Gene duplication as a major force driving the genome expansion in some giant viruses

Talita B Machado, Agnello C R Picorelli, Bruna L de Azevedo, Isabella L M de Aquino, Victória F Queiroz, Rodrigo a L Rodrigues, João Pessoa Araújo, Leila S Ullmann, Thiago M dos Santos, Rafael E Marques, et al.

AMERICAN SOCIETY FOR MICROBIOLOGY | Journal of Virology

Virology | Full-Length Text

Gene duplication as a major force driving the genome expansion in some giant viruses

Talita B. Machado,[1] Agnello C. R. Picorelli,[2] Bruna L. de Azevedo,[1] Isabella L. M. de Aquino,[1] Victória F. Queiroz,[1] Rodrigo A. L. Rodrigues,[1] João Pessoa Araújo Jr.,[3] Leila S. Ullmann,[3] Thiago M. dos Santos,[4] Rafael E. Marques,[5] Samuel L. Guimarães,[5] Ana Cláudia S. P. Andrade,[6] Juliana S. Gularte,[7] Meriane Demoliner,[7] Micheli Filippi,[7] Vyctoria M. A. G. Pereira,[7] Fernando R. Spilki,[7] Mart Krupovic,[8] Frank O. Aylward,[9,10] Luiz-Eduardo Del-Bem,[4] Jônatas S. Abrahão[1]

AUTHOR AFFILIATIONS See affiliation list on p. 14.

ABSTRACT Giant viruses with their gigantic genomes are among the most intriguing components of the virosphere. How these viruses attained such giant genomes remains unclear, despite considerable efforts to understand this phenomenon. Here, we describe the discovery of cedratvirus pambiensis, an amoebal giant virus isolated in Brazil. Although the virion morphology and replication cycle of c. pambiensis are very similar to those described for other cedratviruses, whole genome sequencing revealed the largest cedratvirus genome ever described, with 623,564 base pairs and 842 predicted protein-coding genes (among them, 76 ORFans). Genome analysis has revealed an unprecedented number of paralogous genes, with ~73% of the c. pambiensis genome being composed of genes with two or more copies. Large families of functionally diverse paralogous genes included up to >70 copies and were distributed across the genome. The in-depth investigation of the mechanisms and origins of gene duplications revealed that both tandem-like duplications and distal transfer of syntenic blocks of genes contributed to the c. pambiensis genomic expansion. Finally, a comprehensive genome analysis of viruses from all known giant virus families suggested that gene duplication is one of the key mechanisms underlying genomic gigantism across the phylum *Nucleocytoviricota*. The expansion of viral genomes through successive duplications followed by subfunctionalization and exaptation of the paralogous gene copies may promote the adaptation of giant viruses to a variety of niches.

IMPORTANCE Giant viruses are noteworthy not only due to their enormous particles but also because of their gigantic genomes. In this context, a fundamental question has persisted: how did these genomes evolve? Here we present the discovery of cedratvirus pambiensis, featuring the largest genome ever described for a cedratvirus. Our data suggest that the larger size of the genome can be attributed to an unprecedented number of duplicated genes. Further investigation of this phenomenon in other viruses has illuminated gene duplication as a key evolutionary mechanism driving genome expansion in diverse giant viruses. Although gene duplication has been described as a recurrent event in cellular organisms, our data highlights its potential as a pivotal event in the evolution of gigantic viral genomes.

KEYWORDS giant virus, *Pithoviridae*, cedratvirus pambiensis, genome expansion, paralogous genes, *Nucleocytoviricota*

Gene and genomic segment duplication is a critical mechanism underlying the evolution of cellular organisms by providing raw genetic material for the emergence of new gene functions and pathways (1). Duplicated genes can undergo subfunctionalization by acquiring mutations, resulting in the evolution of new protein

Editor Colin R. Parrish, Cornell University Baker Institute for Animal Health, Ithaca, New York, USA

Address correspondence to Jônatas S. Abrahão, jonatas.abrahao@gmail.com, or Luiz-Eduardo Del-Bem, delbem@ufmg.br.

The authors declare no conflict of interest.

See the funding table on p. 15.

Received 23 August 2023
Accepted 26 October 2023
Published 21 November 2023

functions. The significance of this phenomenon in evolution is evidenced by the widespread occurrence of duplicated genes across all domains of life (2–5). It has been estimated that around 30%–65% of the genes in multicellular eukaryotes, such as humans, have emerged through duplication (6, 7).

Giant viruses of amoeba are characterized by their large genome sizes and complex gene repertoires (8–11). Although the driving forces that led to the genome gigantism of those viruses are not fully understood, horizontal gene transfer, *de novo* gene emergence, and gene duplication have all been hypothesized to have contributed to genome expansion (12–18). Large families of functionally diverse paralogous genes have been identified in giant viruses of amoebas, including genes encoding ankyrin repeat-containing proteins, receptors for ubiquitination targets, and proteins with glycosyltransferase domains. In addition, many of these gene families are composed of unknown proteins or ORFan genes (13, 19). Although studies on gene duplication in giant viruses are scarce, there is convincing evidence showing that approximately one-third of the mimivirus genome and 50% of pandoravirus genomes are composed of multi-copy genes (13, 19).

Here, we report the discovery of cedratvirus pambiensis, a giant amoeba virus with the largest genome size ever described for the cedratvirus group, comprising 623,564 base pairs. The investigation of the architecture of the genome revealed an unprecedented abundance of duplicated genes, which constitute up to 72% of the total genome, a proportion on par with or surpassing that observed in cellular organisms. Most of these genes are grouped into six major gene families. Only 27.7% of the genome is composed of single-copy genes (non-duplicated genes). The expansion of the analyses to other varidnaviruses revealed extensive gene duplications in most groups of giant viruses, most markedly in members of the "Pithoviridae"-related viruses (which includes cedratviruses), pandoraviruses, and some mimiviruses. The discovery of c. pambiensis expands our understanding of the diversity and complexity of giant viruses, emphasizing the role of gene duplication in driving their genome expansion and shaping the genomic content.

RESULTS

Cedratvirus pambiensis particles and replication cycle

As part of our ongoing efforts to characterize the diversity of giant viruses infecting amoeba, we have isolated a new cedratvirus, c. pambiensis, from a water sample collected in a small, forested area at the Federal University of Minas Gerais (UFMG) campus, Belo Horizonte, Brazil. Inoculated amoebas exhibited cytopathic effects such as rounding and lysis, and upon light microscopy examination, it was possible to visualize viral particles. To gain a more comprehensive understanding of the particles' characteristics, we analyzed images obtained by transmission electron microscopy (TEM), negative staining electron microscopy (NSEM) (Fig. 1), and scanning electron microscopy (SEM) (see Supplementary Fig. 1a posted at https://www.giantviruses.com/sup-material-of-papers/sup-material-gene-duplication-as-a-major-force-driving-the-genome-expansion-in-some-giant-viruses). The virus particles were oval-shaped, measuring approximately 1 µm in length and 500 nm in width (Fig. 1a). The capsid is composed of parallel striations (Fig. 1a) and may have one or two apical "corks." Most observed particles had two corks, which is consistent with previous descriptions of cedratviruses. Notably, the capsid of c. pambiensis exhibited surface fibrils, a structure that has not been previously documented in cedratviruses (Fig. 1c). Although not visible by TEM and SEM, these fibrils were present in all images obtained by NSEM.

The replication cycle of c. pambiensis is a complex process that involves several steps (Fig. 1). It begins with the entry of viral particles into the host cell, which probably occurs via phagocytosis. Once inside the cell, the viral particles undergo uncoating, followed by an eclipse period which typically lasts for around 3 hours (Supplementary Fig. 1b posted at https://www.giantviruses.com/sup-material-of-papers/sup-material-gene-duplication-as-a-major-force-driving-the-genome-expansion-in-some-giant-viruses). The uncoating

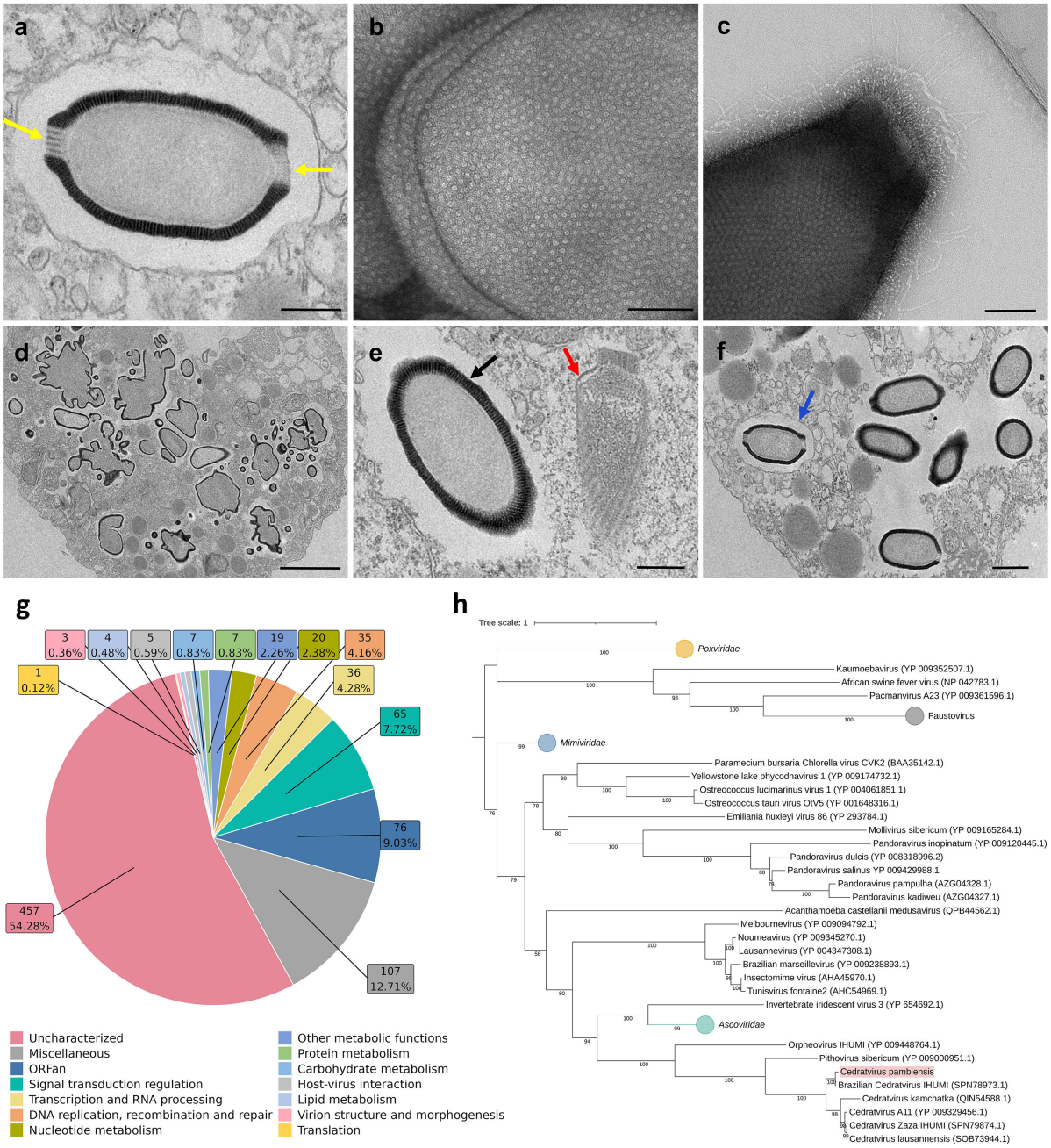

FIG 1 Particle, replication cycle, and genomic features of c. pambiensis. (a) Cedratvirus particle showing an oval shape with two apical corks (yellow arrows). Scale bar: 200 nm. (b) Zoom in the particle, showing the parallel striated structures of the capsid. Scale bar: 100 nm. (c) Observation of surface fibrils on the capsid. Scale bar: 100 nm. (d) Amorphous structures with no defined function in the cytoplasm of an infected amoeba. Scale bar: 2 µm. (e) Assembly of new viral progeny. It is possible to observe structures of early particles being formed (red arrow) and mature particles (black arrow) at the same time inside the cells. Scale bar: 200 nm. (f) A viral particle was observed inside a vesicle after assembly (blue arrow), suggesting release by exocytosis before cell lysis. Scale bar: 500 nm. (g) Functional categories of c. pambiensis predicted genes. The color legend is provided below the graph. (h) Maximum likelihood phylogenetic tree constructed with amino acid sequences from the DNA polymerase subunit B of cedratviruses and other nucleocytoviruses. The new isolate described here, c. pambiensis (highlighted in pink), clustered with other cedratviruses, closer to the Brazilian cedratvirus. The scale bar indicates the genetic distance.

involves the removal of the viral cork to release the viral genome into the cytoplasm of the host cell. The next step is the formation of the viral factory, which is a region within the host cell that supports viral replication. In electron micrographs, the viral factory appears as a space that is unbounded and electron-lucent. Within the factory, amorphous electron-dense structures can be observed (Fig. 1d). Although the function of

TABLE 1 Comparison between the genomes of all cedratviruses with published genomes

Virus	Genome size	Predicted proteins	% Intergenic regions
Cedratvirus A11	589,068 bp	574	18.99
Cedratvirus lausannensis	575,161 bp	643	17.15
Cedratvirus zaza	560,887 bp	636	15.69
Brazilian cedratvirus	460,038 bp	533	14.43
Cedratvirus kamchatka	466,767 bp	545	18.26
Cedratvirus pambiensis	**623,564 bp**	**842**	**10.59**

these structures is unknown, they appear to be composed of material that is similar to that of the viral capsid, as seen in Fig. 1a. During the assembly of new viral particles, initial structures that will contribute to the formation of new virions can be observed, as shown in Fig. 1e. At 12 hours post-infection (hpi), viral production reaches its maximum level, after which it plateaus, as seen in Supplementary Fig. 1b posted at https://www.giantviruses.com/sup-material-of-papers/sup-material-gene-duplication-as-a-major-force-driving-the-genome-expansion-in-some-giant-viruses. The final step in the replication cycle is the release of viral progeny, which occurs via cell lysis. However, exocytosis may also play a role in this process, as depicted in Fig. 1f, where particles can be visualized inside vesicles close to the cell plasma membrane at the end of the infection cycle.

Cedratvirus pambiensis has an unprecedented abundance of paralogous genes

Genomic characterization of c. pambiensis yielded a circular dsDNA molecule of 623,564 bp and encoding 842 predicted proteins. Until then, the largest cedratvirus genome was described for cedratvirus A11, with 589,068 bp, and the largest number of predicted proteins was described for cedratvirus lausannensis, with 643. Thus, c. pambiensis is the cedratvirus with the largest genome and highest number of predicted proteins among all cedratviruses published to date (Table 1). As gene prediction methods may vary among different studies, we performed the gene prediction of all available cedratviruses using the same parameters that we applied to c. pambiensis. Although we observed a general increase in the number of predicted genes for all viruses, c. pambiensis still holds the record for the largest number of predicted genes among the isolated viruses. Functional analysis of the predicted proteins (Fig. 1g) revealed that most of them are uncharacterized (54.24%), and ORFans (9.03%). Proteins related to the regulation of signal transduction; transcription and RNA processing; DNA replication, recombination and repair, and different types of metabolism were also identified. The construction of a phylogenetic tree using amino acid sequences from the family B DNA polymerase showed that the new isolate clustered with other cedratviruses (Fig. 1h), and most closely to the Brazilian cedratvirus. The synteny analysis reinforces the proximity of these two cedratviruses, when compared to the other cedratviruses (see Supplementary Fig. 2 posted at https://www.giantviruses.com/sup-material-of-papers/sup-material-gene-duplication-as-a-major-force-driving-the-genome-expansion-in-some-giant-viruses).

Initially, we hypothesized that the increase in the genome size could be due to the presence of a new class of genes or a substantial number of ORFans. However, the annotation of the c. pambiensis genome revealed a gene content that is similar to that of other cedratviruses. Next, we investigated the intergenic content of c. pambiensis genome and compared it to that of other cedratviruses. However, after analyzing the intergenic content in all cedratviruses with available genomes, we observed that c. pambiensis has the lowest percentage (10.59%) of predicted intergenic regions among all of them (mean of 16.90%) (see Supplementary Table 1 posted at https://www.giantviruses.com/sup-material-of-papers/sup-material-gene-duplication-as-a-major-force-driving-the-genome-expansion-in-some-giant-viruses).

Then, we tested whether the larger genome size of c. pambiensis could be explained by the increase in the number of paralogous genes, which occurs due to gene and genomic segment duplications. We performed an all-against-all BLASTp analysis of the predicted proteins of c. pambiensis, which revealed an unexpectedly large number of paralogous groups with more than three genes, as well as paralogs grouped in pairs and triplets (Fig. 2). Only 27.7% of the c. pambiensis genome consists of single copy genes (Fig. 2). At least six large gene families (>20 genes) were identified in c. pambiensis genome, encoding functionally diverse proteins, including ankyrin-domain containing proteins, collagen-like proteins, serine-threonine protein kinases, hypothetical proteins, proteins containing F-box domain, ORFans, and others. Certain families presented genes with more than one predicted function or domain, suggesting progressive differentiation after duplication (Fig. 2). All the information about the paralogous genes is described in Supplementary Material 1 posted at https://www.giantviruses.com/sup-material-of-papers/sup-material-gene-duplication-as-a-major-force-driving-the-genome-expansion-in-some-giant-viruses.

Of 842 total genes, 609 (72.3%) are part of multi-gene families while only 233 (27.7%) are single-copy genes (Fig. 3a). We observed that a large fraction of these paralogous groups (34.6%) are composed of tandem genomic segment duplications (Fig. 3a and b). We then evaluated the percentage of tandemly duplicated genes across gene families, showing that these events are much more common in the six largest gene clusters (57%), while in low-copy number families present in triplets (9%) or pairs of genes (10%) tandem duplication events are less frequent (Fig. 3c). This finding suggests that tandem duplications are primarily responsible for the formation and expansion of large gene families.

To better understand the tandem duplication events and their effect on the genome evolution of these viruses, we constructed phylogenetic trees with the protein sequences for the six largest gene families (or clusters). Analysis of the phylogenetic trees showed that two types of duplication events occurred: proximal tandem duplications (involving more recent duplications) and chromosomal segment duplications (when an entire block of tandem genes appears to have been copied from one part of the genome and pasted into another, occasionally disrupting the preexisting genes). As an example, these analyses were detailed for the phylogenetic tree of cluster 3 (Fig. 3d), in which tandem genes were identified (Fig. 3e). Both proximal tandem duplication events (Fig. 3f) and chromosome segment duplication events (Fig. 3g) can be observed, and we note that these events can follow each other. Taking the tandem genes of groups A (207-208-209-210-211) and B (241-242-243-244) as an example, two interpretations can be made: (i) a copy-paste event occurred from group A to group B, and a gene was lost afterward; (ii) there was a copy-paste event from group B to group A, and subsequently, a proximal tandem duplication gave rise to gene 208. The phylogenetic trees for the other large gene families can be seen in Supplementary Fig. 3 posted at https://www.giantviruses.com/sup-material-of-papers/sup-material-gene-duplication-as-a-major-force-driving-the-genome-expansion-in-some-giant-viruses. We quantified the two events (proximal tandem duplications and chromosomal segment duplications) for the six largest gene families (see Supplementary Fig. 4 posted at https://www.giantviruses.com/sup-material-of-papers/sup-material-gene-duplication-as-a-major-force-driving-the-genome-expansion-in-some-giant-viruses) and noticed that they are frequent and seem to have a great influence on the genome evolution of this virus.

After a gene duplication event, the copies follow different evolutionary paths. When one of the copies suffers an extreme reduction in its coding sequence (CDS) caused by the emergence of a premature stop codon, we can infer that a pseudogenization event has occurred (resulting in a progressive loss of gene function). Our data suggest that there was a significant difference in CDS length and identity among genes within the

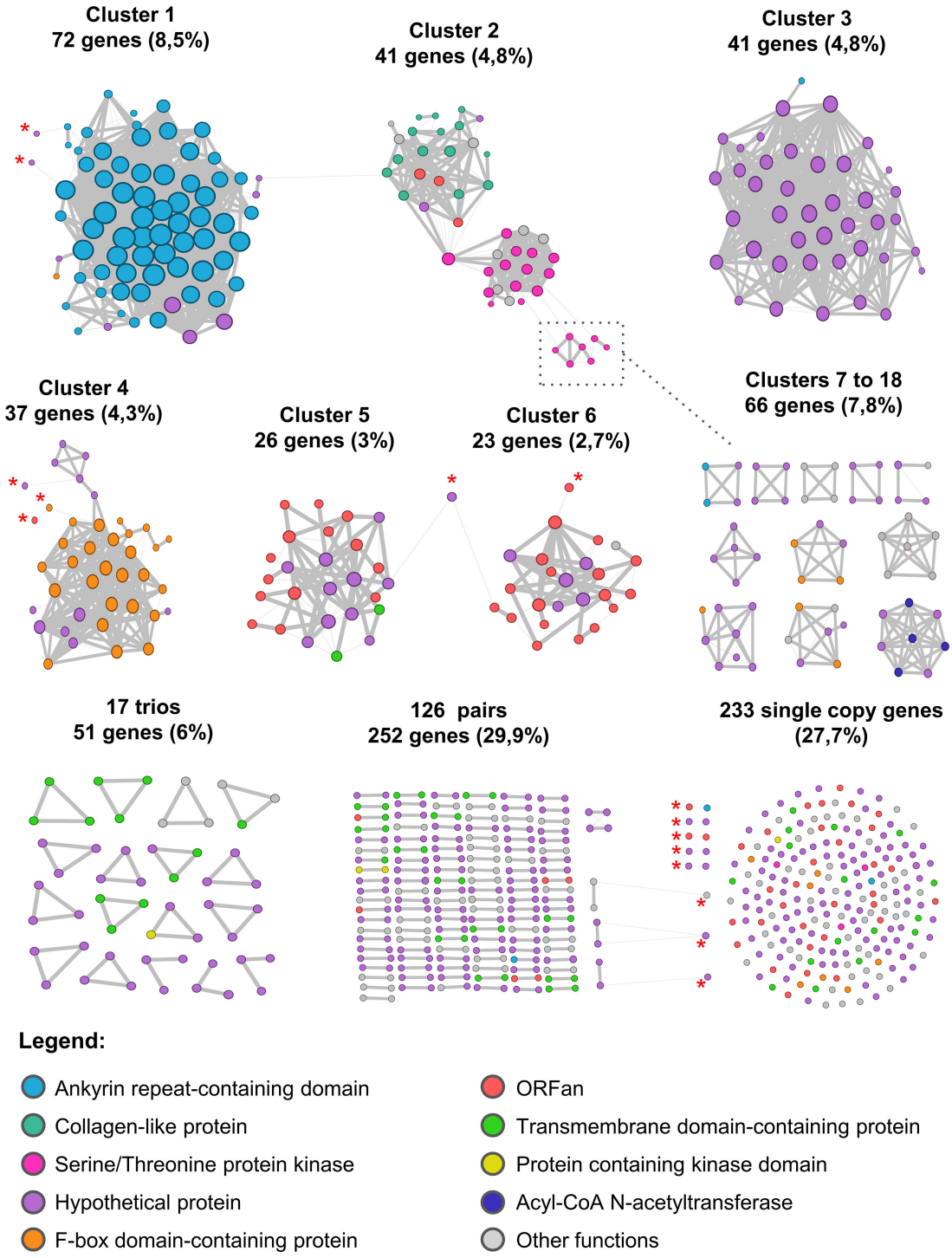

FIG 2 Network of clusters, trios, pairs, and single copy genes in the c. pambiensis genome. Reciprocal BLASTp-hits (coverage ≥30 and e-value < 1e−-4) between two proteins are represented by thicker lines, while non-reciprocal BLASTp-hits are represented by thinner lines. Considering the defined criteria, to be part of the cluster, the gene must: (1) have a reciprocal match with some gene within the cluster (thick line) or (2) have at least two non-reciprocal matches with genes within the cluster (thin line). Genes that only had one non-reciprocal match with a gene within the cluster were not considered part of that cluster. Asterisks indicate single genes, which were not included within the clusters according to the aforementioned criteria. The color legend is provided below the image. The square highlights cluster 10, in which has non-reciprocal hits with cluster 2.

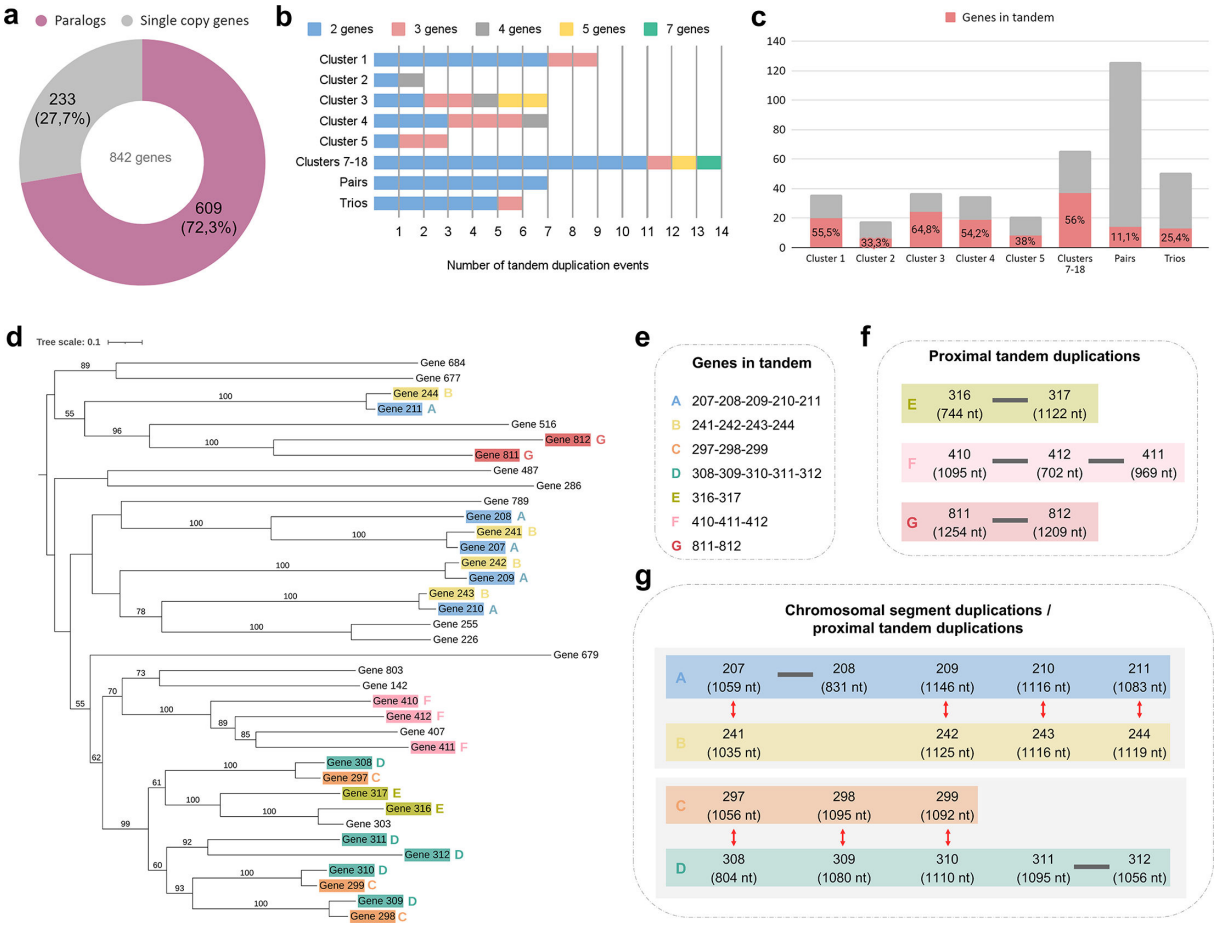

FIG 3 Gene duplication analyses in the c. pambiensis genome. (a) Number and percentage of genes in paralog groups or single copy genes in c. pambiensis genome. (b) Number of tandem duplication events observed within paralog clusters with more than three genes and low-copy number clusters composed of trios and pairs of genes. (c) Percentage of tandem duplication events observed within paralogs clusters, trios, and pairs. (d) Maximum likelihood phylogenetic tree constructed with protein sequences encoded by cluster 3 genes. Tandem genes are highlighted and colored in the tree according to the organization shown in (e). (e) Tandem genes highlighted in (d). They were organized in ascending order. (f) Proximal tandem duplication events are identified in (d). (g) Chromosomal segment duplications plus proximal duplication events identified in (d). In (f) and (g), the length of the gene in nucleotides is depicted below the gene ID.

same family, indicating non-negligible pseudogenization (Fig. 4a, and see Supplementary Fig. 5 and 6 posted at https://www.giantviruses.com/sup-material-of-papers/sup-material-gene-duplication-as-a-major-force-driving-the-genome-expansion-in-some-giant-viruses). In addition to gene size/coverage variation, we also observed considerable sequence divergence of the paralogous proteins, so that some of the cluster members were not reciprocally identifiable as homologous in BLASTp analysis, suggesting independent and progressive evolution after gene duplication (Fig. 3 and 4, and see Supplementary Fig. 5 and 6 posted at https://www.giantviruses.com/sup-material-of-papers/sup-material-gene-duplication-as-a-major-force-driving-the-genome-expansion-in-some-giant-viruses).

The chromosome position of the paralogs can provide clues about how genome expansion has evolved. We observed that some clusters (1, 3, and 4) appear to have more gene copies concentrated in a given region within the genome, while others (2, 5, and 6) are more spread throughout the genome (Fig. 4b). But in general, the paralogs belonging to those six major gene families are scattered throughout the c. pambiensis genome (see Supplementary Fig. 7 posted at https://www.giantviruses.com/sup-material-of-papers/

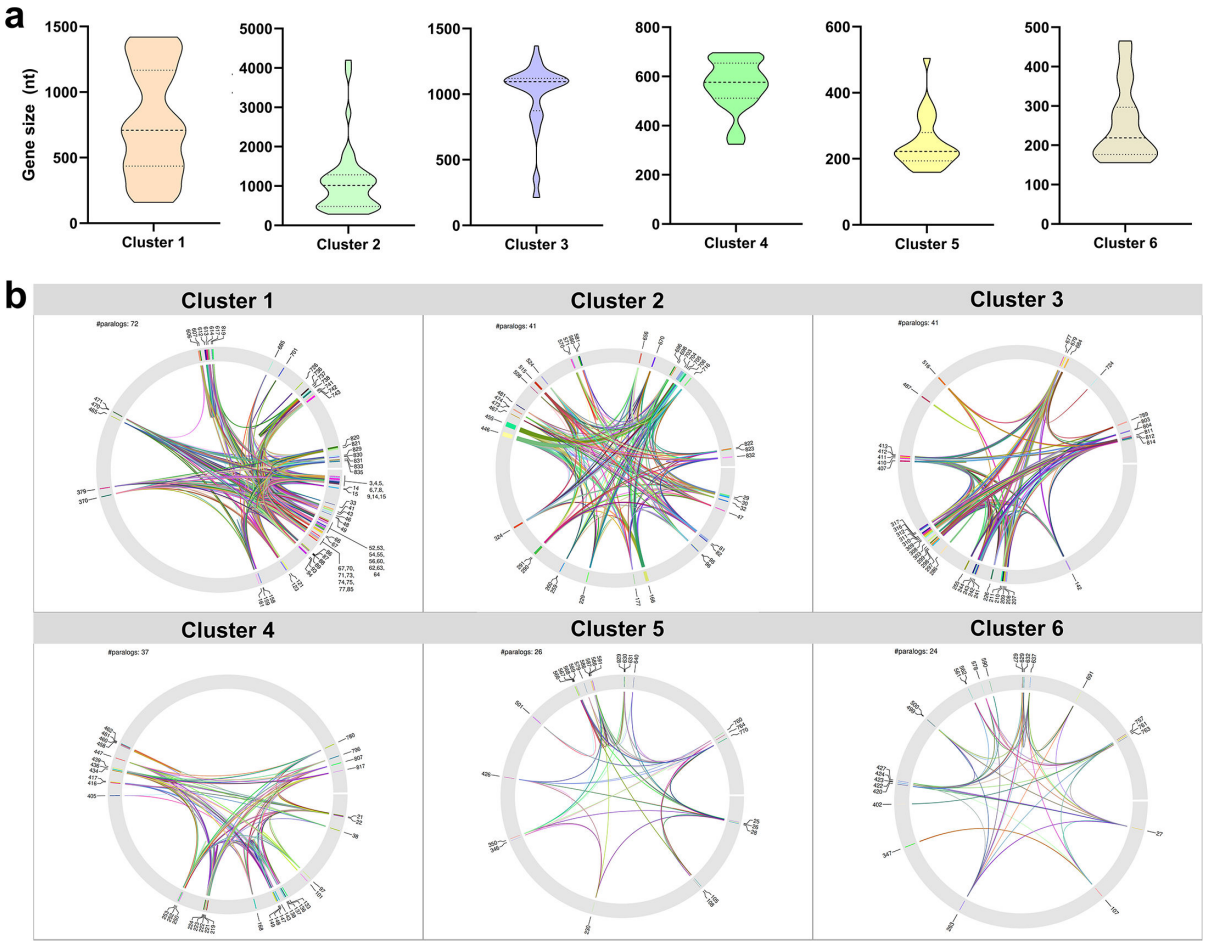

FIG 4 Size and location of paralogous genes comprising each large family of genes (clusters 1 to 6). (a) Violin plots showing gene size variation considering each cluster. Coding sequences far below the mean may suggest possible pseudogenization events. The dashed line represents the mean. Dotted lines delimit the interquartile range. (b). Gene location considering each cluster. Clusters 1, 3, and 4 seem to have some polarity in the genome, while clusters 2, 5, and 6 do not seem to have any pattern and are scattered throughout the genome.

sup-material-gene-duplication-as-a-major-force-driving-the-genome-expansion-in-some-giant-viruses), indicating multiple and successive events of gene/chromosome segment duplication.

Gene duplication as a driving force of genome gigantism in giant viruses

To explore the prevalence and distribution of paralogs in cedratviruses, we expanded our analysis to include all available cedratvirus genomes in public databases (Fig. 5). By performing BLASTp searches of predicted proteins against the complete set of proteins of each cedratvirus, we identified a range of gene families with varying sizes and predicted functions. Firstly, it is noteworthy that c. pambiensis has an atypical and unprecedented relative abundance of duplicated genes compared to other cedratviruses, accounting for 72.3% of its genome. Nevertheless, a significant contribution of paralogs was observed in all cedratvirus genomes, ranging from 43.34% in Brazilian cedratvirus to 52.26% in c. lausannensis. In addition, all cedratviruses share some large gene families, such as families mainly related to ankyrin repeat-containing domain and hypothetical proteins, as well as other functions such as collagen-like proteins, serine/threonine kinases and F-box domain-containing proteins. We further expanded our analysis to include the pithovirus-like group, to which cedratviruses belong. Interestingly, pithovirus and orpheovirus present a similar proportion of duplicated genes in their genomes, at 42.61% and 52.29%, respectively, but other protein domains were

overrepresented, such as collagen-like and MORN-repeat proteins, suggesting that extensive duplications have occurred independently following the radiation of the pithovirus-like group.

Expanding the analysis to other members of the phylum *Nucleocytoviricota* and yaravirus revealed that gene duplications are, again, quite abundant (Fig. 6a). In addition to cedratviruses, gene duplication also seems to be an important factor in the evolution of the genome of other giant viruses, particularly for pandoraviruses (mean of 47.28%) and some mimiviruses (e.g., 49.46% for cotonvirus). Despite the different proportions, all analyzed genomes have duplicated genes. However, it is important to highlight that even after expanding our analysis to other nucleocytoviruses, c. pambiensis remains the virus with the highest percentage of the genome composed of duplicated genes (72.3%). We note, however, that for some viruses, mechanisms other than gene duplication may be acting synergistically or concurrently.

It is generally expected to observe a positive correlation between viral genome size and the number of genes. The isolation of c. pambiensis raised questions about this correlation and the existence of a correlation between genome size and the number of paralogs. To address these questions, we compared genome size with the overall number of genes and the number of duplicated genes across a large sample of giant viruses (Fig. 6b and c). As aforementioned, variations on gene prediction methods must be considered, but the overall available data strongly suggest the presence of a strong positive correlation between genome size and both the total number of genes ($\rho = 0.90$, *P*-value $< 2.2e{-}16$) and the number of paralogs ($\rho = 0.87$, *P*-value = 1.4e−-14) per genome. The linear regression analysis reveals that, although all cedratviruses show a similar correlation between genome size and the number of predicted genes, c. pambiensis stands out as an exceptional case due to the significant contribution of paralogs in its genome (Fig. 6c).

DISCUSSION

The genome gigantism observed in giant viruses represents an intriguing unanswered question. A number of giant viruses have been discovered in recent years, revealing an increasing variety of particles, genome sizes, and predicted genes. Although deserving attention has been given to the functional content of the giant virus genomes, few studies have investigated why the genomes of these viruses are so large, reaching up to 2.8 Mb (8–11, 14, 16). The early efforts to answer this question were hampered by the scarcity of genomic information available at the time, precluding generalizing conclusions. However, the constant efforts of several research teams to isolate novel giant viruses worldwide have now set the stage for a more comprehensive analysis of giant virus genome evolution. Here, we presented the evidence that gene duplication is a primary mechanism for genome expansion among several groups of giant viruses.

Gene duplication has been recognized as an important source of genetic diversity in cellular organisms (2–5). Through gene duplication, new functions can emerge, as duplicated genes typically experience lower negative selection pressures, and the encoded proteins can gain new properties and functions. Gene families within a given organism typically emerge as a result of duplication events followed by divergence (20). In addition to potential functional divergence, the multi-copy gene families can contribute to the increased gene expression via so-called gene dosage phenomenon. Some of the well-known examples of multi-gene families include genes encoding cytomotive filament-forming proteins, globins, and ribosomal units (20–22). As a result, gene duplication is consistently considered a major mechanism in the evolution of cellular organisms. However, the understanding of the consequences of gene duplication in viruses remains limited. There are ample examples showing that gene duplication followed by exaptation of one of the gene copies has played a key role in adaptation and diversification throughout virus evolution (23, 24). Nevertheless, in viruses with small capsids, experimental studies have shown that gene duplications are prohibitive and lead to the loss of infectivity, primarily dictated by the limited packaging capacity of

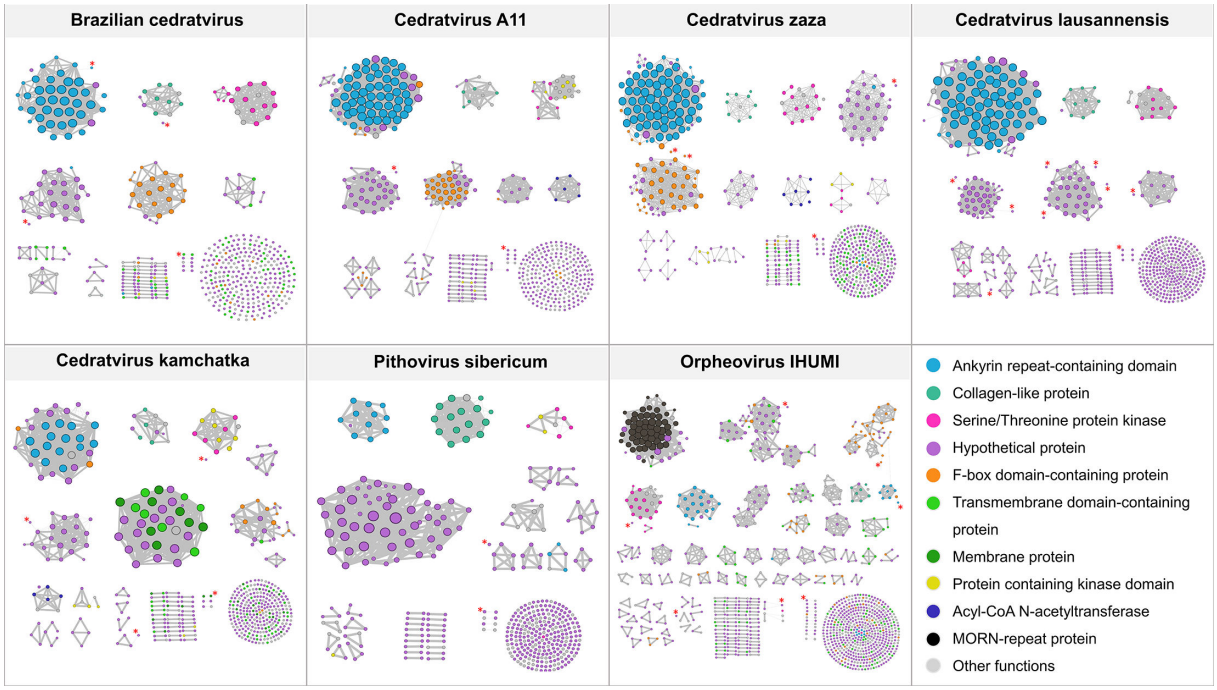

FIG 5 Comparison of gene families in the genome of members of the pithovirus-like group. These networks were made in the same way as for c. pambiensis (Fig. 2). Color legend is provided on the image.

small capsids (25). The situation in giant viruses, which show overall low packaging densities (26), is radically different compared to viruses with small capsids. Thus, genome evolution in giant viruses is apparently not constrained by the capsid size, allowing them to reap the benefits of gene duplication, which has shaped the genomes of cellular organisms.

In this study, we have described that genes in c. pambiensis appear to duplicate through several mechanisms, involving both tandem gene and distal genomic segment duplications. Tandem duplications in cellular organisms may occur through replication slippage, ectopic recombination, or aberrant DNA break repair. Distal duplications typically involve unequal crossing-over rearrangements of gene clusters with similar gene content. This mechanism may become progressively more frequent as the number of paralogs increases, providing more regions with similar content available for recombination (1). It is notable that the largest paralogous gene families encode proteins that themselves consist of repetitive domains, such as ankyrin repeats, leucine-rich repeats, and MORN repeats. Conceivably, the repetitive nucleotide sequences within these genes promote both tandem and long-distance genomic duplications. Furthermore, considering that a substantial number of viral genome copies are produced and compacted within viral factories, and considering the fact that the cedratvirus genome is circular dsDNA, it is reasonable to believe that ectopic recombination and unequal crossing-over may generate both tandem and distal duplications in cedratviruses. Additionally, the role of transposons should be considered in relation to gene duplication, as they have been described in the genomes of amoeba and certain groups of giant viruses (27, 28). Considering that the *Nucleocytoviricota* may have arisen from smaller and simpler viruses infecting early eukaryotes (29–31), the expansion of their genomes might involve a general mechanism conserved in the entire phylum. While genome expansion by extensive gene gain through horizontal gene transfer and *de novo* gene creation seem to be characteristic of only some groups of nucleocytoviruses (13, 15), our data indicate that gene duplication is an evolutionary mechanism common for several groups of giant viruses, especially those capable of infecting amoebas.

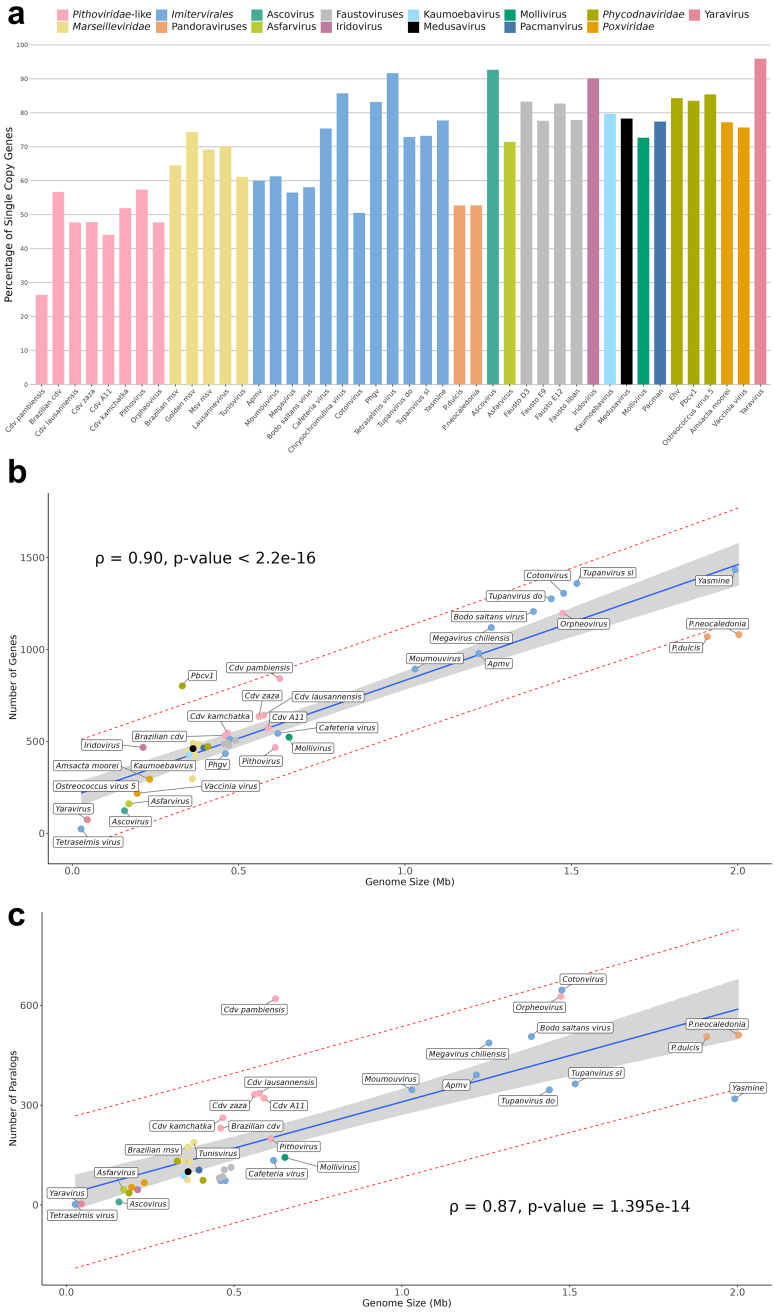

FIG 6 The contribution of paralogs genes in the genomes of giant viruses. (a) Comparison of the percentage of single copy genes in the genome for members of *Nucleocytoviricota* and yaravirus. The representatives of each group were depicted with distinct colors, as described in the legend. The percentage of single-copy genes is shown for each genome. Spearman correlation plots of the relationship between genome size and total number of predicted genes (b) and the relationship between genome size and total number of paralogs (c). Yasminevirus and c. pambiensis are outliers. The solid blue line marks the linear regression while the shaded gray area illustrates the 95% confidence interval associated with the linear regression line. The outer red lines delineate the 95% prediction interval, encapsulating the range within which we anticipate 95% of gene/paralog numbers to fall based on a given genome size. ρ: Spearman correlation coefficient; *p*-value: associated *P*-value.

One question that arises is why cedratviruses have so many duplicated genes in their genome. Maintaining duplicated copies may be important for creating genetic redundancy, and protecting the virus from deleterious mutations in essential genes since

additional copies could maintain the organism's functionality and fitness (32). Indeed, some of the c. pambiensis duplicated genes are potentially essential, such as the major capsid protein, early transcription factors, and transcriptional enzymes. Furthermore, as aforementioned, gene duplication is a phenomenon that provides raw material for evolution. The additional gene copies can be repurposed for functions unrelated to those of the original genes, which appears to be the main trend in virus evolution (23, 33). Furthermore, the duplicated genomic regions provide the raw genetic material for *de novo* gene emergence, a route extensively explored by pandoraviruses (13). Both mechanisms can lead to genetic innovation, increasing the genetic repertoire of these viruses. This is supported by the presence of different identifiable domains/functional categories in certain clusters of paralogous genes in c. pambiensis and other pitho-like viruses (Fig. 2 and 5). Further exploration of the giant virus diversity should further refine our understanding of the mechanisms of genome expansion and evolutionary traits associated with this remarkable group of viruses.

MATERIALS AND METHODS

Sample collection and isolation

To obtain the sample, a plastic container was placed in a small, forested area at the UFMG campus for a few days, collecting rainwater and organic matter present in that environment. From this collected water, the method of co-culture with amoebae of the species *Acanthamoeba castellanii* was carried out, as described in a published protocol (34). Collection authorization: SISBIO 89441-1. Brazilian genetic resources access authorization: SISGEN A2291C9.

Production, purification, and titration

To produce the new isolate, *A. castellanii* were infected with an MOI of 0.01 in glass culture flasks (300 cm^2) with 35 mL of PYG medium and kept at 30°C in a rotary cell oven. After complete lysis of the cells, the contents of the flask were collected. This content was added to a sucrose cushion (40%) and then ultracentrifuged in a Combisorvall Rotor AH-62va centrifuge at 14,000 rotations per minute (AH-629 rotor) for 1 hour, between 4°C and 8°C. The final pellet was resuspended in phosphate-buffered saline. Titration was performed by the limiting dilution method (35) in 96-well plates and the titer was expressed in $TCID_{50}$ per milliliter. The viral stock was kept at −20°C until use.

Electron microscopy and one-step growth curve

Three electron microscopy methods were used during this work, for a better description of the viral particle: TEM, negative contrast electron microscopy, and SEM. For TEM, *A. castellanii* cells cultivated in a PYG medium were infected with an MOI of 0.01 for 24 hours. Cells were then collected and fixed with 2.5% glutaraldehyde + phosphate for 2 hours at room temperature. Subsequently, fixation was performed with 2% osmium tetroxide, and incorporation in EPON resin, in sequence ultrathin sections was made. Image analyses were performed using a transmission electron microscope (FEI SpiritBiotwin 120 kV). For NSEM, the purified virus was diluted 1:10 in water, and 3 µL of this diluted sample was applied onto glow-discharged 400-mesh copper grids covered with a Lacey carbon support film and an ultrathin carbon layer (15 mA, negative charge for 40 seconds, 01824—Ted Pella, USA). After 1 minute, the excess liquid was drained gently touching the edge of the grid with a filter paper. The grid was stained twice with 3 µL uranyl acetate solution (2%) for 30 seconds. The excess solution was drained with filter paper and the grid was allowed to dry at room temperature. Images were collected using a 4k × 4k Ceta CMOS Camera coupled on a Talos F200C Transmission Electron Microscope (200 kV, Thermo Fisher Scientific) at LNNano/CNPEM. For SEM, *A. castellanii* cultivated in PYG medium were infected with an MOI of 0.01 for 24 hours. The cells were then collected, transferred to a coverslip containing poly-L-lysine, and fixed

with 2.5% glutaraldehyde + cacodylate for 2 hours at room temperature. Subsequently, fixation was performed with 1% osmium tetroxide, washing with 0.1 M cacodylate buffer, and immersion in 0.1% tannic acid. Then, dehydration was performed using serial passages in ethanol solutions with different concentrations. Subsequently, a critical point drying process using CO_2 was carried out. Finally, the samples were accommodated in metal supports (called stubs) and metalized with a layer of gold. Image analyses were performed using a scanning electron microscope (FEI Quanta 200 FEG).

For the one-step-growth curve assay, *A. castellanii* cells were infected in duplicates with an MOI of 10 to obtain a synchronous cycle. After 30 minutes of adsorption, the inoculum was removed, and fresh PYG medium was added. The collections of the supernatant with the cells were performed at 0, 1, 3, 6, 9, 12, 24, 48, and 72 hpi, considering the time 0 hpi right after the adsorption. All times were titrated later, and the curve was constructed from the titration result.

Sequencing, assembly, annotation, and phylogenetic analyses

The samples containing the purified virus were sequenced twice using the equipment Illumina MiSeq, with a paired-end library using the kit Illumina DNA Prep (Illumina Inc., San Diego, CA, USA). First, a *de novo* assembly was performed using the SPAdes 3.13.1 software (36). To increase the *de novo* assembly, the SOAPdenovo2 1.12 (37) program was used. Subsequently, a reference genome (the best hit, Brazilian cedratvirus) was used in the Medusa 1.3 program (38) and the final genome was obtained. For the prediction of the open reading frames (ORFs), the GeneMarkS 4.28 software (39) was used, and the sequences smaller than 50 amino acids were removed from the analyses. Gene prediction using GeneMarkS was performed using both prokaryotic and viral parameters. Since the results were very similar, we opted to use data from prokaryotic parameters because several studies on giant viruses employed this strategy. Therefore, some c. pambiensis (and other pitho-like viruses) genes were predicted to start with alternative start codons, different from ATG. The functional annotation of the predicted proteins was performed using BLASTp against the NCBI NR database considering 1e−5 e-value. The annotation was also done for the six largest clusters of genes using HHpred, with similar results.

For the construction of the phylogenetic trees, the amino acid sequences of the viruses of interest were obtained by the BLASTp tool (default parameters) from the NCBI Genbank database (40). These sequences were aligned with that of c. pambiensis using the MUSCLE 3.8.1551 software (41). The phylogenetic tree was built by the IQtree 1.6.12 program (42) using the best-fitted VT+F+R5 model for amino acid substitution and the likelihood-based method aLRT SH-like with 1,000 pseudoreplicates to estimate branch support values. As aforementioned, substantial variation in coverage was observed in genes belonging to the six largest clusters. This result posed challenges to our phylogenetic analyses due to the potential lack of negative selection in pseudogenes that leads to sequence degeneration. To improve the reliability of our analyses, we defined a cut-off point, in which the gene with the longest CDS was considered as a reference and all the other genes with a CDS shorter than half the size of this gene were removed from the alignment. For cluster 2, the procedure was a little different, as it has three genes much longer than all the others (mean of 427.5% larger than the family median). Therefore, for this cluster, the three largest genes were removed from the analysis and the fourth largest one was considered as a reference.

Detection and mapping of duplicated genes

To detect duplicated genes, we BLASTp the predicted proteins of c. pambiensis against themselves, and hits with coverage ≥30% and e-value $<1e^{-4}$ were considered paralogs. Higher stringent cutoffs were evaluated (i.e., 50%) revealing very similar results. To cluster the paralogs, the Gephi 0.9.7 software (43) was used, based on the list of hits obtained in the previous step. To better understand the duplication events happening within the

genome, we constructed phylogenetic trees for the six largest gene families (more than 20 genes) using amino acid sequences. The programs and parameters used to build these trees were the same used previously.

With the groups of paralogs established, we decided to investigate how these genes spread within the genome. The gene and protein prediction steps gave us the coordinates of each gene, so we developed an R script to construct a Circos plot-like using the circlize 0.4.15 package (44) that draws a line between paralog genes.

Statistical analysis

Spearman correlations were used to assess correlations between genome size and total number of paralogs or total number of genes with a significance level of $P < 0.05$. Data distribution was assessed by the Shapiro-Wilk test. Testing and plotting results were all done in Rstudio (45).

ACKNOWLEDGMENTS

We thank the Laboratório de Vírus of Universidade Federal de Minas Gerais for all the support provided and the Microscopy Center of UFMG, in particular the technicians Denilson Cunha, Rodrigo Ferreira, Altair Mendes, Thalita Arantes, Marilene Oliveira, and Breno Moreira, who collaborated from the preparation to the session for observation of the samples.

We would also like to thank the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), and Pró-Reitorias de Pesquisa e Pós-Graduação da UFMG (PRPG-UFMG) for the financial support, and LNNano/CNPEM for access to the EM facility via project 20230751. J.S.A., J.P.A.J., and F.R.S. are CNPq researchers.

T.B.M., L.E.D.B., and J.S.A. designed the study and experiments. All authors performed experiments and/or analyses. All authors wrote the manuscript. The text has been entirely written by the authors, and English revision was partially performed using artificial intelligence. All authors approved the final manuscript.

AUTHOR AFFILIATIONS

[1]Laboratório de Vírus, Departamento de Microbiologia, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Brazil
[2]Laboratório de Genômica Evolutiva, Departamento de Genética, Evolução, Microbiologia e Imunologia, Instituto de Biologia, Universidade Estadual de Campinas (UNICAMP), Campinas, Brazil
[3]Laboratório de Virologia, Departamento de Microbiologia e Imunologia, Instituto de Biotecnologia, Universidade Estadual Paulista (UNESP), Botucatu, Brazil
[4]Del-Bem Lab, Departamento de Botânica, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Brazil
[5]Brazilian Biosciences National Laboratory (LNBio), Brazilian Center for Research in Energy and Materials (CNPEM), Campinas, Brazil
[6]Centre de Recherche du Centre Hospitalier Universitaire de Québec- Université Laval, Laval, Québec, Canada
[7]Laboratório de Microbiologia Molecular, Universidade Feevale, Novo Hamburgo, Brazil
[8]Archaeal Virology Unit, Institut Pasteur, Université Paris Cité, CNRS UMR6047, Paris, France
[9]Department of Biological Sciences, Virginia Tech, Blacksburg, Virginia, USA
[10]Center for Emerging, Zoonotic, and Arthropod-Borne Infectious Disease Virginia Tech, Blacksburg, Virginia, USA

AUTHOR ORCIDs

Rodrigo A. L. Rodrigues http://orcid.org/0000-0001-7148-4012
Frank O. Aylward http://orcid.org/0000-0002-1279-4050
Luiz-Eduardo Del-Bem http://orcid.org/0000-0001-8472-4476
Jônatas S. Abrahão http://orcid.org/0000-0001-9420-1791

FUNDING

Funder	Grant(s)	Author(s)
Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)	303680/2022-9	Jônatas S. Abrahão
Ministério da Ciência, Tecnologia e Inovação (MCTI)	405249/2022-5, 406441/2022-7	Jônatas S. Abrahão
Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES)	88882.348380/2010-1	Jônatas S. Abrahão

AUTHOR CONTRIBUTIONS

Talita B. Machado, Data curation, Formal analysis, Investigation, Methodology, Writing – original draft | Agnello C. R. Picorelli, Data curation, Formal analysis, Investigation | Bruna L. de Azevedo, Formal analysis, Writing – original draft | Isabella L. M. de Aquino, Formal analysis, Investigation, Writing – original draft | Victória F. Queiroz, Data curation, Formal analysis, Investigation, Methodology, Writing – original draft | Rodrigo A. L. Rodrigues, Formal analysis, Methodology, Writing – original draft | João Pessoa Araújo Jr., Methodology, Writing – original draft | Leila S. Ullmann, Methodology | Thiago M. dos Santos, Methodology, Writing – original draft | Rafael E. Marques, Methodology, Writing – original draft | Samuel L. Guimarães, Methodology, Writing – original draft | Ana Cláudia S. P. Andrade, Investigation, Writing – original draft | Juliana S. Gularte, Methodology, Writing – original draft | Meriane Demoliner, Methodology, Writing – original draft | Micheli Filippi, Methodology, Writing – original draft | Vyctoria M. A. G. Pereira, Investigation, Writing – original draft | Fernando R. Spilki, Methodology, Writing – original draft | Mart Krupovic, Investigation, Methodology, Validation, Writing – original draft | Frank O. Aylward, Investigation, Writing – original draft | Luiz-Eduardo Del-Bem, Conceptualization, Funding acquisition, Investigation, Methodology, Validation, Writing – original draft.

DATA AVAILABILITY

The genome of cedratvirus pambiensis is available at GenBank under accession number OR343515. The genome sequence (fasta) is also available at our research group website (https://5c95043044c49.site123.me/sup-material-of-papers/sup-material-gene-duplication-as-a-major-force-driving-the-genome-expansion-in-some-giant-viruses).

REFERENCES

1. Krebs JE, Goldstein ES, Kilpatrick ST. 2017. Lewin's genes twelve. Jones & Bartlett Learning.
2. Cannon SB, Mitra A, Baumgarten A, Young ND, May G. 2004. The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. BMC Plant Biol 4:10. https://doi.org/10.1186/1471-2229-4-10
3. Reams AB, Neidle EL. 2004. Selection for gene clustering by tandem duplication. Annu Rev Microbiol 58:119–142. https://doi.org/10.1146/annurev.micro.58.030603.123806
4. Wapinski I, Pfeffer A, Friedman N, Regev A. 2007. Natural history and evolutionary principles of gene duplication in fungi. Nature 449:54–61. https://doi.org/10.1038/nature06107
5. Persi E, Wolf YI, Karamycheva S, Makarova KS, Koonin EV. 2023. Compensatory relationship between low-complexity regions and gene paralogy in the evolution of prokaryotes. Proc Natl Acad Sci U S A 120:e2300154120. https://doi.org/10.1073/pnas.2300154120
6. Li WH, Gu Z, Wang H, Nekrutenko A. 2001. Evolutionary analyses of the human genome. Nature 409:847–849. https://doi.org/10.1038/35057039
7. Zhang J. 2003. Evolution by gene duplication: an update. Trends Ecol Evol 18:292–298. https://doi.org/10.1016/S0169-5347(03)00033-8
8. Scola BL, Audic S, Robert C, Jungang L, de Lamballerie X, Drancourt M, Birtles R, Claverie J-M, Raoult D. 2003. A giant virus in amoebae. Science 299:2033–2033. https://doi.org/10.1126/science.1081867
9. Philippe N, Legendre M, Doutre G, Couté Y, Poirot O, Lescot M, Arslan D, Seltzer V, Bertaux L, Bruley C, Garin J, Claverie J-M, Abergel C. 2013. Pandoraviruses: amoeba viruses with Genomes up to 2.5 MB reaching that of parasitic eukaryotes. Science 341:281–286. https://doi.org/10.1126/science.1239181

10. Abrahão J, Silva L, Silva LS, Khalil JYB, Rodrigues R, Arantes T, Assis F, Boratto P, Andrade M, Kroon EG, Ribeiro B, Bergier I, Seligmann H, Ghigo E, Colson P, Levasseur A, Kroemer G, Raoult D, La Scola B. 2018. Tailed giant tupanvirus possesses the most complete translational apparatus of the known virosphere. Nat Commun 9:749. https://doi.org/10.1038/s41467-018-03168-1
11. Moniruzzaman M, Martinez-Gutierrez CA, Weinheimer AR, Aylward FO. 2020. Dynamic genome evolution and complex virocell metabolism of globally-distributed giant viruses. Nat Commun 11:1710. https://doi.org/10.1038/s41467-020-15507-2
12. Boyer M, Yutin N, Pagnier I, Barrassi L, Fournous G, Espinosa L, Robert C, Azza S, Sun S, Rossmann MG, Suzan-Monti M, La Scola B, Koonin EV, Raoult D. 2009. Giant marseillevirus highlights the role of amoebae as a melting pot in emergence of chimeric microorganisms. Proc Natl Acad Sci U S A 106:21848–21853. https://doi.org/10.1073/pnas.0911354106
13. Legendre M, Fabre E, Poirot O, Jeudy S, Lartigue A, Alempic J-M, Beucher L, Philippe N, Bertaux L, Christo-Foroux E, Labadie K, Couté Y, Abergel C, Claverie J-M. 2018. Diversity and evolution of the emerging pandoraviridae family. Nat Commun 9:2285. https://doi.org/10.1038/s41467-018-04698-4
14. Filée J. 2015. Genomic comparison of closely related giant viruses supports an accordion-like model of evolution. Front Microbiol 6:593. https://doi.org/10.3389/fmicb.2015.00593
15. Filée J, Pouget N, Chandler M. 2008. Phylogenetic evidence for extensive lateral acquisition of cellular genes by nucleocytoplasmic large DNA viruses. BMC Evol Biol 8:320. https://doi.org/10.1186/1471-2148-8-320
16. Filée J, Chandler M. 2008. Convergent mechanisms of genome evolution of large and giant DNA viruses. Res Microbiol 159:325–331. https://doi.org/10.1016/j.resmic.2008.04.012
17. Yutin N, Wolf YI, Koonin EV. 2014. Origin of giant viruses from smaller DNA viruses not from a fourth domain of cellular life. Virology 466–467:38–52. https://doi.org/10.1016/j.virol.2014.06.032
18. Koonin EV, Yutin N. 2019. Evolution of the large nucleocytoplasmic DNA viruses of eukaryotes and convergent origins of viral gigantism. Adv Virus Res 103:167–202. https://doi.org/10.1016/bs.aivir.2018.09.002
19. Suhre K. 2005. Gene and genome duplication in acanthamoeba polyphaga Mimivirus. J Virol 79:14095–14101. https://doi.org/10.1128/JVI.79.22.14095-14101.2005
20. Moradkhani K, Préhu C, Old J, Henderson S, Balamitsa V, Luo H-Y, Poon M-C, Chui DHK, Wajcman H, Patrinos GP. 2009. Mutations in the paralogous human α-globin genes yielding identical hemoglobin variants. Ann Hematol 88:535–543. https://doi.org/10.1007/s00277-008-0624-3
21. Wickstead B, Gull K. 2011. The evolution of the cytoskeleton. J Cell Biol 194:513–525. https://doi.org/10.1083/jcb.201102065
22. Malik Ghulam M, Catala M, Reulet G, Scott MS, Abou Elela S. 2022. Duplicated ribosomal protein paralogs promote alternative translation and drug resistance. Nat Commun 13:4938. https://doi.org/10.1038/s41467-022-32717-y
23. Koonin EV, Dolja VV, Krupovic M. 2022. The logic of virus evolution. Cell Host Microbe 30:917–929. https://doi.org/10.1016/j.chom.2022.06.008
24. Butkovic A, Dolja VV, Koonin EV, Krupovic M. 2023. Plant virus movement proteins originated from jelly-roll capsid proteins. PLoS Biol 21:e3002157. https://doi.org/10.1371/journal.pbio.3002157
25. Willemsen A, Zwart MP, Higueras P, Sardanyés J, Elena SF. 2016. Predicting the stability of homologous gene duplications in a plant RNA virus. Genome Biol Evol 8:3065–3082. https://doi.org/10.1093/gbe/evw219
26. Chaudhari HV, Inamdar MM, Kondabagil K. 2021. Scaling relation between genome length and particle size of viruses provides insights into viral life history. iScience 24:102452. https://doi.org/10.1016/j.isci.2021.102452
27. Filée J. 2018. Giant viruses and their mobile genetic elements: the molecular symbiosis hypothesis. Curr Opin Virol 33:81–88. https://doi.org/10.1016/j.coviro.2018.07.013
28. Sun C, Feschotte C, Wu Z, Mueller RL. 2015. DNA transposons have colonized the genome of the giant virus pandoravirus salinus. BMC Biol. 13:38. https://doi.org/10.1186/s12915-015-0145-1
29. Krupovic M, Dolja VV, Koonin EV. 2023. The virome of the last eukaryotic common ancestor and eukaryogenesis. Nat Microbiol 8:1008–1017. https://doi.org/10.1038/s41564-023-01378-y
30. Bisio H, Legendre M, Giry C, Philippe N, Alempic J-M, Jeudy S, Abergel C. 2023. Evolution of giant pandoravirus revealed by CRISPR/Cas9. Nat Commun 14:428. https://doi.org/10.1038/s41467-023-36145-4
31. Koonin EV, Krupovic M, Yutin N. 2015. Evolution of double-stranded DNA viruses of eukaryotes: from bacteriophages to transposons to giant viruses. Ann N Y Acad Sci 1341:10–24. https://doi.org/10.1111/nyas.12728
32. Rodrigues RAL, Andreani J, Andrade ACDSP, Machado TB, Abdi S, Levasseur A, Abrahão JS, La Scola B. 2018. Morphologic and genomic analyses of new isolates reveal a second lineage of cedratviruses. J Virol 92:e00372-18. https://doi.org/10.1128/JVI.00372-18
33. Koonin EV, Krupovic M. 2018. The depths of virus exaptation. Curr Opin Virol 31:1–8. https://doi.org/10.1016/j.coviro.2018.07.011
34. Machado TB, de Aquino ILM, Abrahão JS. 2022. Isolation of giant viruses of *Acanthamoeba castellanii*. Curr Protoc 2:e455. https://doi.org/10.1002/cpz1.455
35. Reed LJ, Muench H. 1938. A simple method of estimating fifty per cent endpoints. Am J Epidemiol. 27:493–497. https://doi.org/10.1093/oxfordjournals.aje.a118408
36. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477. https://doi.org/10.1089/cmb.2012.0021
37. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Liu Y, Yu C, Wang B, Lu Y, Han C, Cheung DW, Yiu S-M, Peng S, Xiaoqian Z, Liu G, Liao X, Li Y, Yang H, Wang J, Lam T-W, Wang J. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 1:18. https://doi.org/10.1186/2047-217X-1-18
38. Bosi E, Donati B, Galardini M, Brunetti S, Sagot M-F, Lió P, Crescenzi P, Fani R, Fondi M. 2015. MeDUSa: a multi-draft based scaffolder. Bioinformatics 31:2443–2451. https://doi.org/10.1093/bioinformatics/btv171
39. Besemer J, Borodovsky M. 2005. Genemark: web software for gene finding in prokaryotes, eukaryotes and viruses. Nucleic Acids Res 33:W451–4. https://doi.org/10.1093/nar/gki487
40. Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2014. Genbank. Nucleic Acids Res 42:D32–7. https://doi.org/10.1093/nar/gkt1030
41. Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797. https://doi.org/10.1093/nar/gkh340
42. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol 37:1530–1534. https://doi.org/10.1093/molbev/msaa131
43. Bastian M, Heymann S, Jacomy M. 2009. Gephi: an open source software for exploring and manipulating networks. ICWSM 3:361–362. https://doi.org/10.1609/icwsm.v3i1.13937
44. Gu Z, Gu L, Eils R, Schlesner M, Brors B. 2014. Brors, circlize implements and enhances circular visualization in R. Bioinformatics 30:2811–2812. https://doi.org/10.1093/bioinformatics/btu393
45. Posit team. 2022. Rstudio: integrated development environment for R, Posit software. Available from: http://www.posit.co