NAR Genomics and Bioinformatics, 2024, **6**, lqae044
https://doi.org/10.1093/nargab/lqae044
Advance access publication date: 6 May 2024
Standard Article

DeepMicroClass sorts metagenomic contigs into prokaryotes, eukaryotes and viruses

Shengwei Hou [1,2,*,†], Tianqi Tang [3,†], Siliangyu Cheng [3,†], Yuanhao Liu [1], Tian Xia [1], Ting Chen [4], Jed A. Fuhrman [2] and Fengzhu Sun [3,*]

[1]Department of Ocean Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China
[2]Marine and Environmental Biology, Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089, USA
[3]Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA 90089, USA
[4]Department of Computer Science and Technology, Institute of Artificial Intelligence & BNRist, Tsinghua University, Beijing 100084, China

[*]To whom correspondence should be addressed. Tel: +86 0755 8801 0164; Email: housw@sustech.edu.cn
Correspondence may also be addressed to Fengzhu Sun. Tel: +1 213 740 2413; Email: fsun@usc.edu
[†]The first three authors should be regarded as Joint First Authors.

Abstract

Sequence classification facilitates a fundamental understanding of the structure of microbial communities. Binary metagenomic sequence classifiers are insufficient because environmental metagenomes are typically derived from multiple sequence sources. Here we introduce a deep-learning based sequence classifier, DeepMicroClass, that classifies metagenomic contigs into five sequence classes, i.e. viruses infecting prokaryotic or eukaryotic hosts, eukaryotic or prokaryotic chromosomes, and prokaryotic plasmids. DeepMicroClass achieved high performance for all sequence classes at various tested sequence lengths ranging from 500 bp to 100 kbps. By benchmarking on a synthetic dataset with variable sequence class composition, we showed that DeepMicroClass obtained better performance for eukaryotic, plasmid and viral contig classification than other state-of-the-art predictors. DeepMicroClass achieved comparable performance on viral sequence classification with geNomad and VirSorter2 when benchmarked on the CAMI II marine dataset. Using a coastal daily time-series metagenomic dataset as a case study, we showed that microbial eukaryotes and prokaryotic viruses are integral to microbial communities. By analyzing monthly metagenomes collected at HOT and BATS, we found relatively higher viral read proportions in the subsurface layer in late summer, consistent with the seasonal viral infection patterns prevalent in these areas. We expect DeepMicroClass will promote metagenomic studies of under-appreciated sequence types.

Introduction

Microbes play an essential role in the biogeochemical cycling of elements in diverse ecosystems on the planet (1,2). Microbial communities are a collection of diverse biological entities, including ribosome-encoding cellular organisms, capsid-encoding organisms (i.e., viruses) that can only reproduce within cellular organisms, and orphan replicons (plasmids, transposons, etc.) that parasitize various life forms for propagation (3). Viruses and plasmids are extrachromosomal genetic elements that have important implications for the diversity and function of microbial communities owing to their roles in transferring genetic materials between or within microbes. Thus, together with transposable elements, they are collectively referred to as mobile genetic elements (MGEs). Microbial community diversity can range from a consortium of several dominant strains to a conglomerate of thousands of species, depending on where, when and how metagenomic samples were collected. Thanks to the discovery of the small subunit rRNA gene (SSU) as a universally conserved phylogenetic marker (4), the biodiversity of environmental microbial communities can be easily assessed using the SSU-based amplicon surveys (5,6). Microbial coding potentials and genomic elements can be further probed using cloning libraries of natural microbial assemblages (e.g., cosmid and fosmid libraries) (6–12), which have been transformed by shotgun metagenomes to infer genome-scale functional capabilities of uncultured microbes (13,14). The rapid expansion of metagenomic datasets presents opportunities and challenges. Metagenomics enables the exploration of complex microbial interactions and genetic evolution of individual species (15,16). On the other hand, efficient and reliable retrieval of microbial genomes and MGEs from metagenomic sequence pools requires strategic approaches.

By categorizing metagenomic contigs, consecutive sequences assembled from metagenomic reads, into distinct groups, the complexity of metagenomes can be reduced to certain taxonomic levels, from coarse domains to consensus species or strains. Metagenomic tools to classify contigs can be broadly framed into two categories, supervised contig classification tools (i.e., viral contig predictors) and unsupervised contig clustering tools (i.e. metagenomic binners, see (17) for a review of binning strategies). Metagenomic contig classification has been heavily focused on predicting viral sequences. Viruses are prevalent in aquatic, soil and host-associated systems, and are presumably the most numerous biological entities on Earth (18,19). In marine systems, viral lysis is crucial in redirecting carbon and energy flow to the lower trophic levels (termed 'Viral Shunt'), which has great implications for

Received: September 7, 2023. **Revised:** March 18, 2024. **Editorial Decision:** April 15, 2024. **Accepted:** April 18, 2024

the global biogeochemical cycles (20,21). VirSorter (22) and VirFinder (23) are two pioneer tools to identify viral contigs from metagenomic assemblies. VirSorter predicts viral contigs based on viral signals and categorizes them into three tiers with different confidence levels. VirFinder employs *k*-mer frequencies and a logistic regression model to classify contigs to either viral or host sequences, which outperforms VirSorter for shorter contigs without detectable viral hallmark genes (23). The success of *k*-mer based methods has inspired the application of deep learning in viral sequence discovery, which led to the development of DeepVirFinder (24) and PPR-Meta (25), both of which use one-hot encoding to convert DNA sequences into presence/absence matrices of nucleotides, and use neural networks to train virus-host classifiers at different contig lengths. Besides, PPR-Meta combines both nucleotide path and codon path in the encoding step, and classifies contigs into viruses, host chromosomes and plasmids (25). VIBRANT (26) uses neural networks to distinguish prokaryotic dsDNA, ssDNA and RNA viruses based on 'v-score' metrics, which are calculated from significant protein hits to a collection of Hidden Markov Model (HMM) profiles derived from public databases. The latest VirSorter2 (27) utilizes a multi-classifier design and expands viral identification to under-represented viral groups, such as RNA viruses, giant viruses and virophages. Plasmids are another major type of MGEs heavily studied in environmental microbiome, particularly in host-associated systems or wastewater treatment plants. Via transferring among hosts or exchanging genes with their host genomes, plasmids facilitate the host acquisition of new traits (28). Thus, by carrying genes related to resource utilization, antibiotic, metal resistance, and defense systems, plasmids contribute to the genetic and phenotypic plasticity of hosts, and increase their fitness to changing environments. PPR-Meta (25) and geNomad (29) can be used to simultaneously identify both viral and plasmid sequences from metagenomes in one run. The latter combines the gene-centric and deep-learning approaches, providing a framework for both sequence classification and gene annotation of viruses and plasmids. There are also multiple dedicated tools developed for plasmid identification, such as cBar (30), PlasFlow (31), PlaScope (32) and PlasClass (33). In principle, PlaScope employs a similarity searching approach based on species-specific databases, while cBar, PlasFlow and PlasClass use differential *k*-mer frequencies with different machine-learning methods.

Beyond viruses and plasmids, there is a paucity of applications towards the classification of eukaryotic contigs from metagenomes, though microbial eukaryotes are prevalent in diverse ecosystems such as host-associated habitats (34), deep-sea benthos (35), and geothermal springs (36), etc. Despite achievements in probing eukaryotic diversity using amplicon-based methods (37,38), or investigating metabolic potentials using genomic (39,40) and transcriptomic (41,42) methods, our knowledge is still limited by the availability of diverse microbial eukaryotic genomes (43). Alignment-based applications such as Kaiju (44) and MetaEuk (45) search for close matches in reference databases, thus can be used to assign reads or contigs to taxonomic groups. While the accuracy of these applications depends on the completeness of reference databases, their performance in classifying eukaryotic contigs is impaired due to the lack of a comprehensive microbial eukaryotic database (41). Eukaryotic sequences could also be identified using alignment-independent applications. EukRep (46) is a pioneer application that uses *k*-mer frequency and linear-SVM to classify metagenomic contigs into eukaryotic and prokaryotic sequences. Tiara (47) is a deep-learning based method used for eukaryotic sequence identification in metagenomes, and Whokaryote (48) is a random forest classifier that uses gene-structure based features to distinguish eukaryotic and prokaryotic sequences.

Despite the significant progress made in binary sequence classification, there isn't one tool that could classify eukaryotic/prokaryotic genomes, eukaryotic/prokaryotic viruses, and plasmids in one shot. Here we introduce DeepMicroClass, a versatile multi-class metagenomic contig classifier based on convolutional neural networks (CNN). DeepMicroClass is superior to the other tools by classifying all sequence types simultaneously, which will greatly reduce the time and computational resource usage compared to the conventional workflow of chaining a set of different predictors. We show that DeepMicroClass outperforms all the existing tools by accuracy and F1 score across 20 designed benchmark datasets with variable sequence-type composition. When benchmarked on the CAMI II marine dataset, DeepMicroClass showed comparable performance on viral prediction with geNomad and VirSorter2, but with a significantly reduced running time. Using a coastal marine metagenomic dataset as a case study, we showed that microbial eukaryotes and prokaryotic viruses contributed significantly to the metagenomic sequence pools. By analyzing monthly marine metagenomes of the open ocean, a dynamic viral abundance pattern in the subsurface layer could be detected. These results suggest the importance of accurately predicting different microbial sequence types for ecological studies. The implementation of DeepMicroClass and code for data analysis described in this paper can be accessed at https://github.com/chengsly/DeepMicroClass.

Materials and methods

Dataset preparation

We collected five classes of sequences: prokaryotic chromosomal, eukaryotic chromosomal, prokaryotic plasmid, prokaryotic viral and eukaryotic viral sequences. For prokaryotic chromosomal sequences, we downloaded all the prokaryotic genomes at the assembly level 'Chromosome' or 'Complete' from NCBI Genome database on 22 August 2022. The prokaryotic genomes were cleaned up by removing all sequences annotated as 'Plasmid' according to the assembly reports, and sequences not annotated as plasmids but have identical sequence IDs in the plasmid dataset were also removed. The resulting sample set contains 40,208 sequences. The eukaryotic sequence database includes genomic sequences from the eukaryotic taxa used by Kaiju (44) and the PR2 database (49). Specifically, we downloaded 612 microbial eukaryotic genomic sequences under taxa names: 'Amoebozoa', 'Apusozoa', 'Cryptophyceae', 'Euglenozoa', 'Stramenopiles', 'Alveolata', 'Rhizaria', 'Haptista', 'Heterolobosea', 'Metamonada', 'Rhodophyta', 'Chlorophyta' and 'Glaucocystophyceae' using genome_updater (available at https://github.com/pirovc/genome_updater) on 22 August 2022. In addition to these eukaryotic genomes, we also included 32 073 625 eukaryotic transcripts from the 678 marine eukaryotic transcriptomic re-assemblies (50) generated by the MMETSP project (41), which included 306 pelagic and endosymbiotic marine eukaryotic species representing >40 phyla.

Plasmid sequences and corresponding metadata were retrieved from PLSDB (51) released on 23 June 2021. The dataset contains 34 513 plasmid records. Viral sequences and associated metadata were retrieved from Virus-Host DB (52) released on 1 June 2022, which contains 17 357 nucleic acid records, including 5209 prokaryotic viruses and 12 148 eukaryotic viruses. In all downloaded sequences, we further cross compared sequence IDs in each class, and any sequence with an identical ID occurring in more than one class was removed to reduce potential erroneous taxonomy assignments from the source database.

Training, validation and test dataset preparation

Sequences were split into two parts according to the dates submitted to NCBI, using 1 January 2020 as a cutoff date. Sequences submitted before 1 January 2020 were used for training and validation, with 80% as training and 20% as validation using stratified split, where the split was conducted on each class separately, and the sequences submitted after this date were used for testing. The Mash (53) distance was used to estimate the similarity between sequences among training, validation and test sets. Sequences in the test set with a Mash distance <0.1 to any sequence in the training or validation sets were removed from the test set. Virus-Host DB derived viral sequences (52) and MMETSP derived eukaryotic sequences were not dated. These sequences were randomly split into training, validation and test sets with the proportions of 60%, 20% and 20%, respectively. Similarly, sequences were removed from the test set when the Mash distance to any sequence in the training or validation sets is less than 0.1.

Benchmark dataset preparation

To compare the performance of different predictors on classifying one specific sequence class or multiple sequence classes under different community composition scenarios, we designed 20 equisized (1000 contigs, each 10 kbps long) benchmark datasets with a variable composition of the 5 sequence classes. Briefly, the fractions of PROK (including prokaryotic genomes, prokaryotic viruses, and plasmids) to EUK (including eukaryotic genomes and eukaryotic viruses) sequences were determined using the ratios of 9:1, 7:3, 5:5, 3:7 and 1:9. Then for each fixed PROK:EUK ratio, the PROK fraction was further split into prokaryotic genomes, prokaryotic viruses and plasmids based on the ratios of 5:1:1, 4:1:1, 3:1:1 and 2:1:1; and the EUK fraction was further split into eukaryotic genomes and eukaryotic viruses according to the ratio of 5:1, 4:1, 3:1 and 2:1. Finally, the corresponding number of sequences were drawn from the test sequence pool for each class using the ratios specified above. The actual sequence source composition of the 20 benchmark datasets was shown in Supplementary Figure S1 and Supplementary Table S1 in the Supplementary Material. The CAMI II (Critical Assessment of Metagenome Interpretation II) dataset is a comprehensive resource designed for evaluating metagenomic software. It provides a set of simulated and real-world microbial communities with varying complexity, which are used to benchmark and improve metagenomic analysis tools. This dataset plays a crucial role in advancing the understanding and interpretation of microbial ecosystems by providing standardized challenges for algorithm development and comparison. In this study, we used the CAMI II marine dataset comprising prokaryotic and viral sequences to evaluate the performance of DeepMicroClass and other viral sequence classifiers.

Model design and training

DeepMicroClass employs a di-path convolutional neural network comprising a base-path and a codon-path to classify input sequences into one of the five classes. For the base-path, the input nucleotide sequence was firstly encoded as a one-hot matrix. Specifically, each of the A, C, G and T nucleotides was translated into [1, 0, 0, 0], [0, 1, 0, 0], [0, 0, 1, 0], [0, 0, 0, 1], respectively. Any non-ACGT nucleotide was represented with [0,0,0,0]. The reverse complementary strand of the input sequence is encoded by inverting the order of the rows in the forward one-hot matrix and then reversing the order of elements within each row. For the codon-path, the forward or reverse base-path matrix was first converted into three 64 dimensional one-hot matrices based on three reading frames, and then the three matrices were concatenated into one matrix. Thus, for each strand of an input contig, a di-path incorporating both the base and codon level information was encoded and fed into the convolutional layers (Figure 1). The overview of the network structure of DeepMicroClass is shown in Figure 1. In comparison to the more intricate frameworks of deep learning, such as Recurrent Neural Networks (RNNs) and Transformers, CNNs offer the benefits of simplified and accelerated training processes. Additionally, CNNs exhibit lower computational intensity when processing longer sequences. They further facilitate efficient transformation of codon one-hot matrix from a base one-hot matrix. Consequently, these considerations led to the selection of CNNs for the implementation of DeepMicroClass in our study.

The di-path CNN model was trained by minimizing the cross-entropy loss between the predicted and actual sequence classes of input contigs. The training was run for 3000 epochs with a learning rate of 0.001 and a batch size of 256. The selection of hyperparameters was conducted through a grid search, encompassing learning rates of 0.1, 0.01, 0.001 and 0.0001, alongside batch sizes of 64, 128, 256 and 512. The number of epochs was consistently set at 3000 across all configurations, a point beyond which a visual plateau in performance metrics was observed. For each batch, sequences from the whole training dataset were firstly subsampled with weighted random sampling without replacement within an epoch. The weight for samples of each class i was defined as

$$w_i = \frac{number\ of\ samples}{5 \times number\ of\ samples\ in\ class_i}\},$$

so that the numbers of samples in the five classes were kept the same. After the sequences were sampled, a contig length was chosen from 0.5, 1, 2, 3 and 5 kbps, and a contig with the given length was sampled from the original sequence to construct the batch. In the testing stage, sequences with lengths <5 kbps were fed directly to the model for prediction. For sequences with lengths >5 kbps, each input sequence was first split into multiple non-overlapping 5 kbps chunks, then scores given by the model for each chunk were collected, and the mean score of all chunks was used as the final output of the input sequence.

Use-case data preparation and analysis

The daily time-series metagenomic samples were taken off the coast of Santa Catalina Island (Southern California)

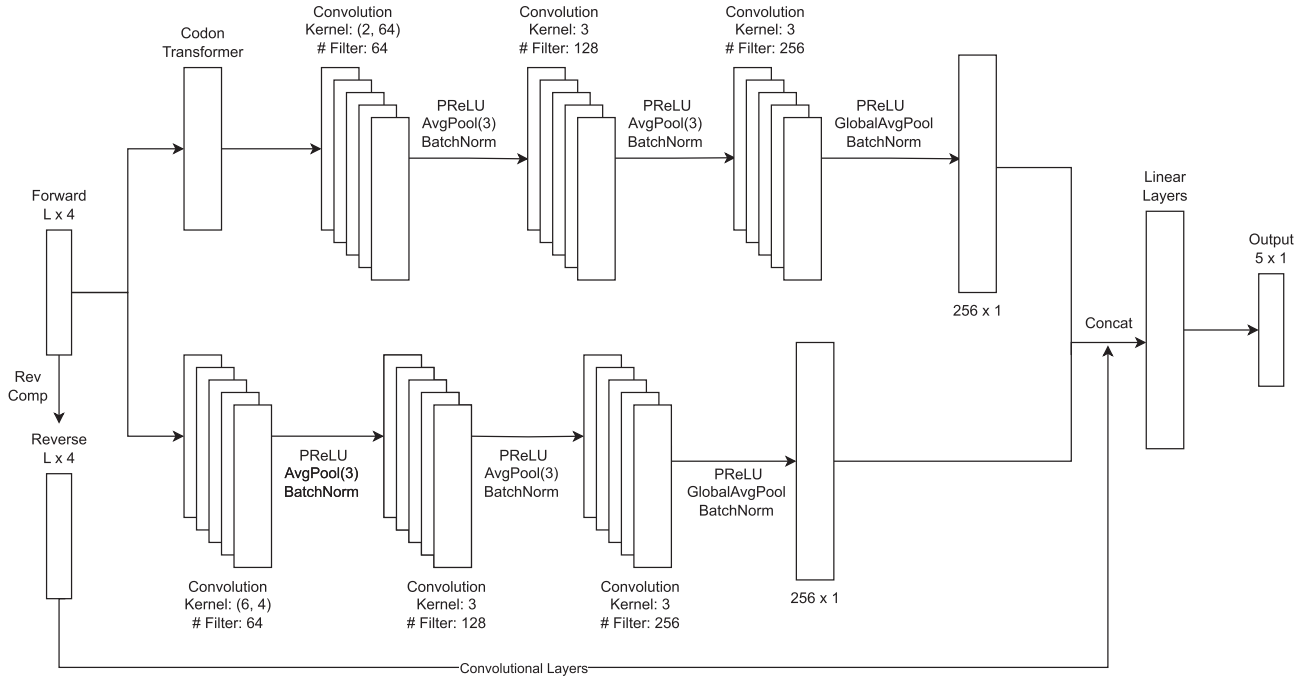

Figure 1. Schematic representation of the multi-class CNN structure used in this study. The network has two convolutional paths, a base-path encodes the nucleotide level information and a codon-path encodes the codon level information. The hyperparameters used for each convolutional layer are marked on the figure. For each strand, the output dimension of base- and codon-paths are 256 and 256, respectively. The di-path outputs of forward and reverse strands are concatenated into a 1024-dimensional vector, which is used as the input of following linear layers. The final linear layer outputs a 5-dimensional vector, with each dimension indicating the probability of the input contig being eukaryotic host, eukaryotic virus, plasmid, prokaryotic host and prokaryotic virus.

near SPOT (San Pedro Ocean Time-series) using an Environmental Sample Processor (ESP). Seawater was prefiltered using a 300 μm mesh, and the 1 μm A/E filters (Pall Gelman) were used to collect microbial cells during day and night (54), and only samples taken during the day were used for DNA extraction and metagenomic sequencing here. Metagenomic libraries were prepared using the Ovation® Ultralow V2 DNA-Seq library preparation kit (NuGEN, Tecan Genomics) under the manufacturer's instruction using 10 ng of starting DNA and amplified for 13 PCR cycles. Metagenomic libraries were sequenced on an Illumina NovaSeq 6000 platform (2 × 150 bp chemistries) at Berry Genomics Co. (Beijing, China). After demultiplexing, the raw reads were first checked with FastQC v0.11.2, then adapter and low-quality regions were trimmed using fastp v0.21.0 (55) with the following parameters: -q 20 -u 20 -l 30 –cut_tail -W 4 -M 20 -c. PhiX174 and sequencing artifacts were removed using bbduk.sh and human genome sequences were removed using bbmap.sh with default parameters, both scripts can be found in the BBTools package v37.24 (https://jgi.doe.gov/data-and-tools/bbtools). Metagenomic samples were assembled independently using metaSPAdes v3.13.0 (56) with a custom kmer set (-k 21,33,55,77,99,127). The assembled contigs were further coassembled as previously described (57). Briefly, all the contigs were pooled and sorted into short (<2 kbps) or long (≥2 kbps) contig sets, the short contig set was first co-assembled using Newbler v2.9 (58), the resulting ≥2 kbps contigs were further co-assembled with the long contig set using minimus2 from the AMOS v3.1.0 toolkit (59). A minimum overlap thresholds of 80 nt and 200 nt were set for Newbler and minimus2, respectively. For both co-assembly steps, a minimum identity cutoff of 0.98 was applied. After co-assembly, contigs were further dereplicated at a 0.98 identity using cd-hit v4.6.8 (60), the resulting contigs were used as reference contigs for sequence classification and read recruitment analysis.

Two monthly time-series metagenomic datasets were used in this study, with samples originally collected at either the Station ALOHA of the Hawaii Ocean Time-series (HOT) program, or the Bermuda-Atlantic Time-series Study (BATS) Station in the Sargasso Sea as previously described (61). Briefly, 500 ml seawater samples were directly filtered onto 0.2 μm polycarbonate filters using a vacuum pump, and preserved at −80°C for metagenomic analysis. Details on DNA extraction and metagenomic sequencing have been described previously (61). These samples were taken from different depths during 2003−2004, with several additional samples collected within 2009. Metagenomes were released as a companion to the bioGEOTRACES datasets. Assemblies and clean reads were downloaded from the NCBI Sequence Read Archive database using accession numbers according to authors' description (61). For both the daily and monthly time-series studies, reads were mapped to reference contigs using bwa mem v0.7.17 with default parameters, and the number of reads aligned >30 nt to reference contigs were counted using bamcov v0.1 (available at https://github.com/fbreitwieser/bamcov) with default parameters. Metagenomic contigs were classified using DeepMicroClass v0.1.0 (in hybrid mode), and read counts assigned to each sequence class were summarized using custom Python scripts (available at https://github.com/chengsly/DeepMicroClass/scripts).

Results

A CNN-based multi-class classifier

Accurate classification of metagenomic contigs of different origins is crucial for gaining a better understanding of microbial community structure and ecological roles of microbes. However, current state-of-the-art sequence classification tools often do not fully appreciate some of the under-represented sequence classes. Here two commonly used viral contig predictors, VirFinder (23) and PPR-Meta (25), were evaluated based on their predicted viral scores. As expected, both predictors gave high scores to prokaryotic viral sequences and low scores to prokaryotic host sequences. However, the scores for eukaryotic host and eukaryotic viral sequences were more evenly distributed (Supplementary Figure S2), revealing an insufficient accuracy in classifying these sequence classes. Out of 500 randomly subsampled genomic sequences for each sequence type of prokaryotes, prokaryotic viruses, microbial eukaryotes, and eukaryotic viruses downloaded from NCBI, 454 prokaryotic viruses and 85 prokaryotic hosts had VirFinder-scores (VF-scores) above 0.5, while 238 eukaryotic viruses and 157 eukaryotic hosts had VF-scores above this value (Supplementary Figure S2A). A similar trend can be observed for PPR-Meta (Supplementary Figure S2B), confirming these tools are not adequately equipped to handle eukaryotic viral and host sequences. This emphasizes the need for novel predictors that consider more sequence types during the model training process.

Here we trained DeepMicroClass as a multi-class predictor to classify five sequence classes, and evaluated its performance on test sequence data of different lengths (0.5, 1, 2, 3, 5, 10, 50 and 100 kbps). The model performance for each sequence type was visualized via bar plots showing the Area under Receiver Operating Characteristics curve (AUROC) score using a one-versus-rest strategy (Figure 2) and line plots showing the corresponding ROC curve (Supplementary Figure S3). Overall, we showed that as the sequence length increased, the model's performance improved across most sequence types, as indicated by the Area Under the Receiver Operating Characteristic (AUC) measurements (Figure 2). DeepMicroClass performed well on all sequence types when the input sequence length was $\geq$1 kbps, with the minimum AUC score being 0.963 on classifying prokaryotic sequences. At the sequence length of 500 bp, DeepMicroClass achieved fairly high AUC scores for eukaryotic (0.944) or prokaryotic (0.96) viruses, whilst the scores for both viral sequence types were always $\geq$0.99 at longer sequence lengths ($\geq$2 kbps) (Figure 2). For non-viral sequences, the AUC scores were highest for eukaryotic sequences, followed by plasmid and prokaryotic genome sequences. However, a slight drop in the True Positive Rate (TPR) could be observed for eukaryotic sequences when the False Positive Rate (FPR) was near 0 (Supplementary Figure S3). With further investigation, the rough curve could be caused by the sharp drop in the number of available eukaryotic sequences in the training dataset, which dropped from 16 002 to 255 when the contig length changed from 10 kbps to 50 kbps.

DeepMicroClass outperforms Tiara and Whokaryote in eukaryotic sequence prediction

In the following three sections, we investigate the performance of DeepMicroClass for particular sequence classes. We used accuracy and F1 score as the metrics to assess the model performance. The former calculates the ratio between the count of correctly predicted sequences and the total number of predictions. The latter is a statistical measure used in binary classification that combines precision and recall into a single metric, calculated as $F1 = 2 \times \frac{precision \times recall}{precision + recall}$, where precision is the ratio of true positive predictions to all positive predictions, and recall is the ratio of true positive predictions to all actual positives. For multi-class predictions, the weighted average of F1 scores for each class was used as the final F1 score. The sequence type composition of 20 benchmark datasets was described in the section 'Benchmark dataset preparation'.

First, we compared the performance of DeepMicroClass with Tiara (47) and Whokaryote (48) on the classification of microbial eukaryotes. Both predictors can be used to identify eukaryotic contigs from metagenomic assemblies without prior knowledge of phylogenetic affiliation. With the compiled benchmark datasets, we showed that DeepMicroClass persistently outcompeted both tools in all scenarios in terms of accuracy and F1 score (Figure 3 and Supplementary Figure S4), and DeepMicroClass was robust to the different compositions of benchmark datasets (Supplementary Figure S4). The average accuracy and F1 score across all benchmark datasets for DeepMicroClass were both 0.99, which were significantly higher than those metrics of Tiara and Whokaryote (pairwise Wilcoxon test *P*-values $\leq$ 9.5e-05 for both accuracy and F1 score). The accuracy of Whokaryote dropped from $\sim$0.95 to $\sim$0.75 as the proportion of eukaryotic sequences increased, and the F1 scores were substantially lower than 0.8 in all test datasets. In contrast, Tiara maintained high accuracy and F1 score across different eukaryotic proportions, though a slight decrease in accuracy could be observed when the eukaryotic proportion was high. DeepMicroClass achieved accuracy and F1 score above 0.98 for all tested scenarios and was robust to variable sequence composition.

A further look into those misclassified sequences revealed that Whokaryote mainly suffered from a lower sensitivity in distinguishing eukaryotic sequences from other sequence types. Tiara showed comparatively lower sensitivity than DeepMicroClass in eukaryotic sequence identification, but received fewer misclassified sequences than Whokaryote (Supplementary Figure S5). Beyond the low sensitivity of identifying eukaryotic sequences, a substantial amount of eukaryotic viruses were mistakenly classified as eukaryotes by Whokaryote, when the proportions of eukaryotic and eukaryotic viral sequences were high in the community. Conversely, when prokaryotic sequences dominated the community, Tiara could potentially classify prokaryotic and plasmid sequences into eukaryotes. Although relatively fewer sequences were misclassified, DeepMicroClass could be further improved by incorporating more eukaryotic viruses during the model training step (Supplementary Figure S5).

DeepMicroClass outcompetes PlasFlow, PPR-Meta and geNomad in plasmid sequence classification

We then compared the performance of DeepMicroClass with PlasFlow (31), PPR-Meta (25) and geNomad (29) in classifying plasmid sequences using the same benchmark datasets described above. DeepMicroClass showed significantly improved results than PlasFlow, PPR-Meta and geNomad in all cases based on both accuracy and F1 score metrics (pairwise Wilcoxon test adj. *P*-value $\leq$ 1.1e-07; Figure 4 and Supplementary Figure S6). Although PlasFlow, PPR-Meta and

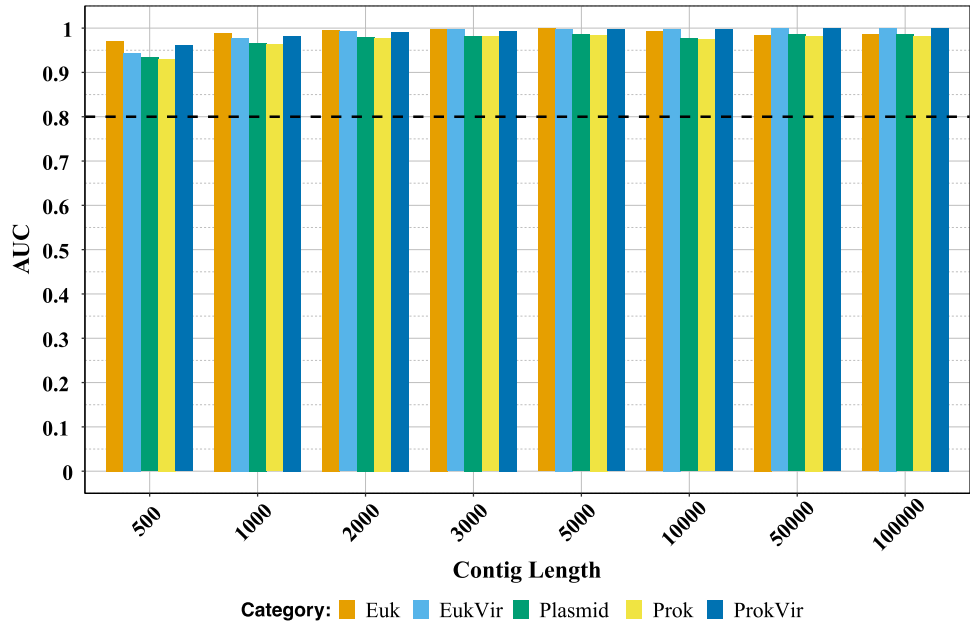

Figure 2. The AUC scores of different length models assessed on test datasets. The barplot shows the AUC scores for five sequence classes at different contig lengths (500 bp, 1 kbps, 2 kbps, 3 kbps, 5 kbps, 10 kbps, 50 kbps and 100 kbps). Euk, eukaryotic sequences; EukVir, eukaryotic viral sequences; Plasmid, plasmid sequences; Prok, prokaryotic genome sequences; ProkVir, prokaryotic viral sequences.

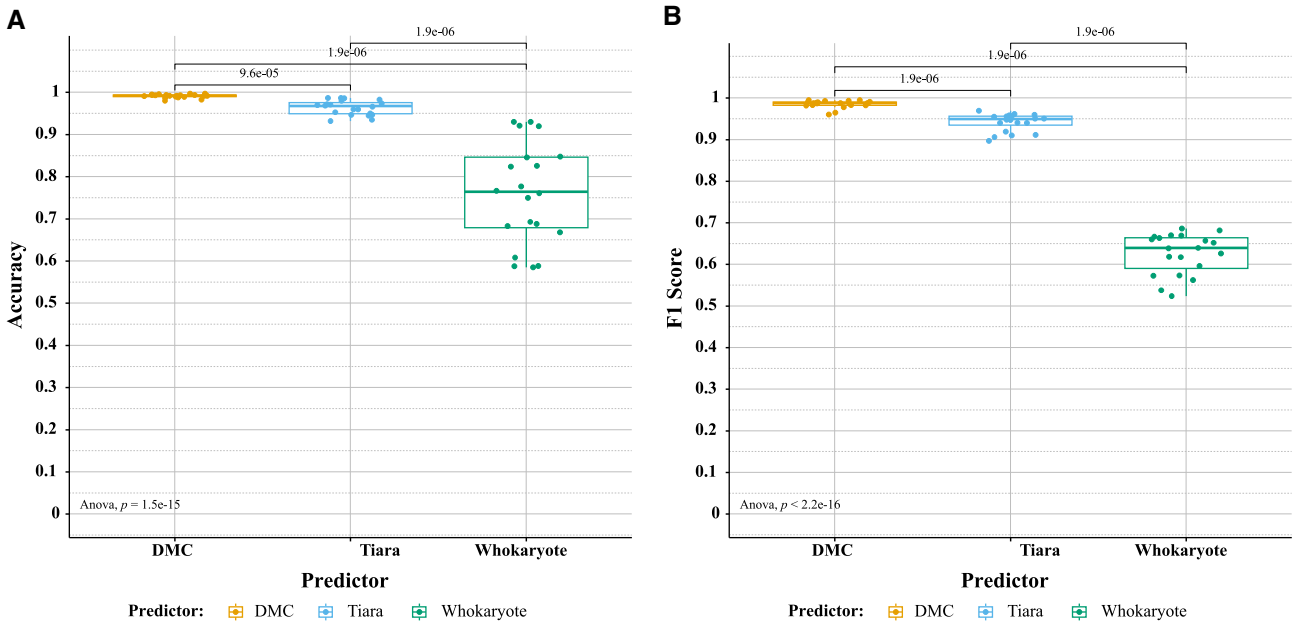

Figure 3. DeepMicroClass outperforms Tiara and Whokaryote on eukaryotic sequence classification. Comparison of (**A**) prediction accuracy and (**B**) F1 score of the three methods on eukaryote sequences. Both the accuracy and F1 score were compared based on 20 designed benchmark datasets. The sequence class composition of the 20 datasets can be found in Supplementary Table S1. Values on top of the pairwise comparisons are Bonferroni adjusted *t*-test *P*-values. The significance of the overall ANOVA test was shown in the bottom left corner.

geNomad were able to achieve a maximum F1 score of 0.68, 0.74 and 0.86, respectively, their performance was severely impaired with increasing proportions of eukaryotic sequences in the benchmark datasets (Supplementary Figure S6). In contrast, the F1 score of DeepMicroClass was constantly higher than the other three tools, though a slight decrease could also be observed with increasing eukaryotic proportions.

We further examined the misclassified sequences and found PlasFlow had high sensitivity but low specificity, and the dominance of misclassified sequence types was in line with the composition of benchmark datasets (Supplementary Figure S7). PPR-Meta might benefit from its modeling of prokaryotic genomes and phages, while it still had a low specificity mainly due to the misclassification of prokaryotic and eukaryotic chromosomal sequences into plasmids (Supplementary Figure S7). On the other hand, geNomad mainly suffered from misclassifying prokaryotic chromosomes into plasmids, though the misclassified eukaryotic se-

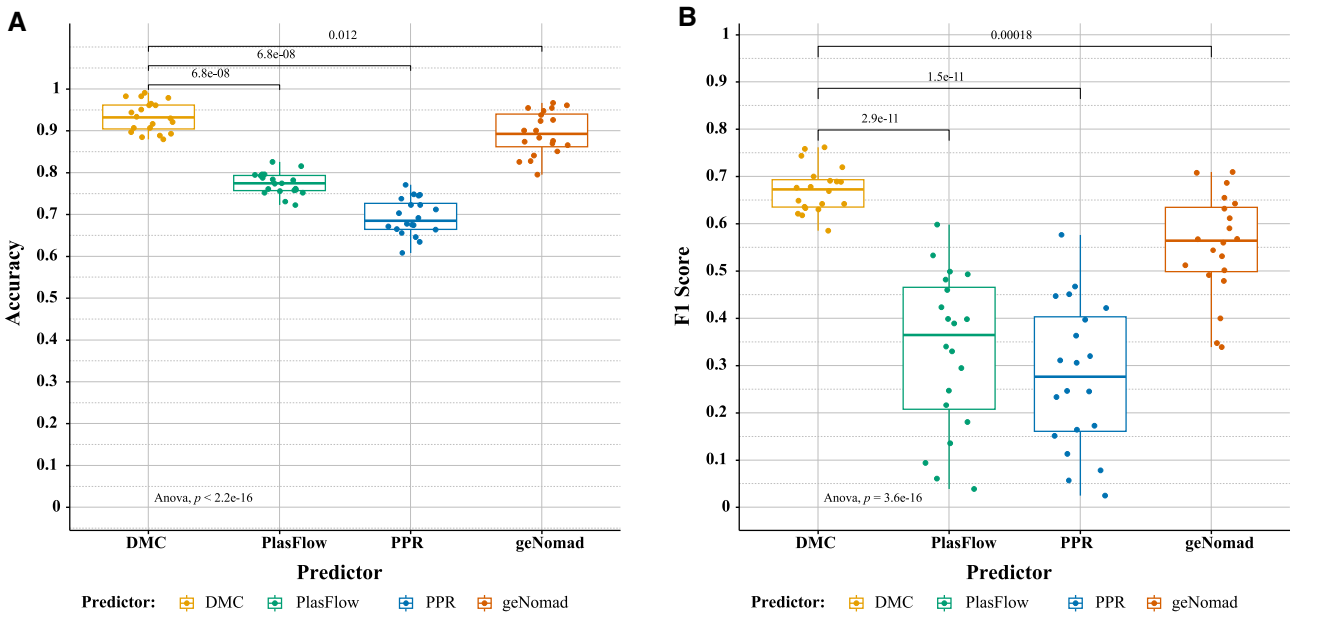

Figure 4. DeepMicroClass outperforms PlasFlow, PPR-Meta and geNomad on plasmid sequence classification. Comparison of (**A**) prediction accuracy and (**B**) F1 score of the four methods on plasmid sequences. The benchmark datasets and the statistical analyses are the same as in Figure 3. DMC, DeepMicroClass; PPR, PPR-Meta.

quences also accounted for a significant share compared to DeepMicroClass (Supplementary Figure S7). It's noteworthy that DeepMicroClass might benefit from its inclusive modeling of eukaryotic genomic and viral sequences since they were rarely misclassified as plasmids, though the misclassification rates between plasmids and prokaryotic chromosomal sequences were still the highest among all misclassifications (Supplementary Figure S8). Probable reasons for such observation are the high affinity and frequent genetic exchange between plasmids and prokaryotic chromosomes. In addition, it should be noted that some of the Integrative and Conjugative Elements (ICEs) on prokaryotic chromosomes may share essential genes with conjugative plasmids or have a plasmid origin, and defective ICEs may have lost their capability to conjugate. Differentiating ICEs and other MGEs from plasmids requires strategical modeling approaches, and further improvements on the neural network structures or incorporating additional features extracted from gene- or operon-centric approaches might yield a more accurate classifier.

DeepMicroClass achieves improved results in viral sequence prediction

Next, we compared the performance of DeepMicroClass with VirSorter2, geNomad, VIBRANT, DeepVirFinder and PPR-Meta on viral contig prediction using the aforementioned benchmark datasets. Among these methods, DeepVirFinder, VIBRANT, PPR-Meta, and geNomad were trained to predict prokaryotic viruses, while VirSorter2 was trained to predict both eukaryotic and prokaryotic viruses. We compared the performance of DeepMicroClass with VirSorter2 on the prediction of both prokaryotic and eukaryotic viruses, and the performance of DeepMicroClass with other predictors on the prediction of prokaryotic viruses. In either case, DeepMicroClass achieved better performance in terms of both accuracy and F1 score than all the other tested tools (Figure 5 and Supplementary Figure S9). VIBRANT and VirSorter2 showed slightly lower accuracy than DeepMicroClass, followed by geNomad, PPR-Meta and DeepVirFinder. More distinct differences were observed in the F1 score metric of these tools across dataset composition, DeepMicroClass achieved an average F1 score of ∼0.96, followed by VirSorter2 and VIBRANT (∼0.90 and ∼0.85, respectively). The F1 score of VIBRANT dropped from 0.94 to <0.80 as increasing proportions of eukaryotic chromosomal and viral sequences in the benchmark datasets (Supplementary Figure S9). Similarly, geNomad, PPR-Meta and DeepVirFinder showed a decreasing tendency in both accuracy and F1 score with the increasing of eukaryotic chromosomal and viral sequences (Figure 5 and Supplementary Figure S9). When considering both prokaryotic and eukaryotic viral sequences as the positive viral set, DeepMicroClass and VirSorter2 were both able to achieve accuracy >0.90 and F1 score >0.80 without being significantly affected by the variations of sequence type composition, and DeepMicroClass constantly outperformed VirSorter2 in both metrics across the benchmark datasets (Figure 5 and Supplementary Figure S9).

The number of misclassified sequences by PPR-Meta, DeepVirFinder, VIBRANT, geNomad and VirSorter2 is shown in Supplementary Figure S10. The distribution of misclassified sequences by PPR-Meta, DeepVirFinder and geNomad showed a similar pattern, that eukaryotic chromosomal and viral sequences were prone to be misidentified as prokaryotic viruses. This indicates tools or models trained without knowledge of eukaryotic sequences are likely to behave similarly when eukaryotes are not rare in the metagenomic community. Although VIBRANT and VirSorter2 had fewer misclassified sequences compared to PPR-Meta, DeepVirFinder and geNomad, both suffered from misclassifying prokaryotic chromosomal or plasmid sequences into prokaryotic viruses (Supplementary Figure S10). Since both VIBRANT and VirSorter2 use a gene-centric approach, it's possible that some of the viral signature genes or fragments could also be widely

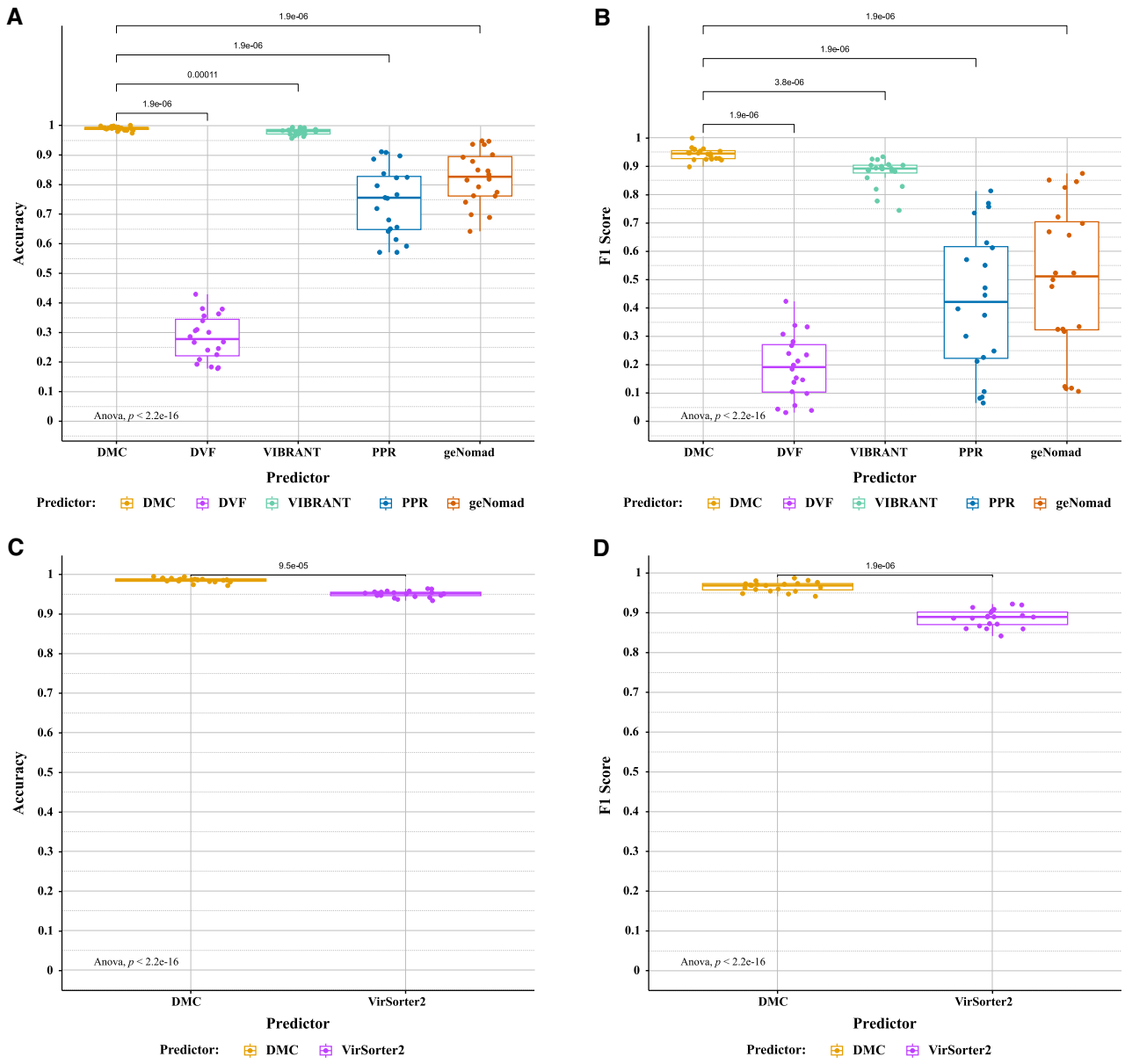

Figure 5. DeepMicroClass outperforms DeepVirFinder, VIBRANT, PPR-Meta and geNomad on prokaryotic viral sequence classification and VirSorter2 on prokaryotic and eukaryotic viral sequence classification. Comparison of (**A**) prediction accuracy and (**B**) F1 score of the five methods on prokaryotic viral sequences. Comparison of (**C**) prediction accuracy and (**D**) F1 score of DeepMicroClass and VirSorter2 on prokaryotic and eukaryotic viral sequences. The benchmark datasets and the statistical analyses are the same as in Figure 3. DMC, DeepMicroClass; DVF, DeepVirFinder; PPR, PPR-Meta.

detected in prokaryotic genomes or plasmids as a result of frequent gene transfer among them. DeepMicroClass could be improved by increasing its sensitivity in prokaryotic viral contig identifications and reducing the misclassification of plasmids as prokaryotic viruses.

Since DeepMicroClass, PPR-Meta and geNomad are multiclass classifiers, here we also compared their performance based on accuracy and F1 score metrics on multiclass sequence classification using the same benchmark datasets (Supplementary Figures S11 and S12). Here we only considered prokaryotic chromosomal, prokaryotic viral and plasmid sequences for comparison with PPR-Meta and geNomad as they were not trained for eukaryotic sequence classification. On the other hand, all five sequence types were considered for the evaluation of DeepMicroClass. In this case, DeepMicroClass still outperformed PPR-Meta and geNomad in all tested scenarios as evaluated by both the accuracy or and F1 score metrics (pairwise Wilcoxon test p-values $\leq$ 1.9e-06; Supplementary Figures S11 and S12). Both accuracy and F1 scores of DeepMicroClass were rarely below 0.95 across the sequence composition of the 20 benchmark datasets, while they were rarely above 0.9 for geNomad, or rarely above 0.8 for PPR-Meta (Supplementary Figure S11). Although the performance of DeepMicroClass was also deteriorated by the misclassification between prokaryotic chromosomal and plasmid sequences (Supplementary Figure S8), the amounts of misclassified sequences were significantly lower than VIBRANT, VirSorter2 or geNomad (Supplementary Figure S10).

Comparison of different viral predictors on the CAMI II marine dataset

We also compared the different viral sequence prediction methods using the CAMI II marine dataset (https://frl.publisso.de/data/frl:6425521/marine/). This dataset contains sequences of different source origins, including archaea, bacteria, viruses and unknown sequences. To preprocess this dataset, we removed the sequences with a Mash distance lower than 0.1 to our training dataset so that the performance of DeepMicroClass would not be inflated due to data overlapping. We employed DeepMicroClass and five other selected viral predictors, PPR-Meta, DeepVirFinder, VIBRANT, VirSorter2, and geNomad, to classify sequences in this dataset and to check their performance on viral prediction. As shown in Figure 6, VirSorter2 and DeepVirFinder were comparable in accuracy (0.845) and were better than those of DeepMicroClass, VIBRANT, and geNomad (0.789, 0.802 and 0.761, respectively)(Figure 6 A). In terms of F1 score, DeepMicroClass, VirSorter2, and geNomad obtained very close F1 scores (0.748, 0.730 and 0.752, respectively), which were substantially higher than those obtained by VIBRANT and DeepVirFinder (0.577 and 0.589, respectively) (Figure 6 B). It should be noted that the F1 scores obtained by different tools based on the CAMI II dataset were much lower than those obtained on our synthetic datasets, possibly due to the presence of unknown sequences in this dataset. Surprisingly, PPR-Meta achieved the best performance with the F1 score ∼0.955 (Figure 6 B). Although the overlap between the training dataset of PPR-Meta and the CAMI II dataset hasn't been examined here, we suggest further verification should be performed to accept this exceptional performance.

In the evaluation of various tools for sequence classification, runtime efficiency is another critical factor. Therefore, we compared the running time of the five tools, DeepMicroClass, geNomad, DeepVirFinder, VIBRANT and VirSorter2, tested on a server running Ubuntu 20.04, equipped with AMD Epyc 7742 (64 cores) and Nvidia T4 GPU, using the latest version of each tool as of 17 January 2024. In the speed test, GPU will be used if the software supports GPU calculation; otherwise, 32 CPU cores will be used. All five tools were tested on a subsampled CAMI II marine dataset consisting of 1084 sequences, and the runtime results for these tools are shown in Supplementary Figure S13. DeepMicroClass, benefiting from its CNN-based structure, achieved the fastest prediction time of 17 seconds. VIBRANT ranked second in terms of running speed with 69 seconds, thanks to fully utilizing multiple CPU cores. Due to additional tasks such as gene annotation and provirus identification, geNomad took 117 seconds to complete the whole process. Despite being a CNN-based tool, DeepVirFinder's less efficient code optimization put it in the fourth place with a runtime of 294 seconds. On the other hand, VirSorter2 performed multiple tasks in its default workflow and exhibited suboptimal CPU core utilization, which took >1600 s, the longest among all the tested predictors. This comparative analysis underscores the efficiency of DeepMicroClass for large-scale metagenomic analyses.

DeepMicroClass complements alignment-based predictors for sequences without close representative sequences

Alignment-based classifiers can suffer from incomplete genomic databases of under-represented sequence types, particularly for complex natural environments such as marine or soil systems. Since some of the eukaryotic sequences in the designed benchmark dataset were not included in the databases of Kaiju (44) and MetaEuk (45), the benchmark dataset allowed us to compare the performance of DeepMicroClass with these alignment-based predictors on eukaryotic sequence classification. The evaluation showed significantly increased accuracy and F1 score of DeepMicroClass over Kaiju and MetaEuk on the synthetic benchmark dataset (Supplementary Figure S14). Therefore, this evaluation distinctly highlighted DeepMicroClass's superior performance compared to Kaiju and MetaEuk on complex metagenomes with uncultured microbial eukaryotes. Thus, DeepMicroClass could complement alignment-based tools to classify novel sequences that are not represented in the current reference database.

DeepMicroClass predicts abundant eukaryotic and viral contigs in real metagenomes

Our results from the synthetic benchmark data and the CAMI II marine data clearly showed the superior performance of DeepMicroClass over other individual or multi-class sequence classification methods. For real metagenomic data such as SPOT, HOT and BATS, the true distributions of different classes of microbes are not known. Thus, they cannot be used to compare the performance of the different sequence classification methods. Instead we use the best performing method, DeepMicroClass to analyze the real data and understand the composition and dynamics of the different microbes in these communities.

We first applied DeepMicroClass to a cell size fractionated daily marine metagenomic dataset sampled near SPOT, off the coast of Southern California (54). The filters could potentially capture microbes within 1-300 μm in size, suggesting that diverse microbial eukaryotes may be retained. The short reads were first assembled using the methods described in the 'Use-case data preparation and analysis' subsection. DeepMicroClass was then used to classify the contigs into different classes. Thirdly, the short reads were mapped to the different contigs to calculate the abundance of different classes of microbes. Using the co-assembled contigs as the reference, DeepMicroClass classified prokaryotes (prokaryotic genomes and plasmids) recruited on average 26.72% of all clean reads, followed by read percentages recruited by eukaryotes (17.86%), prokaryotic viruses (7.89%) and eukaryotic viruses (3.24%) (Figure 7A, B). Reads recruited by prokaryotes or microbial eukaryotes could occasionally account for 35.42% and 21.14% of all clean reads, respectively. Similarly, the maximum read percentages for prokaryotic or eukaryotic viruses were 15.86% and 4.29%, respectively. These percentages suggest microbial eukaryotes and viruses can represent a large fraction of all the microbes and are essential components of natural microbial communities. Accurate prediction of the sequence types is crucial to evaluate their abundance.

We then analyzed the non-fractionated HOT and BATS monthly metagenomes (≥0.2 μm) (61) to understand the variations of abundance levels of various classes of microbes in cellular metagenomes across spatial and temporal scales. Using reference-based taxonomy assignment, reads assigned to viruses contributed only ∼2% of the whole metagenomes, slightly higher than reads assigned to eukaryotes (∼1%) (61).

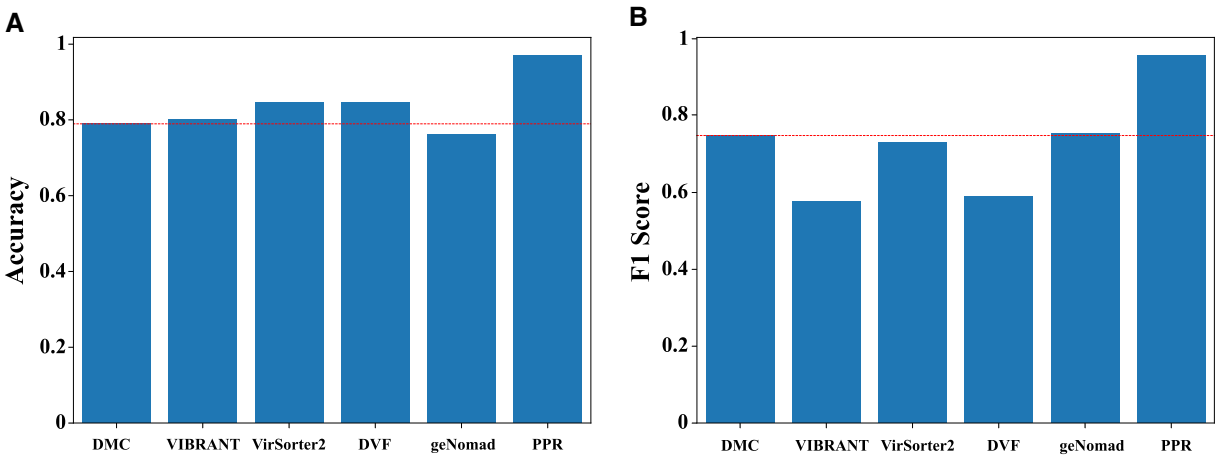

Figure 6. Barplot of prediction accuracy and F1 scores of different sequence prediction tools on the CAMI II marine dataset. The *x*-axis denotes different tools and the *y*-axis denotes accuracy and F1 score. The red dashed line denotes the accuracy and F1 score achieved by DeepMicroClass. DVF denotes DeepVirFinder and PPR denotes PPR-Meta.

In contrast, by classifying the same assemblies using DeepMicroClass at a length cutoff of 2 kbps and mapping reads back to these assemblies, we found that the mean read percentages of all clean reads for eukaryotes and prokaryotic viruses were 3.83% and 3.11% for the HOT metagenomes (Figure 7C–E), or 5.15% and 3.38% for the BATS metagenomes (Figure 7F–H). A higher proportion of prokaryotic viruses could be observed at both the HOT and BATS sites in the subsurface layers (61–125 m) in the late summer or fall seasons (Figure 7C–H). It appeared to be seasonal, with the annual viral contribution peaked from July to October. We further profiled the viral communities using marker genes of known viruses with MetaPhlAn v3 (62), and found that among these known marine viruses, Cyanophage P-RSM6, *Prochlorococcus* phage P-SSM2 and P-SSM7 were enriched in the subsurface layer in late summer at HOT (Supplementary Figure S15A). Similarly, Cyanophage P-RSM6 and *Prochlorococcus* phage P-SSM2 were also abundant at BATS in the subsurface layer in the same month range (Supplementary Figure S15b).

Discussion

The advantages and limitations of DeepMicroClass

DeepMicroClass has demonstrated significant advantages over traditional binary classifiers and other compared tools across all benchmark datasets due to its multi-class model design. Unlike binary classifiers that often suffer from the misclassification of sequence types not modeled by them, DeepMicroClass provides a more reliable and encompassing classification of target sequences by modeling all common sequence types in metagenomes, thereby reducing cross misclassification. This inclusive modeling approach not only enhances the precision and sensitivity of classification but also addresses the propagation of errors to downstream analysis. For instance, recovering eukaryotic genomes requires accurate classification of eukaryotic contigs from metagenomic assemblies (46,63), so that genomic coding potentials, transcriptomic and ecological insights can be derived from these bins (42,64,65). Compared with other multi-class metagenomic sequence classifiers, such as PPR-Meta and geNomad, DeepMicroClass achieved better or comparable performance in plasmid and viral sequence prediction (Figures 4–6 and Supplementary Figures S6, S9), simultaneously providing classification for eukaryotic sequences. Despite using similar CNN models with PPR-Meta (Figure 1), DeepMicroClass excelled at more comprehensive and up-to-date training datasets, in addition to the broader modeled sequence spectrum (25). Compared to geNomad, DeepMicroClass obtained a faster running speed (Supplementary Figure S13). This could be attributed to the additional tasks performed by geNomad as aforementioned, while the underlying model structures may also matter. CNN models are generally simpler and faster than Transformer models used by geNomad, which have been shown when comparing text classification models (66). By leveraging the DeepMicroClass model as a preliminary classification step in metagenomic studies, we advocate for a paradigm shift towards multi-class models in microbial ecology to facilitate a comprehensive understanding of microbial communities.

There are several caveats that should be mentioned while using DeepMicroClass. First, like most kingdom-level sequence classifiers, DeepMicroClass may be complemented by alignment-based classifiers, such as Kaiju or MetaEuk. For sequences with good database coverage, confident alignments to close reference genomes should be used to resolve the disagreements. On the other hand, for unclassified sequences, DeepMicroClass should provide a better broad classification. Second, DeepMicroClass has a relatively lower accuracy in distinguishing plasmids from prokaryotic host genomes (Supplementary Figure S8), when compared to the classification of other sequence classes (Figure 2), despite being the best plasmid classifiers benchmarked (Figure 4 and Supplementary Figure S6). Since the misclassification mostly happens between plasmids and prokaryotic chromosomal sequences, users may simply group them into the same category if this information is not required, or employ an alignment-based method to further refine plasmid classification. Additionally, DeepMicroClass may benefit from modeling the gene contents of plasmids and prokaryotic chromosomes in future implementation. Third, viral sequences predicted by DeepMicroClass may contain proviruses integrated into host genomes. Users may use geNomad, VirSorter2 or CheckV (67) to identify them if needed.

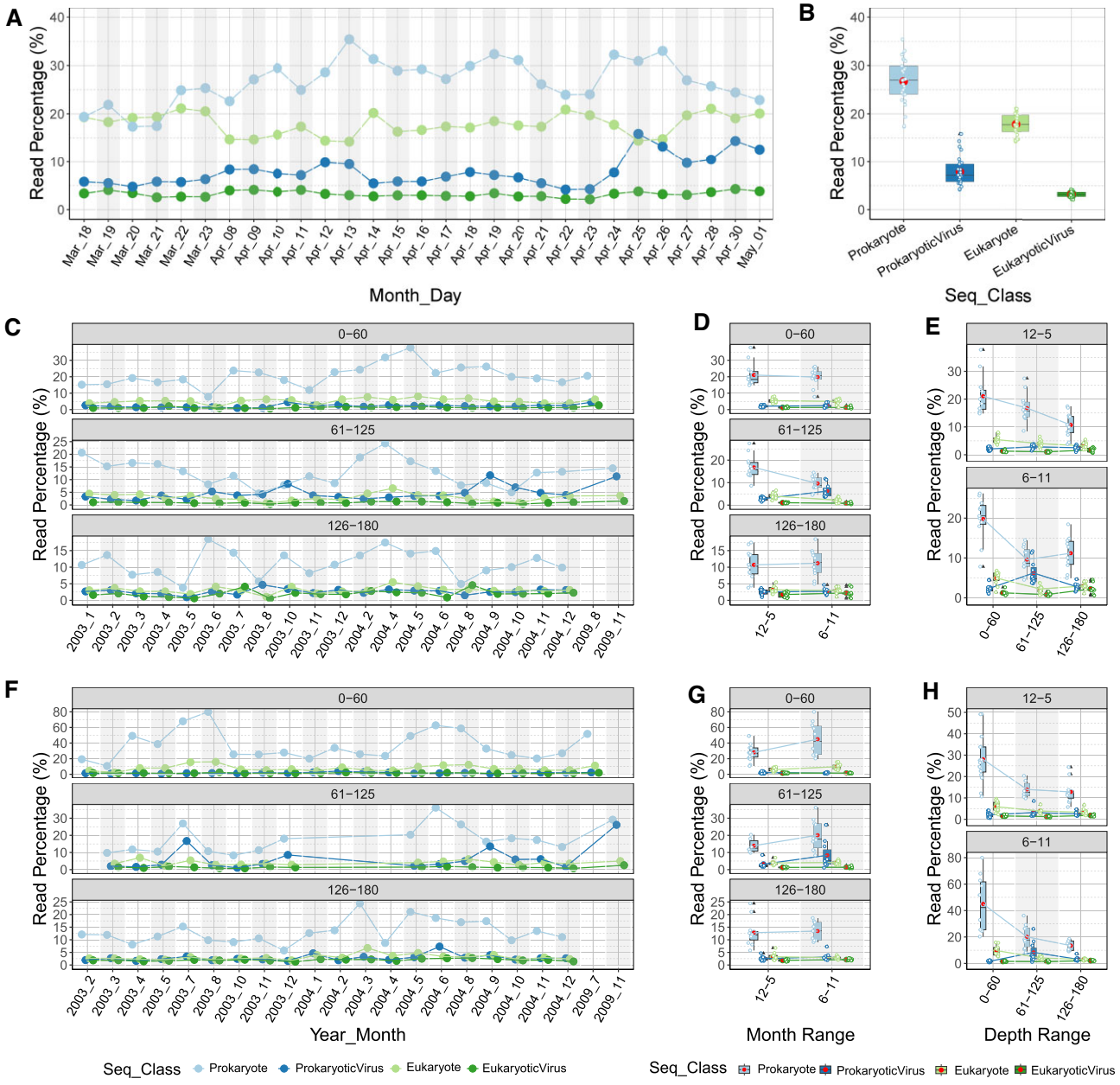

Figure 7. Sequence classification and read abundance analysis of real marine metagenomic samples. Metagenomic samples used in this analysis were taken near SPOT (**A, B**), HOT (**C–E**) and BATS (**F–H**) sampling stations. Metagenomic contigs were classified using DeepMicroClass at a length cutoff of 2 kb, and read percentages of each sequence type were calculated by taking the ratio of read counts assigned to each sequence type and all clean reads of that sample. Samples were taken at a daily frequency near SPOT (A, B), and at a monthly frequency at HOT (C–E) and BATS (F–H). HOT and BATS samples collected at different depths were grouped into three depth ranges, 0-60 m, 61–125 m and 126–180 m. The mean read percentages and standard deviations were summarized for sequence types (B), depth ranges (D, G) and month ranges (E, H). Prokaryote includes reads recruited by prokaryotes and plasmids. Red dots in boxplots indicate the average read percentages for each sequence type. Lines connecting read percentages of same sequence types were drawn for visualization purposes (D, E, G, H).

Implications of increased cellular viruses in marine cellular metagenomes

The abundance of marine viruses varies spatiotemporally, overall showing a nonlinear, power-law relationship with microbial cell abundance (68). Changes in host abundance have a fundamental, disproportionate impact on viral abundance, thus altering the virus-host dynamics and microbial community structure. Studies at SPOT over 5 years of monthly sampling showed a long-term stability of viral communities, with strong seasonal variations imposed (69). On a daily scale, 'boom-and-bust' patterns were also not observed over 40 more days of observation (70), suggesting a community-level resilience dominated by highly abundant viruses, such as cyanophages and pelagiphages (69,71). In this study, the increased relative abundance of viruses in cellular metagenomes at the end of the daily time-series (Figure 7 A) may suggest the recovery of dominant bacterial and viral communities after the phytoplankton bloom (54).

Across the depth gradient, viral abundance decreased from the epipelagic layer to meso- and bathypelagic layers (19,72–74). Viral abundance was also correlated with the chlorophyll

concentration or phytoplankton abundance in the ocean (75), which is subjected to fluctuations of physical processes such as vertical mixing. This is particularly evident for well-studied oceanic regions such as the northwestern Sargasso Sea of the BATS program, where seasonal alternation of winter convective mixing and summer stratification shaped the dynamics of *Prochlorococcus* abundance, resulting in recurring annual maxima in 60-100 m in September (76). Viral abundance was shown to be tightly coupled to *Prochlorococcus* abundance in this layer, and the annual maxima of viral abundance could be well explained by *Prochlorococcus* infecting cyanophages (76,77). Using shotgun metagenomics, we showed that the peaked viral abundance in September 2004 recruited 50% of the mapped cellular metagenomic reads (Figure 7 F-H) and detected close relatives of *Prochlorococcus* phage P-SSM2 were particularly enriched in later summer in the DCM layer (Supplementary Figure S15b). Similarly, previous studies at HOT showed that viral contigs recruited 8% of the cellular metagenomic reads on average across 12 depths, with the maximum being 15% at 100 m depth (78). The high viral contribution to total cellular DNA in the DCM layer is consistent with what we observed here (Figure 7 C-E). In addition, we also detected the seasonal pattern of viral abundance in the DCM layer at HOT, which could be partially explained by enriched relative cyanophages including cyanophage P-RSM6, *Prochlorococcus* phage P-SSM2 and *Prochlorococcus* phage P-SSM7 (Supplementary Figure S15A). These cyanophages were originally isolated from *Prochlorococcus* NATL2A (P-RSM6) and NATL1A (P-SSM2 and P-SSM7), which were members of low-light adapted clade I (LLI) *Prochlorococcus*, the dominant clade at both HOT and BATS in the lower euphotic zone (79). At both stations, the eNATL ecotype of LLI displayed reversed annual sinusoidal patterns in 0–60 m and 60–120 m (79), the superimposed effect might further augment the seasonal dynamics of cyanophages infecting eNATL when assessed using read percentages (Figure 7C–E). Besides, viruses are susceptible to high ultraviolet radiation in surface waters (80,81), which might also contribute to the lower contribution of viral reads to total mapped metagenomic reads. The congruence of these analyses demonstrates that DeepMicroClass can be used to detect robust ecological patterns in natural microbial communities.

Conclusions

DeepMicroClass as a versatile multi-class classifier enables the accurate classification of five different metagenomic sequence types in one shot, meanwhile, it avoids the time-consuming and error-prone preprocessing steps that could potentially propagate errors to the final classification. The inclusive modeling of all common sequence types in metagenomes also makes DeepMicroClass attain better or comparable performance than the other state-of-the-art individual predictors on different benchmark datasets, with faster running speed. Based on DeepMicroClass's classification, we detected a high relative abundance of marine eukaryotes and prokaryotic viruses in a coastal metagenomic dataset. Using two open-ocean metagenomic datasets, monthly dynamic variations of prokaryotic viral abundance in the subsurface layer could be observed, which is consistent with long-term observations at these stations. Our case studies indicate that both host and viral sequences are essential components in the cellular metagenomes, and robust ecological patterns can be obtained with DeepMicroClass, even for coarse sequence types. We argue that by using DeepMicroClass as a preliminary classification step on metagenomic/viromic assemblies, one can further focus on the interested sequence types for the following analysis, such as metagenomic binning of prokaryotic or eukaryotic contigs, comparative genomic analysis of viral or plasmid sequences, etc. We conclude DeepMicroClass is a useful addition to the metagenomic toolbox, and its application can facilitate studies of under-appreciated sequence types, such as microbial eukaryotic or viral sequences.

Data availability

The source code used in this study can be found at https://doi.org/10.5281/zenodo.10989619, and the developing branch can be found at https://github.com/chengsly/DeepMicroClass. Benchmark datasets have been deposited at figshare (available at https://dx.doi.org/10.6084/m9.figshare.14576193). Raw reads for the case study were deposited at NCBI under the umbrella BioProject PRJNA739254. Additional details of data and analysis are available from the corresponding authors upon request.

Supplementary data

Supplementary Data are available at NARGAB Online.

Acknowledgements

We thank Dr David M. Needham, Dr J. Cesar Ignacio-Espinoza and Erin B. Fichot for their help with DNA extraction and metagenomic library preparation. The funders had no roles in study design, data collection or analysis, the decision to publish, and the preparation of the manuscript.

Author contributions: S.H., J.A.F. and F.S. conceived the project; S.H., T.T., S.C. and F.S. designed the neural network structure and model evaluation procedures; S.H. and T.T. designed the training, test datasets and use-case applications; S.H., T.T. and S.C. prepared the training and test datasets; T.T., S.C. and S.H. implemented the software and performed the data analysis; S.H., T.T. and S.C. prepared all the figures and tables; S.H. drafted the manuscript; T.T., S.C., T.C., Y.L., T.X., T.C., J.A.F. and F.S. reviewed and edited the manuscript.

Funding

NSF grant [EF-2125142 to F.S., J.A.F.]; NSFC grant [42276163 to S.H.]; Shenzhen Science, Technology and Innovation Commission Programme [JCYJ20220530115401003 to S.H.]; Simons Collaboration on Computational Biogeochemical Modeling of Marine Ecosystems/CBIOMES) grant [549943 to J.A.F.]; Gordon and Betty Moore Foundation [3779 to J.A.F.]; National Key R&D Program of China [2021YFF1201303, 2022YFC2703105 to T.C.]; Guoqiang Institute of Tsinghua University (to T.C.).

Conflict of interest statement

None declared.

References

1. Falkowski,P.G., Fenchel,T. and Delong,E.F. (2008) The microbial engines that drive Earth's biogeochemical cycles. *Science*, **320**, 1034–1039.
2. Azam,F. and Worden,A.Z. (2004) Oceanography. Microbes, molecules, and marine ecosystems. *Science (New York, N.Y.)*, **303**, 1622–1624.
3. Raoult,D. and Forterre,P. (2008) Redefining viruses: lessons from Mimivirus. *Nat. Rev. Microbiol.*, **6**, 315–319.
4. Woese,C.R. and Fox,G.E. (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci. U.S.A.*, **74**, 5088–5090.
5. Pace,N.R., Stahl,D.A., Lane,D.J. and Olsen,G.J. (1986) The analysis of natural microbial populations by ribosomal RNA sequences. In: Marshall,K.C. (ed.) *Advances in Microbial Ecology, Advances in Microbial Ecology*. Springer, US, pp. 1–55.
6. Olsen,G.J., Lane,D.J., Giovannoni,S.J., Pace,N.R. and Stahl,D.A. (1986) Microbial ecology and evolution: a ribosomal RNA approach. *Annu. Rev. Microbiol.*, **40**, 337–365.
7. Schmidt,T.M., DeLong,E.F. and Pace,N.R. (1991) Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. *J. Bacteriol.*, **173**, 4371–4378.
8. Stein,J.L., Marsh,T.L., Wu,K.Y., Shizuya,H. and DeLong,E.F. (1996) Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. *J. Bacteriol.*, **178**, 591–599.
9. Vergin,K.L., Urbach,E., Stein,J.L., DeLong,E.F., Lanoil,B.D. and Giovannoni,S.J. (1998) Screening of a fosmid library of marine environmental genomic DNA fragments reveals four clones related to members of the order planctomycetales. *Appl. Environ. Microbiol.*, **64**, 3075–3078.
10. Rondon,M.R., August,P.R., Bettermann,A.D., Brady,S.F., Grossman,T.H., Liles,M.R., Loiacono,K.A., Lynch,B.A., MacNeil,I.A., Minor,C., *et al.* (2000) Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl. Environ. Microbiol.*, **66**, 2541–2547.
11. Béjà,O., Suzuki,M.T., Koonin,E.V., Aravind,L., Hadd,A., Nguyen,L.P., Villacorta,R., Amjadi,M., Garrigues,C., Jovanovich,S.B., *et al.* (2000) Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. *Environ. Microbiol.*, **2**, 516–529.
12. Legault,B.A., Lopez-Lopez,A., Alba-Casado,J.C., Doolittle,W.F., Bolhuis,H., Rodriguez-Valera,F. and Papke,R.T. (2006) Environmental genomics of 'Haloquadratum walsbyi' in a saltern crystallizer indicates a large pool of accessory genes in an otherwise coherent species. *BMC Genomics*, **7**, 171.
13. Venter,J.C., Remington,K., Heidelberg,J.F., Halpern,A.L., Rusch,D., Eisen,J.A., Wu,D., Paulsen,I., Nelson,K.E., Nelson,W., *et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science (New York, N.Y.)*, **304**, 66–74.
14. Handelsman,J. (2004) Metagenomics: application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev.*, **68**, 669–685.
15. Xia,L.C., Steele,J.A., Cram,J.A., Cardon,Z.G., Simmons,S.L., Vallino,J.J., Fuhrman,J.A. and Sun,F. (2011) Extended local similarity analysis (eLSA) of microbial community and other time series data with replicates. *BMC Syst. Biol.*, **5**, S15.
16. Schloissnig,S., Arumugam,M., Sunagawa,S., Mitreva,M., Tap,J., Zhu,A., Waller,A., Mende,D.R., Kultima,J.R., Martin,J., *et al.* (2013) Genomic variation landscape of the human gut microbiome. *Nature*, **493**, 45–50.
17. Sedlar,K., Kupkova,K. and Provaznik,I. (2017) Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics. *Comput. Struct. Biotechnol. J.*, **15**, 48–55.
18. Suttle,C.A. (2005) Viruses in the sea. *Nature*, **437**, 356–361.
19. Suttle,C.A. (2007) Marine viruses ' major players in the global ecosystem. *Nat. Rev. Microbiol.*, **5**, 801–812.
20. Fuhrman,J.A. (1999) Marine viruses and their biogeochemical and ecological effects. *Nature*, **399**, 541–548.
21. Wilhelm,S.W. and Suttle,C.A. (1999) Viruses and nutrient cycles in the SeaViruses play critical roles in the structure and function of aquatic food webs. *BioScience*, **49**, 781–788.
22. Roux,S., Enault,F., Hurwitz,B.L. and Sullivan,M.B. (2015) VirSorter: mining viral signal from microbial genomic data. *PeerJ*, **3**, e985.
23. Ren,J., Ahlgren,N.A., Lu,Y.Y., Fuhrman,J.A. and Sun,F. (2017) VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome*, **5**, 69.
24. Ren,J., Song,K., Deng,C., Ahlgren,N.A., Fuhrman,J.A., Li,Y., Xie,X., Poplin,R. and Sun,F. (2020) Identifying viruses from metagenomic data using deep learning. *Quant. Biol.*, **8**, 64–77.
25. Fang,Z., Tan,J., Wu,S., Li,M., Xu,C., Xie,Z. and Zhu,H. (2019) PPR-Meta: a tool for identifying phages and plasmids from metagenomic fragments using deep learning. *GigaScience*, **8**, 6.
26. Kieft,K., Zhou,Z. and Anantharaman,K. (2020) VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome*, **8**, 90.
27. Guo,J., Bolduc,B., Zayed,A.A., Varsani,A., Dominguez-Huerta,G., Delmont,T.O., Pratama,A.A., Gazitúa,M.C., Vik,D., Sullivan,M.B. and et,al. (2021) VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome*, **9**, 37.
28. Hall,J. P.J., Wood,A.J., Harrison,E. and Brockhurst,M.A. (2016) Source-sink plasmid transfer dynamics maintain gene mobility in soil bacterial communities. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, 8260–8265.
29. Camargo,A.P., Roux,S., Schulz,F., Babinski,M., Xu,Y., Hu,B., Chain,P. S.G., Nayfach,S. and Kyrpides,N.C. (2023) Identification of mobile genetic elements with geNomad. *Nat. Biotechnol.*, https://doi.org/10.1038/s41587-023-01953-y.
30. Zhou,F. and Xu,Y. (2010) cBar: a computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data. *Bioinformatics*, **26**, 2051–2052.
31. Krawczyk,P.S., Lipinski,L. and Dziembowski,A. (2018) PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Res.*, **46**, e35.
32. Royer,G., Decousser,J.W., Branger,C., Dubois,M., Médigue,C., Denamur,E. and Vallenet,D. (2018) PlaScope: a targeted approach to assess the plasmidome from genome assemblies at the species level. *Microbial Genom.*, **4**, 9.
33. Pellow,D., Mizrahi,I. and Shamir,R. (2020) PlasClass improves plasmid sequence classification. *PLoS Comput. Biol.*, **16**, e1007781.
34. Parfrey,L.W., Walters,W.A. and Knight,R. (2011) Microbial eukaryotes in the human microbiome: ecology, evolution, and future directions. *Front. Microbiol.*, **2**, 153.
35. Bik,H.M., Sung,W., Ley,P.D., Baldwin,J.G., Sharma,J., Rocha-Olivares,A. and Thomas,W.K. (2012) Metagenetic community analysis of microbial eukaryotes illuminates biogeographic patterns in deep-sea and shallow water sediments. *Mol. Ecol.*, **21**, 1048–1059.
36. Oliverio,A.M., Power,J.F., Washburne,A., Cary,S.C., Stott,M.B. and Fierer,N. (2018) The ecology and diversity of microbial eukaryotes in geothermal springs. *ISME J.*, **12**, 1918–1928.
37. Pawlowski,J., Audic,S., Adl,S., Bass,D., Belbahri,L., Berney,C., Bowser,S.S., Cepicka,I., Decelle,J., Dunthorn,M., *et al.* (2012) CBOL protist working group: barcoding eukaryotic richness beyond the animal, plant, and fungal kingdoms. *PLoS Biol.*, **10**, e1001419.
38. Amaral-Zettler,L.A., McCliment,E.A., Ducklow,H.W. and Huse,S.M. (2009) A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. *PLoS One*, **4**, e6372.
39. Carradec,Q., Pelletier,E., Da Silva,C., Alberti,A., Seeleuthner,Y., Blanc-Mathieu,R., Lima-Mendez,G., Rocha,F., Tirichine,L.,

Labadie,K., *et al.* (2018) A global ocean atlas of eukaryotic genes. *Nat. Commun.*, **9**, 373.
40. Sieracki,M.E., Poulton,N.J., Jaillon,O., Wincker,P., Vargas,C.d., Rubinat-Ripoll,L., Stepanauskas,R., Logares,R. and Massana,R. (2019) Single cell genomics yields a wide diversity of small planktonic protists across major ocean ecosystems. *Sci. Rep.*, **9**, 6025.
41. Keeling,P.J., Burki,F., Wilcox,H.M., Allam,B., Allen,E.E., Amaral-Zettler,L.A., Armbrust,E.V., Archibald,J.M., Bharti,A.K., Bell,C.J., *et al.* (2014) The marine microbial eukaryote transcriptome sequencing project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS Biol.*, **12**, e1001889.
42. Vorobev,A., Dupouy,M., Carradec,Q., Delmont,T.O., Annamalé,A., Wincker,P. and Pelletier,E. (2020) Transcriptome reconstruction and functional analysis of eukaryotic marine plankton communities via high-throughput metagenomics and metatranscriptomics. *Genome Res.*, **30**, 647–659.
43. Burki,F., Roger,A.J., Brown,M.W. and Simpson,A. G.B. (2020) The new tree of eukaryotes. *Trends Ecol. Evol.*, **35**, 43–55.
44. Menzel,P., Ng,K.L. and Krogh,A. (2016) Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat. Commun.*, **7**, 11257.
45. Levy Karin,E., Mirdita,M. and Söding,J. (2020) MetaEuk-sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. *Microbiome*, **8**, 48.
46. West,P.T., Probst,A.J., Grigoriev,I.V., Thomas,B.C. and Banfield,J.F. (2018) Genome-reconstruction for eukaryotes from complex natural microbial communities. *Genome Res.*, **28**, 569–580.
47. Karlicki,M., Antonowicz,S. and Karnkowska,A. (2022) Tiara: deep learning-based classification system for eukaryotic sequences. *Bioinformatics*, **38**, 344–350.
48. Pronk,L.J. and Medema,M.H. (2022) Whokaryote: distinguishing eukaryotic and prokaryotic contigs in metagenomes based on gene structure. *Microb. Genom.*, **8**, 000823.
49. Guillou,L., Bachar,D., Audic,S., Bass,D., Berney,C., Bittner,L., Boutte,C., Burgaud,G., de Vargas,C., Decelle,J., *et al.* (2013) The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids Res.*, **41**, D597–D604.
50. Johnson,L.K., Alexander,H. and Brown,C.T. (2019) Re-assembly, quality evaluation, and annotation of 678 microbial eukaryotic reference transcriptomes. *GigaScience*, **8**, 4.
51. Galata,V., Fehlmann,T., Backes,C. and Keller,A. (2019) PLSDB: a resource of complete bacterial plasmids. *Nucleic Acids Res.*, **47**, D195–D202.
52. Mihara,T., Nishimura,Y., Shimizu,Y., Nishiyama,H., Yoshikawa,G., Uehara,H., Hingamp,P., Goto,S. and Ogata,H. (2016) Linking virus genomes with host taxonomy. *Viruses*, **8**, 66.
53. Ondov,B.D., Treangen,T.J., Melsted,P., Mallonee,A.B., Bergman,N.H., Koren,S. and Phillippy,A.M. (2016) Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.*, **17**, 132.
54. Needham,D.M., Fichot,E.B., Wang,E., Berdjeb,L., Cram,J.A., Fichot,C.G. and Fuhrman,J.A. (2018) Dynamics and interactions of highly resolved marine plankton via automated high-frequency sampling. *ISME J.*, **12**, 2417–2432.
55. Chen,S., Zhou,Y., Chen,Y. and Gu,J. (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, **34**, i884–i890.
56. Nurk,S., Meleshko,D., Korobeynikov,A. and Pevzner,P.A. (2017) metaSPAdes: a new versatile metagenomic assembler. *Genome Res.*, **27**, 824–834.
57. Long,A.M., Hou,S., Ignacio-Espinoza,J.C. and Fuhrman,J.A. (2021) Benchmarking microbial growth rate predictions from metagenomes. *ISME J.*, **15**, 183–195.
58. Margulies,M., Egholm,M., Altman,W.E., Attiya,S., Bader,J.S., Bemben,L.A., Berka,J., Braverman,M.S., Chen,Y.-J., Chen,Z., *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.
59. Treangen,T.J., Sommer,D.D., Angly,F.E., Koren,S. and Pop,M. (2011) Next generation sequence assembly with AMOS. *Curr. Protoc. Bioinform.*. **Chapter 11**, Unit 11.8.
60. Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.
61. Biller,S.J., Berube,P.M., Dooley,K., Williams,M., Satinsky,B.M., Hackl,T., Hogle,S.L., Coe,A., Bergauer,K., Bouman,H.A., *et al.* (2018) Marine microbial metagenomes sampled across space and time. *Scientific Data*, **5**, 180176.
62. Truong,D.T., Franzosa,E.A., Tickle,T.L., Scholz,M., Weingart,G., Pasolli,E., Tett,A., Huttenhower,C. and Segata,N. (2015) MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods*, **12**, 902–903.
63. Olm,M.R., West,P.T., Brooks,B., Firek,B.A., Baker,R., Morowitz,M.J. and Banfield,J.F. (2019) Genome-resolved metagenomics of eukaryotic populations during early colonization of premature infants and in hospital rooms. *Microbiome*, **7**, 26.
64. Duncan,A., Barry,K., Daum,C., Eloe-Fadrosh,E., Roux,S., Tringe,S.G., Schmidt,K., Valentin,K.U., Varghese,N., Grigoriev,I.V., *et al.* (2022) Metagenome-assembled genomes of phytoplankton communities across the Arctic Circle. *Microbiome.*, **10**, 67.
65. Delmont,T.O., Gaia,M., Hinsinger,D.D., Fremont,P., Guerra,A.F., Eren,A.M., Vanni,C., Kourlaiev,A., d'Agata,L., Clayssen,Q., *et al.* (2022) Functional repertoire convergence of distantly related eukaryotic plankton lineages revealed by genome-resolved metagenomics. *Cell Genomics.*, **2**, 5.
66. Lu,H., Ehwerhemuepha,L. and Rakovski,C. (2022) A comparative study on deep learning models for text classification of unstructured medical notes with various levels of class imbalance. *BMC Med. Res. Methodol.*, **22**, 181.
67. Nayfach,S., Camargo,A.P., Schulz,F., Eloe-Fadrosh,E.A., Eloe-Fadrosh,E.A., Roux,S. and Kyrpides,N.C. (2020) checkv assesses the quality and completeness of metagenome assembled viral genomes. *Nat. Biotechnol.*, **39**, 578–585.
68. Wigington,C.H., Sonderegger,D., Brussaard,C. P.D., Buchan,A., Finke,J.F., Fuhrman,J.A., Lennon,J.T., Middelboe,M., Suttle,C.A., Stock,C., *et al.* (2016) Re-examination of the relationship between marine virus and microbial cell abundances. *Nat. Microbiol.*, **1**, 15024.
69. Ignacio-Espinoza,J.C., Ahlgren,N.A. and Fuhrman,J.A. (2020) Long-term stability and Red Queen-like strain dynamics in marine viruses. *Nat. Microbiol.*, **5**, 265–271.
70. Needham,D.M., Chow,C.-E.T., Fuhrman,J.A., Sachdeva,R., Parada,A., Cram,J.A., Needham,D.M., Sachdeva,R., Cram,J.A., Sachdeva,R., *et al.* (2013) Short-term observations of marine bacterial and viral communities: patterns, connections and resilience. *ISME J.*, **7**, 1274–1285.
71. Yeh,Y.-C. and Fuhrman,J.A. (2022) Effects of phytoplankton, viral communities, and warming on free-living and particle-associated marine prokaryotic community structure. *Nat. Commun.*, **13**, 7905.
72. Culley,A.I. and Welschmeyer,N.A. (2002) The abundance, distribution, and correlation of viruses, phytoplankton, and prokaryotes along a Pacific Ocean transect. *Limnol. Oceanogr.*, **47**, 1508–1513.
73. De Corte,D., Sintes,E., Yokokawa,T., Reinthaler,T. and Herndl,G.J. (2012) Links between viruses and prokaryotes throughout the water column along a North Atlantic latitudinal transect. *ISME J.*, **6**, 1566–1577.
74. Lara,E., Vaqué,D., Sà,E.L., Boras,J.A., Gomes,A., Borrull,E., Díez-Vives,C., Teira,E., Pernice,M.C., Garcia,F.C., *et al.* (2017) Unveiling the role and life strategies of viruses from the surface to the dark ocean. *Sci. Adv.*, **3**, e1602565.
75. Danovaro,R., Corinaldesi,C., Dell'Anno,A., Fuhrman,J.A., Middelburg,J.J., Noble,R.T. and Suttle,C.A. (2011) Marine viruses and global climate change. *Fems Microbiol. Rev.*, **35**, 993–1034.
76. Parsons,R.J., Breitbart,M., Lomas,M.W. and Carlson,C.A. (2012) Ocean time-series reveals recurring seasonal patterns of

virioplankton dynamics in the northwestern Sargasso Sea. *ISME J.*, **6**, 273–284.
77. Goldsmith,D.B., Parsons,R.J., Beyene,D., Salamon,P. and Breitbart,M. (2015) Deep sequencing of the viral phoH gene reveals temporal variation, depth-specific composition, and persistent dominance of the same viral phoH genes in the Sargasso Sea. *PeerJ*, **3**, e997.
78. Luo,E., Eppley,J.M., Romano,A.E., Mende,D.R. and DeLong,E.F. (2020) Double-stranded DNA virioplankton dynamics and reproductive strategies in the oligotrophic open ocean water column. *ISME J.*, **14**, 1304–1315.
79. Malmstrom,R.R., Coe,A., Kettler,G.C., Martiny,A.C., Frias-Lopez,J., Zinser,E.R. and Chisholm,S.W. (2010) Temporal dynamics of Prochlorococcus ecotypes in the Atlantic and Pacific oceans. *ISME J.*, **4**, 1252–1264.
80. Wommack,K.E., Hill,R.T., Muller,T.A. and Colwell,R.R. (1996) Effects of sunlight on bacteriophage viability and structure. *Appl. Environ. Microbiol.*, **62**, 1336–1341.
81. Jacquet,S. and Bratbak,G. (2003) Effects of ultraviolet radiation on marine virus-phytoplankton interactions. *FEMS Microbiol. Ecol.*, **44**, 279–289.

Received: September 7, 2023. **Revised:** March 18, 2024. **Editorial Decision:** April 15, 2024. **Accepted:** April 18, 2024