Genomics 114 (2022) 110414

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

VirusTaxo: Taxonomic classification of viruses from the genome sequence using k-mer enrichment

Rajan Saha Raju [a], Abdullah Al Nahid [b], Preonath Chondrow Dev [b], Rashedul Islam [c,d,*]

[a] Department of Computer Science and Engineering, Shahjalal University of Science and Technology, Sylhet 3114, Bangladesh
[b] Department of Biochemistry and Molecular Biology, School of Life Sciences, Shahjalal University of Science and Technology, Sylhet 3114, Bangladesh
[c] Omics Lab, Dhaka, Bangladesh
[d] Bioinformatics Graduate Program, University of British Columbia, Vancouver, BC V5Z 4S6, Canada

ARTICLE INFO

Keywords:
Virus
Taxonomy
Hierarchical classification
k-mer
Genome

ABSTRACT

Classification of viruses into their taxonomic ranks (e.g., order, family, and genus) provides a framework to organize an abundant population of viruses. Next-generation metagenomic sequencing technologies lead to a rapid increase in generating sequencing data of viruses which require bioinformatics tools to analyze the taxonomy. Many metagenomic taxonomy classifiers have been developed to study microbiomes, but it is particularly challenging to assign the taxonomy of diverse virus sequences and there is a growing need for dedicated methods to be developed that are optimized to classify virus sequences into their taxa. For taxonomic classification of viruses from metagenomic sequences, we developed VirusTaxo using diverse (e.g., 402 DNA and 280 RNA) genera of viruses. VirusTaxo has an average accuracy of 93% at genus level prediction in DNA and RNA viruses. VirusTaxo outperformed existing taxonomic classifiers of viruses where it assigned taxonomy of a larger fraction of metagenomic contigs compared to other methods. Benchmarking of VirusTaxo on a collection of SARS-CoV-2 sequencing libraries and metavirome datasets suggests that VirusTaxo can characterize virus taxonomy from highly diverse contigs and provide a reliable decision on the taxonomy of viruses.

1. Introduction

The virus genome consists of either DNA or RNA and is broadly classified as DNA virus or RNA virus [1] respectively. Viruses are classified into taxonomic ranks which play important roles in finding their source, genetic relationship, ancestry, and origin. Taxonomic classification of viruses ensures the consistent and accurate classification of novel viruses [2]. Conventionally, several phenotypic properties of viruses including molecular composition, structure, proteins, host range, and pathogenicity [3] are used to classify taxonomic ranks. Recently, strong relationships between genome sequence and taxonomic assignments of viruses have been reported at family level and inter-family groupings into orders [4]. With the advent of high-throughput sequencing technologies, more viruses have been characterized solely from sequencing data than using phenotypic properties [3]. These newly sequenced viruses are required to be assigned to their taxonomic ranks using automated computational tools. Comparisons of virus sequences using pairwise sequence similarity and phylogenetic relationships have become the major tool to define taxonomic ranks of novel viruses [5,6]. Family of the novel virus SARS-CoV-2 that caused the recent pandemic in 2020 was identified by sequencing and comparing its genome sequence with known virus sequences [7]. Most of the existing computational methods to identify virus taxonomy are based on similarities in genome structure and organization, the presence of homologous gene and protein sequences [8,9,10,11]. Homology based methods require higher computational resources, might produce unreliable alignment for novel viral species and often require human interpretations [12].

For classification of the virus sequences several alignment free supervised machine learning classifiers have been proposed e.g., CASTOR [13], VirFinder [14], DeepVirFinder [15] etc. However, the existing tools notably do not predict hierarchical taxonomic ranks from viral sequences across the diverse virus taxa. These methods were benchmarked on limited datasets of certain well characterized virus families. CASTOR used the features of restriction fragment length polymorphism to train the machine learning models in three virus families. DeepVirFinder uses a convolutional neural network that learns from viral genomic signatures to classify virus sequences from non-virus sequences. VirFinder uses k-mer (i.e., DNA words of length k) frequency

[*] Corresponding author.
E-mail address: rashedul.islam@alumni.ubc.ca (R. Islam).

https://doi.org/10.1016/j.ygeno.2022.110414
Received 4 September 2021; Received in revised form 27 April 2022; Accepted 13 June 2022
Available online 17 June 2022

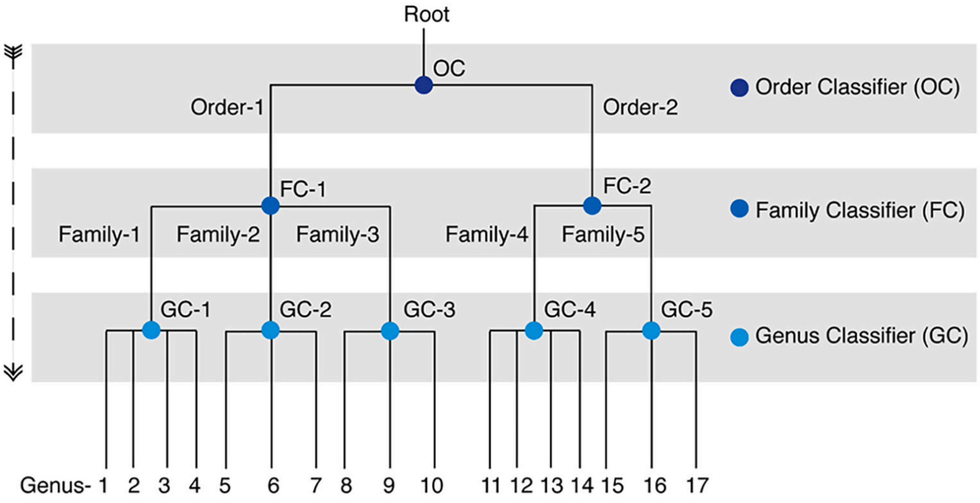


Fig. 1. Multi-class hierarchical classification model. Example of a hierarchical structure of virus taxonomic ranks. Classifier(s) are added at each level of taxonomic ranks. To build the VirusTaxo models, k-mers are extracted from the genomes of each class. Unique k-mers are then indexed and stored in a database to find the k-mer overlap with the query sequence. To measure the confidence of prediction, VirusTaxo provides a ranking of genus prediction using softmax probability and entropy scores (see methods).

features to train the model using a logistic regression model to discriminate virus sequences. For the classification of short metagenomic reads or contigs into the microbial taxa including viruses, k-mer feature is considered as core in some proposed classifiers e.g., Kraken [16], Kraken2 [17], KrakenUniq [18], CLARK [19], CLARK-S [20] and MetaPhlAn [21]. MetaPhlAn uses clade-specific marker genes to assign metagenomic reads to the clades and CLARK uses discriminative k-mers of target sequences e.g., genus-level sequences. Similar to CLARK, we used enrichment of discriminative k-mers to classify taxonomic ranks of viruses using whole genome sequence or contigs in VirusTaxo during hierarchical training. K-mer extraction from virus sequences does not require prior knowledge of sequence homology and coding or non-coding regions at the gene level. Therefore, k-mer based approaches could be more effective at detecting taxonomy of novel viruses that are distantly related to the known virus sequences.

Sequencing of virus genomes has become an essential tool in clinical research, molecular epidemiology and evolutionary genomics. Metagenomic or metavirome sequencing contains sequences of novel or poorly characterized virus genomes [22]. Unassigned virus sequences are required to be classified accurately to their taxonomic ranks and such taxonomic assignment can be done from their sequences alone [4]. Currently, there is a lack of dedicated bioinformatics tools that are optimized for DNA or RNA viruses to assign virus taxonomy from sequences. Here we have developed VirusTaxo which makes decisions based on k-mer overlap (i.e., exact sequence match) of a given sequence with discriminative sets (mutually exclusive sets) of k-mers of known virus genera. VirusTaxo was trained on 6950 virus genomes that encompass 129 families, and 682 genera of DNA and RNA viruses which might help to discover the taxa of uncharacterized viruses that are related to known virus genera. VirusTaxo outperformed other state-of-the-art machine learning methods to accurately assign taxonomic ranks in both DNA and RNA viruses. VirusTaxo has been benchmarked against CLARK [19], Kraken2 [17] and DeepVirFinder [15] to classify virus sequences from metagenomic datasets and outperformed in terms of detecting higher number of viruses in diverse genera. VirusTaxo was applied on 6176 whole and partial genome sequences of SARS-CoV-2 and was able to predict its taxonomy accurately in all cases. The source code of VirusTaxo is publicly available to create and train a classifier on labeled virus sequences. A web application of VirusTaxo is also available for users to predict the taxonomic rank of viruses from genome sequence or contig.

2. Results

2.1. *Classification of taxonomic ranks of viruses using VirusTaxo*

We trained VirusTaxo using DNA and RNA virus genomes to predict their hierarchical taxonomic ranks into order, family and genus. Total 4421 DNA and 2529 RNA virus genomes were used to train the VirusTaxo models that belong to 402 DNA and 280 RNA virus genera **(Supplementary file)**. For the hierarchical classification of virus taxonomic ranks, we trained classifiers at each layer of the taxonomic tree. **(**Fig. 1**)** illustrates an example of total 8 classifiers that were trained for 2 orders, 5 families with 17 genera at three taxonomic ranks. For selecting a classification method to train the VirusTaxo models, we benchmarked the accuracy of k-mer enrichment method used in CLARK along with random forest, gradient boosting, multilayer perceptron, k-nearest neighbors. During method selection, we used a smaller pilot dataset randomly subsampled from the entire RefSeq complete virus genomes **(see method)** and the pilot dataset allowed us to expedite the benchmarking of methods and parameters using lower computational resources. In both DNA and RNA virus datasets, k-mer enrichment outperformed other methods at all taxonomic ranks (e.g., order, family, and genus) whereas k-nearest neighbors and gradient boosting showed the lowest accuracies in DNA and RNA models respectively (Table S1). For the DNA dataset, k-mer enrichment showed on average 1% (order), 6% (family), and 16% (genus) improvement over other four methods. For RNA datasets, k-mer enrichment showed an average of 5.5% (order), 13% (family), and 27% (genus) improvement over other methods. The accuracies of all methods we have tested are relatively lower in RNA dataset compared to DNA dataset. On average RNA virus genome is 5 times smaller than DNA virus and has 43% (2529/4421) less number of species genomes available compared to DNA virus. Potentially for those reasons, higher accuracies could not be achieved in RNA models than DNA models across all the methods. Therefore, we selected the k-mer enrichment method in VirusTaxo with additional modifications such as using it hierarchically, optimizing k-mer length, applying entropy cutoff, and reducing database size for virus sequence classification.

2.2. *Benchmarking of VirusTaxo parameters*

DNA and RNA virus genomes are different in their genome sizes and sequence compositions [1]. The median size of the DNA and RNA virus genomes are 40,562 bp and 4556 bp respectively. We extracted k-mers using different ranges of k-mer lengths e.g., 17–26 bp and 13–22 bp for DNA and RNA viruses respectively and benchmarked the accuracy of models for different k-mer lengths using the pilot dataset. The accuracies

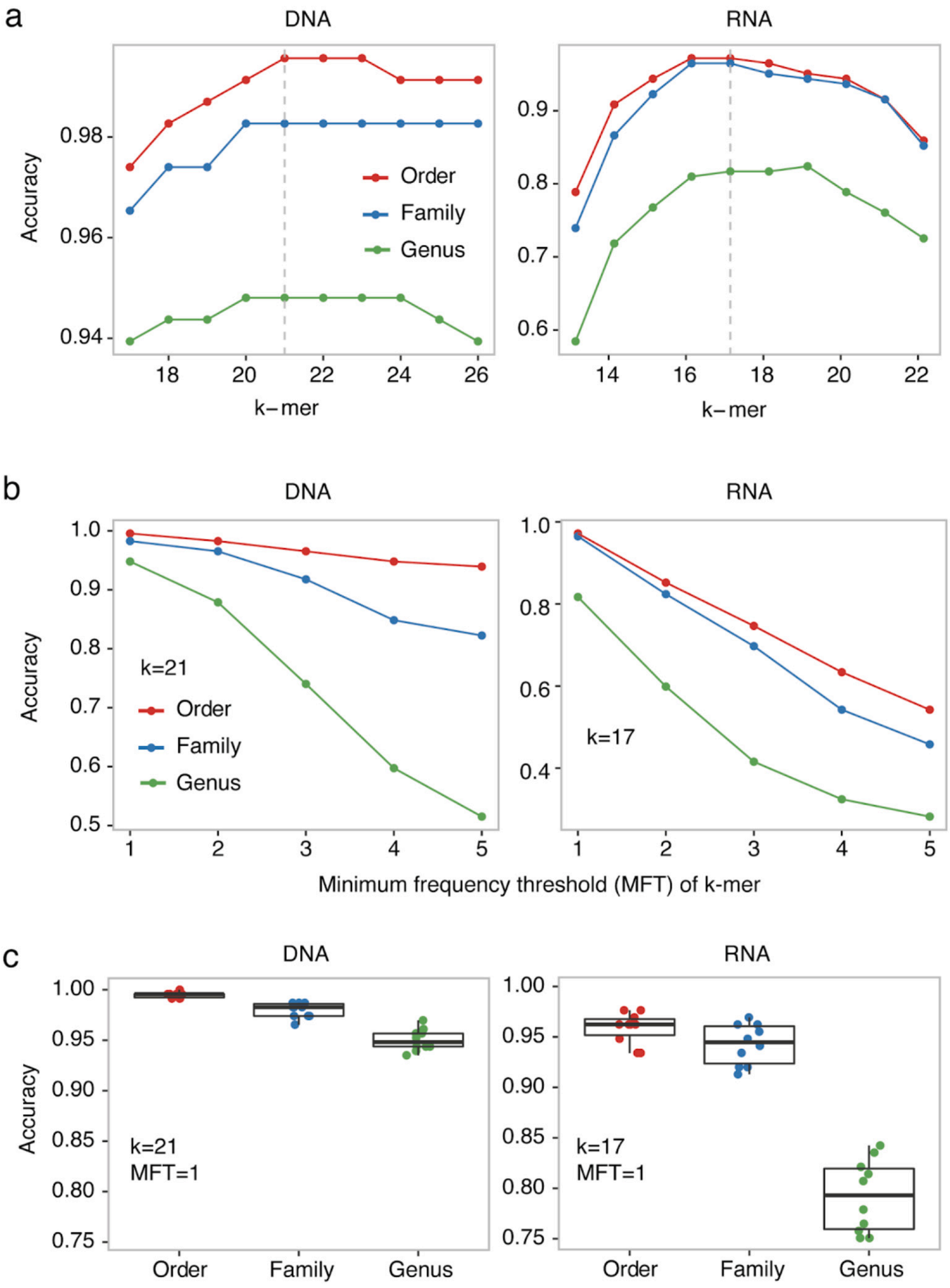


Fig. 2. Accuracy of VirusTaxo for order, family, and genus level classification in the pilot dataset. **a)** Changes of accuracies at different k-mers. For DNA and RNA datasets, 21 and 17 k-mer lengths provided the highest accuracy which are highlighted in gray dotted lines. **b)** Accuracies with different minimum frequency thresholds (MFT) at k-mer length of 21 bp and 17 bp in DNA and RNA viruses respectively. **c)** Accuracies of VirusTaxo for 10 rounds of testing of DNA and RNA models. For each iteration of hierarchical testing, one species genome per genus was randomly selected from the DNA and RNA datasets.

of DNA and RNA models varied at different k-mer lengths. K-mer lengths of 21–23 bp showed the highest accuracies in order (99.57%), family (98.27%), and genus (94.81%) level in the DNA model (Fig. 2a). At the family level, the accuracies did not change between the k-mer lengths of 20–26 bp. For the RNA model, k-mer length of 17 bp provided the maximum accuracies where the accuracies fluctuated with different k-mer lengths (Fig. 2a). This suggests for a given dataset, k-mer length determines the number of distinct genomic k-mers that will be the most discriminative. Our analysis shows that with the increase of k-mer length accuracy increases and after a point, accuracy starts to decrease. At a fixed k-mer length, the accuracies also reduced with the increase of minimum frequency threshold (MFT) of k-mers in both models where MFT value of 1 gave the highest accuracies (Fig. 2b) **(see methods)**. This is suggesting that unique k-mers in each class contribute significantly to discriminate between classes. Using the pilot dataset, we tested the accuracies of DNA and RNA models to predict order, family, and genus by using test datasets that contain one species genome from each genus. From the pilot dataset, 231 DNA and 142 RNA genomes were randomly selected from each genus to generate test datasets and we repeated the testing process 10 times. The average accuracies were 99% (order), 98% (family) and 95% (genus) for the DNA viruses and 97% (order), 96% (family) and 82% (genus) for the RNA viruses (Fig. 2c). Because of fewer branches and larger sample sizes in the higher taxonomic levels, order level accuracies were highest in both models and the accuracies dropped gradually from order to genus level. To build the final prediction models using the entire RefSeq complete virus genomes (RNA = 2529; DNA = 4421), we used k-mer lengths of 21 bp (DNA viruses), 17 bp (RNA viruses) and 20 bp (combining DNA and RNA viruses) with MFT value of 1 using the entire dataset **(see methods)**. Lone taxonomic ranks with only one order, family or genus but with more

Table 1
Virus sequences detected by different methods in metagenomic data.

Libraries	# contigs	VirusTaxo	CLARK (virus db)	Kraken2 (MiniKraken2 db)	Kraken2 (virus db)	DeepVirFinder (DVF)	Overlap between VirusTaxo and DVF
SRR10281034	122,545	39,856 (32.52%)	1268 (1.03%)	18 (0.01%)	117 (0.10%)	38,652 (31.54%)	10,232
SRR10281038	100,894	35,713 (35.4%)	773 (0.77%)	42 (0.04%)	522 (0.52%)	33,231 (32.94%)	13,105
SRR12756394	196,745	55,874 (28.4%)	746 (0.38%)	115 (0.06%)	519 (0.26%)	66,548 (33.82%)	21,750
SRR10971381	23,720	5927 (24.99%)	51 (0.22%)	5 (0.02%)	47 (0.20%)	7486 (31.56%)	1806

Metagenomic contigs are classified using different methods where db means the database.

Table 2
Computational performance of VirusTaxo and other tools.

Tools	Database size	Peak RAM usage	Running time
VirusTaxo	4.6 GB	24.5 GB	4 m 29.423 s
Kraken2 (MiniKraken2)	8 GB	5.9 GB	0 m 9.600 s
Kraken2 (virus)	496.5 MB	3.1 GB	0 m 8.989 s
CLARK	79.1 GB	13.8 GB	0 m 39.458 s
DeepVirFinder	Null	3.2 GB	1086 m 8.795 s

Metavirome dataset (SRR10281034) containing 122,545 contigs was used for benchmarking computational performance.

than one species genomes were also included for the genus level predictions. The singletons with one species genome per genus were removed. VirusTaxo estimates the specificity of prediction using normalized entropy of probability distribution across taxonomic ranks. Higher entropy (>0.5) is considered as undetected and lower entropy (≤0.5) is used to provide level of certainty at the genus level prediction. In the final prediction models, accuracy increases with entire RefSeq data, 1.4% increase in DNA and 8.38% in RNA model compared to the models trained using pilot data because of including more genomes in the training. The final model was trained on 402 DNA and 280 RNA virus genera and have an average accuracy of 90.38% (RNA), 96.2% (DNA) and 92.5% (DNA and RNA) where on average 11.3% of sequences remained undetected. Average accuracy for each model was calculated by repeating the training (with 80% of sequences) and testing (with 20% of sequences) process five times.

2.3. Benchmarking of VirusTaxo using metagenomic datasets

We predicted the accuracies of VirusTaxo, CLARK and Kraken2 using three metavirome (SRR10281034, SRR10281038, SRR12756394) [23] and a metatranscriptome (SRR10971381) [6]. We assembled the metagenomic reads using MEGAHIT [24] following quality trimming by Trimmomatic [25]. The resultant metagenomic contigs were used to classify virus sequences by different methods (Table 1).

VirusTaxo's combined database with DNA and RNA sequences assigned taxonomic ranks of >30% of the metagenomic contigs which is an about 100-fold increase in assigning taxonomy compared to CLARK (0.60%), Kraken2 with MiniKraken2 database (0.033%), and Kraken2 with virus database (0.27%) (Table 1). DeepVirFinder does not classify taxonomy but predicts virus sequences. DeepVirFinder identified >30% of the metagenomic contigs as virus sequences across four metagenomic libraries. A significant portion of the sequences classified by VirusTaxo was also predicted as virus sequence in DeepVirFinder (VirusTaxo = 137,370, DeepVirFinder = 145,917 and intersect = 46,893; Fisher's exact test *p*-value = 2.08853e-33). This is suggesting that a large proportion of virus sequences were not assigned to their taxonomy by CLARK and Kraken2 despite using their latest database **(see Methods)**. By default, CLARK (k-mer = 31 bp) and Kraken2 (k-mer = 35 bp) use larger k-mer sizes whereas VirusTaxo used an optimized k-mer length of 20 bp for its combined model with RNA and DNA sequences. VirusTaxo assigned taxonomic ranks of significantly higher numbers of the contigs from metaviromes and metatranscriptomes.

2.4. Benchmarking of computational performance of VirusTaxo

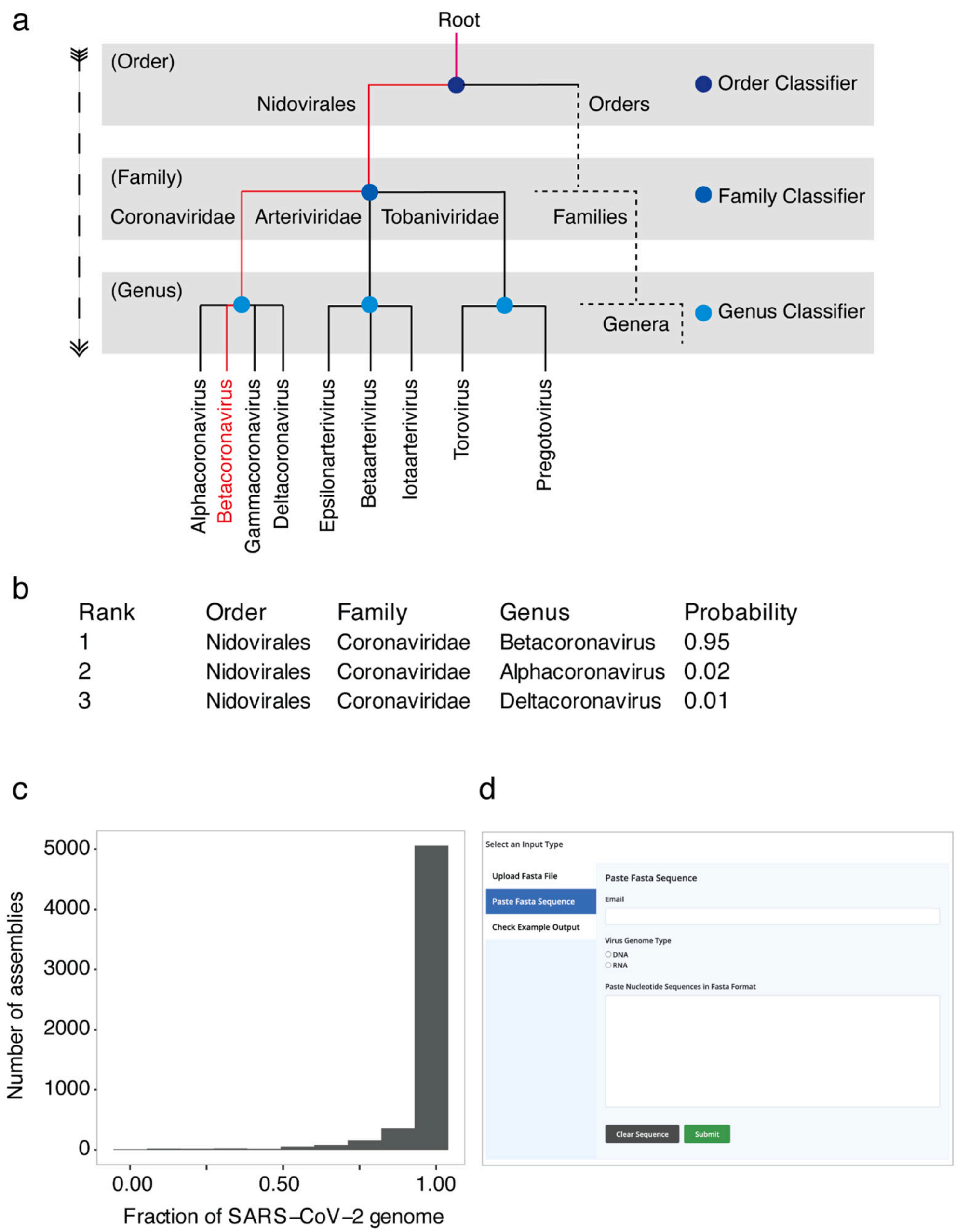
For calculating the Central Processing Unit (CPU) time consumption and Random Access Memory (RAM) usage, we used SRR10281034 metavirome library which has 122,545 contigs. We utilized a single thread on a dedicated computer for all the methods **(see Methods)**. Program running time is represented in wallclock CPU seconds (Table 2). Compared to CLARK and Kraken2, VirusTaxo has a smaller database size of 4.6 GB but requires higher running time and RAM usage. DeepVirFinder took significantly much longer time (18 h) to finish the prediction.

2.5. Predicting hierarchial taxonomy of SARS-CoV-2 from metagenomic assembly using VirusTaxo

SARS-CoV-2 belongs to *Betacoronavirus* genus, *Coronaviridae* family and *Nidovirales* order. **(**Fig. 3a**)** illustrated the taxonomic ranks of SARS-CoV-2 and its hierarchical taxonomic classification by VirusTaxo. The reference genome (MN908947.3) of SARS-CoV-2 was generated from SRR10971381 sequencing library and assembled by MEGAHIT [24] to identify the family of this novel virus that caused the recent pandemic [6]. We downloaded the SRR10971381 library and assembled it with MEGAHIT using the approach described here [6]. The longest contig generated by MEGAHIT was 29,868 bp long and was used as a query sequence in VirusTaxo. To treat SARS-CoV-2 as a novel virus species, we removed its genome from our training dataset to train the RNA model of VirusTaxo. VirusTaxo model predicted the 29,868 bp MEGAHIT contig belongs to *Nidovirales* order, *Coronaviridae* family, *Betacoronavirus* genus (Rank: 1, Entropy: 0.07, Softmax probability: 0.95) **(**Fig. 3b**)**. Accurate assignment of order, family and the ranking of the closest genus by VirusTaxo indicating that the whole process of identifying taxonomy of novel or uncharacterized viruses can be automated without the need for sequence alignment and human interpretation of alignment data given that close relatives of the uncharacterized viruses are present in the database.

2.6. The effect of contig length in VirusTaxo classification

We obtained 6176 de novo assemblies of SARS-CoV-2 genome that were assembled using eight different assemblers [26]. We used BLASTn [27] against a database made of MN908947.3 sequence to obtain the SARS-CoV-2 contigs and selected the largest contig per assembly. These sequences contain full and partial genome assemblies and have diverse variants due to differences in virus samples and assemblers. This dataset contains partial genome assemblies with minimum contig length of 30 bp and 4536 assemblies had <75% of the genome constructed **(**Fig. 3c**)**. Despite the partial genomes provided and variants present in those sequences, VirusTaxo RNA model correctly predicted *Nidovirales* as the order, *Coronaviridae* as the family, and *Betacoronavirus* as the genus for all of the assemblies. In comparison to VirusTaxo, CLARK detected 99.77% (6162/6176) of the contigs and Kraken2 with MiniKraken2 database detected 3521 (57.01%), and Kraken2 with virus database detected 6168 (99.87%) contigs as *Betacoronavirus*. Thus VirusTaxo, CLARK, and Kraken2 that were trained on the virus genomes detected the *Betacoronavirus* contigs and were not affected by contig lengths. For



Rank	Order	Family	Genus	Probability
1	Nidovirales	Coronaviridae	Betacoronavirus	0.95
2	Nidovirales	Coronaviridae	Alphacoronavirus	0.02
3	Nidovirales	Coronaviridae	Deltacoronavirus	0.01

Fig. 3. Benchmarking of VirusTaxo for SARS-CoV-2 genomes. **a)** Schematic representation of hierarchical prediction of taxonomic ranks of SARS-CoV-2 genome using VirusTaxo. VirusTaxo classified the taxonomic ranks of SARS-CoV-2 for its order, family and genus which are highlighted in red color. **b)** Ranking by Softmax probability at genus level prediction for the SRR10971381 assembly. **c)** Distribution of fraction of genome assembled in 5793 assemblies of SARS-CoV-2 genome. **d)** Screenshot of VirusTaxo web interface (https://omics-lab.com/virustaxo). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the prediction of the taxonomic ranks online using VirusTaxo, a web application has been provided with trained models for virus classification (Fig. 3d).

3. Discussion

The International Committee on Taxonomy of Viruses (ICTV) classifies viruses into their taxonomic ranks primarily based on phenotypic properties [28]. However, the ICTV has continually updated its approach to virus taxonomy through incorporation of newer technologies including genome sequence as a property needed for classification [29]. Genomic sequences of viruses show their evolutionary relationships and provide an opportunity to detect virus taxonomy especially for those that lack phenotypic data [3]. Developing and utilizing automated computational methods will facilitate the taxonomic assignment of novel or uncharacterized viruses efficiently and will open the possibility to discover new taxa solely based on the genomic sequences. Supervised machine learning methods can learn from the patterns of existing virus genomes and their taxonomic ranks to assign taxa of novel viruses automatically. Here we proposed VirusTaxo, a machine learning architecture to classify taxonomic ranks (e.g., order, family and genus) using virus genome. Virus taxonomic tree is hierarchically structured with taxonomic ranks at different levels which is challenging for the classifiers to maintain the accuracy towards the low-level taxa. K-mer features of DNA have been shown to contain information about sequence composition and sequence evolution [30,31]. Using k-mer features VirusTaxo obtained >93% overall accuracy in classification at each taxonomic rank. We have shown that RNA and DNA virus classification parameters (e.g., k-mer length) could be different because these two sequence sets are different in their size and composition. Viruses have some exceptions in taxonomic classification by genome sequence and are not always congruent between phenotypic and evolutionary approaches [28]. Despite the challenges in classifying viruses from genome, VirusTaxo showed significant improvement in predicting smaller contigs and classifying taxonomy of more virus sequences from metagenomic datasets compared to other state-of-the-art methods e.g.,

Table 3
Summary of RNA and DNA virus genome sequences.

	DNA genomes	RNA genomes
Family	46	83
Genus	402	280
Total species genomes	4421	2529

CLARK and Kraken2.

High-throughput sequencing of metagenomes or metaviromes can identify the true diversity of viruses in a particular environment sample. Metagenomic sequence assembly creates full or partial genomes of thousands of new viruses that when classified, will contribute to the formation of new virus taxa. Large scale metagenomic studies showed that the vast majority of the identified viruses were unrelated to those in known viruses [32]. Novel viruses that do not have close relationships at the genome sequences with existing taxa pose a particular problem to classify their taxonomic rank using supervised machine learning methods. In such a scenario, the taxonomic assignment could be arbitrary and therefore we introduced probabilistic ranking to assess the certainty using softmax and entropy scores in VirusTaxo. Since virus taxonomy evolves with the addition of new viruses, taxonomy classification methods can be updated over time with new sequences and their taxa. With an increased understanding of newer taxa, machine learning methods can be scaled to learn about those genomes and classify unknown viruses. Overall, our classification results obtained from metagenomic data and SARS-CoV-2 genome assemblies indicate that VirusTaxo is readily capable of distinguishing between viruses belonging to different classes of taxa.

4. Methods

4.1. Datasets

4.1.1. Pilot dataset

RefSeq genomes of all RNA and DNA viruses were downloaded from the NCBI virus database [33]. Taxonomic classification of the viruses was obtained from the International Committee on Taxonomy of Viruses, ICTV Master Species List 2019.v1 release [34]. We chose orders with at least two families, families with at least two genera, and genera with at least three species to ensure sufficient genomes for the RNA and DNA virus classifier models. Pilot dataset contains 2561 DNA and 1480 RNA virus genomes which were used to train the VirusTaxo models that belong to 231 DNA and 142 RNA virus genera.

4.1.2. Entire RefSeq dataset of viruses

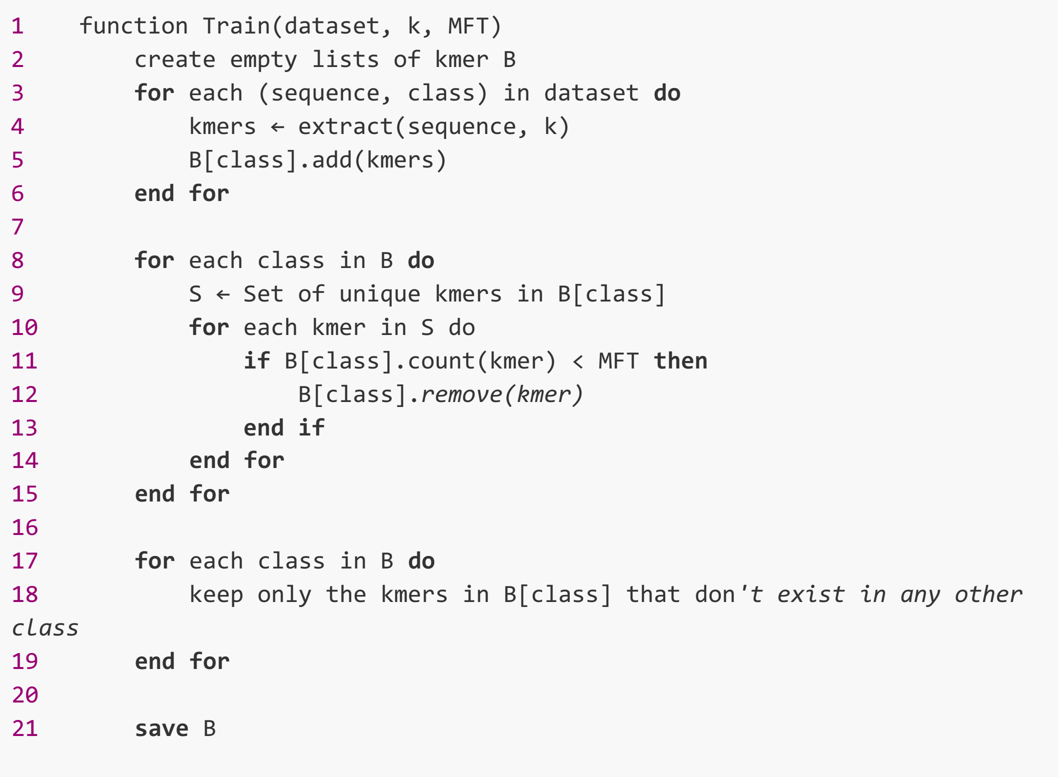
We used all complete virus genomes from NCBI RefSeq consisting of 4421 DNA and 2529 RNA virus genomes. Singletons with one sequence per genus were removed. The summary of the selected datasets is listed in (Table 3**)**. The detailed taxonomic information of the viruses used in the final model is added in the **Supplementary file**.

4.1.3. Metagenomic datasets

Metavirome (SRR10281034, SRR10281038, SRR12756394) [23], SARS-CoV-2 metatranscriptome (SRR10971381) [6] datasets were used in benchmarking with metagenomic classifiers.

4.2. Hierarchical classification architecture of VirusTaxo

VirusTaxo uses a top to bottom approach for the classification of order, family and genus of a virus sequence. For *m* and *n* number of order and family respectively, there will be *m* numbers of family classifiers under respective orders in the 2nd layer and *n* numbers of genus classifiers under respective families in the 3rd layer. There will be a total number of classifiers $= 1 + m + n$ in each model. In addition, taxonomic

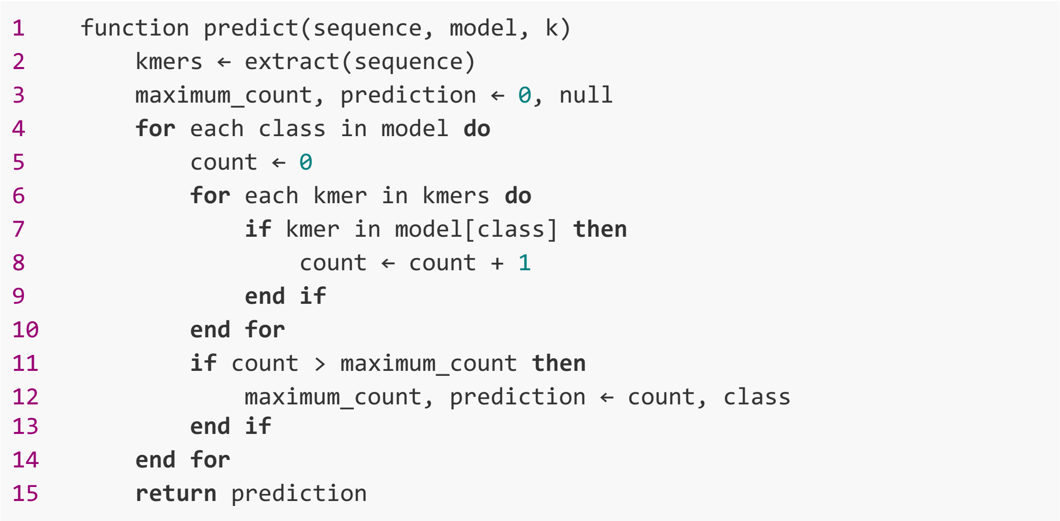


```
function Train(dataset, k, MFT)
    create empty lists of kmer B
    for each (sequence, class) in dataset do
        kmers ← extract(sequence, k)
        B[class].add(kmers)
    end for

    for each class in B do
        S ← Set of unique kmers in B[class]
        for each kmer in S do
            if B[class].count(kmer) < MFT then
                B[class].remove(kmer)
            end if
        end for
    end for

    for each class in B do
        keep only the kmers in B[class] that don't exist in any other class
    end for

    save B
```



```
function predict(sequence, model, k)
    kmers ← extract(sequence)
    maximum_count, prediction ← 0, null
    for each class in model do
        count ← 0
        for each kmer in kmers do
            if kmer in model[class] then
                count ← count + 1
            end if
        end for
        if count > maximum_count then
            maximum_count, prediction ← count, class
        end if
    end for
    return prediction
```

ranks with only one order, family or genus were also included for genus level prediction. We trained the classifiers at different levels of the tree by utilizing Breadth First Search (BFS) [35] graph traversal algorithm for both DNA and RNA datasets. **(**Fig. 1**)** illustrates an example of hierarchical classification by VirusTaxo where the order classifier (OC) in the root is classifying the genomes between two orders (e.g., Order-1 and Order-2). Then all the genomes in an order are split into corresponding families to train the family level models. Two family classifiers (e.g., FC-1 and FC-2) that belong to 2 orders classifying 5 families. Similarly, 5 genus classifiers were built to classify the genomes into 17 genera.

4.3. Hierarchical prediction of taxonomic ranks

We pass a genome sequence through the order classifier and we get the decision for an order. Then we pass it through the family classifier under the predicted order. Finally to get the genus, we go to the genus classifier under the predicted family.

4.3.1. Training algorithm

A dataset and two parameters are required to train the VirusTaxo multi-class classification model. These parameters are k and the minimum frequency threshold (MFT) of k-mers. Here are the training steps:

1. Create empty bags for each class.
2. Iterate over each sequence in the dataset and follow the steps mentioned below.
   a. Generate k-mers by extracting substrings of k bp with k-1 bp overlaps from the sequence.
   b. Add extracted k-mers to a bag in accordance with the class.
3. Discard the k-mers from each bag if the frequencies of the k-mers are less than MFT.
4. Keep only the k-mers in each bag that don’t occur in any other bag to build discriminative (mutually exclusive) bags.
5. Save the bags as a model.

Training pseudocode

4.3.2. Prediction algorithm

For a given input sequence, model and k, the following steps are performed to predict the rank and class.

1. Generate k-mers (same k-mer generation technique that was described in the training part) from the given input sequence.
2. Count the overlap of k-mers between the input sequence derived k-mers and bag of k-mers for each class.
3. Predict the corresponding class as an output for which the overlap count is maximum.

Prediction pseudocode

4.3.3. Determination of confidence level of genus prediction

Firstly, we find the overlap counts of the input sequence with every genus. Suppose, there are n genus and x_(1), x_(2), …, x_(n) denote the overlap count with n genus. We find the probability distribution using x_(1), x_(2), …, x_(n) with the help of the softmax function. The formula of softmax function is given below:

$$p(x_i) = \frac{e^{x_i}}{\sum_{k=1}^{n} e^{x_k}}$$

We rank genus according to the descending value of probabilities. After getting the probability values, we are interested to find how much of our first ranking is reliable. For this, we calculate normalized entropy value which ranges within [0.0,1.0]. The formula of normalized entropy is given value:

$$normalized\ entropy = -\sum_{i=1}^{n} \frac{p(x_i) log_2(p(x_i))}{log_2(n)}$$

When the probabilities are more or less equally distributed across genera, then it is difficult to predict a specific genus. In that scenario, the normalized entropy value will be close to 1.0. On the other hand, if the probabilities distribution for one genus is high and for the rest of all are low, then the entropy value will be close to 0.0.

4.4. Benchmarking of VirusTaxo using metagenomic classifiers

We used CLARK (v1.2.6.1), Kraken2 (v2.1.2) and DeepVirFinder (v.1) to predict the taxonomy from metagenomics datasets (Table S2). CLARK was used against the virus database with ‘clarkdb viruses’. For Kraken2, ‘minikraken2_v2_8GB_201904_UPDATE’ and ‘Viral; 5/17/2021’ database were used. Those tools were run using default parameters, unless otherwise mentioned. For DeepVirFinder, score $>$ 90 and the p-value $<$ 0.05 is used to detect virus sequences.

4.5. Benchmarking of machine learning methods

We compared the performances of VirusTaxo with four other algorithms on RNA and DNA virus datasets. We used word2vec encoding

Table 4
Hyperparameters setup for word2vec training.

Hyperparameters	Value
Method (skip-gram / CBOW)	Skip-gram
Dimension	300
Learning rate	0.025

Table 5
Hyperparameters for four algorithms used to perform benchmarking.

Algorithm	Hyperparameters
Multilayer perceptron	Hidden layer: 1 No of nodes in hidden layer: 100 Optimizer: Stochastic gradient descent Learning rate: 0.01
Random forest	No of trees: 100 Metric: Gini impurity
Gradient boosting	Learning rate: 0.01 Loss function: Deviance
K-nearest neighbors	K: 5

[36] for transforming genome sequences into vectors. To train two word2vec models for DNA and RNA datasets, we generated a stream of k-mers without changing sequence chronology of k-mers from each genome sequence taking 21 bp and 17 bp k-mer lengths respectively. We trained word2vec models using fastText [36] using the following hyperparameters in (Table 4). After completion of word2vec training, we utilize four algorithms (Multilayer perceptron, Random forest, Gradient boosting, KNN) one by one in hierarchical classification of RNA virus and DNA virus. The hyperparameter details of the four algorithms are in (Table 5). Here we randomly choose one species genome from each genus to create the test set.

4.6. *Analysis of metagenomic data*

We have downloaded the Metavirome (SRR10281034, SRR10281038, SRR12756394), SARS-CoV-2 metatranscriptome (SRR10971381) reads from NCBI GEO datasets. Sequencing reads were adaptor and quality trimmed using the Trimmomatic program (v.0.39) [25]. The remaining reads were assembled de novo using MEGAHIT (v.1.2.9) with default parameters. The contigs were used to predict virus taxonomy.

4.7. *CPU and RAM usage*

We used the same procedure described by [26] to measure CPU time and peak RAM usage. We used 32 cores (64 threads) to compare taxonomy classification tools in a AMD EPYC 7502P 2.5 Ghz, 32 cores, 256 GB RAM.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Software availability

Web-based application of VirusTaxo is available at https://omics-lab.com/virustaxo

Source code and dataset are available on Github at https://github.com/omics-lab/VirusTaxo

CRediT authorship contribution statement

Rajan Saha Raju: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing. **Abdullah Al Nahid:** Software, Formal analysis, Investigation, Data curation, Writing – original draft. **Preonath Shuvo:** Conceptualization, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing. **Rashedul Islam:** Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision.

Declaration of Competing Interest

The authors declare that they have no competing interests.

Acknowledgments

We acknowledge Omics-Lab (https://omics-lab.com/) for providing the research platform. We also thank Yusha Araf for his responsibilities on management and coordination of the research team.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2022.110414.

References

[1] K.V. Chaitanya, Structure and organization of virus genomes, Genome Genom. (2019) 1–30, https://doi.org/10.1007/978-981-15-0702-1_1.

[2] C.M. Fauquet, Taxonomy, classification and nomenclature of viruses, Encycl. Virol. (1999) 1730–1756, https://doi.org/10.1006/rwvi.1999.0277.

[3] P. Simmonds, M.J. Adams, M. Benkő, M. Breitbart, J.R. Brister, E.B. Carstens, A. J. Davison, E. Delwart, A.E. Gorbalenya, B. Harrach, et al., Virus taxonomy in the age of metagenomics, Nat. Rev. Microbiol. 15 (2017) 161–168, https://doi.org/10.1038/nrmicro.2016.177.

[4] P. Aiewsakun, P. Simmonds, The genomic underpinnings of eukaryotic virus taxonomy: creating a sequence-based framework for family-level virus classification, Microbiome 6 (2018) 38, https://doi.org/10.1186/s40168-018-0422-7.

[5] H. Guan, A. Shen, X. Lv, X. Yang, H. Ren, Y. Zhao, Y. Zhang, Y. Gong, P. Ni, H. Wu, et al., Detection of virus in CSF from the cases with meningoencephalitis by next-generation sequencing, J. Neuro-Oncol. 22 (2016) 240–245, https://doi.org/10.1007/s13365-015-0390-7.

[6] F. Wu, S. Zhao, B. Yu, Y.-M. Chen, W. Wang, Z.-G. Song, Y. Hu, Z.-W. Tao, J.-H. Tian, Y.-Y. Pei, et al., A new coronavirus associated with human respiratory disease in China, Nature 579 (2020) 265–269, https://doi.org/10.1038/s41586-020-2008-3.

[7] N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang, J. Song, X. Zhao, B. Huang, W. Shi, R. Lu, et al., A novel coronavirus from patients with pneumonia in China, 2019, N. Engl. J. Med. 382 (2020) 727–733, https://doi.org/10.1056/NEJMoa2001017.

[8] T.J. Dougan, S.R. Quake, Viral taxonomy derived from evolutionary genome relationships, PLoS One 14 (2019), https://doi.org/10.1371/journal.pone.0220440 e0220440.

[9] B.M. Muhire, A. Varsani, D.P. Martin, SDT: a virus classification tool based on pairwise sequence alignment and identity calculation, PLoS One 9 (2014), https://doi.org/10.1371/journal.pone.0108277 e108277.

[10] S. Roux, F. Enault, B.L. Hurwitz, M.B. Sullivan, VirSorter: mining viral signal from microbial genomic data, PeerJ 3 (2015), https://doi.org/10.7717/peerj.985 e985.

[11] M. Vilsker, Y. Moosa, S. Nooij, V. Fonseca, Y. Ghysens, K. Dumon, R. Pauwels, L. C. Alcantara, E. Vanden Eynden, A.-M. Vandamme, et al., Genome detective: an automated system for virus identification from high-throughput sequencing data, Bioinforma. Oxf. Engl. 35 (2019) 871–873, https://doi.org/10.1093/bioinformatics/bty695.

[12] A.L. Bazinet, M.P. Cummings, A comparative evaluation of sequence classification programs, BMC Bioinforma. 13 (2012) 92, https://doi.org/10.1186/1471-2105-13-92.

[13] M.A. Remita, A. Halioui, A.A. Malick Diouara, B. Daigle, G. Kiani, A.B. Diallo, A machine learning approach for viral genome classification, BMC Bioinforma. 18 (2017), https://doi.org/10.1186/s12859-017-1602-3.

[14] J. Ren, N.A. Ahlgren, Y.Y. Lu, J.A. Fuhrman, F. Sun, VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data, Microbiome 5 (2017) 69, https://doi.org/10.1186/s40168-017-0283-5.

[15] J. Ren, K. Song, C. Deng, N.A. Ahlgren, J.A. Fuhrman, Y. Li, X. Xie, R. Poplin, F. Sun, Identifying viruses from metagenomic data using deep learning, Quant. Biol. 8 (2020) 64–77, https://doi.org/10.1007/s40484-019-0187-4.

[16] D.E. Wood, S.L. Salzberg, Kraken: ultrafast metagenomic sequence classification using exact alignments, Genome Biol. 15 (2014) R46, https://doi.org/10.1186/gb-2014-15-3-r46.

[17] D.E. Wood, J. Lu, B. Langmead, Improved metagenomic analysis with Kraken 2, Genome Biol. 20 (2019) 257, https://doi.org/10.1186/s13059-019-1891-0.

[18] F.P. Breitwieser, D.N. Baker, S.L. Salzberg, KrakenUniq: confident and fast metagenomics classification using unique k-mer counts, Genome Biol. 19 (2018) 198, https://doi.org/10.1186/s13059-018-1568-0.

[19] R. Ounit, S. Wanamaker, T.J. Close, S. Lonardi, CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers, BMC Genomics 16 (2015) 236, https://doi.org/10.1186/s12864-015-1419-2.

[20] R. Ounit, S. Lonardi, Higher classification sensitivity of short metagenomic reads with CLARK-S, Bioinforma. Oxf. Engl. 32 (2016) 3823–3825, https://doi.org/10.1093/bioinformatics/btw542.

[21] N. Segata, L. Waldron, A. Ballarini, V. Narasimhan, O. Jousson, C. Huttenhower, Metagenomic microbial community profiling using unique clade-specific marker genes, Nat. Methods 9 (2012) 811–814, https://doi.org/10.1038/nmeth.2066.

[22] C.J. Houldcroft, M.A. Beale, J. Breuer, Clinical and biological insights from viral genome sequencing, Nat. Rev. Microbiol. 15 (2017) 183–192, https://doi.org/10.1038/nrmicro.2016.182.

[23] T.V. Butina, I.V. Khanaev, L.S. Kravtsova, O.O. Maikova, Y.S. Bukin, Metavirome datasets from two endemic Baikal sponges Baikalospongia bacillifera, Data Brief 29 (2020) 105260, https://doi.org/10.1016/j.dib.2020.105260.

[24] D. Li, C.-M. Liu, R. Luo, K. Sadakane, T.-W. Lam, MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph, Bioinforma. Oxf. Engl. 31 (2015) 1674–1676, https://doi.org/10.1093/bioinformatics/btv033.

[25] A.M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics 30 (2014) 2114–2120, https://doi.org/10.1093/bioinformatics/btu170.

[26] R. Islam, R.S. Raju, N. Tasnim, I.H. Shihab, M.A. Bhuiyan, Y. Araf, T. Islam, Choice of assemblers has a critical impact on de novo assembly of SARS-CoV-2 genome and characterizing variants, Brief. Bioinform. (2021), https://doi.org/10.1093/bib/bbab102.

[27] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410, https://doi.org/10.1016/S0022-2836(05)80360-2.

[28] P. Simmonds, Methods for virus classification and the challenge of incorporating metagenomic sequence data, J. Gen. Virol. 96 (2015) 1193–1206, https://doi.org/10.1099/jgv.0.000016.

[29] P. Simmonds, M.J. Adams, M. Benkő, M. Breitbart, J.R. Brister, E.B. Carstens, A. J. Davison, E. Delwart, A.E. Gorbalenya, B. Harrach, et al., Consensus statement: virus taxonomy in the age of metagenomics, Nat. Rev. Microbiol. 15 (2017) 161–168, https://doi.org/10.1038/nrmicro.2016.177.

[30] S. Röhling, A. Linne, J. Schellhorn, M. Hosseini, T. Dencker, B. Morgenstern, The number of k-mer matches between two DNA sequences as a function of k and applications to estimate phylogenetic distances, PLoS One 15 (2020), https://doi.org/10.1371/journal.pone.0228070 e0228070.

[31] Z. Yang, H. Li, Y. Jia, Y. Zheng, H. Meng, T. Bao, X. Li, L. Luo, Intrinsic laws of k-mer spectra of genome sequences and evolution mechanism of genomes, BMC Evol. Biol. 20 (2020), https://doi.org/10.1186/s12862-020-01723-3.

[32] D. Paez-Espino, E.A. Eloe-Fadrosh, G.A. Pavlopoulos, A.D. Thomas, M. Huntemann, N. Mikhailova, E. Rubin, N.N. Ivanova, N.C. Kyrpides, Uncovering Earth's virome, Nature 536 (2016) 425–430, https://doi.org/10.1038/nature19094.

[33] J.R. Brister, D. Ako-Adjei, Y. Bao, O. Blinkova, NCBI viral genomes resource, Nucleic Acids Res. 43 (2015) D571–D577, https://doi.org/10.1093/nar/gku1207.

[34] M.J. Adams, E.J. Lefkowitz, A.M.Q. King, B. Harrach, R.L. Harrison, N.J. Knowles, A.M. Kropinski, M. Krupovic, J.H. Kuhn, A.R. Mushegian, et al., Changes to taxonomy and the international code of virus classification and nomenclature ratified by the International Committee on Taxonomy of Viruses (2017), Arch. Virol. 162 (2017) 2505–2538, https://doi.org/10.1007/s00705-017-3358-5.

[35] Edward F. Moore, The shortest path through a maze, in: Proceedings of the International Symposium on the Theory of Switching, Harvard University Press, 1959, pp. 285–292.

[36] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching Word Vectors with Subword Information, 2017. ArXiv160704606 Cs.