

Integrative genomic analyses support the division of the extended *Asfarviridae* clade into multiple viral families

Thiago Mendonça-Santos,^{1,2} Jônatas Abrahão,³ Luiz-Eduardo Del-Bem^{1,4}

AUTHOR AFFILIATIONS See affiliation list on p. 11.

ABSTRACT Giant viruses continue to challenge our understanding of virology, blurring boundaries of what a virus can be. The so-called “extended *Asfarviridae*”—such as African swine fever virus, faustovirus, kaumoebavirus, pacmanvirus, and AbALV—has long puzzled taxonomists. By integrating comparative genomics, pangenomics, and phylogenomics, we show these lineages are deeply divergent, forming multiple families rather than one. This work underscores the huge unexplored diversity of giant viruses and demonstrates the value of integrative genomic analyses for proper taxonomic delineation.

KEYWORDS phylogenomics, viral taxonomy, pangenomics, giant viruses

Giant viruses, such as mimiviruses, exhibit unique characteristics, including virions visible under a light microscope, genomes larger than some unicellular organisms, and a diverse and unprecedented gene repertoire among viruses (1, 2). Their discovery revolutionized virology, challenging established concepts, underscoring the need to revisit conventional definitions, and highlighting the necessity for new isolation and sequencing studies to discover more of these long-overlooked viruses.

The African swine fever virus (ASFV), one of the largest known DNA viruses before the discovery of other giant viruses, remained the sole representative of the *Asfarviridae* family for decades and is the causative agent of African swine fever (ASF). However, other genetically related viruses have been isolated and placed within the *Asfarviridae* clade, leading to the formation of the “extended *Asfarviridae*” group to include ASFV and genetically related viruses such as Faustovirus, Kaumoebavirus, Pacmanvirus, and Abalone asfar-like virus (AbALV). Faustovirus, isolated in 2015 from wastewater (3), and kaumoebavirus, discovered in 2016 (4), both infect *Vermamoeba vermiformis*. Pacmanvirus, identified in 2017 from wastewater (5), replicates in *Acanthamoeba castellanii*. AbALV was first identified in 2020 and is known to infect abalones, causing high mortality rates (6).

Despite the apparent genetic relationship, there is still uncertainty about the inclusion of these associated viruses within the family *Asfarviridae* due to highly divergent characteristics such as differences in host range, environmental niches, genome sizes, and gene repertoire. Questions persist about whether these viruses should remain grouped under a single family or if they represent lineages distinct enough to warrant classification as founding members of entirely new viral families.

To address these issues, we performed comprehensive genomic and functional analyses and constructed a pangenome that reflects the gene content of the extended *Asfarviridae*. This approach provides an *in silico* framework for investigating the evolutionary relationships among these viruses, exploring their genetic similarities and differences, and characterizing their complex genomic architecture and functional diversity. By addressing this knowledge gap, our work contributes to a more precise

Editor Suchetana Mukhopadhyay, Indiana University
Bloomington, Bloomington, Indiana, USA

Address correspondence to Luiz-Eduardo Del-Bem,
delbem@usp.br.

The authors declare no conflict of interest.

See the funding table on p. 11.

Published 13 November 2025

Copyright © 2025 Mendonça-Santos et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](#).

taxonomic classification while revealing aspects of the genetic mechanisms that underlie their diversity, evolution, and environmental adaptation.

METHODOLOGY

Genome acquisition and gene prediction

Publicly available genomes of viral isolates from members of the extended *Asfarviridae* clade were downloaded from the NCBI database. Only complete genomes were downloaded. Gene prediction was performed for each of these genomes to eliminate gene prediction method bias and ensure methodological homogeneity. The genomes were submitted to GeneMarkS v4.28 (7) for the prediction of coding regions.

Predicted protein sequence clustering

To determine the families of orthologous genes among different viral species, the amino acid sequences of predicted proteins were grouped into clusters of potentially orthologous elements using the Proteinortho v6.0.14 tool (8). The parameters of 60% coverage and 20% identity for protein clustering were adopted based on a previous study (5). The pangenome was categorized according to the level of gene sharing as follows: Core, which includes genes present in 100% of the genomes; Soft-core, which includes genes present in 90% of the genomes; Shell, which includes genes present in three or more genomes but less than 90% of the genomes; and Cloud, which includes genes present in two or one genomes. The simplified representation of the pangenome was visualized through a flower plot generated using the custom script available at https://github.com/thiago22yow/flower_plot.

Python scripts were used to convert the table generated by Proteinortho into a binary presence-absence matrix of genes and to create an accumulation curve of the total pangenome and core genes (https://github.com/isabelschober/proteinortho_downstream). Heaps' law was applied to the presence-absence matrix of genes to calculate the α value to predict whether the pangenome is open or closed. The calculation was performed in R using the Micropan package (9) in "random" mode with 1,000 permutations.

Core gene phylogeny

Amino acid sequences for each of the identified core genes were aligned using MAFFT v7 (10) with the L-INS-i strategy. To enhance alignment quality, the resulting alignments were refined with Gblocks (11) to remove poorly aligned regions. These individual alignments were preserved for generating phylogenies of single core genes.

The alignments were then analyzed with IQ-TREE (12) to construct maximum likelihood phylogenetic trees. The LG+F+R4 substitution model, as recommended by ModelFinder (13) based on the Bayesian information criterion, was applied during tree construction. Branch support values (bootstrap) were calculated with 1,000 replicates. This step utilized resources from the Centro Nacional de Processamento de Alto Desempenho em São Paulo (CENAPAD-SP) and the Gloriosos clusters at the Laboratory of Genetics Biochemistry, UFMG.

To infer a consensus species tree while accounting for potential discordance among individual gene trees, we employed Weighted ASTRAL (wASTRAL v1.22.3.7) (14). This method improves phylogenomic inference by integrating both branch support values (bootstrap) and branch lengths to weight the contribution of quartets, effectively reducing the influence of low-confidence signals. Gene trees generated for each core gene by IQ-TREE were used as input. wASTRAL was executed with support-weighted and length-weighted hybrid criteria.

The phylogenetic tree was visualized in FigTree v1.4.4 (<https://tree.bio.ed.ac.uk/software/figtree>) and edited in Inkscape v1.2.1 (2022).

Functional analysis of clustered predicted proteins

To predict the functions of clustered proteins, the amino acid sequences in their respective pangenomic groups were mapped to the Giant Virus Orthologous Groups (GVOGs) (15), a curated set of protein families specific to the *Nucleocyotviricota*. This mapping was performed using Hidden Markov Models (HMMs) with the *hmmScan* tool from the HMMER package (16). The *hmmScan* tool identifies the presence of conserved domains within protein sequences by comparing them to pre-built HMM profiles of known protein families. Subsequently, functional annotations were inferred from the GVOG descriptions, based on the Cluster of Orthologous Groups classification.

Synten analysis

The conservation of gene order is an important indicator of shared evolutionary history. For this analysis, the predicted protein sequences were used in an all-against-all BLASTp (1e-5). The resulting BLASTp table was filtered, selecting hits with p-ident $\geq 20\%$ and $\geq 70\%$. MCScanX, a tool that detects gene collinearity in genome pairs (17), was then used for this purpose. MCScanX utilizes BLASTp outputs along with the GFF files generated by GeneMarkS containing the location of the predicted proteins to assess the presence of syntenic gene blocks among genome pairs. Two analyses were conducted, one for p-ident $\geq 20\%$ and another for p-ident $\geq 70\%$. The results were used to create a synten matrix, visualized and downloaded through Synvisio (18). The two matrices were then combined and edited in Inkscape v1.2.1.

AAI and ANI calculation

To estimate the levels of similarity between genomes in terms of nucleotide and amino acid composition, the genomes and amino acid sequences, respectively, were submitted to the ANI/AAI calculator (19) to calculate average nucleotide identity (ANI) and average amino acid identity (AAI) values. This tool estimates the average identity of nucleotides and amino acids through BLAST using a bidirectional approach where best hits (one-way ANI/AAI) and reciprocal best hits (two-way AAI/ANI) between two sets of genomic protein data are computed.

Matrices were created with the pairwise values and submitted to R for heatmap generation with pheatmap v1.0.12 (20). Hierarchical clusters were then calculated with the hclust function in R, using the complete linkage method, and the genomes were reorganized into dendrograms with the dendsort package (21) and then plotted with ggplot2. For comparison purposes, average AAI values were also calculated for genomes from other NCLDV families (Table S1).

GENOMIC DIVERSITY AND CONTENT PATTERNS REVEAL DISTINCT GROUPINGS WITHIN EXTENDED ASFARVIRIDAE

A search of the NCBI database for complete viral species genomes of extended *Asfarviridae* returned 39 complete genomes as of June 2023 (Table S2). All sequences were obtained in FASTA format. The analyzed attributes included genome length, architecture, and G + C content. Subsequently, the genomes were annotated for coding sequences using the GeneMarkS tool.

No statistically significant differences (Mann-Whitney $P < 0.05$) were found between the genome sizes of pacmanviruses and kaumoebaviruses, but significant divergences were found when comparing the other groups (Fig. S1A). It is noteworthy that there is a significant difference in genome length between ASFV and the faustoviruses, which possess the largest genomes within the clade. Faustoviruses have, on average, a genome 2.5 times larger than ASFV, which has the smallest genomic size in the group.

ASFV exhibits the lowest number of predicted proteins (Fig. S1B), whereas no statistically significant differences were observed among kaumoebaviruses, pacmanviruses, and faustoviruses. Although the genomes of kaumoebaviruses and pacmanviruses

are slightly shorter, they still maintain a number of proteins similar to those of faustoviruses, the group with the largest average genome size.

Faustoviruses and ASFV showed no significant differences in G + C content among their members (Fig. S1C). AbALV has the lowest G + C content in the group, almost 40% lower than the highest G + C content found in kaumobavirus Sc. Faustoviruses are divided into clear three ranges of G + C content: faustovirus D5a, E12, E23, E24, Liban, and ST1, with an average of 36.3%; faustovirus D3, D5b, D6, and VV10, with an average of 37.7%; and faustovirus E9, LCD7, M6, S17, vv57, VV63, with an average of 39.5%.

EXTENDED *ASFARVIRIDAE* PRESENTS AN EXTENSIVE OPEN PANGENOME AND A LIMITED CORE

The clustering process of protein sequences based on similarity resulted in a total of 2,483 clusters of orthologous proteins or singleton genes for extended *Asfarviridae*. Only 37 proteins are found in all the 39 genomes, while 973 proteins were considered singletons (Fig. 1A), part of the cloud group, meaning they were not found in two or more genomes.

For each lineage analyzed, the core genes corresponded to approximately 7%–8% of the total genes in faustoviruses, kaumobaviruses, and pacmanviruses. However, the scenario is different for AbALV, where core genes represent 10% of the total genes, and in ASFV, where this proportion varies from 20% to 25% of the genome.

Conversely, when examining singleton genes, we observe significant variation in proportions across different lineages. Faustoviruses and ASFV display relatively low proportions, ranging from 0 to 5% and 0 to 6%, respectively. In contrast, AbALV exhibits a notably high proportion of 83% singleton genes. Pacmanviruses and kaumobaviruses occupy an intermediate range, with proportions between 14% and 20% and 29% and 35%, respectively.

Gene-sharing patterns within this clade exhibit distinct group-specific dynamics (Fig. 1B). Intragroup variation in ASFVs and faustoviruses ranges from 41% to 49% and 32% to 50%, respectively. Kaumobaviruses, despite their genetic proximity, share only 33% of proteins, representing the highest intergroup sharing. Pacmanviruses share slightly more proteins (41%), while AbALV shows greater proximity to ASFVs (10%–11%). All other intergroup comparisons reveal less than 10% shared proteins.

The sequential inclusion of genomes does not clearly show the formation of a plateau at the end of the curve, suggesting that new inclusions can still increase the number of new genes (Fig. 1C). After applying power regression ($n = \kappa N - \alpha$) to the total number of genes in different lineage combinations, the parameter α was estimated as 0.71. Since $\alpha < 1$ and no plateau is observed, the *Asfarviridae* pangenome is considered open. It is important to note that the core gene curve stabilizes quickly. Unlike the pangenome curve, this core gene group appears conserved with little variation as new genomes are included.

When stratifying orthologous protein clusters according to their pangenomic group, that is, according to the level of protein sharing, the observed patterns are illustrated in Fig. 1D. The histogram depicts the gene frequency of orthologs in each genome, while the pie chart illustrates the proportion of total redundant proteins in each pangenomic group. The histogram indicates that most genes are present in a limited number of genomes (accessory genes), while few genes are present in all genomes (core genes), suggesting a high variation in the gene content of viruses in this clade. Additionally, there is a predominance of singleton/cloud proteins in the clusters, although shell proteins lead in terms of the total number of proteins in the pie chart.

Protein lengths within this clade vary significantly by pangenomic group (Fig. S2). Core group proteins are the largest, surpassing both other virus groups and the average protein length across the entire pangenome. This group is followed by soft core, shell, and cloud proteins. Statistical analysis (Mann-Whitney, $P < 0.05$) confirms significant differences in mean lengths between these groups. Notably, cloud group proteins are consistently the smallest.

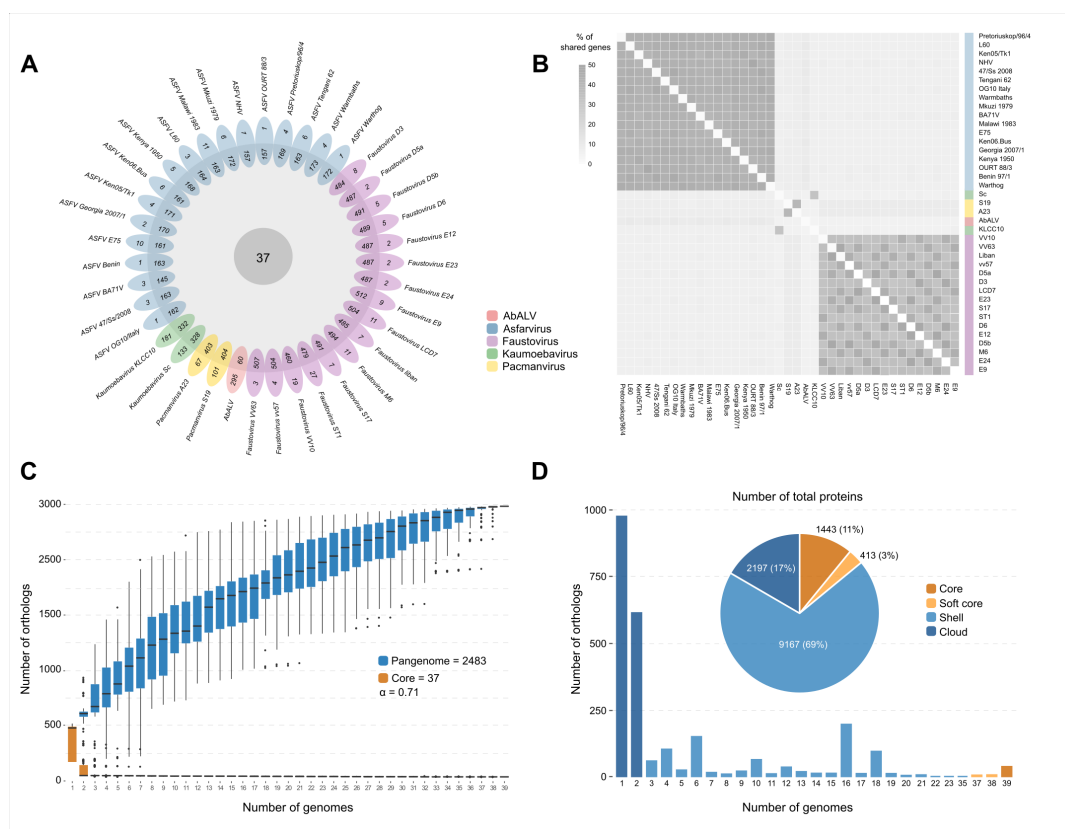


FIG 1 Pangenome structure and gene-sharing patterns in extended *Asfarviridae*. (A) Flower plot illustrating the distribution of core and singleton genes across the 39 genomes. Only 37 proteins are conserved across all genomes, while 973 are singletons, representing lineage-specific or unique genes. (B) Heatmap depicting the percentage of shared proteins between pairs of genomes, highlighting distinct gene-sharing patterns. (C) Gene accumulation curve for the total pangenome and core genes of extended *Asfarviridae*. The α value represents Heaps' law power parameter for the average number of new genes per genome after 1,000 pangenome permutations. Values below 1 indicate an "open" pangenome state. (D) Histogram of gene distributions across the pangenomic groups: core, soft core, shell, and cloud. The pie chart displays the total number of genes in each category, with percentages in parentheses. AbALV: Abalone asfar-like virus; ASFV: African swine fever virus.

CHALLENGES IN FUNCTIONAL ANNOTATION AND MCP FRAGMENTATION IN EXTENDED ASFARVIRIDAE

Functional annotation was achieved for 100% of the orthologous clusters in the core and soft-core groups, while the shell and cloud groups showed lower annotation rates of 60% and 55%, respectively, likely reflecting the abundance of singleton genes in the latter. Proteins of unknown function represent the most common category across all groups, according to the GVOG classification (Fig. S3).

As expected, most of the identified core proteins are functionally associated with DNA maintenance processes, including replication, reinforcing the essential role of these functions in the core genome of the extended *Asfarviridae*. In the soft-core group, the most frequent category is associated with transcription, replication, recombination, and repair, emphasizing the continued importance of these processes even in a less conserved set of genes. The shell group exhibited the greatest functional diversity, with more functions related to metabolism and cellular processes beyond nucleic acids. Aside from the significant proportion of unknown functions, the most represented categories in the shell group are energy production and conversion, secondary metabolite biosynthesis, transport and catabolism, and cell wall/membrane/envelope biogenesis. In the cloud group, the predominant categories are replication, recombination, and repair, and secondary metabolite biosynthesis, transport, and catabolism.

The absence of the structural protein Major Capsid Protein (MCP) in the results is noteworthy. MCP plays a crucial role in forming the viral capsid and is typically present in all viral genomes within the *Bamfordvirae* kingdom, which includes the NCLDV. However, in faustoviruses, this gene is known to be fragmented across the genome (22), potentially impacting the clustering step. To confirm MCP's presence in faustoviruses, a targeted search using BLASTp was conducted, locating 43 transcripts corresponding to MCP in faustoviruses, with an average of 2.6 transcripts per genome. These genes were then aligned with MCP sequences from other members of the group using MAFFT v7 (L-INS-i strategy), and the resulting multiple alignment was visualized in AliView (23) (Fig. S4).

PHYLOGENETIC ANALYSIS REVEALS HIGH GENETIC DIVERGENCE AMONG EXTENDED *ASFARVIRIDAE* GROUPS

The species tree (Fig. 2) was inferred using wASTRAL, based on 37 maximum likelihood core gene trees of the extended *Asfarviridae* group. This coalescent-based approach accounted for potential topological discordance among individual gene trees by integrating both branch support and branch length weighting. The resulting tree, rooted using kaumobavirus as the basal lineage, revealed a well-resolved phylogenetic structure, with strong support for major clades. Faustovirus, pacmanvirus, ASFV, AbALV, and kaumobavirus each formed strongly supported monophyletic groups, with high local posterior probabilities across internal nodes. Despite the shared ancestry among these viruses, the patristic distances separating the major clades were pronounced, typically exceeding 2.0 and reaching over 3.0 in several cases, indicating deep evolutionary divergence within the group.

SYNTENY ANALYSIS REVEALS VARIATIONS IN GENE ORDER CONSERVATION

Synteny analyses using dot plots provide a comprehensive view of genomic organization. BLASTp was used at two identity thresholds (20% and 70%) alongside MCScanX, a tool for detecting synteny. The results showed that gene order conservation varies for each virus group, even with reduced protein similarity. Intergroup comparisons revealed that only AbALV and ASFV showed gene collinearity through segmented gene blocks in the genome (Fig. 3A), especially with reduced protein similarity (p-ident).

Within groups, ASFVs demonstrated high gene order conservation, except for the group formed by ASFV BA71V, Benin 97/1, and E75, which showed no collinearity with the ASFV ken05/tk1, Kenya 1950, and ken06.Bus group. In contrast, faustoviruses showed variations in gene order conservation, ranging from a complete absence to full conservation, depending on the group members. In faustoviruses, genome sequence inversion is observed, as seen in diagonal lines from the bottom left to the top right. White boxes indicate a lack of similarity between the analyzed pairs.

ANI AND AAI VARIABILITY DELINEATES POTENTIAL NEW SPECIES AND GENERA WITHIN EXTENDED *ASFARVIRIDAE*

Analyses of AAI and ANI were performed to compare the amino acid and nucleotide compositions between different viral specimens. The AAI matrix (Fig. 3B) highlights proximity and divergence between groups. Hierarchical clustering revealed distinct patterns of similarity, with ASFVs showing high internal AAI values (>95%) and significantly lower values (~30%) when compared to other groups. AbALV is the closest to ASFVs, with an AAI of 32%. Among faustoviruses, four clusters with high AAI (>95%) were identified: (i) E12, D5a, E23, and E24; (ii) ST1 and Liban; (iii) E9, M6, S17, LCD7, vv57, and VV63; and (iv) D6, D5b, D3, and VV10. AAI values between faustovirus clusters drop to 50%–60%, and comparisons with other groups are around 30%.

Pacmanviruses and kaumobaviruses exhibited high AAI similarity within their respective groups, with kaumobaviruses at 63% AAI and pacmanviruses at 83%. Comparisons between these viruses and ASFV, faustoviruses, and AbALV yielded lower

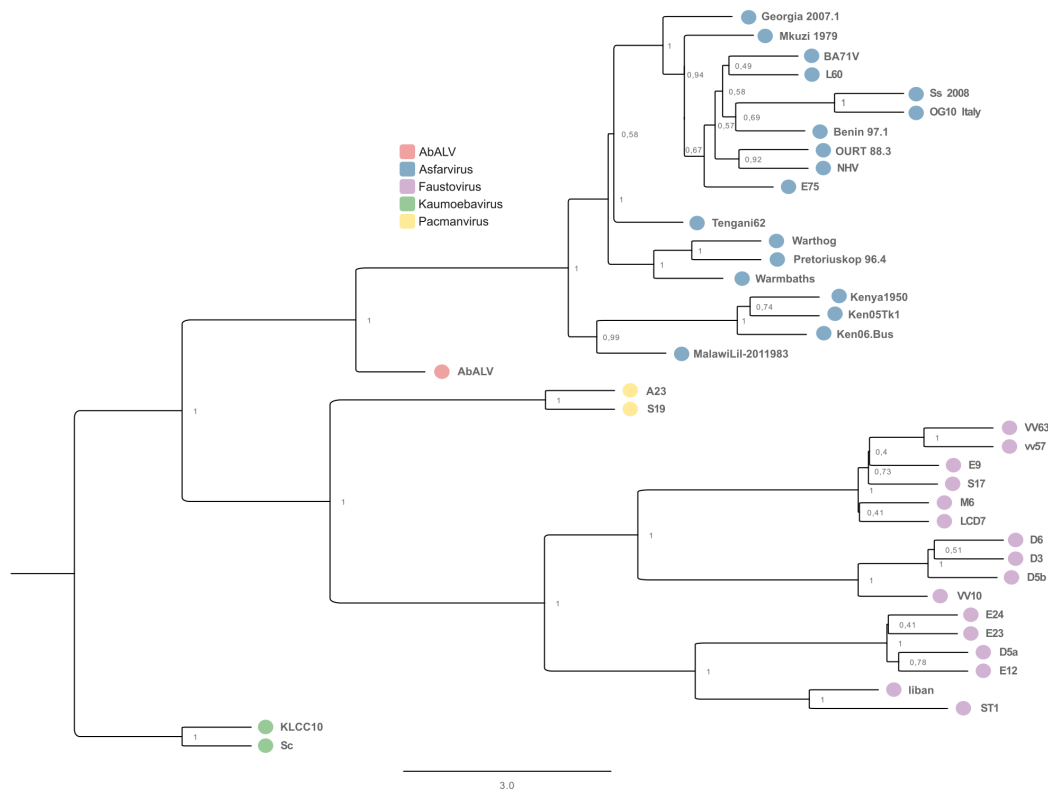
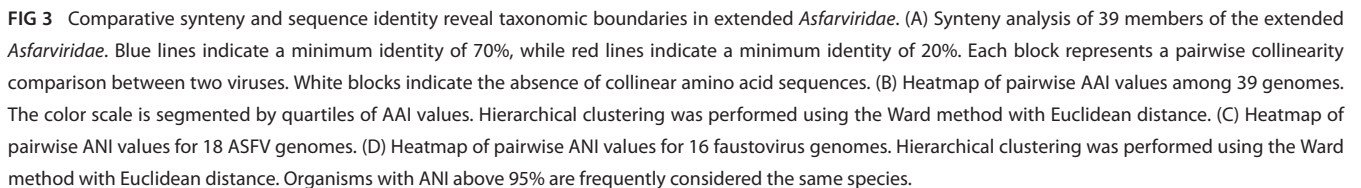


FIG 2 Phylogenetic relationships within the extended *Asfarviridae* based on core genes. Species tree reconstructed using the weighted ASTRAL (wASTRAL) method from individual maximum likelihood (ML) core gene trees, incorporating hybrid support based on bootstrap values and branch lengths. AbALV: Abalone asfar-like virus; ASFV: African swine fever virus.

AAI values, ranging from 29% to 35%. These AAI results suggest a closer relationship among faustoviruses, kaumobaviruses, and pacmanviruses, as compared to ASFV and AbALV, while highlighting notable intragroup variation. The average AAI among the 39 members of the extended *Asfarviridae* was 37%, exceeding that of *Ascoviridae* ($34\% \pm 3.49\%$), *Iridoviridae* ($34\% \pm 4.37\%$), and *Marseilleviridae* ($37\% \pm 8.07\%$), but lower than *Mimiviridae* ($44\% \pm 13.22\%$). However, the extended *Asfarviridae* exhibit greater intra-group variability.

ANI analyses were performed individually due to methodological limitations with highly divergent groups. Organisms with over 95% ANI are frequently considered to belong to the same species. ASFV subgroups identified include the following: (i) ASFV Malawi 1983; (ii) ASFV ken05/tk1, Kenya 1950, ken06.Bus; (iii) ASFV Georgia 2007/1, Mkuzi 1979, Tengani 62, Warmbaths, Pretoriuskop/96/4, and Warthog; and (iv) other ASFVs, all with ANI above 95% (Fig. 3C).

Faustoviruses demonstrate higher divergence in ANI compared to ASFVs, with four distinct groups identified in both AAI and ANI analyses, each showing intragroup ANI above 98% (Fig. 3D), indicating the likely presence of four separate species within this group. Intergroup comparisons between faustoviruses and other viral groups reveal ANI values below 78%. Kaumobavirus strains KLCC10 and Sc show a low ANI of 77.93%, while pacmanvirus strains A23 and S19 exhibit a slightly higher ANI at 85.64%, suggesting closer genomic similarity among pacmanviruses than kaumobaviruses, although both comparisons fall below the 95% ANI threshold. These findings further support the presence of two distinct species within both the pacmanvirus and kaumobavirus groups.



The term “extended *Asfarviridae*” emerged as a temporary classification to encompass a variety of giant viruses, including AbALV, faustovirus, kaumoebavirus, and pacmanvirus, as well as potential other viruses yet to be discovered that exhibit some level of phylogenetic similarity with ASFV, the original member of this clade. The inclusion of these new members results in a broad biological diversity within the extended *Asfarviridae*. These viruses infect a wide range of hosts, from unicellular amoebae to multicellular organisms such as mollusks and vertebrates. While Asfarvirus (ASFV) and AbALV occur naturally in their respective hosts, the true host range of amoeba-associated members (e.g., *faustovirus*, *kaumoebavirus*, *pacmanvirus*) remains uncertain, as the reported hosts may represent systems of isolation.

10.1128/jvi.01337-25 **8**

Marseilleviridae shows notable uniformity in this aspect. Within the extended *Asfarviridae*, faustoviruses are the closest group to ASFV, while others show greater divergence. Additionally, the division of faustovirus genomes into three distinct G + C content bands supports the presence of subgroups (3). Protein-coding patterns reflect that faustoviruses, pacmanviruses, and kaumobaviruses each display a relatively stable gene count (26), whereas ASFV carries far fewer genes, reflecting its smaller genome. Crucially, comparable protein numbers can accompany different genome lengths, highlighting variable gene density. Yet, within an established viral family, genome size itself tends to remain consistent (27, 28). In our analyses, viruses infecting unicellular eukaryotes (e.g., *Vermamoeba*, *Acanthamoeba*) cluster together and are genetically closer to each other, whereas those infecting multicellular hosts (e.g., abalones, swine) form a distinct, more compact group with smaller genomes. This pattern may reflect genome streamlining or gene loss associated with adaptation to more specialized host environments.

A previous analysis of the extended *Asfarviridae* pangenome by Karki et al. (29) primarily focused on investigating aquatic metagenomic-assembled genomes, without stratifying the pangenome into its various groups or calculating whether the pangenome was open or closed. We addressed this gap by applying the Heaps formula, proposed by Tettelin et al. (30, 31), and by constructing a pangenomic accumulation curve to evaluate this aspect. The results indicate that the extended *Asfarviridae* pangenome is considered open. This categorization implies that the discovery and incorporation of new genomes tend to add new genes to the pangenome, suggesting that the diversity of extended *Asfarviridae* is not yet fully known. Unlike many viral groups whose pangenomes are closed due to the limited number of genes, NCLDV may have genomes without this limitation, resulting in open pangenomes, as observed in this study. Examples of other viral groups with open pangenomes include pandoraviruses (32), the *Herpesviridae* family (33), and ASFVs at the species level (34).

The core genome of the extended *Asfarviridae* stabilizes rapidly at only 37 genes, a size close to the 47-gene ancestral NCLDV core (35) yet far below the 86-gene core of the more cohesive *Asfivirus* genus (34). This minimal, highly conserved set, set against a pangenome of 2,483 orthologous clusters, underscores the deep genomic divergence within the clade and suggests the core would expand if its members were more closely related. By comparison, other giant viruses retain much larger cores, such as the pandoravirus group ≈ 425 genes (15%–30% of each genome) and the families *Mimiviridae* 267 genes (22%–28%) and *Marseilleviridae* 202 genes ($\sim 44\%$) (32, 36, 37). Even the debated pithovirus clade (cedratvirus + orpheovirus) carries 52 core genes (38). Thus, extended *Asfarviridae* appears to sit near the minimal core boundary of NCLDVs, highlighting its exceptional heterogeneity. Core group proteins also tend to be longer than accessory ones, a feature linked to functional and evolutionary importance and possibly reflecting a more ancient origin (39). Natural selection is known to suppress changes in longer transcripts while promoting variation in shorter ones (40), which reinforces the stability of this minimal set. In contrast, cloud proteins are generally smaller, more prone to accumulate mutations, and likely represent younger, accessory functions, a pattern also observed in eukaryotes (41) and now evident in viruses.

A significant portion of proteins in the extended *Asfarviridae* pangenome lacks known functions, a recurring issue in NCLDV genomes, where over 80% of proteins can be uncharacterized (42). The core group has the highest proportion of identified proteins due to its conservation and essential genes, aiding functional identification. Core group genes, vital for basic biological functions, include many NCLDV housekeeping genes (35). The MCP gene, critical for all extended *Asfarviridae*, is absent from the core group due to fragmentation in faustoviruses and divergence in other viruses (22). This absence is not uncommon in core genomes; in various studies, it is also not detected in pandoraviruses and some divergent pithoviruses (42). This reflects a limitation of pangenomic techniques in detecting atypical gene structures. Other pangenomic groups showed limited functional annotation due to the high proportion of unknown proteins.

Although these viruses share a common ancestry, the species tree inferred from multi-gene coalescent analysis revealed that kaumobavirus, faustovirus, pacmanvirus, ASFV, and AbALV each form a highly supported and genetically cohesive clade. The patristic distances, that is, the summed branch lengths between taxa, separating these lineages were consistently high, reflecting levels of divergence comparable to those observed among the major clades in this data set. This degree of evolutionary separation, combined with robust monophyly and long internal branches, is inconsistent with a single-family classification.

To further investigate the taxonomic implications of these phylogenetic patterns, we integrated multiple genomic analyses, including phylogeny, AAI, ANI, and synteny. These analyses clearly support the delineation of multiple species within the extended *Asfarviridae*. The clade-wide mean AAI, a key metric for taxonomic classification, is just 37%, yet ASFV isolates share >95% AAI and minor syntenic differences, supporting three species: (i) ken05/tk1, Kenya 1950, and ken06.Bus; (ii) Malawi 1983; and (iii) the remaining 14 isolates. Faustoviruses resolve into four species based on congruent AAI/ANI values and phylogenetic topology: (i) E12, D5a, E23, E24; (ii) ST1 and Liban; (iii) E9, M6, S17, LCD7, vv57, VV63; and (iv) D6, D5b, D3, VV10. Pacmanviruses and kaumobaviruses likely each encompass two species. AbALV lies below the *Asfivirus* AAI threshold, occupying an intermediate position. Collectively, these findings reveal substantial genomic divergence within the extended *Asfarviridae*, indicating evolutionary separations that support taxonomic revision.

Altogether, these findings suggest that these aggregated viruses likely do not belong to a single taxonomic family, especially when compared to the only official member, the *Asfivirus* genus. Instead, faustoviruses, kaumobaviruses, pacmanviruses, and Abalone asfar-like virus represent a cluster of related yet genetically diverse species that may warrant classification into distinct families. Based on our analysis, the 16 faustoviruses likely comprise four species, the 18 ASFV three species, and the pairs of pacmanviruses and kaumobaviruses two species each. The open nature of the pangenome underscores the vast unexplored diversity of these viruses. Additional viral sequences and further studies are essential to deepen our understanding of this clade's genomic complexity and its relationships with hosts.

Taxonomic implications

Based on the strong phylogenetic, genomic, and pangenomic evidence presented, we propose the formal taxonomic reclassification of the so-called extended *Asfarviridae* clade. The deep divergence between ASFV, faustoviruses, kaumobaviruses, pacmanviruses, and AbALV, evidenced by consistently low intergroup AAI values (29%–35%), high patristic distances, and limited gene sharing, exceeds the thresholds typically observed between established viral families. These lineages form well-supported, monophyletic clades with intra-group AAI and ANI values above 95% (consistent with species-level cohesion), while intergroup AAI falls below 40% (supporting family-level separation).

Following established thresholds for *Nucleocytoviricota* taxonomy and the ICTV's genome-based criteria, we propose that the extended *Asfarviridae* be reclassified into five distinct families:

- *Asfarviridae* (retained; genus *Asfivirus*, 3 species),
- *Faustoviridae* fam. nov. (4 species across 4 subclades for all faustovirus species),
- *Kaumobaviridae* fam. nov. (2 genera, 2 species for kaumobavirus),
- *Pacmanviridae* fam. nov. (2 genera, 2 species for pacmanvirus),
- *Abaloneviridae* fam. nov. (genus *Abalonevirus*; pending additional isolates).

Demarcation criteria

Our proposal adheres to the quantitative and evolutionary benchmarks presented in Table 1.

This proposal is based on multiple complementary criteria:

TABLE 1 Proposed demarcation criteria for taxonomic classification within the extended *Asfarviridae*^a

Taxonomic rank	Genomic criteria	Evolutionary criteria
Species	ANI $\geq$ 95% across >75% of genes	Conserved syntenic blocks
Genus	AAI 40–70% + functional divergence (e.g., annotation profiles, gene density shifts)	Monophyly (local PP > 0.80)
Family	AAI < 40% between groups + distinct core genome size	High patristic distance (>2.0) + concordant monophyly (bootstrap > 80%)

^aQuantitative genomic thresholds and evolutionary benchmarks used to define species, genus, and family ranks based on comparative analyses.

- Monophyly with high branch support (>0.95 local posterior probabilities),
- AAI values between groups below 40%,
- ANI values between genera consistently below 70%,
- Patristic distances consistent with or exceeding those between established NCLDV families such as *Marseilleviridae*, *Iridoviridae*, and *Mimiviridae*.

Furthermore, each proposed family exhibits lineage-specific genomic features, including genome length, G + C content, pangenome structure, and functional gene content, that support their distinction at higher taxonomic ranks. These findings reinforce the need to revise the taxonomy of this clade and align with the ICTV's current standards for sequence-based virus classification.

ACKNOWLEDGMENTS

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES)—Finance Code 001. We also acknowledge support from Conselho Nacional de Desenvolvimento Científico e Tecnológico—Brasil (CNPq; grants 401485/2025-0, 303680/2022-9 and 446097/2025-0) and Fundação de Amparo à Pesquisa do Estado de Minas Gerais—Brasil (FAPEMIG; grant APD-01937-25).

AUTHOR AFFILIATIONS

¹Graduate Program in Bioinformatics, Institute of Biological Sciences (ICB), Federal University of Minas Gerais (UFMG), Belo Horizonte, Minas Gerais, Brazil

²Laboratório de Medicina e Saúde Pública de Precisão (MeSP2), Instituto Gonçalo Moniz (IGM), Fundação Oswaldo Cruz (FIOCRUZ), Salvador, Bahia, Brazil

³Department of Microbiology, Federal University of Minas Gerais (UFMG), Belo Horizonte, Minas Gerais, Brazil

⁴Department of Genetics, Luiz de Queiroz College of Agriculture (ESALQ), University of São Paulo (USP), Piracicaba, São Paulo, Brazil

AUTHOR ORCIDs

Thiago Mendonça-Santos  <http://orcid.org/0000-0001-5906-5739>

Jônatas Abrahão  <http://orcid.org/0000-0001-9420-1791>

Luiz-Eduardo Del-Bem  <http://orcid.org/0000-0001-8472-4476>

FUNDING

Funder	Grant(s)	Author(s)
Coordenação de Aperfeiçoamento de Pessoal de Nível Superior	001	Thiago Mendonça-Santos

AUTHOR CONTRIBUTIONS

Thiago Mendonça-Santos, Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft | Jônatas Abrahão, Conceptualization, Writing – review and editing | Luiz-Eduardo Del-Bem, Conceptualization, Project administration, Supervision, Writing – review and editing

ADDITIONAL FILES

The following material is available [online](#).

Supplemental Material

Supplemental material (JV101337-25-S0001.pdf). Tables S1 and S2; Fig. S1 to S4.

REFERENCES

- 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>
- Raoult D, Audic S, Robert C, Abergel C, Renesto P, Ogata H, La Scola B, Suzan M, Claverie J-M. 2004. The 1.2-megabase genome sequence of Mimivirus. *Science* 306:1344–1350. <https://doi.org/10.1126/science.11101485>
- Reteno DG, Benamar S, Khalil JB, Andreani J, Armstrong N, Klose T, Rossmann M, Colson P, Raoult D, La Scola B. 2015. Faustovirus, an asfarvirus-related new lineage of giant viruses infecting amoebae. *J Virol* 89:6585–6594. <https://doi.org/10.1128/JVI.00115-15>
- Bajrai LH, Benamar S, Azhar EI, Robert C, Levasseur A, Raoult D, La Scola B. 2016. Kaumobavirus, a new virus that clusters with Faustoviruses and Asfarviridae. *Viruses* 8:278. <https://doi.org/10.3390/v8110278>
- Andreani J, Khalil JYB, Sevvana M, Benamar S, Di Pinto F, Bitam I, Colson P, Klose T, Rossmann MG, Raoult D, La Scola B. 2017. Pacmanvirus, a new giant icosahedral virus at the crossroads between Asfarviridae and Faustoviruses. *J Virol* 91:e00212-17. <https://doi.org/10.1128/JVI.00212-17>
- Matsuyama T, Takano T, Nishiki I, Fujiwara A, Kiryu I, Inada M, Sakai T, Terashima S, Matsuura Y, Isowa K, Nakayasu C. 2020. A novel Asfarvirus-like virus identified as a potential cause of mass mortality of abalone. *Sci Rep* 10:4620. <https://doi.org/10.1038/s41598-020-61492-3>
- Besemer J, Lomsadze A, Borodovsky M. 2001. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res* 29:2607–2618. <https://doi.org/10.1093/nar/29.12.2607>
- Klemm P, Stadler PF, Lechner M. 2023. Proteinortho6: pseudo-reciprocal best alignment heuristic for graph-based detection of (co-)orthologs. *Front Bioinform* 3:1322477. <https://doi.org/10.3389/fbinf.2023.1322477>
- Snipen L, Liland KH. 2015. Micropan: an R-package for microbial pangenomics. *BMC Bioinformatics* 16:79. <https://doi.org/10.1186/s12859-015-0517-0>
- Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 30:3059–3066. <https://doi.org/10.1093/nar/gkf436>
- Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* 17:540–552. <https://doi.org/10.1093/oxfordjournals.molbev.a026334>
- Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* 32:268–274. <https://doi.org/10.1093/molbev/msu300>
- Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* 14:587–589. <https://doi.org/10.1038/nmeth.4285>
- Zhang C, Mirarab S. 2022. Weighting by gene tree uncertainty improves accuracy of quartet-based species trees. *Mol Biol Evol* 39:msac215. <https://doi.org/10.1093/molbev/msac215>
- Aylward FO, Moniruzzaman M, Ha AD, Koonin EV. 2021. A phylogenomic framework for charting the diversity and evolution of giant viruses. *PLoS Biol* 19:e3001430. <https://doi.org/10.1371/journal.pbio.3001430>
- Eddy SR. 2011. Accelerated profile HMM searches. *PLoS Comput Biol* 7:e1002195. <https://doi.org/10.1371/journal.pcbi.1002195>
- Wang Y, Tang H, DeBarry JD, Tan X, Li J, Wang X, Lee T-h., Jin H, Marler B, Guo H, Kissinger JC, Paterson AH. 2012. MScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res* 40:e49–e49. <https://doi.org/10.1093/nar/gkr1293>
- Bandi V, Gutwin C, Siri JN, Neufeld E, Sharpe A, Parkin I. 2022. Visualization tools for genomic conservation. *Methods Mol Biol Clifton NJ* 2443:285–308. https://doi.org/10.1007/978-1-0716-2067-0_16
- Rodriguez-R LM, Konstantinidis KT. 2016. The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. *PeerJ Preprints*. <https://doi.org/10.7287/peerj.preprints.1900v1>
- Kolde R. 2019. pheatmap: Pretty Heatmaps
- Sakai R, Biederstedt E. 2021. Dendsort: Modular Leaf Ordering Methods for Dendrogram Nodes
- Cherif Louazani A, Baptiste E, Levasseur A, Colson P, La Scola B. 2018. Faustovirus E12 transcriptome analysis reveals complex splicing in capsid gene. *Front Microbiol* 9:2534. <https://doi.org/10.3389/fmicb.2018.02534>
- Larsson A. 2014. AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* 30:3276–3278. <https://doi.org/10.1093/bioinformatics/btu531>
- Meier-Kolthoff JP, Klenk H-P, Göker M. 2014. Taxonomic use of DNA G+C content and DNA-DNA hybridization in the genomic age. *Int J Syst Evol Microbiol* 64:352–356. <https://doi.org/10.1099/ij.s.0.056994-0>
- Witt ASA, Carvalho J, Serafim MSM, Arias NEC, Rodrigues RAL, Abrahão JS. 2024. The GC% landscape of the Nucleocytoviricota. *Braz J Microbiol* 55:3373–3387. <https://doi.org/10.1007/s42770-024-01496-7>
- Rangel-Pineros G, Almeida A, Beracochea M, Sakharova E, Marz M, Reyes Muñoz A, Hölzer M, Finn RD. 2023. VIRify: an integrated detection, annotation and taxonomic classification pipeline using virus-specific protein profile hidden Markov models. *PLoS Comput Biol* 19:e1011422. <https://doi.org/10.1371/journal.pcbi.1011422>
- Yuan W-G, Liu G-F, Shi Y-H, Xie K-M, Jiang J-Z, Yuan L-H. 2022. A discussion of RNA virus taxonomy based on the 2020 international committee on taxonomy of viruses report. *Front Microbiol* 13:960465. <https://doi.org/10.3389/fmicb.2022.960465>
- Eaton HE, Ring BA, Brunetti CR. 2010. The genomic diversity and phylogenetic relationship in the family Iridoviridae. *Viruses* 2:1458–1475. <https://doi.org/10.3390/v2071458>
- Karki S, Moniruzzaman M, Aylward FO. 2021. Comparative genomics and environmental distribution of large dsDNA viruses in the family Asfarviridae. *Front Microbiol* 12:657471. <https://doi.org/10.3389/fmicb.2021.657471>
- Tettelin H, Riley D, Cattuto C, Medini D. 2008. Comparative genomics: the bacterial pan-genome. *Curr Opin Microbiol* 11:472–477. <https://doi.org/10.1016/j.mib.2008.09.006>
- Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Anguoli SV, Crabtree J, Jones AL, Durkin AS, et al. 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial “pan-genome”. *Proc Natl Acad Sci USA* 102:13950–13955. <https://doi.org/10.1073/pnas.0506758102>
- Aherfi S, Andreani J, Baptiste E, Oumessoum A, Dornas FP, Andrade AC dos SP, Chabriere E, Abrahao J, Levasseur A, Raoult D, La Scola B, Colson P. 2018. A large open pangenome and a small core genome for Giant

- Pandoraviruses. *Front Microbiol* 9. <https://doi.org/10.3389/fmicb.2018.01486>
33. Sharma V, Mobeen F, Prakash T. 2016. Comparative genomics of herpesviridae family to look for potential signatures of human infecting strains. *Int J Genomics* 2016:9543274. <https://doi.org/10.1155/2016/9543274>
 34. Wang L, Luo Y, Zhao Y, Gao GF, Bi Y, Qiu H-J. 2020. Comparative genomic analysis reveals an “open” pan - genome of African swine fever virus. *Transbound Emerg Dis* 67:1553–1562. <https://doi.org/10.1111/tbed.13489>
 35. Yutin N, Wolf YI, Raoult D, Koonin EV. 2009. Eukaryotic large nucleocytoplasmic DNA viruses: clusters of orthologous genes and reconstruction of viral genome evolution. *Virology* 6:223. <https://doi.org/10.1186/1743-422X-6-223>
 36. Dornas FP, Assis FL, Aherfi S, Arantes T, Abrahão JS, Colson P, La Scola B. 2016. A Brazilian *Marseillevirus* is the founding member of a lineage in family *marseilleviridae*. *Viruses* 8:76. <https://doi.org/10.3390/v8030076>
 37. Assis FL, Franco-Luiz APM, Dos Santos RN, Campos FS, Dornas FP, Borato PVM, Franco AC, Abrahão JS, Colson P, Scola BL. 2017. Genome characterization of the first mimiviruses of lineage C isolated in Brazil. *Front Microbiol* 8:2562. <https://doi.org/10.3389/fmicb.2017.02562>
 38. Queiroz VF, Carvalho JVRP, de Souza FG, Lima MT, Santos JD, Rocha KLS, de Oliveira DB, Araújo JP Jr, Ullmann LS, Rodrigues RAL, Abrahão JS. 2023. Analysis of the genomic features and evolutionary history of pithovirus-like isolates reveals two major divergent groups of viruses. *J Virol* 97:e0041123. <https://doi.org/10.1128/jvi.00411-23>
 39. Toll-Riera M, Albà MM. 2013. Emergence of novel domains in proteins. *BMC Evol Biol* 13:47. <https://doi.org/10.1186/1471-2148-13-47>
 40. Lopes I, Altav G, Raina P, de Magalhães JP. 2021. Gene size matters: an analysis of gene length in the human genome. *Front Genet* 12:559998. <https://doi.org/10.3389/fgene.2021.559998>
 41. Lipman DJ, Souvorov A, Koonin EV, Panchenko AR, Tatusova TA. 2002. The relationship of protein conservation and sequence length. *BMC Evol Biol* 2:20. <https://doi.org/10.1186/1471-2148-2-20>
 42. Rigou S, Santini S, Abergel C, Claverie J-M, Legendre M. 2022. Past and present giant viruses diversity explored through permafrost metagenomics. *Nat Commun* 13:5853. <https://doi.org/10.1038/s41467-022-33633-x>

AUTHOR BIOS

Thiago Mendonça-Santos holds a Ph.D. in Bioinformatics from the Federal University of Minas Gerais (UFMG, 2024), an M.Sc. in Biochemistry and Molecular Biology (UFBA, 2020), and a B.Sc. in Biotechnology (UFBA/Carleton University, 2017). He is currently a postdoctoral researcher at the Gonçalo Moniz Institute (FIOCRUZ Bahia), where he develops genomic surveillance strategies for respiratory viruses and arboviruses, integrating clinical and environmental monitoring within a One Health framework. His research focuses on viral genomics, evolution, and computational virology, employing bioinformatics and multi-omics approaches to investigate pathogen diversity and host–virus interactions. He is particularly interested in understanding viral evolution and the emergence of new pathogens through comparative genomics and molecular epidemiology.



Jônatas Abrahão is a virologist who has been conducting studies on the discovery, ecology, and evolution of large and giant DNA viruses for 20 years. Together with his research group, he has contributed to the description of entirely new viruses, such as Tupanvirus, Naiavirus, and Yaravirus. He is currently a professor at the Federal University of Minas Gerais, Brazil.



Luiz-Eduardo Del-Bem earned his B.Sc. in Biology (2005) and his M.Sc. (2008) and Ph.D. (2013) in Genetics and Molecular Biology from the State University of Campinas (Unicamp, Brazil). He subsequently conducted postdoctoral research in genomics, gene expression, and evolutionary biology at the Harvard T.H. Chan School of Public Health (HSPH, USA; 2014–2016). He served as a Visiting Professor at the Federal University of Bahia (UFBA, Brazil; 2017–2018) and at Michigan State University (MSU, USA; 2022–2023), and was an Assistant Professor at the Federal University of Minas Gerais (UFMG, Brazil; 2018–2025). He is currently an Assistant Professor in the Department of Genetics at the Luiz de Queiroz College of Agriculture, University of São Paulo (ESALQ/USP, Brazil). His research integrates phylogenomics, molecular biology, and bioinformatics to investigate genome evolution and function in plants and microorganisms.

