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

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 Rigou[1], Sébastien Santini[1], Chantal Abergel[1], Jean-Michel Claverie[1], Matthieu Legendre[1,*]

[1]Aix–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

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

Introduction

The permafrost, soil remaining continuously frozen for at least 2 years, covers 15% of the Northern hemisphere (1) and gathers complex communities of living organisms and variable soil types. The microbial community of the surface cryosol is in some cases subject to freezing and thawing every year (2) whereas communities from deeper layers are trapped in the syngenetic (as old as the sediment) or epigenetic (more recent than the sediment) permafrost. Pleistocene permafrost has been showed to harbor up to $5x10^7$ cells per wet gram of soil of which about a fifth is alive (3). The permafrost has thus the ability to preserve organisms for tens if not hundreds thousands of years and acts as a huge reservoir of ancient microorganisms. It has been shown for instance that numerous bacteria isolated from permafrost samples remained viable (4, 5), even potentially up to 1.1 million years (6). Even in low biomass-containing frozen environments such as glacier ice, metagenomics approaches have recently revealed hundreds of distinct bacterial genera (7). Unicellular (8–10) and even multicellular (11, 12) eukaryotes can also be preserved for thousands of years and be revived from such frozen environments.

Besides cellular organisms, metagenomics studies have revealed bacteriophages communities archived in surface (13) or deeper (7) glacier ice, the majority of which being taxonomically unassigned. Following the high bacterial abundance (14), bacteriophages are expected to be the most abundant viruses in the permafrost. However, in the unfiltered size fraction, the eukaryotic viruses Nucleocytoviricota (formerly known as Nucleocytoplasmic large DNA viruses or NCLDVs) are also highly represented (14). This phylum gathers large double stranded DNA viruses such as Pokkesviricetes (*Poxviridae* and *Asfarviridae*) as well as all the known giant viruses (i.e. viruses visible by light microscopy): the Megaviricetes (*Phycodnaviridae*, *Mimiviridae* and Pimascovirales). Likewise, a handful of scaffolds of potential *Phycodnaviridae*, also belonging to this phylum, were identified in a metagenomic study of glacial environments (13). More importantly, among Nucleocytoviricota, two giant viruses, namely Pithovirus sibericum and Mollivirus sibericum, were reactivated from a 30,000-y-old permafrost sample on *Acanthamoeba castellanii* (15, 16). Together with the presence of numerous protists and in particular amoeba in permafrost (9), this hints at the existence of many more giant viruses in such environments.

Recently, several studies specifically targeting the viral dark matter from environmental metagenomics data have started to grasp the diversity and gene-content of the Nucleocytoviricota (17–19). It became clear that the genomes of these viruses code for various auxiliary metabolic genes, making them capable of reprogramming their host's metabolism and hence are potentially important drivers of global biogeochemical cycles (17, 18, 20). They also seem to be widespread in aquatic environments. More specifically, *Mimiviridae* (in particular the proposed *Mesomimivirinae* sub-family (21)) and *Phycodnaviridae* are major contributors of the marine viromes all over the world, as revealed by thousands of metagenome-assembled viral genome (MAG) sequences (17–19). They also have been found active at the surface layer of the ocean by metatranscriptomics (22). The Nucleocytoviricota ecological functions and diversity in terrestrial samples on the other hand is far less known, with the exception of *Klosneuvirinae* sequences recovered from forest soil samples (23) and of *Pithoviridae* sequences assembled from the Loki's castle deep sea sediments sequences (24). The overwhelming proportion of marine-related as compared to terrestrial Nucleocytoviricota sequences from metagenomic studies is most likely due to the difficulty at revealing their hidden diversity in these environments (23). Soils host highly complex microbial communities making metagenomic studies notoriously challenging as population heterogeneity with closely related strains can hamper sequence assembly (25, 26).

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

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

Results

Cryosol metagenomes assemblies

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

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

Discriminating Nucleocytoviricota in metagenomic samples

From this dataset we then sought to extract Nucleocytoviricota sequences. Our method is based on the detection of both Nucleocytoviricota genes (including the ones specific to the non-icosahedral *Pithoviridae* and *Pandoraviridae*) and cellular ones. Clearly the combination of the two showed a very distinct pattern for Nucleocytoviricota compared to cellular genomic sequences (Fig. 1A), as revealed by a control metagenomic mimicking database containing reference Nucleocytoviricota genomes from (31), cellular genomes randomly sampled from Genbank in addition to amoeba and algae genomes (known to be the hosts of Nucleocytoviricota) as well as amoeba-hosted intracellular bacteria (*Babela massiliensis* and *Parachlamydia acanthamoebae*). The control database was also used to find the optimal parameters discriminating Nucleocytoviricota sequences (slope = 0.1, intercept = 1; Fig. 1), yielding high classification performance (sensitivity = 98.16% and specificity ≥ 99.53%; Fig. S2). For comparison we also tested the Viralrecall tool (35) that confirmed 1848 out of the 1973 (94%) scaffolds we detected. Finally, further controls for contamination in the Nucleocytoviricota dataset involved a search for ribosomal sequences, none of which were found. Manual functional annotation also allowed the identification of 7 scaffolds potentially belonging to intracellular bacteria, a phage and a nudivirus that were removed. At the end, our Nucleocytoviricota identification method on the permafrost dataset resulted in 1966 scaffolds ranging from 10 kb up to 1.6 Mb, corresponding to 1% of all scaffolds over 10 kb in size (Fig. 1B).

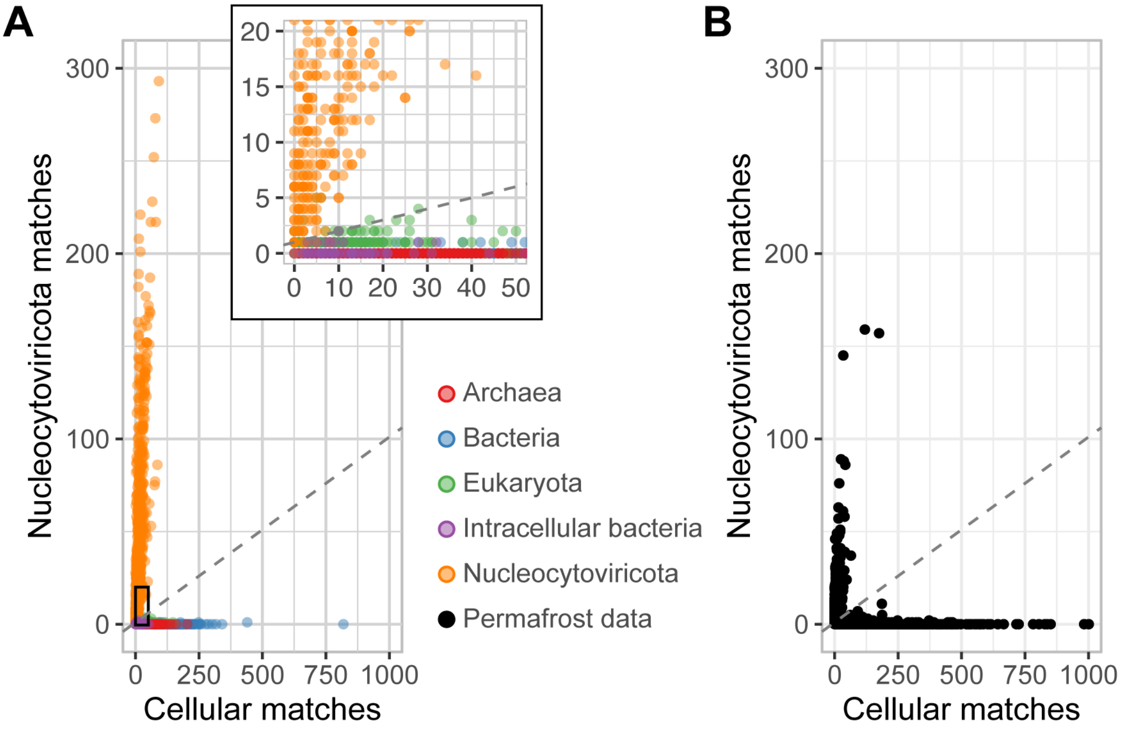

Figure 1: Extraction method of viral scaffolds

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

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

Large viral genomes from deep permafrost

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

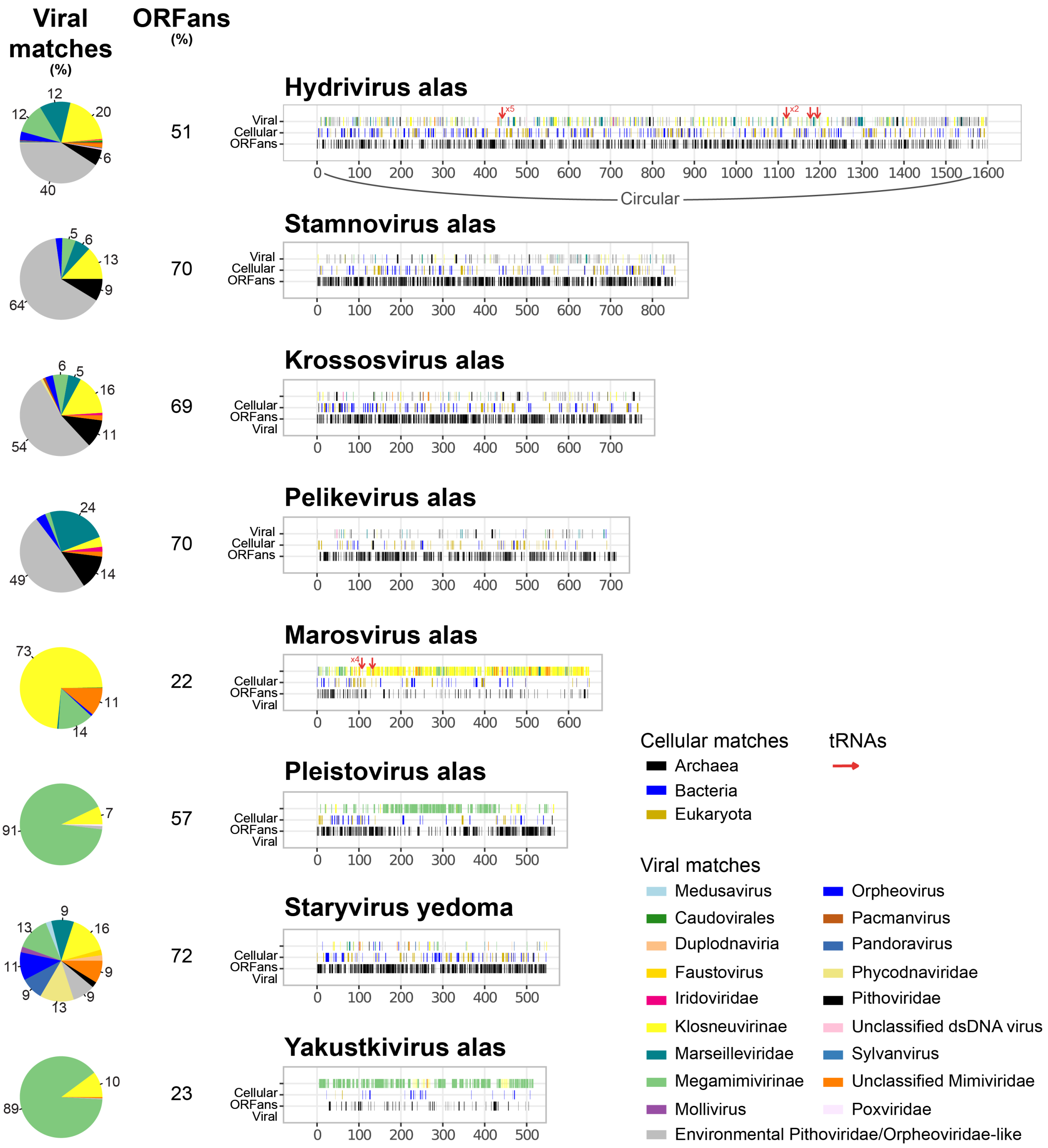

Figure 2: Gene content of the large genomes recovered from ancient permafrost samples

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

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

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

Heterogeneous Nucleocytoviricota abundance of the Russian cryosol

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

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

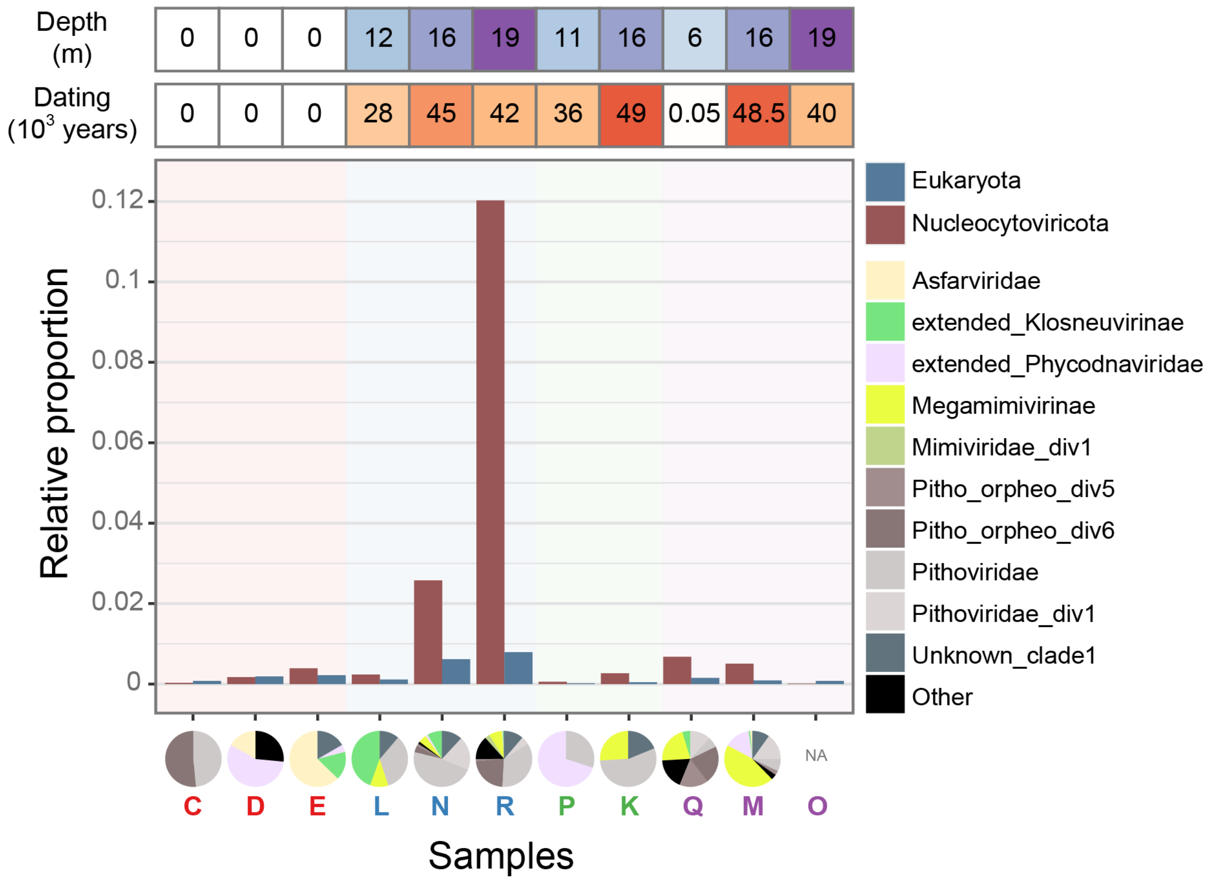

Figure 3: Relative abundance of Nucleocytoviricota and Eukaryota across samples

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

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

Exploration of the sequence diversity

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

With this method, 369 Nucleocytoviricota scaffolds (19%) were taxonomically classified (Fig. 4) corresponding to 40% of the 72 Mb of total Nucleocytoviricota identified sequences. *Pithoviridae* and *Orpheoviridae*-like viral families were clearly the most diverse followed by *Megamimivirinae*. In contrast, *Marseilleviridae*, *Alphairidoviridae*, *Betairidoviridae* and *Ascoviridae* were completely absent from our samples. *Poxviridae* were not included in the phylogeny as they were absent from our samples and adding their marker genes lowered the tree bootstraps values. In addition to our strategy to combine different marker genes, we also computed a phylogenetic tree from a single conserved one, the DNA polymerase, confirming that *Pithoviridae* and *Orpheoviridae*-like sequences were the most diverse families in our samples (Fig. S6).

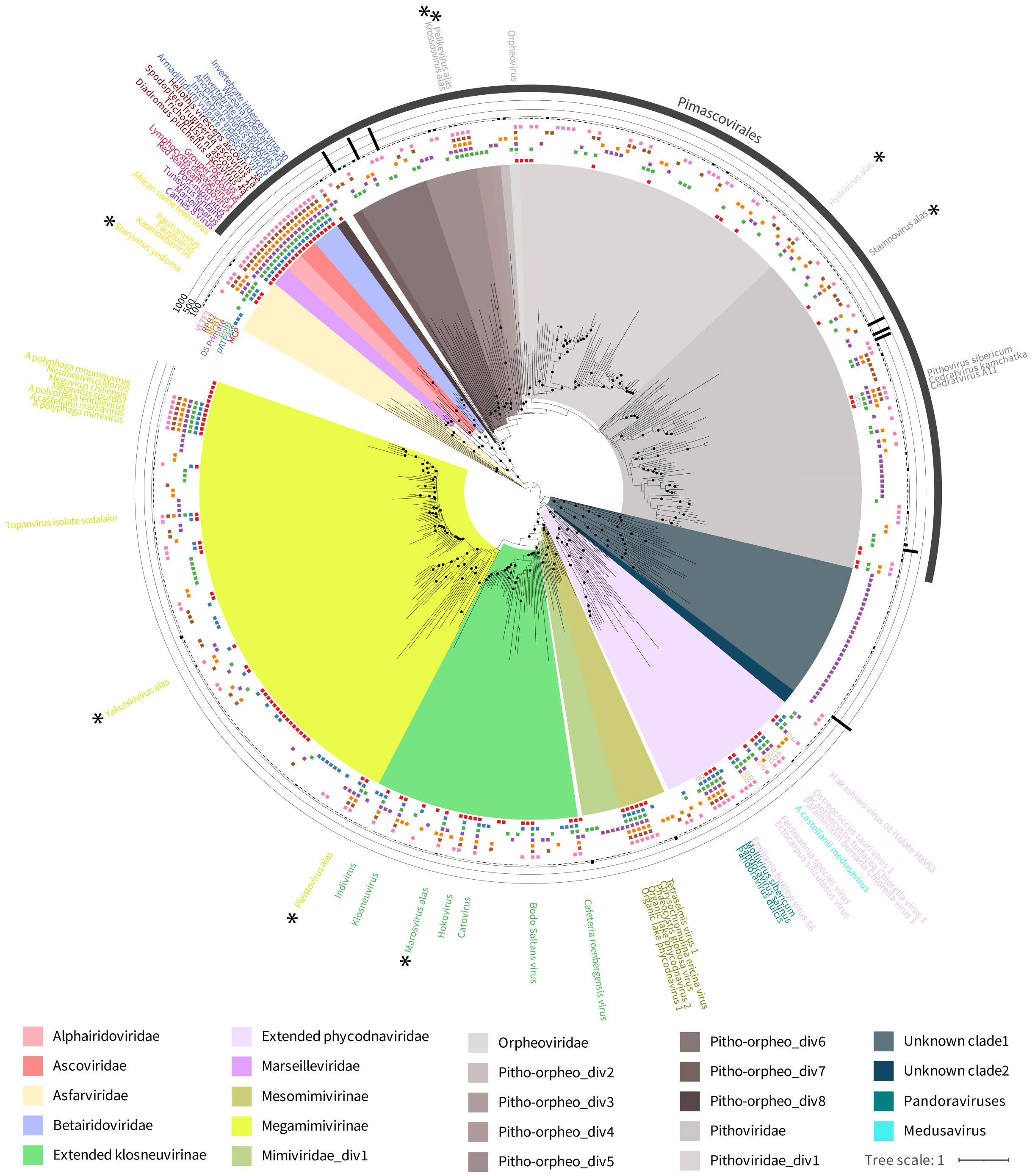

Figure 4: Nucleocytoviricota phylogeny in all samples

Consensus of 1000 bootstrapped trees calculated by a partitioned analysis on 7 marker genes. Black dots represent branch bootstrap support ≥ 90. The colored labels indicate the reference genomes and the large viruses identified in this study (marked with a star). The colored clades were manually created to be monophyletic based on reference genomes. The marker genes used for this phylogeny are indicated as colored squares. Empty squares correspond to marker genes absent from the reference genomes. Black bars show the normalized mean coverage of the scaffold. Pimascovirales are defined as the clade composed of all the *Ascoviridae*, *Iridoviridae*, *Marseilleviridae*, but also *Pithoviridae*, *Orpheoviridae* and the metagenomic intermediate clades. The Extended_phycodnaviridae group includes *Pandoraviridae* and Mollivirus. The Extended_klosneuvirinae group includes the Cafeteria roenbergensis virus.

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

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

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

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

Worldwide Nucleocytoviricota distribution

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

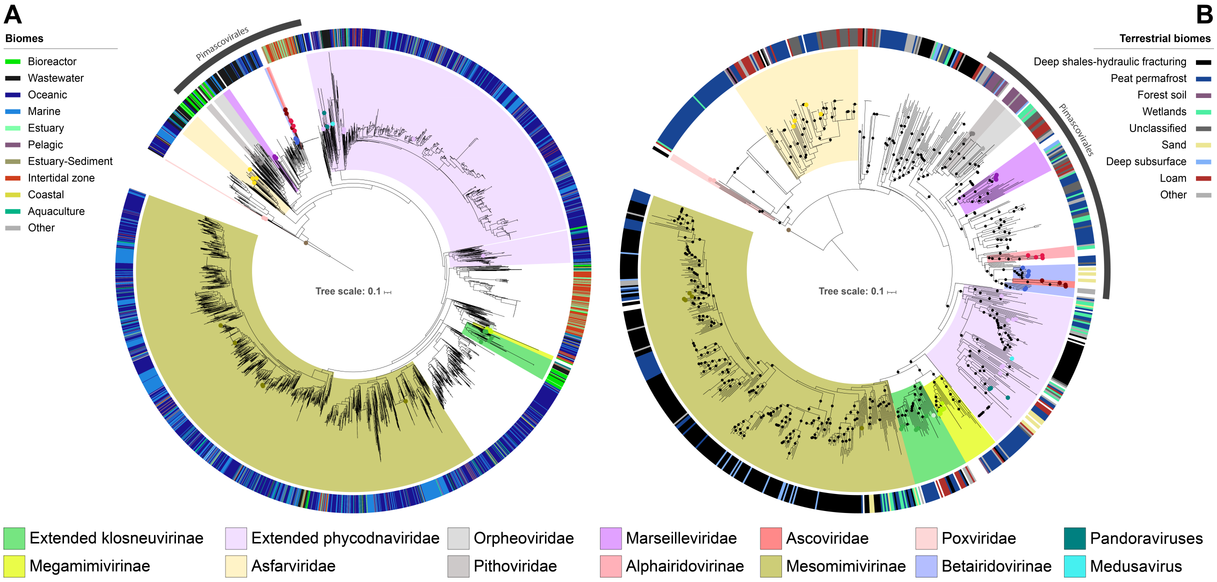

Figure 5: Worldwide Nucleocytoviricota phylogenetic distribution

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

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

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

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

Functions encoded in the permafrost Nucleocytoviricota sequences

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

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

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

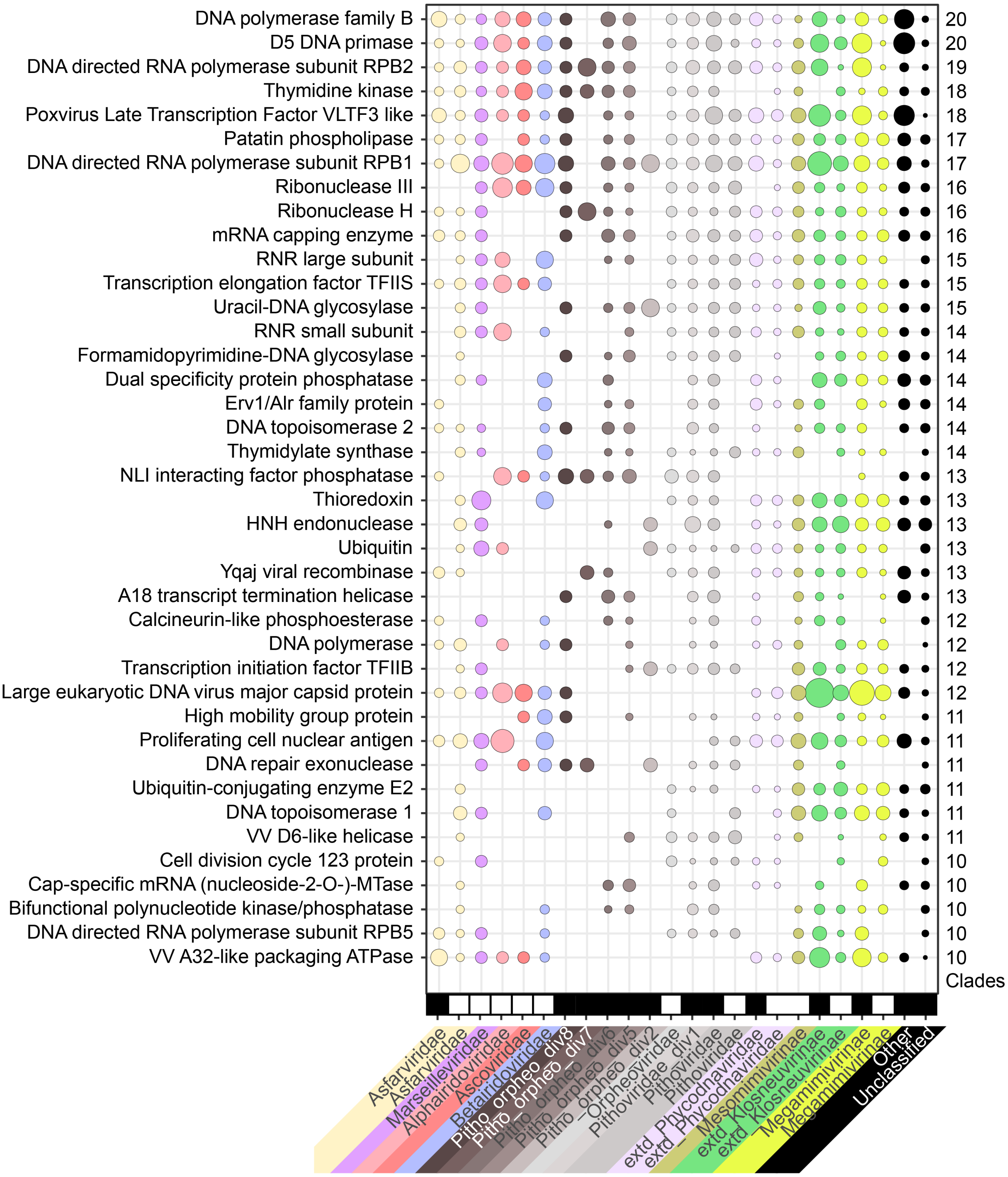

Figure 6: Most shared functions among Nucleocytoviricota families

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

DNA structure-related genes

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

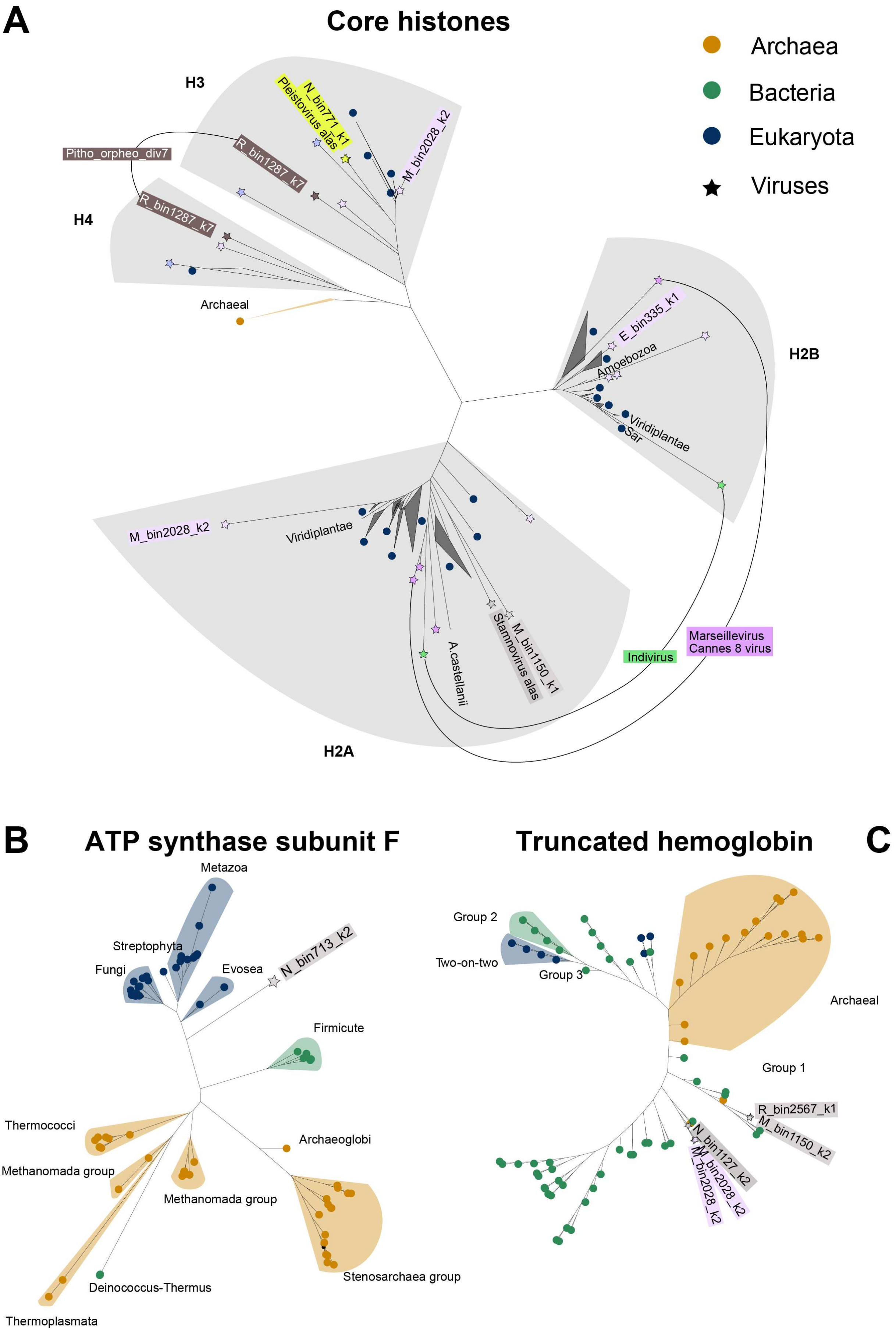
A
Core histones
Archaea
Bacteria
Eukaryota
Viruses
H3
H4
H2B
H2A
Pitho_orpheo_div7
R_bin1287_k7
R_bin1287_k7
N_bin771_k1
Pleistovirus alas
M_bin2028_k2
Archaeal
E_bin335_k1
Amoebozoa
Viridiplantae
Sar
M_bin2028_k2
Viridiplantae
A.castellanii
Stamnovirus alas
M_bin1150_k1
Indivirus
Marseillevirus Cannes 8 virus
B
ATP synthase subunit F
Metazoa
Streptophyta
Fungi
Evosea
N_bin713_k2
Firmicute
Thermococci
Methanomada group
Methanomada group
Archaeoglobi
Stenosarchaea group
Deinococcus-Thermus
Thermoplasmata
Truncated hemoglobin
C
Group 2
Two-on-two
Group 3
Archaeal
Group 1
R_bin2567_k1
M_bin1150_k2
N_bin1127_k2
M_bin2028_k2
M_bin2028_k2

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

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

Auxiliary metabolic genes

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

Other auxiliary metabolic genes found in this study include viral truncated hemoglobins that are absent from reference Nucleocytoviricota. They likely come from three different HGT events (Fig. 7C). A first one occurred between a prokaryote and Pithoviridae_div1 viruses. A second bacteria-to-virus HGT involved a *Pithoviridae* or an Extended_Phycodnaviridae that subsequently exchanged the truncated hemoglobin gene. These proteins are able to bind oxygen and protect cells against oxidative stress from NO or other oxygen reactive molecules (47).

Translation-related genes

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

We investigated the phylogeny of the different types of aaRSs found in our datasets that revealed entangled evolutionary pathways between viruses and cellular organisms (Fig. S14, S15 and S16). In most cases, the viral aaRSs came from a probable HGT from Eukaryotes (tryptophan, leucine, glutamine, threonine, methionine, isoleucine, arginine, aspartate, serine and phenylalanine) (Fig. S14 and S16). One clear example is the exchange of a threonine-tRNA synthetase from Dictyostelia (Amoebozoa) to Hydrivirus alas (Fig. S14). The exchanges concerned both the mitochondrial (for instance arginine, phenylalanine) or the cytoplasmic copies (Fig. S16). There were also some more rare cases of HGT from a Prokaryote to a virus as for the glycine- and tyrosine-tRNA synthetases that were transferred from an Archaea (Fig. S15). Genes have also passed from Bacteria to Nucleocytoviricota as for the glycine-tRNA synthetase of Hydrivirus alas and the valine-tRNA synthetase of a permafrost *Megamimivirinae*. For the latter, the bacterial sources were Rickettsiales that are endosymbionts of amoeba (48), thus probably sharing the same host. The source of the tryptophan-tRNA synthetase in Hydrivirus alas is less clear but one can see that a duplication event occurred at the same locus right after the gene was acquired (Fig. S14).

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

The Major Capsid Protein

Little is known about the structural proteins that constitute the particle of non-icosahedral giant viruses and what function their encoded MCP might have. In *Ascoviridae*, the MCP is still a major protein in the virion (51), while *Pandoraviridae* simply lack the gene. The MCP present in the related Mollivirus sibericum is only the seventh most abundant virion protein and is thought to be involved in scaffolding during virion assembly (52), as observed in *Poxviridae* (53). The annotated MCP-like genes in *Pithoviridae* and *Orpheoviridae* genomes are so divergent that a Blast homology search against the NR database fails to identify other Nucleocytoviricota homologs, even at a low confidence E-value threshold of 10^{-2}. Furthermore, the protein is not detected in Pithovirus sibericum virion proteome (15). Thus, the homology between Nucleocytoviricota MCPs and the MCP-like of *Pithoviridae*/*Orpheoviridae* is worth being explored. We therefore constructed a BlastP network of all annotated large eukaryotic DNA viruses MCPs (Fig. 8A). As expected, the *Pithoviridae* MCPs-like are the most divergent and disconnected from the rest of the network, where icosahedral viruses and *Ascoviridae* form a strong cluster. But when adding the MCPs encoded in the permafrost metagenomics scaffolds, the *Pithoviridae* share connections to the other Nucleocytoviricota (Fig. 8B). More specifically, the MCP encoded in the Pitho-orpheo_div clades (Fig. 4) fill the gap between the *Pithoviridae* and the other Pimascovirales genes (Fig. 8B-C). Increasing the BlastP E-value stringency places the *Pithoviridae*-like MCPs apart from other Megaviricetes except for Pitho-orpheo_div8, indicating that its MCP is closer to *Marseilleviridae* than to *Pithoviridae* (Fig. 8D). From this we can conclude on the homology between the icosahedral Nucleocytoviricota and the non-icosahedral *Pithoviridae* MCPs.

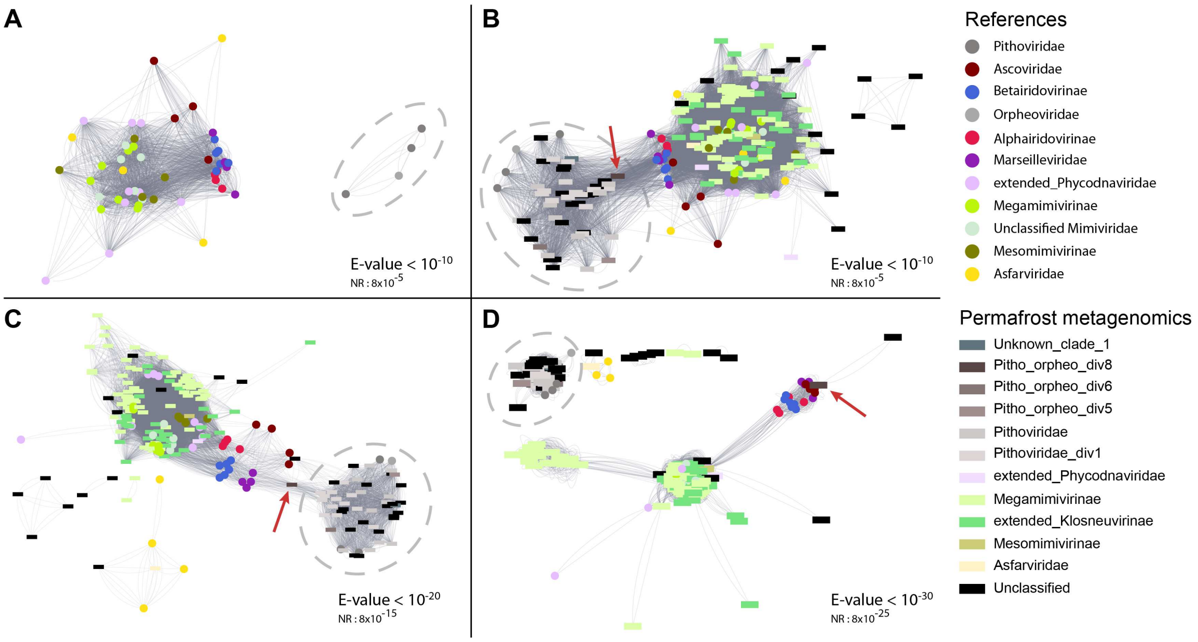

Figure 8: Major capsid protein network

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

Discussion

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

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

Despite the isolation of Mollivirus kamchatka (30) and Cedratvirus kamchatka (29) in the studied samples (samples C-D-E), their genomes were not identified in our assembled metagenomic data. Such a discrepancy had already been observed for Pithovirus sibericum and Mollivirus sibericum, where metagenomic sequence reads confirmed their presence in the samples but at a coverage too low to obtain assembled contigs (16). Concerning the *Pandoraviridae*, although different strains of these viruses were isolated from various geographical locations, including soil (28, 49, 65–68), very few *Pandoraviridae*-like sequences were identified in our cryosol data and in a vast array of environmental samples (Fig. 5), as already noticed from previous metagenomic studies (18). This underlines the importance of exploring complex environment communities with complementary approaches to unravel the true diversity of less studied giant virus families. This includes adapted metagenomic pipelines such as our attempt to reveal non-icosahedral viruses combined to direct isolation.

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

Besides the few functions shared by the Nucleocytoviricota, our work also highlights a patchwork of functions encoded by these genomes. When looking at specific functions, we detected independent cases of HGT from Eukaryotes to viruses but also between viruses belonging to different families (Fig. 7 and Fig. S14-S16). This is probably the testimony of coinfections, as members of the *Marseilleviridae, Mimiviridae, Pithoviridae, Pandoraviridae* and *Molliviridae* families can infect the same host. In line with this hypothesis, we recently showed that DNA methylation, widespread in giant viruses, is mediated by methyltransferases and Restriction-Modification systems that are frequently horizontally exchanged between viruses from different families (29).

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

The primary objective of this study was to assess the diversity of large DNA viruses in permafrost. Our analyses revealed an unexpected number of new viral sub-groups and clades among some of the previously established families of the Nucleocytoviricota phylum, mixing an intricate patchwork of functions amidst a majority of anonymous genes of unknown functions. The in-depth study of these genes will allow to better understand their physiology but also to rule on the existence or not of a common ancestor for its deepest branches.

Materials and Methods

Data preparation

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

Control database preparation

Reference Nucleocytoviricota were chosen following a former phylogenetic study (31). The corresponding genomes were gathered from the NCBI repository. Lausannevirus, Melbournevirus, Ambystoma tigrinum virus, Infectious spleen and kidney necrosis virus, Invertebrate iridovirus 22, Invertebrate iridovirus 25 and Singapore grouper iridovirus were removed to avoid an overrepresentation of their families. We added the genomes of Acanthamoeba castellanii medusavirus (AP018495.1), Bodo saltans virus (MF782455.1), Cedratvirus kamchatka (MN873693.1) and Tetraselmis virus 1 (KY322437.1). Genomes from Archaea, Eukaryota and Bacteria (Table S4) were retrieved from Genbank. For each genome, non-overlapping sequences were cut with an in house script following a distribution similar to our dataset to simulate metagenomic contigs. Genes were then predicted by Genemark (v3.36) (74) using the metagenomic model. For the Nucleocytoviricota phylogeny, core genes previously identified (31) were used in addition to the ones found by Psiblast (from BLAST+ v2.8.1) (75). We also added Amsacta moorei entomopoxvirus (AF250284.1), Variola virus (NC_001611.1) and Cyprinid herpesvirus 2 (MN201961.1) as outgroup.

Nucleocytoviricota specific profiles databases

The database constructed by (18) was completed with specific signatures of *Pithoviridae* using the genomes of Cedratvirus A11 (34), Cedratvirus kamchatka (29), Cedratvirus lausannensis (76), Cedratvirus zaza (77), Brazilian cedratvirus (77), Pithovirus massiliensis (33), Pithovirus sibericum (15), Orpheovirus (32), all the metagenomic *Pithoviridae* released from one study of Loki's Castle hydrothermal vents (24), the divergent *Orpheoviridae*/*Pithoviridae* SRX247688.42 (17), the GVMAG-S-1056828-40 (18) and other Cedratvirus/Pithovirus sequences (supplementary data files). For *Pandoraviridae* we gathered sequences from Pandoravirus braziliensis (78), P. celtis (65), P. dulcis (28), P. inopinatum (67), P. macleodensis (49), P. neocaledonia (49), P. pampulha (78), P. quercus (65), P. salinus (28), Mollivirus kamchatka (30) and M. sibericum (16). The ORFs were then predicted using Genemark (v4.32) with the "--virus" option and ORFs ≥ 50 amino-acids were kept. Orthogroups were calculated with Orthofinder (79) and HMM profiles were built using the Hmmer suite (v3.2.1) (80) for each one. HMMs were further aligned to the Refseq protein database (from March 2020) using the same suite. Only HMMs specific to *Pithoviridae*, *Orpheoviridae*, *Pandoraviridae* or Molliviruses with E-value ≤ 10^{-10} were kept to complete the database. To these were added Nucleocytoviricota-specific VOG orthogroups (https://vogdb.org/).

Retrieving viral sequences

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

Functional annotation

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

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

Contamination control

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

Large genomes assembly verification and circularization

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

Abundance estimation and mapping

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

Phylogenetic analysis

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

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

The 7 marker genes were aligned using PASTA (90), clipped with ClipKIT and concatenated by Catsequences (91). The global tree with ultrafast bootstraps was then inferred by Iqtree with options “-spp, -bb 1000” and “-bi 200 -m MFP” that calculates the best model per marker gene. Tree visualization was handled using Figtree (http://tree.bio.ed.ac.uk/software/figtree/) and the Itol web server (92).

Worldwide Nucleocytoviricota distribution

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

Phylogenetic analyses of selected functions

For each function, a dataset of proteins was built using a combination of Nucleocytoviricota ORFs, corresponding Blast matched proteins from the NR database and reference proteins from specific databases. The latter includes Uniprot reviewed proteins of domains PF01990 (ATP synthase subunit F), IPR001412 (class I aminoacyl-tRNA synthetases), IPR006195 (class II aminoacyl-tRNA synthetases) and IPR001486 (truncated hemoglobin). The reference core histone proteins were also retrieved from the HistoneDB 2.0 database (94) in addition to reviewed archaeal core histones from Uniprot (clustered using CDhit (95)). For all the functions, the multiple alignments were performed using PASTA (90) or MAFFT (84) and trimmed with ClipKit (85). The tree was then computed by Iqtree (87) with options -bb 5000 -bi 200 -m TEST.

Major Capsid Protein network

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

Acknowledgements

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

References

1. Obu J. 2021. How Much of the Earth's Surface is Underlain by Permafrost? Journal of Geophysical Research: Earth Surface 126:e2021JF006123.

2. Mackelprang R, Waldrop MP, DeAngelis KM, David MM, Chavarria KL, Blazewicz SJ, Rubin EM, Jansson JK. 2011. Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. Nature 480:368–371.

3. Burkert A, Douglas TA, Waldrop MP, Mackelprang R. 2019. Changes in the Active, Dead, and Dormant Microbial Community Structure across a Pleistocene Permafrost Chronosequence. Appl Environ Microbiol 85:e02646-18.

4. Vishnivetskaya T, Kathariou S, McGrath J, Gilichinsky D, Tiedje JM. 2000. Low-temperature recovery strategies for the isolation of bacteria from ancient permafrost sediments. Extremophiles 4:165–173.

5. Hinsa-Leasure SM, Bhavaraju L, Rodrigues JLM, Bakermans C, Gilichinsky DA, Tiedje JM. 2010. Characterization of a bacterial community from a Northeast Siberian seacoast permafrost sample. FEMS Microbiology Ecology 74:103–113.

6. Liang R, Lau M, Vishnivetskaya T, Lloyd KG, Wang W, Wiggins J, Miller J, Pfiffner S, Rivkina EM, Onstott TC. 2019. Predominance of Anaerobic, Spore-Forming Bacteria in Metabolically Active Microbial Communities from Ancient Siberian Permafrost. Appl Environ Microbiol 85:e00560-19.

7. Zhong Z-P, Tian F, Roux S, Gazitúa MC, Solonenko NE, Li Y-F, Davis ME, Van Etten JL, Mosley-Thompson E, Rich VI, Sullivan MB, Thompson LG. 2021. Glacier ice archives nearly 15,000-year-old microbes and phages. Microbiome 9:160.

8. Turchetti B, Buzzini P, Goretti M, Branda E, Diolaiuti G, D'Agata C, Smiraglia C, Vaughan-Martini A. 2008. Psychrophilic yeasts in glacial environments of Alpine glaciers. FEMS Microbiol Ecol 63:73–83.

9. Malavin S, Shmakova L, Claverie J-M, Rivkina E. 2020. Frozen Zoo: a collection of permafrost samples containing viable protists and their viruses. Biodivers Data J 8:e51586.

10. Vishnivetskaya TA, Spirina EV, Shatilovich AV, Erokhina LG, Vorobyova EA, Gilichinsky DA. 2003. The resistance of viable permafrost algae to simulated environmental stresses: implications for astrobiology. International Journal of Astrobiology 2:171–177.

11. Yashina S, Gubin S, Maksimovich S, Yashina A, Gakhova E, Gilichinsky D. 2012. Regeneration of whole fertile plants from 30,000-y-old fruit tissue buried in Siberian permafrost. PNAS 109:4008–4013.

12. Shmakova L, Malavin S, Iakovenko N, Vishnivetskaya T, Shain D, Plewka M, Rivkina E. 2021. A living bdelloid rotifer from 24,000-year-old Arctic permafrost. Current Biology 31:R712–R713.

13. Bellas C, Anesio A, Barker G. 2015. Analysis of virus genomes from glacial environments reveals novel virus groups with unusual host interactions. Frontiers in Microbiology 6:656.

14. Rigou S, Christo-Foroux E, Santini S, Goncharov A, Strauss J, Grosse G, Fedorov AN, Labadie K, Abergel C, Claverie J-M. 2021. High prevalence and diversity of beta-lactamase-encoding bacteria in cryosoils and ancient permafrost. bioRxiv 2021.03.17.435775.

15. Legendre M, Bartoli J, Shmakova L, Jeudy S, Labadie K, Adrait A, Lescot M, Poirot O, Bertaux L, Bruley C, Coute Y, Rivkina E, Abergel C, Claverie J-M. 2014. Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses with a pandoravirus morphology. Proc Natl Acad Sci U S A 111:4274–4279.

16. Legendre M, Lartigue A, Bertaux L, Jeudy S, Bartoli J, Lescot M, Alempic J-M, Ramus C, Bruley C, Labadie K, Shmakova L, Rivkina E, Couté Y, Abergel C, Claverie J-M. 2015. In-depth study of Mollivirus sibericum, a new 30,000-y-old giant virus infecting Acanthamoeba. PNAS 112:E5327–E5335.

17. Moniruzzaman M, Martinez-Gutierrez CA, Weinheimer AR, Aylward FO. 2020. Dynamic genome evolution and complex virocell metabolism of globally-distributed giant viruses. 1. Nature Communications 11:1–11.

18. Schulz F, Roux S, Paez-Espino D, Jungbluth S, Walsh DA, Denef VJ, McMahon KD, Konstantinidis KT, Eloe-Fadrosh EA, Kyrpides NC, Woyke T. 2020. Giant virus diversity and host interactions through global metagenomics. Nature 578:432–436.

19. Endo H, Blanc-Mathieu R, Li Y, Salazar G, Henry N, Labadie K, de Vargas C, Sullivan MB, Bowler C, Wincker P, Karp-Boss L, Sunagawa S, Ogata H. 2020. Biogeography of marine giant viruses reveals their interplay with eukaryotes and ecological functions. 12. Nature Ecology & Evolution 4:1639–1649.

20. Needham DM, Yoshizawa S, Hosaka T, Poirier C, Choi CJ, Hehenberger E, Irwin NAT, Wilken S, Yung C-M, Bachy C, Kurihara R, Nakajima Y, Kojima K, Kimura-Someya T, Leonard G, Malmstrom RR, Mende DR, Olson DK, Sudo Y, Sudek S, Richards TA, DeLong EF, Keeling PJ, Santoro AE, Shirouzu M, Iwasaki W, Worden AZ. 2019. A distinct lineage of giant viruses brings a rhodopsin photosystem to unicellular marine predators. Proc Natl Acad Sci USA 116:20574–20583.

21. Gallot-Lavallée L, Blanc G, Claverie J-M. 2017. Comparative Genomics of Chrysochromulina Ericina Virus and Other Microalga-Infecting Large DNA Viruses Highlights Their Intricate Evolutionary Relationship with the Established Mimiviridae Family. J Virol 91.

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

23. Schulz F, Alteio L, Goudeau D, Ryan EM, Yu FB, Malmstrom RR, Blanchard J, Woyke T. 2018. Hidden diversity of soil giant viruses. Nat Commun 9:1–9.

24. Bäckström D, Yutin N, Jørgensen SL, Dharamshi J, Homa F, Zaremba-Niedwiedzka K, Spang A, Wolf YI, Koonin EV, Ettema TJG. 2019. Virus Genomes from Deep Sea Sediments Expand the Ocean Megavirome and Support Independent Origins of Viral Gigantism. mBio 10.

25. Alteio LV, Schulz F, Seshadri R, Varghese N, Rodriguez-Reillo W, Ryan E, Goudeau D, Eichorst SA, Malmstrom RR, Bowers RM, Katz LA, Blanchard JL, Woyke T. 2020.

Complementary Metagenomic Approaches Improve Reconstruction of Microbial Diversity in a Forest Soil. mSystems 5:e00768-19.

26. Sczyrba A, Hofmann P, Belmann P, Koslicki D, Janssen S, Dröge J, Gregor I, Majda S, Fiedler J, Dahms E, Bremges A, Fritz A, Garrido-Oter R, Jørgensen TS, Shapiro N, Blood PD, Gurevich A, Bai Y, Turaev D, DeMaere MZ, Chikhi R, Nagarajan N, Quince C, Meyer F, Balvočiūtė M, Hansen LH, Sørensen SJ, Chia BKH, Denis B, Froula JL, Wang Z, Egan R, Don Kang D, Cook JJ, Deltel C, Beckstette M, Lemaitre C, Peterlongo P, Rizk G, Lavenier D, Wu Y-W, Singer SW, Jain C, Strous M, Klingenberg H, Meinicke P, Barton MD, Lingner T, Lin H-H, Liao Y-C, Silva GGZ, Cuevas DA, Edwards RA, Saha S, Piro VC, Renard BY, Pop M, Klenk H-P, Göker M, Kyrpides NC, Woyke T, Vorholt JA, Schulze-Lefert P, Rubin EM, Darling AE, Rattei T, McHardy AC. 2017. Critical Assessment of Metagenome Interpretation-a benchmark of metagenomics software. Nat Methods 14:1063–1071.

27. Koonin EV, Yutin N. 2019. Evolution of the Large Nucleocytoplasmic DNA Viruses of Eukaryotes and Convergent Origins of Viral Gigantism. Adv Virus Res 103:167–202.

28. Philippe N, Legendre M, Doutre G, Couté Y, Poirot O, Lescot M, Arslan D, Seltzer V, Bertaux L, Bruley C, Garin J, Claverie J-M, Abergel C. 2013. Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. Science 341:281–286.

29. Jeudy S, Rigou S, Alempic J-M, Claverie J-M, Abergel C, Legendre M. 2020. The DNA methylation landscape of giant viruses. Nature Communications 11:2657.

30. Christo-Foroux E, Alempic J-M, Lartigue A, Santini S, Labadie K, Legendre M, Abergel C, Claverie J-M. 2020. Characterization of Mollivirus kamchatka, the First Modern Representative of the Proposed Molliviridae Family of Giant Viruses. J Virol 94.

31. Guglielmini J, Woo AC, Krupovic M, Forterre P, Gaia M. 2019. Diversification of giant and large eukaryotic dsDNA viruses predated the origin of modern eukaryotes. Proc Natl Acad Sci USA 116:19585–19592.

32. Andreani J, Khalil JYB, Baptiste E, Hasni I, Michelle C, Raoult D, Levasseur A, La Scola B. 2017. Orpheovirus IHUMI-LCC2: A New Virus among the Giant Viruses. Front Microbiol 8:2643.

33. Levasseur A, Andreani J, Delerce J, Bou Khalil J, Robert C, La Scola B, Raoult D. 2016. Comparison of a Modern and Fossil Pithovirus Reveals Its Genetic Conservation and Evolution. Genome Biol Evol 8:2333–2339.

34. Andreani J, Aherfi S, Bou Khalil JY, Di Pinto F, Bitam I, Raoult D, Colson P, La Scola B. 2016. Cedratvirus, a Double-Cork Structured Giant Virus, is a Distant Relative of Pithoviruses. Viruses 8.

35. Moniruzzaman M, Weinheimer AR, Martinez-Gutierrez CA, Aylward FO. 2020. Widespread endogenization of giant viruses shapes genomes of green algae. Nature 588:141–145.

36. Aylward FO, Moniruzzaman M. 2021. ViralRecall—A Flexible Command-Line Tool for the Detection of Giant Virus Signatures in ‘Omic Data. 2. Viruses 13:150.

37. Mitchell AL, Almeida A, Beracochea M, Boland M, Burgin J, Cochrane G, Crusoe MR, Kale V, Potter SC, Richardson LJ, Sakharova E, Scheremetjew M, Korobeynikov A, Shlemov A, Kunyavskaya O, Lapidus A, Finn RD. 2020. MGnify: the microbiome analysis resource in 2020. Nucleic Acids Research 48:D570–D578.

38. Chen I-MA, Chu K, Palaniappan K, Ratner A, Huang J, Huntemann M, Hajek P, Ritter S, Varghese N, Seshadri R, Roux S, Woyke T, Eloe-Fadrosh EA, Ivanova NN, Kyrpides NC. 2021. The IMG/M data management and analysis system v.6.0: new tools and advanced capabilities. Nucleic Acids Research 49:D751–D763.

39. Yutin N, Wolf YI, Raoult D, Koonin EV. 2009. Eukaryotic large nucleo-cytoplasmic DNA viruses: clusters of orthologous genes and reconstruction of viral genome evolution. Virol J 6:223.

40. Aurass P, Banerji S, Flieger A. 2014. Loss of a Patatin-Like Phospholipase A Causes Reduced Infectivity of Legionella Pneumophila in Amoeba and Machrophage Infection Models 199–202.

41. Wilson SK, Heckendorn J, Martorelli Di Genova B, Koch LL, Rooney PJ, Morrissette N, Lebrun M, Knoll LJ. 2020. A Toxoplasma gondii patatin-like phospholipase contributes to host cell invasion. PLoS Pathog 16:e1008650.

42. Mozar M, Claverie J-M. 2014. Expanding the Mimiviridae family using asparagine synthase as a sequence bait. Virology 466–467:112–122.

43. Blanca L, Christo-Foroux E, Rigou S, Legendre M. 2020. Comparative Analysis of the Circular and Highly Asymmetrical Marseilleviridae Genomes. 11. Viruses 12:1270.

44. Liu Y, Bisio H, Toner CM, Jeudy S, Philippe N, Zhou K, Bowerman S, White A, Edwards G, Abergel C, Luger K. 2021. Virus-encoded histone doublets are essential and form nucleosome-like structures. Cell 184:4237-4250.e19.

45. Thomas V, Bertelli C, Collyn F, Casson N, Telenti A, Goesmann A, Croxatto A, Greub G. 2011. Lausannevirus, a giant amoebal virus encoding histone doublets. Environ Microbiol 13:1454–1466.

46. Toei M, Saum R, Forgac M. 2010. Regulation and Isoform Function of the V-ATPases. Biochemistry 49:4715–4723.

47. Lama A, Pawaria S, Dikshit KL. 2006. Oxygen binding and NO scavenging properties of truncated hemoglobin, HbN, of Mycobacterium smegmatis. FEBS Letters 580:4031–4041.

48. Schulz F, Martijn J, Wascher F, Lagkouvardos I, Kostanjšek R, Ettema TJG, Horn M. 2016. A Rickettsiales symbiont of amoebae with ancient features. Environmental Microbiology 18:2326–2342.

49. Legendre M, Fabre E, Poirot O, Jeudy S, Lartigue A, Alempic J-M, Beucher L, Philippe N, Bertaux L, Christo-Foroux E, Labadie K, Couté Y, Abergel C, Claverie J-M. 2018. Diversity and evolution of the emerging Pandoraviridae family. Nat Commun 9:2285.

50. Nasir A, Caetano-Anollés G. 2015. A phylogenomic data-driven exploration of viral origins and evolution. Science Advances https://doi.org/10.1126/sciadv.1500527.

51. Chen Z-S, Cheng X-W, Wang X, Hou D-H, Huang G-H. 2019. Proteomic analysis of the Heliothis virescens ascovirus 3i (HvAV-3i) virion. J Gen Virol 100:301–307.

52. Quemin ER, Corroyer-Dulmont S, Baskaran A, Penard E, Gazi AD, Christo-Foroux E, Walther P, Abergel C, Krijnse-Locker J. 2019. Complex Membrane Remodeling during Virion Assembly of the 30,000-Year-Old Mollivirus Sibericum. J Virol 93:e00388-19.

53. Hyun J-K, Accurso C, Hijnen M, Schult P, Pettikiriarachchi A, Mitra AK, Coulibaly F. 2011. Membrane remodeling by the double-barrel scaffolding protein of poxvirus. PLoS Pathog 7:e1002239.

54. Bergh O, Børsheim KY, Bratbak G, Heldal M. 1989. High abundance of viruses found in aquatic environments. Nature 340:467–468.

55. Cochlan WP, Wikner J, Steward GF, Smith DC, Azam F. 1993. Spatial distribution of viruses, bacteria and chlorophyll a in neritic, oceanic and estuarine environments. Marine Ecology Progress Series 92:77–87.

56. Schulz F, Yutin N, Ivanova NN, Ortega DR, Lee TK, Vierheilig J, Daims H, Horn M, Wagner M, Jensen GJ, Kyrpides NC, Koonin EV, Woyke T. 2017. Giant viruses with an expanded complement of translation system components. Science 356:82–85.

57. Raoult D, Audic S, Robert C, Abergel C, Renesto P, Ogata H, La Scola B, Suzan M, Claverie J-M. 2004. The 1.2-megabase genome sequence of Mimivirus. Science 306:1344–1350.

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

59. Filée J. 2015. Genomic comparison of closely related Giant Viruses supports an accordion-like model of evolution. Front Microbiol 6:593.

60. Krupovic M, Koonin EV. 2015. Polintons: a hotbed of eukaryotic virus, transposon and plasmid evolution. Nat Rev Microbiol 13:105–115.

61. Yutin N, Wolf YI, Koonin EV. 2014. Origin of giant viruses from smaller DNA viruses not from a fourth domain of cellular life. Virology 466–467:38–52.

62. Moreira D, Brochier-Armanet C. 2008. Giant viruses, giant chimeras: the multiple evolutionary histories of Mimivirus genes. BMC Evol Biol 8:12.

63. Ohno S. 1970. The Creation of a New Gene from a Redundant Duplicate of an Old Gene, p. 71–82. *In* Ohno, S (ed.), Evolution by Gene Duplication. Springer, Berlin, Heidelberg.

64. Suhre K. 2005. Gene and genome duplication in Acanthamoeba polyphaga Mimivirus. J Virol 79:14095–14101.

65. Legendre M, Alempic J-M, Philippe N, Lartigue A, Jeudy S, Poirot O, Ta NT, Nin S, Couté Y, Abergel C, Claverie J-M. 2019. Pandoravirus Celtis Illustrates the Microevolution Processes at Work in the Giant Pandoraviridae Genomes. Front Microbiol 10:430.

66. Hosokawa N, Takahashi H, Aoki K, Takemura M. 2021. Draft Genome Sequence of Pandoravirus japonicus Isolated from the Sabaishi River, Niigata, Japan. Microbiol Resour Announc 10:e00365-21.

67. Scheid P. 2016. A strange endocytobiont revealed as largest virus. Curr Opin Microbiol 31:58–62.

68. Dornas FP, Khalil JYB, Pagnier I, Raoult D, Abrahão J, La Scola B. 2015. Isolation of new Brazilian giant viruses from environmental samples using a panel of protozoa. Front Microbiol 6:1086.

69. Koonin EV, Yutin N. 2010. Origin and evolution of eukaryotic large nucleo-cytoplasmic DNA viruses. Intervirology 53:284–292.

70. Koonin EV, Dolja VV, Krupovic M, Varsani A, Wolf YI, Yutin N, Zerbini FM, Kuhn JH. 2020. Global Organization and Proposed Megataxonomy of the Virus World. Microbiol Mol Biol Rev 84:e00061-19.

71. Mönttinen HAM, Bicep C, Williams TA, Hirt RP. 2021. The genomes of nucleocytoplasmic large DNA viruses: viral evolution writ large. Microb Genom 7.

72. Boratto PVM, Oliveira GP, Machado TB, Andrade ACSP, Baudoin J-P, Klose T, Schulz F, Azza S, Decloquement P, Chabrière E, Colson P, Levasseur A, La Scola B, Abrahão JS. 2020. Yaravirus: A novel 80-nm virus infecting Acanthamoeba castellanii. Proc Natl Acad Sci U S A 117:16579–16586.

73. Kazlauskas D, Krupovic M, Guglielmini J, Forterre P, Venclovas Č. 2020. Diversity and evolution of B-family DNA polymerases. Nucleic Acids Research 48:10142–10156.

74. Besemer J, Lomsadze A, Borodovsky M. 2001. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Res 29:2607–2618.

75. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25:3389–3402.

76. Bertelli C, Mueller L, Thomas V, Pillonel T, Jacquier N, Greub G. 2017. Cedratvirus lausannensis - digging into Pithoviridae diversity. Environ Microbiol 19:4022–4034.

77. Rodrigues RAL, Andreani J, Andrade AC dos SP, Machado TB, Abdi S, Levasseur A, Abrahão JS, La Scola B. 2018. Morphologic and Genomic Analyses of New Isolates Reveal a Second Lineage of Cedratviruses. J Virol 92:e00372-18.

78. Aherfi S, Andreani J, Baptiste E, Oumessoum A, Dornas FP, Andrade ACDSP, Chabriere E, Abrahao J, Levasseur A, Raoult D, La Scola B, Colson P. 2018. A Large Open Pangenome and a Small Core Genome for Giant Pandoraviruses. Front Microbiol 9:1486.

79. Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biology 20:238.

80. Eddy SR. 2009. A new generation of homology search tools based on probabilistic inference, p. 205–211. *In* Genome Informatics 2009. PUBLISHED BY IMPERIAL COLLEGE PRESS AND DISTRIBUTED BY WORLD SCIENTIFIC PUBLISHING CO.

81. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, Mende DR, Letunic I, Rattei T, Jensen LJ, von Mering C, Bork P. 2019. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Research 47:D309–D314.

82. GitHub - tseemann/barrnap: Bacterial ribosomal RNA predictor. GitHub. https://github.com/tseemann/barrnap. Retrieved 19 August 2021.

83. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. Nat Biotechnol 29:24–26.

84. Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution 30:772–780.

85. Steenwyk JL, Iii TJB, Li Y, Shen X-X, Rokas A. 2020. ClipKIT: A multiple sequence alignment trimming software for accurate phylogenomic inference. PLOS Biology 18:e3001007.

86. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. 6. Nature Methods 14:587–589.

87. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Molecular Biology and Evolution 32:268–274.

88. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Molecular Biology and Evolution 35:518–522.

89. Chernomor O, von Haeseler A, Minh BQ. 2016. Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. Systematic Biology 65:997–1008.

90. Mirarab S, Nguyen N, Warnow T. 2014. PASTA: Ultra-Large Multiple Sequence Alignment, p. 177–191. *In* Sharan, R (ed.), Research in Computational Molecular Biology. Springer International Publishing, Cham.

91. Chris Creevey, Nathan Weeks. 2021. ChrisCreevey/catsequences: Version 1.3. Zenodo. https://zenodo.org/record/4409153. Retrieved 18 August 2021.

92. Letunic I, Bork P. 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics 23:127–128.

93. Arvestad L. 2006. Efficient methods for estimating amino acid replacement rates. J Mol Evol 62:663–673.

94. Draizen EJ, Shaytan AK, Mariño-Ramírez L, Talbert PB, Landsman D, Panchenko AR. 2016. HistoneDB 2.0: a histone database with variants--an integrated resource to explore histones and their variants. Database (Oxford) 2016:baw014.

95. Li W, Jaroszewski L, Godzik A. 2001. Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics 17:282–283.

96. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of 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