Article

Not peer-reviewed version

# Integrating Sequence- and Structure-Based Similarity Metrics for the Demarcation of Multiple Viral Taxonomic Levels

Igor Custódio Santos , Rebecca di Stephano Souza , Igor Tolstoy , Liliane Santana Oliveira , Arthur Gruber *

Posted Date: 4 April 2025

doi: 10.20944/preprints202504.0365.v1

Keywords: viral classification; viral taxonomy; taxa demarcation; sequence similarity; protein structure similarity

 doi:10.20944/preprints202504.0365.v1



*Article*

# Integrating Sequence- and Structure-Based Similarity Metrics for the Demarcation of Multiple Viral Taxonomic Levels

**Igor C. dos Santos [1], Rebecca di Stephano de Souza [2], Igor Tolstoy [3], Liliane S. Oliveira [4] and Arthur Gruber [5,6,*]**

[1] Biotechnology undergraduate course, Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, Brazil; igor223tgec@gmail.com

[2] Biological Sciences undergraduate course, Instituto de Biociências, Universidade de São Paulo, São Paulo, Brazil; rebeccadistephano@usp.br

[3] Argentys Informatics, LLC, 12 South Summit Avenue Suite 200, Gaithersburg, MD 20877, USA; itolstoy@gmail.com

[4] Department of Computer Science, Federal University of Technology of Paraná (UTFPR), Alberto Carazzai Avenue, 1640, Cornélio Procópio 86300-000, PR, Brazil

[5] Department of Parasitology, Instituto de Ciências Biomédicas, Universidade de São Paulo, São Paulo, SP, 05508-000, Brazil; argruber@usp.br

[6] Affiliated member of the European Virus Bioinformatics Center, Leutragraben 1, Jena, 07743, Germany

* Correspondence: argruber@usp.br (AG); Tel.: 55 11 3091-7274

**Abstract:** Viruses exhibit significantly greater diversity than cellular organisms, posing a complex challenge to their taxonomic classification. While primary sequences may diverge considerably, protein functional domains can maintain conserved 3D structures throughout evolution. Consequently, structural homology of viral proteins can reveal deep taxonomic relationships, overcoming limitations inherent in sequence-based methods. In this work, we introduce MPACT (Multimetric Pairwise Comparison Tool), an integrated tool that utilizes both sequence- and structure-based metrics. The program incorporates five metrics: sequence identity, similarity, maximum likelihood distance, TM-score, and 3Di character similarity. MPACT generates heatmaps and distance trees to visualize viral relationships across multiple levels, enabling users to substantiate viral taxa demarcation. Taxa delineation can be achieved by specifying appropriate score cutoffs for each metric, facilitating the definition of viral groups, and storing their corresponding sequence data. By analyzing diverse viral datasets, spanning various levels of divergence, we demonstrate MPACT's capability to reveal viral relationships, even among distantly related taxa. This tool provides a comprehensive approach to assist viral classification, exceeding current methods by integrating multiple metrics and uncovering deeper evolutionary connections.

**Keywords:** viral classification; viral taxonomy; taxa demarcation; sequence similarity; protein structure similarity

## 1. Introduction

Viral diversity far exceeds that of cellular organisms. Viruses exhibit a wide range of genomic compositions, consisting of either DNA or RNA, which may be single- or double-stranded, and can exist in either sense or antisense orientations. These genomes can be organized into single or multiple segments and vary significantly in size, encapsulated within capsids of diverse morphologies [1]. Some viruses may lack capsids entirely, persisting as extrachromosomal elements within host cells [2,3]. While eukaryotic organisms trace their lineage to a last universal common ancestor (LUCA), distinct from bacteria and archaea, viruses exhibit profound evolutionary divergence and may have

 doi:10.20944/preprints202504.0365.v1



emerged multiple times throughout evolutionary history prior to the origin of LUCA [4,5]. Unlike cellular life forms, which possess universally conserved evolutionary markers such as 16S or 18S rRNA genes, viruses lack common genetic markers, complicating phylogenetic reconstruction [6,7].

The classical classification system proposed by David Baltimore over 50 years ago [8] was based on the transmission pathways of viral genome information and initially comprised six groups, later expanded to seven. Despite significant advancements in virology, this classification remains a foundational concept for understanding information transmission pathways in biological systems [9]. Nevertheless, taxonomic assignments frequently rely on variable and inconsistent criteria across diverse viral groups, and continue to be a matter of ongoing debate [1]. In contrast to the well-structured taxonomy of cellular organisms, a formalized binomial Latin nomenclature for viruses was only proposed in 2020 [10], with implementation recommendations still in progress [11].

Traditional viral classification, established in the mid-20th century, included only five taxonomic levels: species, genus, subfamily, family, and order. This model primarily aimed to group closely related viruses. Recently, the International Committee on Taxonomy of Viruses (ICTV) introduced a hierarchical classification system with 15 taxonomic ranks to accommodate the vast genetic diversity of the virosphere, thereby enabling classifications that reflect basal evolutionary relationships among distantly related viruses [12]. The highest rank, 'realm', is analogous to the domain level in cellular life taxonomy and reflects the complex interplay between viral and cellular taxonomies. Realms do not share a universal common ancestor but group viruses with shared genetic traits. Within this framework, members of each realm share sets of ancestral orthologous genes, typically associated with replication or virion formation. This perspective acknowledges that viral classification may extend beyond taxonomy to include alternative classification schemes based on clinical or epidemiological properties.

A panel of virology taxonomy experts proposed four guiding principles for viral classification [13]. The first principle mandates that all recognized taxa must be monophyletic. The second principle suggests that phenotypic and ecological traits may be informative but should not override phylogenetic reconstruction. The third principle asserts that taxonomic classification is just one approach to categorizing viruses, and alternative classifications based on infectivity and virulence may be valuable for disease prevention and treatment, despite often forming polyphyletic groups that do not reflect evolutionary relationships. Lastly, the fourth principle emphasizes the importance of quality control, particularly in metagenomic data, for assigning any viral taxonomic classification.

With the rapid expansion of viral genome sequencing, particularly from metagenomic samples, the development of computational tools for viral sequence classification has become essential. A diverse array of software applications employing various methodological approaches is currently available [14]. One classification approach relies on genetic content conservation and genome organization in gene or protein profiles [15]. Notable platforms utilizing this approach include GRAViTy [16,17] and vConTACT v.2.0 [18]. vConTACT v.2.0 applies network-based whole-genome gene-sharing profiles to perform hierarchical clustering with integrated confidence scores, enabling automated taxonomic classification at the genus and basal taxonomic levels. GRAViTy considers gene content, orientation, and protein-coding signatures, calculating composite generalized Jaccard distances (CGJ) to group viruses into taxonomic categories, spanning from families to inter-family relationships [16,17]. Another category of tools employs phylogenetic reconstruction to establish monophyletic clades for classification. This widely accepted method produces consistent taxonomies, provided that the sequences retain phylogenetic signal. However, due to the high evolutionary rates and substantial divergence among viral families, this approach, as implemented in platforms such as VICTOR [19], is less effective at resolving higher taxonomic ranks, such as families and orders.

Pairwise sequence distance is among the metrics favored by the ICTV for establishing a universal viral classification method. A key challenge in this approach is defining optimal distance thresholds for each taxonomic level, considering the varying genetic divergence across taxa. The PASC (PAirwise Sequence Comparison) tool [20], available through the NCBI (National Center of Biotechnology Information), compares query sequences against a viral sequence database using

 doi:10.20944/preprints202504.0365.v1



BLAST or the Needleman-Wunsch algorithm, generating frequency distributions of genetic identity scores to facilitate taxonomic delineation. The DEmARC (DivErsity pArtitioning by hieRarchical Clustering) tool [21] utilizes multiple sequence alignments and evolutionary distance calculations based on probabilistic models, verifying monophyly and optimizing distance thresholds to classify viruses across all taxonomic levels within a monophyletic group. This tool is particularly suited for nucleotide sequence analysis and is optimized for studying protein sequence domains. The Sequence Demarcation Tool (SDT) performs pairwise alignments using a Needleman-Wunsch (NW) approach that disregards indel-containing positions [22]. It integrates with the Neighbor program of the PHYLIP package to construct a rooted Neighbor-joining phylogenetic tree, organizing sequences based on inferred evolutionary relationships. SDT generates heatmaps to visualize pairwise identity distributions, facilitating intuitive data interpretation. While designed for nucleotide sequence analysis, SDT can also be used with amino acid sequences, although it is then limited to calculating identity percentage, rather than amino acid sequence similarity.

For the detection and classification of novel viruses, pairwise similarity searches, such as BLAST are the most commonly used method, but distant relationships cannot be detected by any pairwise comparison method [23]. Profile-based methods like PSI-BLAST with position-specific scoring matrices (PSSMs) and profile hidden Markov models (HMMs), are more sensitive [6,7,24,25] and can detect distantly related viruses. However, with the increasing influx of metagenomic viral sequences, a vast array of highly divergent viruses is being discovered, often precluding classification restricted to sequence similarity. Protein structure is much more conserved along evolution than primary sequences, allowing to reveal relationships across evolutionary remote organisms [26]. Novel approaches leveraging protein domain structure have gained attention. Advances in artificial intelligence and protein structure prediction, exemplified by AlphaFold [27,28] and ESMFold [29], have led to the development of extensive 3D protein structure databases, such as AlphaFold DB [28] and the ESM Metagenomic Atlas. Efficient structural similarity search tools, including DALI [30] and TM-align [31], have emerged, with the recently developed Foldseek [32] significantly reducing computational requirements while maintaining high sensitivity [33].

The delineation of viral taxonomic levels remains one of the major challenges in virus taxonomy and classification. Currently, there are no simple and versatile tools that integrate multiple metrics for graphical visualization of viral taxonomic distances and classification. The development of a simple and integrated tool that incorporates multiple metrics and provides graphical output in the form of heatmaps would be highly beneficial for the rapid visualization of viral groupings. The integration of various distance metrics—such as nucleotide and amino acid similarity, maximum likelihood (based on evolutionary models), and three-dimensional structural distance—within a single platform is unprecedented and could constitute a significant contribution to the virology research community by adding quantitative and qualitative information to assist viral classification.

In this work, we aimed at developing an integrated tool for the analysis and graphical visualization of biological sequence distance data and 3D protein structures from multiple organisms to support the delineation of viral taxonomic levels. We introduce MPACT, the **M**ultimetric **PA**irwise **C**omparison **T**ool. Applications of this program across diverse viral groups, encompassing varying evolutionary distances, are demonstrated, highlighting its capacity to generate multiple graphical outputs and partition sequence data based on objective parameters.

## 2. Materials and Methods

### *2.1. Data Sources*

#### 2.1.1. RNA Viruses of Orthornavirae

We used 107 sequences of the RNA-directed RNA polymerase representing different families of the *Orthornavirae* kingdom [34], including *Totiviridae*, *Amalgaviridae* [35], *Partitiviridae* [36], *Mitoviridae*, *Botourmiaviridae*, *Narnaviridae* and *Leviviridae* (Supplementary Table S1). Two additional

 doi:10.20944/preprints202504.0365.v1



datasets comprising 66 sequences of the ORF1 protein (Supplementary Table S2) and 67 sequences of the RDRP (Supplementary Table S3) of *Amalgaviridae* viruses were also used. All protein sequences were obtained from public sources. Since *Amalgaviridae* viruses express fusion proteins [37,38], including the ORF1 protein and the RDRP, these sequences were manually separated for the respective datasets.

#### 2.1.2. Microviridae

A dataset composed of 119 sequences of the major capsid protein (VP1) from different subfamilies of the *Microviridae* family were obtained from different sources and reported in the literature. The dataset comprises sequences of *Alpavirinae* [39], *Gokushovirinae* [40,41], *Pichovirinae* [39], *Sukshmavirinae* [42], Group D [43], Parabacteroidetes prophage [44], *Aravirinae* [44], *Stokavirinae* [44], *Liberivirinae* [45], *Amoyvirinae* [46], *Bullavirinae* [47] *Pequeñovirus* [48], CGM group [49], *Tainaviridae* [50], *Occultatumvirinae* [50], *Reekeekeevirinae* [51], and *Roodoodoovirinae* [51]. The accession codes and sources of the sequences are listed in the Supplementary Table S4.

#### 2.1.3. Orthobunyavirus

We used *Orthobunyavirus* datasets containing 55 nucleotide sequences of the large (L) segment and their respective translated amino acid sequences. These datasets comprise representative isolates of 14 distinct serogroups of the genus *Orthobunyavirus* [52] (*Peribunyaviridae* family). NCBI accession codes of the sequences are listed in the Supplementary Table S5.

### *2.2. Multiple Sequence Alignment and Phylogenetic Analysis*

Multiple sequence alignments (MSAs) were generated using MAFFT v7.505 [53]. Phylogenetic reconstruction was performed using IQ-TREE v2.2 [54], with the ModelFinder [55] program used to determine the model that minimizes the Bayesian Information Criterion (BIC) score. Node support values were determined using 1000 pseudoreplicates with the ultrafast bootstrap approximation (UFBoot) method [56].

### *2.3. Three-Dimensional Protein Structure Prediction*

The protein sequences of the datasets were used to predict their respective 3D structures. The AlphaFold2 platform [27,28] was employed, specifically using the ColabFold v1.5.5 server: AlphaFold2 [57]. From the five 3D structures generated by AlphaFold, the top-ranked structure based on the predicted local distance difference test (pLDDT) value [58] was selected for downstream analyses. PDB structure files were converted to 3Di-character files using Foldseek [32].

### *2.4. Implementation of MPACT Program*

MPACT, the **M**ultimetric **PA**irwise **C**omparison **T**ool is an integrated toolbox for pairwise all-against-all comparison of primary nucleotide or amino acid sequences, and 3D protein structures. For nucleotide sequences, the program utilizes identity percentage and maximum likelihood (ML) distance as metrics. In the case of proteins, MPACT determines identity percentage, similarity percentage, and maximum likelihood distance of amino acid sequences, and also obtains structural similarity (TM-scores) and 3Di character similarity of 3D-structures. For each metric, the program performs data clustering and generates heatmaps, frequency distribution plots, and dendrograms. Also, MPACT can partition sequence datasets according to user-defined criteria for each metric. MPACT is written in Python 3 (https://www.python.org/) and utilizes the following third-party programs: Needle (from the EMBOSS package version 6.6.0 [59]), MAFFT [53], IQ-TREE 2 [54], Foldseek [32] and TM-align [31]. The program executable, usage manual, tutorial, and a Docker container image for easy execution are available on GitHub (https://github.com/gruberlab/mpact).

## 3. Results

 doi:10.20944/preprints202504.0365.v1



### *3.1. Workflow of the Program*

The MPACT program's workflow is depicted in Figure 1. The program utilizes two types of input data: protein or nucleotide sequences, in FASTA format, and 3D protein structure files in PDB format. Biological sequences are submitted to all-against-all pairwise alignments using the Needleman-Wunsch global alignment algorithm implemented in the Needle program (EMBOSS package). MPACT extracts global identity and similarity percentage values from these alignments and stores the results as matrices. The input sequences are also submitted to a multiple sequence alignment (MSA) with MAFFT. The resulting MSA is then used for ML phylogenetic analysis using IQ-TREE 2, and the resulting ML phylogenetic tree and distance matrix are stored. Three-dimensional protein structures are submitted to all-against-all pairwise structural comparisons using TM-align. All resulting pairwise scores, normalized by the average length of the compared structures, are used to create a TM-score matrix. PDB files are also converted to 3Di-character sequences by the Foldseek program, and these sequences are aligned using Needle with a 3Di substitution matrix [32]. Data clustering is performed on all matrices using the clustermap function from the Seaborn data visualization library. The resulting clustering trees (Newick format) and heatmap images (JPG and SVG) are stored. Additionally, MPACT converts the matrix data into range-tabulated values and creates frequency distribution plots with Seaborn's lineplot function. The final section of the workflow is devoted to data partitioning. In a first step, all matrices are converted to distance matrices. The data is then submitted to Neighbor-joining (NJ) hierarchical clustering using the Biopython's Bio.Phylo package. The generated NJ tree is rooted at the midpoint, and its nodes are ordered in increasing order of branch length. This order is then used for data partitioning to generate groups of sequences selected within a range of user-defined upper and lower values for each chosen metric. Finally, MPACT stores the resulting tree, heatmap, and sequence files.

### *3.2. Using MPACT on Viral Protein Sequence and 3D-Structure Datasets*

#### 3.2.1. Application on RNA Viral Families of the Orthornavirae Kingdom

To assess the ability of the MPACT program to unravel relationships across different viral families, we chose first to analyze a group of diverse RNA viruses of the Orthornavirae kingdom, comprising eukaryotic dsRNA viruses of the families Totiviridae and Amalgaviridae, presenting monopartite genomes, Partititiviridae, containing bipartite genomes, ssRNA positive-strand [(+)ssRNA] eukaryotic viruses of the families Mitoviridae, Narnaviridae and Botourmiaviridae, and two representatives of Leviridae, (+)ssRNA viruses that infect prokaryotes.

 doi:10.20944/preprints202504.0365.v1



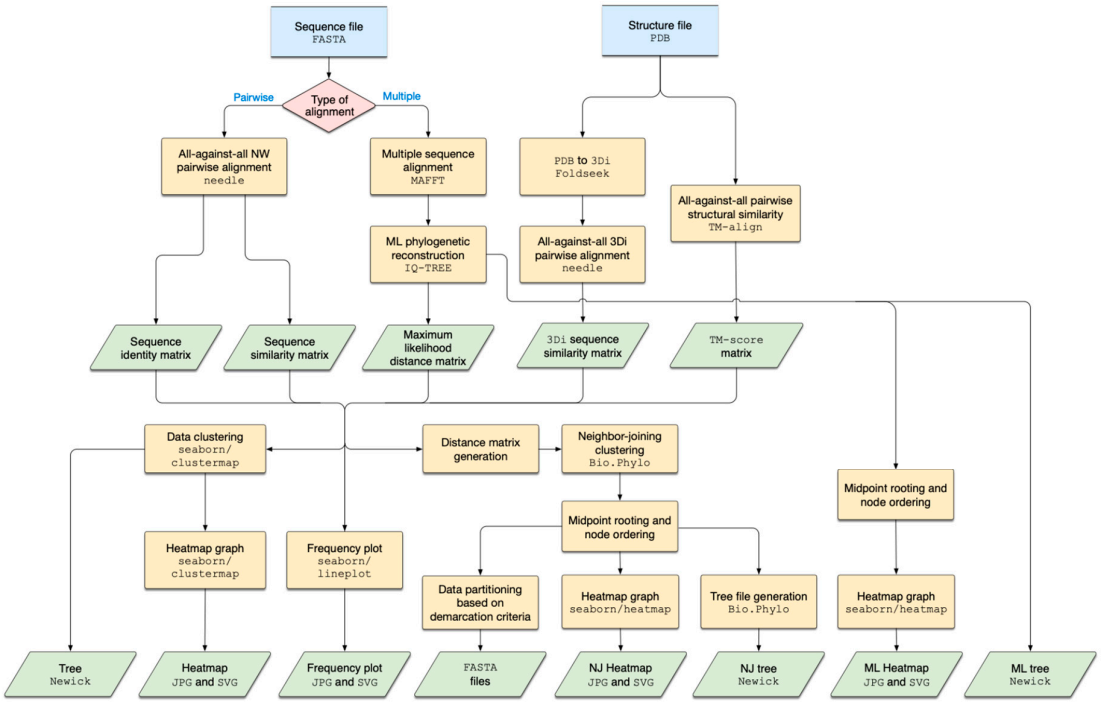



**Figure 1.** Workflow of the MPACT program. The program utilizes two types of input data: biological sequences in FASTA format and 3D protein structures in PDB format. Sequences are submitted to all-against-all pairwise alignments using the needle program (EMBOSS package), and to a multiple sequence alignment (MSA) using the MAFFT program. The MSA is submitted to IQ-TREE program for phylogenetic analysis, generating a maximum likelihood (ML) tree and a heatmap. Three-dimensional protein structures are subjected to all-against-all structural alignments using the TM-align program, and are also converted to 3Di characters by Foldseek, and aligned with needle. Alignment results are clustered and used to generate heatmap and frequency distribution plots. Also, distance matrices are produced and the data clustered by the Neighbor-joining algorithm. The resulting tree is rooted at the midpoint and sorted in ascending order of the nodes. This order is used for data partitioning to generate groups of sequences selected within a range of user-defined values.

To guide our analyses as a golden standard, we performed a phylogenetic reconstruction (Figure 2) of these viral families using RDRP sequences. All families showed monophyly, with the dsRNAviruses presenting a major clade composed of the diverse group of monopartite Amalgaviridae and a sister clade of the bipartite Partitiviridae. A more basal clade comprises monopartite viruses of the Totitiviridae family. The (+)ssRNA viruses constitute a major sister clade containing a large subclade with three families of eukaryotic viruses (Mitoviridae, Narnaviridae and Botourmiaviridae), and a more external clade comprising the prokaryotic Leviviridae phages. These evolutionary relations are in agreement with literature reports [3,35,60,61]. This protein dataset was processed by the MPACT program using the five metrics. Figure 3 shows a typical output generated by MPACT, displaying a heatmap graph of ML distance values, with an upper UPGMA dendrogram. Most of the relationships observed in the phylogenetic analysis (Figure 2) are congruent with the clusters obtained for the ML distance heatmap (Figure 3). MPACT also generates frequency distribution plots (Figure S1) allowing for the comparison of the five metrics, an output that can be used to support data partitioning criteria.

To investigate how different metrics reflect the diversity of viral taxa and protein markers, we restricted our analysis to two proteins from viruses of the *Amalgaviridae* family. This family includes the genus *Amalgavirus*, a group of important plant pathogens, and several Amalga-like viruses found in diverse other hosts, primarily fungi and some insects. The monopartite genome of this family, similar to that of the related sister family *Totiviridae*, comprises two open reading frames (ORFs). A

 doi:10.20944/preprints202504.0365.v1



key difference between *Amalgaviridae* and *Totiviridae* families lies on the first ORF, which encodes a capsid protein in *Totiviridae* but a protein of unknown function in *Amalgaviridae*. The second ORF codes for the RDRP in both families.

The results obtained using MPACT clearly shows that amino acid identity percentage is by far the least conserved metric across amalgaviruses for both proteins, displaying heatmaps with the coolest colors (Figure 4). A higher level of conservation is observed for amino acid similarity (Figure 4B), reflecting evolutionary constraints that conserve residues sharing physicochemical properties. Such residues may play equivalent roles in a functional domain or contribute to protein structure stabilization. Much higher conservation is observed for ML distance (Figure 4C) in both ORF1 protein and RDRP. To understand these results, it is important to consider how the ML distance matrix is calculated. MPACT uses MAFFT to generate an MSA of the viral sequences and then executes IQ-TREE 2 for phylogenetic reconstruction. IQ-TREE 2 employs an ML model to estimate evolutionary relationships, determining the tree topology and branch lengths that maximize the likelihood of the observed data given the model. These branch lengths are then used to calculate pairwise evolutionary distances, which comprise the ML distance matrix. Therefore, ML distance closely reflects phylogenetic analysis. This represents a substantial improvement over simple identity and similarity percentages, which are calculated primarily from pairwise alignments and lack a basis in evolutionary model

 doi:10.20944/preprints202504.0365.v1



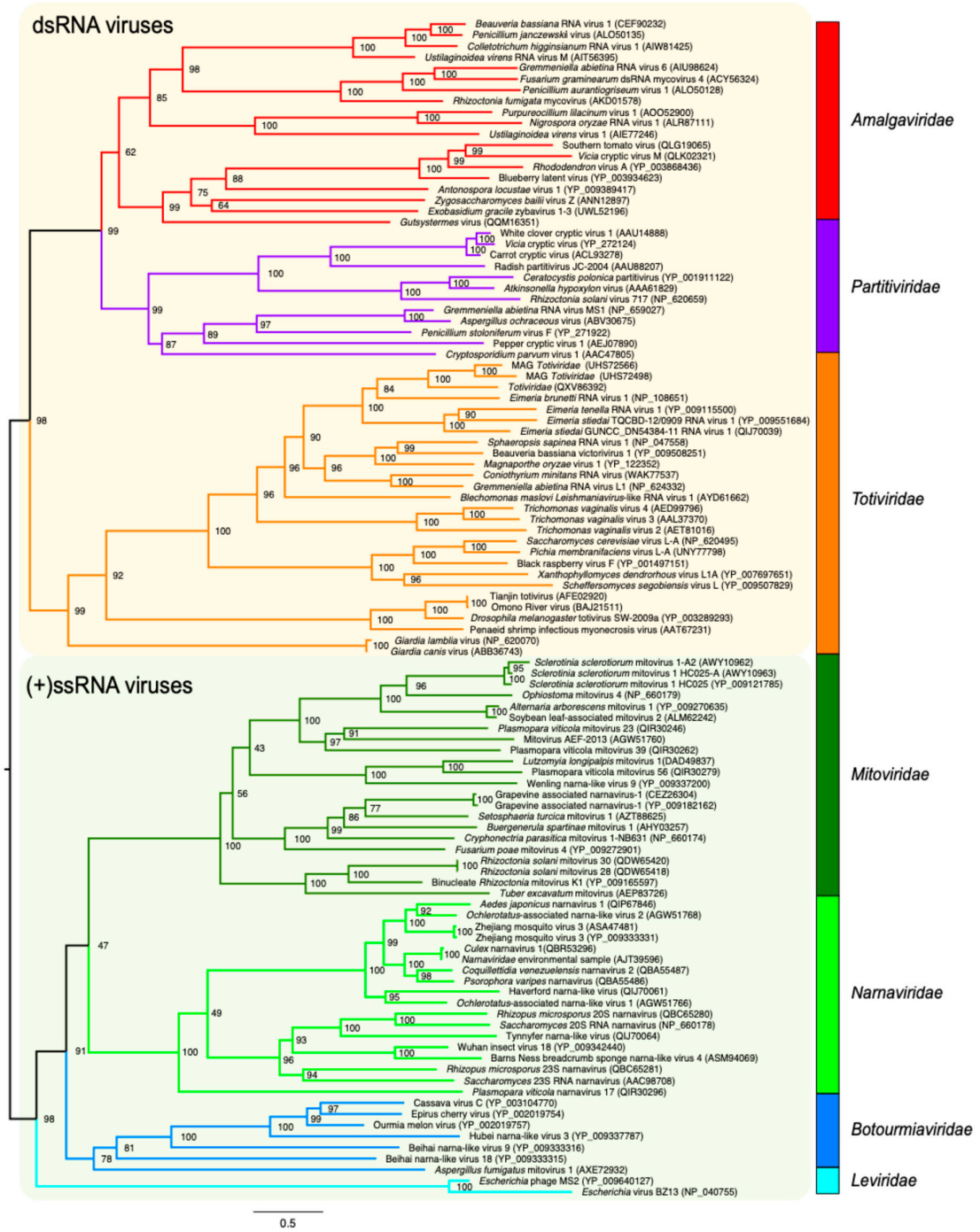



**Figure 2.** Phylogenetic reconstruction of RDRP from representative taxa of different phyla of the *Orthornavirae* kingdom. The maximum likelihood tree was inferred using IQ-TREE 2 with the best-fit model VT+F+R7 on a multiple sequence alignment generated by MAFFT. The tree is rooted at the midpoint and the nodes are sorted in increasing order. Support values are shown at the nodes of the clades. Viral genome composition of the clades is depicted by the colored background. The colored vertical bars indicate the viral families of the respective clades.

 doi:10.20944/preprints202504.0365.v1



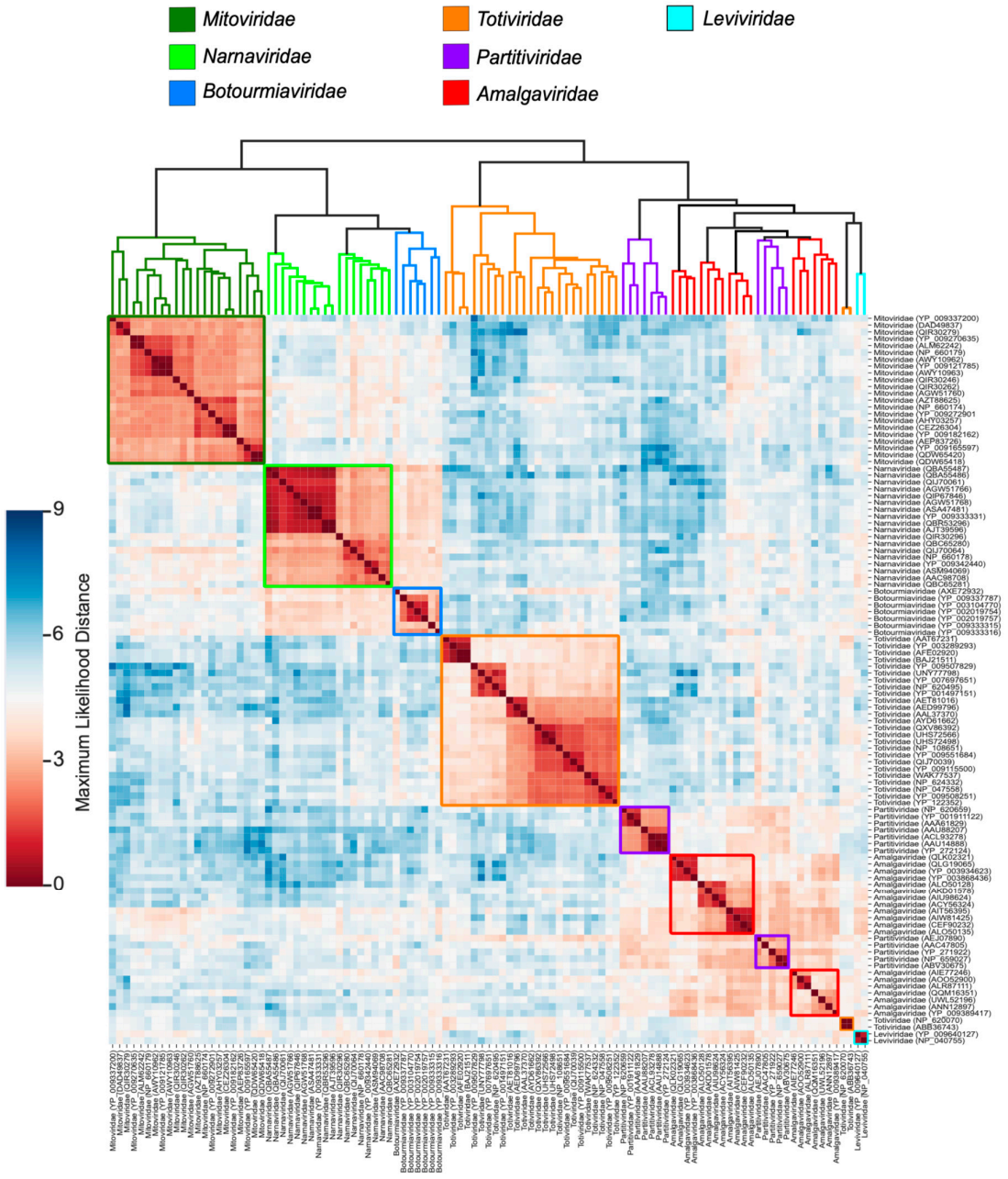



**Figure 3. Heatmap of all-against-all pairwise comparisons of RDRPs from different phyla of the *Orthornavirae* kingdom. A** multiple sequence alignment was performed using MAFFT and an ML distance matrix was obtained with IQ-TREE 2. The resulting ML distance values were clustered by the UPGMA method, and the upper dendrogram represents the clustered taxa. The data is displayed as a heatmap, with the upper right color scale representing the range of ML distance values. The colors of the clades in the tree and the colored squares on the heatmap represent different viral families.

The structural data comparisons across *Amalgaviridae* viruses show discrepant results between the ORF1 protein and RDRP for the TM-score metric (Figure 4D). The ORF1 protein presents an overall low structural similarity characterized by a heatmap with cool colors, whereas RDRP shows much higher TM-scores depicted by warmer colors. This result suggests that ORF1 protein's function does not depend on a very rigid and stable structure. On the other hand, RDRP demonstrates high structural conservation, which is compatible with its highly specialized enzymatic function, the replication of the genetic material of RNA viruses. For both proteins, 3Di-character sequence

 doi:10.20944/preprints202504.0365.v1

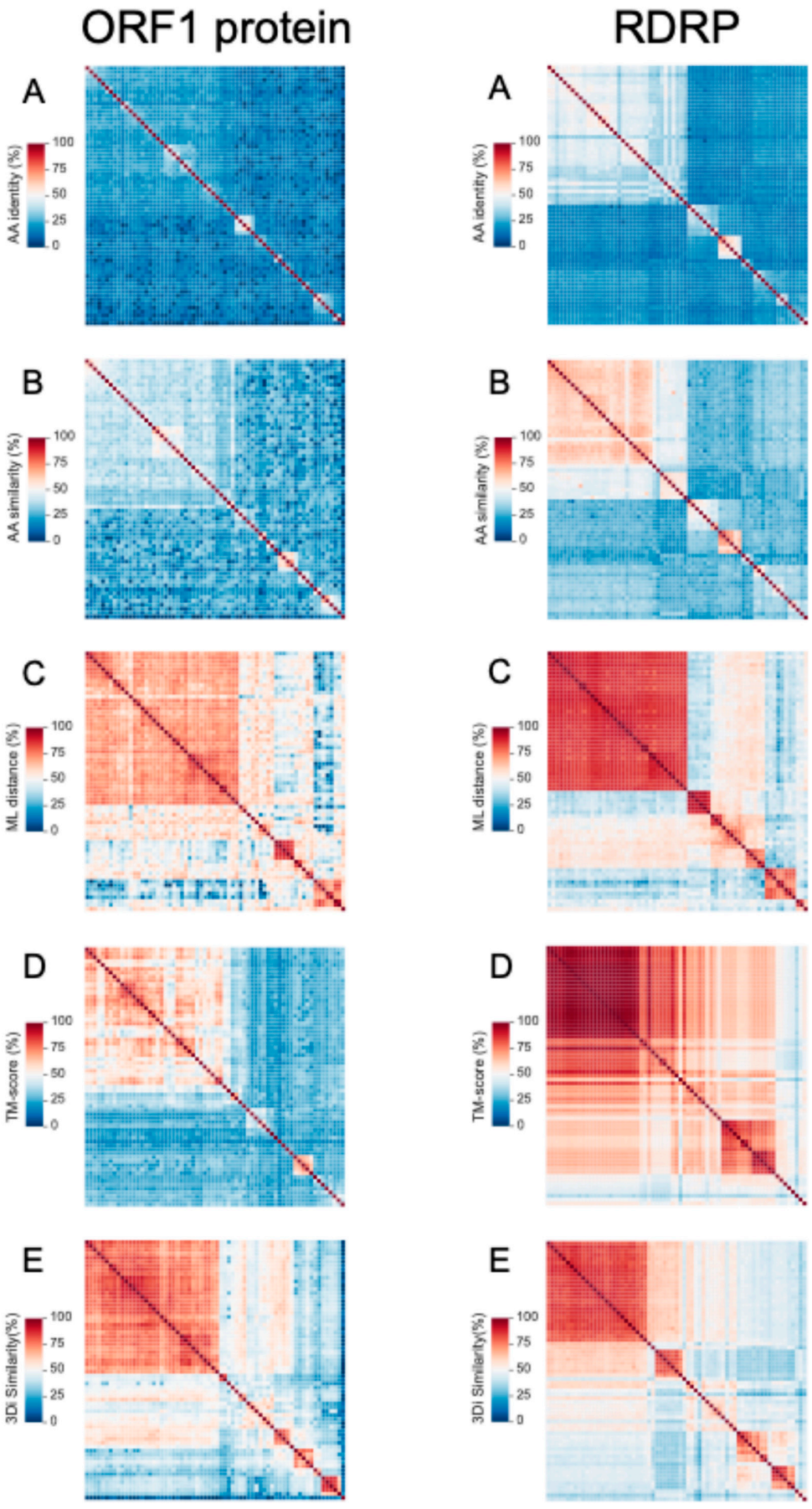


similarity heatmaps (Figure 4E) are relatively similar to the ML distance (Figure 4C) heatmaps. An intriguing observation arises from the unexpected discrepancy between TM-score (Figure 4D) and 3Di similarity heatmaps (Figure 4E) for the ORF1 protein. Given that 3Di characters represent a discretization of protein structures, one would anticipate higher all-against-all conservation in 3Di-character similarity heatmaps compared to amino acid similarity, yet lower than that observed for TM-scores. In fact, the TM-score heatmap of RDRP (Figure 4D) exhibits higher conservation than the corresponding 3Di-character similarity (Figure 4E), which in turn is higher than amino acid similarity (Figure 4B). Conversely, for the ORF1 protein, 3Di-character similarity (Figure 4E) displays significantly greater cross-conservation than both, amino acid similarity (Figure 4B) and TM-score (Figure 4D). This discrepancy can be attributed to the distinct alignment strategies employed in the analyses by TM-align and 3Di character similarity. While TM-align focuses on globally aligning relatively rigid structures, 3Di character analysis emphasizes the geometric relationships between adjacent residues [32]. Consequently, 3Di character alignment tends to prioritize shorter structural segments and the identification of local interactions between spatially proximate amino acids, potentially overlooking the overall 3D structural alignment.

To better understand this result, we performed 3D protein structure predictions of the ORF1 proteins using AlphaFold2 and the results confirmed a high abundance of alpha-helical regions (Figure S2). The pLDDT (predicted local distance difference test) values for the five top-ranked 3D structures ranged from 51.0 to 78.2, while the pTM (predicted template modelling) scores ranged from 0.342 to 0.560. For comparison, the structure predictions of the RDRPs from the same organisms yielded pLDDT values ranging from 59.0 to 86.5 and pTM values ranging from 0.615 to 0.865 (data not shown). These results corroborate that the ORF1 protein presents a more relaxed/flexible 3D structure than the RDRP. The precise function of the ORF1 protein remains unknown and, despite numerous efforts, viral particles have not been observed in amalgaviruses, unlike totiviruses, which exhibit icosahedral capsids [62]. The ORF1 protein shows no similarity to capsid proteins and some evidence point out a possible interaction with the viral RNA genome and a potential protective role [35]. Also, immunogold electron microscopy revealed that amorphous bodies in the cytoplasm of blueberry cells and these aggregates could be involved in viral genome protection [62]. In conclusion, the ORF1 protein may lack a stable, well-defined 3D structure and could exhibit high flexibility and/or a tendency to aggregate. This characteristic, coupled with the abundance of alpha-helical domains, could explain the discrepant results obtained by MPACT using TM-align structural similarity (Figure 4D) and 3Di-character similarity (Figure 4E).

 doi:10.20944/preprints202504.0365.v1





 doi:10.20944/preprints202504.0365.v1



**Figure 4.** Heatmaps of all-against-all pairwise comparisons of the ORF1 protein and RDRP of *Amalgaviridae* viruses derived from different metrics: identity (A) and similarity (B) percentages and maximum likelihood distance (C) of amino acid sequences, TM-scores of 3D structures (D), and 3Di-character sequence similarity (E).

### 3.2.2. Application on Bacteriophages of the Microviridae Family

We extended the validation of the MPACT program to *Microviridae*, a family of phages that infect a wide range of bacterial hosts across various environments. This viral family comprises a growing number of groups, whose taxonomic classification remains incomplete [63]. To establish a golden standard overview of the family's evolutionary relationships, we utilized VP1 (major capsid protein) sequences from representatives of the numerous subfamilies/groups reported in the literature and accessible through public repositories. The ML phylogenetic reconstruction (Figure 5) demonstrates that these groups are monophyletic, with some exhibiting relatively long branches, indicative of high divergence and evolutionary rate. It is noteworthy that phages classified as members of the *Pichovirinae* [39] subfamily and the Parabacteroidetes group [44] are closely related. Likewise, phage belonging to the so-called CGM group [49] are part of the proposed *Occultatumvirinae* subfamily, which constitutes a sister clade to the *Tainavirinae* subfamily, both of them hosted by Alphaproteobacteria [50]. Our findings are in agreement with the original description of the *Reekeekeevirinae* subfamily [51], which constitutes a basal clade to *Liberivirinae* [45] and *Amoyvirinae* [46], and with the phylogenetic location of the *Roodoodoovirinae* subfamily [51], forming a more basal clade to *Bullavirinae* and *Pequeñovirus*. Insertion loops from three VP1 subunits constitute mushroom-like protrusions that are observed in the viral capsids of some *Microviridae* phages [64]. Our 3D protein structure predictions (Supplementary Figure S3) confirm the existence of insertion loops in the capsid protein (VP1), as originally reported for *Alpavirinae*, *Gokushovirinae* and *Pichovirinae* [39], *Stokavirinae* and *Aravirinae* [44], and *Reekeekeevirinae* and *Roodoodoovirinae* [51]. We also observed these loops in members of *Sukshmavirinae*, Group D, Parabacteroidetes prophages, *Occultatumvirinae*, and *Tainavirinae*, and this is the first report of protrusion loops in members of these *Microviridae* subfamilies/groups. Our phylogenetic analysis (Figure 5) reveals an interesting correlation between the main clades and the presence or absence of insertion loops that form mushroom-like protrusions. Beginning from the most basal clades, which correspond in ascending order to *Reekeekeevirinae*, *Amoyvirinae*, and *Liberivirinae*, no mushroom-like structures are observed. Conversely, the sister clade to these subfamilies, comprising *Occultatumvirinae*/CGM group and *Tainavirinae*, exhibits the insertion loop on the VP1 structure. Continuing in ascending order, we observe a clade encompassing *Roodoodoovirinae*, *Pequeñovirus*, and *Bullavirinae*, which lack the mushroom-like structures. Finally, all remaining subfamilies/groups within the upper sister clade possess insertion loops on VP1. These results suggest that mushroom-like structures may have originated independently on multiple occasions throughout the evolutionary history of *Microviridae* phages, but the functional role of these structures remains unknown.

Heatmaps generated by MPACT using five distinct metrics reveal that ML distance (Figure 6C) more effectively highlights the similarities between members of each group compared to amino acid similarity (Figure 6B) or identity (Figure 6A) percentages, consistent with the results observed for the ORF1 protein and RDRP of viruses within the *Orthornavirae* kingdom (Figure 4). As previously mentioned, ML distance closely reflects the phylogenetic analysis and evolutionary history of the viral groups. The TM-align scores (Figure 6D) and 3Di-character similarity (Figure 6E) clearly display the overall strongest relationships across the different taxa, even for those that are more distantly related. This is expected, since both metrics are based on the conservation of 3D structures, which is more conserved than primary sequence identity or similarity percentages [26], and that VP1 presents structural constraints due to its role in capsid formation.

 doi:10.20944/preprints202504.0365.v1



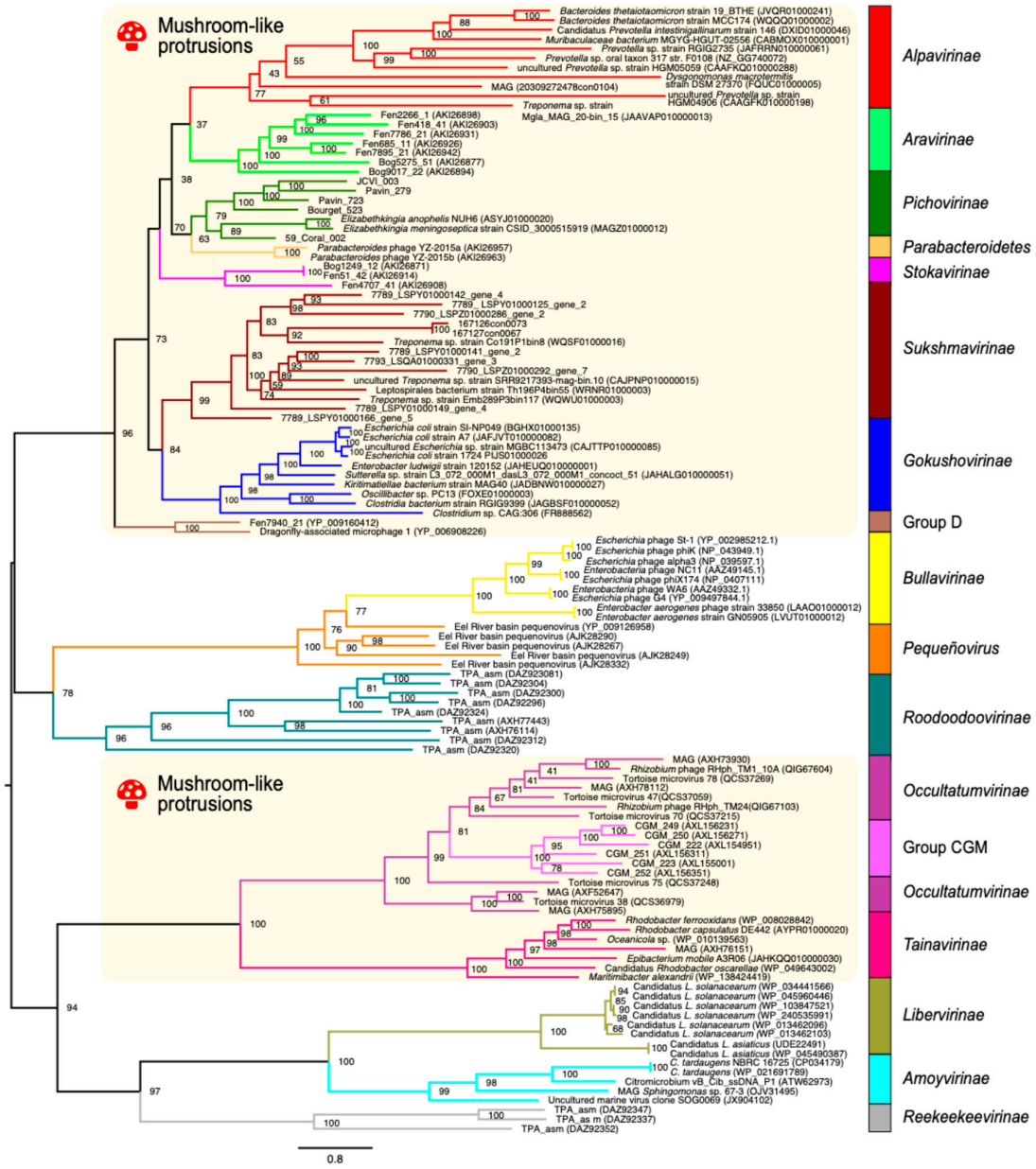



**Figure 5.** Phylogenetic reconstruction. Maximum likelihood tree inferred from amino acid sequences of the major capsid protein (VP1) derived from representative taxa of *Microviridae* viruses. The phylogenetic tree was obtained using IQ-TREE 2 with the best-fit model Q.pfam+F+R7 on a multiple sequence alignment generated by MAFFT. The tree is rooted at the midpoint and the nodes are sorted in increasing order. Support values are shown at the nodes of the clades. The colored background indicates clades whose viral members present mushroom-like protrusions as inferred from 3D structure predictions of the capsid proteins (VP1). The colored vertical bars indicate the viral subfamilies/groups of the respective clades.

 doi:10.20944/preprints202504.0365.v1



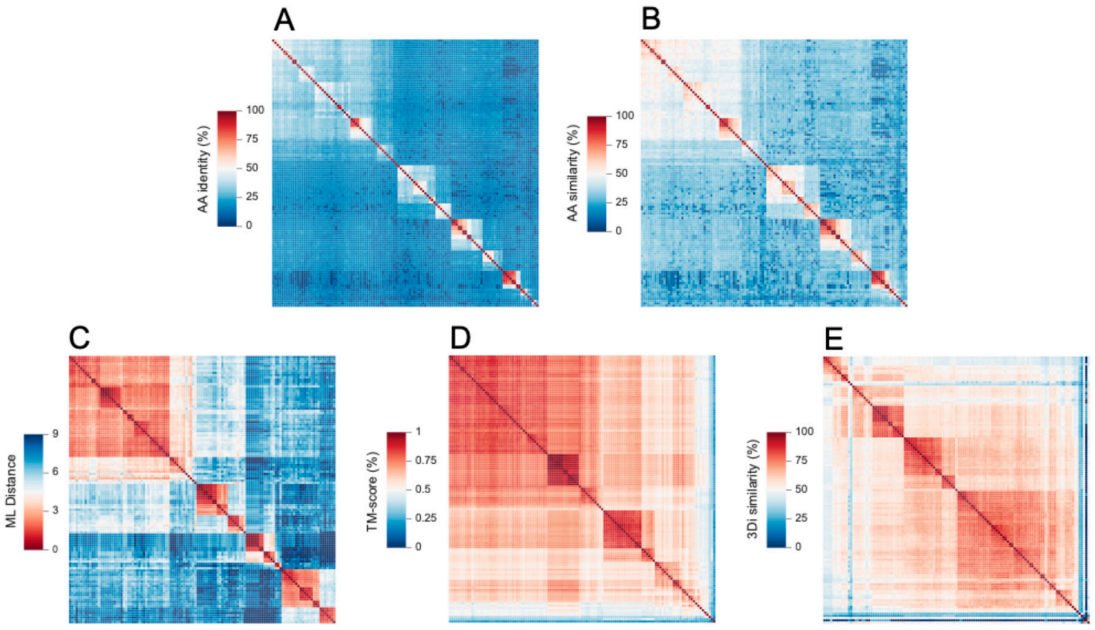



**Figure 6.** Heatmaps of all-against-all pairwise comparisons of VP1 sequences *Microviridae* viruses derived from different metrics: identity (A) and similarity (B) percentages and maximum likelihood distance (C) of amino acid sequences, TM-scores of 3D structures (D), and 3Di-character sequence similarity (E).

### *3.3. Viral Group Demarcation*

One of the primary applications of comparing viral nucleotide or protein sequences across different taxa is to establish criteria for taxonomic delineation. An ideal demarcation criterion should generate reliable viral groups with few or no misplaced taxa within the respective clusters. Phylogenetic analysis, grounded on evolutionary models, provides a robust framework for evaluating and guiding manual assignments. Since MPACT uses IQ-TREE 2 to perform phylogenetic reconstruction, we obtained ML phylogenetic trees for our datasets of *Orthornavirae* RNA viruses (Figure 2) and *Microviridae* phages (Figure 5). Both viral protein datasets yielded well-resolved trees, characterized by monophyletic groups and strong node support values.

As MPACT executes several analyses using various programs on both sequence and 3D-structure data, it enables the selection of the most informative and discriminative metrics to assist taxa demarcation. Starting from the distance matrices obtained for the various metrics, it is necessary to define value range limits for the clustering task. For instance, given a similarity percentage matrix, we need to delineate upper and lower percentage values as inclusion and exclusion criteria for the groups. Since the optimal values are not available a priori, we performed inter- and intragroup comparisons. Members of the viral taxa were defined by the maximum likelihood (ML) phylogenetic reconstructions and data reported in the literature, resulting in *bona fide* subsets. For each subset, sequences were compared in all-against-all fashion using the different metrics, to determine the intragroup value range for each metric. Sequences from each subset were also compared to the remaining sequences of the whole dataset to determine the intergroup value range. This approach was employed for the groups of RNA viruses (Supplementary Table S6) and *Microviridae* phages (Supplementary Table S7).

Across the different families of *Orthornavirae* viruses (Supplementary Table S6), all metrics exhibited a wide range of intragroup values. For example, within the *Amalgaviridae* family, amino acid sequence identity and similarity varied substantially, ranging from 10.4% to 71.7% and 17.0% to 81.9%, respectively. Similar variability was observed for ML distance, TM-score, and 3Di-character sequence similarity. To investigate whether a broad spectrum of intragroup diversity is a common feature across different viral groups, we extended the study to estimate the range of intragroup

 doi:10.20944/preprints202504.0365.v1



diversity values within *Microviridae* (Supplementary Table S7). Like RNA viruses, the different subfamilies/groups of *Microviridae* also revealed a wide range of intragroup variability. For instance, *Gokushovirinae* subfamily exhibited amino acid sequence identity and similarity values ranging from 32.6% to 88.8% and 48.9 to 93.1%, respectively. A comparison of the overall values observed for the five metrics in *Orthornavirae* (Supplementary Table S6) and *Microviridae* (Supplementary Table S7) revealed considerably higher intragroup divergence in the former dataset. The *Orthornavirae* dataset includes dsRNA viruses representatives from the phyla *Duplornaviricota* and *Pisuviricota*, as well as (+)ssRNA viruses belonging to the phylum *Lenarviricota*. In contrast, the *Microviridae* dataset is limited to (+)ssDNA viruses from different subfamilies/group within a single family. This substantial difference in the taxonomic breadth represented by each dataset may explain the greater divergence observed within the *Orthornavirae* dataset compared to the *Microviridae* dataset. In fact, the mean pairwise similarity percentage of *Orthornavirae* viruses is 22.1% ± 10.1 (Supplementary Table S6), whereas this value for *Microviridae* phages is 31.8% ± 11.9 (Supplementary Table S7).

Although *Microviridae* subfamilies show lower intragroup divergence than the different families of the *Orthornavirae* dataset, a considerable range of intragroup diversity is still observed. This feature is likely a consequence of the very high mutation rates of viruses [65–69], leading to high divergence rates even within relatively narrow taxonomic groups. The wide range of intragroup diversity observed across all evaluated metrics in both the *Orthornavirae* (Supplementary Figure S4) and *Microviridae* (Supplementary Figure S5) hinders the clear demarcation of viral taxa. For every metric, regardless of the viral group considered, the range of intragroup values overlaps with the corresponding intergroup value range. For instance, *Alpavirinae* shows intragroup amino acid identity and similarity values ranging from 14.1% to 65.4% and 23.4% to 76.6%, respectively (Supplementary Table S7). The corresponding intergroup ranges are 2.1% to 29.3% and 2.9% to 43.4%, respectively. This indicates that the *Alpavirinae* subfamily includes members whose sequences are more similar to viruses belonging to other subfamilies than they are to members within their own subfamily. This pattern is also observed when analyzing the corresponding value ranges of ML distance, TM scores, and 3Di-character sequence similarity (Supplementary Figure S5). Similarly, overlaps between intragroup and intergroup value ranges are seen for all families of *Orthornavirae* (Supplementary Table S6 and Supplementary Figure S4). These results indicate that high intragroup diversity prevents any metric from discriminating between the various viral groups within the analyzed datasets.

Based on these results, we decided to analyze a taxonomically narrower dataset, focusing on representative viruses from different antigenic groups/serogroups within the genus *Orthobunyavirus* (Supplementary Table S5). As anticipated, the mean pairwise amino acid similarity percentage within *Orthobunyavirus* was 71.85% ± 6.9 (Supplementary Table S8), significantly higher than the values observed for *Orthornavirae* (22.1% ± 10.1 - Supplementary Table S6) and *Microviridae* (31.8% ± 11.9 - Supplementary Table S7). Similarly to the approach adopted for *Orthornavirae* and *Microviridae*, we used a dataset of bona fide representatives from 14 of the 18 antigenic groups/serogroups reported in the literature [70–72] for phylogenetic reconstruction. Since the members of the genus *Orthobunyavirus* are closely related, we used nucleotide sequences for this analysis. The resulting phylogenetic tree (Figure 7) showed monophyly across the different groups and high node support values for most clades. It is noteworthy that some clades display short branch lengths, as observed for California, Gamboa, and Bunyamwera, while other groups show very long branches, such as Nyando, Anopheles A, Simbu, and Maputta. This result can be attributed to a variety of factors, acting either independently or in concert: (1) significantly different evolutionary rates among serogroups, (2) low taxa sampling, and (3) variations in host diversity and population dispersal. Given the relatively close relationship among these viruses, we performed MPACT analyses on both their nucleotide and amino acid sequences. Unlike what has been observed for *Orthornavirae* and *Microviridae*, we obtained for *Orthobunyvirus* clear separation of intragroup and intergroup value ranges for the different analyzed metrics (Supplementary Table S8 and Figure 8). These findings suggest that for *Orthobunyavirus*, arbitrary thresholds may serve as effective parameters for

 doi:10.20944/preprints202504.0365.v1



determining the inclusion or exclusion of viral members within distinct serogroup-related clusters. It is conceivable that similar effective demarcation could be achieved for other viral taxa, provided their members exhibit relatively high similarity (e.g., amino acid sequence similarity above 70%).

Collectively, our findings from the *Orthornavirae*, *Microviridae*, and *Orthobunyavirus* datasets demonstrate that viral group demarcation using fixed upper and lower value limits for any single metric can be misleading. This is due to the potential for greater variation within a viral group than between groups, a phenomenon particularly evident in higher taxonomic ranks, as observed with *Orthornavirae*. Conversely, for closely related viral groups, such as members of different serogroups of the genus *Orthobunyavirus*, clear delineation is both possible and effective, validating the ICTV's species demarcation methodology adopted for many viruses. While single metric value limits are unsuitable for higher taxonomic ranks, MPACT's heatmaps and dendrograms, which closely resemble phylogenetic reconstructions (e.g., Figure 3), offer multiple lines of evidence for delineation, especially when incorporating 3D structure-based comparisons.

### *3.4. Data Partitioning*

To explore the potential of using specific metric values to include or exclude sequences within viral groups, we used MPACT to partition the sequence data of the genus *Orthobunyavirus*. Based on the results obtained for intragroup and intergroup comparisons using the various metrics (Supplementary Tabler S8), we tested different value ranges for data partitioning and determined the resulting number of clusters (Supplementary Table S9). Considering that the original *Orthobunyavirus* dataset comprised representatives of 14 serogroups, we chose partitioning value ranges that produced a cluster number as close as possible to 14. Specifically, we chose 65-100% nucleotide identity percentage (18 clusters – Supplementary Table S9) and 0-1.5 of nucleotide ML distance (13 clusters – Supplementary Table S9). Using ML distance, Cluster 1 contains all representatives of the California serogroup and a unique sequence from Bwamba serogroup (Supplementary Table S10). This clustering is congruent with the phylogenetic analysis of the genus *Orthobunyavirus* (Figure 7), as the Bwamba sequence forms a basal branch to the California clade. Cluster 2 is composed solely of Gamboa sequences, similar to Clusters 9 and 13, which contain representatives of the Tete and Maputta serogroups, respectively, in agreement with the monophyly observed in the phylogenetic tree. Serogroups presenting long branches (Figure 7), as observed for the Nyando (Clusters 3 and 4), Anopheles (Clusters 6 and 7), and Simbu (Clusters 10, 11 and 12) serogroups, were separated into distinct clusters (Supplementary n 10). Conversely, serogroups forming sister clades with short distances between them were included in the same clusters. This result was observed for Bunyamwera and Wyeomyia (Cluster 5), as well as for the Guama, Capim, Group C, and Patois (Cluster 8) serogroups.

 doi:10.20944/preprints202504.0365.v1



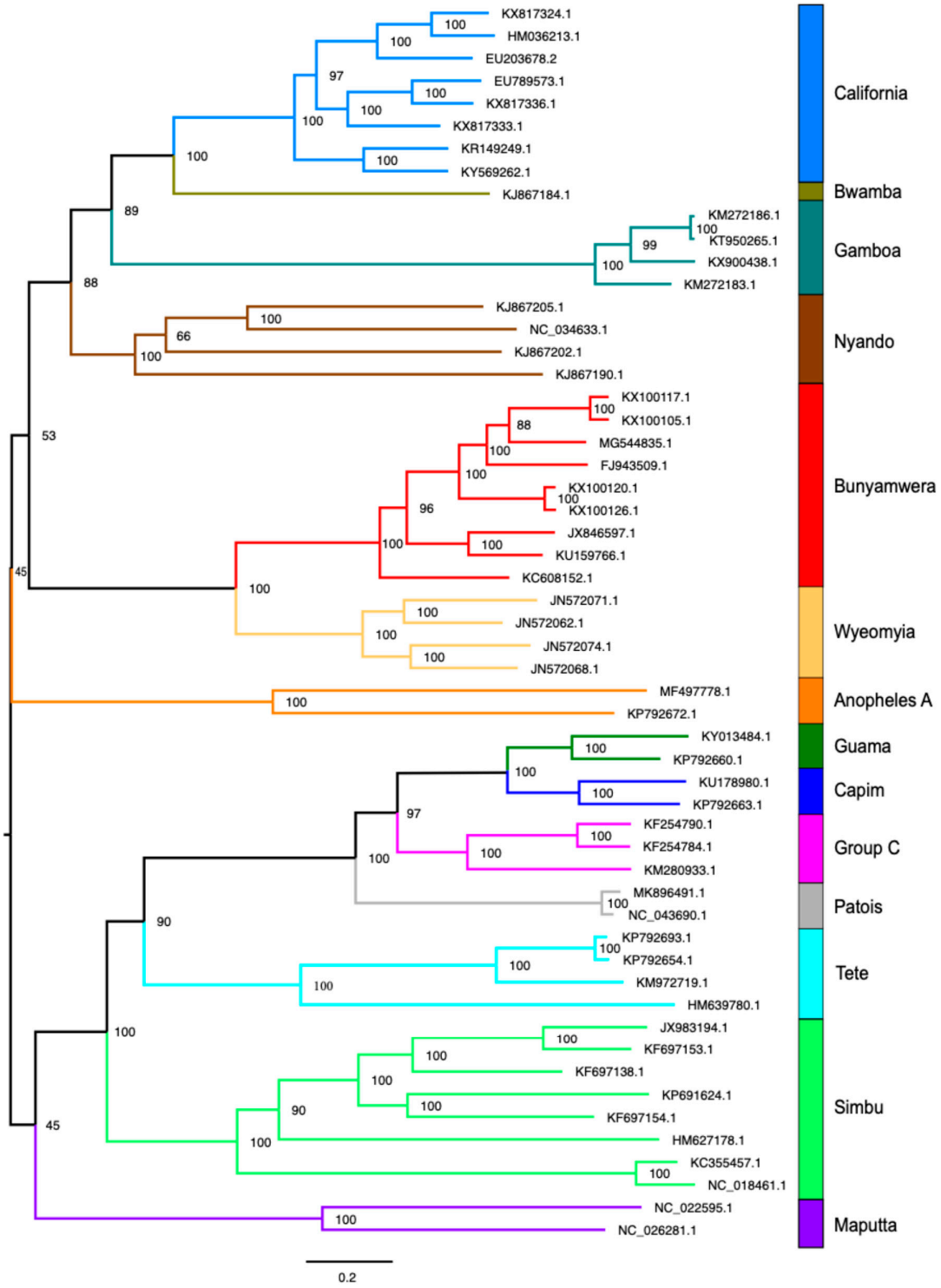



**Figure 7.** Phylogenetic reconstruction. Maximum likelihood tree inferred from nucleotide sequences of the L segment of viruses of the genus *Orthobunyavirus*. The phylogenetic tree was obtained using IQ-TREE with the best-fit model GTR+F+R7 on a multiple sequence alignment generated by MAFFT. The tree is rooted at the midpoint and the nodes are sorted in increasing order. Support values are shown at the nodes of the clades. The colored vertical bars indicate the antigenic groups of the respective clades.

 doi:10.20944/preprints202504.0365.v1



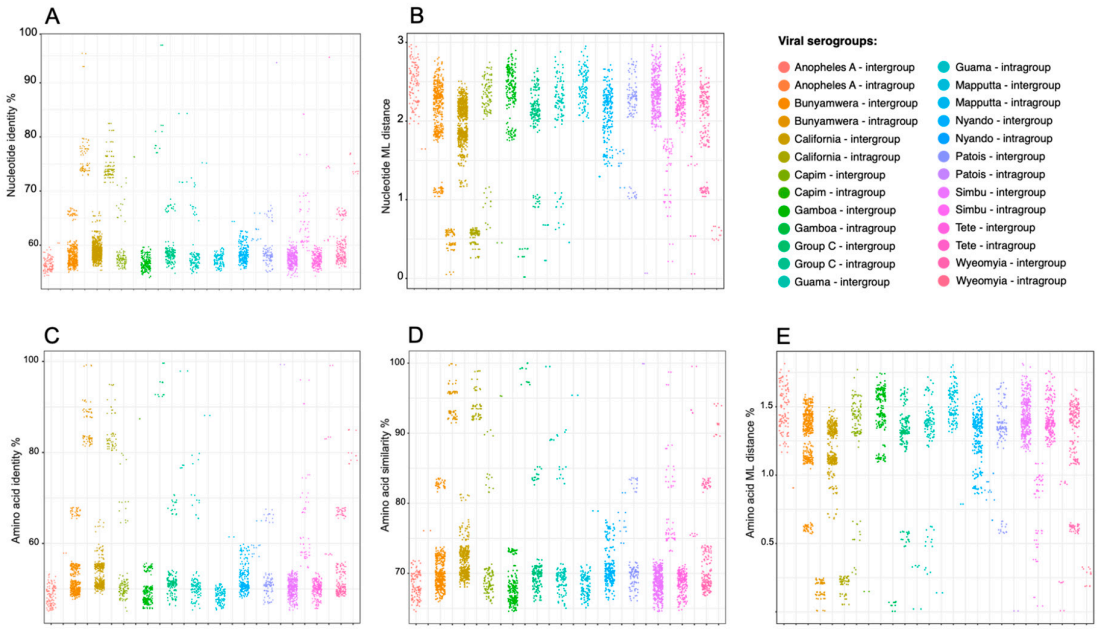



**Figure 8.** All -against-all pairwise nucleotide and amino acid comparisons of sequences derived from the large (L) segment of viruses of the genus Orthobunyavirus. Members of each viral group were defined by a maximum likelihood phylogenetic reconstruction using MAFFT and IQ-TREE, resulting in bona fide subsets. For each subset, sequences were compared all-against-all using pairwise identity percentage (A) and maximum likelihood distance (B) of nucleotide sequences, pairwise identity (C) and similarity (D) percentages and maximum likelihood distance (E) of amino acid sequences. The dots of each column of the scatter plot depict the results obtained for intergroup and intragroup pairwise comparisons, respectively.

In the case of nucleotide pairwise identity percentage, the clustering results were similar, but more stringent, leading to the separation of some clades into additional clusters (Supplementary Table S9). We also tested data partitioning using amino acid sequences with three different metrics to assess whether this type of data would yield clusters more congruent with the serogroup clades obtained in the phylogenetic tree (Figure 7). Our results (Supplementary Table S11) showed no improvement in the partitioning process compared to the results obtained with the corresponding nucleotide sequences (Supplementary Table S10). We conclude that the clustering process itself is functioning properly, but it is not possible to obtain clusters that perfectly match the clades observed in the phylogenetic tree, regardless of the sequence type (nucleotide or amino acid), or metric chosen. The primary reason for this discrepancy is that the clustering process is limited by a single range of metric values applied to the entire dataset. However, as already commented, the evolution rate varies significantly across different clades, resulting in a wide diversity of branch lengths. Consequently, regardless the metric, no single value range can precisely define the clades that correspond to the different serogroups within the genus *Orthobunyavirus*. Within certain ranges, clades encompass all representatives of a serogroup, while other clades may contain a mixture of closely related serogroups. Finally, some serogroups with more distantly related members may be divided into multiple clades, as was the case of the Nyando serogroup. To conclude, although taxa demarcation and data partitioning are feasible with MPACT or comparable tools, users must be keenly aware of the inherent limitations of this methodology.

### *3.5. Comparison of MPACT and Other Similar Tools: SDT and Dali*

Two different publicly available tools are closely related to MPACT: the Sequence Demarcation Tool (SDT) [22] and Dali [30]. SDT was originally developed for all-against-all pairwise alignments of nucleotide sequences but it can also be used for pairwise alignments of amino acid sequences. The program generates color-coded heatmaps according to identity levels and also frequency distribution

 doi:10.20944/preprints202504.0365.v1



plots. Furthermore, SDT allows data partitioning through user-defined lower and upper identity percentage values, enabling easy sequence demarcation. We used a dataset composed of 55 nucleotide sequences of the large (L) segment of viruses of the *Orthobunyavirus* genus to compare MPACT and SDT. Both programs generated comparable heatmaps (Supplementary Figure S6) with different ordering of the taxa, but with very similar clustering, successfully grouping the different serogroups. MPACT incorporates all features available on SDT for nucleotide sequence identity percentage and also ML distance. In addition, MPACT can determine amino acid sequence identity and similarity, ML distance determined from an MSA using evolutionary models, and two metrics based on structural data: the TM-scores calculated from pairwise structural alignments, and 3Di-character pairwise sequence similarity (Supplementary Figure S7). For all these metrics, MPACT can generate the respective heatmaps, UPGMA and NJ dendrograms, and frequency distribution plots, expanding the scope of analyses compared to SDT. In fact, MPACT allows for data partitioning utilizing user-defined upper and lower limits for each metric, providing significantly greater flexibility than SDT for taxa demarcation.

Dali is a program that performs a series of tasks, including structural comparisons of query 3D protein structures against a database of 3D structures. Among its several features, Dali allows to run all-against-all pairwise structural alignments and generates heatmaps and dendrograms derived from average linkage clustering of the structural similarity matrix. The program offers a web server, limited to datasets of up to 64 sequences, or a standalone version that can be installed on a local server. We submitted a dataset of 64 3D structures, predicted by AlphaFold2 from RDRP sequences (Supplementary Table S3) of *Amalgaviridae* viruses, to an all-against-all analysis using MPACT and Dali. Supplementary Figure S8 shows the heatmaps obtained using the TM-score metric by MPACT and the corresponding results produced by Dali. Both programs generated similar heatmaps, featuring a large and easily identifiable cluster composed of members of the *Amalgavirus* genus, as well as several smaller clusters representing minor groups. In addition to Dali, MPACT can also provide comparative all-against-all analyses based of amino acid sequence identity and similarity, and ML distance determined by phylogenetic analyses. Also, 3Di-character sequence alignments can be obtained, increasing the scope of the analyses. Similarly to Dali, MPACT generates an NJ dendrogram, but can also provide dendrograms using a variety of different clustering methods, such as UPGMA and centroid.

## 4. Discussion

### *4.1. Using MPACT for Virus Comparison and Taxa Demarcation*

In this work, we report the development of MPACT, the Multimetric Pairwise Comparison Tool, an integrated program that performs all-against-all pairwise comparisons using both primary biological sequences (nucleotide and amino acid) and 3D protein structures. Unlike existing tools, MPACT is not restricted to a single method or metric; rather, it performs multiple analyses and provides a variety of outputs, including heatmaps, frequency distribution plots, and dendrograms obtained from various clustering methods. This broad set of analyses enables one to apply MPACT to a large gamut of viruses, allowing users to choose the most appropriate metric for determining relationships across viral groups. Notably, 3D structure similarity methods enable the comparison of distantly related viruses, effectively revealing relationships even across highly divergent viral groups, including the different families of the *Orthornavirae* kingdom (Supplementary Figure S1) and the various subfamilies/groups of *Microviridae* phages (Figure 4). Additionally, MPACT implements the use of maximum likelihood (ML) distance, which, despite being a reductionist metric, relies upon phylogenetic reconstruction using evolutionary models. Our results confirmed that ML distance effectively unveils relationships across viruses with greater sensitivity than pairwise alignments, as demonstrated in the resulting heatmaps (Figure 4 – *Amalgaviridae*; Figure 6 – *Microviridae*) and corresponding dendrograms (data not shown). Furthermore, ML distance-derived relationships can be used for data partitioning, with high resemblance to taxa clustering from molecular phylogenetic

 doi:10.20944/preprints202504.0365.v1



methods. For example, data partitioning of *Orthobunyavirus* nucleotide sequences (Supplementary Table 10) using ML distance closely matches the clades from ML phylogenetic reconstruction (Figure 7). Finally, pairwise alignments of either nucleotide or amino acid sequences can be used as effective metrics to reveal relationships among closely related viruses. The capability to utilize multiple metrics based on pairwise alignment (identity and similarity percentage) and multiple sequence alignment (maximum likelihood distance), as well as structural data (TM-score and 3Di similarity percentage), renders the MPACT program significantly more flexible and broadly applicable than other available tools such as SDT and DALI.

### *4.2. Phenotypic and Genotypic Features for Viral Classification*

Viral taxonomy was historically based on the Baltimore classification scheme [8], which initially comprised six viral groups, later expanded to seven, based on genome composition and viral replication strategies. In addition to the classical Baltimore classification, genotypic and phenotypic features can be used for taxonomic delineation.

Phenotypic characterization may more closely correlate with the biological features and have been used for taxa demarcation, especially to define viral species. Viral species have been classically relied upon host specificity or range, cell and tissue tropism, mode of transmission, association to specific pathological/morbid entities, and ecological role of the viruses in a community, a particularly important feature for phages. Taxa demarcation based on phenotypical characteristics is subject to some limitations, including (1) subjectivity; (2) qualitative rather than quantitative features; (3) lack of sufficient knowledge of virus-host associations and specificities; (4) in vitro cultivation not available for many viruses; and (5) unknown environmental factor that may influence viruses.

In the case of genotypic classification, whole-genome sequences and/or their corresponding protein sequences can be employed in phylogenetic reconstruction using evolutionary models, and provide high level of resolution and objectivity, providing a basis for phylogenetic classification. While this approach is being increasingly used, it is also subject to a series of potential limitations: (1) the typical high mutation rates of viruses, which may result in high divergence, leading to inconsistent alignments and trees; (2) no universal markers are available for viruses, that is, no single gene or protein is shared by all viruses, implying that specific markers must be used for particular viral groups; (3) horizontal gene transfer, recombination and segment reassortment events may complicate the analysis and lead to classification inconsistencies; (4) viruses may exhibit distinct evolutionary rates, as well as different molecular markers within the same taxa; (5) consistent evolutionary relationships require good taxa sampling, namely datasets that are representative of viral diversity; (6) the generation of reliable MSAs is increasingly difficult with large datasets and distantly related viruses; (7) establishing reliable thresholds of genetic distance can be tricky, especially for viruses presenting high sequence divergence or distinct evolutionary rates; (8) viruses may mimic their hosts through processes consistent with reticulate evolution; and (9) lack of correlation with phenotype, leading to classifications that show poor correlation with the biology of the viruses and their hosts. Some of these aspects have been used as key arguments in the ongoing debate whether viruses should even be recognized as independent taxonomic units, defined by their unique organization, function and evolution [73,74].

### *4.3. Single Metrics Versus Molecular Phylogeny*

Phylogenetic reconstruction based on primary biological sequences is the golden standard for elucidating the evolutionary relationships, as it is grounded in established evolutionary models. However, the limitations discussed in the previous section present significant challenges to this approach. Achieving comprehensive representation of viral diversity is highly desirable, yet processing large datasets for multiple sequence alignment (MSA) can be computationally demanding and introduce potential inaccuracies, creating an inherent conflict between data breadth and alignment accuracy. As the number of sequences in an MSA increases, the accumulation of indel

 doi:10.20944/preprints202504.0365.v1



events tends to reduce pairwise sequence identity/similarity scores [22] and the overall accuracy of the alignment.

In contrast, methods based on pairwise all-against-all comparisons, while affected by dataset size in terms of computational demand, are not disturbed by the size of input data. For example, SDT implements an all-against-all alignment of nucleotide sequences to determine identity percentage for all sequence pairs, a method proven valuable for taxa demarcation of closely related viruses. Nevertheless, due to the typically high evolutionary rates of viruses, sequence identity percentages become less informative for distantly related taxa. Consequently, the use of more sensitive metrics, such as amino acid similarity percentage, ML distance and, particularly, 3D protein structure alignments, significantly enhances the applicability of pairwise comparisons across a broader range of evolutionary distances.

In this study, we utilized MPACT with diverse metrics, analyzing sequence and structural datasets from highly divergent *Orthornavirae* viruses, moderately divergent *Microviridae* phages, and relatively closely related viruses of the genus *Orthobunyavirus*. As we presented and discussed in the preceding sections, distinct viral groups possess inherent specificities, thereby requiring particular approaches and criteria for the delineation of taxonomic boundaries. By using diverse methodologies, we observed that no single metric provides effective taxa demarcation and data partitioning that are perfectly congruent with classical molecular phylogeny. This is especially true when highly diverse viral groups are analyzed. In this case, phylogenetic analyses based on evolutionary models may provide significantly more information than any metrics discretized to single values.

Classifications based on reductionist metrics, regardless of their nature, may deviate from the true evolutionary relationships of the respective organisms. Consequently, any cutoff value for taxa demarcation is arbitrary and limited, potentially resulting in clusters that do not necessarily reflect evolutionary clades. Therefore, while specific lower and upper range limits for taxa demarcation may be effective for a particular viral group, they should not be indiscriminately extrapolated to diverse viral groups, as evolutionary rates vary among them, and consequently, so do the degrees of divergence across different metrics. We demonstrated that boundary definition becomes increasingly complex for more divergent viral groups, due to varying evolutionary rates across distinct lineages, leading to internal clade members potentially becoming more distant from each other than from external members.

Consequently, reliable demarcation is feasible for closely related viral genomes, such as those within the genus *Orthobunyavirus*, but becomes challenging for more distant groups, as exemplified by *Orthornavirae* (Supplementary Figure S4) and *Microviridae* (Supplementary Figure S5). In light of the debate of whether viruses should even be considered bona fide taxonomic units [74], species demarcation is still a controversial issue and perhaps remains a question that may prove objectively irresolvable. Considering that phylogenetic analysis does not necessarily cover the intricate relationship between viruses and their hosts, we agree with the principle that alternative categorizations based on infectivity and virulence may be more convenient for disease prevention and treatment, even though they do not reflect evolutionary relationships [1].

Although viral taxa demarcation may not perfectly reflect true evolutionary relationships, it does offer significant practical applications and benefits that should not be underestimated. It facilitates standardized viral nomenclature, data sharing, viral database management, and other crucial aspects. Rather than providing rigid foundations, taxa demarcation, based on selected metrics and criteria, should be viewed as a valuable framework for organizing and understanding the viral world, evolving dynamically as new information emerges. In this context, we believe that MPACT serves as a valuable tool for assisting the scientific community in delineating viral taxa across a spectrum of criteria. By providing heatmaps and distance trees derived from sequence and 3D protein structure data, MPACT offers multiple lines of evidence to support taxa demarcation, even for distantly related viral groups where traditional upper and lower threshold values may be ineffective.

### *4.5. Protein Structure Information and MSAs*

 doi:10.20944/preprints202504.0365.v1



Structural similarity, due to its higher conservation compared to primary sequence similarity [26], can assist in identifying distant homologs, as well as improve the quality of MSAs. Two distinct approaches have been reported in the literature to increment MSAs by using structural information. FoldMason [75] performs progressive multiple structural alignment (MSTA) using the 3Di-character alphabet developed for Foldseek [32]. This alphabet, comprising 20 discrete characters representing local 3D arrangement of protein structures, reduces complexity, enabling the alignment of hundreds of thousands of protein structures within reasonable computational times. In additional to increased processing speed, Foldmason generates multiple structural alignment (MSTAs) that can improve the accuracy of phylogenetic analyses for distantly related proteins within the twilight zone. As an alternative to the 3Di-character alphabet, Edgar [76] developed a novel "mega-alphabet" in which each residue in the protein backbone is represented by a unique character, resulting in over 85 billion distinct states. The Reseek program, described in this report, not only generates this alphabet from 3D protein structures but also performs protein homolog detection with higher sensitivity and improved speed compared to other available tools such as TM-align, DALI, and Foldseek. The same author, who developed MUSCLE5 [77], an accurate multiple sequence aligner, recently implemented the "Muscle-3D" feature in MUSCLE5 (https://github.com/rcedgar/muscle), enabling the use of "mega-alphabet" input sequences and the generation of multiple structure alignments [78]. Once these alignments are obtained, MUSCLE5 converts them to conventional amino acid sequence alignments, which can be used by standard phylogenetic reconstruction tools. We intend to incorporate the utilization of FoldMason and Reseek/Muscle-3D in future implementations of the MPAC program, aiming at improving taxonomic classification of evolutionarily distant viruses.

### *4.6. For viral Detection and Classification Go Shorter*

Beyond viral taxonomic studies, viral classification based on genomic and metagenomic sequencing data is also crucial, especially for molecular epidemiology and virus surveillance. Our previous studies have demonstrated that viral detection and discrimination from metagenomic data are more accurate when performed using short, informative protein regions, rather than full-length sequences [6,7,79]. We recently described TABAJARA [25], a tool that implements various methods to select sequence alignment blocks conserved across all sequences of an MSA or, alternatively, to identify blocks that are specific signatures of viral taxa. The program objectively determines such informative blocks, generates profile HMMs, and validates these models against training sets to estimate their sensitivity and specificity. By using customized cutoff scores, these profile HMMs, specific to a variety of viral taxa, can be used to detect and taxonomically classify viral sequences from metagenomic datasets. We envision a combined methodology that merges MPACT and molecular phylogeny to delineate viral taxa groups. For taxa discrimination and inclusion of new viruses into established taxa, the use of informative regions instead of entire sequences or structures makes more sense by using profile HMMs designed for informative regions of each taxon.

**Supplementary Materials:** The following supporting information can be downloaded from website of this paper posted on Preprints.org, Figure S1: Frequency distribution plots of all-against-all pairwise comparisons of RDRP protein sequences of RNA virus families of the *Orthornavirae* kingdom, Figure S2: Three-dimensional structures of the RDRP of representative viruses of the *Amalgaviridae* family, Figure S3: Three-dimensional structures of VP1 (major capsid protein) of the subfamilies/groups of the *Microviridae* family, Figure S4: All-against-all pairwise comparisons of RDRP sequences of viral families of the *Orthornavirae* kingdom, Figure S5: All-against-all pairwise comparisons of VP1 sequences of *Microviridae* phages, Figure S6: Heatmaps generated by MPACT (A) and SDT (B) programs of all-against-all pairwise comparisons of the large (L) segment of viruses of the genus *Orthobunyavirus* using identity percentage of nucleotide sequences, Figure S7: Heatmaps of all-against-all pairwise comparisons of sequences derived from the large (L) segment of viruses of the genus *Orthobunyavirus*, Figure S8: Comparison of heatmaps generated by MPACT (A) and Dali (B), Table S1: Protein sequences of RDRP of viruses of the *Orthornavirae* kingdom, Table S2: ORF 1

 doi:10.20944/preprints202504.0365.v1



protein sequences of *Amalgaviridae* viruses, Table S3: RNA-directed RNA polymerase sequences of *Amalgaviridae* viruses, Table S4: Protein sequences of VP1 proteins of *Microviridae* phages, Table S5: Nucleotide (NT) and amino acid (AA) sequences derived from the L segment of viruses of the genus *Orthobunyavirus*, Table S6: Value ranges of intragroup and intergroup comparisons of viral families of the *Orthornavirae* Kingdom, Table S7: Value ranges of intragroup and intergroup comparisons of distinct *Microviridae* subfamilies/groups, Table S8: Value ranges of intragroup and intergroup comparisons of distinct antigenic groups of *Orthobunyavirus*, Table S9: Data partition of a dataset of 55 sequences of the large (L) segment of *Orthobunyavirus* viruses using different metrics of the MPACT program, Table S10: Data partitioning of a dataset of 55 nucleotide sequences of the large (L) segment of *Orthobunyavirus* viruses using the MPACT program with ranges of 65 to 100% of identity and 0 to 1.5 for maximum-likelihood distance as partitioning criteria, Table S11: Data partition of a dataset of 55 amino acid sequences of the RDRP coded by the large (L) segment of *Orthobunyavirus* viruses using the MPACT program with ranges of 64 to 100% of identity percentage, 80 to 100% of similarity percentage and 0 to 0.8 for maximum-likelihood distance as partitioning criteria.

**Author Contributions:** Conceptualization, A.G.; methodology, I.C.S., R.S.S., I.T. and A.G.; software, I.C.S and A.G.; validation, I.C.S, L.S.O., R.S.S., I.T. and A.G.; formal analysis, A.G.; investigation, I.C.S., A.G. and L.S.O.; resources, A.G.; data curation, I.C.S., L.S.O. and A.G.; writing—original draft preparation, A.G. and L.S.O.; writing—review and editing, I.C.S., R.S.S., I.T., L.S.O., and A.G.; visualization, A.G. and L.S.O.; project administration, A.G.; supervision, A.G.; funding acquisition, A.G. All authors have read and agreed to the published version of the manuscript.

**Funding:** ICS received a Scientific Initiation scholarship from the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP 2023/17992-8); LSO received a post-doctoral fellowship from the Fundação Araucária (Project: NAPI Bioinformática) via grant 66.2021.

**Institutional Review Board Statement:** Not applicable

**Informed Consent Statement:** Not applicable

**Data Availability Statement:** MPACT (Multimetric Pairwise Comparison Tool) is an open-source program available for download in the GitHub repository (https://github.com/gruberlab/mpact), under the terms of the GNU General Public License version 3. The program is fully documented, and a tutorial is provided. Datasets used throughout this work, including nucleotide and protein sequences, and multiple sequence alignments are publicly available in the Supplementary Material).

**Acknowledgments:** ICS and LSO are grateful to FAPESP and Fundação Araucária, respectively, for their scholarships.

**Conflicts of Interest:** The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| 3D | Three-dimensional |
| ML | Maximum likelihood |
| NJ | Neighbor-joining |
| ORF | Open reading frame |
| pLDDT | Predicted local distance difference test |
| MPACT | Multimetric Pairwise Comparison Tool |
| MSA | Multiple sequence alignment |
| RDRP | RNA-dependent RNA polymerase |
| NCBI | National Center of Biotechnology Information |
| UPGMA | Unweighted Pair-Group Method using Arithmetic Averages |

## References

1. Simmonds, P.; Aiewsakun, P. Virus Classification – Where Do You Draw the Line? *Arch Virol* **2018**, *163*, 2037–2046, doi:10.1007/s00705-018-3938-z.

 doi:10.20944/preprints202504.0365.v1




2. Espino-Vázquez, A.N.; Bermúdez-Barrientos, J.R.; Cabrera-Rangel, J.F.; Córdova-López, G.; Cardoso-Martínez, F.; Martínez-Vázquez, A.; Camarena-Pozos, D.A.; Mondo, S.J.; Pawlowska, T.E.; Abreu-Goodger, C.; et al. Narnaviruses: Novel Players in Fungal–Bacterial Symbioses. *ISME J* **2020**, *14*, 1743–1754, doi:10.1038/s41396-020-0638-y.
3. Fonseca, P.; Ferreira, F.; da Silva, F.; Oliveira, L.S.; Marques, J.T.; Goes-Neto, A.; Aguiar, E.; Gruber, A. Characterization of a Novel Mitovirus of the Sand Fly Lutzomyia Longipalpis Using Genomic and Virus–Host Interaction Signatures. *Viruses* **2020**, *13*, 9, doi:10.3390/v13010009.
4. Krupovic, M.; Dolja, V.V.; Koonin, E.V. The LUCA and Its Complex Virome. *Nat Rev Microbiol* **2020**, *18*, 661–670, doi:10.1038/s41579-020-0408-x.
5. Forterre, P. The Origin of Viruses and Their Possible Roles in Major Evolutionary Transitions. *Virus Research* **2006**, *117*, 5–16, doi:10.1016/j.virusres.2006.01.010.
6. Oliveira, L.S.; Gruber, A. Rational Design of Profile HMMs for Viral Classification and Discovery. In *Bioinformatics*; Nakaya, H., Ed.; Exon Publications: Brisbane, Australia, 2021; pp. 151–170 ISBN 978-0-645-00171-6.
7. Reyes, A.; Alves, J.M.P.; Durham, A.M.; Gruber, A. Use of Profile Hidden Markov Models in Viral Discovery: Current Insights. *Adv Genom Genet* **2017**, *Volume 7*, 29–45, doi:10.2147/AGG.S136574.
8. Baltimore, D. Expression of Animal Virus Genomes. *Bacteriol Rev* **1971**, *35*, 235–241, doi:10.1128/br.35.3.235-241.1971.
9. Koonin, E.V.; Krupovic, M.; Agol, V.I. The Baltimore Classification of Viruses 50 Years Later: How Does It Stand in the Light of Virus Evolution? *Microbiol Mol Biol Rev* **2021**, *85*, e00053-21, doi:10.1128/MMBR.00053-21.
10. Siddell, S.G.; Walker, P.J.; Lefkowitz, E.J.; Mushegian, A.R.; Dutilh, B.E.; Harrach, B.; Harrison, R.L.; Junglen, S.; Knowles, N.J.; Kropinski, A.M.; et al. Binomial Nomenclature for Virus Species: A Consultation. *Arch Virol* **2020**, *165*, 519–525, doi:10.1007/s00705-019-04477-6.
11. Zerbini, F.M.; Siddell, S.G.; Mushegian, A.R.; Walker, P.J.; Lefkowitz, E.J.; Adriaenssens, E.M.; Alfenas-Zerbini, P.; Dutilh, B.E.; García, M.L.; Junglen, S.; et al. Differentiating between Viruses and Virus Species by Writing Their Names Correctly. *Arch Virol* **2022**, *167*, 1231–1234, doi:10.1007/s00705-021-05323-4.
12. International Committee on Taxonomy of Viruses Executive Committee; Gorbalenya, A.E.; Krupovic, M.; Mushegian, A.; Kropinski, A.M.; Siddell, S.G.; Varsani, A.; Adams, M.J.; Davison, A.J.; Dutilh, B.E.; et al. The New Scope of Virus Taxonomy: Partitioning the Virosphere into 15 Hierarchical Ranks. *Nat Microbiol* **2020**, *5*, 668–674, doi:10.1038/s41564-020-0709-x.
13. Simmonds, P.; Adriaenssens, E.M.; Zerbini, F.M.; Abrescia, N.G.A.; Aiewsakun, P.; Alfenas-Zerbini, P.; Bao, Y.; Barylski, J.; Drosten, C.; Duffy, S.; et al. Four Principles to Establish a Universal Virus Taxonomy. *PLoS Biol* **2023**, *21*, e3001922, doi:10.1371/journal.pbio.3001922.
14. Gorbalenya, A.E.; Lauber, C. Bioinformatics of Virus Taxonomy: Foundations and Tools for Developing Sequence-Based Hierarchical Classification. *Current Opinion in Virology* **2022**, *52*, 48–56, doi:10.1016/j.coviro.2021.11.003.
15. Evseev, P.; Gutnik, D.; Shneider, M.; Miroshnikov, K. Use of an Integrated Approach Involving AlphaFold Predictions for the Evolutionary Taxonomy of Duplodnaviria Viruses. *Biomolecules* **2023**, *13*, 110, doi:10.3390/biom13010110.
16. Aiewsakun, P.; Simmonds, P. The Genomic Underpinnings of Eukaryotic Virus Taxonomy: Creating a Sequence-Based Framework for Family-Level Virus Classification. *Microbiome* **2018**, *6*, 38, doi:10.1186/s40168-018-0422-7.
17. Aiewsakun, P.; Adriaenssens, E.M.; Lavigne, R.; Kropinski, A.M.; Simmonds, P. Evaluation of the Genomic Diversity of Viruses Infecting Bacteria, Archaea and Eukaryotes Using a Common Bioinformatic Platform: Steps towards a Unified Taxonomy. *Journal of General Virology* **2018**, *99*, 1331–1343, doi:10.1099/jgv.0.001110.
18. Bin Jang, H.; Bolduc, B.; Zablocki, O.; Kuhn, J.H.; Roux, S.; Adriaenssens, E.M.; Brister, J.R.; Kropinski, A.M.; Krupovic, M.; Lavigne, R.; et al. Taxonomic Assignment of Uncultivated Prokaryotic Virus Genomes Is Enabled by Gene-Sharing Networks. *Nat Biotechnol* **2019**, *37*, 632–639, doi:10.1038/s41587-019-0100-8.
19. Meier-Kolthoff, J.P.; Göker, M. VICTOR: Genome-Based Phylogeny and Classification of Prokaryotic Viruses. *Bioinformatics* **2017**, *33*, 3396–3404, doi:10.1093/bioinformatics/btx440.

 doi:10.20944/preprints202504.0365.v1




20. Bao, Y.; Chetvernin, V.; Tatusova, T. PAirwise Sequence Comparison (PASC) and Its Application in the Classification of Filoviruses. *Viruses* **2012**, *4*, 1318–1327, doi:10.3390/v4081318.
21. Lauber, C.; Gorbalenya, A.E. Partitioning the Genetic Diversity of a Virus Family: Approach and Evaluation through a Case Study of Picornaviruses. *J Virol* **2012**, *86*, 3890–3904, doi:10.1128/JVI.07173-11.
22. Muhire, B.M.; Varsani, A.; Martin, D.P. SDT: A Virus Classification Tool Based on Pairwise Sequence Alignment and Identity Calculation. *PLoS ONE* **2014**, *9*, e108277, doi:10.1371/journal.pone.0108277.
23. Brenner, S.E.; Chothia, C.; Hubbard, T.J.P. Assessing Sequence Comparison Methods with Reliable Structurally Identified Distant Evolutionary Relationships. *Proceedings of the National Academy of Sciences* **1998**, *95*, 6073–6078, doi:10.1073/pnas.95.11.6073.
24. Park, J.; Karplus, K.; Barrett, C.; Hughey, R.; Haussler, D.; Hubbard, T.; Chothia, C. Sequence Comparisons Using Multiple Sequences Detect Three Times as Many Remote Homologues as Pairwise Methods. *Journal of Molecular Biology* **1998**, *284*, 1201–1210, doi:10.1006/jmbi.1998.2221.
25. Oliveira, L.S.; Reyes, A.; Dutilh, B.E.; Gruber, A. Rational Design of Profile HMMs for Sensitive and Specific Sequence Detection with Case Studies Applied to Viruses, Bacteriophages, and Casposons. *Viruses* **2023**, *15*, 519, doi:10.3390/v15020519.
26. Caetano-Anolles Benefits of Using Molecular Structure and Abundance in Phylogenomic Analysis. *FGENE* **2012**, doi:10.3389/fgene.2012.00172.
27. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596*, 583–589, doi:10.1038/s41586-021-03819-2.
28. Varadi, M.; Velankar, S. The Impact of AlphaFold Protein Structure Database on the Fields of Life Sciences. *Proteomics* **2023**, *23*, 2200128, doi:10.1002/pmic.202200128.
29. Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; et al. Evolutionary-Scale Prediction of Atomic-Level Protein Structure with a Language Model. *Science* **2023**, *379*, 1123–1130, doi:10.1126/science.ade2574.
30. Holm, L. Dali Server: Structural Unification of Protein Families. *Nucleic Acids Research* **2022**, *50*, W210–W215, doi:10.1093/nar/gkac387.
31. Zhang, Y.; Skolnick, J. TM-Align: A Protein Structure Alignment Algorithm Based on the TM-Score. *Nucleic Acids Res* **2005**, *33*, 2302–2309, doi:10.1093/nar/gki524.
32. Van Kempen, M.; Kim, S.S.; Tumescheit, C.; Mirdita, M.; Lee, J.; Gilchrist, C.L.M.; Söding, J.; Steinegger, M. Fast and Accurate Protein Structure Search with Foldseek. *Nat Biotechnol* **2023**, doi:10.1038/s41587-023-01773-0.
33. Steinegger, M.; Söding, J. MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets. *Nat Biotechnol* **2017**, *35*, 1026–1028, doi:10.1038/nbt.3988.
34. Kuhn, J.H.; Abe, J.; Adkins, S.; Alkhovsky, S.V.; Avšič-Županc, T.; Ayllón, M.A.; Bahl, J.; Balkema-Buschmann, A.; Ballinger, M.J.; Kumar Baranwal, V.; et al. Annual (2023) Taxonomic Update of RNA-Directed RNA Polymerase-Encoding Negative-Sense RNA Viruses (Realm Riboviria: Kingdom Orthornavirae: Phylum Negarnaviricota). *Journal of General Virology* **2023**, *104*, doi:10.1099/jgv.0.001864.
35. Martin, R.R.; Zhou, J.; Tzanetakis, I.E. Blueberry Latent Virus: An Amalgam of the Partitiviridae and Totiviridae. *Virus Research* **2011**, *155*, 175–180, doi:10.1016/j.virusres.2010.09.020.
36. Vainio, E.J.; Chiba, S.; Ghabrial, S.A.; Maiss, E.; Roossinck, M.; Sabanadzovic, S.; Suzuki, N.; Xie, J.; Nibert, M.; ICTV Report Consortium ICTV Virus Taxonomy Profile: Partitiviridae. *Journal of General Virology* **2018**, *99*, 17–18, doi:10.1099/jgv.0.000985.
37. Nibert, M.L.; Pyle, J.D.; Firth, A.E. A +1 Ribosomal Frameshifting Motif Prevalent among Plant Amalgaviruses. *Virology* **2016**, *498*, 201–208, doi:10.1016/j.virol.2016.07.002.
38. Depierreux, D.; Vong, M.; Nibert, M.L. Nucleotide Sequence of Zygosaccharomyces Bailii Virus Z: Evidence for +1 Programmed Ribosomal Frameshifting and for Assignment to Family Amalgaviridae. *Virus Research* **2016**, *217*, 115–124, doi:10.1016/j.virusres.2016.02.008.
39. Roux, S.; Krupovic, M.; Poulet, A.; Debroas, D.; Enault, F. Evolution and Diversity of the Microviridae Viral Family through a Collection of 81 New Complete Genomes Assembled from Virome Reads. *PloS one* **2012**, *7*, e40418, doi:10.1371/journal.pone.0040418.

 doi:10.20944/preprints202504.0365.v1




40. Brentlinger, K.L.; Hafenstein, S.; Novak, C.R.; Fane, B.A.; Borgon, R.; McKenna, R.; Agbandje-McKenna, M. Microviridae, a Family Divided: Isolation, Characterization, and Genome Sequence of phiMH2K, a Bacteriophage of the Obligate Intracellular Parasitic Bacterium Bdellovibrio Bacteriovorus. *J Bacteriol* **2002**, *184*, 1089–1094, doi:10.1128/jb.184.4.1089-1094.2002.
41. Carstens, E.B. Ratification Vote on Taxonomic Proposals to the International Committee on Taxonomy of Viruses (2009). *Arch Virol* **2010**, *155*, 133–146, doi:10.1007/s00705-009-0547-x.
42. Tikhe, C.V.; Husseneder, C. Metavirome Sequencing of the Termite Gut Reveals the Presence of an Unexplored Bacteriophage Community. *Front. Microbiol.* **2018**, *8*, 2548, doi:10.3389/fmicb.2017.02548.
43. Rosario, K.; Dayaram, A.; Marinov, M.; Ware, J.; Kraberger, S.; Stainton, D.; Breitbart, M.; Varsani, A. Diverse Circular ssDNA Viruses Discovered in Dragonflies (Odonata: Epiprocta). *Journal of General Virology* **2012**, *93*, 2668–2681, doi:10.1099/vir.0.045948-0.
44. Quaiser, A.; Dufresne, A.; Ballaud, F.; Roux, S.; Zivanovic, Y.; Colombet, J.; Sime-Ngando, T.; Francez, A.-J. Diversity and Comparative Genomics of Microviridae in Sphagnum- Dominated Peatlands. *Front. Microbiol.* **2015**, *6*, doi:10.3389/fmicb.2015.00375.
45. Zhang, L.; Li, Z.; Bao, M.; Li, T.; Fang, F.; Zheng, Y.; Liu, Y.; Xu, M.; Chen, J.; Deng, X.; et al. A Novel Microviridae Phage (CLasMV1) From "Candidatus Liberibacter Asiaticus." *Front. Microbiol.* **2021**, *12*, 754245, doi:10.3389/fmicb.2021.754245.
46. Zheng, Q.; Chen, Q.; Xu, Y.; Suttle, C.A.; Jiao, N. A Virus Infecting Marine Photoheterotrophic Alphaproteobacteria (Citromicrobium Spp.) Defines a New Lineage of ssDNA Viruses. *Front. Microbiol.* **2018**, *9*, 1418, doi:10.3389/fmicb.2018.01418.
47. Krupovic, M.; Dutilh, B.E.; Adriaenssens, E.M.; Wittmann, J.; Vogensen, F.K.; Sullivan, M.B.; Rumnieks, J.; Prangishvili, D.; Lavigne, R.; Kropinski, A.M.; et al. Taxonomy of Prokaryotic Viruses: Update from the ICTV Bacterial and Archaeal Viruses Subcommittee. *Arch Virol* **2016**, *161*, 1095–1099, doi:10.1007/s00705-015-2728-0.
48. Bryson, S.J.; Thurber, A.R.; Correa, A.M.S.; Orphan, V.J.; Vega Thurber, R. A Novel Sister Clade to the Enterobacteria Microviruses (Family *Microviridae* ) Identified in Methane Seep Sediments: DNA Phages Associated with Methane Seeps. *Environ Microbiol* **2015**, *17*, 3708–3721, doi:10.1111/1462-2920.12758.
49. Creasy, A.; Rosario, K.; Leigh, B.; Dishaw, L.; Breitbart, M. Unprecedented Diversity of ssDNA Phages from the Family Microviridae Detected within the Gut of a Protochordate Model Organism (Ciona Robusta). *Viruses* **2018**, *10*, 404, doi:10.3390/v10080404.
50. Zucker, F.; Bischoff, V.; Olo Ndela, E.; Heyerhoff, B.; Poehlein, A.; Freese, H.M.; Roux, S.; Simon, M.; Enault, F.; Moraru, C. New *Microviridae* Isolated from *Sulfitobacter* Reveals Two Cosmopolitan Subfamilies of Single-Stranded DNA Phages Infecting Marine and Terrestrial Alphaproteobacteria. *Virus Evolution* **2022**, *8*, veac070, doi:10.1093/ve/veac070.
51. Olo Ndela, E.; Roux, S.; Henke, C.; Sczyrba, A.; Sime Ngando, T.; Varsani, A.; Enault, F. Reekeekee- and Roodoodooviruses, Two Different *Microviridae* Clades Constituted by the Smallest DNA Phages. *Virus Evolution* **2023**, *9*, veac123, doi:10.1093/ve/veac123.
52. De Souza, W.M.; Calisher, C.H.; Carrera, J.P.; Hughes, H.R.; Nunes, M.R.T.; Russell, B.; Tilson-Lunel, N.L.; Venter, M.; Xia, H. ICTV Virus Taxonomy Profile: Peribunyaviridae 2024: This Article Is Part of the ICTV Virus Taxonomy Profiles Collection. *Journal of General Virology* **2024**, *105*, doi:10.1099/jgv.0.002034.
53. Katoh, K.; Standley, D.M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* **2013**, *30*, 772–780, doi:10.1093/molbev/mst010.
54. Nguyen, L.-T.; Schmidt, H.A.; von Haeseler, A.; Minh, B.Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution* **2015**, *32*, 268–274, doi:10.1093/molbev/msu300.
55. Kalyaanamoorthy, S.; Minh, B.Q.; Wong, T.K.F.; von Haeseler, A.; Jermiin, L.S. ModelFinder: Fast Model Selection for Accurate Phylogenetic Estimates. *Nat Methods* **2017**, *14*, 587–589, doi:10.1038/nmeth.4285.
56. Hoang, D.T.; Chernomor, O.; von Haeseler, A.; Minh, B.Q.; Vinh, L.S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Molecular Biology and Evolution* **2018**, *35*, 518–522, doi:10.1093/molbev/msx281.
57. Mirdita, M.; Schütze, K.; Moriwaki, Y.; Heo, L.; Ovchinnikov, S.; Steinegger, M. ColabFold: Making Protein Folding Accessible to All. *Nat Methods* **2022**, *19*, 679–682, doi:10.1038/s41592-022-01488-1.

 doi:10.20944/preprints202504.0365.v1




58. Mariani, V.; Biasini, M.; Barbato, A.; Schwede, T. lDDT: A Local Superposition-Free Score for Comparing Protein Structures and Models Using Distance Difference Tests. *Bioinformatics* **2013**, *29*, 2722–2728, doi:10.1093/bioinformatics/btt473.
59. Rice, P.; Longden, I.; Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics* **2000**, *16*, 276–277, doi:10.1016/S0168-9525(00)02024-2.
60. Sabanadzovic, S.; Valverde, R.A.; Brown, J.K.; Martin, R.R.; Tzanetakis, I.E. Southern Tomato Virus: The Link between the Families Totiviridae and Partitiviridae. *Virus Research* **2009**, *140*, 130–137, doi:10.1016/j.virusres.2008.11.018.
61. Ghabrial, S.A.; Nibert, M.L. Victorivirus, a New Genus of Fungal Viruses in the Family Totiviridae. *Arch Virol* **2009**, *154*, 373–379, doi:10.1007/s00705-008-0272-x.
62. Isogai, M.; Nakamura, T.; Ishii, K.; Watanabe, M.; Yamagishi, N.; Yoshikawa, N. Histochemical Detection of Blueberry Latent Virus in Highbush Blueberry Plant. *J Gen Plant Pathol* **2011**, *77*, 304–306, doi:10.1007/s10327-011-0323-0.
63. Kirchberger, P.C.; Ochman, H. Microviruses: A World Beyond phiX174. *Annu Rev Virol* **2023**, *10*, 99–118, doi:10.1146/annurev-virology-100120-011239.
64. Lee, H.; Baxter, A.J.; Bator, C.M.; Fane, B.A.; Hafenstein, S.L. Cryo-EM Structure of Gokushovirus ΦEC6098 Reveals a Novel Capsid Architecture for a Single-Scaffolding Protein, Microvirus Assembly System. *J Virol* **2022**, *96*, e0099022, doi:10.1128/jvi.00990-22.
65. Gago, S.; Elena, S.F.; Flores, R.; Sanjuan, R. Extremely High Mutation Rate of a Hammerhead Viroid. *Science* **2009**, *323*, 1308, doi:10.1126/science.1169202.
66. Sanjuán, R.; Nebot, M.R.; Chirico, N.; Mansky, L.M.; Belshaw, R. Viral Mutation Rates. *JVI* **2010**, *84*, 9733–9748, doi:10.1128/JVI.00694-10.
67. Holland, J.; Spindler, K.; Horodyski, F.; Grabau, E.; Nichol, S.; VandePol, S. Rapid Evolution of RNA Genomes. *Science* **1982**, *215*, 1577–1585, doi:10.1126/science.7041255.
68. Drake, J.W. Rates of Spontaneous Mutation among RNA Viruses. *Proceedings of the National Academy of Sciences* **1993**, *90*, 4171–4175, doi:10.1073/pnas.90.9.4171.
69. Peck, K.M.; Lauring, A.S. Complexities of Viral Mutation Rates. *J Virol* **2018**, *92*, e01031-17, /jvi/92/14/e01031-17.atom, doi:10.1128/JVI.01031-17.
70. Dias, H.G.; Dos Santos, F.B.; Pauvolid-Corrêa, A. An Overview of Neglected Orthobunyaviruses in Brazil. *Viruses* **2022**, *14*, 987, doi:10.3390/v14050987.
71. Briese, T.; Calisher, C.H.; Higgs, S. Viruses of the Family Bunyaviridae: Are All Available Isolates Reassortants? *Virology* **2013**, *446*, 207–216, doi:10.1016/j.virol.2013.07.030.
72. Calisher, C.H. History, Classification, and Taxonomy of Viruses in the Family Bunyaviridae. In *The Bunyaviridae*; Elliott, R.M., Ed.; Springer US: Boston, MA, 1996; pp. 1–17 ISBN 978-1-4899-1366-1.
73. Caetano-Anollés, G.; Claverie, J.-M.; Nasir, A. A Critical Analysis of the Current State of Virus Taxonomy. *Front. Microbiol.* **2023**, *14*, 1240993, doi:10.3389/fmicb.2023.1240993.
74. Caetano-Anollés, G. Are Viruses Taxonomic Units? A Protein Domain and Loop-Centric Phylogenomic Assessment. *Viruses* **2024**, *16*, 1061, doi:10.3390/v16071061.
75. Gilchrist, C.L.M.; Mirdita, M.; Steinegger, M. Multiple Protein Structure Alignment at Scale with FoldMason 2024.
76. Edgar, R.C. Protein Structure Alignment by Reseek Improves Sensitivity to Remote Homologs. *Bioinformatics* **2024**, *40*, btae687, doi:10.1093/bioinformatics/btae687.
77. Edgar, R.C. Muscle5: High-Accuracy Alignment Ensembles Enable Unbiased Assessments of Sequence Homology and Phylogeny. *Nat Commun* **2022**, *13*, 6968, doi:10.1038/s41467-022-34630-w.

 doi:10.20944/preprints202504.0365.v1



78. Edgar, R.C.; Tolstoy, I. Muscle-3D: Scalable Multiple Protein Structure Alignment 2024.
79. Alves, J.M.P.; de Oliveira, A.L.; Sandberg, T.O.M.; Moreno-Gallego, J.L.; de Toledo, M.A.F.; de Moura, E.M.M.; Oliveira, L.S.; Durham, A.M.; Mehnert, D.U.; Zanotto, P.M. de A.; et al. GenSeed-HMM: A Tool for Progressive Assembly Using Profile HMMs as Seeds and Its Application in Alpavirinae Viral Discovery from Metagenomic Data. *Front. Microbiol.* **2016**, *7*, doi:10.3389/fmicb.2016.00269.