VirTaK (Virus Taxonomy K-mer-based)

Alejandro Miguel Cisneros-Martínez

March 7th 2025

Metagenomic viral contigs (mVCs) can only be characterised by features inferred from their genetic sequence, so it is necessary to employ metrics of genetic relatedness in order to assign them a taxonomic classification (Simmonds et al. 2017; Simmonds and Aiewsakun 2018). Although different virus taxonomic classification tools exist (see a review in Gorbalenya and Lauber 2022), we decided to develop VirTaK (Virus Taxonomy from Kmers, available at: https://github.com/AleCisMar/VirTaK) as a new accessible and resource-efficient tool that could rapidly assess the similarity between a query viral genome and an extensive database of officially recognised viruses. This was achieved by creating a database derived from the ICTV's Virus Metadata Resource (VMR), which includes examples of well-characterised virus isolates for each virus species, and requires only 5.1 gigabytes of space, including a copy of the Pfam-A database.

Given the changing dynamics of virus taxonomy (Gorbalenya 2018; Siddell et al. 2019), the VirTaK repository includes an additional script called build_virtak_db.py, which enables the construction of an updated database based on more recent VMR releases (Fig. S1).

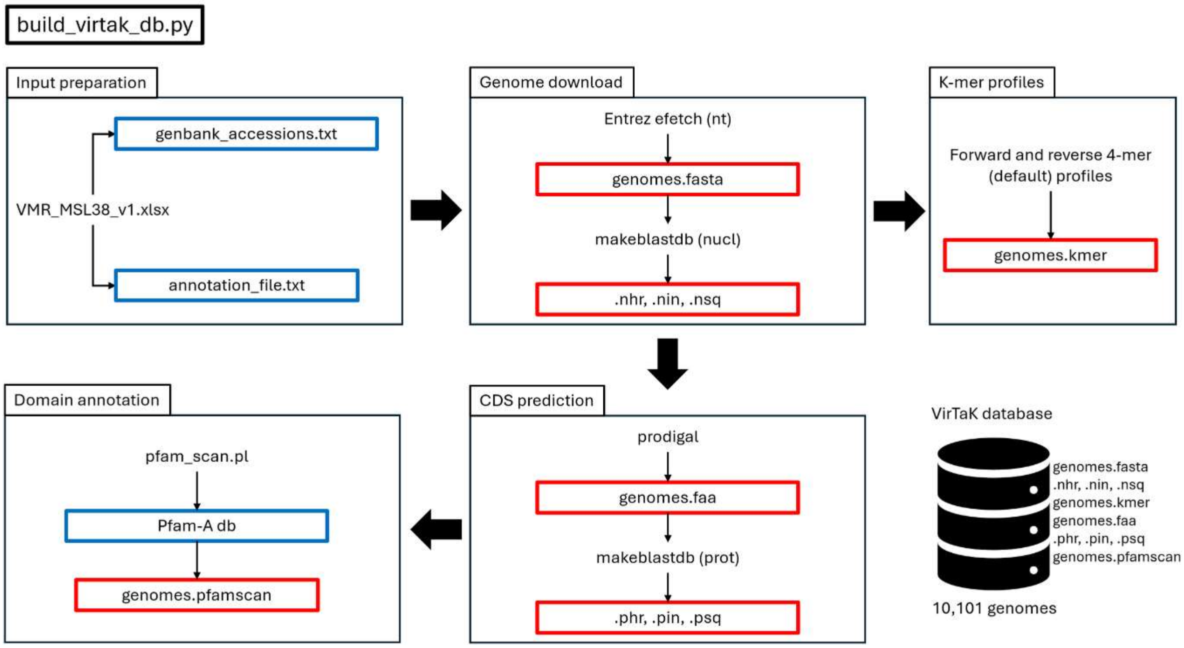

Figure S1. The pipeline for building the VirTaK database. Blue boxes indicate input files. Red boxes indicate either output files or output data.

Which viruses are related to the query genome?

Briefly, VirTaK is based on a composite score that gives increasing weight to six similarity criteria: genome length, number of CDSs, k-mer profiles, protein domain content, amino acid sequence and nucleotide sequence. In addition, the score value is associated with five different suggestions of possible degrees of relatedness: "not related", "similar genomes", "distantly related", "related" and "closely related".

More specifically, for a set of genomes in fasta format, VirTaK calculates: i) length proportion; ii) average percentage of nucleotide identity; iii) proportion of the number of protein coding sequences (CDS); iv) average percentage of amino acid identity normalised by the maximum number of matches; v) Bray-Curtis similarity of protein domain profiles; and vi) Bray-Curtis similarity of 4-mer profiles. Each value is multiplied by a linearly increasing weight such that the sum of the products gives a score ranging from 0 to 1. The highest weight is assigned to nucleotide identity (0.2857), followed by amino acid identity (0.2381), protein domain profile similarity (0.1905), k-mer profile similarity (0.1429), CDS proportion (0.0952), and genome length proportion (0.0476) (Fig. S2).

Weights were assigned according to assumptions about whether the similarities were meaningful in terms of homology. For example, highly significant nucleotide matches (Blastn evalue cutoff of 1e-20) are expected between closely related genomes, whereas significant amino acid matches (Blastp evalue cutoff of 1e-10), as well as similar protein domain profiles may occur between more distantly related genomes. Similar k-mer profiles, CDS counts, and genome lengths may be found between closely related genomes, but are more likely to occur by chance rather than common ancestry.

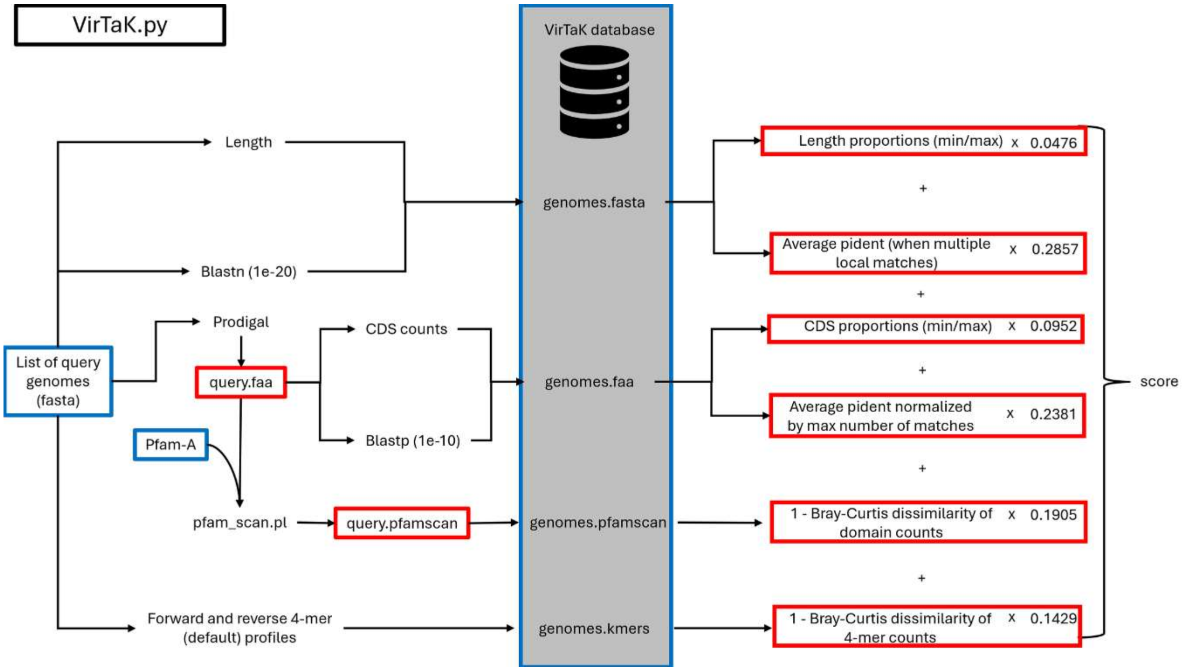
VirTaK.py
VirTaK database
Length
Blastn (1e-20)
Prodigal
List of query genomes (fasta)
query.faa
CDS counts
Blastp (1e-10)
Pfam-A
pfam_scan.pl
query.pfamscan
Forward and reverse 4-mer (default) profiles
genomes.fasta
genomes.faa
genomes.pfamscan
genomes.kmers
Length proportions (min/max) x 0.0476
+
Average pident (when multiple local matches) x 0.2857
+
CDS proportions (min/max) x 0.0952
+
Average pident normalized by max number of matches x 0.2381
+
1 - Bray-Curtis dissimilarity of domain counts x 0.1905
+
1 - Bray-Curtis dissimilarity of 4-mer counts x 0.1429
score

Figure S2. Diagram depicting the VirTaK pipeline. Blue boxes indicate input files. Red boxes indicate either output files or output data.

After running VirTaK, the result file contains a list of the top N similar genomes for each query mVC, with a suggestion of the degree of relatedness with cutoffs based on the score value. If two genomes have at least the same length and the same number of CDS, we will get a score equal to or greater than 0.1429, suggesting with certainty that we have "similar genomes", although not necessarily related. Pairs of genomes with lower scores should be considered as "not related". If a pair of genomes has at least the same length, the same number of CDS and the same k-mer profiles, we get a score equal to or greater than 0.2858, where at best, without homology information, we can suggest that they are "distantly related" genomes. Furthermore, if the pair also has at least the same set of protein domains, we get a score equal to or greater than 0.4762, suggesting that the genomes are "related". Finally, pairs with scores equal to or greater than 0.7143 are proposed as "closely related" genomes, which must have a high percentage of amino acid and nucleotide identities at that point.

To evaluate the accuracy of VirTaK, we selected one virus for each of the 107 DNA (88 dsDNA and 19 ssDNA) and 83 RNA virus families (190 virus families in total) represented by at least two members in the VirTaK database. For each virus, VirTaK was run with the -n 2 option to obtain the two most similar genomes in the database (as the first match is always a self-match, the second most similar genome was used for evaluation). To evaluate accuracy, we used two approaches, one that directly assigns the genus, family, order, class, phylum, kingdom, or domain of the most similar virus, and another that evaluates the accuracy of the suggested degrees of relatedness (closely related, related, distantly related, similar genomes, and not related).

When directly assigning the genus, family, order, class, phylum, kingdom or domain of the most similar virus, we obtained an overall precision of 64.21% (59.81% for DNA and

69.88% for RNA viruses) at the genus level (Fig. S3). However, this precision increased rapidly at the family level (90% overall, 88.79% for DNA and 91.57% for RNA viruses), and was greater than 90% in all cases from order onwards. False positives (the most similar virus was actually not related) remained below 2% in all cases.

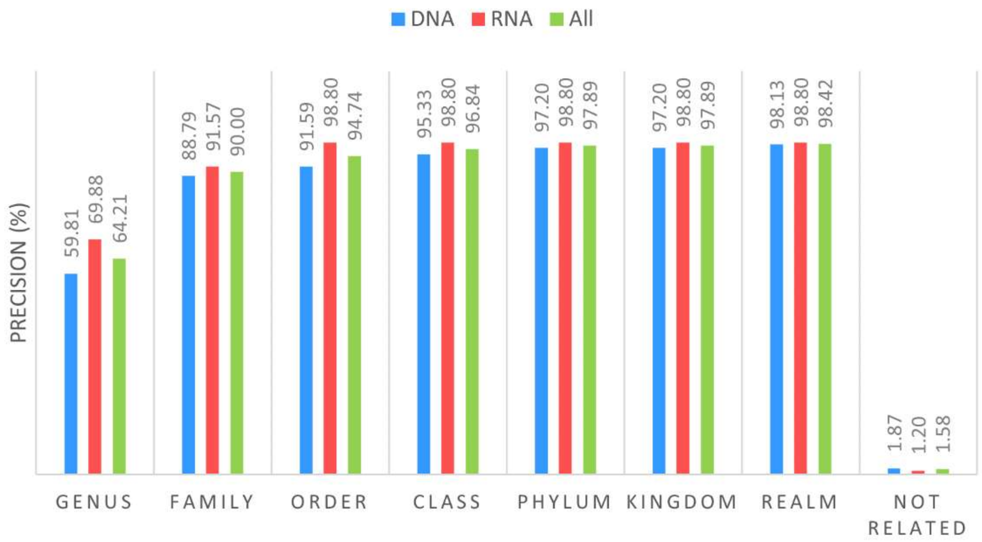

Figure S3. Precision as percentage of viruses whose most similar genome had the same taxonomic rank (genus, family, order, class, phylum, kingdom or real) or was not related.

Although directly assigning the taxonomic rank of the most similar virus showed a somewhat high precision, such a greedy approach may be less reliable than using the degree of relatedness suggestions as a guide to decide the taxonomy of the newly assembled virus. When the most similar virus was suggested to be "closely related", the precision reached 87.87% at the genus level (Fig. S4). At higher taxonomic levels the accuracy reached 100%. When the most similar virus was predicted to be "related", we observed a precision of 94.73% at the family level, which increased at higher taxonomic levels. The "distantly related" predictions had a precision of 80% at the order level and 90% at the class level, with precisions of over 90% at higher taxonomic levels. "Similar genomes" had an overall low precision of only 33.33% at the order, class, phylum and kingdom level, and 66.66% at the kingdom level. Finally, while "not related" predictions were 100% correct, there were 33.33% of "similar genomes" and 3.33% of "distantly related genomes" that were actually unrelated (false positives).

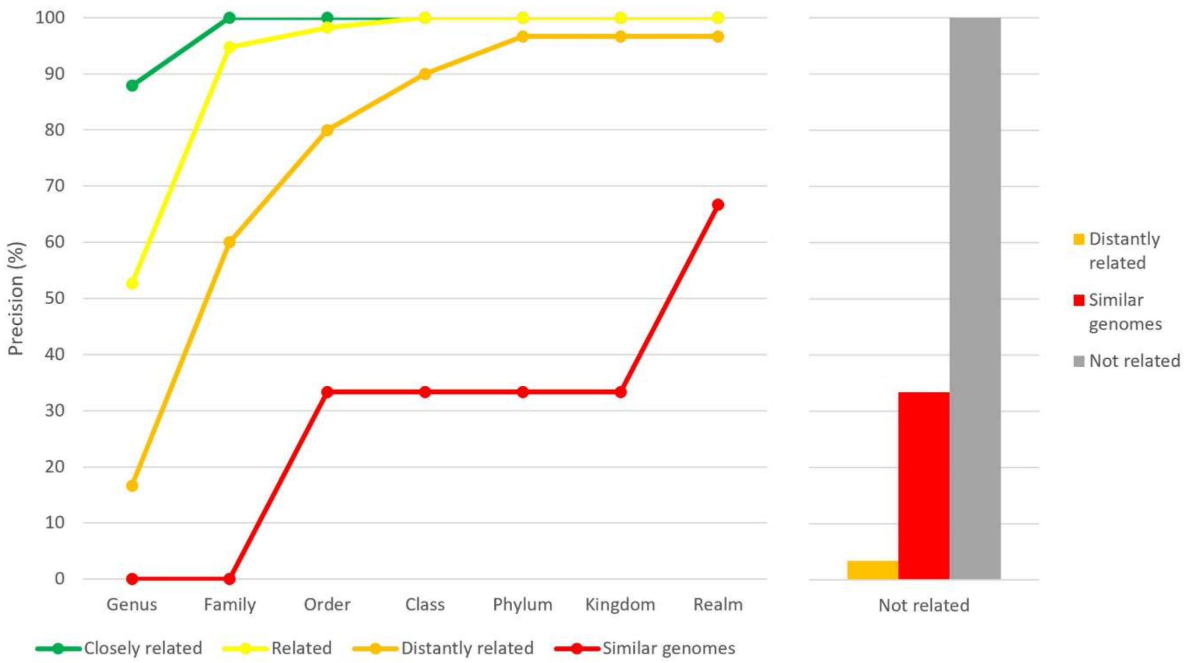

Figure S4. Precision as percentage of predictions (closely related, related, distantly related, similar genomes) that had the same taxonomic rank (genus, family, order, class, phylum, kingdom, and realm) between query mVCs and their most similar genomes (left). Percentage of distantly related, similar genomes and not related predictions that were actually not related (right).

The evaluations indicate that "closely related" predictions are suitable for assigning genus-level classifications (87.87% precision), "related" predictions for family-level classifications (94.73% precision), and "distantly related" predictions for order (80% precision) or class-level (90% precision) classifications. It should be noted that although "distantly related" predictions may include viruses with some shared domains, they may also include similar viruses solely on the basis of genome length, number of CDS and k-mer profile similarity, which are homology-free criteria. Consequently, there was a false positive rate of 3.33%.

Using protein folds and structural motifs may be a more robust approach for detecting distant relationships through deep homology signals (Simmonds et al. 2023). It is also important to note that "related" predictions are predominantly made for viruses with similar domain content, while "closely related" predictions are primarily made for viruses with high amino acid level similarities. However, some "related" predictions also include viruses with amino acid level similarities, while some "closely related" predictions also include viruses with nucleotide level similarities.

It is noteworthy that the above distributions are largely consistent with the current taxonomy of viruses (Simmonds et al. 2017), and with a recently proposed scheme, which is likely to become the standard in the near future, that suggests similar approaches and methodologies to use at different taxonomic ranks (Simmonds et al. 2023).

How does VirTaK compare to other popular virus classification tools?

We compared the performance of VirTaK with that of vConTACT2 (Jang et al. 2019), which classifies viral genomes based on shared protein clusters (PCs). To do so, we used the dataset of 107 DNA viruses, described in the previous section, each representing a different viral family, and vConTACT2 v0.11.3 was executed with default parameters. Of the 107 viruses, 38 were catalogued as singletons (similarity score < 1, meaning no significant similarity to any virus), leaving 69 that were assigned to the categories outlier (sub-family or family level relationship to similar virus), overlap (belong to two or more clusters) or clustered (genus level relationships). However, 15 viruses, which corresponded mostly to nucleo-cytoplasmic large DNA viruses (NCLDV) were clustered among themselves and with no other virus in the database, which was to be expected given that vConTACT2 focuses on classifying prokaryotic viruses or phages (Jang et al. 2019). Of the 54 remaining viruses, only 48 were used to compare VirTaK with vConTACT2 at the genus, family and class levels, as six viruses showed similarities with viruses with undetermined taxonomies. In the case of 23 viruses the match with the highest score was identified as self-match, for which the second match with the highest score was taken as the most similar virus.

Both algorithms performed well (precision > 90%) at the family and class level. However, VirTaK had a better precision at the genus level (62.5% compared to 50%) (Fig. S5).

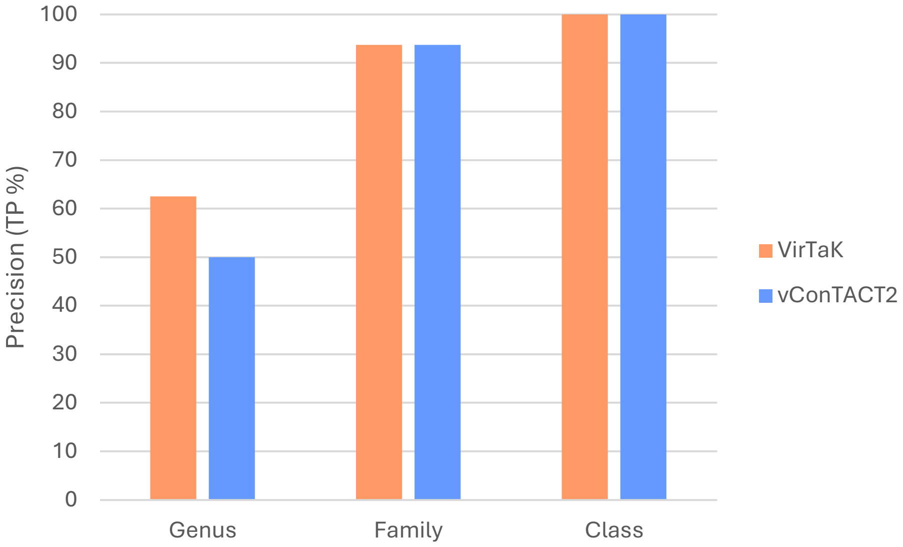

Figure S5. Comparison of VirTaK and vConTACT2 precision for assigning genus, family and class level taxonomic ranks to 48 DNA viruses.

This brief comparison highlights different advantages of using VirTaK for virus taxonomic classification. First, VirTaK performs in a highly competitive manner, which makes it a viable alternative to other known algorithms. Second, VirTaK can be used to classify any type of virus present in its database (based on the ICTV's VMR), including RNA viruses and NCLDV, which means it is not limited to classifying DNA prokaryotic viruses as vConTACT2 is. Finally, VirTaK appears to be a highly sensitive classification tool that managed to assign a classification to all viruses in the original dataset of 107 DNA viruses with arguably high precision (59.81%; Fig. S3). It is worth noting that vConTACT2 not only focuses on prokaryotic DNA viruses but also seems to be more stringent when filtering significant similarities. VirTaK's overall sensitivity may be associated with the implementation of more sensitive homology-based similarity metrics, such as the similarity in Pfam domain profiles (Fig. S2) (as opposed to vConTACT2 which defines PCs with Blastp (Altschul et al. 1997)) and the inclusion of non-homology-based similarity metrics (similarity in k-mer profiles, CDS counts, and genome lengths).

One advantage of using vConTACT2 is its network representation of genome similarities. However, as described in the following section, we take advantage of the domain annotations already created by VirTaK (with pfam_scan) (Fig. S2) and jackhammer-based clustering of proteins without domain annotation, to create a distance matrix based on domain and protein clusters profiles which can be represented as a dendrogram or easily processed to create edge files that can be loaded to Cytoscape (Shannon et al. 2003).

Which genes support the relationship between the query genome and their predicted relatives?

To analyse the evolutionary context of the relationships between query genomes and the putative related viruses, we developed another Python script called PanPhylo (Pangenomic and Phylogenomic analysis), which is also available in the VirTaK repository.

The principal advantage of PanPhylo over other methods based on pairwise distance and protein content (Gorbalenya and Lauber 2022) is that it utilises files generated by the VirTaK pipeline to construct profiles of domain-annotated and unannotated proteins, thereby streamlining time-consuming steps (Fig. S6). Furthermore, the analysis is intended to be restricted to groups of viruses related to the query mVCs, as suggested by VirTaK. This avoids the construction of large and cumbersome polyphyletic phylogenomic trees that include unrelated viruses. This is more consistent with the current notion about the multiple independent origins of viruses (Koonin et al. 2020).

Briefly, PanPhylo takes as input a list of query genomes and optionally a list of strings to search in the VirTaK database (e.g., *Alphacoronavirus*, *Betacoronavirus*, etc.). If no strings are provided, PanPhylo will perform the analysis for the query genomes only. On the one hand, it extracts the domain profiles from the pfam_scan (Finn et al. 2010) files generated during the execution of the VirTaK pipeline. On the other hand, it tries to identify proteins without domain annotations, in order to create clusters of unannotated proteins with jackhmmer (Johnson et al. 2010). To save time and avoid redundancy, it only runs jackhmmer on proteins that do not already exist in any of the resulting clusters. To further remove redundancy, it identifies overlapping clusters and takes the largest cluster as the representative. After removing redundancy, it generates unannotated cluster profiles for query mVCs and related genomes (if a list of strings to search in the VirTaK database is provided). Finally, if unannotated proteins are found, it merges the unannotated cluster profiles with the protein domain profiles to create a gene content table that can be displayed as a heatmap and used to perform pangenomic analyses. Such a table is also used to generate a Bray-Curtis distance matrix, which in turn is used to compute a distance tree using the neighbor joining algorithm (Fig. S6).

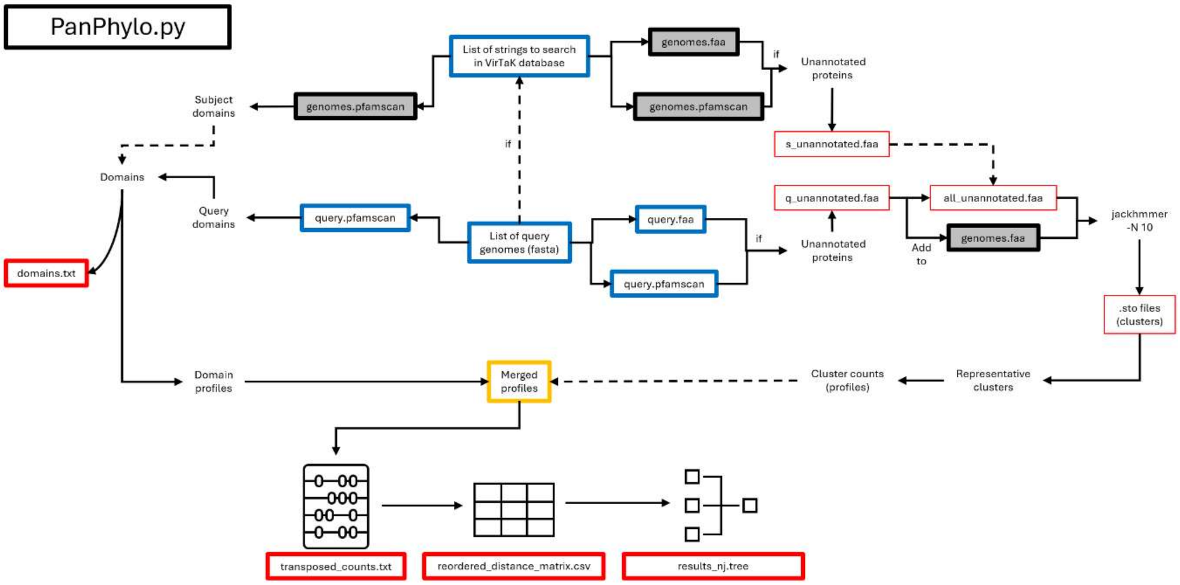

Figure S6. Diagram depicting the PanPhylo pipeline. Blue boxes indicate input files. Grey boxes indicate VirTaK database files. Red boxes indicate output files. Dashed lines indicate processes that occur only if previous conditions are met.

PanPhylo was tested with viral families represented by at least three members (151 families) in the VirTaK database. These are 79 DNA (63 dsDNA and 16 ssDNA) and 72 RNA viral families. For each analysed family, we estimated the average distance between its members to ascertain whether viral families could be delineated by a common average pairwise distance, or whether such a threshold would be specific to each viral family. Overall, the average distance was 0.42 for all 151 families. However, the average distance was highly variable (Fig. S7-9), ranging from 0.022 in the RNA virus family *Togaviridae* (n=46) (Fig. S7) to 0.946 in the archaea-infecting DNA virus family *Verdandiviridae* (n=3). While the average distance for all RNA virus families was 0.337 (Fig. S8), the average distance for all DNA virus families was 0.497.

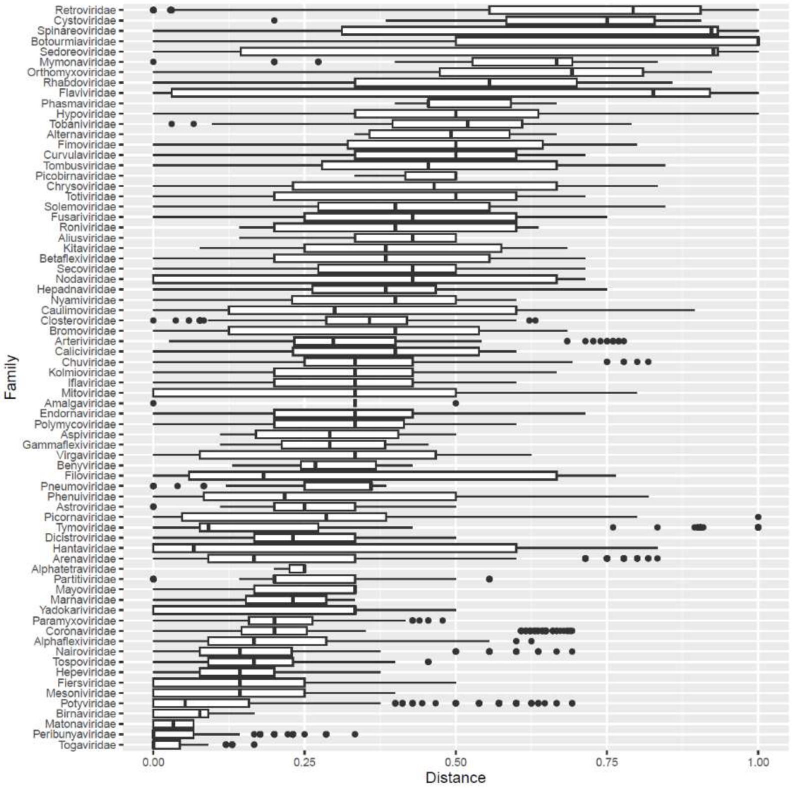

Figure S7. Average distance between viruses of different RNA virus families as calculated by PanPhylo.

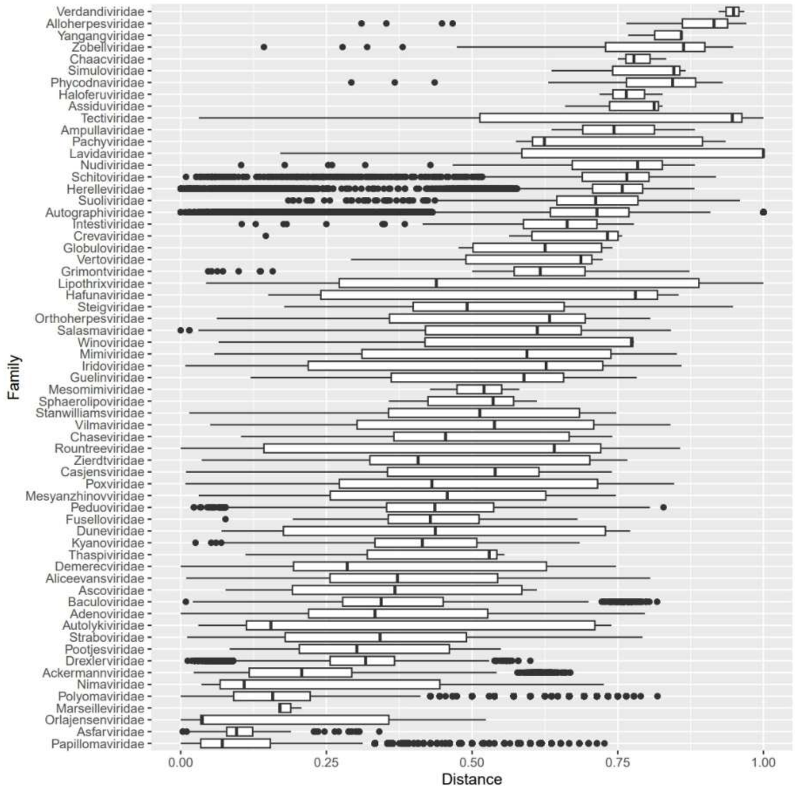

Figure S8. Average distance between viruses of different dsDNA virus families as calculated by PanPhylo.

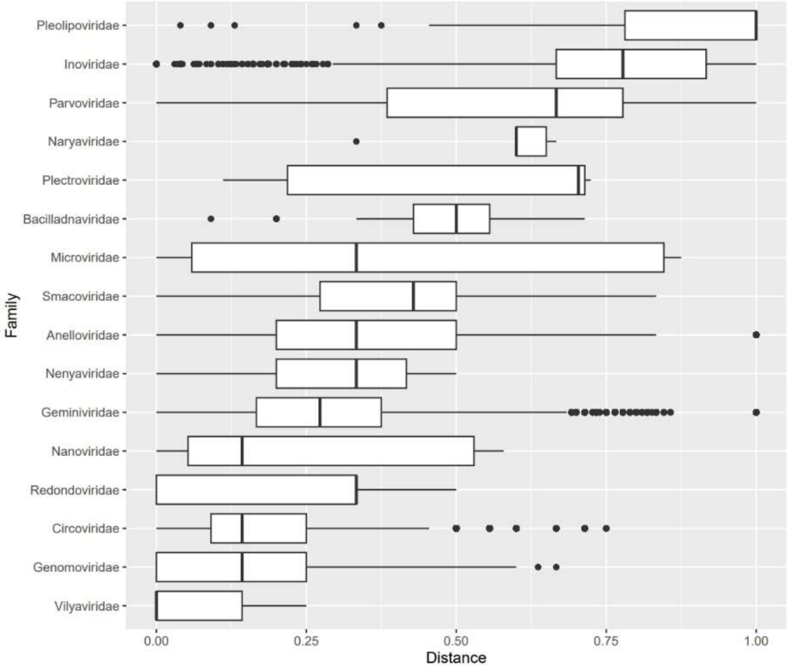

Figure S9. Average distance between viruses of different ssDNA virus families as calculated by PanPhylo.

One could posit that an overall cutoff (0.42) could serve as a general guideline for determining whether a virus should be considered a member of a given family. However, given the high variability of the average distance between families, it would be advisable to carefully evaluate each specific case. Similar scenarios have been observed for other methodologies based on pairwise distance and protein content (Aiewsakun et al. 2018).

It is noteworthy that several families, most notably *Pleolipoviridae*, *Botourmiaviridae* and *Lavidaviridae*, exhibited distances of 1, indicating a total lack of shared domains. This indicates that the criteria used to classify viruses are not entirely consistent, allowing the formation of incoherent genetic groups (Simmonds and Aiewsakun, 2018). Consequently, it is clearly necessary to re-evaluate some families to ensure that all members comply with some minimal metrics of genetic relatedness, as would be expected for a monophyletic group, so that the virus taxonomy reflects the evolutionary history of viruses (Simmonds et al. 2023).

Conclusions

VirTaK, is accessible and resource-efficient tool that can rapidly assess the similarity between a query viral genome and an extensive database of officially recognised viruses. Performance evaluations demonstrated that the integration of diverse approaches into an unified metric facilitates the establishment of confident associations across diverse taxonomic ranks, particularly from the genus to the class level. An additional tool, PanPhylo, which was designed to perform phylogenetic and pangenomic analyses based on protein content, demonstrated a considerable degree of heterogeneity in the average distance between viruses across different viral families. This suggests that there are still some incoherent genetic groups that require re-evaluation in order to ensure that the virus taxonomy accurately reflects the evolutionary history of viruses.

References

Aiewsakun, P. et al. (2018) Evaluation of the genomic diversity of viruses infecting bacteria, archaea and eukaryotes using a common bioinformatic platform: steps towards a unified taxonomy. *J. Gen. Virol.*, 99: 1331–1343.

Altschul, S. F. et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, *25*, 3389–3402.

Finn, R. D. et al. (2010) The Pfam protein families database. *Nucleic Acids Research*, 38: D211–22.

Gorbalenya, A. E. (2018) Increasing the number of available ranks in virus taxonomy from five to ten and adopting the Baltimore classes as taxa at the basal rank. *Arch. Virol.*, 163: 2933–2936.

Gorbalenya, A. E., and Lauber, C. (2022) Bioinformatics of virus taxonomy: foundations and tools for developing sequence-based hierarchical classification. *Current Opinion in Virology*, 52: 48–56.

Jang, H. B. et al. (2019). Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat Biotechnol*, *37*, 632–639.

Johnson, L. S., Eddy, S. R., and Portugaly E. (2010) Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics*, 11: 431.

Koonin, E. V. et al. (2020) Global Organization and Proposed Megataxonomy of the Virus World. *Microbiol. Mol. Biol. Rev.*, 84: e00061–19.

Shannon, P. et al. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res*., *13*, 2498–2504/

Siddell, S. G. et al. (2019) Additional changes to taxonomy ratified in a special vote by the International Committee on Taxonomy of Viruses (October 2018). *Arch. Virol.*, 164: 943–946.

Simmonds, P. et al. (2017) Virus taxonomy in the age of metagenomics. *Nat. Rev. Microbiol.*, 15: 161–168.

Simmonds, P., and Aiewsakun, P. (2018) Virus classification – where do you draw the line? *Arch. Virol.*, 163: 2037–2046.

Simmonds, P., et al. (2023) Four principles to establish a universal virus taxonomy. *PLoS Biol.*, 21: e3001922.