Microbiome

SOFTWARE

Open Access

Bidirectional subsethood of shared marker profiles enables accurate virus classification

Christopher Riccardi[1,2†], Yuqiu Wang[1†], Shibu Yooseph[3] and Fengzhu Sun[1*]

Abstract

Background Due to the impact of viral metagenomic sequencing, the official virus taxonomy is updated several times a year, with labels being renamed even substantially across releases. While this helps reveal newer aspects on the classification of viruses, existing bioinformatic methods for classification struggle to stay in sync with this ever-improving resource.

Results We developed a new computer program, named Virgo, that is able to correctly predict virus families from metagenomic data with an F1 score above 0.9 using a novel viral sequence similarity metric proposed in this work. Moreover, it ensures compatibility with any version of the official taxonomy of viruses.

Conclusions Virgo is designed to easily incorporate newer releases of the official taxonomy, thus representing a valuable resource in the virology community while raising awareness to develop computational methods that evolve alongside manually curated resources.

Keywords Classification and taxonomy, Virology, Software, Metagenomics, Bidirectional subsethood, ICTV

Introduction

Virus detection and classification have benefited tremendously from viral metagenomics and from the computational methods developed around it [1, 2]. The International Committee on Taxonomy of Viruses (ICTV [3]) has recently begun accepting complete molecular sequence analysis and annotation as a sufficient requirement for entries inclusion and potential ratification, resulting in a taxonomic architecture that is a better reflection of the polyphyletic nature of viral evolution [4]. The ICTV has also created a repository of metadata and lineage information for all recognized virus species, now organized in a 15-rank classification hierarchy that mirrors the Linnaean taxonomy system [5]. The number of species listed in the Virus Metadata Resource (VMR) has more than doubled in the last five years (since release MSL35, Fig. 2a) and with 3468 new entries added between the last two versions of the master species list (Fig. 2a). Contingent on the growth of this valuable resource (which also encompasses satellite nucleic acids, viriforms, and viroids, highly relevant in plant biology [6]), bioinformatic methods for virus prediction and viral genome analysis have shifted their source of classification, from the Baltimore classes or NCBI lineages towards ICTV-ratified taxa [7, 8]. As the ICTV expands, it inevitably becomes more complex, but this expansion ensures greater precision and thoroughness, albeit at the cost of introducing new taxa and renaming existing ones in subsequent releases [4, 9, 10]. These updates can have a profound effect on computational prediction methods, especially when trained on specific versions of the ICTV, thus challenging the labeling process and

†Christopher Riccardi and Yuqiu Wang contributed equally to this work.

*Correspondence:
Fengzhu Sun
fsun@usc.edu
[1] Quantitative and Computational Biology Department, University of Southern California, 1050 Childs Way, Los Angeles 90089, CA, USA
[2] Department of Biology, University of Florence, Via Madonna del Piano 6, Sesto Fiorentino I-50019, Italy
[3] Kravis Department of Integrated Sciences, Claremont McKenna College, Claremont 91711, CA, USA

calling for awareness when using such software. This is a "good" problem, that has been circumvented by tools such as vConTACT2 [11], which does not directly output a taxonomic lineage, but rather makes inference of the taxonomic context (via RefSeq) a query sequence is more likely associated with, using a network-based approach on protein clusters.

Nonetheless, other computational frameworks rely exclusively on past releases of the ICTV and their predicted labels are somewhat "crystallized" to a specific release, thus making it laborious now, and in the future, to not only trace back the correct viral prediction that might have changed over time, but also to make a fair comparison when benchmarking new software. Other approaches embed taxonomic information directly within the features used in the training, then use majority rule-based scoring systems or membership ratios to infer the taxonomy. Even this approach makes synchronization to an ever-improving resource challenging. For example, PhaGCN2 [12] (a recent virus classification program that combines convolutional neural networks) relies on taxonomic labels from a pre-trained version of the ICTV. Similarly, TIGTOG [13], which uses random forests to classify giant viruses using DNA and amino acid sequence features, does not allow to update training set labels. VPF-Class [14] relies on a set of viral protein families that were pre-annotated and assigned to specific taxonomic levels using purity thresholds to enhance the classification of viral genomes. A similar strategy was adopted by geNomad [15], where the encoded genes in a query sequence are aligned to a set of 227,897 markers which may contain taxonomic information, and a single taxonomic assignment is emitted according to a weighted scheme based on the bitscore of the taxonomically informed matched marker profiles. While these strategies have demonstrated exceptional power in identifying viruses from metagenomic data, they are not exactly compatible with an ever-refining official taxonomy (as of 2025).

Here we present a straightforward and effective program for virus classification, Virgo, that infers the ICTV-ratified taxonomic lineage of a given set of query sequences. Our approach finds similarity between the query and a database of ICTV viruses using a *bidirectional subsethood* metric, which is used to score the way two sequences independently align to a set of virus specific markers. Genomic sequences are modeled using unordered collections of matched marker profiles, with markers coming from a recently published, large environmental metagenomic survey [15].

The more two such representations resemble each other in terms of markers distributions, the higher the score (closer to 1), or closer to 0 otherwise. The taxonomic lineage is then drawn from the ICTV entry with the highest similarity. A formal presentation of the algorithm is provided in the "Materials and methods" section.

We designed a computational framework that uses the ICTV-ratified lineage labels, and is compatible with different releases of the virus metadata resource, thus allowing the program to operate on fresh updates. Unlike the other tools that explicitly embed taxonomic information within the features, we let the features aggregate fluidly and autonomously in sync with the ICTV version used as reference, thus ensuring reproducibility and usability.

We benchmarked our tool with state-of-the-art virus-detection and classification programs, that range from the specific detection of giant viruses and prokaryotic viruses, to omni-comprehensive virus prediction tools. During benchmarking we realized that the evaluation of the classification results is often hampered by the variability of labels across releases of the reference taxonomy. Therefore, we addressed the necessity to develop software that can digest newer releases of an ever-growing taxonomy by making publicly available the source code for an ICTV sequence dump program. This program, which we named ICTVdump, connects to any version of the Virus Metadata Resource release and downloads sequences, metadata and taxonomic lineage associated to every sequence listed.

We report several metrics to show Virgo's performance on both metagenomic and reference viruses in relation to other software and investigate the potential reasons behind the fraction of incorrectly classified viruses. Aware of the fact that the programs used for benchmarking rely on taxonomic labels tied to previous versions of the ICTV, we accessed the past releases and ran Virgo on the same version as those programs. This was possible using ICTVdump, and it ensured a fair and consistent comparison to existing programs for virus classification. Overall, Virgo exhibits consistent and high accuracy in resolving the family level of viruses, even when those are fragmented or incomplete, and it is among the fastest in terms of speed, compared to the tested tools. Virgo is written in python and it requires a database which we distribute together with the source code at https://github.com/christopher-riccardi/Virgo.

Results

Virgo overview and workflow

Virgo's workflow consists of three main steps shown in Fig. 1: (1) downloading sequences and metadata from the ICTV online resource, (2) creating the Virgo database, and (3) performing virus classification. In Step 1, the user retrieves viral sequences and metadata for a specific ICTV release. To facilitate this, we developed a

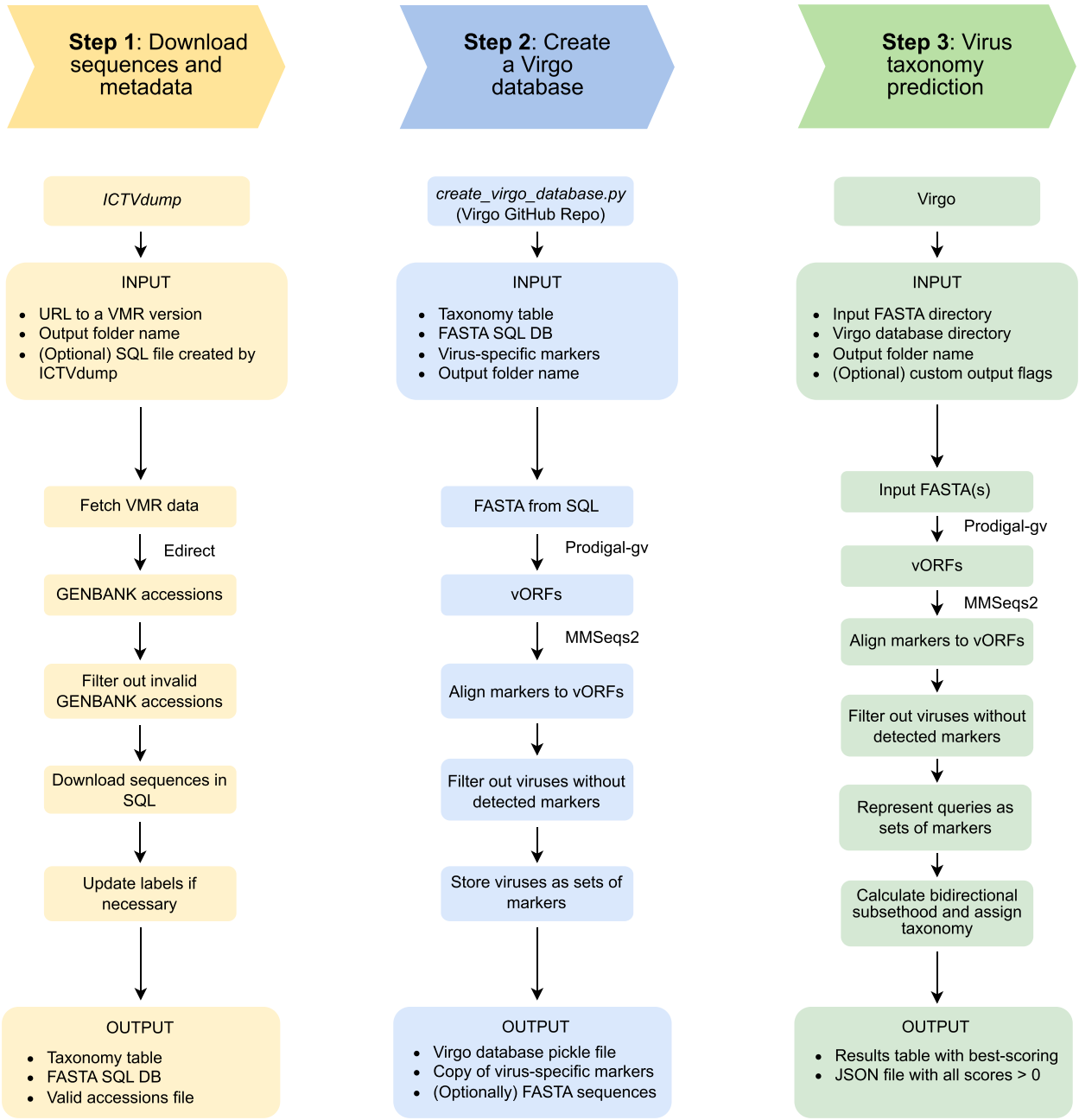


Fig. 1 Workflow overview. Virgo operates on a database tailored to a specific release of the ICTV. In Step 1, the user provides a URL to a specific release of the VMR, and ICTVdump collects the necessary information for every virus. In Step 2, the user runs the script create_virgo_database.py to package this information into a computer-readable format; it also writes to a folder with the necessary files in order to run the actual classification, which happens in Step 3 (i.e., running Virgo)

companion tool, ICTVdump, which downloads nucleotide sequences in FASTA format along with the corresponding metadata. In Step 2, the user can run the script create_virgo_database.py which will infer the viral open reading frames (vORFs), align them to the virus-specific markers released with geNomad and then package the computational representation of each virus together with their ICTV-ratified labels, all in a fully automated fashion. In Step 3, taxonomic classification is performed. For each query, Virgo computes the bidirectional subsethood score against all reference viruses in the Virgo database. The query is then assigned the taxonomic lineage of the reference virus with the highest score. This modular workflow allows Virgo to remain compatible with any

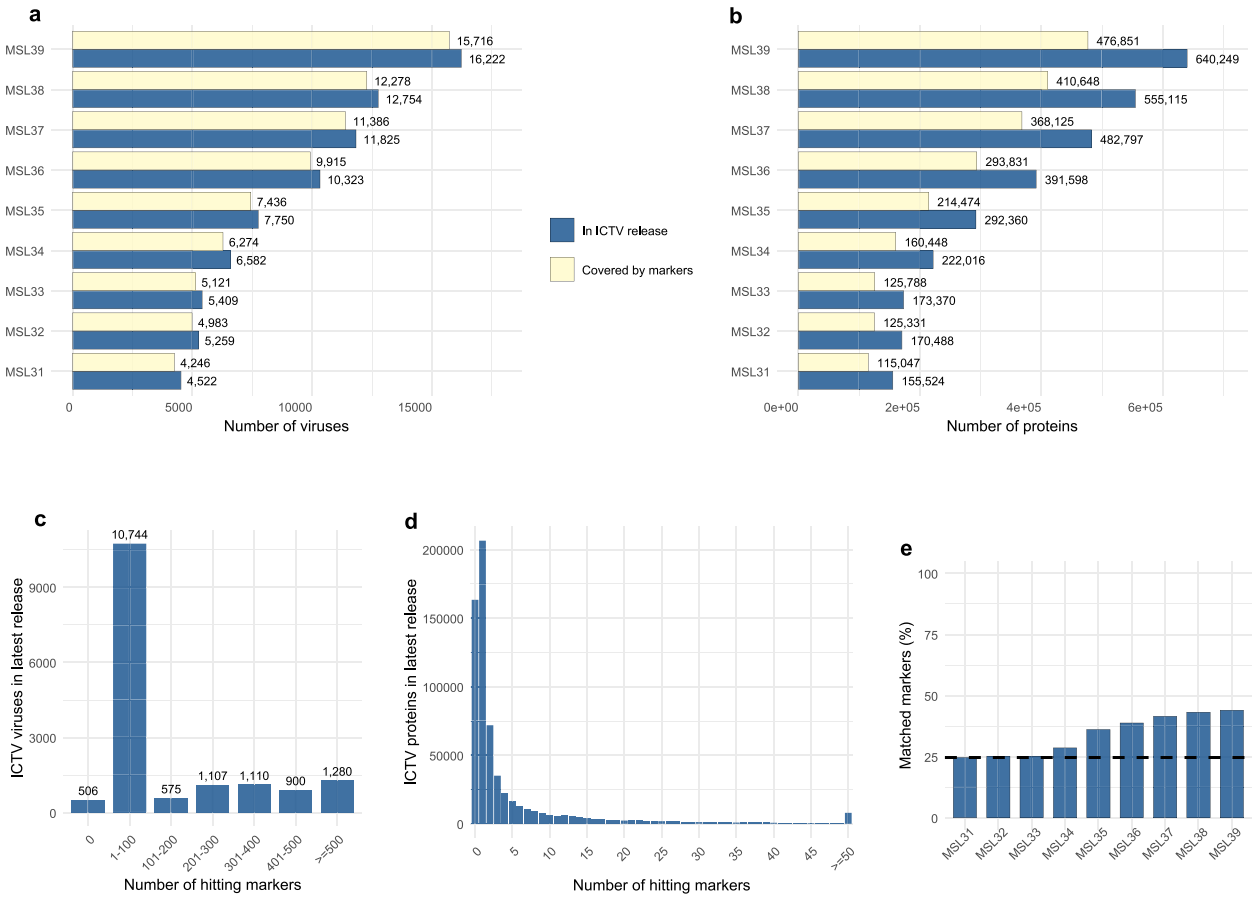


Fig. 2 **a** ICTV sequences and virus-specific markers information. The volume of species in the VMR over different versions (blue), and those covered by the virus-specific markers data set (yellow). MSL39 indicates the release that is currently available (#39 version 4). **b** Similar representation as subfigure **a**, having the number of proteins instead of the number of viruses. **c** Zoom on the latest release: number of viruses that are covered by increasing intervals of protein markers. The number of viruses is specified on top of each bar, and the number of hitting markers are binned in groups of 100, except for the first (zero markers) and last (anything matching 500 or more markers). Five hundred six viruses are excluded from our analysis since they match exactly zero markers. **d** Similar visualization as panel **c**, where the proteins are shown instead of the number of viruses. The horizontal axis reports individual values and then ICTV proteins hitting 50 or more markers are grouped together (163,398 proteins hit exactly 0 markers, while the last grouping counts 8106 proteins). **e** Percentage of all virus-specific marker profiles aligned to viral sequences in the ICTV over the releases. The number almost doubles between the first (dashed line) and current release (1.78 fold increase), but it stays below 50%, indicating that a larger fraction of sequences still needs to make entry in the ICTV

version of ICTV taxonomies and provides an interpretable framework for virus classification.

The ICTV sequences cover roughly 44% of the marker profiles database

Virgo implements a system that attributes the taxonomic lineage by maximizing a coverage score between two sequences, calculated on the degree of subsethood between unordered collections of sets of matched *marker profiles*. More details and examples are provided in the "Materials and methods" section. The default markers deployed with the software are an extract of the virus-specific sequences collected by Camargo et al. [15] for geNomad's marker-based classification, and they represent an invaluable data resource derived from many and diverse biological and environmental contexts [16, 17]. The extract is composed of 161,862 markers with high specificity for viruses. The selection criteria are explained in the "Materials and methods" section. During a preliminary exploratory analysis we mapped these marker profiles to the sequences in the ICTV to quantify the fraction of markers that are currently represented in the official virus taxonomy.

Interestingly, at the time of writing (2025), the ICTV captures up to 71,279 distinct markers, representing 44.05% of the total. Prompted by this observation we aligned the virus-specific markers to every historical record of the ICTV, selecting the very first

available (MSL31) and then every last version of each release ($n = 9$) and noticed that the number of matched markers increases together with the number of added sequences across the ICTV releases (Fig. 2). As the number of genomes almost quadruples (Fig. 2a), the amount of matched markers nearly doubles (Fig. 2e). For the MSL39 release consisting of 16,222 virus sequences, about 3% of the viruses (506) do not contain any markers, 66 % (10,744) contain 1–100 markers, yet 8 % contain over 500 markers (Fig. 2c). We performed similar analyses for the viral proteins and the results are shown in Fig. 2b,d. This trend is a direct indicator of the dynamic nature of the body of sequences in the ICTV that grows not just in size, but also in terms of genetic variability, and highlights the importance of ensuring compatibility between classification software and virus metadata resources.

Virgo accurately classifies phages from human gut metaviromes

The known viral sequence clusters (kVSCs, $n = 2232$) dataset consists of viral sequences derived from several hundreds of highly enriched human gut metavirome samples, assembled by Zolfo et al. [18]. In terms of composition, the kVSCs dataset contains exclusively bacteria-infecting viruses, mostly *Caudoviricetes* (98.83%), and it includes 15 single-stranded DNA bacteriophages (*Malgrandaviricetes*) and 11 single-stranded DNA filamentous bacteriophages (*Faserviricetes*). As pointed out by others [11], many viruses belonging to the class *Caudoviricetes* are unclassified at the order and family levels. Because of this, we evaluated performance using two criteria: (i) a stringent criterion, which deems a prediction correct only if it accurately identifies the family-level taxon, provided that the true label includes a family-level classification, and (ii) a loose criterion, which considers a prediction correct if it correctly identifies the taxon at the order or class level only when the true label does not specify a family-level classification. Virgo, geNomad, PhaGCN2, VPF-Class and vConTACT2 were tested. Note that VPF-Class, vConTACT2, geNomad and PhaGCN2 are based on ICTV release MSL 33, 36, 37, and 39, respectively. For fair comparison, we ran Virgo based on the corresponding releases when comparing with the other tools.

The benchmarking results for both criteria are visualized in Fig. 3a, using the notation *a* and *b*, for the stringent and loose criterion, respectively. Detailed metrics results are reported in Supplementary Table S1. Among the tested tools, Virgo performed the best on both criteria with a perfect classification based on the stringent, and F1 score above 0.99 for the loose criterion. The loose criterion included many more sequences, hence the difference. Very close to Virgo, vConTACT2 reached an F1 score above 0.99 on the loose and stringent evaluation criteria. Also vConTACT2 was the program that assigned the least amount of viruses to a cluster (hence classify). This is due to the fact that the unclassified *Caudoviricetes* were previously assigned to the *Caudovirales* order and *Siphoviridae* family, both of which were abolished after vConTACT2's publication.

geNomad and PhaGCN2 reached an F1 score consistently above 0.9 (0.918 for both criteria for the former, 0.914 and 0.968 for the latter). VPF-Class performed identically on the stringent and loose criteria, with an F1 score of 0.99. Virgo achieved a perfect classification using both criteria when ran on the same ICTV version as VPF-Class. Moreover, VPF-Class classified more sequences than Virgo in the loose evaluation of the kVSCs (2207 sequences against 2144).

Benchmarking results on the kVSCs dataset indicate superior performance by Virgo in terms of classification (Fig. 3a–d) on gut metagenomic sequences compared to other tools, highlighting its potential applicability in clinical settings. Additional benchmarking results at the genus level are provided in Supplementary Fig. S1a–b.

Virgo generalizes on unseen data with performance comparable to state-of-the-art

To further evaluate Virgo's classification performance across a broader range of viral taxonomic affiliations, we selected $n = 860$ exemplar viral genomes through stratified sampling at the family taxonomic level, ensuring that their full taxonomic lineage is defined across releases (refer to "Materials and methods").

These viruses represent a heterogeneous dataset (hereafter, ICTV dataset) that encompasses 485 distinct genera and 192 families, whose taxonomic labels can be correctly pinpointed across releases for fair benchmarking.

Virgo's performance was compared to that of geNomad, PhaGCN2, and VPF-Class.

Virgo achieved a higher F1 score (0.982) compared to geNomad (0.768), PhaGCN2 (0.974 and 0.909, respectively) and VPF-Class (0.964 and 0.959, respectively), as well as every other metric (Supplementary Table S1). A visual representation of performance evaluation and sample sizes are shown in Fig. 3e-g.

As with every other benchmarking dataset used in this study, Virgo was run using the `-with_replacement` option which skips the database search for queries identical to a database entry, forcing the next best-scoring virus to be used as taxonomic reference (see "Materials and methods"). This option was particularly relevant in testing the ICTV dataset since it consisted in a Leave-One-Out type of study. The same approach could not

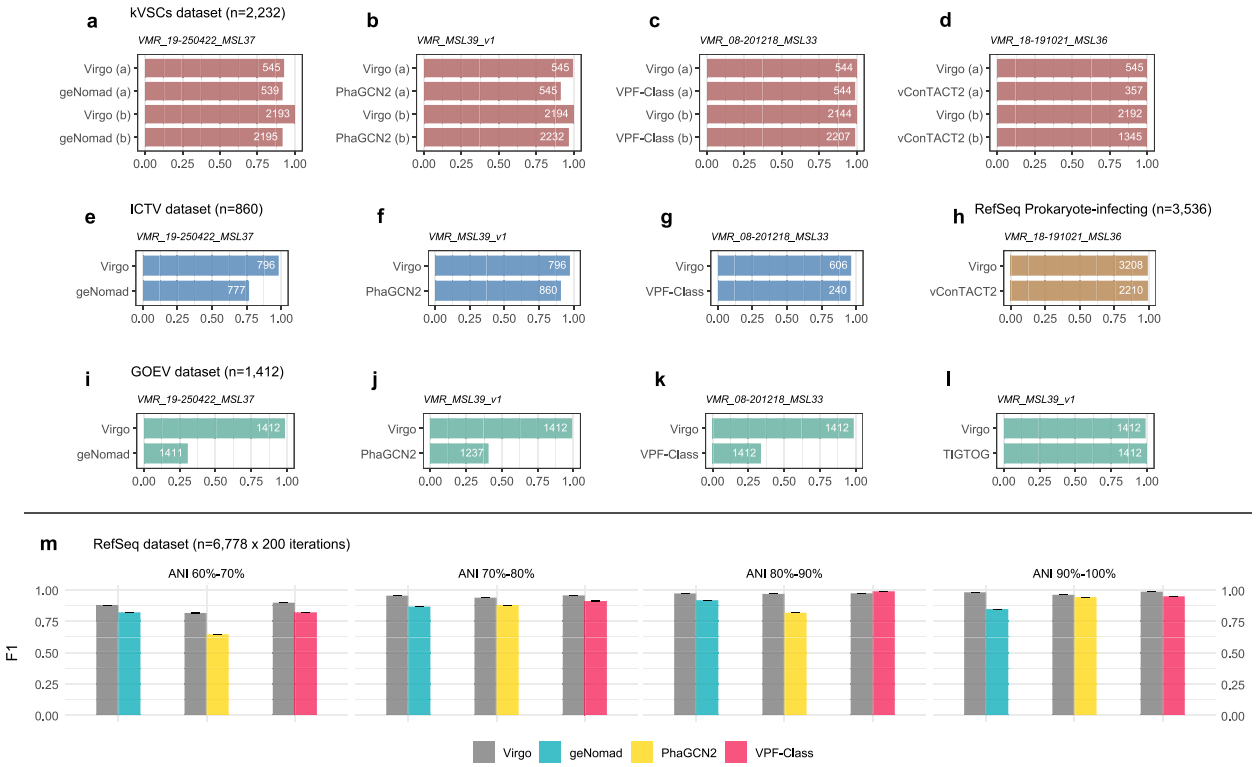


Fig. 3 Benchmarking results across datasets. Panels **a** through **d** show software performance (F1 score) for the kVSCs (human gut metagenomic viruses). We compared Virgo with geNomad, PhaGCN2, VPF-Class, vConTACT2, and TIGTOG based on ICTV releases MSL37, 39, 33, 36, and 39, respectively. To evaluate performance for this dataset we applied two different criteria: a stringent criterion (a) and a loose criterion (b). The stringent criterion only considers a prediction correct if it accurately identifies the true family-level taxon, provided that the family-level classification is available in the true label. Instead, the loose criterion allows for a correct prediction if the tool correctly identifies the taxon at the order or class level when the true label does not specify a family-level classification; the family is compared otherwise. Panels **e–g** show the results for the selected ICTV viruses of broad taxonomic coverage. Panel **h** shows the benchmarking results comparing Virgo and vConTACT2 on the prokaryote-infecting fraction of the RefSeq dataset. Panels **i–l** show the results on the GOEV ocean metagenomic dataset. Numbers inside bars indicate the number of viruses classified by each program. Panel **m** shows the results for the RefSeq dataset random iterations across various ANI levels between training and testing. All numerical values are reported in Supplementary Table 1

be performed for the remaining programs since it is not possible to access and modify their training data.

We then expanded our benchmarking dataset by incorporating a larger body of viruses present in publicly-available databases, and accessed the entire NCBI Virus resource (the RefSeq dataset). Given the predominant presence of *Coronaviridae* in this dataset, we randomly sampled 6,778 sequences (several times, see "Materials and methods" for details) and compared Virgo's performance to that of geNomad, PhaGCN2 and VPF-Class using the corresponding ICTV releases.

We also considered the maximum Average Nucleotide Identity (ANI) that exists between these viruses and the ICTV sequences that Virgo uses as reference database. High maximum ANI values indicate high similarity between the query and references, and progressively lower ANI values represent a more challenging testing dataset. Computing identity below 60% was not possible, therefore the RefSeq data were split into four distinct ANI groups, in increasing levels of 10%. Results for this part of the analysis are shown in Fig. 3m, and the detailed information relative to performance, central tendency and dispersion are reported in Supplementary Table 1.

Virgo generalizes better than current software at all ANI intervals, with VPF-Class performing better than Virgo on sequences in the ANI slot 80–90%, where the former achieved an average F1 score of 0.996 ± 0.0 versus 0.974 ± 0.0. It is worth mentioning that the ICTV release used for this specific comparison, MSL33, came out in 2018 when a substantially lower number of sequences were present in the database, therefore limiting the number of viruses classifiable for both programs (113.52 ± 0.784 for Virgo and 110.65 ± 0.771 for VPF-Class).

For reference PhaGCN2, which classifies all input viruses and contains pre-trained labels for a much newer release, produced a prediction for an average of 754.42 ± 1.365, and showed an average F1 score of 0.819 ± 0.001.

On the harder-to-classify ANI group (60–70% maximum ANI), Virgo achieved an average F1 score of 0.899 ± 0.001 compared to VPF-Class (0.833 ± 0.001), an average F1 score of 0.817 ± 0.001 compared to PhaGCN2 (0.647 ± 0.001) and an average F1 score of 0.874 ± 0.001 compared to geNomad (0.822 ± 0.001).

Taken together, these results indicate that the bidirectional subsethood metric applied to a database-lookup system may have robust applications when lower similarity exists between query and reference.

Among the RefSeq viruses were 3536 prokaryote-infecting viruses which we used for testing Virgo's performance against vConTACT2 (Fig. 3h). The latter achieved an overall higher performance on most metrics compared to the former, albeit with a very similar classification ability (vConTACT2 F1 score 0.992, Virgo F1 score 0.991). Virgo classified 3208 viruses and vConTACT2 2210. No additional considerations (e.g., in terms of ANI) were made for this subset of the data given the elevated performance achieved by both programs.

We further identified 400 viruses that belong to 40 new families in Version MSL39_v4 but are connected at higher hierarchical levels to version MSL_37. We were able to compare Virgo and geNomad on this rather substantial sample size, and try to understand what happens when novel families are presented to both tools. The rationale behind looking at these two specific versions lies in the fact that, as stated, geNomad performs classification using labels from MSL_37, but the current taxonomy is at version MSL_39. This indicates that several new viruses were added and thus it was possible to compare both Virgo and geNomad using a reference built on MSL_37, to predict labels in MSL_39. However, neither of the two programs are able to discard the input viruses under new families. Presumably, the sequence signals are strong for viral components, and the ability to identify viruses that are incomplete or fragmented trades off with a lower capability of rejecting viruses that belong to a new taxonomic lineage. Nonetheless, the F1 metric for being able to predict the class taxonomic level (higher level) remains high for both (Virgo: 0.936, geNomad: 0.9) but lower for the order level (Virgo: 0.854, geNomad: 0.751. Data not shown for visualization). Additional benchmarking results at the genus level are provided in Supplementary Fig. S1c–e.

Virgo can classify giant viruses from metagenomic datasets

The Global Ocean Eukaryotic Viral database (GOEV) [19] is a resource of MAGs enriched in large and giant marine viruses belonging to the phylum *Nucleocytoviricota*. We extracted 1412 sequences with well-defined order taxonomy from the original publication's metadata to compare our program's results against geNomad, PhaGCN2, VPF-Class and TIGTOG [13].

TIGTOG uses a machine learning approach based on protein family profiles to classify giant virus genomes at the ICTV order level. Given the high diversity and distinct signatures of protein content among different taxonomic groups within the *Nucleocytoviricota*, TIGTOG leverages the unique composition of giant virus orthologous groups within each lineage for classification. To avoid reliance on a fixed set of marker genes, it applies a random forest algorithm to model taxonomic classification at these levels, using features such as the presence of ortholog groups and G+C content, with pre-established taxonomic labels guiding the model. Benchmarking results for six distinct evaluation metrics are reported in Supplementary Table S1; a visual representation of the differences in terms of F1 score is depicted in Fig. 3i–l.

Interestingly, three out of four tested programs yield accuracies below 0.5. More specifically, Virgo ranks second in terms of accuracy (0.984), with TIGTOG achieving an almost perfect classification accuracy of 0.995. Differences in terms of F1 score are less pronounced, with Virgo scoring 0.991 and TIGTOG 0.998.

However, geNomad showed an F1 score of 0.307, the lowest across all comparisons, despite Virgo and geNomad using the same marker dataset. After further investigation of the erroneously attributed orders, we concluded that geNomad did not generalize well the taxonomic branch of the *Imitervirales*. This is likely due to a poor representation of the reference order *Imitervirales* in release #37 ($n = 2$) compared to the later releases (e.g., $n = 22$ in release #39). The GOEV benchmarking dataset is composed of 32 *Asfuvirales*, 54 *Chitovirales*, 226 *Algavirales* and 1100 *Imitervirales*, thus probably representing a considerable challenge for geNomad. We note that Virgo was also run using a reference database with the same taxonomic labels as geNoma's; however, Virgo was able to correctly classify this taxonomic branch better than any other program. More specifically, Virgo misclassified 1.63% of the input data. Most errors involved confusing *Algavirales* with *Imitervirales* ($n = 16$) and failing to recognize one third of the *Asfuvirales* sequences ($n = 7$). A possible explanation for this is in the low degree of overlapping marker patterns between the ICTV database and the incorrectly assigned viruses.

With reference to the meta-analysis reported in this study, there is strong evidence indicating that the misclassified giant viruses tend to have lower scores compared to the correctly classified ones, despite the overall average score being already low at 0.472 (S.D. 0.19). Therefore, the low fraction of incorrectly classified viruses is presumably due to the inherent biology of these

large viruses. Their exceptional genomic complexity, coupled with dynamic gene exchanges between these viruses and their hosts [20], likely contributed to the errors, especially given the sporadic representation in the virus-specific marker dataset.

Virgo is robust to incomplete data

We investigated Virgo's robustness against genome fragmentation and mutations by performing an artificial reduction of 1000 genomes randomly sampled from release #39 over a progression of 9 random fragmentation percentages, with 10 replicates each (Fig. 4). The genomes were allowed to reduce in size until the required fragmentation or a minimum size of 1 kbp were obtained, whichever condition was met first. In this instance, we used Virgo masking the query and reference genome to ensure replacement in a Leave-One-Out strategy. The average size of the initial 1000 ICTV genomes was 46,234.3 bp (S.D. 70,730.94) and it was reduced to an average of 4799.19 bp (S.D 6977.82) during the last iteration. The reduction in genome completeness was confirmed by CheckV [21] (Fig. 4a). Random mutations at rate of 0.01 were also introduced at the nucleotide level in order to add an additional layer of difficulty; note that the sequencing error rates reported on Illumina platforms is in the order of 0.001 [22]. The results show a general decrease in the average F1 score (vertical axis) as the reduction gets stronger (horizontal axis), with a more rapid decay between a reduction of 80% and 90%. However, a consistent performance indicated by an F1 score > 0.9 is always observable across all fragmentation levels, further corroborating Virgo's ability to correctly classify viral sequences as long as viral open reading frames are still detectable. Figure 4b shows the benchmarking results of Virgo versus two alternate classification approaches—one that predicts a query's virus family using Mash [23], and another using nucleotide BLAST [24]. Briefly, these programs classify a query sequence by selecting the virus family with the highest ANI (Mash) or highest bitscore (BLAST) between the query and the ICTV reference viruses, allowing for replacement just as performed with Virgo. Virgo consistently outperforms the alignment-free and alignment-based approaches. More details for these comparisons are provided in the "Materials and methods" section.

Tie score differs across orders

As noted above, running Virgo with replacement did not achieve a 100% accurate classification on the ICTV benchmarking dataset, indicating that a fraction of viruses remained classified incorrectly. Therefore, we also conducted a detailed analysis of the correctly versus incorrectly classified viruses ($n = 30$), focusing specifically on those viruses for which Virgo calculated a tie score less than 1.

We suspected that, in sporadic cases, some viral families were "interfering" with correct taxonomic attribution by exhibiting very high bidirectional subsethood scores but lower tie scores, meaning that more than one family had equal chance of being chosen as the predicted. We focused on a group of viruses ($n = 7$) that were consistently attributed to a different family within the same order (*Mononegavirales*) and computed the bidirectional subsethood scores through pairwise comparisons among all viruses in this order. For visualization, we selected the incorrectly assigned viruses along with other randomly

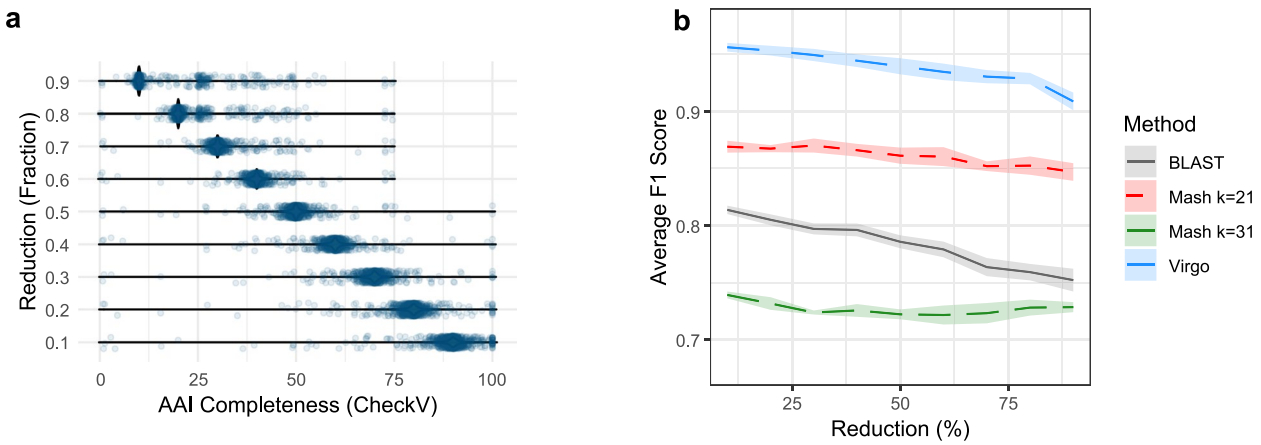


Fig. 4 Sequence fragmentation study on 1000 random ICTV viruses. The genome reduction is confirmed through Average Amino acid Identity (AAI) completeness, confirmed by CheckV (panel **a**). Panel **b** shows the average F1 score (vertical axis) with respect to the percentage reduction (horizontal axis) for Virgo and two alternate classification approaches that rely on alignment-free (Mash) and alignment-based (BLAST) scoring engines

sampled members from the same family as both the tested and predicted viruses. Similarly, we randomly selected an order from the group of correctly classified viruses (*Crassvirales*) and computed the same metric for comparison. The results are shown in Fig. 5. The seven incorrectly classified viruses, highlighted in yellow in panel Fig. 5a, exhibit elevated scores when compared to viruses from different families within the same order (indicated by a more intense color), often matching or exceeding the scores seen when compared to members of their own family. In contrast, the bidirectional subsethood patterns in a correctly predicted order, as shown in Fig. 5b, display a more regular pattern, with members of the same family showing higher similarity to each other compared to members of other families.

Meta-analysis of the effect of several factors on classification performance

We synthesized the association between predictions (correctly and incorrectly classified) and three factors, namely AAI completeness, bidirectional subsethood score and tie score, using a meta-analysis.

For this analysis, in order to capture the true sources of error we ran Virgo one more time, this time without the replacement mode that was used in the other performance evaluation steps (see sections above). At this stage, the ICTV benchmarking dataset was excluded, since running Virgo without replacement produced a perfect classification. We considered the kVSCs, GOEV and one random iteration of the RefSeq viruses, then ran Virgo keeping track of the correct and incorrect classifications. The resulting data table is available as Supplementary Table S2. The three studies were meta-analyzed comparing two groups, the correct and incorrect classifications, with respect to each of the three quantitative dependent variables whose effect sizes were expressed in terms of raw mean differences (MD). Details for the meta-analysis are described in the "Materials and methods" section, and the resulting forest plots are illustrated in Fig. 6d–f.

The meta-analysis result for AAI completeness is an estimated MD of 10.97% (CI 95% [7.37, 13.93]) indicating that, on average, correctly classified viruses are nearly 11% more complete compared to the incorrectly classified portion. This is reflected by all three studies lying on the right side of the central vertical solid line centered at 0 in Fig. 6d, thus showing a positive mean difference, although the kVSCs study showed an MD of 14.48 (CI 95% [−13.66, 42.62]) with a wide confidence interval

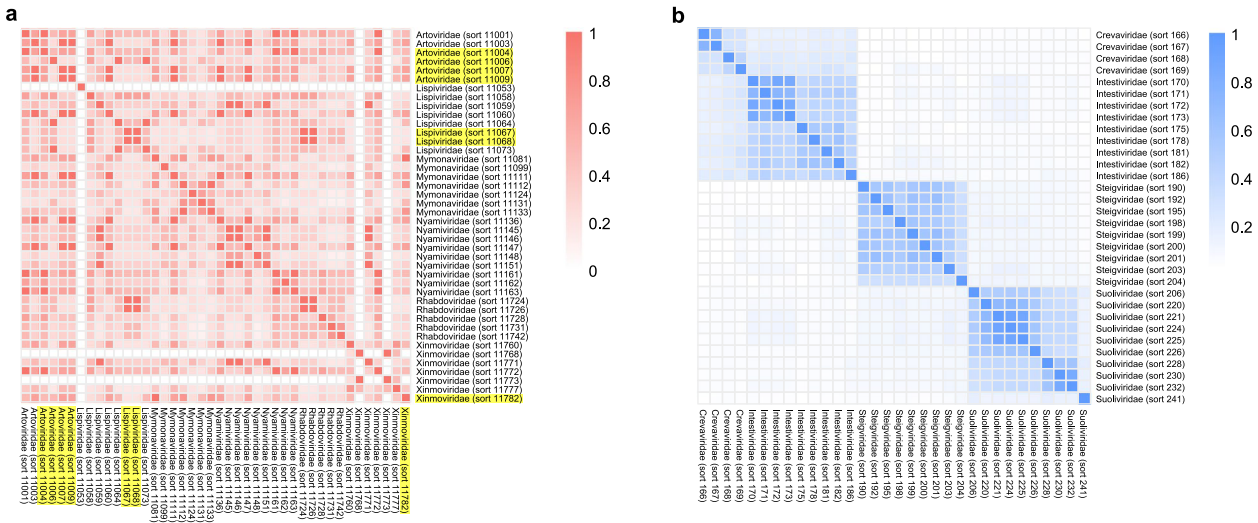
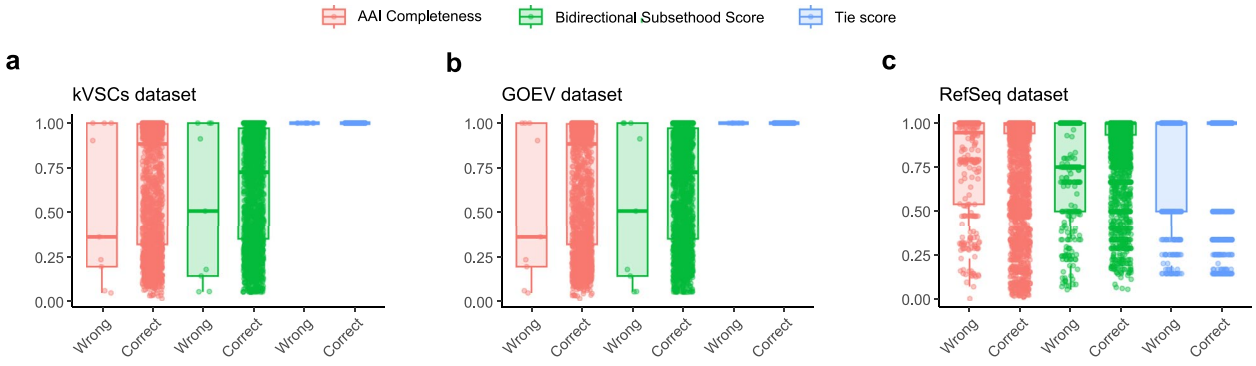


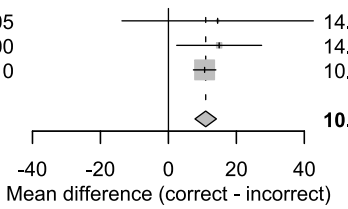
Fig. 5 Heatmaps displaying bidirectional subsethood scores among viruses in two distinct orders. Panel **a** shows the bidirectional subsethood scores for viruses belonging to six different families of the order *Mononegavirales*. Viruses highlighted in the column and row labels are those that were erroneously predicted as belonging to a different family within the same order. The heatmap also includes a random selection of viruses from the same family as both the tested and the predicted viruses. The score patterns for these viruses are generally irregular, with viruses from different families scoring comparably in terms of subsethood, indicating a similarity level comparable to that observed within the same family cluster. Panel **b** displays the bidirectional subsethood scores for a random selection of viruses from different families within the order *Crassvirales*, with all viruses correctly predicted to their respective families. The patterns in panel **b** are more regular, with viruses within the same family sharing markers more similarly than they do with viruses from other family clusters. Notably, there is a faint hue, in panel **b**, corresponding to a light similarity between *Intestiviridae* and *Suoliviridae*. However, this mild similarity does not affect the prediction accuracy because other members of each respective family exhibit substantially higher scores. The bidirectional subsethood metric, ranging from 0 (no shared markers) to 1 (all markers shared identically), is used to quantify the similarity between viruses. The family of each virus is reported as well as the virus number identifier as of release #39 version 1



d

AAI completeness analysis

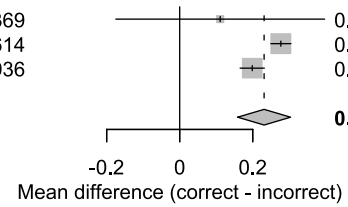
Study	Experimental Total	Mean	SD	Control Total	Mean	SD	Mean Difference	MD	95%-CI	Weight
kVSCs	2185	67.84	34.1193	9	53.36	43.0105		14.48	[-13.66; 42.62]	1.3%
GOEV	1388	60.61	26.4640	23	45.68	30.3400		14.93	[2.45; 27.40]	6.4%
RefSeq	5508	88.16	24.0940	310	77.51	28.9110		10.65	[7.37; 13.93]	92.4%
								10.97	**[7.82; 14.12]**	**100.0%**



e

Bidirectional subsethood score analysis

Study	Experimental Total	Mean	SD	Control Total	Mean	SD	Mean Difference	MD	95%-CI	Weight
kVSCs	2185	0.65	0.3173	9	0.54	0.4369		0.11	[-0.17; 0.40]	5.7%
GOEV	1388	0.47	0.1993	23	0.20	0.0614		0.28	[0.25; 0.30]	47.9%
RefSeq	5508	0.91	0.1806	310	0.71	0.2936		0.20	[0.17; 0.23]	46.4%
								0.23	**[0.16; 0.30]**	**100.0%**



f

Tie score analysis

Study	Experimental Total	Mean	SD	Control Total	Mean	SD	Mean Difference	MD	95%-CI	Weight
kVSCs	2185	1.00	0.0000	9	1.00	0.0000		0.00		0.0%
GOEV	1388	1.00	0.0000	23	1.00	0.0000		0.00		0.0%
RefSeq	5508	0.97	0.1437	310	0.71	0.3310		0.26	[0.22; 0.30]	100.0%
								0.26	**[0.22; 0.30]**	**100.0%**

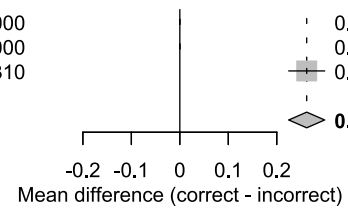


Fig. 6 Meta-analysis of three studies examining the association between accuracy of classification and three factors. Panels **a** through **c** show individual data points relative to correct and incorrect predictions across the three studies, with respect to AAI completeness (**a**), bidirectional subsethood score (score for short in the legend, panel **b**) and tie score (**c**). Panels **d** through **f** show forest plots with the effect sizes (horizontal axes) expressed as raw mean difference. The solid vertical line centered at 0 indicates the significance threshold. The boxes are bounded by confidence interval at 95% and their size is proportional to the precision of their estimate. The diamonds in each panel represent the final estimates for random effect models

which suggests substantial uncertainty (due to the small sample size for the incorrect classifications $n = 9$). The GOEV dataset yielded an MD of 14.93 (CI 95% [2.45, 27.40]), indicating a statistically significant difference in completeness between correctly and incorrectly classified sequences. The RefSeq dataset, which had the largest sample size, exhibited a mean difference of 10.65% (CI 95% [7.37, 13.93]), demonstrating a strong and significant effect with a relatively narrow confidence interval.

Regarding the bidirectional subsethood score, the kVSCs dataset exhibited a mean difference of 0.11 (CI 95% [−0.17, 0.40]), with a wide confidence interval, which includes zero, suggests that this difference is not statistically significant. In contrast, the GOEV dataset showed a statistically significant mean difference of 0.28 (CI 95% [0.25, 0.30]), indicating that correctly classified sequences had substantially higher bidirectional subsethood scores than incorrectly classified sequences. The RefSeq study also demonstrated a significant mean difference of 0.20 (CI 95% [0.17, 0.23]), further supporting the trend that correct classifications correspond to higher scores.

The random-effects model produced an estimate of 0.23 (CI 95% [0.16, 0.30]). This suggests that, on average, correctly classified sequences had a bidirectional subsethood score that was 0.23–0.24 higher than incorrectly classified sequences. The GOEV dataset's relatively larger effect size drives the majority of the effect and causes significant heterogeneity ($I^2 = 85.4\%$, $\tau^2 = 0.0027$ and $p = 0.0011$). This is the dataset with the overall lowest AAI completeness (60.61% in the correct, Fig. 6e) which (in this case) is also responsible for the lowest average bidirectional subsethood score (0.47 in the correct group, Fig. 6d) since it is composed of sequences with the largest genomes among the tested data (they are giant viruses). As shown in a later paragraph, viruses with greater number of vORFs tend to be classified more accurately, which is consistent with seeing a dataset with such low average bidirectional subsethood score being classified with high accuracy (reported to be 0.984 in the Supplementary Table S1).

As per the tie score, interestingly the kVSCs and GOEV datasets do not contribute to the meta analysis since the standard deviation is exactly 0 between the two groups making the effect size to become undefined. The RefSeq dataset is the only contributor to the pooled effect, which indicates that the tie score tends to be 0.26 lower in the incorrect fraction of RefSeq viruses (CI 95% [0.22, 0.30], Fig. 6f). Despite not being able to fully determine the heterogeneity for the tie score metric, we are able to explain why the pooled effect derives from RefSeq only. The kVSCs and GOEV datasets are composed of sequences with generally larger viruses (prokaryote-infecting bacteria in the former, and giant viruses in the latter) compared to the more taxonomically broad dataset that is the RefSeq. Since our framework bases its computations on the number of vORFs that are present in the query and reference viruses, it is less likely for longer viruses to produce tying scores compared to shorted sequences (see later paragraphs for a more in depth explanation).

Calibration of prediction accuracy using vORF-based metrics

In addition to the meta-analysis, factors that may affect prediction accuracy were further investigated using a calibration dataset. The calibration dataset was pooled from the benchmarking datasets: ICTV, kVSCs, GOEV, and one random iteration of the 200-iteration RefSeq set as described in Data retrieval and preprocessing section. We hypothesized that the bidirectional subsethood score may oversimplify sequence relationships if the number of viral open reading frames (vORFs) in either of the pair of sequences is low. We studied the relationship between classification accuracy and varying number of vORF for the query and reference sequences in the calibration dataset. Results are shown in Fig. 7a. Classification accuracy is 0.892 when either the query or the most similar reference are composed of just one vORF. When the number of vORFs is at least 2 for both the query and the most similar reference sequence, the classification accuracy rises substantially, to 0.987.

We next looked at the classification accuracy as a function of the tie score and bidirectional subsethood score for query sequences with 1, 2, or at least 3 vORFs, respectively, as shown in Fig. 7b and c. When the query sequence just contains one vORF, the prediction accuracy is lower than 0.933 regardless of the bidirectional subsethood score value. When the query sequence contains two vORFs, the classification accuracy shows a general increasing trend and is 0.978 when the bidirectional subsethood score is above 0.6. For query sequences with at least 3 vORFs, the prediction accuracy is above 0.99 when the subsethood score is above 0.2. Figure 7b shows that the prediction accuracy is above 0.97 when tie score is 1 and is lower than 0.677 when tie score is less than 1.

While we could apply a strict filter requiring both the query and the most similar reference sequence to have at least two vORFs, this filter would discard 2280 potential predictions. Instead, we explored a refined filtering approach that leverages both bidirectional subsethood score and tie score. We found that for predictions where either the query or the most similar reference sequence has only one vORF, for a subsethood score above 0.8 and a tie score of 1, the accuracy reaches 0.977. In contrast, classifications performed without applying both criteria

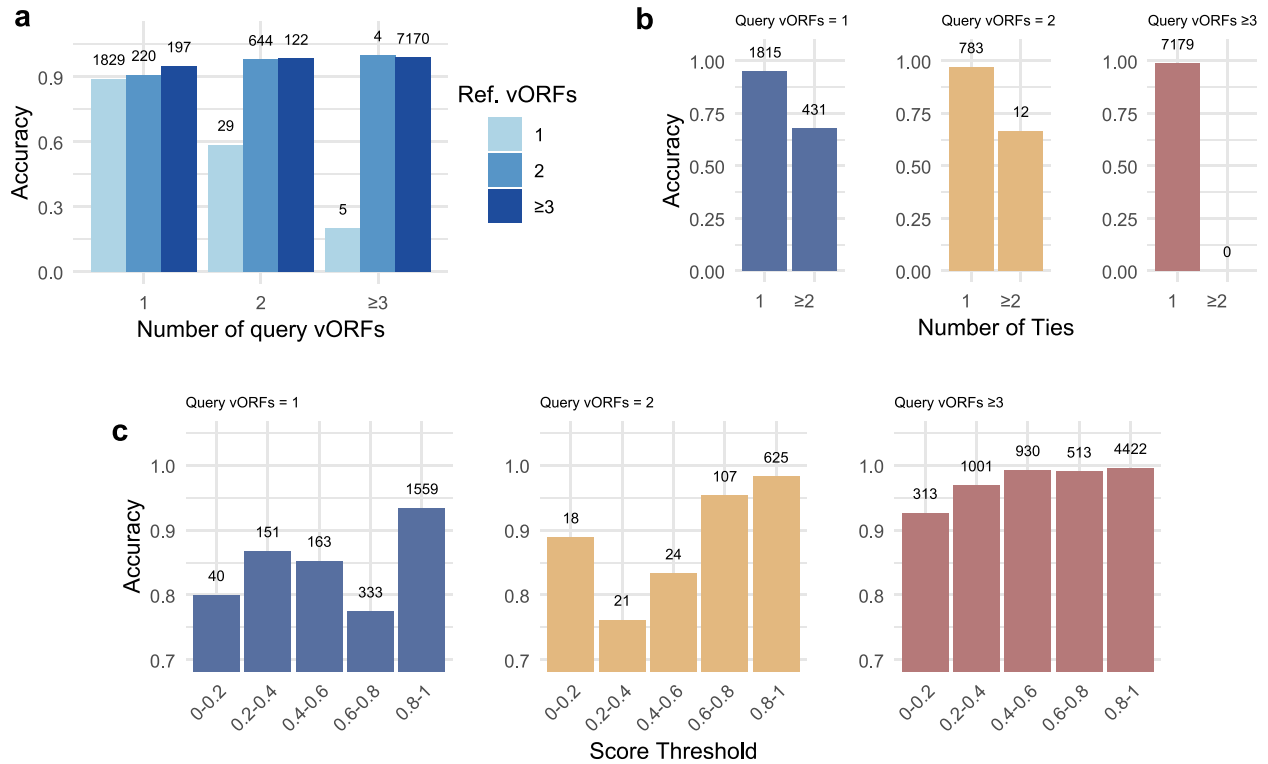


Fig. 7 Prediction accuracy across vORF counts, tie scores, and subsethood score thresholds. Panel **a** shows the relationship between prediction accuracy and the numbers of vORFs for the query and most similar reference sequences. Panel **b** shows the accuracy as a function of the number of additional tying families (reciprocal of tie score) for the same query vORF categories. The numbers on top of each bar indicate the sample sizes in each category. Panel **c** shows the accuracy trends based on subsethood score thresholds and the number of query vORFs. The number on the top of each bar is the number of sequences in each category

(bidirectional subsethood score > 0.8 and tie score = 1) show a much lower accuracy of 0.794.

In the light of these observations, we include in the results table a confidence label (0 or 1) based on the following two criteria: (i) both the query and most similar reference sequence have at least 2 vORFs or (ii) either the query or reference sequence contain just one vORF, and the bidirectional subsethood score is at least 0.8 and tie score equals 1. Using these two criteria, we discard roughly 12 % of the sequences and obtain a prediction accuracy at about 0.98.

Materials and methods

The bidirectional subsethood scoring metric

The core principle by which we obtain good classification power is reliant upon the assumption that related protein-coding sequences exhibit specificity for protein domains or families that perform similar functions. Overall, this assumption is made by most virus classification software that use information from the viral open reading frames to establish connections between viruses.

Here we perform a pairwise comparison between a query sequence and viruses in the ICTV that are representative of each family, to infer its possible taxonomic affiliation. However, we do not compare the two sequences directly. Instead, we analyze the patterns each sequence independently forms in relation to a set of virus-specific markers.

We utilize a recently published, comprehensive protein marker dataset that includes multiple sequence alignments of protein families derived from de novo-generated profiles, external profiles, and non-redundant profiles specific to viruses. This dataset is sourced from the largest collections of viral and microbial sequences obtained from single genomes and metagenomes, as well as the NCBI database. It also encompasses 25,729 protein markers across *Nucleocytoviricota*, Asgard archaea viruses, archaeal tailed viruses, and unannotated domains of polyproteins [16, 17]. These profiles are functionally annotated and used for classifying sequences, providing functional insights, and enabling taxonomic assignments. We broaden the application of these markers, without considering their detailed biological functions or the specific taxonomic categories, which are hidden and not accounted for. Briefly, to build the marker profiles, Carmago et al. first

retrieved a large number of protein sequences from various databases. These sequences were de-replicated and then clustered to form different clusters. They then performed multiple sequence alignment of sequences in each cluster to form the protein marker profiles. Detailed steps to obtain these profiles are given in the "Methods" section of their paper [15]. In this study, we only use marker profiles related to viruses.

We provide a formal definition of the bidirectional subsethood, including one example, then proceed to explain the implementation details in the following section. More examples are provided in the Supplementary materials 2. Just like a genomic DNA sequence encodes for a collection of open reading frames, we model a virus as an unordered collection of sets. Each vORF is represented by a set, and each element of the set corresponds to one virus-specific marker profile that aligns to that particular vORF. Since a vORF can match multiple marker profiles, each set may contain multiple markers, theoretically as many as there are in the database. We define a measure of similarity between two such objects, ranging between 0 and 1, with 0 indicating no shared markers and 1 indicating perfect sharing. Let A and B be two unordered collections of sets. Virgo first computes a similarity matrix S, where each element S_{ij} is the Jaccard similarity between the i-th set in A and the j-th set in B:

$$S_{ij} = J(A_i, B_j)$$

with A_i and B_j representing sets from A and B, respectively.

The bidirectional subsethood metric aggregates the similarity matrix into a single value. It is computed by first finding the best match (i.e., the maximum similarity) for each set in A against all sets in B, and vice versa. Then, the average of these best matches is taken.

$$\text{Best match for } A_i = \max_j S_{ij}$$

$$\text{Best match for } B_j = \max_i S_{ij}$$

The final bidirectional subsethood s is calculated as

$$s(A,B) = \frac{1}{|A|+|B|}\left(\sum_{i=1}^{|A|}\max_j S_{ij} + \sum_{j=1}^{|B|}\max_i S_{ij}\right) \quad (1)$$

where $|A|$ is the number of sets in A, and $|B|$ is the number of sets in B.

Example. Let us model two genomic sequences as the collections of sets of markers, A and B, containing two and three vORFs, respectively:

$$A = \{\{a,b\},\{d\}\}$$
$$B = \{\{d\},\{b,c,d,e\},\{a\}\}$$

We begin by calculating the similarity matrix S where each element $S_{i,j}$ is the Jaccard similarity between the i-th set in A and the j-th set in B.

$$S = \begin{pmatrix} S_{1,1} & S_{1,2} & S_{1,3} \\ S_{2,1} & S_{2,2} & S_{2,3} \end{pmatrix} = \begin{pmatrix} 0 & 0.2 & 0.5 \\ 1 & 0.25 & 0 \end{pmatrix}$$

The best match for every set in A versus all sets in B and each set in B against all sets in A are extracted from the similarity matrix and then averaged out, to yield

$$s(A,B) = \frac{1}{2+3}\cdot(0.5+1+1+0.25+0.5) = 0.65$$

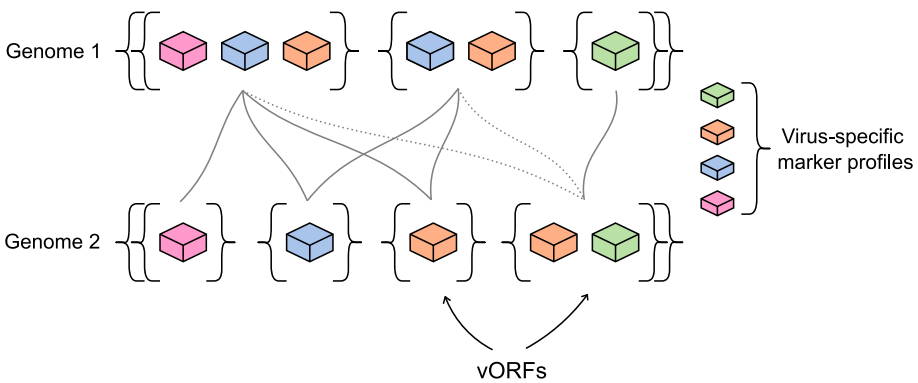


Fig. 8 Representation of two viral genomes using unordered collections of sets. The top and bottom rows show three and four distinct vORFs, each matching one or more virus-specific marker profiles (colored bricks). The gray solid lines connecting the vORFs maximize the local Jaccard coefficients between two sets, as opposed to the dashed lines, that indicate a lower similarity that will not be chosen for the calculation of the bidirectional subsethood. The final bidirectional subsethood for this specific toy configuration equals 0.452

Therefore, the bidirectional subsethood measure for the two mock genomes would be 0.65 given the marker profiles.

We illustrate an artistic representation of the collections of sets in Fig. 8.

Virgo's computational framework

Virgo transforms each query sequence into a multiset representation, compares them to a database of reference viruses, also represented as multisets (precomputed), and assigns the taxonomic lineage of the database entry with the highest similarity based on the bidirectional subsethood of shared markers.

We describe here the feature extraction procedure needed to capture the distribution of viral-specific marker genes across a set of DNA sequences. We first infer vORFs using Prodigal V2.11.0-gv [25] with enhanced specificity for virus nucleotide translation, parameters (-p: meta/anon to apply pre-calculated training to the input sequence) and align them to the viral-specific subset of marker genes using MMSeqs2 Version: 15.6f452 [26] in easy-search mode (parameters: −s 7.5 for high sensitivity, −e 1e−3 −c 0.2 –cov-mode 1 to enable the coverage of the target to be at least 20%, with an *E*-value of at most 0.001). Note that we utilize this procedure and parameters in order to adhere to the specifics used by the authors of the virus-specific marker profiles database. The alignment file listing, among others, the query (each individual vORF) and the target (virus-specific marker genes) is then converted to an unordered collection of sets of matched markers for each vORF.

Upon release of a new ICTV VMR resource, the user may use our freely distributed source code to generate a Virgo database, or download it through the GitHub page. The database serves as input to the computer program for virus identification through the command line flag `-data`. The same procedure is also embedded in Virgo to convert the input query sequences into an unordered collection representation, to then maximize the bidirectional subsethood and assign the taxonomic lineage up to the family level. When the user runs Virgo on a query virus, the program reports the taxonomic lineage as well as the bidirectional subsethood score between the query and the database entry with the highest match. Additionally, it reports the number of database entries that achieved the maximum bidirectional subsethood, along with the *tie score*, which is inversely related to the number of distinct families associated with those top-scoring viruses. The tie score is calculated as

$$t = \frac{1}{n} \quad \text{for} \quad n \in \{1, 2, 3, \ldots n_{\text{families}}\}$$

A tie score of 1 indicates that all the top-scoring database entries belong to the same family, a tie score of 0.5 or lower indicates otherwise.

By default, the current implementation employs a straightforward approach that chooses the database entry with the most similar G+C content, to prevent random selection in the event of ties. The G+C fraction information is derived from the gene calling procedure and it does not add additional overhead on the overall program execution. To facilitate Leave-One-Out studies on Virgo, we also included a flag (`-with_replacement`) that masks the database entries that identical to the query during classification. This method is meaningful when testing viruses from the ICTV downloaded using our program ICTVdump.

Virgo operates in multi-threading using python's native libraries. Moreover, Virgo can accept multiple genomes at once through the `-input` command line option. We also provide options to narrow down the results according to the user's need: the flag `-min_score` only saves viruses with bidirectional subsethood score above a certain value; `-drop_ties` omits viruses with any ties in the final results table; `-virus-by-virus` allows a more granular and verbose output, by writing a JSON file with all query-db comparisons that yield a score greater than zero. The GitHub page provides further information on usage, examples and access to a repository of pre-calculated VMR releases, eliminating the need for users to construct the database themselves. Runtime and memory consumption metrics are reported in Supplementary Table S3.

Data retrieval and preprocessing

The kVSCs dataset is an extract of the representative viral sequences included in the MetaPhlan 4.1 release [27]. To construct the kVSC dataset, Zolfo et al. selected 5651 sequences, including 699 bacteriophage genomes from RefSeq with known taxonomic labels and 4952 viral contigs identified from high-quality metaviromes. These sequences were clustered into 3944 viral sequence clusters (VSCs) using VSEARCH (version 2.14.2) at 90% sequence identity. The clustering process was performed with the following parameters: –cluster_fast –id 0.9 –strand both –maxseqlength 200,000. Clusters that contained at least one RefSeq viral genome were classified as known viral sequence clusters (kVSCs) [18]. In total, 588 VSCs contained a viral reference genome and were labeled as kVSCs. The DNA sequences were obtained from the files VSC5_rep_fnas_nr99_45k_metaphlanDB.fna.gz and VSCs_groups.csv, downloaded from Zenodo (https://zenodo.org/records/10512460) on June 28th, 2024. From the 45,872 representative sequences included in the MetaPhlan 4.1 module, we selected those

clustering with a RefSeq representative, yielding the kVSCs. The selection process further involved matching RefSeq accession names in the metadata with those in the ICTV Release #38 to ensure accurate labeling, resulting in 2232 eligible samples. Taxonomic assignments were based on the linked RefSeq accessions present in ICTV Release #38.

ICTV viral exemplar genomes were extracted from the ICTV Release #39. The data were downloaded on July 17, 2024, using ICTVdump with default parameters. A total of 1000 viral genomes, representing 119 different viral families, were randomly selected using equal probability weighting in pandas (python version 3.12.3). The genomic DNA sequences were then artificially fragmented from both ends in a random manner using the script fragment_dna.py, available at https://doi.org/10.6084/m9.figshare.28730093.v1 (under the "Reduction" subfolder). The fragmentation process involved a mutation parameter of 0.01, a lower bound of 1000 bp, and an increment of the reduction parameter by 10% each time (from 0.1 to 0.9), resulting in 9 different reduction settings. This was performed in 10 replicates across the 1000 genomes, yielding a total of 90,000 fragmented sequences. The SLURM script used to generate these fragments is also provided in the same FigShare repository, under the "Reduction" subfolder.

An additional set of $n = 860$ genomes was extracted using stratified sampling at the family level, selecting up to 5 viruses per family. This data subset (the ICTV benchmarking dataset) was assembled by comparing two ICTV releases (versions release #37 and #39) and extracting viruses that (i) shared the same GenBank accession, (ii) had a family assigned (though not necessarily with consistent naming), and (iii) included at least two representatives per family. The taxonomic assignment for these genomes follows the ICTV-ratified lineage. The complete list of all ICTV versions are kept at https://ictv.global/ as of 2025.

We further used sequences from RefSeq via the NCBI Virus resource (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/), last accessed on January 27, 2025. Initially, all 3,376,487 viral sequences were downloaded via the standard web interface. We then filtered the viruses to only include those that have a family taxon assigned. This reduced the data from 3,376,487 to 3,342,047. We further investigated the stratification of the dataset to account for family imbalance, finding that out of the 258 families, the most represented were *Coronaviridae* with over 3M records. 254/258 family labels that were attached to the RefSeq sequences matched the ICTV family labels in release VMR_MSL38_v3, therefore these were kept for further consideration (n=3,342,022). Since there were a median of 43 viruses per family in this dataset (data not shown), we performed 200 random samplings of the 3,342,022 initial viruses, pulling up to 43 viruses per family in every cycle. The final RefSeq dataset consisted of 200 folders containing each exactly 6,778 viruses, for a total of 1,355,600 sequences. We computed the ANI ($1 - \text{Mash distance} * 100$) between all sequences in the RefSeq benchmarking dataset and those listed in the ICTV Release #39. We then grouped these sequences into four different slots, according to their ANI value in descending blocks of 10%, from 100% down to 60% ANI. Mean and standard deviation of all ANI groups are reported in the Supplementary Table 1. The RefSeq dataset is composed of both eukaryotic and prokaryotic viruses.

Finally, we used a dataset extracted from the Global Ocean Eukaryotic Viral database, as detailed by Gaïa et al. [19]. The dataset comprises eukaryotic double-stranded DNA viruses: 591 MAGs from Schulz et al. (2020) [28], 445 MAGs from Sunagawa et al. (2020) [29], and 218 MAGs from Moniruzzaman et al. (2020) [30], along with 158 reference viral assemblies, with the most recent access being on July 20, 2024. Data were sourced from GOEV_DB_CONTIGS.db.zip (https://doi.org/10.6084/

Table 1 Summary of benchmarking datasets. Each dataset is characterized by its size, description, dominant viral taxonomic groups, number of unique virus families, and the number of viruses with labels

Dataset	Size	Description	Main virus group	# of families	Viruses with labels
ICTV	860	Viruses extracted from ICTV VMR	Caudoviricetes, Pisoniviricetes, Alsuviricetes, Monjiviricetes, Bunyaviricetes	192	860
RefSeq	6,778	RefSeq viruses that exhibit a minimum MASH distance greater than 0.1 from any virus present in ICTV	Caudoviricetes, Pisoniviricetes, Alsuviricetes, Monjiviricetes, Bunyaviricetes, Arfiviricetes	270	5,870
kVSCs	2,232	A known viral sequence clusters dataset from human gut	Caudoviricetes	17	732
GOEV	1,412	A dataset of viruses related to giant viruses	Megaviricetes	N/A	1,412

m9.figshare.20284713), with selection criteria focusing on data labeled at the order taxon level. The taxonomic assignments from the GOEV database were found to be consistent with ICTV release VMR_MSL38_v3.

An overview of these datasets is provided in Table 1. RefSeq data can be downloaded via the NCBI Virus resource as indicated above. The ICTV virus metadata resource releases used in the benchmarking stage of this study were VMR_08-201218_MSL33 (release #33), VMR_18-191021_MSL36 (release #36), VMR_19-250422_MSL37 (release #37), VMR_MSL38_v3 (release #38), and VMR_MSL39_v1 (release #39).

Additional software used in our study

vConTACT2 v0.11.3 was run using the required amino acid FASTA file containing the input sequences, and the gene-to-genome mapping file generated using the vcontact2_gene2genome.py script (parameters: -s Prodigal-FAA) from MetaPhage v0.3.3, last accessed on Aug 8, 2024 via https://github.com/MattiaPandolfoVR/MetaPhage.

geNomad v1.8.0 was used in end-to-end mode using its default parameters across all benchmarking runs.

TIGTOG v0.1 was run using an E-value threshold of at most 1e−10 as suggested by the authors (parameters: −e 1e−10); last accessed on July 19, 2024 via https://github.com/anhd-ha/TIGTOG.

PhaGCN2 was run through Phabox (v2.1.9) [31] using minimum length 1500 (parameters: –len 1500).

VPF-Class v0.1 was run using the default parameters across all benchmarking datasets; last accessed on January 27, 2025 via https://github.com/biocom-uib/vpf-tools/tree/master.

Nucleotide BLAST from the BLAST+ suite installed locally (version 2.16.0) with parameter `outfmt` 6 and default word size was used to evaluate nucleotide sequence identity between 90,000 fragmented viruses and all ICTV viruses during the genome reduction study. Similarly, Mash version 2.3 was used to compute the average nucleotide identity between all RefSeq and all ICTV viruses. Sequences were first sketched using Mash `sketch` with a *k*-mer size of 21, then Mash `dist` was used to compute the distance. Mash was also used to compute the distance between 90,000 fragmented viruses and all ICTV viruses during the genome reduction study. *K*-mer values of $k=21$ and $k=31$ were used. Mash default value of k is 21, hence the choice to use such k, and a larger value of k which is a standard value used in bioinformatics genome analyses ($k = 31$, e.g., sequence assemblers).

CheckV v0.7.0 was used in completeness mode to assess viral genome completeness throughout the study, and to study the relationship of the wrongly predicted taxa with genome completeness. In three separate analyses, the values for (i) AAI completeness, (ii) bidirectional subsethood, and iii) tie score were compared between two groups: correctly classified viruses and incorrectly classified viruses. The virus-specific marker profiles are the fraction of markers that are labeled as being at least specific for viruses, therefore being labeled as "*V' or "V*" in the MMSeqs2 profile database file genomad_db_v1.7.tar.gz from https://zenodo.org/records/10594875 CC by 4.0, Antonio Camargo.

ICTVdump is a simple program for the automatic retrieval of genome sequences and metadata linked through the virus metadata resource. It automatically connects to any VMR version, retrieving the GenBank identifiers, then downloads them in batch from the NCBI and generates an SQL database of DNA sequences and one taxonomy table in CSV format. The program is written in python and is freely available at https://github.com/christopher-riccardi/ICTVdump. All software used in our study were run using 40 cores (AMD EPYC 9354 32-Core Processors) using the resources provided by USC Center for Advanced Research Computing (CARC).

Three distinct meta-analytic random-effects models were fit using the function `meta` package version 8.0.2 [32] in R version 4.2.2. The mean difference was chosen as the effect size since all studies were expressed on a common scale (AAI completeness, ranging between 0 and 100, bidirectional subsethood and tie score ranging between 0 and 1). Let $\overline{Y}_1$ and $\overline{Y}_2$ be the sample means of two independent groups of viruses, in terms of AAI completeness for the correctly (1) and incorrectly (2) classified. The raw mean difference is given by

$$D = \overline{Y}_1 - \overline{Y}_2$$

Let S_1 and S_2 be the sample standard deviations of the two groups, and n_1 and n_2 be the sample size in the two groups. Here we don't assume that the two population standard deviations are the same, and the variance of D is estimated by

$$v_D = \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}$$

We calculated D and v_D in three separate instances, one for every meta-analyzed factor and confirmed that the above-mentioned R package calculated these values as required.

Although the viral sequences from the three datasets do not overlap, as they are collected from different data sources, they may be related through their taxonomic relationships. Consequently, the three substudies may not be independent, which violates the assumptions of the random effects model. Therefore, the results from

the random effects model analyses should be interpreted with caution.

Evaluation metrics

The evaluation metrics used in this study include accuracy, precision, recall (sensitivity), specificity, F1 score, and Matthews Correlation Coefficient (MCC). Briefly, we calculated the accuracy score as the number of correct predictions divided by the total number of predictions:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

Precision was calculated as the ratio of true positive predictions to the total number of positive predictions (true positives and false positives). Therefore, for a given class i:

$$\text{Precision}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i}$$

Recall was calculated as the ratio of true positive predictions to the total number of actual positives (true positives and false negatives). For the i-th class:

$$\text{Recall}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i}$$

The proportion of actual negative instances that are correctly identified, for each class i, was calculated as

$$\text{Specificity}_i = \frac{\text{TN}_i}{\text{TN}_i + \text{FP}_i}$$

The F1 score, also for each class i, was calculated using the formula for the harmonic mean of precision and recall:

$$\text{F1 score}_i = 2 \times \frac{\text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}$$

MCC was calculated using all the elements in the confusion matrix, as

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

All the above metrics were calculated using python's scikit-learn library within a multi-class classification context. The overall metrics were derived as a weighted average to ensure that each class was considered proportionately to its distribution in the dataset.

Discussion

We propose a virus classification method based on a similarity metric between sequences, and show that using a recently published, comprehensive database of virus-specific marker profiles, our method can accurately predict the family taxonomic label with an overall average accuracy above 0.96. Benchmarking and performance analysis across all tested data sources demonstrate that Virgo can generalize and accurately predict viral family taxonomy with precision comparable to more sophisticated tools. Arguably, flexibility and tolerance in dealing with imprecise data are valuable characteristics for any virus classification software, and the bidirectional subsethood metric we propose for this type of classification may exhibit such characteristics.

Nonetheless, in at least two instances, the incorrect predictions were significantly associated to lower tie scores. Further investigation of these cases revealed that the bidirectional subsethood score, though robust for the majority of the viruses tested, can be shared at comparable levels across distinct families of the *Mononegavirales* order, thus making resolution at family level problematic.

Moreover, our method assumes that the user-provided sequences are viral; therefore, pre-filtering non-viral sequences using methods like geNomad [15], DeepVirFinder [33], or DeepMicroClass [34] is advised prior to classification.

The metagenomic datasets used in our study pay special attention to viruses of microorganisms, since viruses within the *Caudoviricetes* order and *Nucleocytoviricota* phylum are abundant in the sunlit ocean, where they play a crucial role in regulating plankton community composition and controlling bloom dynamics, and whereas gut phages have been shown to play a fundamental role in shaping and modulating bacterial growth by lysis and lysogeny, thus directly affecting human health since early age [35, 36].

Virgo showed excellent results in the classification of giant viruses derived from ocean metagenomes, but it was generally outperformed by a dedicated program, TIGTOG. It is fair to mention, however, that TIGTOG is in fact trained on a set of data that included sequences from GOEV itself, and this likely contributed to the higher performance on this dataset. Compared to the several other programs tested, we were surprised to observe that a general virus classification program like Virgo achieved comparable results on a dataset characterized by such a remarkably broad phylogenetic diversity.

Virgo exhibited a relatively stable performance on the kVSCs dataset, while the other methods exhibited an unexpectedly poorer performance. Although this dataset represents a novel resource of viruses associated with the human gut microbiome, deriving from high quality metaviromes, these viruses are still phylogenetically close to viruses in the current data banks (recall how the kVSCs are the known viral sequence clusters). Virgo was able to

capture their taxonomic lineage with great precision and recall, perceptibly better than any other method tested.

It is interesting to note that the viruses in the ICTV cover around 44% of the 161,862 virus-specific markers. On the one hand, this suggests that a significant portion of the markers are well represented in the ICTV. On the other hand, the remaining 56% of the markers are found in viral sequences that have been captured and are present in publicly available datasets, but have yet to be classified and named. In the light of this observation, we highlight the necessity to develop computational methods that can easily integrate new versions of the manually curated taxonomic labels.

We also provide an accessory program, ICTVdump, to facilitate and automate this operation, given the utmost importance of manually curated reference taxonomic lineages. Our method is designed to deal with an ever-improving resource such as the ICTV. Moreover, while our method is generally robust to fragmentation, as confirmed by the non-parametric tests of completeness across datasets and, even more prominently, by the artificial genome mutation and reduction study, the meta-analysis revealed that viruses correctly classified by Virgo tend to be 10% more complete than the misclassified ones. This suggests a potential limitation in cases where genomes are highly fragmented, since our method relies on the indirect comparison of marker gene sharing, and genome completeness is closely tied to viral sequence size. Nonetheless, compared to straightforward metrics of similarity scores (e.g., BLAST, ANI, or Jaccard coefficient, see "Results"), our method has shown better performance in dealing with incomplete data. These are desirable features that could position Virgo as a valuable tool for metagenomic studies or reclassification efforts in light of new ICTV updates.

Taken together, the presented results favor the hypothesis that the bidirectional subsethood of matched marker profiles is a viable metric on which accurate virus classification can rely on. Nonetheless, we can devise ways to further enhance methods using this similarity measure. For example, when attempting to distinguish harder-to-classify viruses, such as those in the order *Mononegavirales*, the metric could be further refined by re-architecturing the bidirectional subsethood and transforming the collections of sets from unordered to ordered. Modeling the order in which the two genomic sequences individually map to the markers profiles (a synteny-like approach) could help reduce ties and improve classification accuracy. We would like to emphasize that the classification robustness achieved is likely a testament to the extensive effort put into the collection and validation of the markers by the authors of geNomad, and we gratefully acknowledge their contributions.

Supplementary information

The online version contains supplementary material available at https://doi.org/10.1186/s40168-025-02159-x.

Supplementary material 1.

Supplementary material 2.

Supplementary material 3.

Supplementary material 4.

Acknowledgements
Not applicable.

Authors' contributions
Conceptualization: C.R. and F.S.; data collection and curation: C.R. and Y.W; source code writing and analysis: C.R., Y.W., and S.Y.; validation and testing: S.Y., C.R., and Y.W.; writing—original draft preparation: C.R. and Y.W; S.Y. helped with data analysis, presentation, and finalization of the manuscript. Writing—review and editing: all authors; supervision: F.S.; funding acquisition: F.S. All authors have read and agreed to the published version of the manuscript.

Funding
NSF grant [EF-2125142 to F.S.]; University of Southern California; University of Florence.

Data availability
The data supporting the findings of this study are available on FigShare and can be accessed at https://doi.org/10.6084/m9.figshare.28730093.v1. Source code for ICTVdump and Virgo are provided under an MIT license and are available at https://github.com/christopher-riccardi/ICTVdump and https://github.com/christopher-riccardi/ Virgo, respectively.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 28 September 2024 Accepted: 4 June 2025
Published online: 24 July 2025

References

1. Kuhn JH, Wolf YI, Krupovic M, Zhang YZ, Maes P, Dolja VV, et al. Classify viruses – the gain is worth the pain. Nature. 2019;566(7744):318–20. https://doi.org/10.1038/d41586-019-00599-8.
2. Dutilh BE, Varsani A, Tong Y, Simmonds P, Sabanadzovic S, Rubino L, et al. Perspective on taxonomic classification of uncultivated viruses. Curr Opin Virol. 2021;51:207–15. https://doi.org/10.1016/j.coviro.2021.10.011.
3. Lefkowitz EJ, Dempsey DM, Hendrickson RC, Orton RJ, Siddell SG, Smith DB. Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV). Nucleic Acids Res. 2018;46:D708–17. https://doi.org/10.1093/nar/gkx932.
4. Siddell SG, Smith DB, Adriaenssens E, Alfenas-Zerbini P, Dutilh BE, Garcia ML, et al. Virus taxonomy and the role of the International Committee on Taxonomy of Viruses (ICTV). J Gen Virol. 2023;104(5):001840. https://doi.org/10.1099/jgv.0.001840.
5. Gorbalenya AE, Krupovic M, Mushegian A, Kropinski AM, Siddell SG, Varsani A, et al. The new scope of virus taxonomy: partitioning the virosphere into 15 hierarchical ranks. Nat Microbiol. 2020;5(5):668–74. https://doi.org/10.1038/s41564-020-0709-x.

6. Steger G, Riesner D, Prusiner SB. Viroids, Satellite RNAs and Prions: Folding of Nucleic Acids and Misfolding of Proteins. Viruses. 2024;16(3):360. https://doi.org/10.3390/v16030360.
7. Koonin EV, Krupovic M, Agol VI. The Baltimore Classification of Viruses 50 Years Later: How Does It Stand in the Light of Virus Evolution? Microbiol Mol Biol Rev. 2021;85(3):10.1128/mmbr.00053–21. https://doi.org/10.1128/mmbr.00053-21.
8. Hull R, Rima B. Virus taxonomy and classification: naming of virus species. Arch Virol. 2020;165(11):2733–6. https://doi.org/10.1007/s00705-020-04748-7.
9. Simmonds P, Adriaenssens EM, Zerbini FM, Abrescia NGA, Aiewsakun P, Alfenas-Zerbini P, et al. Four principles to establish a universal virus taxonomy. PLoS Biol. 2023;21(2):e3001922. https://doi.org/10.1371/journal.pbio.3001922.
10. Zerbini FM, Siddell SG, Lefkowitz EJ, Mushegian AR, Adriaenssens EM, Alfenas-Zerbini P, et al. Changes to virus taxonomy and the ICTV Statutes ratified by the International Committee on Taxonomy of Viruses (2023). Arch Virol. 2023;168(7):175. https://doi.org/10.1007/s00705-023-05797-4.
11. Bin Jang H, Bolduc B, Zablocki O, Kuhn JH, Roux S, Adriaenssens EM, et al. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. Nat Biotechnol. 2019;37(6):632–9. https://doi.org/10.1038/s41587-019-0100-8.
12. Jiang JZ, Yuan WG, Shang J, Shi YH, Yang LL, Liu M, et al. Virus classification for viral genomic fragments using PhaGCN2. Brief Bioinform. 2022:bbac505. https://doi.org/10.1093/bib/bbac505.
13. Ha AD, Aylward FO. Automated classification of giant virus genomes using a random forest model built on trademark protein families. NPJ Viruses. 2024;2(1):1–9. https://doi.org/10.1038/s44298-024-00021-9.
14. Pons JC, Paez-Espino D, Riera G, Ivanova N, Kyrpides NC, Llabrés M. VPF-Class: taxonomic assignment and host prediction of uncultivated viruses based on viral protein families. Bioinformatics. 2021;37(13):1805–13. https://doi.org/10.1093/bioinformatics/btab026.
15. Camargo AP, Roux S, Schulz F, Babinski M, Xu Y, Hu B, et al. Identification of mobile genetic elements with geNomad. Nat Biotechnol. 2023:1–10. https://doi.org/10.1038/s41587-023-01953-y.
16. Chen IMA, Chu K, Palaniappan K, Ratner A, Huang J, Huntemann M, et al. The IMG/M data management and analysis system vol 7: content updates and new features. Nucleic Acids Res. 2023;51:D723–32. https://doi.org/10.1093/nar/gkac976.
17. Camargo AP, Nayfach S, Chen IMA, Palaniappan K, Ratner A, Chu K, et al. IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata. Nucleic Acids Res. 2023;51:D733–43. https://doi.org/10.1093/nar/gkac1037.
18. Zolfo M, Silverj A, Blanco-Míguez A, Manghi P, Rota-Stabelli O, Heidrich V, et al.. Discovering and exploring the hidden diversity of human gut viruses using highly enriched virome samples. bioRxiv; 2024. https://doi.org/10.1101/2024.02.19.580813.
19. Gaïa M, Meng L, Pelletier E, Forterre P, Vanni C, Fernandez-Guerra A, et al. Mirusviruses link herpesviruses to giant viruses. Nature. 2023;616(7958):783–9. https://doi.org/10.1038/s41586-023-05962-4.
20. Koonin EV, Dolja VV, Krupovic M, Varsani A, Wolf YI, Yutin N, et al. Global Organization and Proposed Megataxonomy of the Virus World. Microbiol Mol Biol Rev. 2020;84(2):10.1128/mmbr.00061–19. https://doi.org/10.1128/mmbr.00061-19.
21. Nayfach S, Camargo AP, Schulz F, Eloe-Fadrosh E, Roux S, Kyrpides NC. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nat Biotechnol. 2021;39(5):578–85. https://doi.org/10.1038/s41587-020-00774-7.
22. Glenn TC. Field guide to next-generation DNA sequencers. Mol Ecol Resour. 2011;11(5):759–69. https://doi.org/10.1111/j.1755-0998.2011.03024.x.
23. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 2016;17(1):132. https://doi.org/10.1186/s13059-016-0997-x.
24. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 10:421. https://doi.org/10.1186/1471-2105-10-421.
25. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11(1):119. https://doi.org/10.1186/1471-2105-11-119.
26. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017;35(11):1026–8. https://doi.org/10.1038/nbt.3988.
27. Blanco-Míguez A, Beghini F, Cumbo F, McIver LJ, Thompson KN, Zolfo M, et al. Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. Nat Biotechnol. 2023;41(11):1633–44. https://doi.org/10.1038/s41587-023-01688-w.
28. Schulz F, Roux S, Paez-Espino D, Jungbluth S, Walsh DA, Denef VJ, et al. Giant virus diversity and host interactions through global metagenomics. Nature. 2020;578(7795):432–6. https://doi.org/10.1038/s41586-020-1957-x.
29. Sunagawa S, Acinas SG, Bork P, Bowler C, Eveillard D, Gorsky G, et al. Tara Oceans: towards global ocean ecosystems biology. Nat Rev Microbiol. 2020;18(8):428–45. https://doi.org/10.1038/s41579-020-0364-5.
30. Moniruzzaman M, Weinheimer AR, Martinez-Gutierrez CA, Aylward FO. Widespread endogenization of giant viruses shapes genomes of green algae. Nature. 2020;588(7836):141–5. https://doi.org/10.1038/s41586-020-2924-2.
31. Shang J, Peng C, Liao H, Tang X, Sun Y. PhaBOX: a web server for identifying and characterizing phage contigs in metagenomic data. Bioinforma Adv. 2023;01:3. https://doi.org/10.1093/bioadv/vbad101.
32. Balduzzi S, Rücker G, Schwarzer G. How to perform a meta-analysis with R: a practical tutorial. Evid-Based Ment Health. 22(4):153–160. https://doi.org/10.1136/ebmental-2019-300117.
33. Ren J, Song K, Deng C, Ahlgren NA, Fuhrman JA, Li Y, et al. Identifying viruses from metagenomic data using deep learning. Quant Biol. 2020;8(1):64–77. https://doi.org/10.1007/s40484-019-0187-4.
34. Hou S, Tang T, Cheng S, Liu Y, Xia T, Chen T, et al. DeepMicroClass sorts metagenomic contigs into prokaryotes, eukaryotes and viruses. NAR Genomics Bioinforma. 2024;6(2):lqae044. https://doi.org/10.1093/nargab/lqae044.
35. Yeh YC, Fuhrman JA. Effects of phytoplankton, viral communities, and warming on free-living and particle-associated marine prokaryotic community structure. Nat Commun. 2022;13(1):7905. https://doi.org/10.1038/s41467-022-35551-4.
36. Yi X, Lu H, Liu X, He J, Li B, Wang Z, et al. Unravelling the enigma of the human microbiome: evolution and selection of sequencing technologies. Microb Biotechnol. 2024;17(1):e14364. https://doi.org/10.1111/1751-7915.14364.

Publisher's Note