Evolutionary history of the transposon-invaded *Pithoviridae* genomes

Sofia Rigou[1], Alain Schmitt[1], Jean-Marie Alempic[1], Audrey Lartigue[1], Peter Vendloczki[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), IM2B, IOM, 13288 Marseille Cedex 9, France

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

Abstract

Pithoviruses are amoeba-infecting giant viruses possessing the largest viral particles known so far. Since the discovery of *Pithovirus sibericum*, recovered from a 30,000-y-old permafrost sample, other pithoviruses, and related cedratviruses, were isolated from various terrestrial and aquatic samples. Here we report the isolation and genome sequencing of two strains of *Pithoviridae* from soil samples, in addition to three other recent isolates. Using the 12 available genome sequences, we conducted a thorough comparative genomics study of the *Pithoviridae* family to decipher the organization and evolution of their genomes. Our study reveals a non-uniform genome organization in two main regions: one concentrating core genes, and another gene duplications. We also found that *Pithoviridae* genomes are more conservative than other families of giant viruses, with a low and stable proportion (5% to 7%) of genes originating from horizontal transfers. Genome size variation within the family is mainly due to variations in gene duplication rates (from 14% to 28%) and massive invasion by miniature inverted-repeats transposable elements (MITEs). While these repeated elements are absent from cedratviruses, repeat-rich regions cover as much as a quarter of the pithoviruses genomes. These regions, identified using a dedicated pipeline, are hotspots of mutations, gene capture events and genomic rearrangements, that likely contribute to their evolution.

Introduction

Pithoviridae are amoeba-infecting giant viruses possessing the largest known viral particles. The prototype of the family, *Pithovirus sibericum*, was recovered almost 10 years ago from a 30'000-y-old permafrost sample (Legendre et al. 2014). Following this discovery, 6 additional isolates, all infecting *Acanthamoeba castellanii*, have been sequenced (Andreani et al. 2016; Levasseur et al. 2016; Bertelli et al. 2017; Rodrigues et al. 2018; Jeudy et al. 2020). Their dsDNA circular genomes range from 460 to 686 kb. The *Pithoviridae* are composed of two main clades: the pithoviruses and the cedratviruses. Both possess ovoid-shaped virions, capped by a cork-like structure at one extremity for the former and at both extremities for the latter. *Orpheovirus*, the closest, although distant, relative to the family, infecting *Vermamoeba vermiformis*, also has an ovoid-shaped virion but a much larger 1.6 Mb genome (Andreani et al. 2018).

Pithoviridae have mostly been isolated from permafrost (Legendre et al. 2014; Jeudy et al. 2020; Alempic et al. 2023) and sewage (Levasseur et al. 2016; dos Santos Silva et al. 2018) samples. Metagenomic surveys have also revealed *Pithoviridae*-like sequences in deep-sea sediments (Bäckström et al. 2019), in forest soil samples (Schulz et al. 2018), and their high abundance in permafrost (Rigou et al. 2022). In every case, a phylogeny of the metagenomic viral sequences showed that they are related to the isolated *Pithoviridae* while branching outside the clade, suggesting that new viral species are yet to be discovered (Rigou et al. 2022).

Genomic gigantism has been observed several times in the virosphere, among viruses infecting prokaryotes, such as "huge" (Al-Shayeb et al. 2020) and "jumbo" phages (Yuan and Gao 2017), or eukaryotes, as in the *Nucleocytoviricota* phylum to which the *Pithoviridae* family belongs. But its origin remains a mystery as most giant viruses' genes have no known origin. In *Nucleocytoviricota*, massive horizontal gene transfers from their host (Moreira and Brochier-Armanet 2008) and gene duplications (Filée and Chandler 2008) have been proposed as the driving force behind their expanded genome size. Another mechanism proposed in *Pandoraviridae* is *de novo* gene creation from intergenic regions (Legendre et al. 2018). Whatever the main evolutionary process at play, different families of giant viruses exhibit inhomogeneity in their genomes, by having a "creative" part and a "conservative" one. This pattern is revealed by an unequal distribution of core genes, duplicated genes and genomic rearrangements, preferentially concentrated in one half of the genome (Legendre et al. 2018; Blanca et al. 2020; Christo-Foroux et al. 2020).

Another factor that might shape giant viruses' genomes are transposons. For instance different *Pandoraviridae* are known to harbor Miniature Inverted Transposable Elements (MITEs) (Zhang et al. 2018). These non-autonomous class II transposable elements are composed of terminal inverted repeats separated by an internal sequence that lacks the transposase gene. Thus, they rely on an autonomous transposon for transposition (Zhang et al. 2001). Their target sites are often as simple as AT dinucleotides that give rise to target site duplication (TSD) (Ge et al. 2017). In *Pandoravirus salinus*, the transposon probably associated to these MITEs has been found in the *A. castellanii* cellular host genome (Sun et al. 2015). The *Pithovirus sibericum* genome also contains many copies of a 140-nucleotides-long palindromic repeated sequence in non-coding regions (Legendre et al. 2014). The nature of these repeated sequences, also found in *Pithovirus massiliensis* (Levasseur et al. 2016) remains unknown. Surprisingly, cedratviruses are completely devoid of such sequences (Andreani et al. 2016).

In this study, we report the genome sequences of two new *Pithoviridae* viruses isolated from soil samples (cedratvirus borely and cedratvirus plubellavi), in addition to the recently isolated cedratvirus lena (strain DY0), cedratvirus duvanny (strain DY1) and pithovirus mammoth (strain Yana14) (Alempic

et al. 2023). The comparative analysis of these sequenced genomes, complemented with previously published *Pithoviridae* sequences (Legendre et al. 2014; Levasseur et al. 2016; Bertelli et al. 2017; Rodrigues et al. 2018; Jeudy et al. 2020), provides insight into the gene distribution and the evolution of the family. In addition, an in-depth study of pithoviruses' repeats using a dedicated tool reveals that they correspond to highly structured MITEs that massively colonized their genomes and likely influenced their evolution.

Results

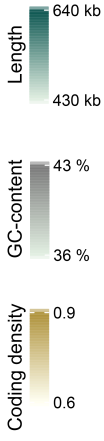
Pithoviridae isolation from soil samples and genome sequencing

We isolated two strains of cedratviruses (cedratvirus borely and cedratvirus plubellavi), both infecting *A. castellanii*, from two soil samples located 10m away in a French park (43°15'34.0"N, 5°22'58.9"E and 43°15'34.3"N, 5°22'59.2"E, respectively). As shown in Fig. S1, they possess a typical lemon-like cedratvirus morphology with two corks, one at each apex of the particle. We next sequenced their genomes. In addition we assembled and annotated the ones of three recently reported *Pithoviridae* isolated from various Siberian environments (Alempic et al. 2023), including a pithovirus from frozen soil containing mammoth wool (*Pithovirus mammoth*), a cedratvirus from the Lena river in Yakoutsk (*Cedratvirus lena*) and another cedratvirus (*Cedratvirus duvanny*) from a melting ice wedge in the Duvanny yar permafrost exposure (Table 1). All included, 12 *Pithoviridae* genome sequences are now available (Table S1) for a comparative study of the family.

Table 1. Genome metrics of sequenced *Pithoviridae* from this study compared to previously published isolates

The names of the *Pithoviridae* sequenced in this study are written in italic while the names in bold represent the mean and standard deviation of the group considering all isolates. Cedratvirus clades follow the ones defined in (Jeudy et al. 2020) and are shown in Figure 1. The right part of the table shows the genome metrics after removal of the repeats identified by our pipeline (see further).

	Real genome			Without repeats		
	Length (kb)	GC%	Coding density	Length (kb)	GC%	Coding density
Pithovirus mammoth	610	35.8	0.7	469	39.5	0.9
Pithoviruses	637 ± 40.15	35.6 ± 0.13	0.6 ± 0.03	485 ± 0.04	39.5 ± 0.04	0.9 ± 0.02
Cedratvirus borely	570	42.8	0.8	553	42.8	0.8
Cedratvirus plubellavi	568	42.8	0.8	552	42.8	0.9
Cedratviruses clade A	573 ± 10.49	42.8 ± 0.02	0.8 ± 0.01	556 ± 0.01	42.8 ± 0.01	0.8 ± 0.01
Cedratvirus lena	466	40.8	0.8	434	40.7	0.9
Cedratvirus duvanny	472	40.8	0.8	440	40.8	0.9
Cedratviruses clade B	468 ± 3.50	40.7 ± 0.10	0.8 ± 0.02	441 ± 0.00	40.7 ± 0.09	0.9 ± 0.02
Cedratviruses clade C	460	43.0	0.8	445	42.9	0.9



Pithoviridae phylogeny

To get insight into the *Pithoviridae* family evolution, we next performed a phylogenetic reconstruction of the 12 genomes in addition to the more distantly-related *Orpheovirus* (Andreani et al. 2018) and *Hydrivirus*, the only complete *Pithoviridae*-like genome assembled from metagenomics data (Rigou et al. 2022). As shown in Figure 1 *Orpheovirus* and *Hydrivirus* are the most divergent, pithoviruses and

cedratviruses split into two well established clades, and cedratviruses can be further divided into 3 previously defined clades (Jeudy et al., 2020). Although *Hydrivirus* and *Orpheovirus* cluster in a well-supported clade, they diverge from each other (Average Amino-acid Identity, AAI = 31%) more than cedratviruses from pithoviruses (AAI = 42.2% ± 0.2). In addition, *Hydrivirus* and *Orpheovirus* only share 140 HOGs (Hierarchical Orthologous Groups, see Methods), as compared to the more than 1400 genes identified in their respective genomes. This suggests that the group will likely split into better defined clades as new related viruses are added.

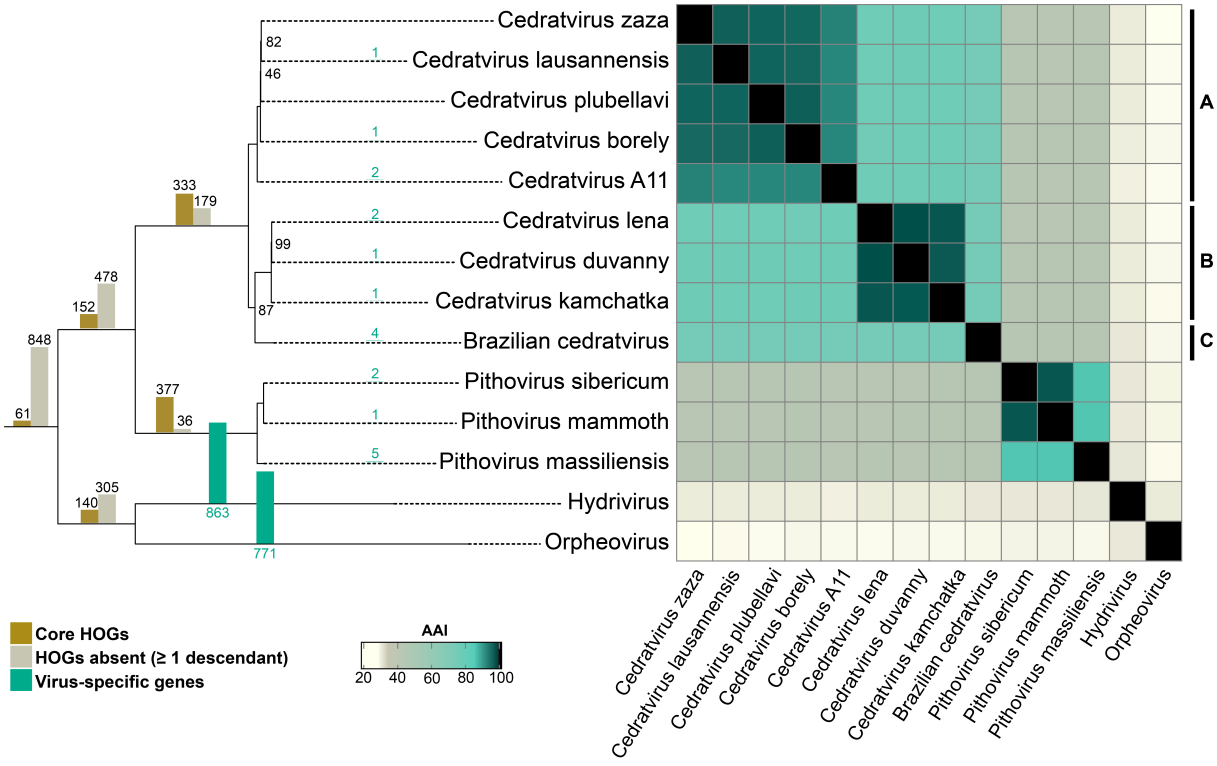


Figure 1. Phylogeny and average amino-acid identity of the *Pithoviridae* and their closest relatives

The phylogeny (left) was built from the concatenation of shared single copy HOGs (Hierarchical Orthologous Groups) applying the LG+F+G4 evolutionary model. Bootstrap values are indicated or are 100% otherwise. The bars on each branch represent the number of shared HOGs and other HOGs that were recomputed by OrthoFinder according to this tree. The heatmap (right) shows the average amino-acid identity (AAI) between viruses. The right-most bars (labeled A, B and C) indicate previously determined cedratvirus clades (Jeudy et al. 2020).

Consistent with the phylogeny, the codon usage pattern shows a similar trend, with cedratviruses tightly clustered together, as for pithoviruses, and *Orpheovirus* being the most distant (Fig. S2). This is in line with the fact that the *Pithoviridae* and *Orpheovirus* infect different laboratory hosts (Andreani et al. 2018).

Cedratviruses or pithoviruses exhibit sequence conservation and gene collinearity despite several rearrangements (Fig. S3). *Pithovirus massiliensis* shows one major inversion and one translocation compared to the two other pithoviruses. Both *Cedratvirus kamchatka* and *Brazilian cedratvirus* exhibit many rearrangements compared to clade A.

Heterogeneity within the genomes of *Pithoviridae*

The comparative genomics studies of other giant virus families previously highlighted a biased evolution of their genomes with a “creative” and a “conservative” part (Legendre et al. 2018; Blanca

et al. 2020). We thus looked for a similar trend in the *Pithoviridae* genomes. As shown in Figure 2A, core genes are not uniformly distributed along the artificially linearized pithoviruses' genomes, with a high concentration at the extremities and a lower density at the center. Likewise, core genes are also very scarce in the mid-right portion of the cedratviruses genomes, except for *Brazilian cedratvirus* who has undergone rearrangements in this region (Fig. S3). This pattern contrasts with gene duplications that seem to occur in specific hotspots preferentially located in the center (Fig. 2B). Altogether, this data shows a shared non-uniform architecture of the *Pithoviridae* genomes.

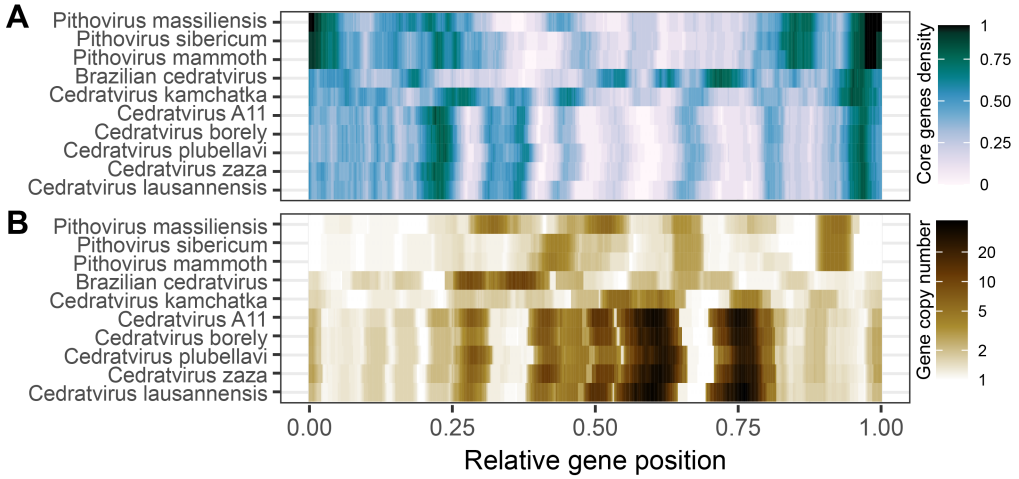


Figure 2. Non-uniform distribution of core and duplicated genes

(A) Average density of core genes within a sliding window of 21 ORFs. (B) Average gene copy number within the HOGs containing each of the genes of the sliding window.

Pithoviridae are conservative compared to other *Nucleocytoviricota*

We next quantified the *Pithoviridae* core- and pan-genomes and compared them to other viral families. The core-genome of cedratviruses is of 333 ORFs (Open Reading Frames) over 100 amino-acids (Fig. 3A-B) while the one of the whole *Pithoviridae* family is twice as small with an asymptote at 152 ORFs (Fig. 3B).

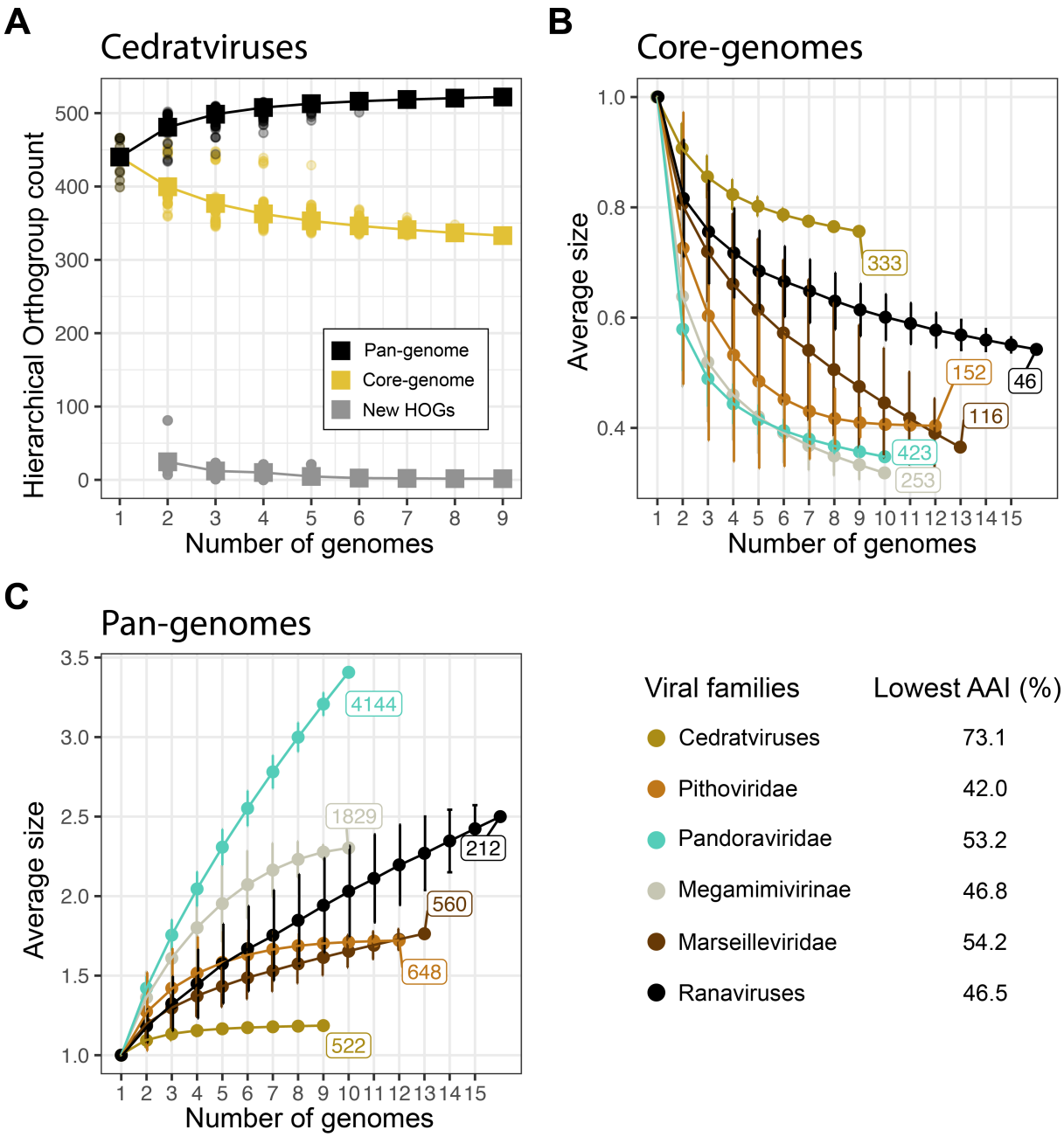


Figure 3. The core and pan-genomes of *Pithoviridae* and other *Nucleocytoviricota*
(A) Pan-genome, core-genome and new HOGs have been estimated for cedratviruses by adding new genomes to a set of previously sequenced genomes in an iterative way. For comparison, the core-genome (B) and pan-genome (C) sizes of other *Nucleocytoviricota* have been estimated in the same iterative way. The pan-genome and core-genome sizes are defined as the relative size in comparison to their initial mean size. The lowest AAI shown in the legend indicates the AAI of the most distant viruses within the set of genomes used for this analysis.

The pan-genomes of cedratviruses and *Pithoviridae* have both reached a plateau (Fig. 3C), suggesting a so-called "closed" pan-genome. In agreement with this, each new genome brings less than two new HOGs to the cedratviruses (Fig. 3A). In other words, *Pithoviridae* appear to be much more conservative (*i.e.* closer pan-genome) than other *Nucleocytoviricota* (Fig. 3C). Thus, unlike for *Pandoraviridae,* continuous *de novo* gene creation might not be a significant process in the evolution of *Pithoviridae* (Legendre et al. 2018). Furthermore, in concordance with this apparent conservative evolution, cedratviruses and pithoviruses specific genes are mostly shared within their respective genomes, in

contrast to *Pandoraviridae* and *Marseilleviridae* that exhibit a much larger fraction of accessory genes within their clades (Fig. S4).

Gene duplication and HGT in *Pithoviridae*

Next, we investigated gene duplication as a possibly important cause of viral genome gigantism (Filée and Chandler 2008). Gene duplications occurred all along the history of *Pithoviridae*, even during the short divergence time separating the closely related *Pithovirus sibericum* and *Pithovirus mammoth*. They mostly occurred in the vicinity of their original copy with a median distance of 6872 bp in cedratviruses and 1575 bp in pithoviruses. Overall, from 14 % to 28 % (median = 19 %) of the *Pithoviridae* genes come from a duplication event (Fig. 4), in line with other *Nucleocytoviricota* such as *Marseilleviridae* (16 %), *Pandoraviridae* (15 %) and *Megamimivirinae* (14 %). Within cedratviruses, gene duplications largely explain genome size variations between clade A and clades B-C, with 27.4 ± 0.9 % in clade A and 18.5 ± 1.7 % in clades B-C (Fig. 4). Consistently, the most duplicated gene, coding for an ankyrin-repeat protein, is present in 50 copies in clade A cedratviruses and only 20 copies in clades B-C. Likewise, the related *Orpheovirus* and *Hydrivirus* very large genomes exhibit high rates of gene duplications, with 42% and 27% respectively (Fig. 4). By contrast, the larger genome size of pithoviruses compared to that of cedratviruses does not correlate with a higher rate of gene duplication, suggesting that another factor is at play.

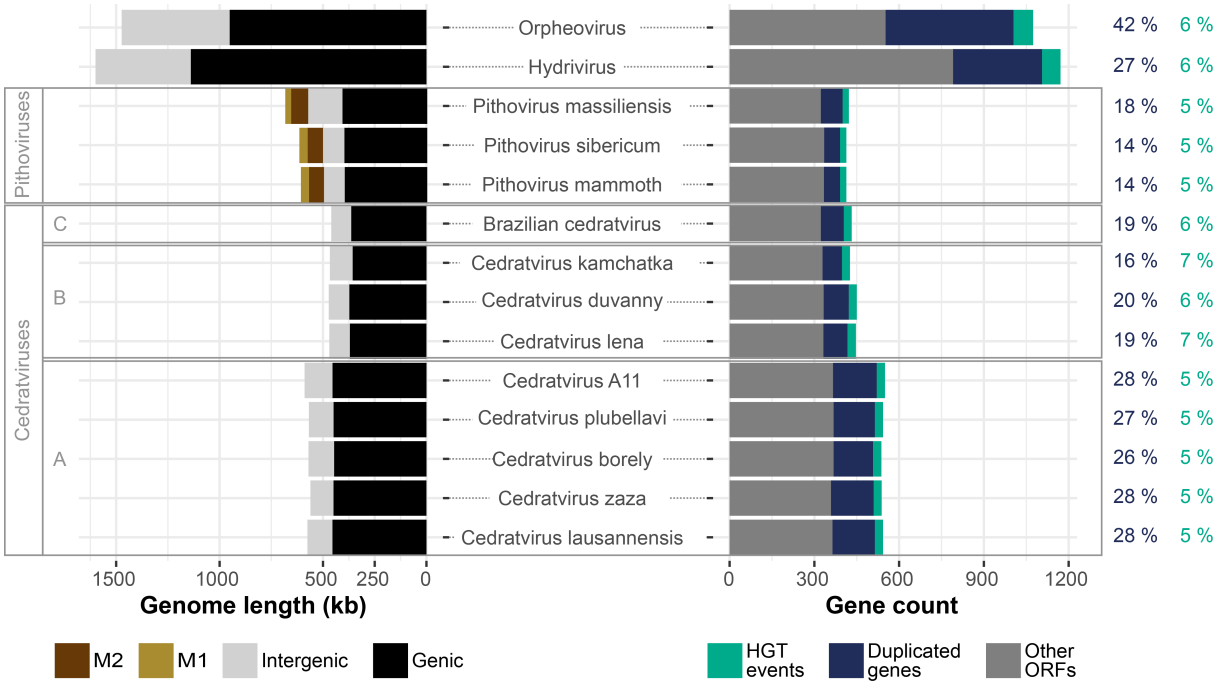


Figure 4. Genome and gene content statistics of *Pithoviridae* and relatives

The left panel presents the nucleotide content of the different genomes with clades labels on the left (see Fig. 1). M1 and M2 correspond to repeat MITEs (see further). The right panel shows their composition in ORFs. The percentage of genes that arose from a duplication event (in blue) and the percentage of HGT events toward each genome (in green) are shown on the right.

We next investigated horizontal gene transfers (HGTs) towards our viruses based on the HOG phylogenetic trees complemented with homologous sequences (see Methods), as a possible source of genome size increase. It turned out that HGTs are far less frequent than gene duplications with a stable fraction of 5% to 7% of the gene content across *Pithoviridae* and in *Orpheovirus* and *Hydrivirus* (Fig. 4).

The largest proportion of *Pithoviridae* HGTs come from eukaryotes (42 ± 2 %) closely followed by those originating from Bacteria (41 ± 3 %), (Fig. S5). The HGT from Eukaryota do not clearly point to known

hosts. Most often, the root of the HGT is ancient, branching before or in-between Discosea and Evosea, two classes of amoebas (Fig. S5). We also estimate that 10 % of the HGT events came from another virus.

Overall, the low rate of HGT in *Pithoviridae* is coherent with the closeness of their pan-genome and thus cannot account for the difference in genome sizes between cedratviruses and pithoviruses, hinting again at a different factor.

Two MITEs massively colonized the genomes of pithoviruses after they diverged from cedratviruses

Repeat content is another factor that could strongly influence genome size. Indeed, it has been shown that pithoviruses genomes are shaped by intergenic interspersed palindromic repeat sequences (Legendre et al. 2014). These are present in clusters and usually separated by 140 nucleotides (median). After masking these sequences (Fig. 5A-B) from the genomes, we identified additional repeats close to the masked regions (Fig. 5C-D). By running the MUST (Ge et al. 2017) and MITE-Tracker (Crescente et al. 2018) tools, we found that both types of repeats were distinct MITEs that we referred to as M1 and M2.

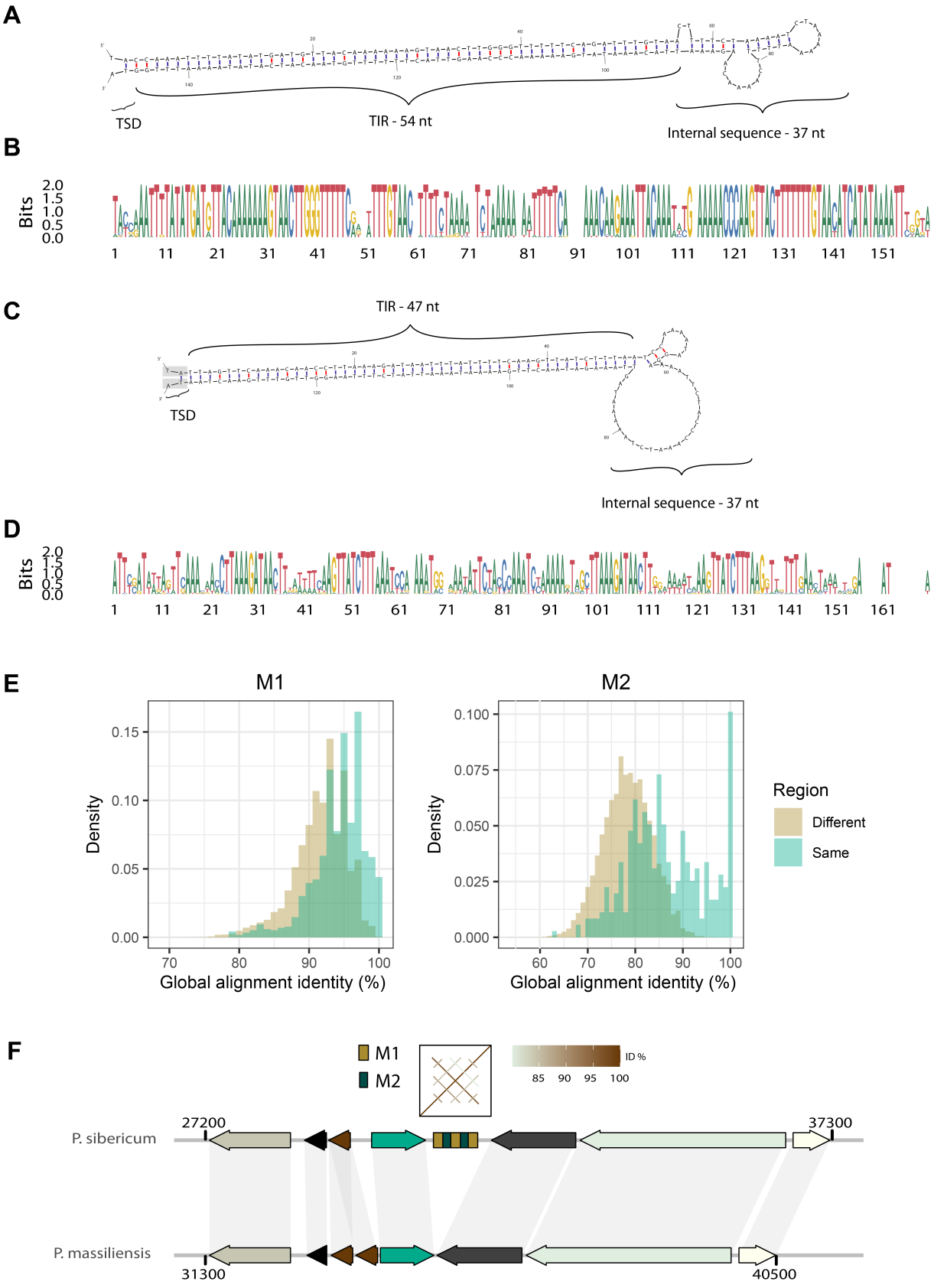


Figure 5. MITEs found in pithoviruses

DNA folding structures of the reference sequence for M1 (A) and M2 (C) MITEs clusters respectively. Their free energy ΔG is of -79.2 and -65.5 kcal/mol. TSD is for Target Site Duplication and TIR for Terminal Inverted Repeat. The TSD highlighted in grey in (C) indicates that the dinucleotide is shared between M1 and M2 when the MITEs are next to each other. (B) and (D) are the alignment logos of all

the sequences in the clusters of M1 and M2 respectively. (E) Pairwise identity percentage in-between M1 (left) and M2 (right) repeats retrieved from the same (green) and from distinct (brown) regions. The pairwise identity percentages were calculated using the needle tool from the EMBOSS package. Both distributions are significantly different (*P*values < 10^{-15}, wilcoxon test). The pairs coming from different regions were randomly subsampled to match the number of pairs in the other distribution. (F) Example of a repeated region found in *Pithovirus sibericum* but absent from *Pithovirus massiliensis* in a syntenic region of their genomes.

We designed a pipeline dedicated to automatically identify and cluster repeat sequences (see Methods and Fig. S6). When combined together, the M1 and M2 sequences represent as much as 18.4%, 18.2% and 16.1% of the genomes of *Pithovirus sibericum*, *Pithovirus mammoth* and *Pithovirus massiliensis*, respectively (Fig. 4), and 21 to 24% when all kinds of repeats are considered. Unlike duplicated and core genes (Fig. 3), repeats are not concentrated in specific genomic regions. Instead, we found that they were uniformly distributed along the pithoviruses genomes (Kolmogorov-Smirnov test against uniform distribution *P*value = 0.6). Our pipeline also provided an extensive description of the structure of the repeated regions resulting in the following rules:

1) M2 can never be seen in a repeat region without M1
2) M1 can be seen without M2
3) When several M1 are present in a region, they are always separated by a sequence of about 140 bases, whether M2 is present or not
4) When several M2 are present in a region, they are separated by M1

The most common structure of the repeated regions in the three pithoviruses genomes is: (M1-M2){1 to 8 times}-M1. In *Pithovirus sibericum*, M1 is present 515 times and M2, 371 times. *Pithovirus mammoth* has a very similar number of regions containing M1 and M2 (Table S2) but the number of M1 or M2 copies per orthologous region is often different. Thus, the differences most often correspond to the extension or contraction of existing repeated regions rather than insertions of a MITE in a repeat-free region. The extension of existing repeat regions is supported by the fact that repeats from the same region are more similar to each other than repeats from different regions (Fig. 5E, *P*values < 10^{-15}). However, insertion of repeats in repeat-free regions is also necessary to explain the observed distribution of repeat regions. Such insertions and/or excisions have happened several times since the divergence of *Pithovirus sibericum* and *Pithovirus massiliensis*, as exemplified by a repeated region in *Pithovirus sibericum* but absent from the cognate syntenic orthologous region in *Pithovirus massiliensis* (Fig. 5F). We also found slightly more M1-containing regions in *Pithovirus massiliensis* compared to *Pithovirus sibericum* (Table S2).

In search for an autonomous transposon associated to these MITEs, we screened the non-redundant NCBI database (that includes the genome of *Acanthamoeba castellanii*) and metagenomic *Nucleocytoviricota* sequences. No corresponding transposon or M1 and M2 MITEs were found. However, we found a few *Pithoviridae*-like genomes from metagenomic data that were highly structured by direct repeats (Fig. S7A-B). These constitute 13% of the LCPAC302 pithovirus-like partial genome sequenced from deep-sea sediments (Bäckström et al. 2019). Overall, *Pithoviridae* and *Pithoviridae*-like genomes are highly diverse in repeat content, ranging from none to almost a quarter of their genomes.

Pithoviruses' repeat-rich regions are hotspots of genetic variability

As repeats constitute a large proportion of pithoviruses' genomes, we further investigated the genes located in those regions. Although HGTs are not abundant in pithoviruses (Fig. 4), they are sightly but significantly enriched in repeats regions: 9.9% within *versus* 5.2% outside (chi² test *P*value = 9.9 x 10^{-4},

and individual *P*values of 0.04, 0.35 and 0.05 in *Pithovirus sibericum*, *Pithovirus massiliensis* and *Pithovirus mammoth*, respectively).

We also estimated the ancestry of the genes present within these regions compared to other regions. This was performed considering the last common ancestor of all species within each HOG. From that, we observed a significant trend (Cochran-Armitage test *P*value = 5 x 10^{-4}) whereby newly acquired genes appeared more frequent than ancestral genes in these regions (Fig. 6A). In other words, repeated regions are more prone to gene novelty.

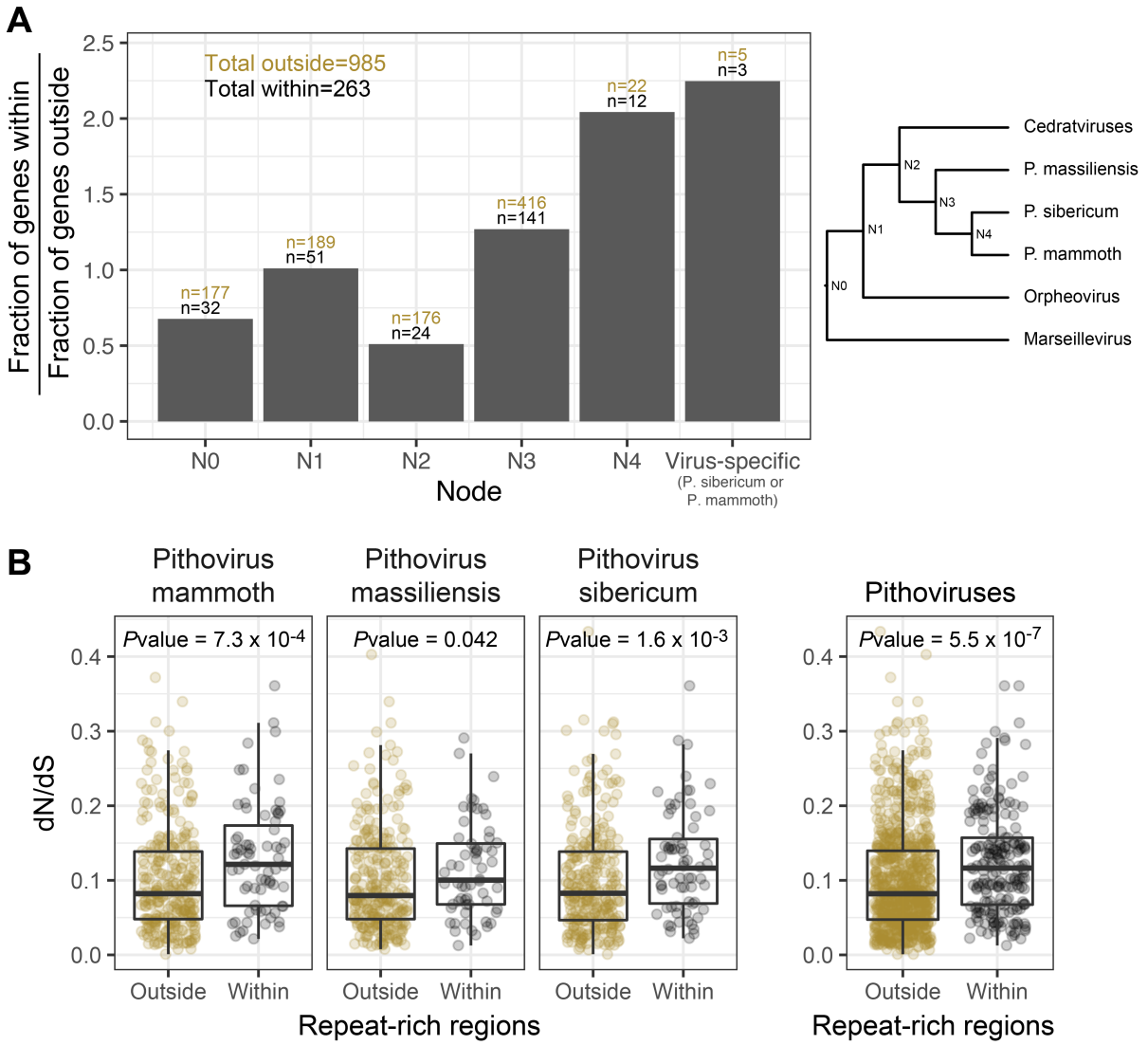


Figure 6. Evolution of ORFs within and outside repeat-rich regions

(A) Ancestry of genes within and outside repeat-rich regions. The ancestry of each gene was estimated considering the last common ancestor of the species present in the cognate HOG. Nodes are ordered from the most ancient to the most recent, as shown in the cladogram next to the plot. (B) dN/dS values of each gene within (gray) or outside (brown) repeat-rich regions detected by our pipeline. *P*values were calculated using Wilcoxon rank tests.

The rates of mutation in repeat-rich *versus* repeat-free regions were compared using orthologous pithovirus genes. We found that genes located in repeat-rich regions tended to have higher mutation rates for both, synonymous (*P*value = 7.7 x 10^{-6}) and non-synonymous (*P*value = 6.3 x 10^{-13}) positions (Fig. S8). In addition, the genes within repeat-rich regions also exhibit higher dN/dS values, thus are less evolutionary constrained (Fig. 6B, *P*value = 5.5 x 10^{-7}).

Finally, we investigated the frequency of genomic rearrangements located in repeat-rich compared to repeat-free regions. We took advantage of the fact that two pithoviruses (*Pithovirus sibericum* and *Pithovirus mammoth*) were sequenced using long reads and exhibited mostly colinear genomes. We manually inspected orthologous regions of these two viruses to spot potential rearrangement and mutational events. Again, we found that repeat-rich regions were highly enriched in several types of rearrangements compared to repeat-free regions. This includes insertions/deletions, inversions, duplications and substitutions affecting genes, accounting for a total 28 events in repeat-rich regions for only 13 in repeat-free regions (chi^2 *P*value = 1.42×10^{-11}, Table S3).

Altogether, these various results establish that pithoviruses repeat-rich regions are hotspots of genetic novelty and undergo relaxed evolutionary constraints.

Discussion

Here we reported the isolation and genome sequences of two strains of cedratviruses (*Cedratvirus borely* and *Cedratvirus plubellavi*) from soil samples. We also assembled and annotated the genome of *Pithovirus mammoth* recently isolated from 27000-y-old permafrost (Alempic et al. 2023), of a cedratvirus from fresh water (*Cedratvirus lena*) (Alempic et al. 2023) and another one from melting ice (*Cedratvirus duvanny*) (Alempic et al. 2023). Along with previously described *Pithoviridae*, mostly originating from permafrost (Legendre et al. 2014) and sewage water (Levasseur et al. 2016; Bertelli et al. 2017; dos Santos Silva et al. 2018), these new isolates confirm the ubiquity of this viral family, members of which are present within various aquatic and soil environments. This is also consistent with recent metagenomic surveys exhibiting the presence of *Pithoviridae* in permafrost, forest soils and deep-sea sediments (Bäckström et al. 2019; Rigou et al. 2022).

These 5 additional sequenced strains were combined to 7 previously published genomes to perform a thorough comparative analysis of the *Pithoviridae* family, revealing the organization of their circular genomes. We found that their genes are broadly distributed in two distinct regions, one enriched in core genes and the other in gene duplications (Fig. 2). This type of non-uniform genome partition with a “creative” and a conserved region is reminiscent of what has been observed in *Marseilleviridae* (Blanca et al. 2020), a viral family belonging to the same order (*Pimascovirales*) and whose genomes are also circular. However, the two regions are more clearly defined in *Marseilleviridae*, where duplications and accessory genes are evenly dispersed in the “creative” region, while they occur in hotspots in *Pithoviridae* (Fig. 2).

Other viral families, that only share a handful of genes (Mönttinen et al. 2021) and have various virion morphology and genome organization (linear or circular), also exhibit this non-uniform distribution of their genes. In *Poxviridae* for instance, core genes are concentrated in the central part of the genome while accessory genes, mostly involved in host-virus interactions, are located at the genome termini (Senkevich et al. 2021). It has been proposed that this accessory partition is a hotspot of frequent gene loss and gain through HGTs (Senkevich et al. 2021), but the few HGT identified in *Pithoviridae* does not support this model. In *Pandoraviridae*, core and essential genes, and those whose proteins are identified in the viral particle, are mostly localized in the left part of the genome, while accessory genes are located on the right part (Legendre et al. 2018; Bisio et al. 2022). This likely reflects ongoing genome increase involving *de novo* gene creation (Legendre et al. 2018) and accelerated gene duplications (Bisio et al. 2022). In all cases, the dichotomous genome partitioning might also be linked to the epigenetic transcriptional regulation of core genes, through a peculiar 3D structure of this genomic region. A hypothesis that remains to be tested.

Even though *Pithoviridae* genomes are conservative, the cedratviruses and pithoviruses clades exhibit large differences in genome sizes correlated with their repeat contents. These repeats correspond to two types of non-autonomous transposable elements (M1 and M2 MITEs) frequently organized as M1{M2-M1} repeated patterns. It is not uncommon for MITEs to insert next to another MITE, like in the rice (*Oryza sativa*) genome where this event occurred several times (Tarchini et al. 2000) and where 11% of MITEs exist in multimers (Jiang and Wessler 2001). The repeat-rich regions structure in pithoviruses suggests that M1 and M2 mostly move together. However, the fact that M1 can be seen without M2 and is also more frequent (Table S2), suggests that they were once independent.

It has been shown that *submariner* MITEs could colonize the *Pandoraviridae* genomes (Sun et al. 2015; Zhang et al. 2018). However, their presence in up to 30 copies only represents a minute fraction of their genomes (a few kb in up to 2.5 Mb). In pithoviruses, repeats regions cover as much as a quarter of the genomes. A comparison of the *Pithovirus sibericum* and *Pithovirus massiliensis* genomes shows that the transposition of their MITEs has been ongoing since the two species diverged, more than 30,000 years ago (Legendre et al. 2014; Levasseur et al. 2016). The primo-invasion followed by drastic expansion occurred after the pithovirus/cedravirus clade divergence. Was this the result of an explosive event or that of a gradual invasion remains to be determined. More deeply branching pithoviruses would be needed to settle this question. We have looked for an autonomous transposon that would be responsible for the M1 and M2 MITEs transposition, but none was found, either in the viral or known host genomes. Since pithoviruses genomes are circularized, it is unlikely that the transposon is lacking from the assembly. One could hypothesize that after drastic repeat expansion it became detrimental and thus, viral strains that lost the transposon were favored. We also cannot exclude that this sequence is still present in an unknown *A. castellanii* strain or a non-assembled region of the available genome sequence. In addition to transposition, our data suggests that M1 and M2 MITEs could locally multiply within repeats region, probably through recombination events.

If repeat-rich regions constitute a large fraction of pithoviruses genomes, they are also the source of genetic innovations. By comparing genes localized in repeat-rich regions with those in other regions, we found that they are more divergent and less evolutionary constrained (Fig. 6 and Fig. S8). We also found that repeat-rich regions are prone to gene capture of cellular and viral origins, and undergo many genomic rearrangements. One could hypothesize that the high conservation of MITEs sequences triggers genomic recombination and gene exchange between co-infecting viral strains. As previously stated, our comparative analysis of the *Pithoviridae* family shows that they are conservative compared to other giant viruses' families. These genomic islands might thus provide an opportunity for them to promote genetic diversity and raw genetic material for evolution to work on.

Materials & Methods

Isolation of cedratviruses

Cedratvirus borely and cedratvirus plubellavi were isolated in February 2017 from soil samples from Marseilles, France (Parc Borély). The isolation and cloning of viruses were performed as previously described (Christo-Foroux et al. 2020).

Genome sequencing, assembly and annotation

Pithovirus mammoth sequence data was assembled using a combination of Illumina MiSeq short reads and Oxford Nanopore Technologies long reads. *Cedratvirus borely* and *Cedratvirus plubellavi* were sequenced using the Illumina MiSeq technology. Finally, the genomes of *Cedratvirus lena* and *Cedratvirus duvanny* were sequenced using Illumina MiSeq and NovaSeq technologies. *Cedratvirus lena* was assembled after removing reads mapped to a contaminant pandoravirus using Bowtie 2.

Cedratvirus lena and *Cedratvirus duvanny* reads were assembled using SPAdes v 3.14 (Prjibelski et al. 2020) with options --careful and -k 15,17,19,21,29,33,41,55,63,71,91,101,115. The scaffolding was then performed by RaGOO (Alonge et al. 2019) using *Cedratvirus kamchatka* as template. *Cedratvirus borely* and *Cedratvirus plubellavi* were assembled using SPAdes v 3.9.1 and v 3.9.0, respectively, with the –careful option. The *Pithovirus mammoth* sequence was assembled using Unicycler (Wick et al. 2017) v 0.4.8 with Illumina short reads and nanopore long reads larger than 40 kb.

The 3 pithoviruses and the 9 cedratviruses genomic sequences were then artificially linearized to start at the same position for comparative analyses. The accessions of the previously published genomes used in this study can be found in Table S1A and S1B.

For functional annotation, genes were predicted using Genemark (Besemer et al. 2001) with option –virus. ORFs over 50 amino acids were kept for publication and ORFs over 100 amino acids were used for comparative genomic analysis.

ORFs were annotated using InterProScan (v5.39-77.0, databases PANTHER-14.1, Pfam-32.0,ProDom-2006.1, ProSitePatterns-2019_01, ProSiteProfiles-2019_01, SMART-7.1, TIGRFAM-15.0) (Jones et al. 2014) and CDsearch (Conserved Domain Database) (Lu et al. 2020). We also searched for viral specific functions using hmmsearch on the virus orthologous groups database (https://vogdb.org/). ORFs were compared to the nr and swissprot databases using BLASTP (Altschul et al. 1990). Transmembrane domains were identified with Phobius (Käll et al. 2004).

Relative synonymous codon usage as well as other gene and genome metrics were calculated using an in-house script relying on Biopython (Cock et al. 2009).

Computation of orthologous gene groups and phylogeny

A phylogenetic tree was computed by OrthoFinder (v2.5.4) (Emms and Kelly 2019) using all available *Pithoviridae* genomes in addition to the *Orpheovirus* (Andreani et al. 2018), *Hydrivirus* (Rigou et al. 2022) and *Marseillevirus* genomes (Table S1). The tree was then rooted using the distantly related *Marseillevirus* (Boyer et al. 2009) as an outgroup. Hierarchical Orthologous Groups (HOGs) were then determined by OrthoFinder (v2.5.4) using this rooted tree. A final phylogeny was inferred on the concatenated alignment of single copy core HOGs by IQ-TREE (Nguyen et al. 2015) with the LG+F+G4 model and options -bb 5000 -bi 200.

Selection pressure on genes was estimated by the ratios of non-synonymous substitution rates (dN) to synonymous substitution rate (dS), calculated by codeml of the PAML v4.9 package (Yang 1997). All pairs of single copy orthologues as defined by OrthoFinder were retrieved and aligned with T-Coffee (Notredame et al. 2000). Codeml was given the sequence pairs alignments and the resulting dN/dS ratio was considered only if $dS < 1.5$, $dS > 0.1$ and $dN/dS < 10$. Later, the dN and dS values for each gene was estimated as the mean of all value calculated on gene pairs.

Estimation of cedratviruses core and pan-genomes

Core/pan-genomes sizes were calculated on HOGs (Hierarchical Orthologous Groups) at the root node. Genomes were iteratively added with all possible combinations to simulate a dataset with 1 to 9 genomes. We used the presence/absence matrix of HOGs instead of gene counts as in the original method (Tettelin et al. 2005). Data were processed using R (v4.04 (R Core Team 2021)).

For comparison, the ORF predictions, orthology analyses and core/pan-genome estimations were performed on other viral families: *Pandoraviridae* (Table S1C), *Marseilleviridae* (Table S1D), ranaviruses (Table S1E), *Megavirinae* (Table S1F). The outgroups used were respectively *Mollivirus sibericum*, *Ambystoma tigrinum virus*, *Red seabream iridovirus* and *Chrysochromulina ericina virus*.

Identification of horizontal gene transfers

HGTs were identified based on phylogenetic trees of each HOG complemented with homologous sequences that were retrieved using a two steps procedure. First, the sequences of each HOG were aligned using DIAMOND BLASTP (Buchfink et al. 2015) against the RefSeq database (from March 2019 (O'Leary et al. 2016)) with an e-value threshold of 1e-5, keeping only matches covering more than 50 % of the query. Up to 10 matches per domain (Bacteria, Archaea, Eukaryota and Viruses) were kept for each query and CD-hit was applied on the retrieved sequences. Secondly, the resulting sequences were queried again against the RefSeq using DIAMOND with the same e-value threshold. A maximum of two proteins per domain, whose matches covered more than 80 % of the query, were kept at this point. The HOGs and selected sequences from the first and second rounds were aligned using MAFFT v7.475 (Katoh and Standley 2013) and phylogenetic trees were built using IQ-TREE with options -bb 1000 -bi 200 -m TEST. Each resulting phylogenetic tree was rooted by mad v2.2 (Tria et al. 2017). Trees were finally visually inspected and HGT events counted when one or several *Pithoviridae* genes were within a bacterial, eukaryotic, archaeal or different viral clade.

Detection and classification of genomic repeats

A pipeline was developed to retrieve repeat-rich regions and map individual repeats from pithoviruses' genomes. The steps were: (1) genome-wide alignment, (2) flattened dotplot calculation, (3) repeat-rich regions mapping, (4) individual repeats retrieval, (5) repeat clustering.

(1) genomes were aligned against themselves by BLASTN with an e-value threshold of 1e-10.
(2) for each position of the genome, the number of times it was aligned was counted resulting in a vector (y); similar to a flattened dotplot.
(3) A smooth vector (y_{Ss}) was first estimated by sliding mean filtering with a window size of 500 nt. A detection threshold (τ) was calculated as $\tau = \overline{ys} * sensitivity^{-1}$, with a sensitivity coefficient set to 2.5. Repeat-rich regions were detected by comparing the vector y_s to τ. Repeat-rich regions were defined as regions where y_s is above the threshold τ. Each region's start and stop are thus the positions of intersections of y_s and τ.
(4) For each previously detected region, individual repeats were extracted using a smoothed derivative of y. Smoothing was applied before and after the derivation, this time with a window size of 20 nt. Then, the absolute value was taken in order to obtain the vector $|y_s|$. Then the local maxima were considered as repeat delimitations if above a cutoff set to 10.
(5) repeats are globally aligned to each other by needle of the EMBOSS suite (Rice et al. 2000). They are then ordered according to the mean distance (100 – needle identity percentage) to their 10 closest neighbors. The first sequence becomes a reference sequence. Then, sequences are clustered together if they are at least 70 % identical to a reference or they become themselves a reference. Finally, clusters are merged together if over half of their respective sequences are at least 70 % identical. For visual inspection to infer repeat types and similarity in-between clusters, a matrix of dotplots presenting the alignments of reference sequences is drawn.

For an in-depth analysis of pithoviruses' MITEs, the sequences from the cluster of repeats corresponding to the main MITE were aligned with MAFFT and sequences were trimmed were most had their terminal inverted repeat TA. The reference sequence (see step 5) of M1 and M2 were folded by mFold (Zuker 2003). To retrieve divergent M1 and M2 clusters, the dotpots of reference sequences was visually inspected. Reference sequences aligned to the reference of M1 or M2 clusters were annotated as M1 or M2-like (example given by cluster 3 in step 5, Fig. 1).

MUST v2-4-002 (Ge et al. 2017) and MITE-Tracker (Crescente et al. 2018) were used to confirm the nature of the repeats.

Acknowledgments

This work was supported by the Agence Nationale de la Recherche grant (ANR-22-CE12-0041) to ML, (ANR-10-INBS-09-08) to J-MC and CNRS Projet de Recherche Conjoint (PRC) grant (PRC1484-2018) to CA. S.R. was supported by a doctoral fellowship obtained from Aix-Marseille University. We thank the PACA Bioinfo platform for computing support and Hugo Bisio for carefully reading the manuscript.

Availability of Data and Materials

Genome sequences and annotations of the following five *Pithoviridae* have been deposited to GenBank: *cedratvirus borely* (OQ413575), *cedratvirus plubellavi* (OQ413576), *cedratvirus lena* (OQ413577, OQ413578, OQ413579, OQ413580), *cedratvirus duvanny* (OQ413581) and *pithovirus mammoth* (OQ413582).

The code for pithovirus repeats detection and clustering is available at: https://src.koda.cnrs.fr/igs/genome-repeats-detection.git.

None of the authors have any competing interests.

References

Alempic J-M, Lartigue A, Goncharov AE, Grosse G, Strauss J, Tikhonov AN, Fedorov AN, Poirot O, Legendre M, Santini S, et al. 2023. An Update on Eukaryotic Viruses Revived from Ancient Permafrost. *Viruses* 15:564.

Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, Sedlazeck FJ, Lippman ZB, Schatz MC. 2019. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biol.* 20:224.

Al-Shayeb B, Sachdeva R, Chen L-X, Ward F, Munk P, Devoto A, Castelle CJ, Olm MR, Bouma-Gregson K, Amano Y, et al. 2020. Clades of huge phages from across Earth's ecosystems. *Nature* 578:425–431.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403–410.

Andreani J, Aherfi S, Khalil JYB, 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-Basel* 8:300.

Andreani J, Khalil JYB, Baptiste E, Hasni I, Michelle C, Raoult D, Levasseur A, La Scola B. 2018. Orpheovirus IHUMI-LCC2: A New Virus among the Giant Viruses. *Front. Microbiol.* 8:2643.

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* [Internet] 10. Available from: https://mbio.asm.org/content/10/2/e02497-18

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

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.

Bisio H, Legendre M, Giry C, Philippe N, Alempic J-M, Jeudy S, Abergel C. 2022. Evolution of giant pandoravirus from small icosahedral viruses revealed by CRISPR/Cas9. Available from: https://hal.archives-ouvertes.fr/hal-03810585

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

Boyer M, Yutin N, Pagnier I, Barrassi L, Fournous G, Espinosa L, Robert C, Azza S, Sun S, Rossmann MG, et al. 2009. Giant Marseillevirus highlights the role of amoebae as a melting pot in emergence of chimeric microorganisms. *Proc. Natl. Acad. Sci.* 106:21848–21853.

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12:59–60.

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.

Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, et al. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinforma. Oxf. Engl.* 25:1422–1423.

Crescente JM, Zavallo D, Helguera M, Vanzetti LS. 2018. MITE Tracker: an accurate approach to identify miniature inverted-repeat transposable elements in large genomes. *BMC Bioinformatics* 19:348.

Darling AE, Mau B, Perna NT. 2010. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PloS One* 5:e11147.

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

Filée J, Chandler M. 2008. Convergent mechanisms of genome evolution of large and giant DNA viruses. *Res. Microbiol.* 159:325–331.

Ge R, Mai G, Zhang R, Wu X, Wu Q, Zhou F. 2017. MUSTv2: An Improved De Novo Detection Program for Recently Active Miniature Inverted Repeat Transposable Elements (MITEs). *J. Integr. Bioinforma.* 14:/j/jib.2017.14.issue-3/jib-2017-0029/jib-2017-0029.

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

Jiang N, Wessler SR. 2001. Insertion Preference of Maize and Rice Miniature Inverted Repeat Transposable Elements as Revealed by the Analysis of Nested Elements [W]. *Plant Cell* 13:2553–2564.

Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al. 2014. InterProScan 5: genome-scale protein function classification. *Bioinforma. Oxf. Engl.* 30:1236–1240.

Käll L, Krogh A, Sonnhammer ELL. 2004. A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.* 338:1027–1036.

Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* 30:772–780.

Legendre M, Bartoli J, Shmakova L, Jeudy S, Labadie K, Adrait A, Lescot M, Poirot O, Bertaux L, Bruley C, et al. 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.

Legendre M, Fabre E, Poirot O, Jeudy S, Lartigue A, Alempic J-M, Beucher L, Philippe N, Bertaux L, Christo-Foroux E, et al. 2018. Diversity and evolution of the emerging Pandoraviridae family. *Nat. Commun.* 9:2285.

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.

Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Marchler GH, Song JS, et al. 2020. CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res.* 48:D265–D268.

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

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

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.* 32:268–274.

Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* 302:205–217.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 44:D733-745.

Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. 2020. Using SPAdes De Novo Assembler. *Curr. Protoc. Bioinforma.* 70:e102.

R Core Team. 2021. R: A language and environment for statistical computing. Vienna, Austria Available from: https://www.R-project.org/

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet. TIG* 16:276–277.

Rigou S, Santini S, Abergel C, Claverie J-M, Legendre M. 2022. Past and present giant viruses diversity explored through permafrost metagenomics. *Nat. Commun.* 13:5853.

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.Sandri-Goldin RM, editor. *J. Virol.* 92:e00372-18, /jvi/92/13/e00372-18.atom.

dos Santos Silva LK, dos Santos Pereira Andrade AC, Dornas FP, Lima Rodrigues RA, Arantes T, Kroon EG, Bonjardim CA, Abrahao JS. 2018. Cedratvirus getuliensis replication cycle: an in-depth morphological analysis. *Sci. Rep.* 8:4000.

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.

Senkevich TG, Yutin N, Wolf YI, Koonin EV, Moss B. 2021. Ancient Gene Capture and Recent Gene Loss Shape the Evolution of Orthopoxvirus-Host Interaction Genes. *mBio* 12:e0149521.

Sun C, Feschotte C, Wu Z, Mueller RL. 2015. DNA transposons have colonized the genome of the giant virus Pandoravirus salinus. *BMC Biol.* [Internet] 13. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4495683/

Tarchini R, Biddle P, Wineland R, Tingey S, Rafalski A. 2000. The complete sequence of 340 kb of DNA around the rice Adh1-adh2 region reveals interrupted colinearity with maize chromosome 4. *Plant Cell* 12:381–391.

Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, et al. 2005. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome." *Proc. Natl. Acad. Sci.* 102:13950–13955.

Tria FDK, Landan G, Dagan T. 2017. Phylogenetic rooting using minimal ancestor deviation. *Nat. Ecol. Evol.* 1:1–7.

Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLOS Comput. Biol.* 13:e1005595.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Bioinformatics* 13:555–556.

Yuan Y, Gao M. 2017. Jumbo Bacteriophages: An Overview. *Front. Microbiol.* 8:403.

Zhang H-H, Zhou Q-Z, Wang P-L, Xiong X-M, Luchetti A, Raoult D, Levasseur A, Santini S, Abergel C, Legendre M, et al. 2018. Unexpected invasion of miniature inverted-repeat transposable elements in viral genomes. *Mob. DNA* 9:19.

Zhang XY, Feschotte C, Zhang Q, Jiang N, Eggleston WB, Wessler SR. 2001. P instability factor: An active maize transposon system associated with the amplification of Tourist-like MITEs and a new superfamily of transposases. *Proc. Natl. Acad. Sci. U. S. A.* 98:12572–12577.

Zuker M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* 31:3406–3415.

Supplementary material

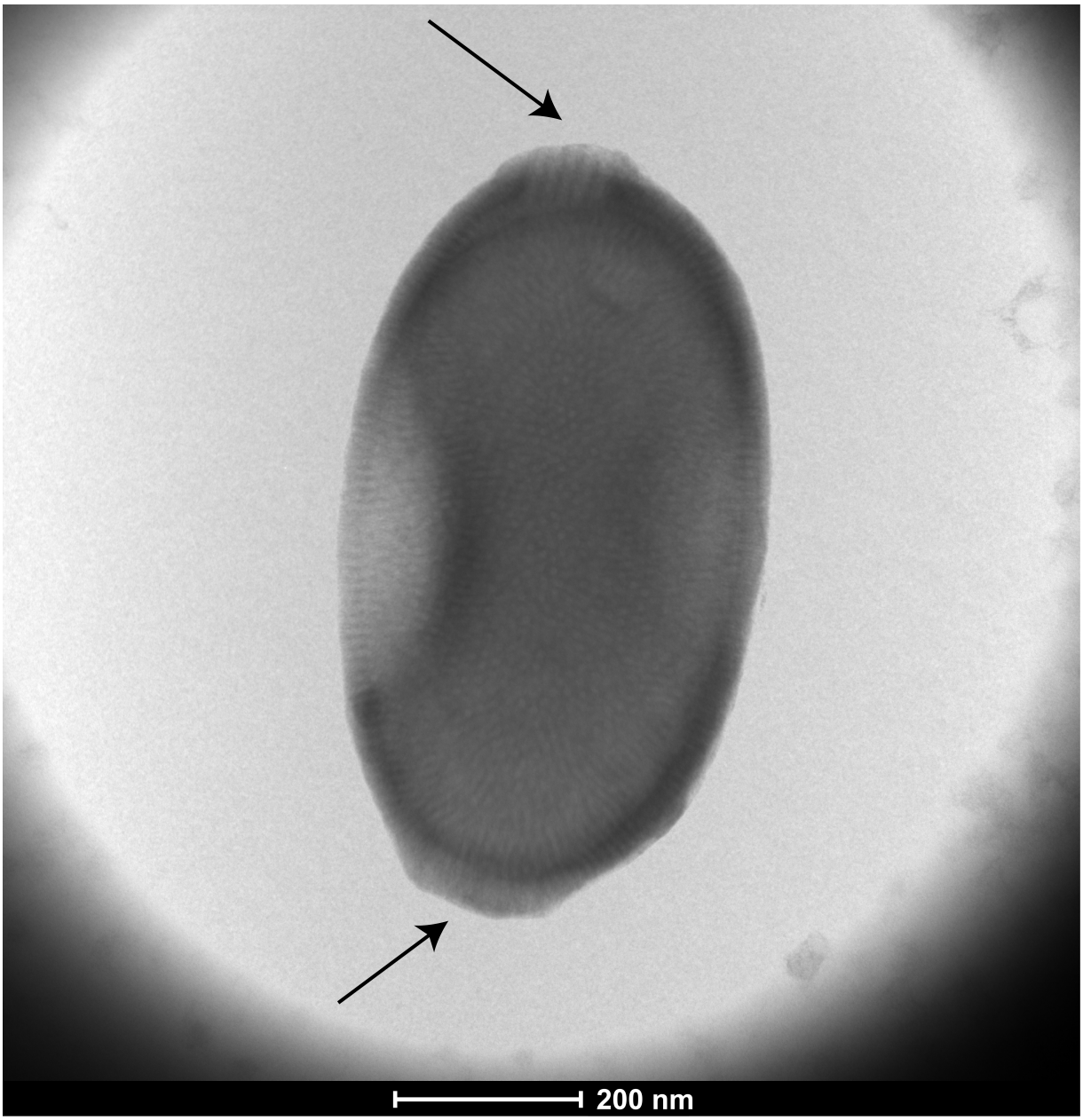


Figure S1. Negative staining microscopy of cedratvirus plubellavi
Corks at each apex of the viral particle are shown with arrows.

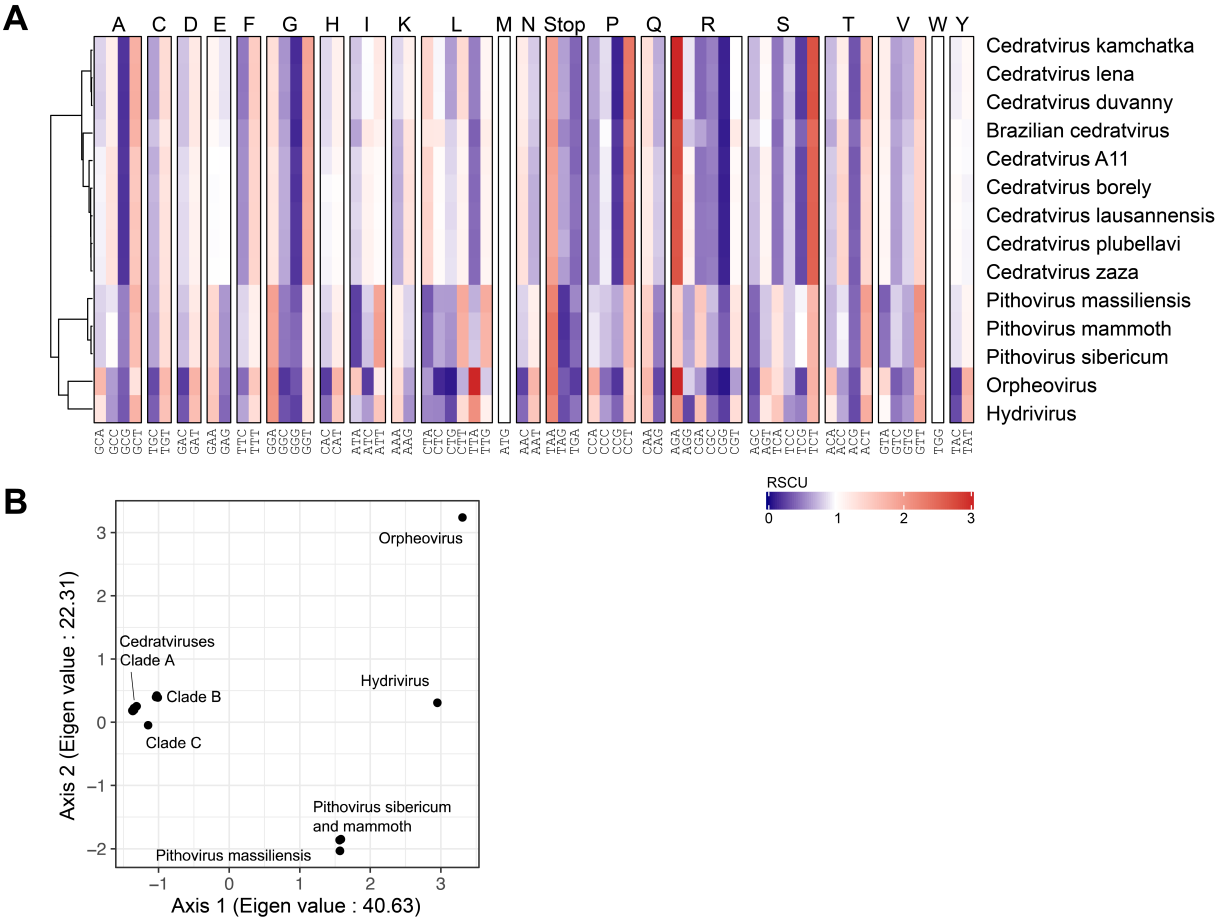


Figure S2. Comparison of relative synonymous codon usages

(A) Codon usage bias for each amino acid represented by the RSCU value. (B) PCOA analysis of the RSCU values without the stop codons, the tryptophan and the methionine codons.

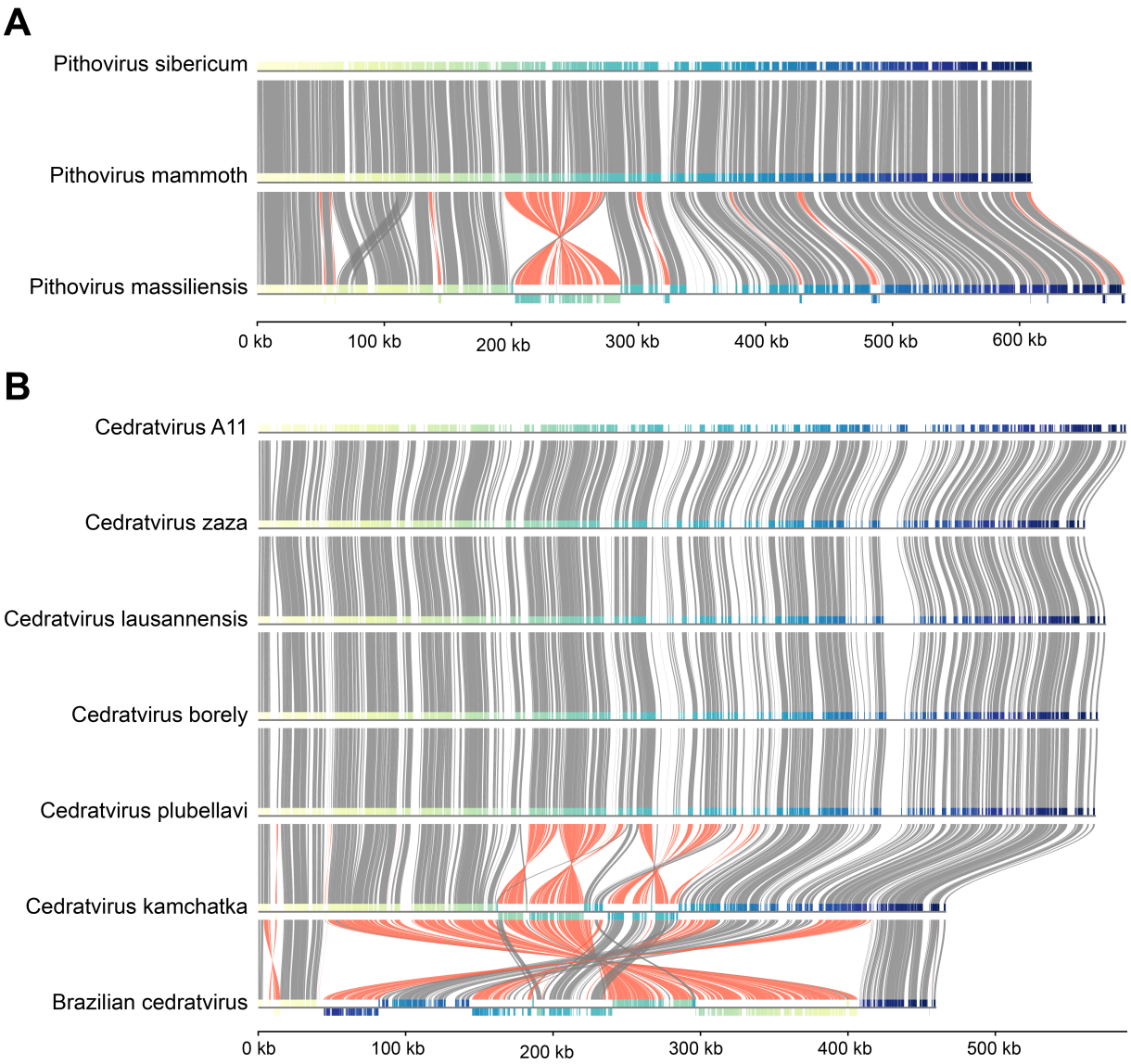


Figure S3. Genome alignment of *Pithoviridae*

Shared nucleotide sequence blocks within families were drawn based on the alignment by progressive-mauve (Darling et al. 2010) of (A) the three pithoviruses and (B) seven cedratviruses. *Cedratvirus lena* and *Cedratvirus duvanny* have been excluded since the assembly with incomplete (multiple contigs). Syntenic regions are shown in gray and large inversions in red. ORFs are color-coded from yellow to blue according to their genomic positions. A scale bar of genome sizes is shown at the bottom.

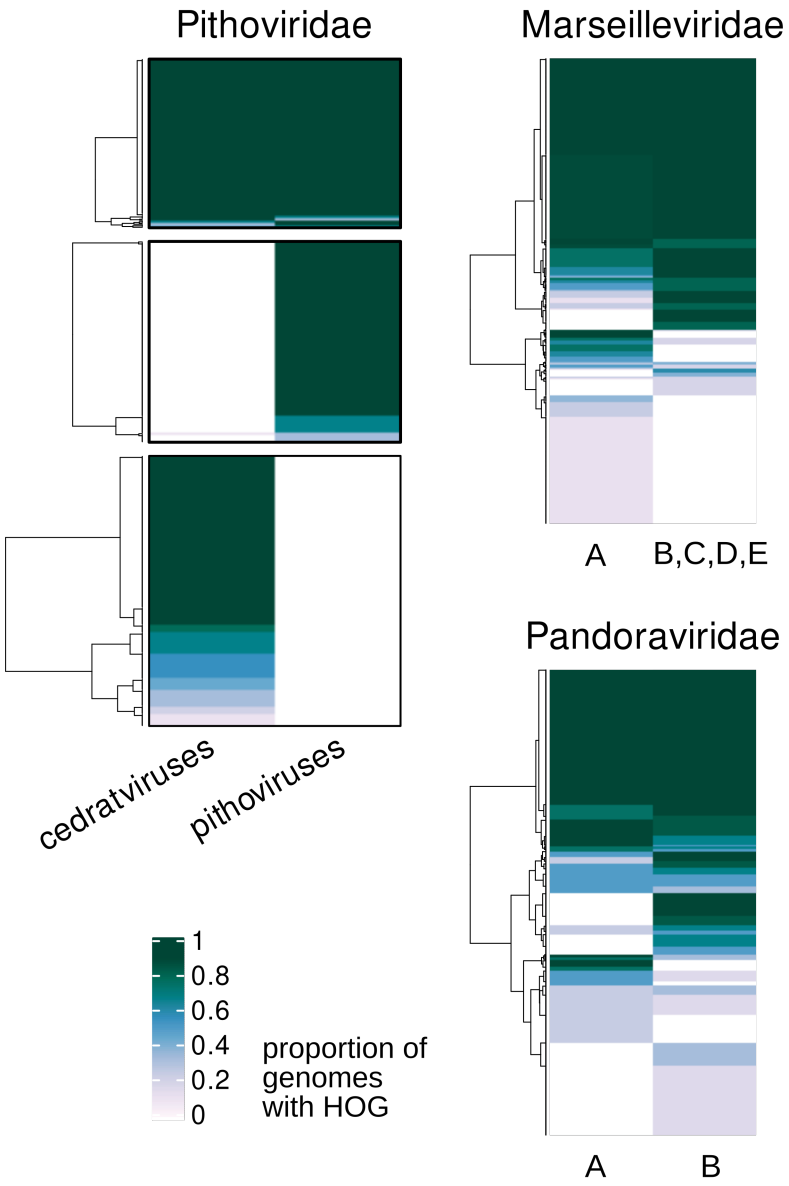


Figure S4. Patterns of presence/absence of HOGs within viral family's sub-clades
Species were divided into clades ignoring the outgroup. The number of species from each clade that appeared in each HOG was then counted.

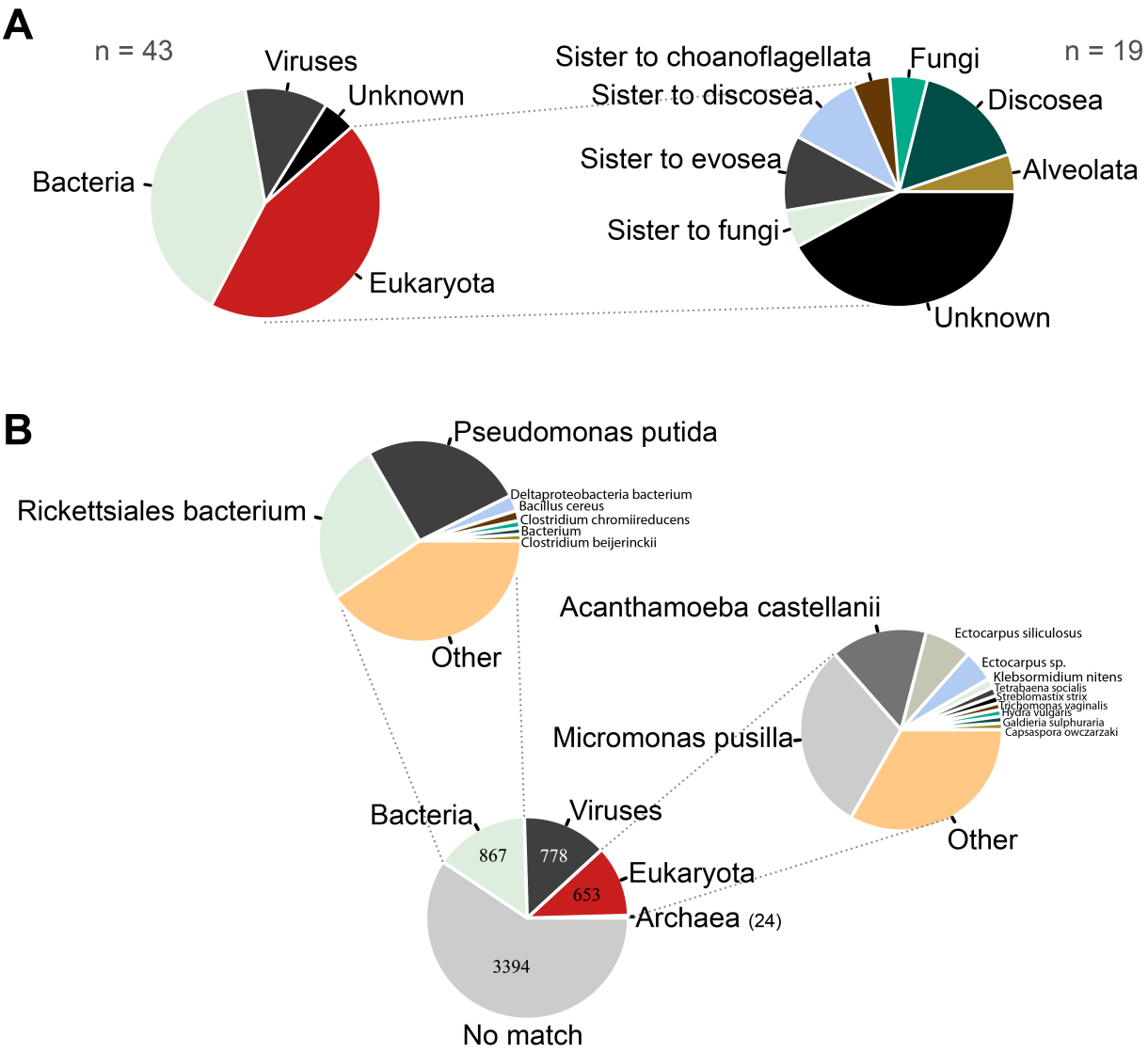


Figure S5. Horizontal Gene Transfer events in *Pithoviridae* and BLASTp control

For each HGT event, the likely origin as estimated from the visualization of phylogenetic trees (A) and best BLASTP results (*E*value ≤ 10^{-5}) from the nr database free of *Pithoviridae* (B) are shown. From eukaryotes, "sister to" is short for "sister group of…". Bacterial and eukaryotic species with more than 1% of matches in their respective category are shown.

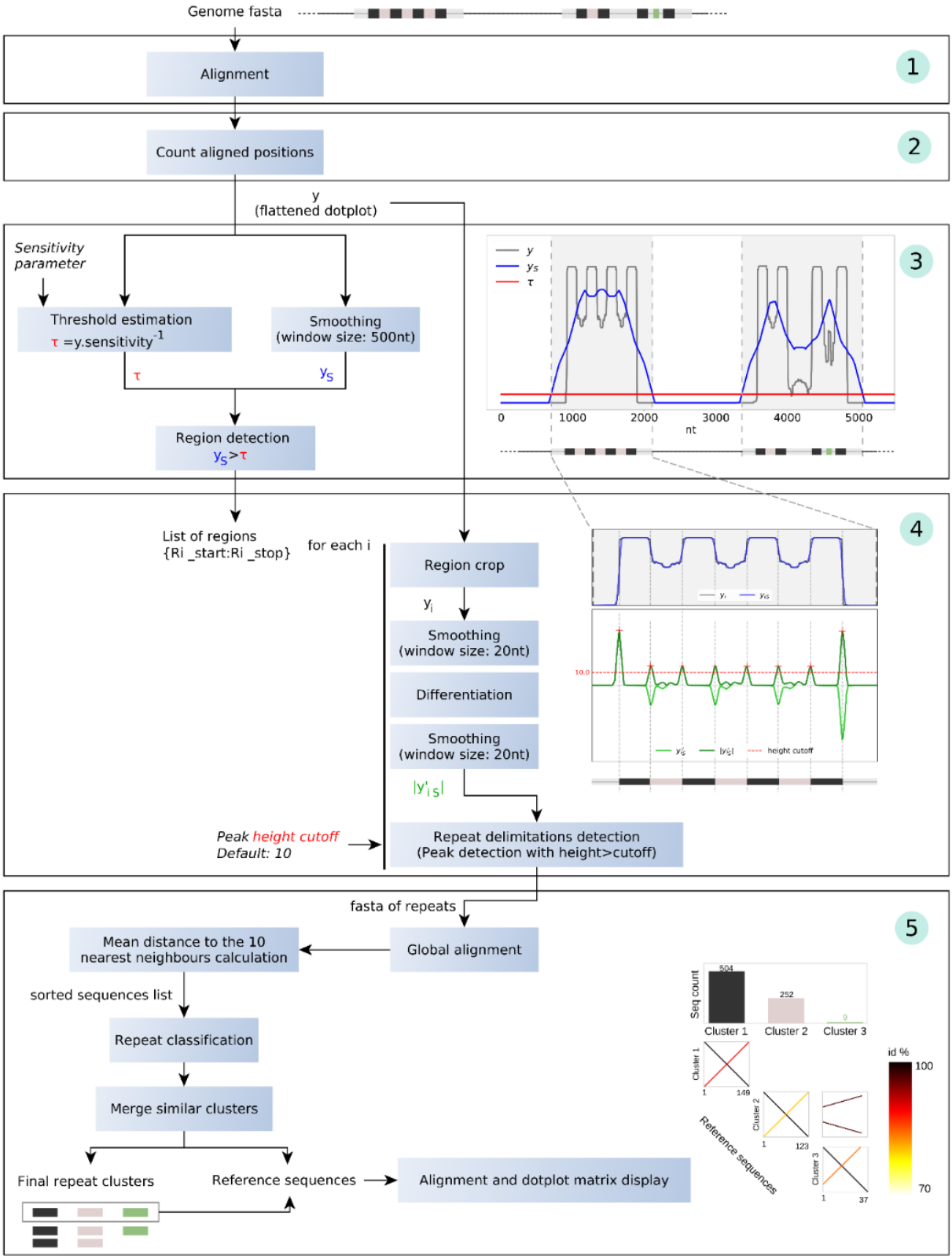


Figure S6. Workflow for repeat analysis

Steps one to five are represented within large boxes. Operations are in blue boxes while objects are shown as black text. Besides "Genome fasta" is schematized a portion of the genome containing repeats as colored boxes. The slightly grey boxed represent unclustered sequences.

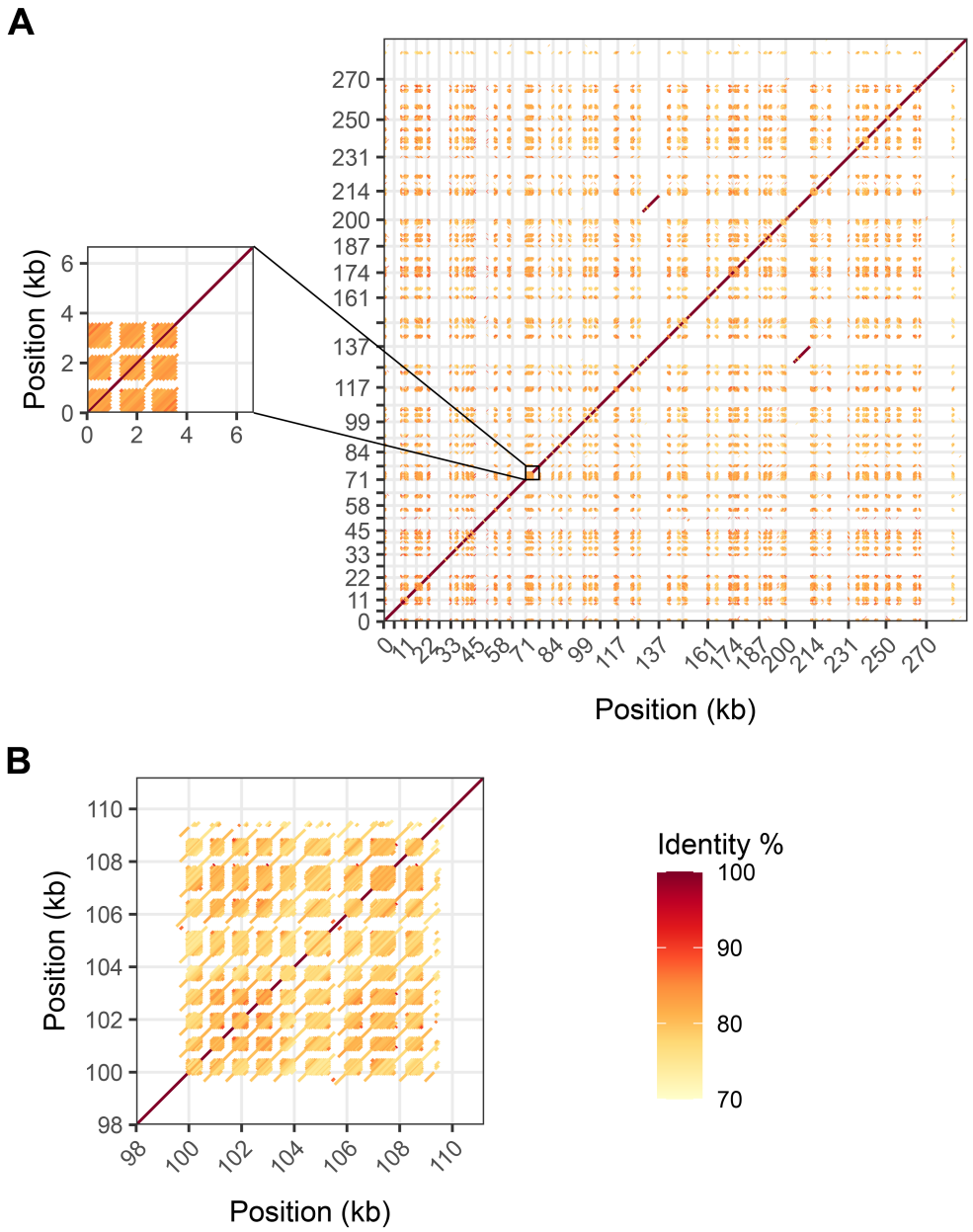


Figure S7. Repeats found in *Pithoviridae*-like metagenomes

(A) *Pithovirus LCPAC302* (Bäckström et al. 2019) presents numerous direct repeats. In some rare cases, these repeats are interspersed by a similar sequence as shown in the zoom. X-axis and y-axis breaks correspond to the delimitation of contigs. (B) Regularly interspersed direct repeats from a permafrost *Pithoviridae*-like metagenome (K_bin2137_k1) (Rigou et al. 2022).

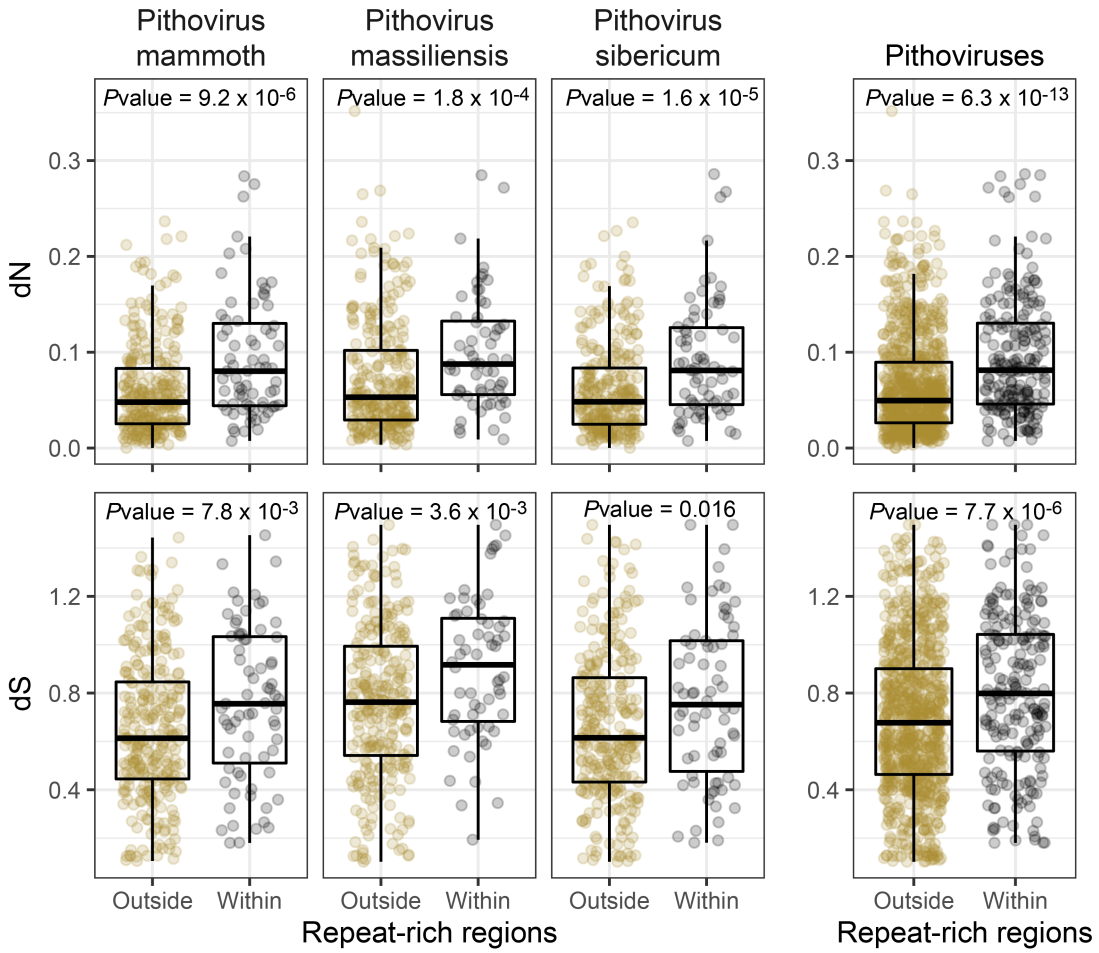


Figure S8. Mutation rates in pithoviruses repeat-rich regions

The dN and dS values were calculated from orthologous single copy genes and divided into two groups in respect to the repeated regions given by our repeat pipeline. The given *P*values were calculated using Wilcoxon rank tests.

Table S1. Assemblies used for comparative genome size analysis

A) Previously published	**NCBI accessions**	**E) Ranaviruses**	
Pithovirus sibericum	NC_023423.1	Ambystoma tigrinum virus	GC_ 000841005.1
Pithovirus massiliensis	SAMEA4074172	Bohle iridovirus	GCF_002826565.1
Cedratvirus A11	NC_032108.1	Common midwife toad virus	GCF_003033105.1
Cedratvirus lausannensis	LT907979.1	Epizootic haematopoietic necrosis virus	GCF_000897115.1
Cedratvirus zaza	LT994652.1	European catfish virus	GCF_000897115.1
Brazilian cedratvirus	LT994651.1	Frog virus 3	GCF_001717415.1
Cedratvirus kamchatka	MN873693.1	Infectious spleen and kidney necrosis virus	GCF_000848865.1
Orpheovirus (outgroup)	NC_036594.1	Lymphocystis disease virus 1	GCF_000839605.1
Hydrivirus (outgroup)	GCA_943296135.1	Lymphocystis disease virus-isolate China	GCF_000844885.1
Marseillevirus (outgroup)	NC_013756.1	Lymphocystis disease virus Sa	GCF_001974475.1
		Ranavirus maximus	GCF_001717415.1
B) New *Pithoviridae*		Largemouth bass virus	GCA_013122655.1
Cedratvirus borely	OQ413575	Scale drop disease virus	GCF_001274405.1
Cedratvirus plubellavi	OQ413576	Short-finned eel ranavirus	GCF_001678255.2
Cedratvirus lena	OQ413577 OQ413578 OQ413579 OQ413580	Singapore grouper iridovirus	GCF_000846905.1
Cedratvirus duvanny	OQ413581	Grouper iridovirus	GCA_006465545.1
Pithovirus mammoth	OQ413582	Red seabream iridovirus (outgroup)	GCA_011894875.1

C) *Pandoraviridae*		**F) *Megavirinae***	
Pandoravirus braziliensis	LT972217.1	Acanthamoeba polyphaga lentillevirus	GCA_000320725.1
Pandoravirus celtis	MK174290.1	Mamavirus	GCA_002966335.1
Pandoravirus dulcis	GCA_000911655.1	Megavirus chilensis	GCF_000893915.1
Pandoravirus inopinatum	GCA_000928575.1	Megavirus courdo7	GCF_000893915.1
Pandoravirus macleodensis	GCA_003233935.1	Megavirus vitis	GCA_004156275.1
Pandoravirus massiliensis	MZ384240.1	Mimivirus	GCA_024266865.1
Pandoravirus neocaledonia	GCA_003233915.1	Moumouvirus australiensis	GCA_004156295.1
Pandoravirus pampulha	OFAJ00000000.1	Moumouvirus	GCF_000904035.1
Pandoravirus quercus	GCA_003233895.1	Tupanvirus deep ocean	GCA_002966475.2
Pandoravirus salinus	GCA_000911955.1	Tupanvirus soda lake	GCA_002966485.2
Mollivirus sibericum (outgroup)	NC_027867.1	Chrysochromulina ericina virus (outgroup)	GCF_001399245.1

D) *Marseilleviridae* as in Blanca et al., 2020 (doi: 10.3390/v12111270)	
Marseillevirus	GU071086
Lausannevirus	HQ113105
Cannes 8 virus	KF261120
Insectomime virus	HG428764
Tunisvirus	KF483846
Brazilian marseillevirus	KT752522
Melbournevirus	KM275475
Port-miou virus	KT428292
Tokyovirus	Reassembled in (Blanca et al. 2020)
Noumeavirus	KX066233
Golden marseillevirus	KT835053
Kurlavirus	KY073338
Marseillevirus shanghai	MG827395
Ambystoma tigrinum virus (outgroup)	MK580533.2

Table S2. Pithoviruses' MITEs occurrences

A region is defined as a genomic sequence with a high density of repeats within a sliding window of 500 bp. Within each region, the number of M1 and M2 MITEs was counted. The clusters containing divergent M1 and M2 sequences were included in these results.

		P. sibericum		P. mammoth		P. massiliensis	
		M1	**M2**	**M1**	**M2**	**M1**	**M2**
Regions	**Total**	110	100	109	100	115	79
	M1 or M2	10	0	9	0	36	0
Per region	**Min count**	1	1	1	1	1	1
	Max count	11	12	13	17	13	8
	Mean	4.68	3.71	4.58	4	5.05	3.01
	Sd	2.12	2.26	2.31	3.06	2.98	1.64

Table S3. Genomic rearrangements and mutations between *Pithovirus sibericum* and *Pithovirus mammoth*

	Rearrangements types							Total orthologous pairs with rearrangement events	Conserved orthologous pairs without rearrangement event
Repeats regions	Insertions/ deletions	Single nucleotide insertions/ deletion	Substitutions	Inversions	Duplications in P. sibericum	Duplication in P. mammoth	Complex events		
Within	9	5	2	5	1	4	2	228	109
Outside	5	2	1	1	1	3	0	13	41
								Chi^2 *P*value = 5.5×10^{-7}	