (Fig. S13). The
339 distribution of functional categories found in the permafrost is the same as in the
340 Nucleocytoviricota references (Fig. S12). Overall, our analysis highlights a patchwork
341 of functions encoded by these viruses (Fig. S13).

342 Looking at the most shared functions (i.e. present in most families) among the
343 reference genomes and permafrost MAGs, we identified the known core genes (Fig.
344 6). Interestingly, the highly conserved mRNA capping enzyme is absent from the

345 *Iridoviridae/Ascoviridae* clade. The patatin phospholipase, suspected to be conserved
346 among Nucleocytoviricota (39), is confirmed as a core gene, only absent from
347 *Alphairidoviridae* (Fig. 6). Its role in viral infection is still unclear but such proteins
348 participate to cell invasion in parasitic bacteria and eukaryotes (40, 41). Also,
349 according to our data, the A32-like packaging ATPase is no longer a universal
350 Nucleocytoviricota marker gene, as it is not only lacking from the reference
351 *Pithoviridae* genomes but also absent from all clades ranging from Pitho-orpheo_div8
352 to *Pithoviridae* (Fig. 6). Surprisingly, the Glutamine and Glutamine-dependent
353 asparagine synthases known to characterize *Mimiviridae* (42) were also found in a
354 permafrost *Pithoviridae*.

355



356

357 **Figure 6: Most shared functions among Nucleocytoviricota families**

358 Functions were selected among the annotations found in at least 10 clades. Metagenomic sequences
 359 are marked as black rectangles at the bottom of the plot while blank spaces correspond to reference
 360 genomes. Groups with less than 300 ORFs were marked as "Other". The size of the dots represents
 361 the normalized ORFs counts (i.e ORF counts/total number of ORFs in the group). The right-most

362 column indicates the number of distinct clades having the function. The lines are sorted according to
363 this value.

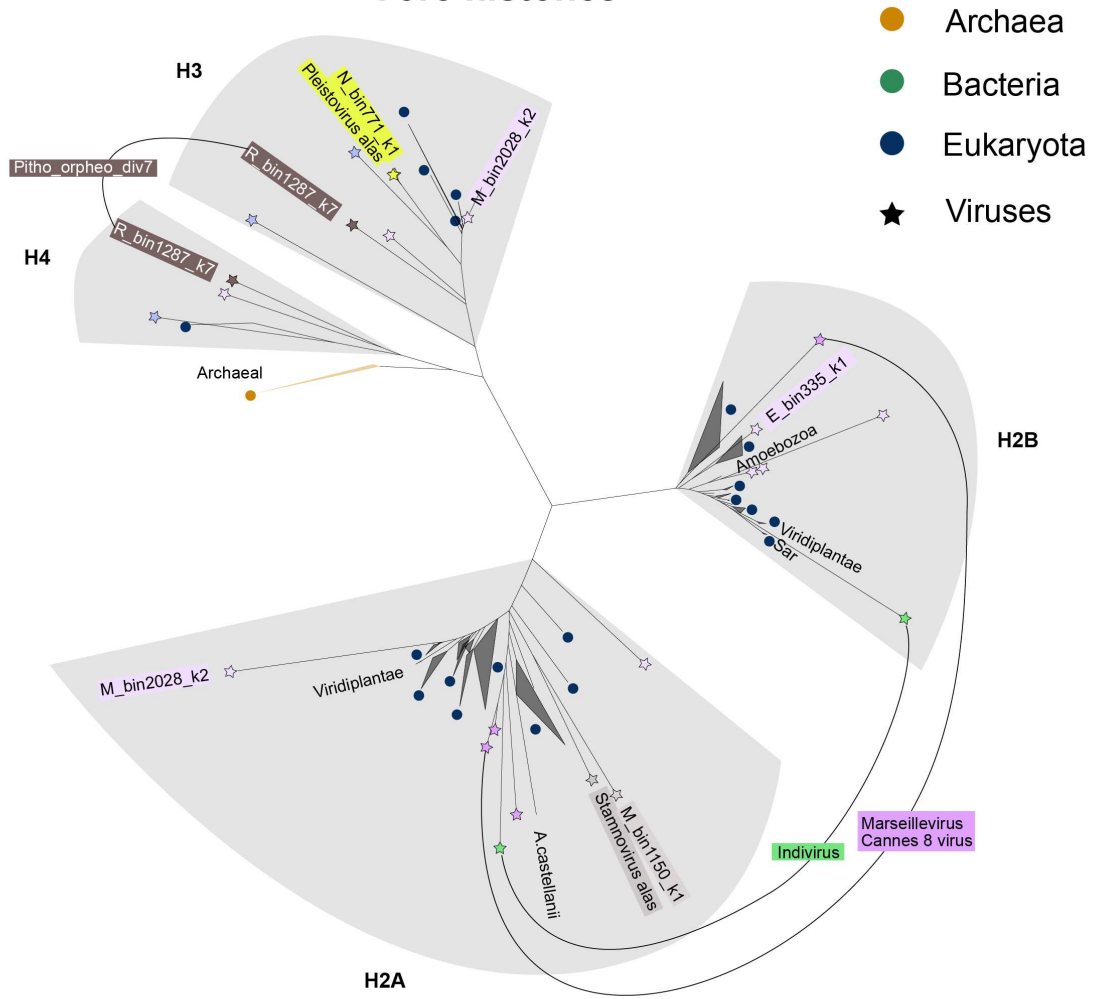
364 **DNA structure-related genes**

365 Giant DNA viruses from different clades exhibit either circular (15, 43) or linear (16,
366 28) genome structures. *Marseilleviridae* have a chromatin-like genome organized
367 around virally encoded histones (44). Here we expanded the range of viral histones,
368 identifying them in *Pithoviridae*, *Megamimivirinae*, *Pithoviridae_div1*, *Pitho-*
369 *orpheo_div7* and *extended_Phycodnaviridae* sequences. We reconstructed their
370 phylogenetic histories which turned out to involve many independent HGTs of different
371 ages (Fig. 7A). In some cases, such as for *Pleistovirus* *alaskae* and a *Pitho-orpheo_div7*
372 scaffold, the viral H3 histone is of ancient origin with a deep branching before the
373 eukaryotic ones. The latter also forms a histone doublet with fused H3-H4 domains as
374 already observed for the *Marseilleviridae* H2A-H2B and H3-H4 histones (44, 45).
375 Other viral histones appear more recently acquired from eukaryotes, like for the 400
376 kb *M_bin2028_k2* scaffold belonging to the *extended_Phycodnaviridae* (close to
377 *Pandoraviridae* and *Molliviridae*) that encodes a H2A histone from *Viridiplantae* and a
378 H3 one from an unknown eukaryote (Fig. 7A). Even more recently, the reference
379 *Pandoraviruses* (*P. salinus* and *P. dulcis*) and *Medusavirus* acquired H2B histones of
380 amoebic origin.

381

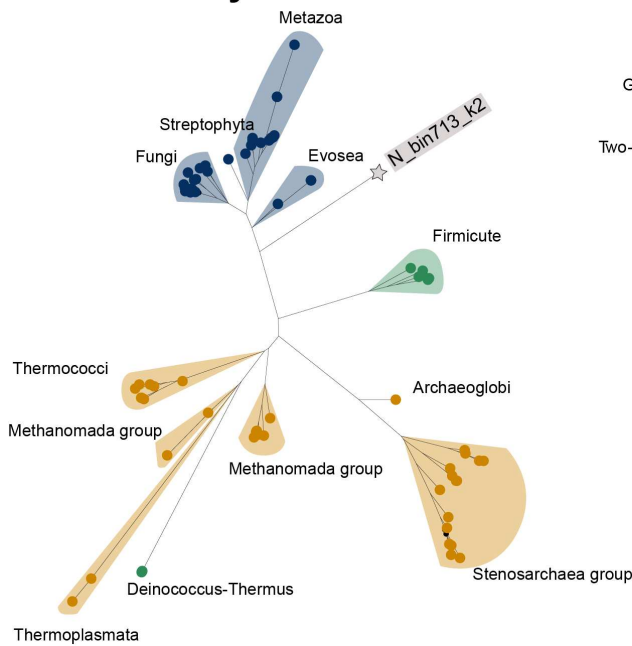
A

Core histones

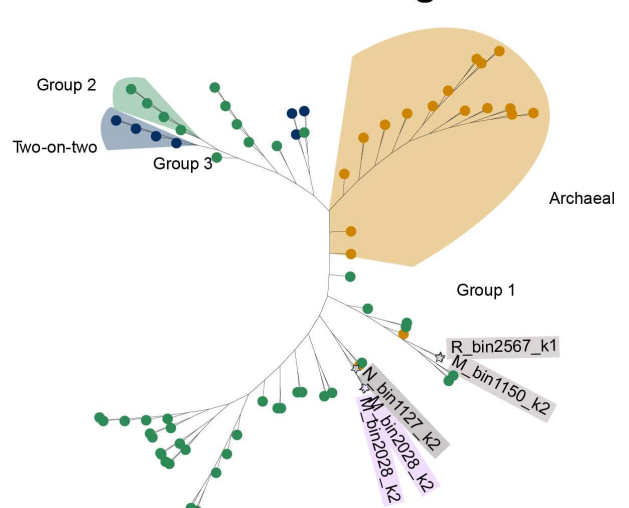


B

ATP synthase subunit F



Truncated hemoglobin



C

383 **Figure 7: Phylogeny of three functions found in the permafrost**
384 **Nucleocytoviricota.**

385 All trees were computed by Iqtree. Only ORFs of permafrost classified scaffolds were used in the trees.
386 (A) Sequences of core histones were retrieved from the HistoneDB database. The types of the viral
387 histones were confirmed through an Hmsearch alignment of reference histone on the viral ORFs. (B)
388 The ATP synthase subunit F tree was build using sequences matching the PF1990 Pfam domain as
389 well as Pithoviridae_div1 proteins with this annotated function and the best BlastP matching proteins
390 against NR. (C) The truncated hemoglobin tree was computed using the proteins from this study
391 combined with the IPR00146 Interpro domain sequences and BlastP matches of Nucleocytoviricota
392 sequences.

393 **Auxiliary metabolic genes**

394 Unexpectedly, an ATP synthase subunit F was found in a Pithoviridae_div1 sequence
395 of nearly 200kb (Fig. 7B). The viral ORF matches the PF01990 Pfam domain that
396 gathers prokaryotic ATP synthases as well as subunits of the eukaryotic vacuolar
397 ATPase. In eukaryotes, these proteins can serve many roles depending on the
398 organism and cell type but a common function is to acidify cellular compartments such
399 as lysosomes (46). The ATP synthase subunit found in Pithoviridae_div1 appears to
400 be of ancient origin (Fig. 7B). Two other subunits of the ATP synthase (one
401 Delta/Epsilon and one Beta) were also found in unclassified Nucleocytoviricota from
402 this study.

403 Other auxiliary metabolic genes found in this study include viral truncated hemoglobins
404 that are absent from reference Nucleocytoviricota. They likely come from three
405 different HGT events (Fig. 7C). A first one occurred between a prokaryote and
406 Pithoviridae_div1 viruses. A second bacteria-to-virus HGT involved a *Pithoviridae* or
407 an Extended_Phycodnaviridae that subsequently exchanged the truncated

408 hemoglobin gene. These proteins are able to bind oxygen and protect cells against
409 oxidative stress from NO or other oxygen reactive molecules (47).

410 **Translation-related genes**

411 We found 20 different types of virally-encoded aminoacyl-tRNA synthetases (aaRSs)
412 in the permafrost metagenomic scaffolds. *Klosneuvirinae* is the clade with the most
413 translation-related gene content followed by *Megamimivirinae*. For instance the
414 *Klosneuvirinae* Marosvirus alas found in this study (Fig 2 and Fig. 4) contains an
415 expanded translation-related gene repertoire (10 translation initiation factors, 4
416 translation elongation factors, a translation termination factor and as much as 11
417 different aaRSs) as well as 5 tRNAs clustered together (Fig. 2). Besides *Mimiviridae*,
418 ten different types of aaRSs were found in the Pithoviridae_div1 clade, including 7
419 different ones in Hydrivirus alas (Fig 2 and Fig. 4) that also encodes 9 tRNA, 3
420 translation initiation and elongation factors, and a translation termination factor.

421 We investigated the phylogeny of the different types of aaRSs found in our datasets
422 that revealed entangled evolutionary pathways between viruses and cellular
423 organisms (Fig. S14, S15 and S16). In most cases, the viral aaRSs came from a
424 probable HGT from Eukaryotes (tryptophan, leucine, glutamine, threonine,
425 methionine, isoleucine, arginine, aspartate, serine and phenylalanine) (Fig. S14 and
426 S16). One clear example is the exchange of a threonine-tRNA synthetase from
427 Dictyostelia (Amoebozoa) to Hydrivirus alas (Fig. S14). The exchanges concerned
428 both the mitochondrial (for instance arginine, phenylalanine) or the cytoplasmic copies
429 (Fig. S16). There were also some more rare cases of HGT from a Prokaryote to a virus
430 as for the glycine- and tyrosine-tRNA synthetases that were transferred from an
431 Archaea (Fig. S15). Genes have also passed from Bacteria to Nucleocytoviricota as

432 for the glycine-tRNA synthetase of Hydrivirus alas and the valine-tRNA synthetase of
433 a permafrost *Megamimivirinae*. For the latter, the bacterial sources were Rickettsiales
434 that are endosymbionts of amoeba (48), thus probably sharing the same host. The
435 source of the tryptophan-tRNA synthetase in Hydrivirus alas is less clear but one can
436 see that a duplication event occurred at the same locus right after the gene was
437 acquired (Fig. S14).

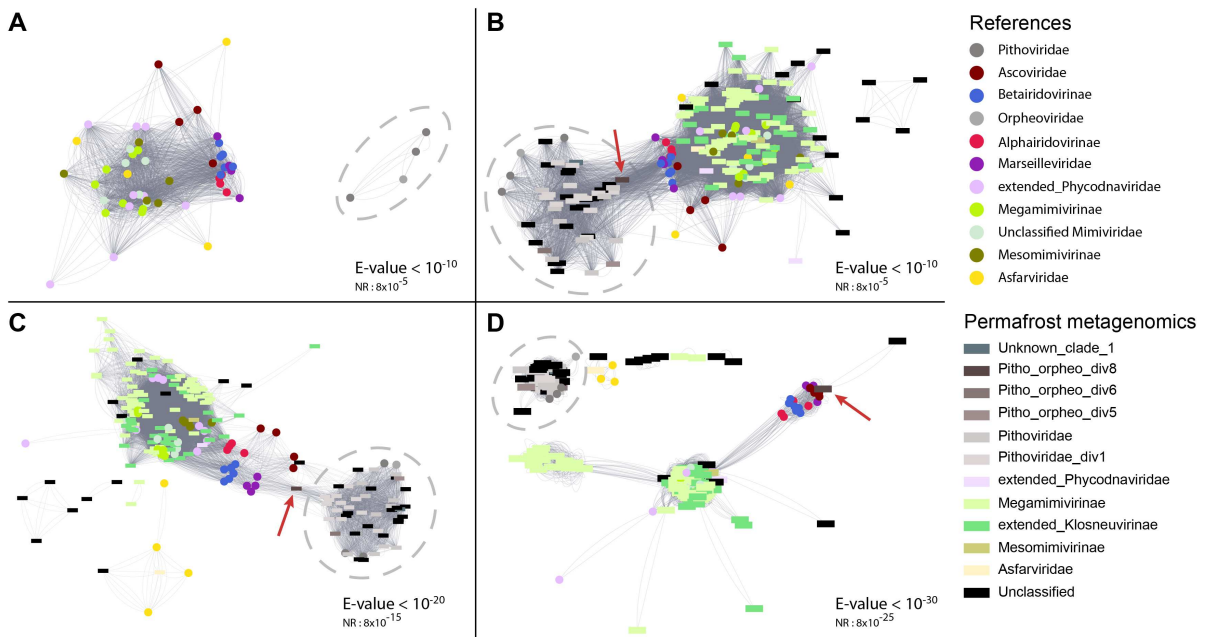
438 While the vast majority of Nucleocytoviricota genes have no identifiable homologs, the
439 ones with cellular homologs usually deeply branch in the phylogenetic trees (17, 49),
440 in accordance with their suspected ancient origin (31, 50). We found here several viral
441 aaRSs that belong to divergent families tightly clustered together within the cellular
442 homologs (Fig. S16). So not only viral aaRSs are of cellular origin, spanning all
443 domains of life, they were also probably exchanged between viruses of different
444 families.

445 **The Major Capsid Protein**

446 Little is known about the structural proteins that constitute the particle of non-
447 icosahedral giant viruses and what function their encoded MCP might have. In
448 *Ascoviridae*, the MCP is still a major protein in the virion (51), while *Pandoraviridae*
449 simply lack the gene. The MCP present in the related Mollivirus sibericum is only the
450 seventh most abundant virion protein and is thought to be involved in scaffolding
451 during virion assembly (52), as observed in *Poxviridae* (53). The annotated MCP-like
452 genes in *Pithoviridae* and *Orpheoviridae* genomes are so divergent that a Blast
453 homology search against the NR database fails to identify other Nucleocytoviricota
454 homologs, even at a low confidence E-value threshold of 10^{-2} . Furthermore, the protein
455 is not detected in Pithovirus sibericum virion proteome (15). Thus, the homology

456 between Nucleocytoviricota MCPs and the MCP-like of *Pithoviridae/Orpheoviridae* is
 457 worth being explored. We therefore constructed a BlastP network of all annotated
 458 large eukaryotic DNA viruses MCPs (Fig. 8A). As expected, the *Pithoviridae* MCPs-
 459 like are the most divergent and disconnected from the rest of the network, where
 460 icosahedral viruses and *Ascoviridae* form a strong cluster. But when adding the MCPs
 461 encoded in the permafrost metagenomics scaffolds, the *Pithoviridae* share
 462 connections to the other Nucleocytoviricota (Fig. 8B). More specifically, the MCP
 463 encoded in the Pitho-orpheo_div clades (Fig. 4) fill the gap between the *Pithoviridae*
 464 and the other Pimascovirales genes (Fig. 8B-C). Increasing the BlastP E-value
 465 stringency places the *Pithoviridae*-like MCPs apart from other Megaviricetes except
 466 for Pitho-orpheo_div8, indicating that its MCP is closer to *Marseilleviridae* than to
 467 *Pithoviridae* (Fig. 8D). From this we can conclude on the homology between the
 468 icosahedral Nucleocytoviricota and the non-icosahedral *Pithoviridae* MCPs.

469



470

471 **Figure 8: Major capsid protein network**

472 The network was made with a BlastP of all vs all annotated large eukaryotic DNA virus major capsid
473 proteins with varying E-value cutoffs and visualized in cytoscape. The edges were calculated from the
474 bitscore. Circles correspond to reference genomes and rectangles to MAGs from this study. The red
475 arrows depict the MCP identified in a *Pitho-orpheo_div8* scaffold (see Fig. 4 for phylogeny) and the
476 dashed ellipses highlight the *Pithoviridae/Orpheoviridae*-like MCPs. The NR E-value was calculated
477 based on the difference of database size.

478 **Discussion**

479 Recent large scale metagenomic data analyses strikingly revealed that
480 Nucleocytoviricota are widespread in various environments (17, 18, 23, 24). Our
481 analysis of the cryosol and permafrost samples, as well as other datasets (JGI IMG/M
482 and EBI Mgnify databases), confirms this ubiquity. Nevertheless we pointed out an
483 important heterogeneity in Nucleocytoviricota proportion across environments. Some
484 of the permafrost datasets appeared to be among the most enriched in
485 Nucleocytoviricota, reaching up to 12% of the sequenced organisms. The relative DNA
486 sequence coverage (Fig. 3) even suggests that they outnumber their hosts, in the
487 same way bacteriophages often outnumber bacteria in the ocean (54, 55). This high
488 abundance is also the result of the high Nucleocytoviricota diversity in the samples, as
489 it does not come from a single virus. Furthermore, by taking advantage of the
490 permafrost ability to preserve ancient organisms, we showed that some
491 Nucleocytoviricota strains were not only abundant but had been present in the active
492 community for a long time (Fig. S9). Considering only syngenetic permafrost samples,
493 we found Nucleocytoviricota shared in samples of up to 14,000 years difference. This
494 indicates that they are important players of this particular area of central Yakutia.

495 The Nucleocytoviricota diversity explored in this study strikingly revealed many
496 *Pithoviridae*-like sequences that are very divergent from the reference genomes and
497 constitute new clades within the Pimascovirales. This includes large genomes, in
498 particular the complete 1.6 Mb Hydrivirus alas genome. So, next to Pandoraviruses
499 (28), Orpheoviruses (32), Klosneuviruses (56) and Mimiviruses (57), this provides with
500 yet another example of a viral genome largely over 1 Mb. The nature of the
501 evolutionary forces pushing some viruses to retain or acquire so many genes remains
502 a matter of debate (58–61). Horizontal gene transfers from cellular hosts is
503 hypothesized by some authors to account for their large gene content (56, 62). We
504 indeed found examples of cellular genes gained by HGT in this study (Fig. 7 and Fig.
505 S14-S16) but this only accounts for a small proportion of their gene content, the vast
506 majority having no identifiable cellular homologs. Gene duplication, on the other hand,
507 a well-known source of functional innovation since the pioneering work of Susumu
508 Ohno (63), may contribute to the genome inflation of giant viruses (49, 64). Another
509 possible source of genetic innovation is the *de novo* gene creation from intergenic
510 regions (49, 65). The present work expanded the Nucleocytoviricota families'
511 pangenomes, in particular the *Pithoviridae*-like and *Mimiviridae*, with an overwhelming
512 proportion of ORFans. This militates for the *de novo* gene creation hypothesis that
513 remains to be further tested.

514 Despite the isolation of Mollivirus kamchatka (30) and Cedratvirus kamchatka (29) in
515 the studied samples (samples C-D-E), their genomes were not identified in our
516 assembled metagenomic data. Such a discrepancy had already been observed for
517 Pithovirus sibericum and Mollivirus sibericum, where metagenomic sequence reads
518 confirmed their presence in the samples but at a coverage too low to obtain assembled

519 contigs (16). Concerning the *Pandoraviridae*, although different strains of these
520 viruses were isolated from various geographical locations, including soil (28, 49, 65–
521 68), very few *Pandoraviridae*-like sequences were identified in our cryosol data and in
522 a vast array of environmental samples (Fig. 5), as already noticed from previous
523 metagenomic studies (18). This underlines the importance of exploring complex
524 environment communities with complementary approaches to unravel the true
525 diversity of less studied giant virus families. This includes adapted metagenomic
526 pipelines such as our attempt to reveal non-icosahedral viruses combined to direct
527 isolation.

528 The functional annotation performed in this work highlights the paucity of functions
529 strictly shared between Nucleocytoviricota. Even a central protein like the A32
530 Packaging ATPase is absent from the entire *Pithoviridae*-like clade (Fig. 8). Likewise
531 the MCP is not encoded in the *Pandoraviridae* genomes. Regarding the highly
532 divergent *Pithoviridae/Orpheoviridae* MCP-like genes, our analysis helped to reveal
533 their homology with the other Pimascovirales (Fig. 8). These genes could then either
534 come from a shared ancestor (69), as suggested by the core genes phylogenies (Fig.
535 S5), or was acquired very early in the *Pithoviridae/Orpheoviridae* evolution. We can
536 speculate that the MCP quickly started to lose or change its function, before the
537 divergence of Pitho-orpheo_div7 and the other family members (Fig. 4), perhaps with
538 a progressive change in virion morphology.

539 Besides the few functions shared by the Nucleocytoviricota, our work also highlights
540 a patchwork of functions encoded by these genomes. When looking at specific
541 functions, we detected independent cases of HGT from Eukaryotes to viruses but also
542 between viruses belonging to different families (Fig. 7 and Fig. S14-S16). This is

543 probably the testimony of coinfections, as members of the *Marseilleviridae*,
544 *Mimiviridae*, *Pithoviridae*, *Pandoraviridae* and *Molliviridae* families can infect the same
545 host. In line with this hypothesis, we recently showed that DNA methylation,
546 widespread in giant viruses, is mediated by methyltransferases and Restriction-
547 Modification systems that are frequently horizontally exchanged between viruses from
548 different families (29).

549 The functional patchwork, the gene exchanges between viruses of different families,
550 together with the very few shared genes, may challenge the monophyly of the recently
551 established Nucleocytoviricota phylum by the International Committee on Taxonomy
552 of Viruses (ICTV) (70). Except for the DNA primase of Cedratviruses, our trees of
553 seven marker genes would indeed indicate a shared ancestry of the different
554 Nucleocytoviricota families analyzed in this work (Fig. S5). However, when cellular
555 genes are integrated to the phylogenetic trees, only three of the five most shared
556 genes strictly support the monophyly of the Nucleocytoviricota (71): the viral late
557 transcription factor 3, the Holliday junction resolvase and the A32 packaging ATPase.
558 The latter has also been shown to be exchanged between *Mimiviridae* and Yaravirus,
559 an Acanthamoeba infecting virus that does not belong to the phylum (71, 72). The
560 other core genes such as the DNA polymerase is separated by several cellular clades
561 between Pokkesviricetes and Megaviricetes (73). Likewise the two largest subunits of
562 the RNA polymerase of *Asfarviridae* and *Mimiviridae* have a different history than the
563 other Nucleocytoviricota (31). These examples question the consistency of the
564 phylum.

565 The primary objective of this study was to assess the diversity of large DNA viruses in
566 permafrost. Our analyses revealed an unexpected number of new viral sub-groups

567 and clades among some of the previously established families of the
568 Nucleocytoviricota phylum, mixing an intricate patchwork of functions amidst a majority
569 of anonymous genes of unknown functions. The in-depth study of these genes will
570 allow to better understand their physiology but also to rule on the existence or not of
571 a common ancestor for its deepest branches.

572 **Materials and Methods**

573 **Data preparation**

574 Illumina sequencing reads from all samples (Table S1) were assembled into contigs
575 using Spades (v3.14) (74) and then binned using Metabat2 (v2.15) (75) with a minimal
576 contig length ≥ 1500 and bin length $\geq 10,000$. Reads corresponding to each contig
577 were retrieved and gathered from their respective bins using an in house script. The
578 read subsets were then reassembled using Spades (v3.14) in default mode or with the
579 “--meta” option. Reads were mapped on the resulting scaffolds $\geq 10\text{kb}$ using Bowtie2
580 (v2.3.4.1) (76) with the “--very-sensitive” option. Scaffold relative coverage was
581 computed as the mean scaffold coverage divided by the total sample coverage. Bins,
582 contigs and scaffolds were verified with Checkm (v1.1.2) (77) using the lineage
583 workflow.

584 **Control database preparation**

585 Reference Nucleocytoviricota were chosen following a former phylogenetic study (31).
586 The corresponding genomes were gathered from the NCBI repository. Lausannevirus,
587 Melbournevirus, Ambystoma tigrinum virus, Infectious spleen and kidney necrosis
588 virus, Invertebrate iridovirus 22, Invertebrate iridovirus 25 and Singapore grouper
589 iridovirus were removed to avoid an overrepresentation of their families. We added the

590 genomes of *Acanthamoeba castellanii* medusavirus (AP018495.1), Bodo saltans virus
591 (MF782455.1), Cedratvirus kamchatka (MN873693.1) and Tetraselmis virus 1
592 (KY322437.1). Genomes from Archaea, Eukaryota and Bacteria (Table S4) were
593 retrieved from Genbank. For each genome, non-overlapping sequences were cut with
594 an in house script following a distribution similar to our dataset to simulate
595 metagenomic contigs. Genes were then predicted by Genemark (v3.36) (74) using the
596 metagenomic model. For the Nucleocytoviricota phylogeny, core genes previously
597 identified (31) were used in addition to the ones found by Psiblast (from BLAST+
598 v2.8.1) (75). We also added *Amsacta moorei* entomopoxvirus (AF250284.1), Variola
599 virus (NC_001611.1) and Cyprinid herpesvirus 2 (MN201961.1) as outgroup.

600 **Nucleocytoviricota specific profiles databases**

601 The database constructed by (18) was completed with specific signatures of
602 *Pithoviridae* using the genomes of Cedratvirus A11 (34), Cedratvirus kamchatka (29),
603 Cedratvirus lausannensis (76), Cedratvirus zaza (77), Brazilian cedratvirus (77),
604 Pithovirus massiliensis (33), Pithovirus sibericum (15), Orpheovirus (32), all the
605 metagenomic *Pithoviridae* released from one study of Loki's Castle hydrothermal
606 vents (24), the divergent *Orpheoviridae/Pithoviridae* SRX247688.42 (17), the
607 GVMAG-S-1056828-40 (18) and other Cedratvirus/Pithovirus sequences
608 (supplementary data files). For *Pandoraviridae* we gathered sequences from
609 Pandoravirus braziliensis (78), *P. celtis* (65), *P. dulcis* (28), *P. inopinatum* (67), *P.*
610 *macleodensis* (49), *P. neocaledonia* (49), *P. pampulha* (78), *P. quercus* (65), *P.*
611 *salinus* (28), Mollivirus kamchatka (30) and *M. sibericum* (16). The ORFs were then
612 predicted using Genemark (v4.32) with the "--virus" option and ORFs ≥ 50 amino-acids
613 were kept. Orthogroups were calculated with Orthofinder (79) and HMM profiles were
614 built using the Hmmer suite (v3.2.1) (80) for each one. HMMs were further aligned to

615 the Refseq protein database (from March 2020) using the same suite. Only HMMs
616 specific to *Pithoviridae*, *Orpheoviridae*, *Pandoraviridae* or Molliviruses with E-value \leq
617 10^{-10} were kept to complete the database. To these were added Nucleocytoviricota-
618 specific VOG orthogroups (<https://vogdb.org/>).

619 **Retrieving viral sequences**

620 The Nucleocytoviricota-specific profile database was searched against the control and
621 permafrost ORFs using Hmsearch. To check for cellular signatures, all the ORFs
622 were aligned to the Refseq protein database using Diamond blastp (v0.9.31.132) with
623 the “--taxonlist 2,2759,2157” option and hits \geq 35% sequence identity were checked.
624 On the control metagenomic simulated dataset, the amount of false positives and false
625 negatives were assessed according to the cellular and viral matches for each group
626 (Nucleocytoviricota, Archaea, Bacteria, Eukaryota). We set the threshold at less than
627 1% of false eukaryotic positives. The same threshold was applied to the permafrost
628 data to retrieve viral contigs.

629 **Functional annotation**

630 All the ORFs \geq 50 amino-acids were queried against the NR database (from June
631 2020) using Blastp, the VOG database using Hmsearch, the Pfam database using
632 Interproscan (v.5.39-77) and against EggNOG (81) using the online version of
633 Emapper-1.03. For all, the E-value threshold was set to 10^{-5} . Functional annotations
634 of each predicted protein were defined manually, first based on the matching domains
635 annotations, then by considering the full sequence alignments (Blast, EggNOG and
636 VOG). EggNOG categories were also set manually for each gene. When existing, the
637 functional annotations of reference viral genomes (see control database preparation)
638 were retrieved from Genbank. Grouper iridovirus, *Heliothis virescens* ascovirus 3e and

639 Invertebrate iridescent virus 6 were manually reannotated using the same protocol as
640 for the permafrost ORFs.

641 **Contamination control**

642 The functional annotation step helped to remove non-Nucleocytoviricota scaffolds
643 based on the presence of typical viral/phage genes or with ORFs consistently
644 matching cellular organisms. The scaffolds were checked for the presence of
645 ribosomes using Barrnap (v0.9) (82). Finally, we checked for possible GEVEs (Giant
646 Endogenous viral elements) in our curated scaffolds. We made pseudo-contigs from
647 the GEVEs identified by (35) and applied our method on them. As 57% (193 out of
648 338) of the GEVEs pseudo-contigs were caught, we proceeded to check for
649 endogenization signs in our permafrost scaffolds. This was done by plotting the
650 domain of the Blastp hits as well as the VOG matches for each scaffold with the results
651 of the Viralrecall (v2.0) rolling score (36). Scaffolds with at least one region with a
652 negative Viralrecall score were visually inspected. For comparison, we also tested
653 Viralrecall with the "--contiglevel" option.

654 **Large genomes assembly verification and circularization**

655 The eight largest MAGs ($\geq 500\text{kb}$) were scrutinized for possible chimeric assemblies.
656 We used the Integrative Genome Viewer (83) to assess potential coverage drops
657 (mainly due to ambiguous bases added during scaffolding), but in each case read pairs
658 overlapped the low coverage intervals. For circularization, we created a model contig
659 concatenating both ends of the MAG, mapped the reads using Bowtie2 and checked
660 the uniformity of the coverage at the junctions.

661 **Abundance estimation and mapping**

662 Metagenomics reads were mapped to the viral scaffolds using Bowtie2 with the –very-
663 sensitive option and filtered with Samtools (-q 3 option). Reads ≤ 30 nucleotides were
664 discarded. The relative mean coverage of the scaffolds were then used as estimators
665 of the scaffold abundance in the sample. For in-between sample comparisons, reads
666 were size-filtered and then mapped to the viral scaffold with a minimum quality filter of
667 30. Then, only scaffold ≥ 10 kb in size were considered.

668 **Phylogenetic analysis**

669 For the selected marker genes, individual gene trees were built from reference
670 genomes only. Multiple alignments were performed using MAFFT (v7.407) (84),
671 removal of divergent regions with ClipKIT (85) and models estimations (86) and tree
672 inference using Iqtree (v1.6.12) (87) (options “-bb 1000” (88), “-bi 100” and “-m MFP”).
673 The best model was VT+F+R4 for the TFIS tree, LG+F+G4 for the MCP and
674 LG+F+R5 for all the other marker genes. A global tree was calculated by a partitioned
675 analysis (89) to include genomes with missing data.

676 To identify the marker genes in the permafrost data, Psiblast was used to align
677 reference marker genes to the viral ORFs (initial E-value $\leq 10^{-5}$). Next, in order to avoid
678 using a paralog of the marker genes, we defined a second stringent E-value threshold
679 the following way: E-values of all second matches for scaffolds with multiple copies
680 were sorted in ascending order, then the stringent threshold was defined based on the
681 first quartile (Table S5). Finally, only the best match per scaffold was kept for
682 phylogenetic reconstruction if it was better than the stringent threshold for this gene.

683 The 7 marker genes were aligned using PASTA (90), clipped with ClipKIT and
684 concatenated by Catsequences (91). The global tree with ultrafast bootstraps was

685 then inferred by Iqtree with options “-spp, -bb 1000” and “-bi 200 -m MFP” that
686 calculates the best model per marker gene. Tree visualization was handled using
687 Figtree (<http://tree.bio.ed.ac.uk/software/figtree/>) and the ItoI web server (92).

688 **Worldwide Nucleocytoviricota distribution**

689 The EBI Mgnify (37) protein database from March 11th 2021 was downloaded and
690 proteins from environmental or engineered biomes were extracted for further analysis.
691 We only selected proteins from contigs ≥ 10 kb in size. We also downloaded 1835
692 terrestrial assemblies from the JGI IMG/M (38) database (Table S6), of which 1502
693 exhibited at least one contig ≥ 10 kb. The ORFs were predicted using Metagenemark
694 as previously. Nucleocytoviricota sequences were extracted from both databases as
695 described above (see Retrieving viral sequences). The same method than previously
696 described (see phylogenetic analysis) was applied to search for marker genes for
697 phylogeny. Reference and metagenomic marker genes were aligned using MAFFT
698 with the “—auto” option. *Amsacta moorei* entomopoxvirus, Variola virus and Cyprinid
699 herpesvirus 2 were included in the analysis. The alignments were clipped with ClipKIT
700 and concatenated for a partitioned analysis. Empirical models for each partition were
701 inferred by ModelEstimator (93). Finally, the trees were computed using Iqtree (with -
702 bb 1000 -bi 200).

703 **Phylogenetic analyses of selected functions**

704 For each function, a dataset of proteins was built using a combination of
705 Nucleocytoviricota ORFs, corresponding Blast matched proteins from the NR
706 database and reference proteins from specific databases. The latter includes Uniprot
707 reviewed proteins of domains PF01990 (ATP synthase subunit F), IPR001412 (class
708 I aminoacyl-tRNA synthetases), IPR006195 (class II aminoacyl-tRNA synthetases)

709 and IPR001486 (truncated hemoglobin). The reference core histone proteins were
710 also retrieved from the HistoneDB 2.0 database (94) in addition to reviewed archaeal
711 core histones from Uniprot (clustered using CDhit (95)). For all the functions, the
712 multiple alignments were performed using PASTA (90) or MAFFT (84) and trimmed
713 with ClipKit (85). The tree was then computed by Iqtree (87) with options -bb 5000 -bi
714 200 -m TEST.

715 **Major Capsid Protein network**

716 All proteins annotated as “Large eukaryotic DNA virus major capsid protein” or
717 “Divergent major capsid protein” were gathered with the reference MCPs and aligned
718 against each other with BlastP (E-value $\leq 10^{-5}$). The network was created using
719 Cytoscape (v3.8.2) (96). The edge-weighted Spring Embedded layout was used and
720 the bitscores were chosen as weights in the heuristic mode. The E-value threshold
721 was progressively decreased to 10^{-30} and changes in the network were observed along
722 the way.

723 **Acknowledgements**

724 We would like to thank Alexander Morawitz for collecting the Kamchatka soil samples.
725 We thank the PACA Bioinfo platform for computing support. We also thank Eugène
726 Christo-Foroux for processing the sample and performing DNA extraction, Dr. Jens
727 Strauss, Dr. Guido Grosse, Prof. N. Fedorov for providing the Ukechi permafrost
728 samples and Dr Karine Labadie for supervising the sequencing on the Genoscope
729 platform.

730 **References**

- 731 1. Obu J. 2021. How Much of the Earth's Surface is Underlain by Permafrost? *Journal*
732 *of Geophysical Research: Earth Surface* 126:e2021JF006123.
- 733 2. Mackelprang R, Waldrop MP, DeAngelis KM, David MM, Chavarria KL, Blazewicz SJ,
734 Rubin EM, Jansson JK. 2011. Metagenomic analysis of a permafrost microbial
735 community reveals a rapid response to thaw. *Nature* 480:368–371.
- 736 3. Burkert A, Douglas TA, Waldrop MP, Mackelprang R. 2019. Changes in the Active,
737 Dead, and Dormant Microbial Community Structure across a Pleistocene
738 Permafrost Chronosequence. *Appl Environ Microbiol* 85:e02646-18.
- 739 4. Vishnivetskaya T, Kathariou S, McGrath J, Gilichinsky D, Tiedje JM. 2000. Low-
740 temperature recovery strategies for the isolation of bacteria from ancient
741 permafrost sediments. *Extremophiles* 4:165–173.
- 742 5. Hinsa-Leasure SM, Bhavaraju L, Rodrigues JLM, Bakermans C, Gilichinsky DA,
743 Tiedje JM. 2010. Characterization of a bacterial community from a Northeast
744 Siberian seacoast permafrost sample. *FEMS Microbiology Ecology* 74:103–113.
- 745 6. Liang R, Lau M, Vishnivetskaya T, Lloyd KG, Wang W, Wiggins J, Miller J, Pfiffner S,
746 Rivkina EM, Onstott TC. 2019. Predominance of Anaerobic, Spore-Forming Bacteria
747 in Metabolically Active Microbial Communities from Ancient Siberian Permafrost.
748 *Appl Environ Microbiol* 85:e00560-19.

- 749 7. Zhong Z-P, Tian F, Roux S, Gazitúa MC, Solonenko NE, Li Y-F, Davis ME, Van Etten
750 JL, Mosley-Thompson E, Rich VI, Sullivan MB, Thompson LG. 2021. Glacier ice
751 archives nearly 15,000-year-old microbes and phages. *Microbiome* 9:160.
- 752 8. Turchetti B, Buzzini P, Goretti M, Branda E, Diolaiuti G, D'Agata C, Smiraglia C,
753 Vaughan-Martini A. 2008. Psychrophilic yeasts in glacial environments of Alpine
754 glaciers. *FEMS Microbiol Ecol* 63:73–83.
- 755 9. Malavin S, Shmakova L, Claverie J-M, Rivkina E. 2020. Frozen Zoo: a collection of
756 permafrost samples containing viable protists and their viruses. *Biodivers Data J*
757 8:e51586.
- 758 10. Vishnivetskaya TA, Spirina EV, Shatilovich AV, Erokhina LG, Vorobyova EA,
759 Gilichinsky DA. 2003. The resistance of viable permafrost algae to simulated
760 environmental stresses: implications for astrobiology. *International Journal of*
761 *Astrobiology* 2:171–177.
- 762 11. Yashina S, Gubin S, Maksimovich S, Yashina A, Gakhova E, Gilichinsky D. 2012.
763 Regeneration of whole fertile plants from 30,000-y-old fruit tissue buried in
764 Siberian permafrost. *PNAS* 109:4008–4013.
- 765 12. Shmakova L, Malavin S, Iakovenko N, Vishnivetskaya T, Shain D, Plewka M, Rivkina
766 E. 2021. A living bdelloid rotifer from 24,000-year-old Arctic permafrost. *Current*
767 *Biology* 31:R712–R713.
- 768 13. Bellas C, Anesio A, Barker G. 2015. Analysis of virus genomes from glacial
769 environments reveals novel virus groups with unusual host interactions. *Frontiers*
770 *in Microbiology* 6:656.

- 771 14. Rigou S, Christo-Foroux E, Santini S, Goncharov A, Strauss J, Grosse G, Fedorov AN,
772 Labadie K, Abergel C, Claverie J-M. 2021. High prevalence and diversity of beta-
773 lactamase-encoding bacteria in cryosoils and ancient permafrost. bioRxiv
774 2021.03.17.435775.
- 775 15. Legendre M, Bartoli J, Shmakova L, Jeudy S, Labadie K, Adrait A, Lescot M, Poirot O,
776 Bertaux L, Bruley C, Coute Y, Rivkina E, Abergel C, Claverie J-M. 2014. Thirty-
777 thousand-year-old distant relative of giant icosahedral DNA viruses with a
778 pandoravirus morphology. Proc Natl Acad Sci U S A 111:4274–4279.
- 779 16. Legendre M, Lartigue A, Bertaux L, Jeudy S, Bartoli J, Lescot M, Alempic J-M, Ramus
780 C, Bruley C, Labadie K, Shmakova L, Rivkina E, Couté Y, Abergel C, Claverie J-M.
781 2015. In-depth study of Mollivirus sibericum, a new 30,000-y-old giant virus
782 infecting Acanthamoeba. PNAS 112:E5327–E5335.
- 783 17. Moniruzzaman M, Martinez-Gutierrez CA, Weinheimer AR, Aylward FO. 2020.
784 Dynamic genome evolution and complex virocell metabolism of globally-
785 distributed giant viruses. 1. Nature Communications 11:1–11.
- 786 18. Schulz F, Roux S, Paez-Espino D, Jungbluth S, Walsh DA, Denef VJ, McMahon KD,
787 Konstantinidis KT, Eloë-Fadrosch EA, Kyrpides NC, Woyke T. 2020. Giant virus
788 diversity and host interactions through global metagenomics. Nature 578:432–436.
- 789 19. Endo H, Blanc-Mathieu R, Li Y, Salazar G, Henry N, Labadie K, de Vargas C, Sullivan
790 MB, Bowler C, Wincker P, Karp-Boss L, Sunagawa S, Ogata H. 2020. Biogeography of
791 marine giant viruses reveals their interplay with eukaryotes and ecological
792 functions. 12. Nature Ecology & Evolution 4:1639–1649.

- 793 20. Needham DM, Yoshizawa S, Hosaka T, Poirier C, Choi CJ, Hehenberger E, Irwin NAT,
794 Wilken S, Yung C-M, Bachy C, Kurihara R, Nakajima Y, Kojima K, Kimura-Someya T,
795 Leonard G, Malmstrom RR, Mende DR, Olson DK, Sudo Y, Sudek S, Richards TA,
796 DeLong EF, Keeling PJ, Santoro AE, Shirouzu M, Iwasaki W, Worden AZ. 2019. A
797 distinct lineage of giant viruses brings a rhodopsin photosystem to unicellular
798 marine predators. *Proc Natl Acad Sci USA* 116:20574–20583.
- 799 21. Gallot-Lavallée L, Blanc G, Claverie J-M. 2017. Comparative Genomics of
800 Chrysochromulina Ericina Virus and Other Microalga-Infecting Large DNA Viruses
801 Highlights Their Intricate Evolutionary Relationship with the Established
802 Mimiviridae Family. *J Virol* 91.
- 803 22. Ha AD, Moniruzzaman M, Aylward FO. 2021. High Transcriptional Activity and
804 Diverse Functional Repertoires of Hundreds of Giant Viruses in a Coastal Marine
805 System. *mSystems* 6:e0029321.
- 806 23. Schulz F, Alteio L, Goudeau D, Ryan EM, Yu FB, Malmstrom RR, Blanchard J, Woyke
807 T. 2018. Hidden diversity of soil giant viruses. *Nat Commun* 9:1–9.
- 808 24. Bäckström D, Yutin N, Jørgensen SL, Dharamshi J, Homa F, Zaremba-Niedwiedzka
809 K, Spang A, Wolf YI, Koonin EV, Ettema TJG. 2019. Virus Genomes from Deep Sea
810 Sediments Expand the Ocean Megavirome and Support Independent Origins of
811 Viral Gigantism. *mBio* 10.
- 812 25. Alteio LV, Schulz F, Seshadri R, Varghese N, Rodriguez-Reillo W, Ryan E, Goudeau
813 D, Eichorst SA, Malmstrom RR, Bowers RM, Katz LA, Blanchard JL, Woyke T. 2020.

- 814 Complementary Metagenomic Approaches Improve Reconstruction of Microbial
815 Diversity in a Forest Soil. *mSystems* 5:e00768-19.
- 816 26. Sczyrba A, Hofmann P, Belmann P, Koslicki D, Janssen S, Dröge J, Gregor I, Majda S,
817 Fiedler J, Dahms E, Bremges A, Fritz A, Garrido-Oter R, Jørgensen TS, Shapiro N,
818 Blood PD, Gurevich A, Bai Y, Turaev D, DeMaere MZ, Chikhi R, Nagarajan N, Quince
819 C, Meyer F, Balvočiūtė M, Hansen LH, Sørensen SJ, Chia BKH, Denis B, Froula JL,
820 Wang Z, Egan R, Don Kang D, Cook JJ, Deltel C, Beckstette M, Lemaitre C, Peterlongo
821 P, Rizk G, Lavenier D, Wu Y-W, Singer SW, Jain C, Strous M, Klingenberg H, Meinicke
822 P, Barton MD, Lingner T, Lin H-H, Liao Y-C, Silva GGZ, Cuevas DA, Edwards RA, Saha
823 S, Piro VC, Renard BY, Pop M, Klenk H-P, Göker M, Kyrpides NC, Woyke T, Vorholt
824 JA, Schulze-Lefert P, Rubin EM, Darling AE, Rattei T, McHardy AC. 2017. Critical
825 Assessment of Metagenome Interpretation-a benchmark of metagenomics
826 software. *Nat Methods* 14:1063–1071.
- 827 27. Koonin EV, Yutin N. 2019. Evolution of the Large Nucleocytoplasmic DNA Viruses
828 of Eukaryotes and Convergent Origins of Viral Gigantism. *Adv Virus Res* 103:167–
829 202.
- 830 28. Philippe N, Legendre M, Doutre G, Couté Y, Poirot O, Lescot M, Arslan D, Seltzer V,
831 Bertaux L, Bruley C, Garin J, Claverie J-M, Abergel C. 2013. Pandoraviruses: amoeba
832 viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. *Science*
833 341:281–286.
- 834 29. Jeudy S, Rigou S, Alempic J-M, Claverie J-M, Abergel C, Legendre M. 2020. The DNA
835 methylation landscape of giant viruses. *Nature Communications* 11:2657.

- 836 30. Christo-Foroux E, Alempic J-M, Lartigue A, Santini S, Labadie K, Legendre M,
837 Abergel C, Claverie J-M. 2020. Characterization of Mollivirus kamchatka, the First
838 Modern Representative of the Proposed Molliviridae Family of Giant Viruses. *J*
839 *Virol* 94.
- 840 31. Guglielmini J, Woo AC, Krupovic M, Forterre P, Gaia M. 2019. Diversification of
841 giant and large eukaryotic dsDNA viruses predated the origin of modern
842 eukaryotes. *Proc Natl Acad Sci USA* 116:19585–19592.
- 843 32. Andreani J, Khalil JYB, Baptiste E, Hasni I, Michelle C, Raoult D, Levasseur A, La
844 Scola B. 2017. Orpheovirus IHUMI-LCC2: A New Virus among the Giant Viruses.
845 *Front Microbiol* 8:2643.
- 846 33. Levasseur A, Andreani J, Delerce J, Bou Khalil J, Robert C, La Scola B, Raoult D. 2016.
847 Comparison of a Modern and Fossil Pithovirus Reveals Its Genetic Conservation
848 and Evolution. *Genome Biol Evol* 8:2333–2339.
- 849 34. Andreani J, Aherfi S, Bou Khalil JY, Di Pinto F, Bitam I, Raoult D, Colson P, La Scola B.
850 2016. Cedratvirus, a Double-Cork Structured Giant Virus, is a Distant Relative of
851 Pithoviruses. *Viruses* 8.
- 852 35. Moniruzzaman M, Weinheimer AR, Martinez-Gutierrez CA, Aylward FO. 2020.
853 Widespread endogenization of giant viruses shapes genomes of green algae. *Nature*
854 588:141–145.
- 855 36. Aylward FO, Moniruzzaman M. 2021. ViralRecall—A Flexible Command-Line Tool
856 for the Detection of Giant Virus Signatures in ‘Omic Data. 2. *Viruses* 13:150.

- 857 37. Mitchell AL, Almeida A, Beracochea M, Boland M, Burgin J, Cochrane G, Crusoe MR,
858 Kale V, Potter SC, Richardson LJ, Sakharova E, Scheremetjew M, Korobeynikov A,
859 Shlemov A, Kunyavskaya O, Lapidus A, Finn RD. 2020. MGnify: the microbiome
860 analysis resource in 2020. *Nucleic Acids Research* 48:D570–D578.
- 861 38. Chen I-MA, Chu K, Palaniappan K, Ratner A, Huang J, Huntemann M, Hajek P, Ritter
862 S, Varghese N, Seshadri R, Roux S, Woyke T, Eloë-Fadrosch EA, Ivanova NN, Kyrpides
863 NC. 2021. The IMG/M data management and analysis system v.6.0: new tools and
864 advanced capabilities. *Nucleic Acids Research* 49:D751–D763.
- 865 39. Yutin N, Wolf YI, Raoult D, Koonin EV. 2009. Eukaryotic large nucleo-cytoplasmic
866 DNA viruses: clusters of orthologous genes and reconstruction of viral genome
867 evolution. *Virology* 6:223.
- 868 40. Aurass P, Banerji S, Flieger A. 2014. Loss of a Patatin-Like Phospholipase A Causes
869 Reduced Infectivity of *Legionella Pneumophila* in *Amoeba* and *Machrophage*
870 *Infection Models* 199–202.
- 871 41. Wilson SK, Heckendorn J, Martorelli Di Genova B, Koch LL, Rooney PJ, Morrissette
872 N, Lebrun M, Knoll LJ. 2020. A *Toxoplasma gondii* patatin-like phospholipase
873 contributes to host cell invasion. *PLoS Pathog* 16:e1008650.
- 874 42. Mozar M, Claverie J-M. 2014. Expanding the Mimiviridae family using asparagine
875 synthase as a sequence bait. *Virology* 466–467:112–122.
- 876 43. Blanca L, Christo-Foroux E, Rigou S, Legendre M. 2020. Comparative Analysis of the
877 Circular and Highly Asymmetrical Marseilleviridae Genomes. 11. *Viruses* 12:1270.

- 878 44. Liu Y, Bisio H, Toner CM, Jeudy S, Philippe N, Zhou K, Bowerman S, White A,
879 Edwards G, Abergel C, Luger K. 2021. Virus-encoded histone doublets are essential
880 and form nucleosome-like structures. *Cell* 184:4237-4250.e19.
- 881 45. Thomas V, Bertelli C, Collyn F, Casson N, Telenti A, Goesmann A, Croxatto A, Greub
882 G. 2011. Lausannevirus, a giant amoebal virus encoding histone doublets. *Environ*
883 *Microbiol* 13:1454–1466.
- 884 46. Toei M, Saum R, Forgac M. 2010. Regulation and Isoform Function of the V-
885 ATPases. *Biochemistry* 49:4715–4723.
- 886 47. Lama A, Pawaria S, Dikshit KL. 2006. Oxygen binding and NO scavenging properties
887 of truncated hemoglobin, HbN, of *Mycobacterium smegmatis*. *FEBS Letters*
888 580:4031–4041.
- 889 48. Schulz F, Martijn J, Wascher F, Lagkouvardos I, Kostanjšek R, Ettema TJG, Horn M.
890 2016. A Rickettsiales symbiont of amoebae with ancient features. *Environmental*
891 *Microbiology* 18:2326–2342.
- 892 49. Legendre M, Fabre E, Poirot O, Jeudy S, Lartigue A, Alempic J-M, Beucher L, Philippe
893 N, Bertaux L, Christo-Foroux E, Labadie K, Couté Y, Abergel C, Claverie J-M. 2018.
894 Diversity and evolution of the emerging Pandoraviridae family. *Nat Commun*
895 9:2285.
- 896 50. Nasir A, Caetano-Anollés G. 2015. A phylogenomic data-driven exploration of viral
897 origins and evolution. *Science Advances* <https://doi.org/10.1126/sciadv.1500527>.

- 898 51. Chen Z-S, Cheng X-W, Wang X, Hou D-H, Huang G-H. 2019. Proteomic analysis of the
899 *Heliothis virescens* ascovirus 3i (HvAV-3i) virion. *J Gen Virol* 100:301–307.
- 900 52. Quemin ER, Corroyer-Dulmont S, Baskaran A, Penard E, Gazi AD, Christo-Foroux E,
901 Walther P, Abergel C, Krijnse-Locker J. 2019. Complex Membrane Remodeling
902 during Virion Assembly of the 30,000-Year-Old Mollivirus Sibericum. *J Virol*
903 93:e00388-19.
- 904 53. Hyun J-K, Accurso C, Hijnen M, Schult P, Pettikiriarachchi A, Mitra AK, Coulibaly F.
905 2011. Membrane remodeling by the double-barrel scaffolding protein of poxvirus.
906 *PLoS Pathog* 7:e1002239.
- 907 54. Bergh O, Børsheim KY, Bratbak G, Heldal M. 1989. High abundance of viruses found
908 in aquatic environments. *Nature* 340:467–468.
- 909 55. Cochlan WP, Wikner J, Steward GF, Smith DC, Azam F. 1993. Spatial distribution of
910 viruses, bacteria and chlorophyll a in neritic, oceanic and estuarine environments.
911 *Marine Ecology Progress Series* 92:77–87.
- 912 56. Schulz F, Yutin N, Ivanova NN, Ortega DR, Lee TK, Vierheilig J, Daims H, Horn M,
913 Wagner M, Jensen GJ, Kyrpides NC, Koonin EV, Woyke T. 2017. Giant viruses with
914 an expanded complement of translation system components. *Science* 356:82–85.
- 915 57. Raoult D, Audic S, Robert C, Abergel C, Renesto P, Ogata H, La Scola B, Suzan M,
916 Claverie J-M. 2004. The 1.2-megabase genome sequence of Mimivirus. *Science*
917 306:1344–1350.

- 918 58. Claverie J-M, Abergel C. 2016. Giant viruses: The difficult breaking of multiple
919 epistemological barriers. *Stud Hist Philos Biol Biomed Sci* 59:89–99.
- 920 59. Filée J. 2015. Genomic comparison of closely related Giant Viruses supports an
921 accordion-like model of evolution. *Front Microbiol* 6:593.
- 922 60. Krupovic M, Koonin EV. 2015. Polintons: a hotbed of eukaryotic virus, transposon
923 and plasmid evolution. *Nat Rev Microbiol* 13:105–115.
- 924 61. Yutin N, Wolf YI, Koonin EV. 2014. Origin of giant viruses from smaller DNA viruses
925 not from a fourth domain of cellular life. *Virology* 466–467:38–52.
- 926 62. Moreira D, Brochier-Armanet C. 2008. Giant viruses, giant chimeras: the multiple
927 evolutionary histories of Mimivirus genes. *BMC Evol Biol* 8:12.
- 928 63. Ohno S. 1970. The Creation of a New Gene from a Redundant Duplicate of an Old
929 Gene, p. 71–82. *In* Ohno, S (ed.), *Evolution by Gene Duplication*. Springer, Berlin,
930 Heidelberg.
- 931 64. Suhre K. 2005. Gene and genome duplication in *Acanthamoeba polyphaga*
932 *Mimivirus*. *J Virol* 79:14095–14101.
- 933 65. Legendre M, Alempic J-M, Philippe N, Lartigue A, Jeudy S, Poirot O, Ta NT, Nin S,
934 Couté Y, Abergel C, Claverie J-M. 2019. Pandoravirus *Celtis* Illustrates the
935 Microevolution Processes at Work in the Giant Pandoraviridae Genomes. *Front*
936 *Microbiol* 10:430.

- 937 66. Hosokawa N, Takahashi H, Aoki K, Takemura M. 2021. Draft Genome Sequence of
938 Pandoravirus japonicus Isolated from the Sabaishi River, Niigata, Japan. *Microbiol*
939 *Resour Announc* 10:e00365-21.
- 940 67. Scheid P. 2016. A strange endocytobiont revealed as largest virus. *Curr Opin*
941 *Microbiol* 31:58–62.
- 942 68. Dornas FP, Khalil JYB, Pagnier I, Raoult D, Abrahão J, La Scola B. 2015. Isolation of
943 new Brazilian giant viruses from environmental samples using a panel of protozoa.
944 *Front Microbiol* 6:1086.
- 945 69. Koonin EV, Yutin N. 2010. Origin and evolution of eukaryotic large nucleo-
946 cytoplasmic DNA viruses. *Intervirology* 53:284–292.
- 947 70. Koonin EV, Dolja VV, Krupovic M, Varsani A, Wolf YI, Yutin N, Zerbini FM, Kuhn JH.
948 2020. Global Organization and Proposed Megataxonomy of the Virus World.
949 *Microbiol Mol Biol Rev* 84:e00061-19.
- 950 71. Mönttinen HAM, Bicep C, Williams TA, Hirt RP. 2021. The genomes of
951 nucleocytoplasmic large DNA viruses: viral evolution writ large. *Microb Genom* 7.
- 952 72. Boratto PVM, Oliveira GP, Machado TB, Andrade ACSP, Baudoin J-P, Klose T, Schulz
953 F, Azza S, Decloquement P, Chabrière E, Colson P, Levasseur A, La Scola B, Abrahão
954 JS. 2020. Yaravirus: A novel 80-nm virus infecting *Acanthamoeba castellanii*. *Proc*
955 *Natl Acad Sci U S A* 117:16579–16586.

- 956 73. Kazlauskas D, Krupovic M, Guglielmini J, Forterre P, Venclovas Č. 2020. Diversity
957 and evolution of B-family DNA polymerases. *Nucleic Acids Research* 48:10142–
958 10156.
- 959 74. Besemer J, Lomsadze A, Borodovsky M. 2001. GeneMarkS: a self-training method
960 for prediction of gene starts in microbial genomes. Implications for finding
961 sequence motifs in regulatory regions. *Nucleic Acids Res* 29:2607–2618.
- 962 75. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997.
963 Gapped BLAST and PSI-BLAST: a new generation of protein database search
964 programs. *Nucleic Acids Research* 25:3389–3402.
- 965 76. Bertelli C, Mueller L, Thomas V, Pillonel T, Jacquier N, Greub G. 2017. Cedratvirus
966 lausannensis - digging into Pithoviridae diversity. *Environ Microbiol* 19:4022–
967 4034.
- 968 77. Rodrigues RAL, Andreani J, Andrade AC dos SP, Machado TB, Abdi S, Levasseur A,
969 Abrahão JS, La Scola B. 2018. Morphologic and Genomic Analyses of New Isolates
970 Reveal a Second Lineage of Cedratviruses. *J Virol* 92:e00372-18.
- 971 78. Aherfi S, Andreani J, Baptiste E, Oumessoum A, Dornas FP, Andrade ACDS, P,
972 Chabriere E, Abrahao J, Levasseur A, Raoult D, La Scola B, Colson P. 2018. A Large
973 Open Pangenome and a Small Core Genome for Giant Pandoraviruses. *Front*
974 *Microbiol* 9:1486.
- 975 79. Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for
976 comparative genomics. *Genome Biology* 20:238.

- 977 80. Eddy SR. 2009. A new generation of homology search tools based on probabilistic
978 inference, p. 205–211. *In* Genome Informatics 2009. PUBLISHED BY IMPERIAL
979 COLLEGE PRESS AND DISTRIBUTED BY WORLD SCIENTIFIC PUBLISHING CO.
- 980 81. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H,
981 Mende DR, Letunic I, Rattei T, Jensen LJ, von Mering C, Bork P. 2019. eggNOG 5.0: a
982 hierarchical, functionally and phylogenetically annotated orthology resource based
983 on 5090 organisms and 2502 viruses. *Nucleic Acids Research* 47:D309–D314.
- 984 82. GitHub - tseemann/barrnap: Bacterial ribosomal RNA predictor. GitHub.
985 <https://github.com/tseemann/barrnap>. Retrieved 19 August 2021.
- 986 83. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G,
987 Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol* 29:24–26.
- 988 84. Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software
989 Version 7: Improvements in Performance and Usability. *Molecular Biology and*
990 *Evolution* 30:772–780.
- 991 85. Steenwyk JL, Iii TJB, Li Y, Shen X-X, Rokas A. 2020. ClipKIT: A multiple sequence
992 alignment trimming software for accurate phylogenomic inference. *PLOS Biology*
993 18:e3001007.
- 994 86. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermini LS. 2017.
995 ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature*
996 *Methods* 14:587–589.

- 997 87. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: A Fast and
998 Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies.
999 *Molecular Biology and Evolution* 32:268–274.
- 1000 88. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2:
1001 Improving the Ultrafast Bootstrap Approximation. *Molecular Biology and Evolution*
1002 35:518–522.
- 1003 89. Chernomor O, von Haeseler A, Minh BQ. 2016. Terrace Aware Data Structure for
1004 Phylogenomic Inference from Supermatrices. *Systematic Biology* 65:997–1008.
- 1005 90. Mirarab S, Nguyen N, Warnow T. 2014. PASTA: Ultra-Large Multiple Sequence
1006 Alignment, p. 177–191. *In* Sharan, R (ed.), *Research in Computational Molecular*
1007 *Biology*. Springer International Publishing, Cham.
- 1008 91. Chris Creevey, Nathan Weeks. 2021. ChrisCreevey/catsequences: Version 1.3.
1009 Zenodo. <https://zenodo.org/record/4409153>. Retrieved 18 August 2021.
- 1010 92. Letunic I, Bork P. 2007. Interactive Tree Of Life (iTOL): an online tool for
1011 phylogenetic tree display and annotation. *Bioinformatics* 23:127–128.
- 1012 93. Arvestad L. 2006. Efficient methods for estimating amino acid replacement rates. *J*
1013 *Mol Evol* 62:663–673.
- 1014 94. Draizen EJ, Shaytan AK, Mariño-Ramírez L, Talbert PB, Landsman D, Panchenko AR.
1015 2016. HistoneDB 2.0: a histone database with variants--an integrated resource to
1016 explore histones and their variants. *Database (Oxford)* 2016:baw014.

- 1017 95. Li W, Jaroszewski L, Godzik A. 2001. Clustering of highly homologous sequences to
1018 reduce the size of large protein databases. *Bioinformatics* 17:282–283.
- 1019 96. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski
1020 B, Ideker T. 2003. Cytoscape: a software environment for integrated models of
1021 biomolecular interaction networks. *Genome Res* 13:2498–2504.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [NCLDVmetaGpaperV13supplementarytables.xlsx](#)
- [NCLDVmetaGpaperV13supplementaryfigures.pdf](#)