Virus classification for viral genomic fragments using PhaGCN2

Jing-Zhe Jiang (jingzhejiang@gmail.com)
Chinese Academy of Fishery Sciences
Wen-Guang Yuan
Guangdong Pharmaceutical University
Jiayu Shang
City University of Hong Kong
Ying-Hui Shi
Guangdong Pharmaceutical University
Li-Ling Yang
Tianjin Agricultural University
Min Liu
Shanghai Ocean University
Peng Zhu
Shanghai Ocean University
Tao Jin
Guangdong Magigene Biotechnology Co., Ltd
Yanni Sun
City University of Hong Kong
Li-Hong Yuan
Guangdong Pharmaceutical University

software

Keywords: graph convolutional network, semi-supervised machine learning, Virus classification, ICTV, PhaGCN2

Posted Date: December 12th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1658089/v2

Abstract

Viruses are the most ubiquitous and diverse entities in the biome. Due to the rapid growth of newly identified viruses, there is an urgent need for accurate and comprehensive virus classification, particularly for novel viruses. Here, we present PhaGCN2, which can rapidly classify the taxonomy of viral sequences at family level and supports the visualization of the associations of all families. We evaluate the performance of PhaGCN2 and compare it with the state-of-the-art virus classification tools, such as vConTACT2, CAT, and VPF-Class, using the widely accepted metrics. The results show that PhaGCN2 largely improves the precision and recall of virus classification, increases the number of classifiable virus sequences in the Global Ocean Virome dataset (v2.0) by 4 times, and classifies more than 90% of the Gut Phage Database. PhaGCN2 makes it possible to conduct high-throughput and automatic expansion of the database of the International Committee on Taxonomy of Viruses. The sourcecode is freely available at https://github.com/KennthShang/PhaGCN2.0 .

Key Points

- PhaGCN2 can rapidly classify the taxonomy of viral sequences at family level and supports the visualization of the associations of all families.
- PhaGCN2 largely improves the precision and recall of virus classification, increases the number of classifiable virus sequences in the Global Ocean Virome dataset (v2.0) by 4 times, and classifies more than 90% of the Gut Phage Database.
- PhaGCN2 makes it possible to conduct high-throughput and automatic expansion of the database of the International Committee on Taxonomy of Viruses.

Background

As the most abundant biological entities on Earth, viruses can hijack organisms from every branch of the tree of life. They play critical roles in host mortality, metabolism, physiology, and evolution, impacting marine biogeochemical cycling and shaping the Earth's microbiomes [1-5]. David Baltimore established a virus classification system based on messenger RNA (mRNA) synthesis-the Baltimore classification system [6]. Similarly, based on the virus host, viruses can be classified into four types, namely, animal viruses, fungi viruses, plant viruses, or bacteriophages [7]. Based on these different classification features, some virus databases have been established, such as plant and fungi virus database—DPVweb [8], *coronavirus* database—ViPR [9], influenza and *coronavirus* database—GISAID [10], and comprehensive virus databases that are publicly available resource and updated weekly—ViralZone [11] and Virxicon [12].

Culture-independent next-generation sequencing technologies have recently been used to explore the tremendous diversity of the virosphere from multiple samples [13-15]. With rapid expansion of viral genome databases, these advances have led the International Committee on Taxonomy of Viruses (ICTV) to present a consensus statement suggesting a shift from the "traditional" classification criteria—for example, virion morphology and single- or multiple-gene phylogenies—toward a genome-centered, and perhaps one day largely automated, viral taxonomy [16].

The virus classification mainly relies on the manual classification and definition of virologists, which is too slow to classify millions of viral genome sequences. For example, despite millions of virus sequences in IMG/VR [15, 17], there are only about 10,550 types of classified viruses in the ICTV 2021 report (hereafter ICTV2021). Therefore, there is an urgent need for a virus classification method that can rapidly and accurately classify these new viral genome sequences and align computational classifications with ICTV-ratified taxa [18].

In our previous work, we present a semi-supervised machine learning model, named PhaGCN [19], based on a graph convolutional network (GCN). There are two main components in PhaGCN: convolutional neural network (CNN) encoder and GCN classifier. First, the CNN encoder will encode contigs from different lengths into 256-dimensional embedding vectors. Each vector represents the motif-related patterns captured from the DNA sequences. Second, a knowledge graph is built to connect known phages in the RefSeq database and the test phages. Each node in the graph represents a phage, and the edges between phages represent sequence and protein-composition based similarity. We use the embedding vectors outputted from the CNN encoder as the node features and apply protein organization and protein similarity to define the edges. Finally, the semi-supervised GCN is applied on the knowledge graph to utilize both known phages and test phages for training. However, the current version can only conduct the classification virus under *Caudovirales* [19]. More importantly, ICTV will frequently adjust its taxonomy criteria according to the progress of research, such as deleting old families, adding new families, and moving members from one family to another. The continuous change of the reference and the emergence of novel viruses are impeding the accuracy and sensitivity of automatic prediction. In particular, most learning-based models must specify the label set (e.g. family labels), which will not accommodate viruses from new families [20]. Thus, a method that can possibly recognize new families is needed to support automatic virus classification.

Here, we present PhaGCN2 to align computational classifications with ICTV-ratified taxa by automatically upgrading the database. PhaGCN2 can predict the taxonomy of viral sequences at the family level and accurately identify the members of the novel virus families that have not yet been defined in ICTV. We compare PhaGCN2 with the state-of-the-art virus classification tools (vConTACT2 [21], CAT [22], and VPF-Class [23]) using widely accepted metrics such as precision, recall, and required computing resources. The experimental results show that our method is superior to the existing methods.

Material And Methods

Datasets and benchmarked tools

The main datasets and tools used or evaluated in this paper are listed as follows (table 1 and table 2):

Table 1.Datasets

Datasets	Years	Habits	Description
ICTV2021	2021		The 2021 ICTV Virus Metadata Resource https://ictv.global/filebrowser/download/468
ICTV2020	2020		The 2020 ICTV Virus Metadata Resource https://ictv.global/filebrowser/download/467
GPD (Gut Phage Database)	2021	Human gut	Lawley et al. (2021) created the Gut Phage Database (GPD), a collection of 142,809 non-redundant viral genomes (length>10 kb) obtained by mining 28,060 globally distributed human gut metagenomes and 2,898 reference genomes of cultured gut bacteria [14]. http://ftp.ebi.ac.uk/pub/databases/metagenomics/genome_sets/gut_phage_database/
GOV2.0 (global ocean DNA virome dataset)	2019	Ocean	Gregory et al. (2021) established an ~12-fold expanded global ocean DNA virome dataset (GOV2.0) of 195,728 viral populations, now including the Arctic Ocean, and validated that these populations form discrete genotypic clusters [13]. https://data.iplantcollaborative.org/dav/iplant/commons/community_released/iVirus/GOV2.0/
MGV (Metagenomic Gut Virus)	2021	Human stool	Nayfach, et al. (2021) assembled the Metagenomic Gut Virus catalogue that comprises 189,680 viral genomes from 11,810 publicly available human stool metagenomes, naming the dataset as MGV [28]. https://github.com/snayfach/MGV
DOV (Dataset of Oyster Virome)	2021	Oyster	Jiang et al. (2021) established a Dataset of Oyster Virome (DOV) that contains 728,784 non-redundant viral operational taxonomic unit (vOTU) contigs and 3,473 high-quality viral genomes, enabling the first comprehensive overview of viral communities in oysters [27]. https://ngdc.cncb.ac.cn/gsub/submit/bioproject/subPRO010366/overview
Test RNA database	2016 and 2018	invertebrate and vertebrate	Shi et al. 2016 [29] profiled the transcriptomes of over 220 invertebrate species sampled across nine animal phyla and reported the discovery of 1,445 RNA viruses, including some that are sufficiently divergent to comprise new families. And in 2018, using a large-scale meta-transcriptomic approach, they discovered 214 vertebrate-associated viruses in reptiles, amphibians, lungfish, ray-finned fish, cartilaginous fish and jawless fish [30]. https://static-content.springer.com/esm/art%3A10.1038%2Fnature20167/MediaObjects/41586_2016_BFnature20167_MOESM439_ESM.xlsx https://static-content.springer.com/esm/art%3A10.1038%2Fs41586-018-0012-7/MediaObjects/41586_2018_12_MOESM3_ESM.xlsx

Table 2. The benchmarked tools

Tools	vConTACT2	CAT	VPF-Class
Years	2019	2019	2021
Author	Sullivan, M. B.et.al.	Dutilh, B. E.et.al.	Pons, J. C.et.al
Description	vConTACT2 is a tool to perform taxonomy classification of viral genomic sequence data. It is designed to cluster and provide the taxonomic context of viral metagenomic sequencing data [21].	CAT is a comparison-based species classification tool for metagenomic contigs. It first conducts gene calling, then maps the predicted ORFs against the nr protein database, and finally classifies entire contigs based on classification of the individual ORFs [22].	VPF-Class is a tool that can conduct host prediction and taxonomic classification of viruses. It is a comparison-based metagenomic contig annotation tool [23].
	(https://bitbucket.org/MAVERICLab/vcontact2/wiki/Home)	(https://github.com/dutilh/CAT)	(https://github.com/biocom-uib/vpf-tools)

Sequences preprocessing before building the protein database

When training the CNN [24] model, to ensure that the number of samples sequence at the family-level is enough, we need to remove small families before building the database. The filter condition is length≧1700bp (To make sure the sequence contains enough information), family members≧8 (To ensure that each family contains at least seven training sequences and one validation sequence), and ACGT contigs (skipping contigs with non-ACGT characters (e.g NNN gaps)).

Statistical information

We select random number of sequences to quantify the usage of computing resources. The sequences are randomly chosen using a random number generator in python. Run time was measured with the "/usr/bin/time" command available in Linux. Peak memory was measured with the "/usr/bin/free -h" command available in Linux. The Knowledge Graph network was visualized with Gephi [25] (v.0.9.2; https://gephi.org/) software. The others are drawn by R.

Method optimization of PhaGCN and PhaGCN2

In addition to predicting families under only *Caudoviruses*, there are still some important limitations in PhaGCN that have not been addressed. Because ICTV will frequently adjust its taxonomy criteria according to the progress of research, such as deleting old families, adding new families, and moving members from one family to another. The continuous change of the reference and the emergence of novel viruses are impeding the accuracy and sensitivity of automatic prediction. In particular, most learning-based models must specify the label set (e.g. family labels), which will not accommodate viruses from new families [20]. In view of this, we have made the following improvements to PhaGCN, including (1) updating with ICTV and using prodigal to build reference database under the entire virus realm, (2) using network graph to show the clustering relationship among family members, and (3) the prediction of novel viral families (*family_like*) based on the topology of the network (outlier nodes).

Identification of *family_like* nodes in the PhaGCN2 network

Despite the efforts of ICTV in providing continuous updates of taxonomic classification system for viruses, existing taxonomic groups are not adequate to keep pace with fast accumulation of diverse viruses. Thus, PhaGCN2 allows users to use '*family_like*' classification to discovery possibly new taxonomic groups. Specifically, PhaGCN2 applies GCN to conduct prediction for all nodes in the graph. If a test node has a predicted family label 'A' and is not a one-hop neighbor of any training node, it is defined as "A_like". Thus, we have *family_like* nodes for different families. A closer look shows that many of these nodes can form connected components by themselves that are not part of the bigger network. Examples of those clusters (clusters -) are shown in Fig. S1. These components tend to represent new taxonomic groups that have not been included in the current classification system. Some '*family_like*' nodes can also form paths with characterized viruses (training node) as exemplified by cluster in Fig. S1. Following the definition of '*family_like*' nodes, the distance between them and any training node must be bigger than 2 (i.e. 2 edges at least).

Results

Improvements of PhaGCN2

In summary, PhaGCN2 contains three major improvements comparing to the previous version (Table S1), including (1) updating with ICTV and using prodigal to build reference database under the entire virus realm (Table S2), (2) using network topology to assist outlier recognition, and (3) assigning outlier nodes to *family_like*. The improvements in (2) and (3) enable PhaGCN2 to automatically suggest new families, which removes the limitation on fixed set of labels in commonly used supervised learning models. These improvements allow PhaGCN2 to obtain more accurate predictions than the original version, with the *precision* (equation (1)) increased from 73.19% to 83.91%, the *recall* (equation (2)) increased from 87.92% to 89.30%, and F_1 (equation (3)) increased from 79.88% to 86.52% (Table S3). The detailed descriptions can be found in the following sections.

$$Precision = \frac{TP\ (True\ Positive)}{TP\ (True\ Positive)+FP\ (False\ Positive)} \quad (1)$$

$$Recall = \frac{TP\ (True\ Positive)}{TP\ (True\ Positive)+FN\ (False\ Negative)} \quad (2)$$

$$F_1 = 2 * \frac{Recall*Precision}{Recall+Precision} \quad (3)$$

Database construction. The PhaGCN protein database is constructed by manually downloading protein sequences from National Center for Biotechnology Information (NCBI). There are two potential disadvantages to use the old database. First, the number of proteins is limited by the update of the RefSeq protein database. Second, users need to map the proteins to their original genomes sequence-by-sequence, which is tedious and error-prone. To establish a faster and more user-friendly pipeline to construct the database, we apply Prodigal [26] to conduct gene finding and protein translation based on the up-to-date ICTV database, with the latest ICTV2021 containing 10,550 viruses. PhaGCN2 with the database constructed by Prodigal was compared with the original PhaGCN database using 8,760 virus sequences (length>8000bp) in DOV (Dataset of Oyster Virome) [27]. The results reveal that 98.46% of the predictions are consistent, indicating that using Prodigal to establish a protein database is reliable (Table S2). Now, users can align computational classifications with ICTV-ratified taxa by the function of training virus classification database in PhaGCN2.

Network visualization. Similar to vConTACT2, PhaGCN2 can also output the virus family clustering network. This gives us an intuitive understanding of the relationship between different virus families and family members. In addition to visualizing the family relationship, we also use the network topology to identify possible new families, which consist of subgraphs with weak connection with nodes from ICTV. First, we identify *outliers*, which are test viruses (nodes) not connected to any viruses from ICTV (Figure S1A, red dots). Often these *outliers* are from new families but they were assigned to the predefined families (Figure S1B, green dots) due to the design limitation of the supervised learning algorithm.

*Family-like prediction***.** To support the automatic identification of new families, we assign these *outliers* as *family_like* (probably belong to another family which is close to a reference family). For instance, if a node is predicted to be *Lipothrixviridae_like*, it means that this node is close to *Lipothrixviridae*, but it is not recommended to be cluster it into the same family. To verify the feasibility of predicting *outlier* as *family_like*, we use the ICTV2020 virus to build a protein database, and use the newly added viruses from ICTV2021 (including 2,636 viral reference genomes after filtrating) as the test data. Detailed prediction results are shown in Table S3. The *precision* and *recall* after integrating this function for each family is shown in Table S4.

Among the 2,636 newly added viruses, 339 of them belong to families that are not defined in ICTV2020 and thus their labels do not exist in our training data. PhaGCN2 assigned 204 viruses as *family_like* in total. Among these sequences, 167 test sequences are members of real novel families of ICTV2021 or the families not included during ICTV2020 training. Therefore, the *precision* of *family_like* label is 81.86% (167/204), and the *recall* is 49.26% (167/339). Among the 167 true *family_like* labels, 153 viruses are defined in ICTV2021 as *Genomoviridae* (a novel family in ICTV2021), but they were predicted as *Geminiviridae* (the same order under *Geplafuvirales* with *Genomoviridae*) in PhaGCN. Now, PhaGCN2 predicts them as *Geminiviridae_like*, which means these viruses probably belong to a family closely related to *Geminiviridae*. The other 37 test sequences were mistakenly annotated as *family_like*, as they are family members in the ICTV2020 list according to ICTV2021. For example, some viruses are *Myoviridae* in ICTV2021, but were predicted as *Drexlerviridae* (the same order under *Caudovirales* with *Myoviridae*) by PhaGCN. Now, PhaGCN2 recognize them as *Drexlerviridae_like.* Notably, although they are classified under *Myoviridae* according to the ICTV2021 criteria, they belong to a new genus under the family, which have no edges to the members of *Myoviridae* in ICTV2020. In fact, most of the 37 test sequences are classified a new genus in ICTV2021.

Comparison with the state-of-the-art tools

In order to have a comprehensive evaluation of PhaGCN2, we compare PhaGCN2 with vConTACT2, CAT, and VPF-Class using six widely used metrics, *precision* (equation (1)), *recall* (equation (2)), *F1-score* (Balanced Score, equation (3)), *consistency* (equation (4)) computing speed, and peak memory.

$$consistency = \frac{\text{The same prediction by two tools}}{\text{the number of viruses predicted by both tools}} \quad (4)$$

Table 3. Comparison of PhaGCN2 with the state-of-the-art virus classification tools

Tools	PhaGCN2	vConTACT2[1]	CAT	VPF-Class[2]
Test Data	9603[3](ICTV2021)			
True Positive	8379	1616	6928	3803
False Positive	260	704	825	1026
False Negative	965	3098	1851	855
Precision	96.99%	69.65%	89.36%	78.75%
Recall	89.67%	34.28%	78.92%	81.64%
F1-score	93.19%	45.95%	83.82%	80.17%

[1]RNA virus genomes were excluded from vConTACT2 test data evaluation as it was designed for only DNA virus classification.

[2]The *Orthornavirae* virus genomes were excluded from VPF-Class test data evaluation as it was designed for only DNA virus and RT virus classification. We only count the result that both the membership ratio and confidence score are high than 0.2 as the positive result of VPF-Class.

[3]Virus genomes longer than 1700 bp in the ICTV2021 were used as the test data for the evaluation of all the software.

Consistency: To compare the *consistency* of the prediction made by the three tools, we take the ICTV2021 data (9603 viral genomes sequence, known reference viruses) as test data. As show in Figure 1A, the number of viruses of predicted by both vConTACT2 and PhaGCN2 are 2248, and 1494 of them are identical, with a consistency value of 66.46% ((739+755)/(1199+1049)) (Detailed information is listed in Table S5). The number of viruses predicted by both PhaGCN2 and CAT are 6752, and 5090 of them are identical, with a consistency value of 75.39% ((739+4351)/(1199+5553)). The number of viruses predicted by both vConTACT2 and CAT are 1266, and 777 of them are identical, with a consistency value of 61.37% ((739+38)/(1199+67)). There are 1199 sequences predicted by all three tools with 739 having the same prediction, leading to a consistency value of 61.63% (739/1199).

Then we further take GOV2.0 (including 482,522 virus genome sequences and most of them are novel viruses) as the test data. The paper of GOV2.0 provided the ready-to-use results of vConTACT2 prediction [13]. Thus, we only ran PhaGCN2 and CAT to predict the GOV2.0 (VPF-Class is not included in the test as its slow calculation). vConTACT2 only acquired 47,839 predictions (9.91%), CAT predicted 170,200 viruses (35.27%), and PhaGCN2 acquired 199,833 predictions (41.41%). As shown in Figure 1B, the number of viruses predicted by both vConTACT2 and PhaGCN2 are 20,287, and 16,958 of them are identical, with a *consistency* value of 83.59% ((3205+13753)/(5441+14846)) (Detailed information is listed in Table S6). The number of viruses predicted by both PhaGCN2 and CAT are 13,996, and 5,694 of them are identical, with a *consistency* value of 40.68% ((3205+2489)/(5441+8555)). The number of viruses predicted by both vConTACT2 and CAT are 10780, and 5,893 of them are identical, with a consistency value of 54.67% ((3205+2688)/(5441+5339)). There are 5,441 sequences predicted by all three tools, and 3,205 sequences have the same results, with a *consistency* of 58.90% (3205/5441). These results show that these tools have similar *consistency* for known viruses. But when focusing on unknown viruses, alignment-based classification methods such as CAT has lower *consistency* with other tools.

Precision, recall, and F1-score: As mentioned above, when using the newly added viruses from ICTV2021 as the test data. the *recall* and *precision* of PhaGCN2 are 89.30% and 83.91%, respectively (Table S1). Here, we further tested PhaGCN2, vConTACT2, CAT, and VPF-class on all the ICTV2021 sequences collected in the PhaGCN2 database, and compared the obtained predictions with the classification of ICTV2021 (Table 3). As shown in table 3, PhaGCN2 achieves the best performance, with 10% higher F1-score than the second-best tool CAT. In particular, PhaGCN2's precision and recall are 7% and 11% higher than CAT, respectively. The results show that PhaGCN2 largely improves the precision and recall of virus classification over the state-of-the-art tools. The detailed results are shown in Table S5.

The elapsed time and peak memory: In addition, we recorded the elapsed time and peak memory of the three tools. We randomly selected 1000, 5000, and 10000 sequences from GPD [14] for testing (Figure 2). PhaGCN2 is faster than vConTACT2 but slower than CAT. vConTACT2 has a high memory usage in the step of calculating similarity networks, while PhaGCN2 and CAT consumes less memory.

Analysis of the sequences without predictions

As mentioned above, while using the newly added virus from ICTV2021 as the test data, there are 1,492 sequences with predictions and 1,142 sequences without prediction. In our analysis of 1142 sequences without predictions (Table S7), 992 of them are from newly added families by ICTV2021 and thus cannot be predicted by PhaGCN2. Of the remaining 150 sequences, 80 are new genera under known families. We speculate that these new genera cannot be predicted because the different genera in these families are of low similarity. In addition, 49 sequences are missed. Although they are not new genera, they are not trained by PhaGCN2 because the sample size of this genus in the 2020 training set was too small (genera member < 8). For the remaining 21 sequences, we cannot determine the cause for the time being. However, compared with the total 2,634 test sequences, the number is acceptable.

Furthermore, we examined the protein-level similarity between newly added sequences in ICTV2021 with and without predictions against the reference genomes (ICTV2020 training data) using Diamond blastx, and compared their similarity distributions. As shown in Figure S2, the protein sequence

identity distributions are significantly different between the two groups. Among them, virus sequences with relatively low variability and identity about 54.8% are likely to be predicted by PhaGCN2. However, highly variable sequences with identity lower than 37.4% have a low probability of prediction. Detailed results are shown in Table S8.

Possibility of genus-level prediction

Same as vConTACT2, PhaGCN2 can also draw the network diagram. We use the metagenomes of about 1700 human gut microbiome DNA viruses [28] as the test data and map the network with the results of PhaGCN2. Due to the space limitation, we only show the results of the 10 largest families in the database (Figure 3A). It is obvious that virus nodes of the same families cluster closely. To visualize clusters at genus-level, we investigated the genera in the most abundant family—*Siphoviridae*. Again, the top 16 genus members in *Siphoviridae* were visualized using different colors in Figure 3B. We can see that some genera, *Pahexavirus*, *Skunavirus,* and *Ceduovirus*, were clustered within themselves. However, some genera (such as *Triavirus, Phietavirus*, *Bioseptimavirus*, *Dubowvirus,* and *Peeveelvirus*) were mixed together (Figure 3B). This suggests that they are not different enough for PhaGCN2 to predict them as different genus.

Investigation of public data using PhaGCN2

GPD and GOV2.0 represent two completely different viral habitats. In this section, we use PhaGCN2 to classify the GPD and GOV2.0 database. After removal of the ineligible sequences, they are left with 142,333 (in all 142809) and 328,173 (in all 482522), respectively. As shown in Figure 4, the overall recall of GPD and GOV2.0 is 91.9% and 40.8% respectively. The higher proportion of the unknown viruses in GOV2.0 is far more than GPD, indicating that viruses in the ocean has not been fully explored, with a large portion still under the iceberg. When only focusing on the classified categories (without unknown), *Siphoviridae,* and *Myoviridae* account for 54.5%, and *Siphoviridae_like* and *Myoviridae_like* account for 31.1% in GPD. In contrast to GPD, *Siphoviridae*, and *Myoviridae* account for 28.9%, and *Siphoviridae_like* and *Myoviridae_like* account for 40.4% in GOV2.0. If other families under *Caudovirales*, such as *Podoviridae* and *Herelleviridae,* are included, 99.16% of the phages in the human gut are *Caudovirales*, while 94.8% in the ocean. It means that *Caudovirales* is the majority of both GPD and GOV2.0 at the order level, but GPD and GOV2.0 is quite different at the family level. Detailed results are shown in Table S9 and Table S5.

We further applied PhaGCN2 to classify 2202 qualified RNA virus genomes from the study of invertebrate and vertebrate viromes [29, 30]. There are 1094 sequences with predictions, and only six virus genomes are predicted to be non-RNA viruses. The top 3 families are *Marnaviridae*, *Dicistroviridae*, and *Nodaviridae*, and they account for 18.7% in total. However, there are up to 52.5% of viruses cannot be taxonomically classified to a known viral family, which shows that our understanding of RNA virosphere is still very limited. The detailed results are shown in Table S10 and Figure S3.

Furthermore, according to the classification and site information of GOV2 at the family level, we plotted the distribution abundance maps of *Myoviridae* and *Siphoviridae* at different sites and depths (Figure 5). As shown in Figure 5, the closer to the equatorial region and upper ocean, the higher the proportion of *Myoviridae* is. In contrast, the proportion of *Siphoviridae* in the two poles is higher than in the equator. This means that viruses from different families may have evolved unique adaptations to the different niches over a long period. The detailed longitude, latitude, and content data are shown in Table S11.

Discussion

vConTACT2 is a widely recognized tool for virus classification using a combination of ClusterONE [31], hierarchical clustering [32], and Markov cluster algorithm (MCL)-generated protein clusters [21]. The advantage of this method is that it can accurately predict the genome classification of large DNA phages with multiple ORFs and frequent recombination. However, its performance deteriorates for phage contigs that contain fewer protein clusters. PhaGCN integrates the protein-cluster-based features into a more powerful machine learning model based on graph convolutional network and thus achieves higher accuracy with less computing resources. However, PhaGCN is limited to only phages, limiting its utility to comprehensive virus taxonomic classification. PhaGCN2 removes this limitation by augmenting the learning model and reference database. PhaGCN2 can be applied to all types of viral metagenomic data and automatically produces family-level taxonomic classification of both DNA and RNA viruses. In addition, it can suggest new viral families based on the network topology. Alignment-based classification methods such as CAT or comprehensive BLAST [33] rely only on the alignment result, and simply infer species' classification based on majority votes. Although CAT is the second accurate tool in identifying known viruses (Table 3), alignment-based tools are not optimized for classifying novel or highly diverged viruses. VPF-Class can conduct host prediction and taxonomic classification of viruses, but it only can classify dsDNA, ssDNA, and retroviruses.

Compared to CAT, vConTACT2, and VPF-Class, one limitation of PhaGCN2 is that is cannot conduct taxonomic classification below family [34]. Although our method can be extended to genus-level prediction, the small number of members of many genera are not sufficient to train a generalized learning model. Removing this limitation is our future work. An ideal genus-level classification tool should address some additional challenges. First, the number of genera is significantly larger than the families. Currently, there are about 2200 genera based on ICTV's report. In addition, the number of genomes in the genera form a highly imbalanced distribution, posing challenges for rare genus classification. Third, some genera under the current ICTV standard are too similar to be distinguished effectively (Figure 3B). Because our work focuses on family-level classification, presenting the detailed comparison of current tools at genus level is beyond the scope of this work. With the continuous growth of the ICTV reference data set and the adjustment of ICTV on close-related genera, prediction at the genes-level will be more feasible.

Like other learning-based models, PhaGCN2's performance also relies on the quality of the training data. Due to the bias in sequencing, current training data does not systematically cover different taxonomic groups. Although PhaGCN2 leverages network topology to suggest novel families, its prediction ability on new families is limited. The detection rate of unknown virus sequences with identity less than 37.4% is usually very low (Figure S2). One possible strategy to enhance classification of new viral families is to conduct iterative prediction using PhaGCN2. First, we can conduct predictions on all viral genome data (such as IMG/VR [15]) using PhaGCN2. Then, we can add the newly predicted *Family_like* members into the training data to increase the capacity of PhaGCN2 on identifying more members of new families. The iterative training and searching is likely to increase the ability of PhaGCN2 on new family detection. We will investigate this in our future work.

However, for those "dark matter" sequences with no or very low similarity, it may be an impossible task to do a *de novo* viral classification. First, we can't evaluate the accuracy of predictions. Second, without any homologs, it is difficult to characterize the structure or function of their genomes. No matter how many sequences are identified, they are still "dark matter".

Finally, as PhaGCN2 does not predict whether the input sequence belongs to the virus or the host cell, we strongly recommend using viral sequences as input to PhaGCN2. In other words, virus identification tools (such as DIAMOND [35], Virsorter2 [36], etc.) should be used to remove non-viral sequences before PhaGCN2 is applied.

Declarations

AVAILABILITY

The source code of PhaGCN2 is available via: https://github.com/KennthShang/PhaGCN2.0.

SUPPLEMENTARY DATA

Figure 1. Venn diagram of consistency among the results of three virus classification software. A: The test data: 9603 ICTV2021 sequences. **B:** The test data :482,522 GOV2.0 sequences [13]. The number without parentheses is the number of sequences with predictions, the number in parentheses is the number of sequences with the same prediction by the corresponding tools, the percentage is the *consistency* between two tools or among three tools.

Figure 2. Comparison of the elapsed time and peak memory between PhaGCN2, vConTACT2, and CAT. 1000, **5000**, and **10000** represent the number of test genome sequences.

Figure 3. The clustering effect of PhaGCN2 network diagram at family-level and genes-level. The topological structures of A and B are identical. The test data is MGV [28]. The top 10 families in the A are marked with different colors. low_abundance without staining represents other low abundance families. B specifically shows different genera in *Siphoviridae*. High-abundance genera (number of members≧10) are marked with different colors, low_abundance is marked with light green, representing other low-abundance genera in *Siphoviridae*; other_family represents non-*Siphoviridae*.

Figure 4. Comparison of family-level composition in GPD and GOV2.0 based on PhaGCN2 predictions. The pie chart shows the percentage of each family in the **GPD** and **GOV2.0** database based on PhaGCN2's results. Low_abundance represents the total number of families with a low number (less than 0.5% of the total number); unknown represents the unpredicted number; the others represent each section. The total test sequences of GPD and Gov2.0 are 142809 and 482522, respectively.

Figure 5. Distribution comparison of the *Myoviridae* and *Siphoviridae* viral populations of GOV2 with latitude and depth.The color depth in the map represents the percentage of ***Myoviridae*** and ***Siphoviridae*** in the total virus species in the sample site. The sampling depth from left to right is **5m-15m**, **15m-150m**, and **150m-1000m** respectively.

Figure S1. The network diagram comparison before and after applying "family_like". A: visualization of the network of PhaGCN2. **B:** visualization of the network of PhaGCN. The enlarged portion of the red box is in the upper right corner. Database_virus (purple dots): the reference genomes from ICTV 2020. Test_virus (green dots) and family_like (red dots) are the test data (2,636 newly added viruses in ICTV2021 comparing to ICTV2020). Database_virus (purple dots): training nodes. Test_virus (green dots): a test node that is directly connected with training node. 'Family_like' (red dots): a test node that is not adjacent to any training node.

Figure S2. Comparison of protein sequence identity between test sequences (with and without predictions by PhaGCN2) and the reference genomes. N: sequences without predictions; Y: sequences with predictions.

Figure S3. Family-level composition of RNA viruses based on PhaGCN2 predictions. The pie chart shows the percentage of each family in 2202 novel RNA virus genomes from the study of invertebrate and vertebrate viromes [29, 30] based on PhaGCN2 results. Low_abundance represents the total number of families with a low number; unknown represents the sequences without predictions; the others represent each section. The total test sequences of RNA virus datatabase is 2202.

Supplementary Table 1: Method optimization of PhaGCN2 comparing to PhaGCN

Supplementary Table 2: Comparison of the database construction methods between PhaGCN and PhaGCN2

Supplementary Table 3: Comparison of PhaGCN2 with PhaGCN

Supplementary Table 4: Comparison of PhaGCN2 with PhaGCN for each virus family (ICTV2020)

Supplementary Table 5: Comparison among the results of three virus classification software (vConTACT2, PhaGCN2, and CAT)

Supplementary Table 6: Detailed predictions using four tools (vConTACT2, PhaGCN2, CAT, and VPF-tools)

Supplementary Table 7: Analysis of sequences without predictions

Supplementary Table 8: Diamond blastx results of ICTV2021 newly added genomes against the ICTV2020 reference genomes

Supplementary Table 9: Detailed predictions of GPD using PhaGCN2

Supplementary Table 10: Detailed predictions of 2202 novel RNA virus genomes from the study of invertebrate and vertebrate viromes

Supplementary Table 11: Station information of GOV2.0 and its corresponding virus family percentage

Supplementary Data are available at BIB online.

FUNDING

This project was supported by the Key-Area Research and Development Program of Guangdong Province (2022B0202110001), the Natural Science Foundation of China (nos. 31872499 and 31972847) to Yuan LH and Jiang JZ; the Central Public-Interest Scientific Institution Basal Research Fund, CAFS (nos. 2020TD42 and 2021SD05) to Jiang JZ; the Guangdong Provincial Special Fund for Modern Agriculture Industry Technology Innovation Teams (no. 2019KJ141) to Jiang JZ. The funders had no role in the study design, data collection, and analysis, decision to publish, or manuscript preparation.

CONFLICT OF INTEREST

The authors declare that they have no competing interests.

References

1. Gelderblom HR. Structure and Classification of Viruses. In: Baron S, editor. Medical Microbiology. 4th edition: Galveston (TX): University of Texas Medical Branch at Galveston, 1996. Chapter 41.

2. Suttle CA. Marine viruses–major players in the global ecosystem, Nature Reviews Microbiology 2007;5:801-812.

3. Geoghegan JL, Holmes EC. Predicting virus emergence amid evolutionary noise, Open Biol 2017;7:170189.

4. Asokan GV, Kasimanickam RK. Emerging Infectious Diseases, Antimicrobial Resistance and Millennium Development Goals: Resolving the Challenges through One Health, Cent Asian J Glob Health 2013;2:76.

5. Grant W. Hypothesis—Ultraviolet-B Irradiance and Vitamin D Reduce the Risk of Viral Infections and thus Their Sequelae, Including Autoimmune Diseases and some Cancers, Photochemistry and photobiology 2008;84:356-365.

6. Baltimore D. Expression of animal virus genomes, Bacteriol Rev 1971;35:235-241.

7. Bhat AI, Rao GP. Host Range of Viruses. In: Bhat A. I., Rao G. P. eds). Characterization of Plant Viruses : Methods and Protocols. New York, NY: Springer US, 2020, 29-31.

8. Adams MJ, Antoniw JF. DPVweb: a comprehensive database of plant and fungal virus genes and genomes, Nucleic Acids Res 2006;34:D382-385.

9. Pickett BE, Greer DS, Zhang Y et al. Virus pathogen database and analysis resource (ViPR): a comprehensive bioinformatics database and analysis resource for the coronavirus research community, Viruses 2012;4:3209-3226.

10. Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health, Glob Chall 2017;1:33-46.

11. Masson P, Hulo C, De Castro E et al. ViralZone: recent updates to the virus knowledge resource, Nucleic Acids Res 2013;41:D579-583.

12. Kudla M, Gutowska K, Synak J et al. Virxicon: A Lexicon Of Viral Sequences, Bioinformatics 2020;36:5507-5513.

13. Gregory AC, Zayed AA, Conceicao-Neto N et al. Marine DNA Viral Macro- and Microdiversity from Pole to Pole, Cell 2019;177:1109-1123 e1114.

14. Camarillo-Guerrero LF, Almeida A, Rangel-Pineros G et al. Massive expansion of human gut bacteriophage diversity, Cell 2021;184:1098-1109 e1099.

15. Roux S, Paez-Espino D, Chen IA et al. IMG/VR v3: an integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses, Nucleic Acids Res 2021;49:D764-D775.

16. Simmonds P, Adams MJ, Benko M et al. Consensus statement: Virus taxonomy in the age of metagenomics, Nat Rev Microbiol 2017;15:161-168.

17. Paez-Espino D, Chen IA, Palaniappan K et al. IMG/VR: a database of cultured and uncultured DNA Viruses and retroviruses, Nucleic Acids Res 2017;45:D457-d465.

18. Dutilh BE, Varsani A, Tong Y et al. Perspective on taxonomic classification of uncultivated viruses, Curr Opin Virol 2021;51:207-215.

19. Shang J, Jiang J, Sun Y. Bacteriophage classification for assembled contigs using graph convolutional network, Bioinformatics 2021;37:i25-i33.

20. Y.S. Abu-Mostafa MM-I, H.-T. Lin. Learning from data: a short course. [United States] : AMLBook, 2012, ISBN-13: 978-1600490064.

21. Bin Jang H, Bolduc B, Zablocki O et al. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks, Nat Biotechnol 2019;37:632-639.

22. von Meijenfeldt FAB, Arkhipova K, Cambuy DD et al. Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT, Genome biology 2019;20:1-14.

23. Pons JC, Paez-Espino D, Riera G et al. VPF-Class: Taxonomic assignment and host prediction of uncultivated viruses based on viral protein families, Bioinformatics 2021;37:1805-1813.

24. Shang J, Sun Y. CHEER: HierarCHical taxonomic classification for viral mEtagEnomic data via deep leaRning, Methods 2021;189:95-103.

25. M B, S H, M J. Gephi: an open source software for exploring and manipulating networks, International AAAI Conference on Weblogs and Social Media 2009;3:361-362.

26. Hyatt D, Chen GL, Locascio PF et al. Prodigal: prokaryotic gene recognition and translation initiation site identification, BMC Bioinformatics 2010;11:119.

27. Jiang J-Z, Fang Y-F, Wei H-Y et al. Dataset of Oyster Virome and the Remarkable Virus Diversity in Filter-Feeding Oysters, Research Square 2021.

28. Nayfach S, Páez-Espino D, Call L et al. Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome, Nature Microbiology 2021;6:960-970.

29. Shi M, Lin X-D, Tian J-H et al. Redefining the invertebrate RNA virosphere, Nature 2016;540:539-543.

30. Shi M, Lin X-D, Chen X et al. The evolutionary history of vertebrate RNA viruses, Nature 2018;556:197-202.

31. Nepusz T, Yu H, Paccanaro A. Detecting overlapping protein complexes in protein-protein interaction networks, Nat Methods 2012;9:471-472.

32. Lima-Mendez G, Van Helden J, Toussaint A et al. Reticulate representation of evolutionary and functional relationships between phage genomes, Mol Biol Evol 2008;25:762-777.

33. Altschul SF, Gish W, Miller W et al. Basic local alignment search tool, J Mol Biol 1990;215:403-410.

34. Yilin Zhu JS, Yanni Sun. Phage taxonomic classification: challenges, current tools, and limitations 2022:arXiv: 2209.01942.

35. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND, Nat Methods 2015;12:59-60.

36. Guo J, Bolduc B, Zayed AA et al. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses, Microbiome 2021;9:37.

Figures

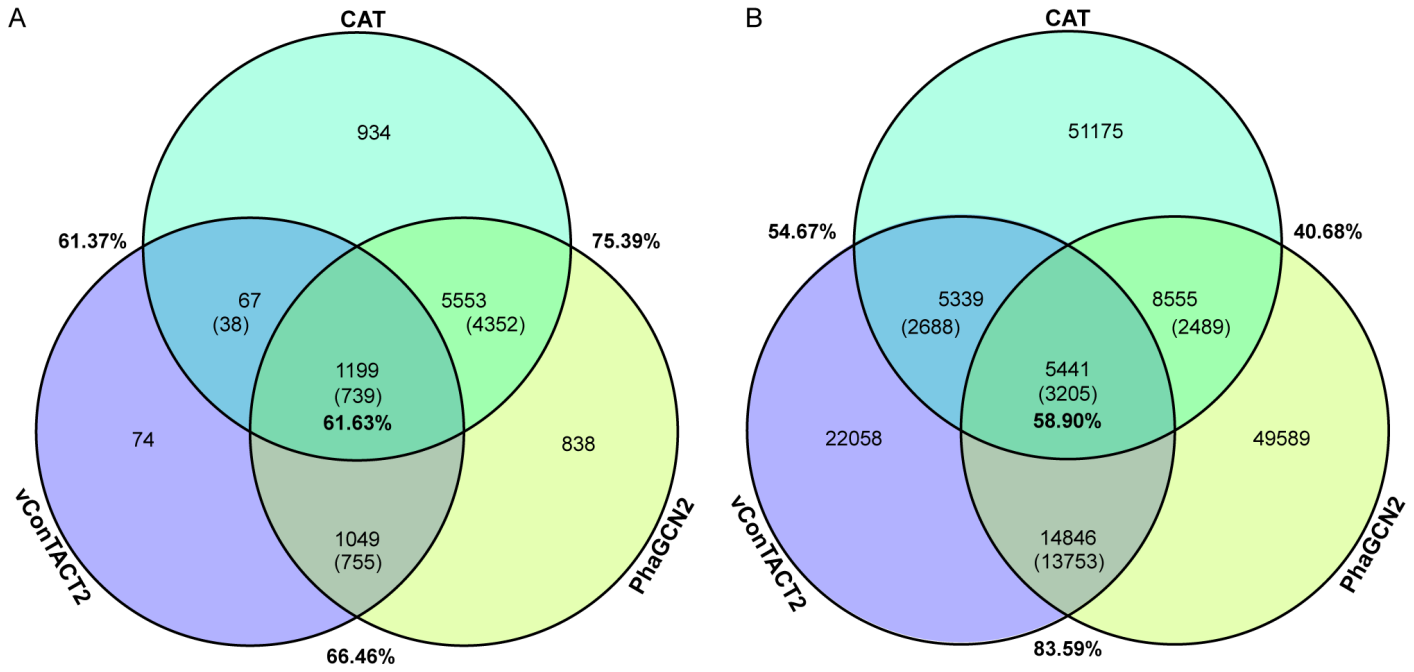

Figure 1. Venn diagram of consistency among the results of three virus classification software. A: The test data: 9604 ICTV2021 sequences. **B:** The test data :482,522 GOV2.0 sequences(6). The number without parentheses is the number of sequences with predictions, the number in parentheses is the number of sequences with the same prediction by the corresponding tools, the percentage is the consistency between two tools or among three tools.

Figure 1

Venn diagram of consistency among the results of three virus classification software. A: The test data: 9603 ICTV2021 sequences. **B** The test data :482,522 GOV2.0 sequences . The number without parentheses is the number of sequences with predictions, the number in parentheses is the number of sequences with the same prediction by the corresponding tools, the percentage is the *consistency* between two tools or among three tools.

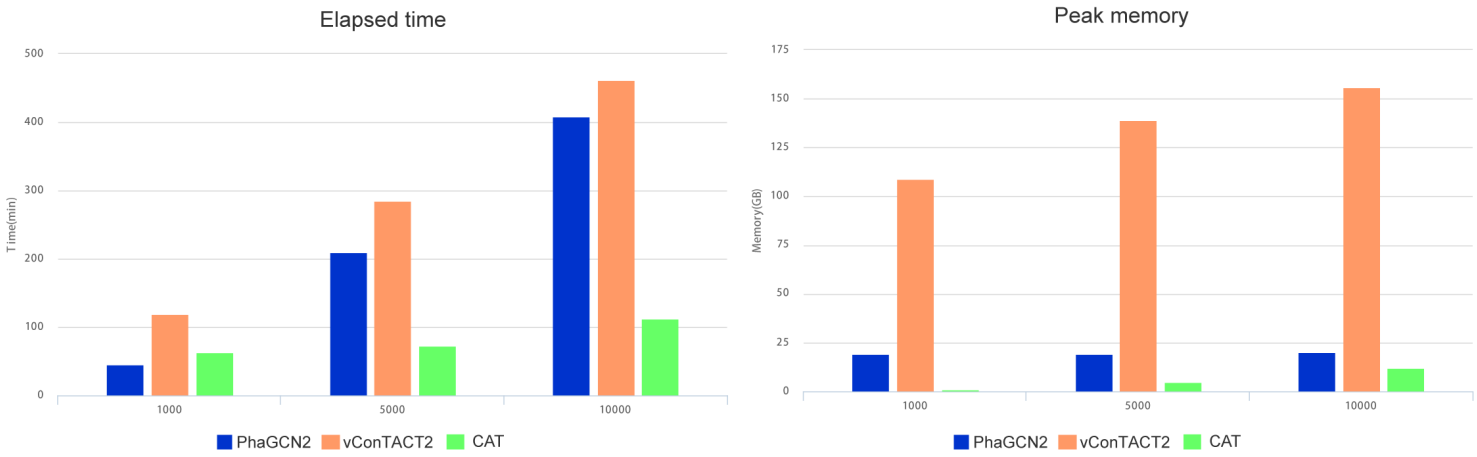

Figure 2. Comparison of the elapsed time and peak memory between PhaGCN2, vConTACT2, and CAT. 1000, 5000, and **10000** represent the number of test genome sequences.

Figure 2

Comparison of the elapsed time and peak memory between PhaGCN2, vConTACT2, and CAT. 1000, **5000**, and **10000** represent the number of test genome sequences.

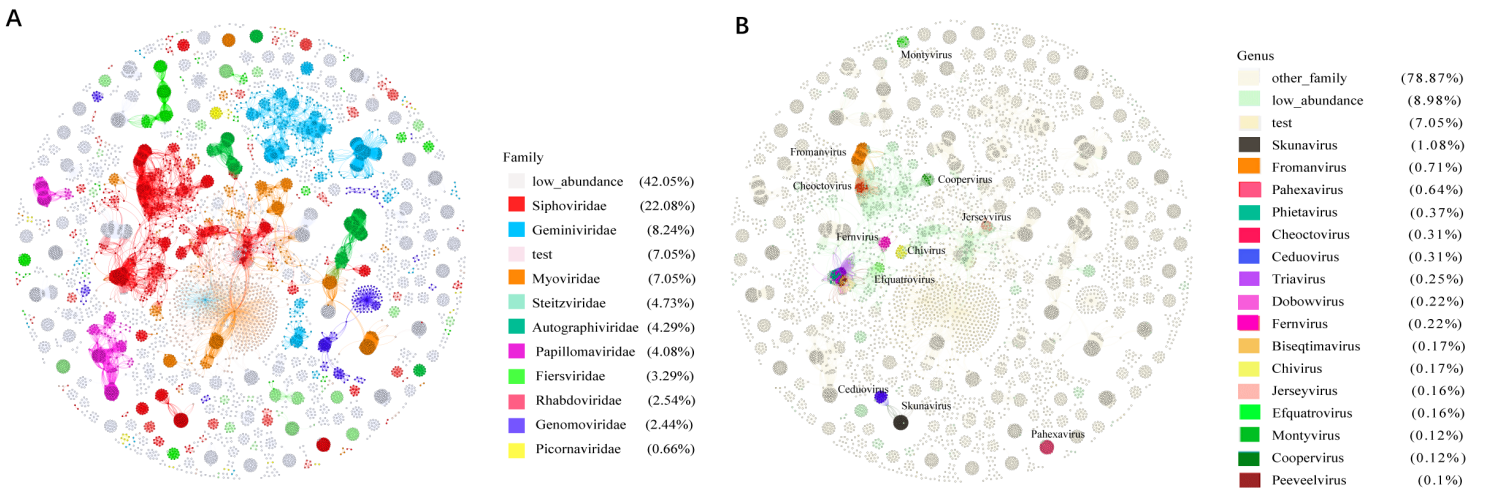

Figure 3. The clustering effect of PhaGCN2 network diagram at family-level and genes-level.The topological structures of A and B are identical. The test data is MGV(16). The top 10 families in the A are marked with different colors. low_abundance without staining represents other low abundance families. B specifically shows different genera in *Siphoviridae*. High-abundance genera (number of members≧10) are marked with different colors, low_abundance is marked with light green, representing other low-abundance genera in *Siphoviridae*; other_family represents non-*Siphoviridae*.

Figure 3

The clustering effect of PhaGCN2 network diagram at family-level and genes-level. The topological structures of A and B are identical. The test data is MGV . The top 10 families in the A are marked with different colors. low_abundance without staining represents other low abundance families. B specifically shows different genera in *Siphoviridae*. High-abundance genera (number of members≧10) are marked with different colors, low_abundance is marked with light green, representing other low-abundance genera in *Siphoviridae*; other_family represents non-*Siphoviridae*.

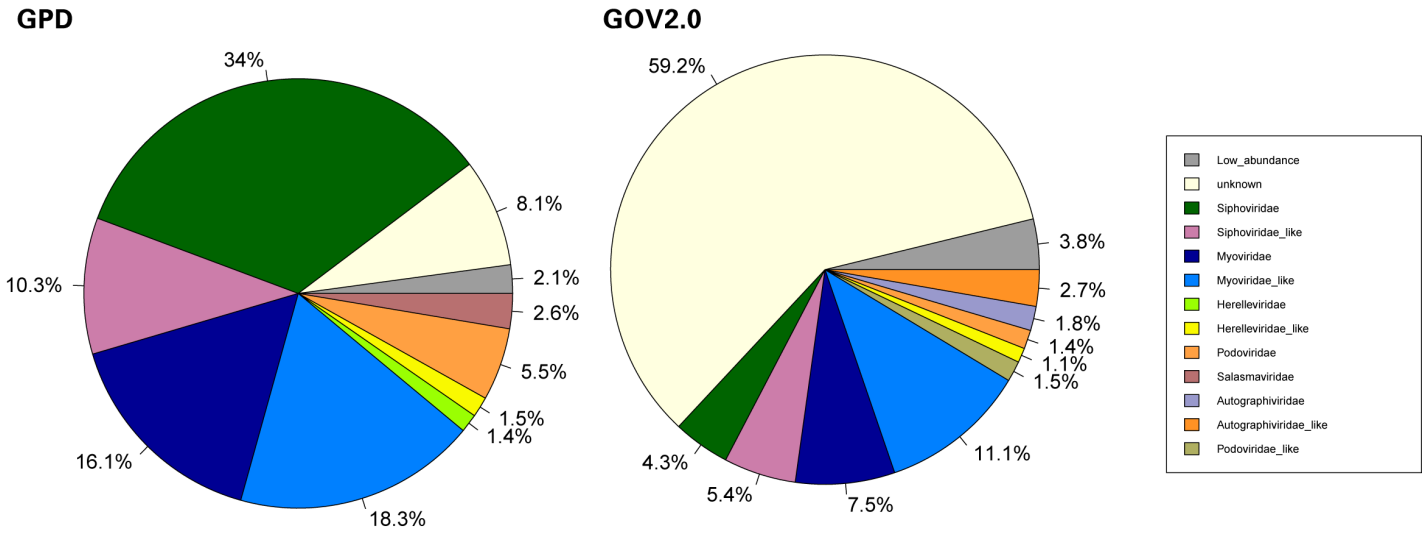

Figure 4. Comparison of family-level composition in GPD and GOV2.0 based on PhaGCN2 predictions. The pie chart shows the percentage of each family in the **GPD** and **GOV2.0** database based on PhaGCN2's results. Low_abundance represents the total number of families with a low number (less than 0.5% of the total number); unknown represents the unpredicted number; the others represent each section. The total test sequences of GPD and Gov2.0 are 142809 and 482522, respectively.

Figure 4

Comparison of family-level composition in GPD and GOV2.0 based on PhaGCN2 predictions. The pie chart shows the percentage of each family in the **GPD** and **GOV2.0** database based on PhaGCN2's results. Low_abundance represents the total number of families with a low number (less than 0.5% of the total number); unknown represents the unpredicted number; the others represent each section. The total test sequences of GPD and Gov2.0 are 142809 and 482522, respectively.

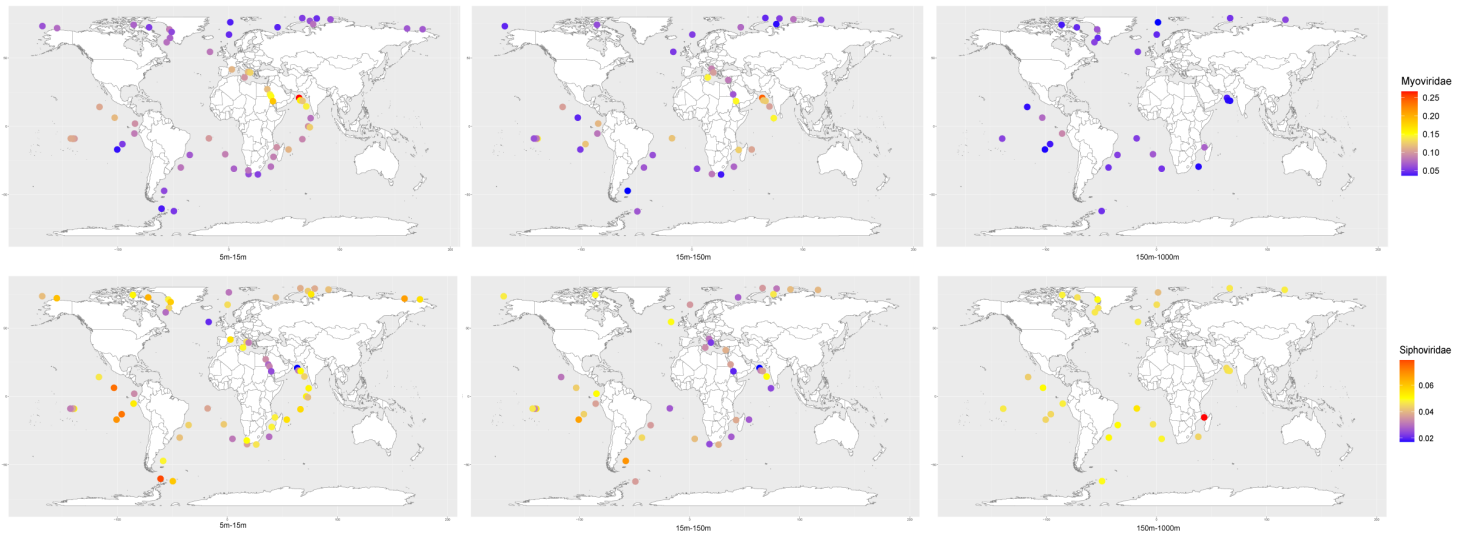

Figure 5. Distribution comparison of the Myoviridae and Siphoviridae viral populations of GOV2 with latitude and depth. The color depth in the map represents the percentage of ***Myoviridae*** and ***Siphoviridae*** in the total virus species in the sample site. The sampling depth from left to right is **5m-15m, 15m-150m, and 150m-1000m** respectively.

Figure 5

Distribution comparison of the *Myoviridae* and *Siphoviridae* viral populations of GOV2 with latitude and depth. The color depth in the map represents the percentage of ***Myoviridae*** and ***Siphoviridae*** in the total virus species in the sample site. The sampling depth from left to right is **5m-15m**, **15m-150m**, and **150m-1000m** respectively.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarydata.xlsx
- FigureS1.pdf
- FigureS2.pdf
- FigureS3.pdf