Microbes Environ. 41(1), 2026
https://www.jstage.jst.go.jp/browse/jsme2 doi:10.1264/jsme2.ME25085

Genome-wide Characterization of Non-shared Sequences among Amphora-shaped Giant Viruses

Motohiro Akashi[1]*, Masaharu Takemura[2], and Seiichi Suzuki[1]

[1]*Department of Science and Technology, Faculty of Science and Technology, Seikei University, 3–3–1 Kichijojikitamachi, Musashino city, Tokyo 180–8633, Japan; and* [2]*Graduate School of Science, Tokyo University of Science, 1–3 Kagurazaka, Shinjuku, Tokyo 162–8601, Japan*

(Received November 28, 2025—Accepted February 23, 2026—Published online March 18, 2026)

OPEN ACCESS

Citation: Akashi, M., Takemura, M., and Suzuki, S. (2026) Genome-wide Characterization of Non-shared Sequences among Amphora-shaped Giant Viruses. *Microbes Environ* **41**: ME25085.
https://doi.org/10.1264/jsme2.ME25085

***Corresponding author.** E-mail: motohiro-akashi@st.seikei.ac.jp

Copyright: © Japanese Society of Microbial Ecology / Japanese Society of Soil Microbiology / Taiwan Society of Microbial Ecology / Japanese Society of Plant Microbe Interactions / Japanese Society for Extremophiles

https://creativecommons.org/licenses/by/4.0/

Giant viruses are distinguished not only by their large particle size, but also by their extensive genomes, often reaching megabase levels. Many sequences within these genomes are considered to have been introduced by hosts, surrounding organisms, or other viruses. Since the natural hosts of many giant viruses remain unidentified, analyzing sequences potentially derived from other organisms may aid in clarifying their hosts. In the present study, we identified eukaryote-homologous sequences by isolating those not shared among viruses, an aspect previously overlooked. Our primary focus was on pandoravirus, which, with a genome size of ~2 Mb, is the largest among giant viruses. We obtained 375 BLAST hits with an average sequence identity of ~90%. Among the 102 detected species, those with higher hits included *Mus musculus*, *Lampetra planeri*, *Melanogrammus aeglefinus*, *Lampetra fluviatilis*, *Scylla paramamosain*, *Cardiocondyla obscurior*, *Monodelphis domestica*, *Vespula pensylvanica*, *Micromonas pusilla*, *Physcomitrium patens*, *and Peromyscus californicus*. Similar analyses of *Cedratvirus* and *Pithovirus*, which share an amphora-shaped particle structure with pandoraviruses, yielded fewer data (48 and 5 hits, respectively), with no common taxa at the order level. Thirteen BLAST hits exceeded 100 bp, including conserved non-coding elements (CNEs) in fish and other taxa, along with sequences of unknown functions. These results indicate the presence of short regions with sequence similarity in non-shared sequences, although direct host identification proved difficult.

Key words: giant virus, Pandoravirus, comparative genomic analysis, *Nucleocytoviricota*

Giant viruses possess significantly larger genome sequences than those of other viruses (Bosmon *et al.*, 2025). Viruses often possess highly sophisticated short genomes that enable them to replicate their copies rapidly and cost-effectively. They sometimes harbor auxiliary metabolic genes (AMGs), which are known to modify or confer functions in host metabolism (Breitbart, 2012). For example, MazG is considered to be involved in transcriptional regulation and the control of DNA replication during starvation in bacteria. However, because the *mazG* gene in phages is closely related to a different bacterium than that of the host cyanobacteria, genes acquired by the virus and genomic fragments do not necessarily originate from the primary host (Bryan *et al.*, 2008). Although the natural hosts of giant viruses have not been captured thus far, the core genes shared by giant viruses are considered to have been acquired independently by different viral families from different species, while their genomes have grown in length, and their association with natural hosts is suspected (Yutin *et al.*, 2014). Major virion protein 2 is another good example of horizontal gene transfer (HGT), which is widely conserved among a member of the giant virus, pandoraviruses. However, this factor was found to be derived from inactivated glycoside hydrolase 16 originally encoded in bacteria (Krupovic *et al.*, 2020). Long genomes contain not only genes that are functional for the giant virus and host, but also self-serving sequences, such as transfer factors (Aherfi *et al.*, 2018). Furthermore, interviral HGT has been reported in giant viruses (Wu *et al.*, 2024). Since some giant viruses possess the ability to inhibit superinfection, it is plausible that specific conditions must be satisfied for viral HGT to occur (Aquino *et al.*, 2024).

Among giant viruses, pandoraviruses are notable for their extensive genomes, with the largest reaching 2.5 Mb (Philippe *et al.*, 2013). Furthermore, pandoraviruses exhibit markedly low genome conservation within the same genus, as evidenced by a core genome/pangenome ratio <0.05 (Aherfi *et al.*, 2018). However, it is important to note that not all giant viruses possess genomes of several megabases. For example, the genome sizes of pithoviruses and cedratviruses, which feature amphora-shaped viral particles similar to those of pandoraviruses, are approximately 600 kb (Legendre *et al.*, 2014; Andreani *et al.*, 2016). In an experiment involving the genetic engineering of a segment of the

pandoravirus genome, a 60-kb deletion was successfully achieved, but resulted in a reduced growth rate. This finding suggests that the giant virus has undergone gene-accordion evolution, consistent with the hypothesis that giant viruses have experienced genome enlargement through acquisition (Elde *et al.*, 2012; Filée, 2015; Bisio *et al.*, 2023). The genomic accordion-like model proposes that giant virus genomes expanded through a continuous process of gene acquisition and loss. This model was proposed by Jonathan Filée in 2013 and subsequent comparative genomic analyses confirmed that genome size expansion results from mutations including gene acquisition, gene loss, and gene duplication (Filée, 2013, 2015). Conversely, an experimental evolution of *Mimivirus* showed a reduction in genome size from 1.2 to 0.99 Mb after 150 passages with *Acanthamoeba* under experimental conditions (Boyer *et al.*, 2011).

Research on the presence of core genes of giant viruses within eukaryotic genomes or transcriptome datasets has identified one or more homologous genes across 66 distinct datasets (Gallot-Lavallée and Blanc, 2017). It is hypothesized that the core genes of giant viruses are acquired through HGT with proto-eukaryotes during an earlier period, coinciding with the establishment of the Last Eukaryotic Common Ancestor (LECA) (Guglielmini *et al.*, 2019). The genome of *Acanthamoeba castellanii* strain Neff revealed 267 genes originating from various giant viruses. However, a molecular phylogenetic analysis of these genes suggests that the specific giant virus from which they are directly derived remains unidentified (Maumus and Blanc, 2016). Furthermore, analyses of HGT using global metagenomic data have indicated that giant viruses are capable of infecting major eukaryotic organisms (Schulz *et al.*, 2020). These findings suggest that giant viruses have engaged in a co-evolutionary relationship with HGT since the advent of eukaryotes.

In the present study, we provided insights into the recent interactions between giant viruses and eukaryotes, with a focus on identifying traces of the host and its associated organisms within the viral genome. Specifically, we conducted a comparative analysis of pandoravirus genomes, isolating sequences that were not shared among them, and assessed eukaryotic species compositions using a BLAST analysis. Our objective was to catalog organisms related to these viruses by examining the species composition of eukaryotic organisms using a BLAST analysis. We focused on non-shared sequences because the presence of homologous sequences in multiple species, despite their absence in pithoviruses and cedratviruses, suggests that these regions were recently acquired prior to isolation and culture. Given that the morphology of viral particles may reflect the virus's evolutionary ecological strategy, we extended our analysis to pithoviruses and cedratviruses, which although phylogenetically distant from pandoraviruses, exhibit similar particle shapes. We compared these results with those obtained for pandoraviruses.

Materials and Methods

Viruses targeted for analysis

The genomes of amphora-shaped giant viruses, as obtained from the National Center for Biotechnology Information (NCBI), comprise nine pandoraviruses, four pithoviruses, and five cedratviruses. The accession numbers of each genome are listed in Table S1. In addition, we conducted a genomic analysis of two pandoravirus strains isolated by Akashi and Takemura (2019), six pandoravirus strains and one pithovirus and one cedratvirus strain each isolated by Akashi *et al.* (2024). Therefore, the genomes of 17 pandoraviruses, 4 pithoviruses, and 7 cedratviruses—a total of 28 isolates—were utilized in the analysis. As a comparative reference, we analyzed the non-shared regions of five *Nucleocytoviricota* viruses exhibiting icosahedral symmetry: *Mimivirus*, 15; *Moumouvirus*, 6; *Megavirus*, 12; *Faustovirus*, 17; *Marseillevirus*, 29 (Table S1).

Whole-genome analysis of amphora-shaped viruses

The present study analyzed the complete genomic sequences of amphora-shaped viruses already reported (Akashi *et al.*, 2024). These viruses were isolated following the infection of *Acanthamoeba castellanii* cultured in PYG medium, which comprised the following components: 4 mM $CaCl_2$, 4 mM $MgSO_4 \cdot 7H_2O$, 2.5 mM $Na_2HPO_4 \cdot 7H_2O$, 2.5 mM KH_2PO_4, 0.1% (w/v) sodium citrate$\cdot 2H_2O$, 0.05 mM $Fe(NH_4)_2(SO_4)_2 \cdot 6H_2O$, and 100 mM glucose, with a pH of 6.5. Viral particles were harvested by centrifuging the virus-infected culture medium at 500×*g* for 1 min, followed by centrifugation of the resulting supernatant at 8,000×*g* for 35 min.

Viral DNA was extracted using phenol-chloroform-isoamyl alcohol (PCI) and ethanol precipitation. The procedure was as follows: the viral solution was centrifuged at 15,000×*g* for 1 min to remove the supernatant, followed by the addition of 400 µL TES buffer (25% [w/v] sucrose, 50 mM Tris-HCl [pH 7.6], and 20 mM EDTA [pH 8.0]), 20 µL lysozyme (100 mg mL^{-1}), and 4 µL RNase (10 mg mL^{-1}). The mixture was then incubated at 37°C for 30 min. Subsequently, 20 µL of 10% SDS solution and 4 µL of 20 mg mL^{-1} Proteinase K were added and the mixture was incubated at 37°C for an additional 30 min. Thereafter, 400 µL of PCI was added and the sample was vortexed. The sample was then centrifuged at 15,000×*g* for 5 min and the aqueous layer was collected. The PCI treatment was repeated twice. Ethanol precipitation was conducted as follows: a 1/10 volume of both 3M sodium acetate and 2.5× volume of 99% ethanol was added to the recovered DNA solution, followed by centrifugation at 15,000×*g* at 4°C for 5 min. After the supernatant was removed, 500 µL of 70% ethanol was added and the mixture was centrifuged at 15,000×*g* at 4°C for 5 min. The supernatant was removed and the sample was dried. One hundred microliters of Tris-EDTA Buffer (10 mmol L^{-1} Tris-HCl [pH 8.0] and 1 mM EDTA·2Na [pH 8.0]) was added and the sample was stored at 4°C.

A sequencing analysis was conducted in accordance with the manufacturer's guidelines using a Revio Polymerase Kit (PacBio) from Bioengineering Lab. Co., Ltd. Library preparation was performed using SMRTbell Prep Kit 3.0 (PacBio) and SMRTbell gDNA Sample Amplification Kit (PacBio). Ultra-low PCR adapters were excised from HiFi reads using lima (version 2.9.0). PCR duplicates were eliminated using pbmarkdup (version 1.0.3). Sequences under 1,000 bases were removed using Filtlong (version 0.2.1). Sequences of 1,000 bases or more were assembled using Flye (version 2.9.3-b1797) in meta-mode under default conditions. Gene prediction was performed using Prokka (version 1.14.6). The sequencing analysis was performed by Bioengineering Lab.

Non-shared region extraction

Homologous regions of the genomic DNA of interest were identified using an all-vs-all BLASTN (Nucleotide-Nucleotide BLAST 2.15.0+) analysis with default parameters (output format 6). The genome analysis utilized 17 pandoraviruses, 4 pithoviruses, and 7 cedratviruses, totaling 28 isolates. Based on the data obtained from the all-vs-all analysis, the coordinates of genomic regions not shared with any virus were examined for each virus genus using a custom Python script. Regions that were non-shared, with a length exceeding 100 bases, and devoid of the N nucleotide were extracted.

Online BLASTN search and visualization of eukaryotes with sequences homologous to unique regions

A comprehensive BLASTN search was conducted using the NCBI web-based BLAST interface against the core nucleotide (core_nt) database, with non-shared regions being used as queries (search date: 20 February 2025). The queries were submitted in batches as separated multi-FASTA files. BLASTN searches were performed with low-complexity filtering (DUST) and soft masking enabled, such that low-complexity regions were excluded from seed generation and allowed during alignment extension. The specified search parameters were as follows: organisms: eukaryota (taxid: 2759), program selection: somewhat similar sequences (BLASTN),% identity ≥80%, e-value <1e-4. Taxonomic information was derived from the subject sequence ID. The results obtained from the BLAST search were integrated with the taxonomy information of each query and subject and subsequently visualized using the Python program. Data were then subjected to the following filtering process. Regarding hits with identical query sequence IDs, filtering was performed sequentially based on the following priority order: the highest bit score, highest% identity, lowest e-value, and greatest alignment length. At each step, only hits meeting the optimal criterion were retained for the next filtering step. If multiple hits remained after applying all criteria, all of these hits were retained. Filtering was then applied across the entire dataset (not per query), retaining only taxonomic groups represented by two or more records at each rank. Filtering was performed sequentially at the kingdom, phylum, class, order, family, and genus levels, and records belonging to groups with only a single representative were excluded at each step before proceeding to the next rank. Using the detected non-shared region as the query sequence, homologous sequences were similarly searched for within viral genomes (Program: blastn, Percent Identity: 70–100%, Query Coverage: 80–100%, including Viruses (taxid:10239) and excluding Pandoravirus [taxid:2060084] data). At this stage, the Percent Identity threshold was lowered from 80%, which was used in other analyses, to 70% due to the low number of detections. The best hit sequences within non-shared regions exceeding 100 base pairs (bp) were identified via BLAST searches against the Protist 10,000 Genomes Project database (P10K) (Program: blastn, Percent Identity: 80–100%, Query Coverage: 80–100%, Gao *et al.*, 2024).

Phylogenetic analysis

Twelve pandoravirus isolates with non-shared regions were subjected to a molecular phylogenetic analysis using the nucleotide sequences of their family-B DNA polymerase. Nucleotide sequences were aligned using MAFFT (v7.520) with the options '--genafpair --maxiterate 1000' (Katoh *et al.*, 2018). Alignments were trimmed using Trimal v1.4.rev22 with the '-strictplus' option, followed by the construction of a molecular phylogenetic tree using RapidNJ (bootstrap 10000; Capella-Gutiérrez *et al.*, 2009; Simonsen *et al.*, 2008). The tree was drawn using FigTree v1.4.4.

Statistical analysis

A chi-square goodness-of-fit test was conducted based on the observed counts of eukaryotic species with the highest number of hits in the non-shared region, assuming a uniform expected distribution across taxa (degrees of freedom=k–1). The uniform distribution was used as a simplified baseline and does not incorporate potential biases arising from unequal taxonomic representation in public databases. Pearson's standardized residuals were calculated, and $|d_{ij}|>1.96$ was used as a two-sided significance threshold (α=0.05). The Mann-Whitney U test was employed when comparing either the best-hit sequence length on the non-shared region of Pandoravirus between intergenic and intragenic regions, or the percentage of the genome occupied by CDSs in cases where the best-hit sequence was detected on the non-shared region versus cases where it was not. To investigate the complexity of base compositions in the best-hit sequences, the overall frequency of each 2-mer was calculated for each sequence. The seaborn clustermap function in Python was used for visualization.

Results

This study examined three types of giant viruses—*Pandoravirus*, *Pithovirus*, and *Cedratvirus*—which share a unique amphora-shaped morphology despite being evolutionarily distinct. We extracted non-shared sequences and conducted database searches to identify eukaryotic organisms with homologous sequences. Using this methodology, we aimed to compile a comprehensive list of organisms that exhibit some form of relationship with these viruses.

The sequence profiles of the ten isolates across the three genera of viruses analyzed in this study are presented in Table S2. Fundamental information on the genomes sequenced in this study is as follows: pandoraviruses, 1.8 Mb genome size, 1,600 CDSs, 62% GC content; pithovirus: 612 kb genome size, 479 CDSs, 36% GC content; cedratvirus: 472 kb genome size, 599 CDSs, 42% GC content.

Following an all-vs-all BLASTN analysis of the genomes of pandoraviruses, sequences exceeding 100 bp in length were extracted, resulting in 2,148 sequences. The mean sequence length was approximately 1,037 bp (SD: 1,405 bp) with an average GC content of 55.5% (SD: 5.51). Utilizing these sequences as queries in a subsequent BLASTN search against the eukaryotic nucleotide database, 185 query sequences produced hits, with an average length of 1,393 bp (SD: 1,642) and a mean GC content of 56% (SD: 5.4). Two of these queries produced hits at distinct positions. The total number of subject sequences identified was 10,594, with an average length of 77.8 bp (SD: 36.2) and an average identity percentage with query sequences of 88% (SD: 5.22). By eliminating redundancy among the same query sequences and selecting only the best-hit data that contained at least two closely related taxonomic groups at the genus level, the dataset was refined to 375 entries with an average identity percentage of 89.1% (SD: 5.47) and an average sequence length of 59 bp (SD: 17.9). An examination of histograms depicting the lengths and GC contents of the query sequences indicated that noise-like peaks were effectively removed both prior to and following the BLASTN analysis (Fig. 1A). Although the number of eukaryotic hits significantly decreased after filtering, the overall shapes of the histograms remained consistent with sequence lengths, but not GC contents (Fig. 1A and B). We confirmed the presence of the 187 best-hit sequences identified in the non-shared region of the pandoravirus within the CDS of the pandoravirus genome from which they were detected. We found that 99 sequences (53%) were located in intergenic

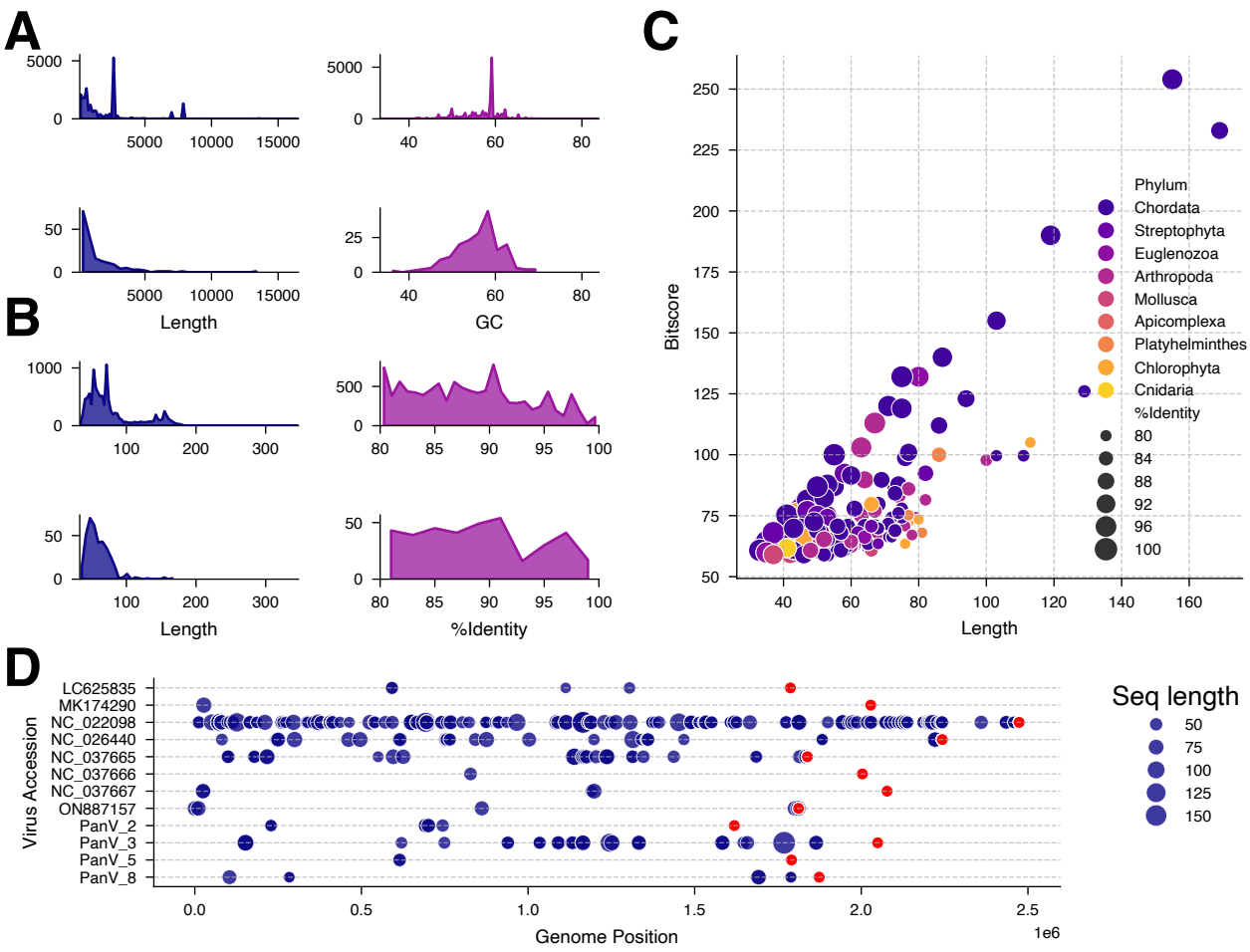


Fig. 1. Detection of virus-specific sequences homologous to eukaryotes.
A. Histogram of query sequence lengths (left) and GC contents (right) for all unique sequences (top) and BLAST hit sequences (bottom). B. Histogram of sequence lengths (left) and% identity (right) for all BLAST hits (top) and best hits (bottom). C. Relative plot for the best-hit sequence regions and bit scores. D. Dot plot for the position of the best hit sequence regions on the virus genome. Red dot: end position of the genome.

regions, while 88 (47%) were located in intragenic regions. The sequences in the intergenic regions were shorter in length (intergenic: 55.9 bp; intragenic: 60.2 bp; the Mann–Whitney U-test: P=0.028; Fig. S1A). The average CDS density (the percentage of the total CDS length relative to the pandoravirus genome size) was 0.688 (SD=0.060) in samples containing detected non-shared regions, and 0.729 (SD=0.048) in samples in which no best-hit sequences were detected (the Mann–Whitney U-test: P=0.250; Fig. S1B). This suggests that more than half of the non-shared regions were located within approximately 30% of non-coding regions. Consequently, non-shared regions appeared to be slightly biased toward intergenic regions. To investigate the extent to which homopolymers and short tandem repeats (STRs) were present in the 187 sequences contained within the non-shared region detected as best-hit sequences, the frequency of the 2-mer nucleotide composition for each sequence was calculated (Table S3). The results revealed that approximately half of the sequences contained prominent repeat sequences composed of AG, TC, AC, or TG (Fig. S2). Sequences containing numerous AA, TT, or AT repeats were limited to a subset of sequences. Although pandoravirus has a GC content of approximately 60%, the frequency of CG, GG, CC, and GC repeats was lower than that of other 2-mers. As described in the Materials and Methods section, STRs were filtered and masked as low-complexity regions during the seed construction process at the beginning of the BLAST search; therefore, STRs did not initiate BLAST hits. Consequently, STRs only appeared in the BLAST search extension step after high-complexity nucleotide regions had been identified.

To assess potential biases in the data of species that yielded hits in BLAST searches, we visualized hit sequences based on their length, bit score, percentage identity, and taxonomic classification at the phylum level. The analysis revealed that subject sequences with long lengths, high bit scores, and an elevated percentage identity were predominantly classified as Chordata (vertebrates) (Fig. 1C). We then examined the relationship between the distribution and length of the hit sequences within the viral genomes. Hits were identified in the genomes of 12 pandoravirus species (71%), and irrespective of the genome size, hit sequences were distributed across various genomic locations (Fig. 1D). The graph for each virus indicates that *Pandoravirus salinus* (NC_022098) exhibited a significantly higher number of hits than the other viruses.

To identify which species exhibited the highest frequency of hits, we analyzed the taxonomic composition of the best-hit data at the Species, Class, and Order levels. Among the 102 species identified, 11 demonstrated a significantly elevated number of hits: *Mus musculus*, *Lampetra planeri*, *Melanogrammus aeglefinus*, *Lampetra fluviatilis*, *Scylla paramamosain*, *Cardiocondyla obscurior*, *Monodelphis domestica*, *Vespula pensylvanica*, *Micromonas pusilla*, *Physcomitrium patens*, and *Peromyscus californicus* (Fig. 2A). The chi-square test was performed using a uniform

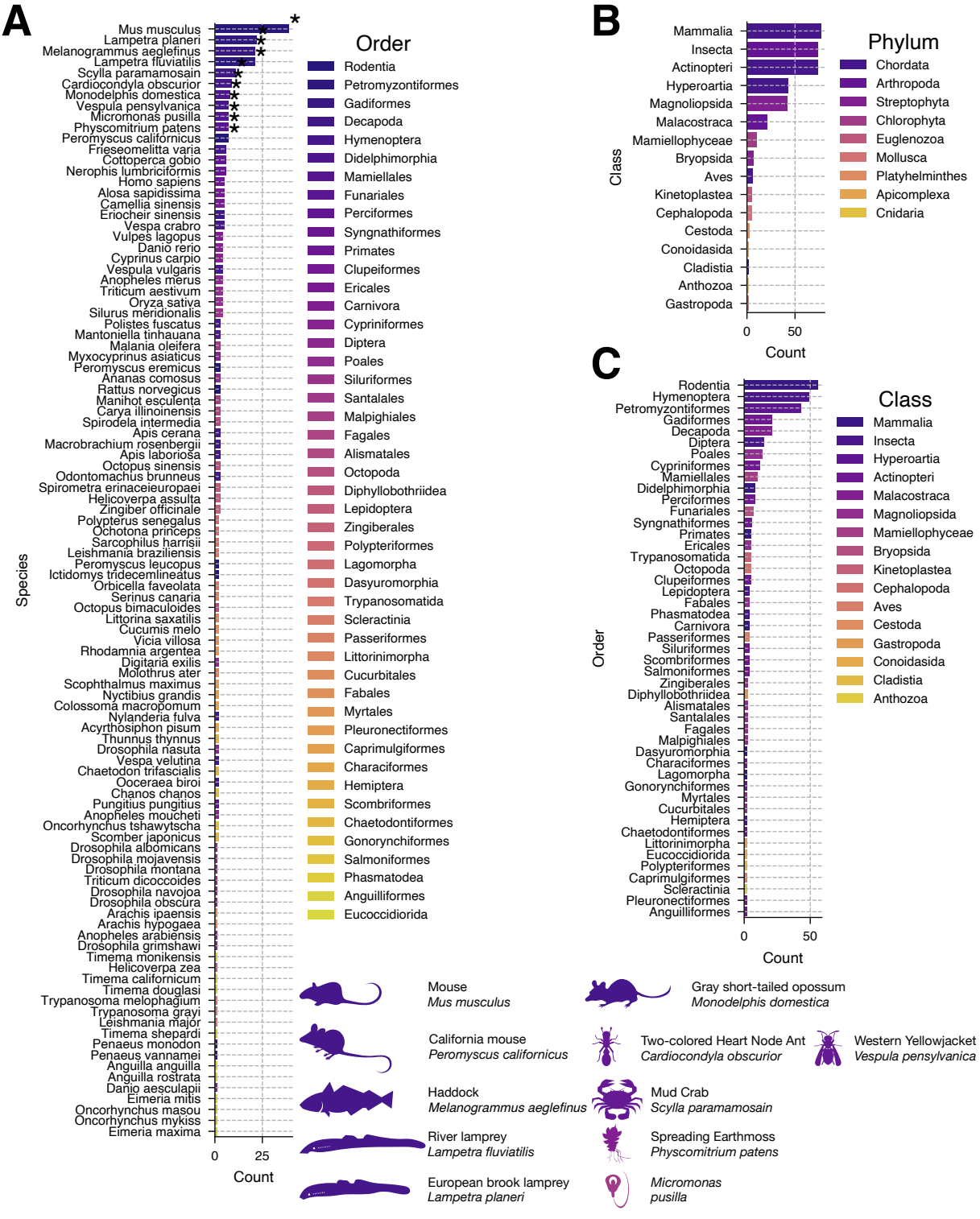


Fig. 2. Best hit taxonomy assignment for virus-unique sequences.
Each panel indicates the taxonomy assignment at the Species (A), Class (B), and Order levels (C). The asterisk in panel A indicates significance at the 5% level based on Pearson's standardized residuals from the chi-square goodness-of-fit test. The chi-square test was performed only at the species rank. A uniform expected distribution was used as a simplified baseline and does not account for unequal taxonomic representation in public databases.

expected distribution as a simplified baseline; therefore, the results obtained need to be interpreted with caution regarding unequal taxonomic representation in public databases. At broader taxonomic levels, the predominant classes were Mammal, Insecta, and Actinopteri, whereas the dominant Orders were Rodentia, Hymenoptera, and Petromyzontiformes (Fig. 2B and C).

Considering potential variations in organisms with hit sequences across different virus species, we analyzed the number of hits to each viral genome at the Class level for each species (Fig. 3). Sequences corresponding to eight virus species were identified in Magnoliopsida, seven species were identified in Hyperoartia, and six species were identified in Mammalia and Insecta. The virus exhibiting hits across the greatest number of classes was *P. salinus* (NC_022098), which was detected in 13 classes, followed by *Pandoravirus inopinatum* (NC_026440) in 11 classes and PanV_3 in seven classes. Considering the possibility that non-shared regions proliferated during the evolutionary process of pandoraviruses, we compared the phylogenetic relationships of pandoraviruses possessing non-shared regions with the actual number of detected non-shared regions. However, no clear correlation was found between the number of detected non-shared regions and the phylogenetic relationships between pandoraviruses and species possessing homologous sequences. Pandoraviruses are generally classified into A-clade and B-clade (*e.g.*, Poirot *et al.*, 2019). We found no extreme bias in the number of non-shared regions between these clades; only a few isolates sporadically possessed a large number of non-shared regions.

To identify non-shared regions contributed by Pandoravirus, BLASTN searches were also performed against viral genomes excluding Pandoravirus. Under conditions of >70% identity and >80% coverage, 16 best hit sequences originating entirely from *Nucleocytoviricota* were obtained from six pandoraviruses (Fig. S3). The average% identity at this stage was 76.17%. Only two hits exceeded 80% identity, originating from *Oceanusvirus* sp. (% identity: 81.25%) and *Cedratvirus lena* (% identity: 82.26%).

A comparable analysis was performed for *Cedratvirus* and *Pithovirus*; however, the number of best-hit data obtained was lower than that for pandoravirus, with 48 for cedratvirus and 5 for pithovirus (Fig. S4 and 5). Given the morphological similarities among viruses, genomic links to closely related organisms were postulated. Consequently, the origins of the sequences identified among the three viral genera were compared at the Class and Order levels. The analysis revealed that Actinopteri was the sole taxon that exhibited a common signal at the class level across all three genera, whereas no common taxonomic group was identified at the order level (Fig. 4).

As a comparative reference for amphora-shaped viruses, the non-shared region was also analyzed with a giant virus with a regular icosahedral shape. The best hit sequences for the 23 non-shared regions in *Mimivirus* were detected in three of the fifteen isolates (average length: 66.39 bp, SD: 14.93; Fig. S6). The distribution of non-shared regions was biased toward both ends of the genome (Fig. S3D). The isolates with the highest detection numbers were from *Aphis gossypii* (cotton aphid) (Fig. S6F). In *Moumouvirus*,

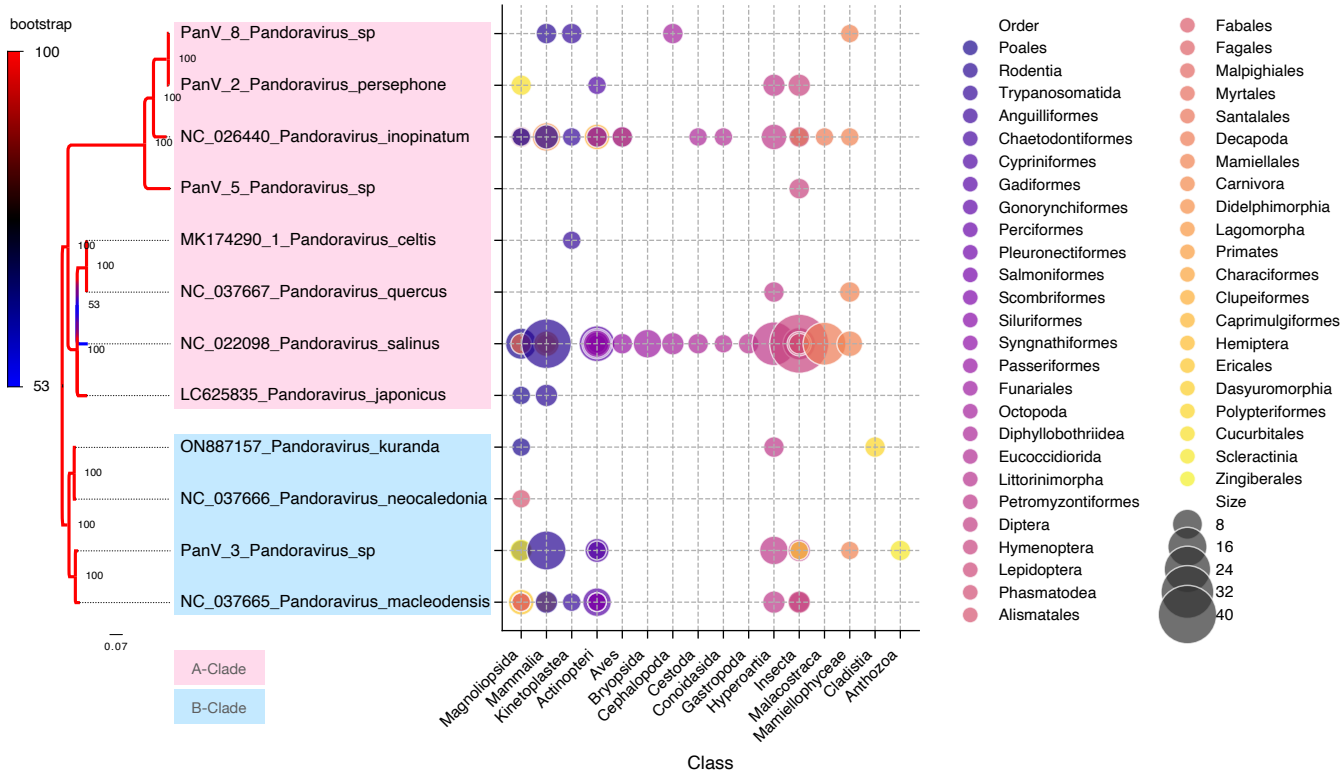


Fig. 3. Taxonomic composition of eukaryotic sequences detected in pandoraviruses.
X-axis: accession ID of pandoraviruses; y-axis: class-level taxonomy of detected eukaryotic sequences; bubble size: number of sequences detected. The phylogenetic tree next to the bubble chart indicates the phylogenetic relationship of each pandoravirus using the nucleotide sequences of family B DNA polymerase (neighbor-joining tree, bootstrapping: 10,000).

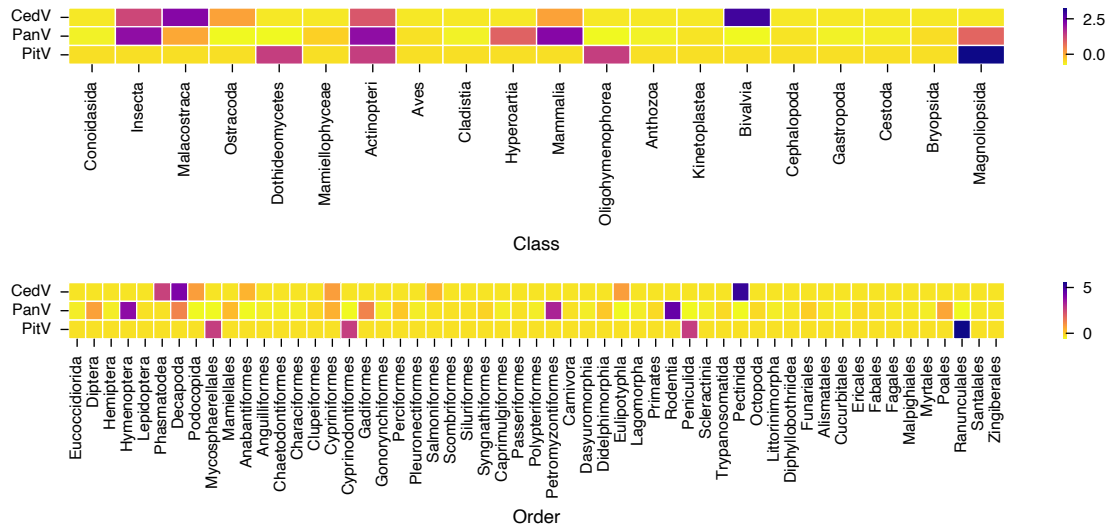


Fig. 4. Cross-validation of the taxonomic composition of amphora-like giant viruses.
Top: Class level composition; bottom: order-level composition. Count data for each virus were normalized using z-scores.

the best-hit sequences for 90 non-shared regions were detected in all six isolates (average length: 63.34 bp, SD: 11.80, Fig. S7). The species with the highest detection counts were *Mytilus trossulus* (Pacific blue mussel), *Chelonus insularis* (wasp), *Cannabis sativa*, and *Strongylocentrotus droebachiensis* (green sea urchin) (Fig. S7F). In *Megavirus*, the best hit sequences for the 55 non-shared regions were detected in eight of the 12 isolates (average length: 71.16 bp, SD: 23.35, Fig. S8). The two species with the highest detection frequencies were *Homo sapiens* and *Glossina fuscipes* (riverine fly) (Fig. S8F). Among the 17 *Faustovirus* isolates, one isolate yielded a best hit sequence for one non-shared region, which matched two *Flustra foliacea* sequences (both with% identity: 84%; e-value: 3.12E-08; Fig. S9). In *Marseillevirus*, we examined 29 isolates comprising Groups A, B, and C, but no data exceeding the threshold were obtained.

To elucidate the characteristics of nucleotide sequences homologous to viral sequences, we extracted data from query-subject pairs belonging to Eukaryote taxonomic groups above the genus level, each containing at least one record and measuring 100 bp or longer in length (Table S4). Consequently, 13 hits were identified, exhibiting percentage identities ranging from 80 to 95% and e-values between 1.1E-15 and 2.76E-62. Each nucleotide sequence comprised several tens of bps of specific homologous sequences and repetitive sequences–2–3 bp in length (homopolymer, STR). The organisms harboring these homologous sequences included *Melanogrammus aeglefinus* (six hits), *Vespula vulgaris* (four hits), *L. fluviatilis* (one hit), *Micromonas pusilla* (one hit), and *Oncorhynchus mykiss* (one hit). The corresponding viruses were *P. salinus* (10 hits), *Pandoravirus sp.* (PanV_3, two hits), and *P. inopinatum* (one hit). Among the 13 hit sequences, four were located on open reading frames (ORFs) in eukaryotic genomes, specifically formin-J-like isoforms X2, 3, and 4 and zygotic gap protein knirps-like isoform X1. The remaining sequences were either intergenic (four cases) or intronic (four cases). The hit sequence of *O. mykiss* (rainbow trout) against *P. salinus* is part of a non-coding sequence (conserved non-coding element; CNE 782–785), and it has been reported that CNE 782–785 is conserved in zebrafish, medaka, human, rainbow trout, and Atlantic salmon (Moghadam *et al.*, 2009). This CNE sequence was found within an intron of a hypothetical protein gene in *P. salinus* (locus tag "psal_cds_664" on NC_022098.1, from 1,164,857 to 1,165,010 bp; Table S4). Using the nine types of non-shared region sequences of 100 bp or longer listed in Table S4, BLASTN was performed against the protozoa-specific P10K database (thresholds: ≥80% identity and ≥80% coverage). A total of 293 hit sequences were obtained from eight of the nine query sequences (Table S5). The average coverage of these hit sequences was 96.7%, and the average% identity was 87%. The top three organisms with the highest number of hits when classified at the genus level or above were *Acanthamoeba* (92 hits), *Botryococcus* (43 hits), and *Symbiodinium* (24 hits). At the species level, the top three species by hit count were *Acanthamoeba healyi* (60 hits), *Botryococcus braunii* (43 hits), and *Acanthamoeba lenticulata* (25 hits), with the genus *Acanthamoeba* accounting for approximately the top 30% of hits. However, when examined by sequence, all *Acanthamoeba* hits corresponded to the same query sequence (Query ID: 5 on Table S5). The% identity values indicated by the best hit sequences in these BLASTN results were largely consistent with those for hits against the NCBI BLASTN eukaryote database (Table S4).

Discussion

In examining the presence of repeat sequences within both virus- and eukaryote-derived sequences identified in this study, several plausible scenarios warrant consideration: HGT via recombination or transposable elements (TEs), conserved sequences originating from a common ancestor. We acknowledge that species-level identification based on sequence similarity searches may be affected by a database size-driven bias because certain taxa are disproportionately

represented in public databases. Accordingly, the observed taxonomic distribution of hits needs to be interpreted with caution. Notably, the sequence similarities detected were not driven by tandem repeat–like sequences. Instead, significant matches were established by non-repetitive regions, with tandem repeats appearing only as part of the BLAST extension step. The conservation of such repeat structures across distant taxa may itself reflect underlying biological constraints rather than computational artifacts (Fig. S2, Table S3 and S4). The virus contains sequences homologous to certain non-coding regions previously reported in rainbow trout (Table S4; Moghadam *et al.*, 2009). The result showing that conserved CNEs were not perfectly conserved among rainbow trout implies that CNE-homologous sequences found among non-shared sequences of viruses reflect this variability. In essence, unless there is strong selective pressure that results in the survival of only a few individuals within a species, a diverse array of sequences from various origins may be maintained within a population. CNE sequences function as transcriptional regulatory elements, are ubiquitously present across species, and have undergone multiple diversification processes throughout their evolutionary history (Gonzalez *et al.*, 2024). According to these findings, CNEs are frequently located on TEs, with approximately 5–10% of CNEs found in arthropods, cnidarians, and mollusks derived from TEs. Furthermore, *P. salinus* contains 30 copies of a TE known as a submariner, a miniature inverted-repeat transposable element (MITEs), whereas its congener *Pandoravirus dulcis* does not retain any of them (Sun *et al.*, 2015). In the present study, *P. salinus* conserved the largest number of non-shared sequences homologous to those in eukaryotes, whereas no non-shared sequences were detected in *P. dulcis* (Fig. 1D and 3). Consequently, the presence or abundance of non-shared sequences in each pandoravirus may be associated with TEs, which may facilitate the acquisition or donation of new DNA sequences during viral infection. In summary, these results are consistent with the possibility that CNE-like sequences are horizontally transferred across species, potentially through viral infection and TE transposition within genomes.

The initial aim of this study was to investigate whether traces of HGT remained in regions not shared by viruses (non-shared regions). However, a relationship with evolutionarily ancient CNE (conserved NCE)-like sequences was actually discovered on the intron of the hypothetical protein-coding gene in *P. salinus*. No regulatory or replication-related functions have been reported for non-coding regions of pandoraviruses to date. Some CNEs are considered to be derived from TEs, such as long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs), and are conserved as sequences of tens to hundreds of bps across a wide range of vertebrates, including humans, birds, and coelacanths (Bejerano *et al.*, 2006; Kamal *et al.*, 2006; Xie *et al.*, 2006). The lengths of the non-shared regions detected in this study were similar in scale to these sequences (Table S4), and it is of interest that they shared structural similarities with them. Furthermore, recent studies reported that STRs associated with Alzheimer's disease were concentrated within SINE-VNTR-Alu (SVA)-type reverse transposons (Guo *et al.*, 2025). The detection of STRs within the non-shared region in this study (Table S4) suggests that these regions are associated with transposon-derived sequences. Conversely, CNE sequences are frequently functionally constrained, and studies on *Drosophila* reported a low frequency of TE insertion into CNEs (Manee *et al.*, 2018). Consequently, if CNE-like sequences exist in pandoraviruses, they may have remained intact for long periods due to functional constraints. The correspondence observed in this study between CNE-like sequences and non-shared regions may represent an older evolutionary signature from the time of the co-evolution of eukaryotes and viruses, rather than a recent HGT event. Due to the short lengths of the detected similarities (average 59 bp), these matches need to be interpreted with caution and do not, on their own, constitute definitive evidence of HGT. Further comparative analyses of these regions may reveal phylogenetic relationships predating the TEs found in vertebrates and insects.

Given that numerous giant viruses, such as pandoraviruses, have been isolated using *Acanthamoeba*, it was initially anticipated that the genomes of protozoa, specifically sequences from their natural hosts, may be frequently observed. However, the results obtained present interpretative challenges. Assuming that the identified sequences were acquired through HGT, two scenarios were considered. In the first scenario, the detected sequence belongs to the direct host and parts of the sequences are recombined when the virus infects it. In the second scenario, the genome of the detected organism is incorporated into the host cell along with the virus and recombination occurs within the cell. It has been demonstrated that giant virus sequences may undergo recombination within *Acanthamoeba*. For example, recombinant viruses may be produced by infecting *A. castellanii* with melbournevirus and simultaneously transfecting DNA with homologous sequences (Liu *et al.*, 2021). Nevertheless, in a year-long serial passage experiment conducted to investigate the genome evolution of giant viruses in *Acanthamoeba*, where *Acanthamoeba* and *Lausannevirus* were co-cultured with the bacterium *Estrella lausannensis*, which inhabits amoeba cells, HGT was not detected (Mueller *et al.*, 2017). The genomes acquired by viruses are not necessarily derived from their primary hosts (Bryan *et al.*, 2008). Furthermore, since many giant viruses do not replicate in the nucleus, but in virus factories (Queiroz *et al.*, 2022), HGT will not occur unless there is at least an opportunity for transfected naked DNA to enter the viral factory. *Medusavirus* is an exception to this rule because it replicates inside the nucleus and harbors DNA polymerase B (*polB*), which is phylogenetically closer to eukaryotic DNA polymerase delta than to *polB* of *Nucleocytoviricota*. This suggests that replication relies on the eukaryotic nuclear machinery, highlighting a unique interaction between the virus and host (Yoshikawa *et al.*, 2019). When sequences from non-shared regions exceeding 100 bp were submitted to the protozoa-specific P10K database, one query yielded hits for the genus *Acanthamoeba* (*A. healyi* and *A. lenticulata*, Table S5). This result suggests, as one possibility, that these sequences were introduced by *Acanthamoeba* during infection. However, because the

viruses were propagated in *Acanthamoeba* culture systems, contamination by host DNA during library preparation cannot be ruled out, and this needs to be considered as an alternative explanation. In addition, the percentage identity value indicated by the best hit in the P10K database was similar to that observed for hits against NCBI BLASTN eukaryote data and, thus, the origin of this non-shared region sequence was not definitively identified. Furthermore, we performed homology searches against the viral genome using the same non-shared region as the query sequence. However, the number of best hits was extremely low at 16, with maximum homology of approximately 82% (Fig. S3). Therefore, the probability of novel sequence exchange occurring beyond the genus level in Pandoravirus is low. Sequences homologous to various eukaryotes were detected not only from Pandoravirus, but also from the non-shared region of certain *Mimiviridae* with icosahedral shapes. Conversely, virtually no sequences were detected from *Marseillevirus* or *Faustovirus*. This suggests that factors other than viral shape are responsible for the retention of these diverse non-shared region sequences within giant viruses. Infection experiments involving the capture of natural hosts are necessary to ascertain the conditions under which this infection occurs in natural environments.

In this study, we concentrated on non-conserved sequences among virus species, which have not been extensively examined to date, through a comparative genomic analysis of amphora-shaped viruses. We analyzed the homology of these sequences with those of eukaryotes. The primary objective of this study was to establish insights into the natural hosts of these viruses in their natural environments. However, they unexpectedly indicated matches with genome sequences from a diverse array of species, including arthropods, fish, and mammals. Although the detection rate was markedly high for certain species, approximately 90% of the detected species exhibited a low frequency, complicating the identification of direct host clues. Nevertheless, the identification of a CNE-like sequence in *P. salinus*, together with the widespread distribution of CNEs across diverse taxa, suggests the involvement of these elements in virus–host evolutionary interactions. However, given the limited number of detected cases and the short lengths of the similarities, these results need to be interpreted cautiously, and further investigations will be necessary to clarify their biological significance.

Data availability

The non-shared region extractor may be found at the following GitHub link (https://github.com/mothir-a/Blast-Gap-Extractor). Newly sequenced data from this study are available in GenBank under the following accession IDs: CedV_1, LC877910; PanV_1, LC877911; PanV_2, LC877912; PanV_3, LC877913; PanV_4, LC877914; PanV_5, LC877915; PanV_6, LC877916; PanV_7, LC877917; PanV_8, LC877918; and PitV_1, LC877919.

Acknowledgements

We would like to thank Bioengineering Lab Co., Ltd. for performing the sequence analysis. We express our gratitude to Shigeru Kato and Christine Dwi A.P.W for their advice during the course of our research.

Sources of funding

This study was conducted with the support of a Grant-in-Aid for Scientific Research (JSPS KAKENHI for MA; No. 23K14167).

Conflict of interest

The authors affirm that the research was conducted without any commercial or financial affiliations that may be perceived as potential conflicts of interest.

References

Aherfi, S., Andreani, J., Baptiste, E., Oumessoum, A., Dornas, F.P., Andrade, A.C.D.S.P., *et al.* (2018) A large open pangenome and a small core genome for giant Pandoraviruses. *Front Microbiol* **9**: 1486.

Akashi, M., and Takemura, M. (2019) Co-isolation and characterization of two Pandoraviruses and a Mimivirus from a riverbank in Japan. *Viruses* **11**: 1123.

Akashi, M., Takemura, M., and Suzuki, S. (2024) Continuous year-round isolation of giant viruses from brackish shoreline soils. *Front Microbiol* **15**: 1402690.

Andreani, J., Aherfi, S., Bou Khalil, J.Y., Di Pinto, F., Bitam, I., Raoult, D., *et al.* (2016) Cedratvirus, a double-cork structured giant virus, is a distant relative of pithoviruses. *Viruses* **8**: 300.

Aquino, I.L.M., Reis, E.S., Moreira, R.O.A.M., Arias, N.E.C., Barcelos, M.G., Queiroz, V.F., *et al.* (2024) Giant viruses inhibit superinfection by downregulating phagocytosis in Acanthamoeba. *J Virol* **98**: e0104524.

Bejerano, G., Lowe, C.B., Ahituv, N., King, B., Siepel, A., Salama, S.R., *et al.* (2006) A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature* **441**: 87–90.

Bisio, H., Legendre, M., Giry, C., Philippe, N., Alempic, J.-M., Jeudy, S., and Abergel, C. (2023) Evolution of giant pandoravirus revealed by CRISPR/Cas9. *Nat Commun* **14**: 428.

Bosmon, T., Abergel, C., and Claverie, J.M. (2025) 20 years of research on giant viruses. *npj Viruses* **3**: 9.

Boyer, M., Azza, S., Barrassi, L., Klose, T., Campocasso, A., Pagnier, I., *et al.* (2011) Mimivirus shows dramatic genome reduction after intraamoebal culture. *Proc Natl Acad Sci U S A* **108**: 10296–10301.

Breitbart, M. (2012) Marine viruses: Truth or dare. *Ann Rev Mar Sci* **4**: 425–448.

Bryan, M.J., Burroughs, N.J., Spence, E.M., Clokie, M.R.J., Mann, N.H., and Bryan, S.J. (2008) Evidence for the intense exchange of MazG in marine cyanophages by horizontal gene transfer. *PLoS One* **3**: e2048.

Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**: 1972–1973.

Elde, N.C., Child, S.J., Eickbush, M.T., Kitzman, J.O., Rogers, K.S., Shendure, J., *et al.* (2012) Poxviruses deploy genomic accordions to adapt rapidly against host antiviral defenses. *Cell* **150**: 831–841.

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

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

Gallot-Lavallée, L., and Blanc, G. (2017) A glimpse of nucleo-cytoplasmic large DNA virus biodiversity through the eukaryotic genomics window. *Viruses* **9**: 17.

Gao, X., Chen, K., Xiong, J., Zou, D., Yang, F., Ma, Y., *et al.* (2024) The P10K database: a data portal for the protist 10 000 genomes project. *Nucleic Acids Res* **52**: D747–D755.

Gonzalez, P., Hauck, Q.C., and Baxevanis, A.D. (2024) Conserved noncoding elements evolve around the same genes throughout metazoan evolution. *Genome Biol Evol* **16**: evae052.

Guglielmini, J., Woo, A.C., Krupovic, M., Forterre, P., and Gaia, M. (2019) Diversification of giant and large eukaryotic dsDNA viruses predated the origin of modern eukaryotes. *Proc Natl Acad Sci U S A* **116**: 19585–19592.

Guo, M.H., Lee, W.P., Vardarajan, B., Schellenberg, G.D., and Phillips-Cremins, J.E. (2025) Polygenic burden of short tandem repeat expansions promotes risk for Alzheimer's disease. *Nat Commun* **16**: 1–15.

Kamal, M., Xie, X., and Lander, E.S. (2006) A large family of ancient repeat elements in the human genome is under strong selection. *Proc Natl Acad Sci U S A* **103**: 2740–2745.

Katoh, K., Rozewicki, J., and Yamada, K.D. (2018) MAFFT online service: Multiple sequence alignment, interactive sequence choice and visualization. *Briefings Bioinf* **20**: 1160–1166.

Krupovic, M., Yutin, N., and Koonin, E. (2020) Evolution of a major virion protein of the giant pandoraviruses from an inactivated bacterial glycoside hydrolase. *Virus Evol* **6**: 1–8.

Legendre, M., Bartoli, J., Shmakova, L., Jeudy, S., Labadie, K., Adrait, A., *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.

Liu, Y., Bisio, H., Toner, C.M., Jeudy, S., Philippe, N., Zhou, K., *et al.* (2021) Virus-encoded histone doublets are essential and form nucleosome-like structures. *Cell* **184**: 4237–4250.e19.

Manee, M.M., Jackson, J., and Bergman, C.M. (2018) Conserved noncoding elements influence the transposable element landscape in Drosophila. *Genome Biol Evol* **10**: 1533–1545.

Maumus, F., and Blanc, G. (2016) Study of gene trafficking between acanthamoeba and giant viruses suggests an undiscovered family of amoeba-infecting viruses. *Genome Biol Evol* **8**: 3351–3363.

Moghadam, H.K., Ferguson, M.M., and Danzmann, R.G. (2009) Comparative genomics and evolution of conserved noncoding elements (CNE) in rainbow trout. *BMC Genomics* **10**: 1–13.

Mueller, L., Bertelli, C., Pillonel, T., Salamin, N., and Greub, G. (2017) One year genome evolution of Lausannevirus in allopatric versus sympatric conditions. *Genome Biol Evol* **9**: 1432–1449.

Philippe, N., Legendre, M., Doutre, G., Couté, Y., Poirot, O., Lescot, M., *et al.* (2013) Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. *Science* **341**: 281–286.

Poirot, O., Jeudy, S., Abergel, C., and Claverie, J.M. (2019) A puzzling anomaly in the 4-Mer composition of the giant pandoravirus genomes reveals a stringent new evolutionary selection process. *J Virol* **93**: e01206–19.

Queiroz, V.F., Rodrigues, R.A.L., De Miranda Boratto, P.V., La Scola, B., Andreani, J., and Abrahao, J.S. (2022) Amoebae: Hiding in plain sight: Unappreciated hosts for the very large viruses. *Annu Rev Virol* **9**: 79–98.

Schulz, F., Roux, S., Paez-Espino, D., Jungbluth, S., Walsh, D.A., Denef, V.J., *et al.* (2020) Giant virus diversity and host interactions through global metagenomics. *Nature* **578**: 432–436.

Simonsen, M., Mailund, T., and Pedersen, C.N.S. (2008) Rapid neighbour-joining. In *Lecture Notes in Computer Science (LNBI), vol. 5251*. Crandall, K.A., and Lagergren, J. (eds). Berlin: Springer, Berlin, Heidelberg, pp. 113–122.

Sun, C., Feschotte, C., Wu, Z., and Mueller, R.L. (2015) DNA transposons have colonized the genome of the giant virus Pandoravirus salinus. *BMC Biol* **13**: 38.

Wu, J., Meng, L., Gaïa, M., Hikida, H., Okazaki, Y., Endo, H., and Ogata, H. (2024) Gene transfer among viruses substantially contributes to gene gain of giant viruses. *Mol Biol Evol* **41**: msae161.

Xie, X., Kamal, M., and Lander, E.S. (2006) A family of conserved noncoding elements derived from an ancient transposable element. *Proc Natl Acad Sci U S A* **103**: 11659–11664.

Yoshikawa, G., Blanc-Mathieu, R., Song, C., Kayama, Y., Mochizuki, T., Murata, K., *et al.* (2019) Medusavirus, a novel large DNA virus discovered from hot spring water. *J Virol* **93**: JVI.02130-18.

Yutin, N., Wolf, Y.I., and Koonin, E.V. (2014) Origin of giant viruses from smaller DNA viruses not from a fourth domain of cellular life. *Virology* **466–467**: 38–52.