bioRxiv preprint doi: https://doi.org/10.1101/2021.10.26.466018; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

# DeepMicrobeFinder sorts metagenomes into prokaryotes, eukaryotes and viruses, with marine applications

Shengwei Hou[1¶$*], Siliangyu Cheng[2$], Ting Chen[3], Jed A. Fuhrman[1], Fengzhu Sun[2*]

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

[$] These authors contributed equally to this work.
[*] Correspondence: shengwei@usc.edu (S. Hou) and fsun@usc.edu (F. Sun)

**Abstract**

Sequence classification is valuable for reducing the complexity of metagenomes and providing a fundamental understanding of the composition of metagenomic samples. Binary metagenomic classifiers offer an insufficient solution because metagenomes of most natural environments are typically derived from multiple sequence sources including prokaryotes, eukaryotes and the viruses of both. Here we introduce a deep-learning based (not reference-based) sequence classifier, DeepMicrobeFinder, that classifies metagenomic contigs into five sequence classes, e.g., viruses infecting prokaryotic or eukaryotic hosts, eukaryotic or prokaryotic chromosomes, and prokaryotic plasmids. At different sequence lengths, DeepMicrobeFinder achieved area under the receiver operating characteristic curve (AUC) scores >0.9 for most sequence classes, the exception being distinguishing prokaryotic chromosomes from plasmids. By benchmarking on 20 test datasets with variable sequence class composition, we showed that DeepMicrobeFinder obtained average accuracy scores of 0.94, 0.87, and 0.92 for eukaryotic, plasmid and viral contig classification respectively, which were significantly higher than the other state-of-the-art individual predictors. Using a 1-300 µm daily time-series metagenomic dataset sampled from coastal Southern California as a case study, we showed that metagenomic read proportions recruited by eukaryotic contigs could be doubled with DeepMicrobeFinder's classification compared to the counterparts of other reference-based classifiers. In addition, a positive correlation could be observed between eukaryotic read proportions and potential prokaryotic community growth rates, suggesting an enrichment of fast-growing copiotrophs with increased eukaryotic particles. With its inclusive modeling and unprecedented performance, we expect DeepMicrobeFinder will be a useful addition to the toolbox of microbial ecologists, and will promote metagenomic studies of under-appreciated sequence types.

**keywords**: metagenomic contig classification, microbial eukaryotes, eukaryotic viruses, phages, plasmids

## Introduction

Microbes are omnipresent in all conceivable systems on earth, be it extreme environment such as deep-sea hydrothermal vent or host-associated ecosystem like human gut, exerting immense influence on the global biogeochemical cycles (Falkowski et al., 2008; Azam & Worden, 2004) and host physiology (Turnbaugh et al., 2007; Berendsen et al., 2012). Studies of microbial diversity and evolution were pioneered by the discovery of three domains of life using the universally conserved small subunit rRNA gene

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.26.466018; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

sequences (SSUs) as the phylogenetic marker (Woese & Fox, 1977), which enabled biodiversity surveys of environmental microbial communities (Pace et al., 1986; Olsen et al., 1986) and gave rise to the discovery of abundant archaea lineages in the open ocean (Fuhrman, 1992; DeLong, 1992). Microbial coding potentials were further probed using cloning libraries of natural microbial assemblages (e.g., cosmid and fosmid libraries) (Olsen et al., 1986; Schmidt et al., 1991; Stein et al., 1996; Vergin et al., 1998; Rondon et al., 2000; Béjà et al., 2000; Legault et al., 2006), which were revolutionized by shot-gun metagenomes to infer ecological roles of uncultured microbes (Venter et al., 2004; Handelsman, 2004). Depending on where, when and how metagenomic samples were collected, the microbial richness within a sample can range from a consortium of several dominant strains to a conglomerate of thousands of species. The tremendous amount of ever-growing metagenomic data compound with its inherent heterogeneous nature not only provides opportunities to decipher the cryptic interactions of complex microbial communities and the genome-level evolutionary trajectory of individual species, but also poses challenges on how to reliably extract genomic/transposable fragments from metagenomic sequence pools. By assigning metagenomic contigs into distinct groups, the complexity of metagenomes can be reduced to certain taxonomic levels, from coarse domains to fine-grained consensus species or strains. Metagenomic applications developed with the objective of computationally retrieving intended contigs can be briefly framed into two categories, supervised contig classification tools (i.e., viral contig predictors), and unsupervised contig clustering tools (i.e., metagenomic binners, see Sedlar et al., 2017 for a review of binning strategies).

Microbial communities are a collection of diverse biological entities including the ribosome-encoding cellular organisms (REOs), the capsid-encoding organisms (CEOs, e.g., viruses) that can only reproduce within cells of REOs, and orphan replicons (plasmids, transposons, etc) that parasitize REOs or CEOs for propagation (Raoult & Forterre, 2008). Viruses are one of the most abundant entities on earth (Suttle, 2005, 2007), playing a crucial role in the global biogeochemical cycles by controlling nutrient flow via viral shunt (Fuhrman, 1999; Wilhelm & Suttle, 1999). Metagenomic contig classification has been heavily focused on the prediction of viral sequences. One of the state-of-the-art tools to identify viral contigs from metagenomic assemblies is VirSorter (Roux et al., 2015), which predicts viral contigs based on viral signal and categorizes them into three tiers with different confidence levels. VirFinder (Ren et al., 2017) is another viral contig predictor that employs k-mer frequencies and logistic regression to classify contigs to either viral or host sequences, which outperforms VirSorter at different contig lengths, especially for shorter contigs without detectable viral genes. The success of k-mer based methods has inspired the application of deep learning in viral sequence discovery, which leads to the development of DeepVirFinder (Ren et al., 2020) and PPR-Meta (Fang et al., 2019), 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 (Fang et al., 2019). VIBRANT (Kieft et al., 2020) is a recently published tool that uses neural networks to distinguish prokaryotic dsDNA, ssDNA and RNA viruses based on "v-score" metrics, which were calculated from significant protein hits to a collection of Hidden Markov Model (HMM) profiles derived from public databases. Most of the aforementioned tools target bacteriophages. Eukaryotic virus predictors are emerging in recent years, and one of such tools is Host Taxon Predictor (HTP) (Gałan et al., 2019), which utilizes four machine learning methods to classify viral sequences to eukaryotic viruses or bacteriophages based on sequence features including mono-, dinucleotide absolute frequencies and di-trinucleotide relative frequencies. Beyond viruses, plasmids are also important players of shaping

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.26.466018; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

microbial genome evolution and environmental adaptation. Plasmids are mobile genetic elements that can exchange genes with host chromosomes and shuttle between different hosts, leading to gene flows within microbial communities. Thus, by carrying genes related to environmental adaptations and defense systems, plasmids play a pivotal role in maintaining the host genetic and phenotypic plasticity, and increase the host fitness to the changing environments. This also poses a challenge in classifying chromosomal and plasmid contigs from metagenomes, which is particularly true for plasmids sharing a significant amount of genes with their host genomes. There are multiple dedicated tools developed besides PPR-Meta, such as cBar (Zhou & Xu, 2010), PlasFlow (Krawczyk et al., 2018), PlaScope (Royer et al., 2018) and PlasClass (Pellow et al., 2020). 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 targeting the classification of eukaryotic contigs from metagenomes, while eukaryotes are indispensable to the ecological functioning of natural microbial communities. Reference-based applications such as Kaiju (Menzel et al., 2016) and MetaEuk (Levy Karin et al., 2020a) 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 arguable due to the lack of a comprehensive microbial eukaryotic database (Keeling et al., 2014). EukRep (West et al., 2018) is a reference-independent application that uses k-mer frequency and linear-SVM to classify metagenomic contigs into eukaryotic and prokaryotic sequences. It has been proven that when combined with the conventional metagenomic and metatranscriptomic analyses, such as reconstructing eukaryotic bins and gene co-abundance analysis, biological and ecological insight can be readily obtained for uncultured eukaryotes (Vorobev et al., 2020; West et al., 2018).

Despite the significant progress made in the past years, there isn't one tool that can classify eukaryotic/prokaryotic genomes, eukaryotic/prokaryotic viruses, and plasmids in one shot. In fact, all these binary classifiers suffer from sequence types that are not modeled, such as eukaryotic contigs or plasmids can be misclassified as viruses by viral predictors, and viral contigs can be misclassified as plasmids by plasmid predictors, etc. Thus, in order to achieve a more reliable classification of the target sequences, one has to run several of these tools consecutively, each suffers from its own sensitivity and specificity, and the error rates propagate throughout the workflow, resulting in less accurate and biased classification. Here we introduce DeepMicrobeFinder, a versatile multi-class metagenomic contig classifier based on convolutional neural networks (CNN). We show that DeepMicrobeFinder outperforms all the existing tools by precision and sensitivity across all test datasets with different sequence type compositions. More importantly, DeepMicrobeFinder is superior to the other tools by classifying all sequence types simultaneously, which will greatly reduce the time and computation resource usage compared to the conventional way of pipelining a set of different predictors. Using a coastal marine metagenomic dataset as a case study, we show that DeepMicrobeFinder captures more eukaryotic contigs than reference-based classifiers. The higher eukaryotic read proportion is positively correlated with prokaryotic community growth rates, indicating the higher abundance of fast-growing copiotrophs might be involved in the recycling of particular nutrients during the spring bloom.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.26.466018; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

# Materials and methods

## Training dataset preparation

For prokaryotic chromosome sequences, we downloaded all the prokaryotic genomes from NCBI RefSeq on Jan 17, 2020. The prokaryotic genomes were cleaned up by removing all the sequences annotated as "Plasmid" according to the assembly reports. Plasmid sequences were downloaded from NCBI RefSeq plasmid database (available at ftp://ftp.ncbi.nlm.nih.gov/refseq/release/plasmid/) on the same day. For eukaryotic hosts, we created a microbial genomic sequence database including all the bacteria and archaea sequences and selected microbial eukaryotic sequences based on the eukaryotic taxa used by Kaiju (Menzel et al., 2016) (available at https://github.com/bioinformatics-centre/kaiju/blob/master/util/kaiju-taxonlistEuk.tsv) and PR2 database ](Guillou et al., 2013) (https://pr2-database.org/). Specifically, we downloaded all the bacteria and archaea genomes, and selected microbial eukaryotic genomes 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 Jan 19, 2020. In addition to these eukaryotic genomes, we also included 48,961,036 eukaryotic host sequences from the 678 marine eukaryotic transcriptomic re-assemblies (Johnson et al., 2019) of cultured samples generated by the MMETSP project (Keeling et al., 2014), which included 306 pelagic and endosymbiotic marine eukaryotic species representing more than 40 phyla (re-assemblies are available at https://zenodo.org/record/1212585).

Complete and draft viral sequences and associated metadata were retrieved from the NCBI Viruses database (Brister et al., 2015) (https://www.ncbi.nlm.nih.gov/genome/viruses) on Jan 17, 2020, which contains 3,214,806 nucleic acid records in total. To include more viral sequences from the increasing metagenomic and single-cell genomic datasets, we also included more than 760,000 viral sequences from IMG VR2 (Paez-Espino et al., 2019). In total, there are 3,975,259 viral sequences included in this study. For viruses from NCBI, we classified them into eukaryotic or prokaryotic viruses according to their host domains based on the host taxonomy lineages using taxonkit v0.3.0 (Shen & Ren, 2021). Similarly, the IMG VR2 viruses were also classified into eukaryotic or prokaryotic viruses based on the "Host_domain" field in the sequence information file.

## Sequence preprocessing, model selection, training and validation

Training sequences were randomly selected using seqtk (available at https://github.com/lh3/seqtk) from the collected host, virus and plasmid sequences, which contained 3,404 prokaryotic chromosome sequences (Prok), 5,515 prokaryotic virus sequences (ProkVir), 10,952 eukaryotic chromosome sequences (Euk), 173,082 eukaryotic virus sequences (EukVir), and 2,390 plasmid sequences (Plas). Validation sequences were randomly selected from the same original dataset after removing the training sequences. Training and validation sequences were labeled as one of the five classes, e.g., Prok, ProkVir, Euk, EukVir, and Plas. In order to train models at different sequence lengths (500 bp, 1 kb, 2 kb, 3 kb, 5 kb), we first cut training sequences into fixed-length fragments for each model. This resulted in roughly an equal number of chopped sequence fragments for each class for different length models. Specifically, the numbers of training fragments are 400,000 at 500 bp, 200,000 at 1 kb, 90,000 at 2 kb, 66,000 at 3 kb, and 38,000 at 5 kb, respectively. The validation dataset was $\sim 1/10$ of the training dataset for each length model. Both the training and validation sequence fragments were used as input to train a fully connected 3

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.26.466018; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

layer one dimension multi-class CNN. Sequence fragments were first one-hot encoded strandwise into a binary matrix with the dimension of $4 \times L$ where $L$ is the length of the fragment, which was used as the input layer of the neural network. The neural network comprises three convolutional layers, with 64, 128 and 256 filters and kernel sizes of 6, 3, and 2, accordingly. To improve the model robustness and reduce overfitting, a max pooling layer and a batch normalization layer were added after each of the first two convolutional layers, and a global max pooling layer, a dropout layer and a flatten layer were added after the third convolutional layer. Two dense layers were connected after the convolutional layers, with the first one containing 500 hidden units and the second one containing 5 hidden units, corresponding to 5 input classes (**Fig. S1**). Input sequences were encoded in both forward and reverse directions, and the predicted classes were determined by the average scores of both results. When preparing training and validation data, first the training data was sampled from the entire downloaded data, then from the left-over non-training data, we sampled the validation and test datasets with the size of 1/10 of the training data, respectively. The validation and test datasets were used to compute the model performance metrics.

Two prediction modes were provided for user input sequences, the single mode and the hybrid mode. The single mode allows users to select a specific length model, then to cut input sequences into non-overlapping fixed-length chunks to fit the selected model, and finally to make predictions based on the cumulative scores of all chunks for each input sequence. Chunks smaller than the half of the model length will be discarded. In the hybrid mode, when possible, models with longer sequence length have the highest priority. Input sequences were first cut into chunks corresponding to longer models, the remaining part of the sequence were further cut into chunks corresponding to shorter sequence models if possible. This way, most part of the input sequences will be used for the prediction, and the longer models will be always preferred to maximize prediction accuracy. The final prediction scores will be the sum of predicted scores for all chunks.

## Custom benchmark dataset preparation

To compare with the other state-of-the-art individual predictors, we have generated 20 equal-sized (1000 contigs) test datasets with a variable composition of the 5 sequence classes (**Supplemental Table S1**). Briefly, the fractions of PROK (including prokaryotic hosts, prokaryotic viruses, and plasmids) to EUK (including eukaryotic hosts 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 hosts, 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 hosts 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.

## Use-case data preparation and analysis

The daily time-series metagenomic samples were taken off the coast of Southern California using an Environmental Sample Processor (ESP), and the 1 µm A/E filters (Pall Gelman) collected during the day were used for DNA extraction as described previously (Needham et al., 2018). 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 13 PCR cycles. Metagenomic libraries were sequenced on an Illumina NovaSeq 6000 platform ($2 \times 150$ bp chemistries) at Berry Genomics Co. (Beijing, China). After demultiplexing, the raw reads were first checked with

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.26.466018; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

FastQC v0.11.2, then adapter and low quality regions were trimmed using fastp v0.21.0 (Chen et al., 2018) 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 (Nurk et al., 2017) with a custom kmer set (-k 21,33,55,77,99,127). The assembled contigs were further coassembled as previously described (Long et al., 2021). Briefly, all the contigs were pooled and sorted into short (<2kb) or long (≥2kb) contig sets, the short contig set was first coassembled using Newbler v2.9 (Margulies et al., 2005), the resulting ≥2kb contigs were further coassembled with the long contig set (Treangen et al., 2011). A minimum overlap thresholds of 80 nt and 200 nt were set for Newbler and minimus2, respectively. For both coassembly steps, a minimum identity cutoff of 0.98 was applied. After co-assembly, contigs were further dereplicated at 0.98 identity using cd-hit v4.6.8 (Li & Godzik, 2006), the resulting contigs were used as reference contigs for sequence classification and read recruitment analysis. Reference contigs were classified using Kaiju v1.7.3 (Menzel et al., 2016) and MetaEuk v1 (Levy Karin et al., 2020b), as well as DeepMicrobeFinder v0.1.0 (in hybrid mode), read counts assigned to each sequence class were summarized using custom Python scripts. 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.

### Community doubling time and growth rate calculation

The prokaryotic community microbial growth rates were calculated using gRodon (Weissman et al., 2021) in weighted metagenomic mode with temperature adjustment. Specifically, prokaryotic contigs from each individual assembly were predicted using DeepMicrobeFinder and annotated using Prokka v1.14.5 (Seemann, 2014). Reads were mapped to predicted coding genes to get the coverage information, and genes encoding ribosomal proteins were used as highly expressed gene sets for growth rate prediction. The ambient temperature recorded by the sampler was used for growth rate prediction according to the user manual. The direct output of gRodon is doubling time, which was converted to growth rate per day using equation 1,

$$\mu = 24 \times \frac{ln(2)}{d} \tag{1}$$

where $\mu$ and $d$ stand for maximal growth rate per day and minimal doubling time, respectively.

## Results

### A CNN-based multi-class classifier

Microbial eukaryotes and viruses infecting them are not dispensable but indigenous in microbial communities of diverse ecosystems. Confidently identifying these sequences in metagenomes is crucial to understanding their ecological roles. Unfortunately, most of the eukaryotic viruses and hosts are underappreciated by current state-of-the-art tools. For instance, assessed by the predicted viral scores, the two popular viral contig predictors, VirFinder (Ren et al., 2017) and PPR-Meta (Fang et al., 2019), 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 (**Fig. S2**).

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.26.466018; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 VF-score above 0.5, while 238 eukaryotic viruses and 157 eukaryotic hosts had VirFinder-score (VF-score) above this value (**Fig. S2a**). A similar trend can be observed for PPR-Meta (**Fig. S2b**), indicating these tools are not well trained to handle eukaryotic virus and host sequences, which calls for the development of novel predictors that take more sequence types into consideration in the model training step.

Here we compiled a collection of training datasets and trained several CNN-based multi-class models using one-hot encoding at variable lengths (500 bp, 1 kb, 2 kb, 3 kb, and 5 kb) to simultaneously classify eukaryotic host, eukaryotic virus, prokaryotic host, prokaryotic virus and plasmid sequences in one shot (**Fig. S1**). Using test sequences randomly sampled from the datasets that were not used for training, we evaluated the model performance using the Receiver Operating Characteristics (ROC) curve for each sequence type for each trained model (**Fig. 1**). Overall, we showed that with the sequence length increased, the model performance improved based on the Area Under the Receiver Operating Characteristic (AUC or AUROC) measurements for most sequence types (**Fig. 1**). The AUC scores were higher for eukaryotic viruses and eukaryotic hosts, followed by prokaryotic viruses, plasmids, and lastly prokaryotic hosts (**Fig. 1**). When prokaryotic host and plasmid classes collapsed into one sequence type, the performance of all length models improved, indicating the misclassification between prokaryotic hosts and plasmids is a major caveat of DeepMicrobeFinder (**Fig. 1**).

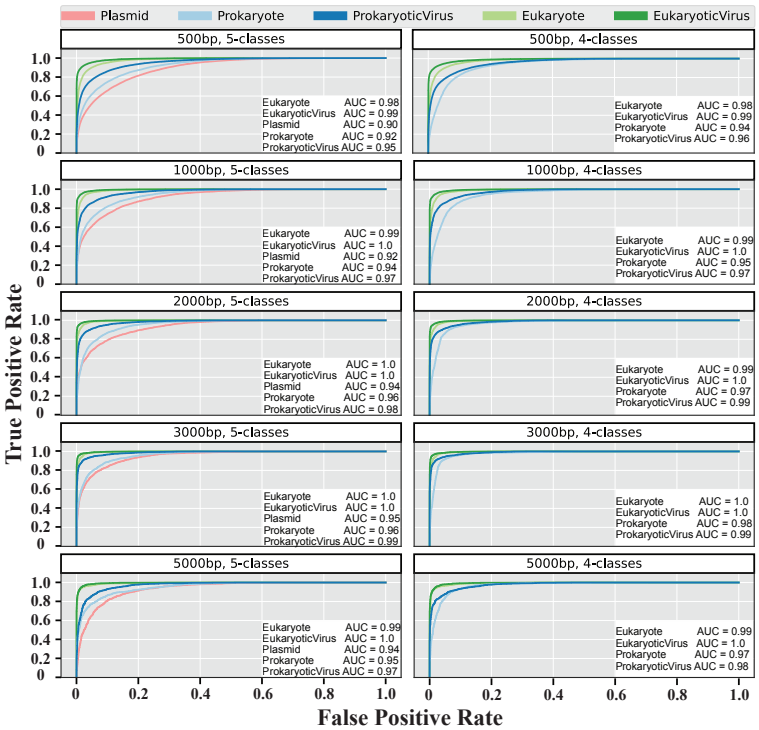


**Fig 1. The ROC curves and AUC scores of different length models assessed on validation datasets.** Left panel shows the ROC curves for 5 sequence classes at different model lengths (500bp, 1k, 2k, 3k and 5k), and the right panel shows the performance when prokaryotic hosts and plasmids were collapsed into the "Prokaryotes" class.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.26.466018; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

## DeepMicrobeFinder outperforms EukRep in eukaryotic host sequence prediction

We compiled a list of benchmark datasets to cover 20 composition scenarios of different sequence types (**Table S1**). The five sequence types were first grouped into two umbrella classes: PROK (including prokaryotic hosts, prokaryotic viruses and plasmids) and EUK (eukaryotic hosts and eukaryotic viruses). Five large groups were first determined using the PROK:EUK ratios of 9:1, 7:3, 5:5, 3:7, 1:9, then within each group, the fractions of host sequences decreased gradually (the details of benchmark dataset preparation can be found in the **Materials and methods** section). This allowed us to compare the performance of DeepMicrobeFinder with the other state-of-the-art predictors under variable sequence composition of metagenomes.

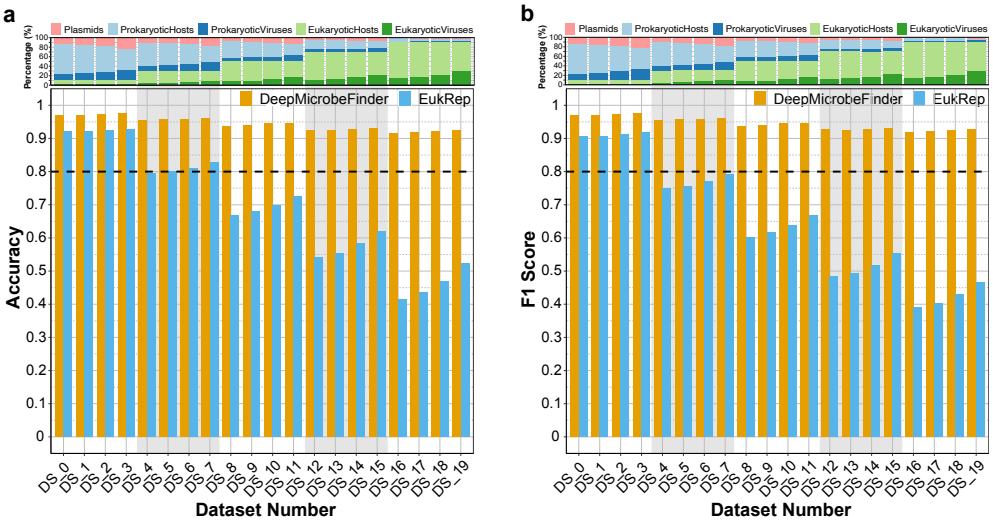


**Fig 2. Distribution of (a) accuracies and (b) F1 scores across 20 test datasets for DeepMicrobeFinder and EukRep.** The top panel shows the sequence composition of 20 test datasets, the composition ratios of different sequence types can be found in Table S1. The dashed black lines indicate where accuracy or F1 score equals to 0.8. Please note a decreasing tendency in both accuracy and F1 score for EukRep along with the increasing eukaryotic fractions.

Microbial eukaryotes are understudied by current metagenomic studies, and EukRep (West et al. 2018) is currently the most popular tool to identify eukaryotic contigs from metagenomic assemblies without prior knowledge of microbial phylogenetic affiliation. With the compiled test datasets, we first benchmarked the performance of DeepMicrobeFinder and EukRep in classifying eukaryotic contigs (**Fig. 2**). DeepMicrobeFinder persistently outcompeted EukRep in all scenarios in terms of accuracy (**Fig. 2a, S3a**) and F1 score (**Fig. 2b, S3b**), and DeepMicrobeFinder was robust to the different compositions of test datasets (**Fig. 2**). The average accuracy and F1 score across all test datasets for DeepMicrobeFinder were both 0.94, which are significantly higher than these metrics of EukRep (0.69 and 0.65, pairwise Wilcoxon test $p$-values $\leq$ 1.9e-06 for both accuracy and F1 score; **Fig. S3**). EukRep had accuracy or F1 score lower than 0.8 in most cases (**Fig. 2**), and for those datasets EukRep achieved accuracy or an F1 score higher than 0.8, the EUK fraction (including eukaryotic hosts and viruses, Table S1) of the test datasets were less than 10% (**Fig. 2, Table S1**), suggesting the performance of EukRep decreases with the increase of the eukaryotic contig proportion. This trend also holds for each fixed EUK fraction, the performance of EukRep increases with the eukaryotic host:virus ratios decrease from 5:1 to 2:1 (**Fig. 2, Table S1**). In contrast, the accuracy and F1 score of DeepMicrobeFinder are higher than 0.9 in all tested scenarios with smaller standard deviations (accuracy/F1 score: 0.02/0.019)

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.26.466018; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

compared to EukRep (accuracy/F1 score: 0.17/0.18), indicating DeepMicrobeFinder is accurate and robust to all the tested scenarios.

A further look into those misclassified sequence types revealed that the poor performance of EukRep is mainly due to its low sensitivity in recognizing eukaryotic contigs, though some plasmids and prokaryotic viruses were also misclassified into eukaryotes when the PROK fractions were high (**Fig. S4**).

## DeepMicrobeFinder outcompetes PlasFlow and PPR-Meta in plasmid sequence prediction

Plasmids are mobile genetic elements of diverse prokaryotes and are one of the major agents of horizontal gene transfer (HGT) among hosts. Here we compared the performance of DeepMicrobeFinder to PlasFlow (Krawczyk et al., 2018) and PPR-Meta (Fang et al., 2019) using the same benchmark datasets. The F1 scores of DeepMicrobeFinder and PPR-Meta were higher than those of PlasFlow in all tested scenarios (pairwise Wilcoxon test adj. $p$-values $\leq$ 5.7e-06; **Fig. S5**), and DeepMicrobeFinder showed significantly improved results than PPR-Meta in all tested cases (pairwise Wilcoxon test adj. $p$-value $\leq$ 1.7e-05; **Fig. 3 & S5**). Moreover, DeepMicrobeFinder showed a conspicuous increment in performance metrics when the EUK fractions were high, while the performance of PPR-Meta and PlasFlow were severely impaired, which is particularly perceptible for PlasFlow (**Fig. 3**). For each fixed PROK fraction, the performance of DeepMicrobeFinder was roughly the same or slightly decreased with the increase of non-host sequences, while marginally improved for PPR-Meta and PlasFlow.

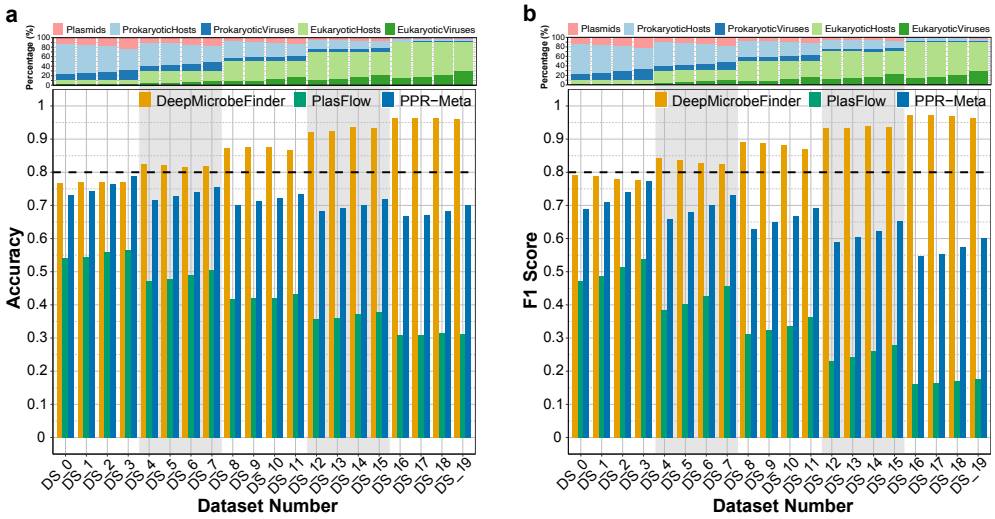


**Fig 3. Distribution of (a) accuracy and (b) F1 scores across 20 test datasets for DeepMicrobeFinder, PlasFlow and PPR-Meta on plasmid classification.** DeepMicrobeFinder achieved the highest performance in respect to accuracy (a) and F1 scores (b). Same benchmarking datasets were used as in **Fig. 2**.

We further examined the misclassified sequences and found PlasFlow had high sensitivity but low specificity, the dominance of misclassified sequence types was in line with the composition of benchmark datasets (**Fig. S6a**). PPR-Meta might benefit from its modeling of chromosomes and phages, while it still had a low specificity mainly due to the misclassification of prokaryotic and eukaryotic host sequences into plasmids (**Fig. S6b**). It's noteworthy that DeepMicrobeFinder might further benefit from its modeling of eukaryotic hosts and viruses since eukaryotic host sequences were rarely classified

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.26.466018; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

as plasmids, though the misclassification rates between plasmids and prokaryotic hosts were still high (**Fig. S7a**). The performance of DeepMicrobeFinder can be improved by collapsing prokaryotic hosts and plasmids into one sequence class as demonstrated previously (**Fig. 1**), which also greatly reduced the total misclassified cases (**Fig. S7b**). This suggests that current tools are inefficient in distinguishing plasmids from prokaryotic hosts, and further improvements on the neural network structures or using additional features extracted from gene or operon centric approaches might yield a better classifier.

## DeepMicrobeFinder achieves improved results in viral sequence prediction

Viruses are ubiquitously found in every natural system where cellular organisms colonize. Significant advances have been made in recent years in developing tools to identify viral contigs from metagenomic assemblies, using essentially gene-centric (e.g., VirSorter, VIBRANT) or oligonucleotide-centric (e.g., VirFinder, DeepVirFinder, PPR-Meta) approaches. Here we compared the performance of DeepMicrobeFinder to VirSorter, VIBRANT, and PPR-Meta on viral contig prediction using the aforementioned benchmark datasets. DeepMicrobeFinder achieved better performance in terms of accuracy and F1 score than all the other tools in all tested datasets (pairwise Wilcoxon test $p$-values $\leq$ 1.1e-05; **Fig. 4 & S8**). The accuracies and F1 scores of DeepMicrobeFinder were rarely lower than 0.9, while none of the other tested tools had an accuracy or F1 score higher than 0.9. In addition, with the share of EUK sequences increasing, the accuracies and F1 scores of VIBRANT, VirSorter, and PPR-Meta decreased, while DeepMicrobeFinder kept constant or slightly increased. VIBRANT and PPR-Meta showed slightly higher accuracy scores than VirSorter in most cases, while within each fixed PROK fraction, both of them showed a decreasing tendency in both accuracies and F1 scores with the increasing of non-host sequences, suggesting higher proportions of virus and plasmid sequences can degrade the performance of VIBRANT and PPR-Meta (**Fig. 4a**). In contrast, the performance of VirSorter was less variable within each fixed PROK:EUK group, and the accuracy and F1 scores could be higher than VIBRANT and PPR-Meta in cases where the host percentages were lowest (host:virus:plasmid ratio of 2:1:1 for PROK, and host:virus ratio of 2:1 for EUK) (**Fig. 4**). This is particularly true for PPR-Meta in terms of F1 scores, which could rapidly decline to lower than 0.7 when the host sequences were less dominant (**Fig. 4b**). A previous benchmarking study showed that VirSorter had a slightly higher specificity than VIBRANT on distinguishing plasmid sequences (Kieft et al., 2020), indicating the higher plasmid fraction in the benchmark datasets might degenerate the performance of VIBRANT and PPR-Meta. The misclassified sequences by VirSorter were mainly bacteriophages and eukaryotic viruses, indicating it suffered from low sensitivity (**Fig. S9a**). VIBRANT also showed low sensitivity in predicting bacteriophages and eukaryotic viruses, while it also frequently classified plasmids and prokaryotic genomes as viruses (**Fig. S9b**). The rapid deterioration of PPR-Meta with increasing EUK fraction can be attributed to its low specificity by misclassifying eukaryotes to phages and its low sensitivity by misclassifying eukaryotic viruses to chromosomes (**Fig. 6b**).

Both DeepMicrobeFinder and PPR-Meta are multiclass classifiers, here we also compared the accuracies and F1 scores of them on these custom test datasets (**Fig. S10**). Although DeepMicrobeFinder can classify more sequence classes than PPR-Meta, it still outperformed PPR-Meta in all tested cases in both accuracies and F1 scores (pairwise Wilcoxon test $p$-values $\leq$ 1.9e-06; **Fig. S10 & S11**).

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.26.466018; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

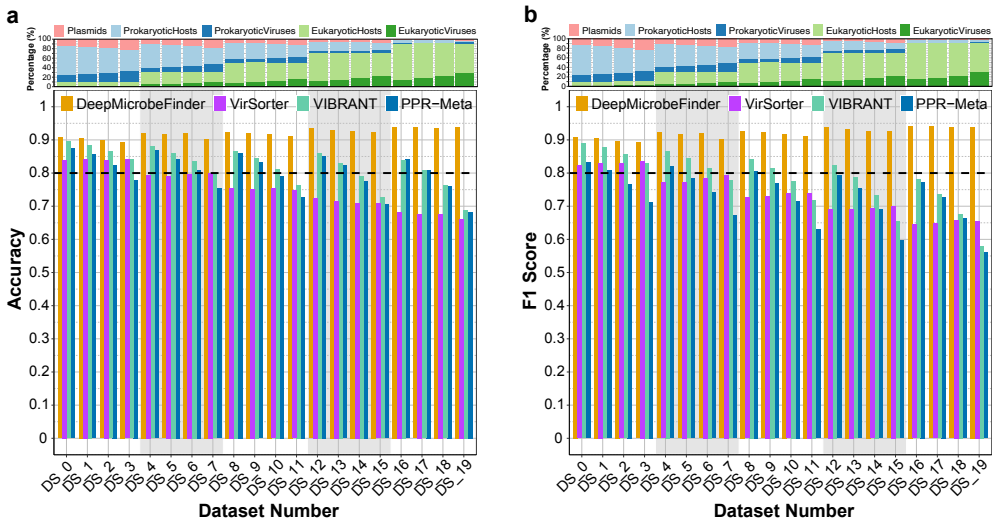


**Fig 4. Distribution of (a) accuracy and (b) F1 scores across 20 test datasets for DeepMicrobeFinder, VirSorter, VIBRANT and PPR-Meta on viral contig classification.** DeepMicrobeFinder received the highest scores in both accuracy and F1 score in all tested scenarios compared to the other predictors. Increasing the fraction of eukaryotic related sequences didn't impaired the performance of DeepMicrobeFinder, but did for the other tools. The dashed black lines indicate where accuracy or F1 score equals to 0.8. Same benchmarking datasets were used as in **Fig. 2**.

## DeepMicrobeFinder predicted more eukaryotic and viral contigs than reference-based predictors

Reference-based classifiers can suffer from incomplete genomic databases, particularly for complex natural environments such as marine or soil systems. To test the performance of DeepMicrobeFinder in real metagenomic context, here we examined its performance with the other two sequence classifiers, Kaiju (Menzel et al., 2016) and MetaEuk (Levy Karin et al., 2020a), using a 1-300 µm size fraction marine metagenomic dataset sampled off the coast of Southern California (Needham et al., 2018). Using the co-assembled contigs as the reference, we show DeepMicrobeFinder classified less prokaryotic but more eukaryotic, eukaryotic viral and prokaryotic viral contigs than Kaiju and MetaEuk (**Fig. 5a**). Among all the prokaryotic contigs classified by both Kaiju and MetaEuk, 73.6% of them were predicted to be prokaryotic by DeepMicrobeFinder, and 11.88%, 10.39%, and 4.14% of them were predicted to be eukaryotic, prokaryotic viral and eukaryotic viral sequences, respectively (**Fig. 5b**). Contigs that couldn't be taxonomically determined by Kaiju (16.41%) or MetaEuk (10.01%) are mainly dominated by eukaryotic sequences (57.13% / 38.3%) as predicted by DeepMicrobeFinder (**Fig. 5c & 5d**). Although MetaEuk classified more eukaryotic contigs than Kaiju (21.88% vs 15.26%, **Fig. 5a**), the latter classified more prokaryotic viral contigs (4.38% vs 1.51%, **Fig. 5a**). This is consistent with the higher percentage of prokaryotic viral sequences in the unclassified contigs of MetaEuk than Kaiju (28.86% vs 14.87%, **Fig. 5c & 5d**). By mapping reads to reference contigs, we calculated the read percentages recruited by different sequence types. The average eukaryotic read percentage recruited by DeepMicrobeFinder (17.86%) is considerably higher than by MetaEuk (13.74%) or Kaiju (9.88%), at the expense of lower prokaryotic read percentages (17.24%, 34.36% and 33.71%, respectively, **Fig. 5f-h**). Similarly, the average read percentages of prokaryotic viral and eukaryotic viral sequences recruited by DeepMicrobeFinder (7.89%/3.24%) are also higher than MetaEuk (0.75%/0.31%) and Kaiju (2.49%/0.64%) (**Fig. 5f-h**). Notably, though DeepMicrobeFinder assigned less prokaryotic and more eukaryotic reads than other classifiers, the relative abundance profiles across the whole time series are highly correlated (**Fig. 12a**

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.26.466018; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

& **12b**), and to a less extent for the prokaryotic viral read percentage profiles (**Fig. 12c**). This is not the case for eukaryotic viral read abundance profiles, where Kaiju and MetaEuk are highly correlated, but not to DeepMicrobeFinder (**Fig. 12d**). To sum up, DeepMicrobeFinder is more correlated with MetaEuk in eukaryotic read profiles, and more correlated with Kaiju in prokaryotic and prokaryotic viral read profiles.

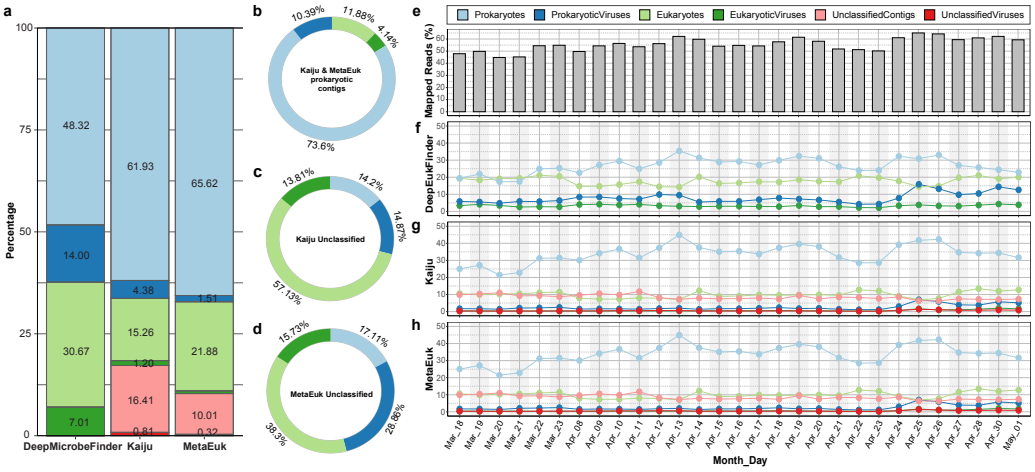


**Fig 5. Sequence classification and read abundance of a 1-300 µm size fraction marine metagenomic dataset sampled off the coast of Southern California.** Metagenomic contigs were classified using DeepMicrobeFinder, Kaiju and MetaEuk at a length cutoff of 2 kb, and percentages of different sequence types were calculated (a). Contigs predicted as Prokaryotes by both Kaiju and MetaEuk (b), and contigs that were not classified by Kaiju (c) or MetaEuk (d) were further broken down into DeepMicrobeFinder's classification. Clean reads were aligned to metagenomic contigs and percentages of mappable reads were calculated (e). Mapped read percentages were further summarized according to sequence types of reference contigs as predicted by DeepMicrobeFinder (f), Kaiju (g) and MetaEuk (h). Prokaryotes included both prokaryotic hosts and plasmids. UnclassifiedViruses were sequences predicted to be viruses but their taxonomy couldn't be further resolved by Kaiju or MetaEuk. Same benchmarking datasets were used as in **Fig. 2**.

## Abundant microbial eukaryotes are correlated with potential prokaryotic community growth rates

If we assume higher eukaryotic read percentages can be a proxy of higher eukaryotes-derived particles, will higher percentages of eukaryotic reads result in faster prokaryotic community growth? Here we calculated the prokaryotic community maximum doubling time using gRodon with species abundance correction (Weissman et al., 2021) and found a significant positive correlation between centered log-ratio (CLR) transformed eukaryotic read percentages and prokaryotic potential community growth rates (**Fig. 6a & 6b**), suggesting higher relative abundances of eukaryotes might brew more fast-growing particle-attached copiotrophs. Significant correlations ($p$-values $< 0.01$) can also be observed between the relative abundance profiles of prokaryotic and eukaryotic viral sequences, as well as between eukaryotic and prokaryotic viral read profiles (**Fig. 6a**). In contrast, the relative read abundance profiles of prokaryotic and prokaryotic viral sequences were negatively correlated with prokaryotic community potential growth rates, though this correlation relationship was less significant (**Fig. 6a & 6b**).

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.26.466018; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

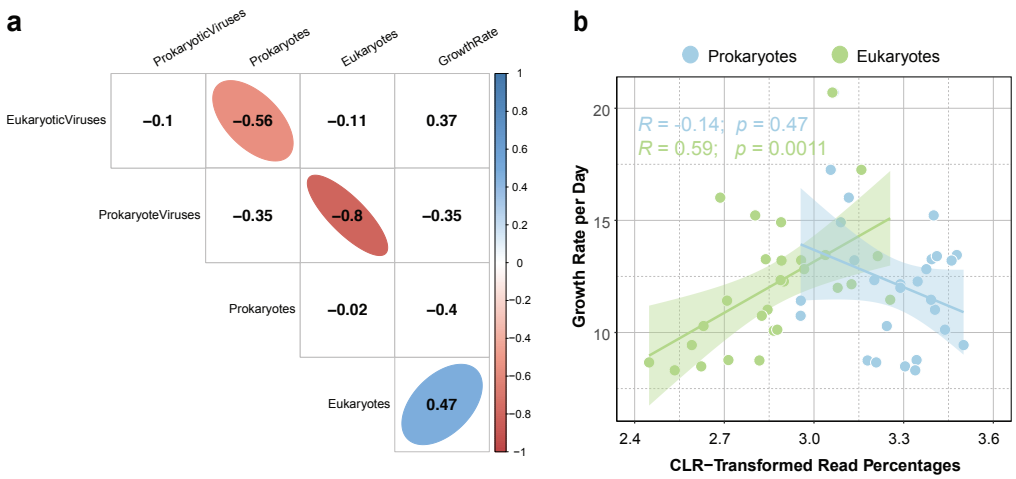


**Fig 6. Positive correlation of prokaryotic community potential growth rates and eukaryotic read percentages.** Correlation coefficients were calculated between relative abundances of different sequence types and prokaryotic community potential growth rates (a). Prokaryotic community potential growth rates (b) as a linear function of centered log-ratio (CLR) transformed read percentages of eukaryotes and prokaryotes. Numbers in (a) were Pearson's $r$ correlation coefficients, only significant ($p$-values $\leq 0.01$) correlations were highlighted in colors. CLR transformation was performed using the mclr function from the SPRING package (Yoon et al., 2019) before performing correlation or regression studies. The potential growth rate of sample Mar_20 were determined to be outliers based on the Grubbs test ($p$-value = 5.4e-08), thus excluded from the regression analysis.

# Discussion

## Microbial eukaryotes and viruses infecting them are understudied

Microbial eukaryotes are prevalent in diverse ecosystems such as host-associated habitats (Parfrey et al., 2011), deep-sea benthos (Bik et al., 2012), and geothermal springs (Oliverio et al., 2018), etc. Due to challenges in cultivation and whole genome-sequencing of microbial eukaryotes, biodiversity surveys of microbial eukaryotes were commonly performed using marker genes, such as the 18S rDNA hypervariable V4 or V9 regions (Pawlowski et al., 2012; Amaral-Zettler et al., 2009). The amplicon-based analysis provides valuable information on the taxonomy of microbial eukaryotes, while in order to probe their metabolic potentials or ecological functions, genomic and transcriptomic information are essential. Despite several achievements in collecting microbial eukaryotic genes (Carradec et al., 2018; Vorobev et al., 2020), transcripts (Keeling et al., 2014) or single-cell amplified genomes (SAGs) (Sieracki et al., 2019) towards a comprehensive microbial eukaryotic database, our knowledge are still limited by the availability of diverse microbial eukaryotic genomes ](Burki et al., 2020). With the rapid accumulation of metagenomic datasets and availability of binning software, it's appealing to recover eukaryotic genomes from natural microbial communities. EukRep was developed in such a context to identify eukaryotic contigs for metagenomic binning (West et al., 2018). This approach has enabled the genome-resolved analysis of fungi, protists, and rotifers from human microbiome studies (West et al., 2018; Olm et al., 2019). Similar approaches have been applied to marine microbiome studies (Duncan et al., 2020; Delmont et al., 2020), which recovered hundreds of microbial eukaryotic MAGs and provided insight into the functional diversity and evolutionary histories of microbial eukaryotes beyond the taxonomic information.

Beyond microbial eukaryotes, current viromic studies are biased towards viruses infecting prokaryotes. This could be introduced by the skewed distribution of viral genomes in the RefSeq database, which is dominated by phages and pathogenic viruses. By Jan 23, 2021, among 10,161 viral reference

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.26.466018; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

genomes, there are only 33 records belonging to algae-infecting Phycodnaviridae and 6 belonging to protists-infecting Mimiviridae. Both of the two viral families are subgroups of the Nucleocytoplasmic Large DNA Viruses (NCLDV) (Iyer et al., 2001). Since most of the commonly used viral predictors are trained on the RefSeq viral database, it's expected that these tools suffered from identifying eukaryotic viruses from the test datasets (**Fig. 4 & S2**). Given the high diversity of protists (Foissner, 1999; Slapeta et al., 2005), high throughput metagenomes and single-cell genomes are expected to offer a culture-independent solution to rapidly expand the coverage of viral database. For instance, two recent studies reconstructed 2,074 and 501 NCLDV MAGs from global environmental metagenomes (Schulz et al., 2020; Moniruzzaman et al., 2020), dramatically increased the phylogenetic and functional diversity of NCLDVs. Single-cell metagenomics was also employed to identify viruses infecting marine microbial eukaryotes (Needham et al., 2019b,a), these studies provided insightful findings of the viral encoded proteins and metabolic pathways.

These studies demonstrated that metagenomics and single-cell genomics can be promising in studying microbial eukaryotes and viruses infecting them. While most commonly used tools are not optimized in classifying eukaryotes (**Fig. 2 & S3**) or eukaryotic viruses (**Fig. 4 & S2**). Given the high performance of DeepMicrobeFinder and the evidence of abundant eukaryotic contigs in marine ecosystems (**Fig. 5 & 6**), we expect it will be a valuable addition to the toolbox of marine ecologists.

## The challenge of classifying prokaryotic host and plasmid sequences

DeepMicrobeFinder has a relatively lower accuracy in classifying plasmids when compared to the classification of eukaryotic or viral contigs (**Fig. 2, 3 & 4**). The majority of the sequences that were misclassified as plasmids were from prokaryotic host genomes (**Fig. S7**), confirming classifying prokaryotic chromosomal and plasmid sequences is a caveat of DeepMicrobeFinder (**Fig. 1**). In comparison, the other tested plasmid classifiers suffered from both prokaryotic and eukaryotic sequences as we have benchmarked (**Fig. 3, S5 & S6**). It's noteworthy that this marginal advantage can be crucial in natural environments, such as the marine environments as we mentioned here (**Fig. 5 & 6**), where eukaryotic sequences can have a substantial impact on the performance of plasmid sequence classifiers. This also indicates that it is achievable to separate plasmid sequences from eukaryotic sequences solely based on patterns of oligonucleotides, and current plasmid predictors can benefit from using a more comprehensive training dataset including eukaryotic sequences.

It is understandable given the higher genome complexity of eukaryotes than prokaryotes (Lynch & Conery, 2003), such as the coding density, prevalence of introns and repetitive sequences, etc. In contrast, it's challenging to classify plasmids and prokaryotic chromosomal sequences for all the tested plasmid predictors (**Fig. 3**). The reasons can be manifold, but plasmid transmission among microbial hosts and plasmid-chromosome gene shuffling can be two fundamental ones. The host range of plasmids is variable, it can be within closely related species for narrow host range plasmids, or across distant phylogenetic groups for broad host range plasmids (Jain & Srivastava, 2013). Broad host range plasmids can be important drivers of the gene flux among host microbes in natural environments (Heuer & Smalla, 2007; Wolska, 2003; Davison, 1999). For instance, in natural soil microbial communities, the IncP- and IncPromA-type broad host range plasmids were found to be able to transfer from proteobacteria to diverse bacteria belonging to 11 bacterial phyla (Klümper et al., 2015). When plasmid carriage could increase the fitness of the hosts, such as improving host survival with antibiotic resistance, it can be rapidly

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.26.466018; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

adopted and persistently maintained in natural microbial communities (Li et al., 2020; Bellanger et al., 2014). On the other hand, when the maintenance of plasmids imposed a high fitness cost on the hosts, plasmids or plasmid-borne genes could be lost as in the process of purifying selection (Hall et al., 2016). Interestingly, studies also suggested sometimes this fitness cost could be ameliorated by compensatory evolution (Millan et al., 2014; Harrison et al., 2015; Loftie-Eaton et al., 2017), which was hypothesized to be the major factor of plasmid survival and persistence (Hall et al., 2017). Plasmid carriage also increases the chance of plasmid-chromosome genetic exchange mediated by SOS-induced mutagenesis citeprodriguez-beltranHorizontalGeneTransfer2021 or mobile genetic elements such as transposons and integrons, etc citepfrostMobileGeneticElements2005,rodriguez-beltranHorizontalGeneTransfer2021. For instance, genes carried by transposons or in the variable regions were also frequently found on plasmids (Eberhard, 1990; Zheng et al., 2015). Thus, the permissive transfer of plasmids across diverse hosts and the plasmid-chromosome gene flow pose a challenge for current plasmid classifiers. The oligonucleotide-based approaches might be complemented by gene-centric approaches using plasmid signature genes or enriched gene functions, such as genes involved in mobilization or conjugation. In addition, a comprehensive plasmid database is also crucial for model training, and plasmid enriched metagenomics (plasmidome) can be a promising way to screen plasmids from environmental samples (Shi et al., 2018).

## Ecological influence of microbial eukaryotes on the prokaryotic community

Marine water is a continuum of particles, which are aggregates of diverse planktonic detritus and minerals, providing nutrient rich microenvironments in the oligotrophic oceans (Simon et al., 2002). Size-fractionated metagenomes are commonly used to study microbial lifestyles in aquatic environments. Prokaryotes commonly found in larger size fractions (>1 µm) prefer a particle-associated (PA) lifestyle, while microbes dominating the smaller size fraction lead a free-living (FL) lifestyle (Grossart, 2010). Bacterial production experiments showed that though PA microbes were less abundant, their growth rates could be 20-fold higher than their FL counterparts (Friedrich et al., 1999). Fast-replicating PA bacteria were also tightly linked to chlorophyll a, suggesting their growth might be fueled by particulate organic matter (POM) derived from phytoplankton (Friedrich et al., 1999). PA microbes are also found to be associated with the degradation of hydrocarbons and lipid materials derived from eukaryotic plankton (Yoshimura et al., 2009; Wei et al., 2013; Fontanez et al., 2015), playing an important role in the POM remineralization and biogeochemical cycles. Nutrient-demanding copiotrophs (here *Vibrio*, *Roseovarius*, *Polaribacter*, etc, according to Needham et al. (2018)) are usually PA microbes, while oligotrophs (here SAR11, *Prochlorococcus*, etc) are FL adapted to nutrient poor environments (Giovannoni et al., 2014). Copiotrophs encode more genes involved in carbohydrate and amino acid transport and metabolism than oligotrophs (Weissman et al., 2021), and can be classified solely based on the minimum doubling time (<5 hours), which correlates with the codon usage bias (CUB) of highly expressed genes (such as ribosomal genes) due to the selection for translational efficiency (Vieira-Silva & Rocha, 2010; Long et al., 2021; Weissman et al., 2021). Thus, the increase of eukaryotic read proportion can be used as a proxy of higher POM availability, which promotes the growth of fast-growing particle degraders and changes the prokaryotic community composition. Conversely, using a species abundance aware community growth rate prediction method (Weissman et al., 2021), as we have done here, one can also probe the relative nutrient or POM status of given samples based on the predicted potential community growth rates from metagenomes.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.26.466018; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

# Conclusions

DeepMicrobeFinder 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 DeepMicrobeFinder attain better performance than the other state-of-the-art individual predictors due to reduced cross misclassifications. We also detected high abundance of marine eukaryotes in a daily time-series dataset, and further showed that eukaryotic read fractions were positively correlated with prokaryotic community growth rates. Our case study indicates that both host and viral sequences are essential components in the cellular metagenomes, and robust ecological patterns can be obtained with DeepMicrobeFinder even for coarse sequence types. We argue that by using DeepMicrobeFinder 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 DeepMicrobeFinder achieves higher performance than the other benchmarked predictors, and its application can facilitate studies of under-appreciated sequence types, such as microbial eukaryotic or viral sequences.

# Availability of data and materials

The source code and user guide are available at https://github.com/chengsly/DeepMicrobeFinder. Test datasets and scripts used to run different predictors have been deposited at figshare (available at dx.doi.org/10.6084/m9.figshare.14576193). Raw reads for 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.

# Competing interests

The authors declare that they have no competing interests.

# Authors' contributions

SH, JAF, and FS conceived the project; SH, SC, and FS designed the neural network structure and model evaluation procedures; SH designed the training, test datasets and use-case applications; SH and SC prepared the training and test datasets; SC implemented the software with the help of SH; SH and SC performed the data analysis; SH and SC prepared all the figures and tables; SH drafted the manuscript with the help of SC; SC, JAF and FS reviewed and edited the manuscript.

# Acknowledgements

This study was supported by the NIH grants (Grant ID:R01GM120624, 1R01GM131407) to F. Sun, the Simons Collaboration on Computational Biogeochemical Modeling of Marine Ecosystems/CBIOMES) grant (Grant ID: 549943) and the Gordon and Betty Moore Foundation (Grant Number: 3779) to J. Fuhrman, and the NSFC grant (Grant ID: 61872218, 61721003) and National Key R&D Program of China (Grant ID: 2019YFB1404804) to T. Chen. The funders had no roles in study design, data

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.26.466018; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

collection or analysis, the decision to publish, and preparation of the manuscript. 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.

# List of abbreviations

Abbreviations used in this manuscript:

| Abbreviations | Definition |
|---|---|
| AUC | area under the ROC curve |
| ROC | receiver operating characteristics |
| SSU rRNA | small subunit ribosomal RNA |
| REOs | ribosome-encoding cellular organisms |
| CEOs | capsid-encoding organisms |
| HMM | hidden Markov model |
| CNN | convolutional neural network |
| ESP | environmental sample processor |
| HGT | horizontal gene transfer |
| CLR | centered log-ratio |
| NCLDV | nucleocytoplasmic large DNA viruse |
| MAG | metagenome-assembled genome |
| FL | free-living |
| PA | particle-associated |
| CUB | codon usage bias |
| POM | particulate organic matter |

# Supporting information

**Supplemental Table S1. The composition of test datasets used in this study for benchmarking different tools.** PROK includes prokaryotic genomes, plasmids and prokaryotic viruses; EUK includes eukaryotic genomes and viruses. Prok: prokaryotic genomes, ProkVir: prokaryotic viruses/phages, Plas: plasmids, Euk: eukaryotic genomes, EukVir: eukaryotic viruses. Test files were named using the number of sampled sequences from each sequence class following the rule: Prok_ProkVir_Plas_Euk_EukVir_test.fasta, which can be found at dx.doi.org/10.6084/m9.figshare.14576193.

**Supplemental Figure S1. Schematic representation of the multi-class CNN structure used in this study.** The hyperparameters used for each layer are: a) 64 filters with a kernel size of 6 were used for convolution layer 1, followed by relu activation, b) the stride and pooling size were set to 2 for max pooling layer 1, c) 128 filters with a kernel size of 3 were used for convolution layer 2, followed by relu activation, d) the stride was set to 1 and the pooling size was set to 2 for max pooling layer 2, e) 256 filters with a kernel size of 2 were used for convolution layer 3, followed by relu activation, f) a dropout threshold of 0.1 was used for drop out layers, g) 500 hidden units were chosen for the first dense layer, followed by relu activation, h) 5 hidden units were chosen for the last dense layer, followed by softmax activation.

**Supplemental Figure S2. The distribution of viral confidence scores for (a) VirFinder and (b) PPR-Meta.** For both predictors, the same dataset was used and the predictions were performed

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.26.466018; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

with default parameters. VirFinder uses VF-Scores to determine the likelihood of input sequences to be viral or not, and PPR-Meta uses phage scores to discern viruses from host chromosomes and plasmids. Both predictors achieved a high recall for prokaryotic viruses, while the confidence scores of eukaryotic viruses were more evenly spreaded across all confidence regions. Besides, both predictors achieved a high performance in distinguishing prokaryotic host sequences from prokaryotic viruses, but less so for eukaryotic host sequences.

**Supplemental Figure S3. Performance of DeepMicrobeFinder and EukRep on eukaryotic sequence classification.** Both the accuracy (a) and F1 score (b) were compared based on 20 designed test datasets. The sequence class composition of the 20 test datasets can be found in supplemental table 1. Values on top of the pairwise comparisons are Bonferroni adjusted t-test $p$-values.

**Supplemental Figure S4. The distribution of misclassified sequence types by EukRep.** The sequence composition of these datasets can be found in supplementary table Supplemental Table S1.

**Supplemental Figure S5. Performance of DeepMicrobeFinder, PlasFlow, and PPR-Meta on plasmid sequence classification.** Both the accuracy (a) and F1 score (b) were compared based on 20 designed test datasets. The sequence class composition of the 20 test datasets can be found in supplemental table 1. Values on top of the pairwise comparisons are Bonferroni adjusted t-test $p$-values. The significance of the overall ANOVA test was shown on the bottom left corner.

**Supplemental Figure S6. The distribution of misclassified sequence types by PlasFlow (a) and PPR-Meta (b).** The sequence composition of these datasets can be found in supplementary table S1. For each dataset, the total number of test sequences is 1000.

**Supplemental Figure S7. The distribution of misclassified sequence types by DeepMicrobeFinder.** Sequences were classified into 5 classes (a) or 4 classes by collapsing prokaryotic hosts and plasmids into prokaryotes (b). The sequence composition of these datasets can be found in supplementary table Supplemental Table S1. For each dataset, the total number of test sequences is 1000.

**Supplemental Figure S8. Performance of DeepMicrobeFinder, VirSorter, VIBRANT, and PPR-Meta on viral sequence classification.** Both the accuracy (a) and F1 score (b) were compared based on 20 designed test datasets. The sequence class composition of the 20 test datasets can be found in supplemental table 1. Values on top of the pairwise comparisons are Bonferroni adjusted t-test $p$-values. The significance of the overall ANOVA test was shown on the bottom left corner.

**Supplemental Figure S9. The distribution of misclassified sequence types by VirSorter (a) and VIBRABT (b).** The sequence composition of these datasets can be found in supplementary table Supplemental Table S1. For each dataset, the total number of test sequences is 1000.

**Supplemental Figure S10. Distribution of (a) accuracy and (b) F1 scores across 20 test datasets for DeepMicrobeFinder and PPR-Meta on multiclass contig classification.** DeepMicrobeFinder received higher scores in both accuracy and F1 score in all tested scenarios compared to PPR-Meta. DeepMicrobeFinder showed improved performance with increasing fractions of eukaryotic related sequences, while the performance of PPR-Meta severely degraded.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.26.466018; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

**Supplemental Figure S11. Performance of DeepMicrobeFinder and PPR-Meta on multiclass sequence classification.** Both the accuracy (a) and F1 score (b) were compared based on 20 designed test datasets. The sequence class composition of the 20 test datasets can be found in supplemental table Supplemental Table S1.

**Supplemental Figure S12. Correlation coefficients of Prokaryotic (a), Eukaryotic (b), ProkaryoticViral (c), and EukaryoticViral (d) sequence relative abundances of different sequence classifiers.** Coefficients highlighted in colors are significant ones ($p$-value $< 0.01$).

## References

Amaral-Zettler, L. A., McCliment, E. A., Ducklow, H. W., & 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*(7), e6372.

Azam, F., & Worden, A. Z. (2004). Oceanography. microbes, molecules, and marine ecosystems. *Science (New York, N.Y.)*, *303*(5664), 1622–1624.

Bellanger, X., Guilloteau, H., Breuil, B., & Merlin, C. (2014). Natural microbial communities supporting the transfer of the incp-1 plasmid pb10 exhibit a higher initial content of plasmids from the same incompatibility group. *Frontiers in Microbiology*, *0*.
URL `https://www.frontiersin.org/articles/10.3389/fmicb.2014.00637/full`

Berendsen, R. L., Pieterse, C. M. J., & Bakker, P. A. H. M. (2012). The rhizosphere microbiome and plant health. *Trends in Plant Science*, *17*(8), 478–486.

Bik, H. M., Sung, W., Ley, P. D., Baldwin, J. G., Sharma, J., Rocha-Olivares, A., & Thomas, W. K. (2012). Metagenetic community analysis of microbial eukaryotes illuminates biogeographic patterns in deep-sea and shallow water sediments. *Molecular Ecology*, *21*(5), 1048–1059.

Brister, J. R., Ako-adjei, D., Bao, Y., & Blinkova, O. (2015). Ncbi viral genomes resource. *Nucleic Acids Research*, *43*(Database issue), D571–D577.

Burki, F., Roger, A. J., Brown, M. W., & Simpson, A. G. B. (2020). The new tree of eukaryotes. *Trends in Ecology & Evolution*, *35*(1), 43–55.

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. *Environmental Microbiology*, *2*(5), 516–529.

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. *Nature Communications*, *9*(11), 373.

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one fastq preprocessor. *Bioinformatics (Oxford, England)*, *34*(17), i884–i890.

Davison, J. (1999). Genetic exchange between bacteria in the environment. *Plasmid*, *42*(2), 73–91.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.26.466018; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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. (2020). Functional repertoire convergence of distantly related eukaryotic plankton lineages revealed by genome-resolved metagenomics. *bioRxiv*, (p. 2020.10.15.341214).

DeLong, E. F. (1992). Archaea in coastal marine environments. *Proc. Natl. Acad. Sci. U. S. A.*, *89*(12), 5685–5689.

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. (2020). Metagenome-assembled genomes of phytoplankton communities across the arctic circle. *bioRxiv*, (p. 2020.06.16.154583).

Eberhard, W. G. (1990). Evolution in bacterial plasmids and levels of selection. *The Quarterly Review of Biology*, *65*(1), 3–22.

Falkowski, P. G., Fenchel, T., & Delong, E. F. (2008). The microbial engines that drive earth's biogeochemical cycles. *Science*, *320*(5879), 1034–1039.

Fang, Z., Tan, J., Wu, S., Li, M., Xu, C., Xie, Z., & Zhu, H. (2019). Ppr-meta: a tool for identifying phages and plasmids from metagenomic fragments using deep learning. *GigaScience*, *8*(6).
URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6586199/`

Foissner, W. (1999). Protist diversity: estimates of the near-imponderable. *Protist*, *150*(4), 363–368.

Fontanez, K. M., Eppley, J. M., Samo, T. J., Karl, D. M., & DeLong, E. F. (2015). Microbial community structure and function on sinking particles in the north pacific subtropical gyre. *Frontiers in Microbiology*, *6*.
URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4436931/`

Friedrich, U., Schallenberg, M., & Holliger, C. (1999). Pelagic bacteria-particle interactions and community-specific growth rates in four lakes along a trophic gradient. *Microbial Ecology*, *37*(1), 49–61.

Fuhrman, J. A. (1992). Novel major archaebacterial group from marine plankton. *Nature*, *356*(6365), 148–149.

Fuhrman, J. A. (1999). Marine viruses and their biogeochemical and ecological effects. *Nature*, *399*(67366736), 541–548.

Gałan, W., Bąk, M., & Jakubowska, M. (2019). Host taxon predictor - a tool for predicting taxon of the host of a newly discovered virus. *Scientific Reports*, *9*(1), 3436.

Giovannoni, S. J., Cameron Thrash, J., & Temperton, B. (2014). Implications of streamlining theory for microbial ecology. *The ISME journal*, *8*(8), 1553–1565.

Grossart, H.-P. (2010). Ecological consequences of bacterioplankton lifestyles: changes in concepts are needed. *Environmental Microbiology Reports*, *2*(6), 706–714.

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 Research*, *41*(D1), D597–D604.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.26.466018; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Hall, J. P. J., Brockhurst, M. A., Dytham, C., & Harrison, E. (2017). The evolution of plasmid stability: Are infectious transmission and compensatory evolution competing evolutionary trajectories? *Plasmid*, *91*, 90–95.

Hall, J. P. J., Wood, A. J., Harrison, E., & Brockhurst, M. A. (2016). Source-sink plasmid transfer dynamics maintain gene mobility in soil bacterial communities. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(29), 8260–8265.

Handelsman, J. (2004). Metagenomics: application of genomics to uncultured microorganisms. *Microbiology and molecular biology reviews: MMBR*, *68*(4), 669–685.

Harrison, E., Guymer, D., Spiers, A. J., Paterson, S., & Brockhurst, M. A. (2015). Parallel compensatory evolution stabilizes plasmids across the parasitism-mutualism continuum. *Current biology: CB*, *25*(15), 2034–2039.

Heuer, H., & Smalla, K. (2007). Horizontal gene transfer between bacteria. *Environmental Biosafety Research*, *6*(1–21–2), 3–13.

Iyer, L. M., Aravind, L., & Koonin, E. V. (2001). Common origin of four diverse families of large eukaryotic dna viruses. *Journal of Virology*, *75*(23), 11720–11734.

Jain, A., & Srivastava, P. (2013). Broad host range plasmids. *FEMS Microbiology Letters*, *348*(2), 87–96.

Johnson, L. K., Alexander, H., & Brown, C. T. (2019). Re-assembly, quality evaluation, and annotation of 678 microbial eukaryotic reference transcriptomes. *GigaScience*, *8*(4).
URL `https://doi.org/10.1093/gigascience/giy158`

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 Biology*, *12*(6), e1001889.

Kieft, K., Zhou, Z., & Anantharaman, K. (2020). Vibrant: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome*, *8*(1), 90.

Klümper, U., Riber, L., Dechesne, A., Sannazzarro, A., Hansen, L. H., Sørensen, S. J., & Smets, B. F. (2015). Broad host range plasmids can invade an unexpectedly diverse fraction of a soil bacterial community. *The ISME Journal*, *9*(4), 934–945.

Krawczyk, P. S., Lipinski, L., & Dziembowski, A. (2018). Plasflow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Research*, *46*(6), e35–e35.

Legault, B. A., Lopez-Lopez, A., Alba-Casado, J. C., Doolittle, W. F., Bolhuis, H., Rodriguez-Valera, F., & 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*(1), 171.

Levy Karin, E., Mirdita, M., & Söding, J. (2020a). Metaeuk-sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. *Microbiome*, *8*(1), 48.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.26.466018; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Levy Karin, E., Mirdita, M., & Söding, J. (2020b). Metaeuk—sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. *Microbiome*, *8*(1), 48.

Li, L., Dechesne, A., Madsen, J. S., Nesme, J., Sørensen, S. J., & Smets, B. F. (2020). Plasmids persist in a microbial community by providing fitness benefit to multiple phylotypes. *The ISME Journal*, *14*(5), 1170–1181.

Li, W., & Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22*(13), 1658–1659.

Loftie-Eaton, W., Bashford, K., Quinn, H., Dong, K., Millstein, J., Hunter, S., Thomason, M. K., Merrikh, H., Ponciano, J. M., & Top, E. M. (2017). Compensatory mutations improve general permissiveness to antibiotic resistance plasmids. *Nature Ecology & Evolution*, *1*(9), 1354–1363.

Long, A. M., Hou, S., Ignacio-Espinoza, J. C., & Fuhrman, J. A. (2021). Benchmarking microbial growth rate predictions from metagenomes. *The ISME Journal*, *15*(11), 183–195.

Lynch, M., & Conery, J. S. (2003). The origins of genome complexity. *Science*, *302*(5649), 1401–1404.

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*(7057), 376–380.

Menzel, P., Ng, K. L., & Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with kaiju. *Nature Communications*, *7*, 11257.

Millan, A. S., Peña-Miller, R., Toll-Riera, M., Halbert, Z. V., McLean, A. R., Cooper, B. S., & MacLean, R. C. (2014). Positive selection and compensatory adaptation interact to stabilize non-transmissible plasmids. *Nature Communications*, *5*(1), 5208.

Moniruzzaman, M., Martinez-Gutierrez, C. A., Weinheimer, A. R., & Aylward, F. O. (2020). Dynamic genome evolution and complex virocell metabolism of globally-distributed giant viruses. *Nature Communications*, *11*(11), 1–11.

Needham, D. M., Fichot, E. B., Wang, E., Berdjeb, L., Cram, J. A., Fichot, C. G., & Fuhrman, J. A. (2018). Dynamics and interactions of highly resolved marine plankton via automated high-frequency sampling. *The ISME Journal*, (p. 1).

Needham, D. M., Poirier, C., Hehenberger, E., Jiménez, V., Swalwell, J. E., Santoro, A. E., & Worden, A. Z. (2019a). Targeted metagenomic recovery of four divergent viruses reveals shared and distinctive characteristics of giant viruses of marine eukaryotes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *374*(1786), 20190086.

Needham, D. M., Yoshizawa, S., Hosaka, T., Poirier, C., Choi, C. J., Hehenberger, E., Irwin, N. A. T., Wilken, S., Yung, C.-M., Bachy, C., & et al. (2019b). A distinct lineage of giant viruses brings a rhodopsin photosystem to unicellular marine predators. *Proceedings of the National Academy of Sciences*, (p. 201907517).

Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). metaspades: a new versatile metagenomic assembler. *Genome Research*, *27*(5), 824–834.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.26.466018; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Oliverio, A. M., Power, J. F., Washburne, A., Cary, S. C., Stott, M. B., & Fierer, N. (2018). The ecology and diversity of microbial eukaryotes in geothermal springs. *The ISME Journal*, *12*(88), 1918–1928.

Olm, M. R., West, P. T., Brooks, B., Firek, B. A., Baker, R., Morowitz, M. J., & Banfield, J. F. (2019). Genome-resolved metagenomics of eukaryotic populations during early colonization of premature infants and in hospital rooms. *Microbiome*, *7*(1), 26.

Olsen, G. J., Lane, D. J., Giovannoni, S. J., Pace, N. R., & Stahl, D. A. (1986). Microbial ecology and evolution: a ribosomal rna approach. *Annual Review of Microbiology*, *40*, 337–365.

Pace, N. R., Stahl, D. A., Lane, D. J., & Olsen, G. J. (1986). The analysis of natural microbial populations by ribosomal rna sequences. In K. C. Marshall (Ed.) *Advances in Microbial Ecology*, Advances in Microbial Ecology, (p. 1–55). Springer US.
URL `https://doi.org/10.1007/978-1-4757-0611-6_1`

Paez-Espino, D., Roux, S., Chen, I.-M. A., Palaniappan, K., Ratner, A., Chu, K., Huntemann, M., Reddy, T. B. K., Pons, J. C., Llabrés, M., & et al. (2019). Img/vr v.2.0: an integrated data management and analysis system for cultivated and environmental viral genomes. *Nucleic Acids Research*, *47*(D1), D678–D686.

Parfrey, L. W., Walters, W. A., & Knight, R. (2011). Microbial eukaryotes in the human microbiome: ecology, evolution, and future directions. *Frontiers in Microbiology*, *2*, 153.

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 Biology*, *10*(11), e1001419.

Pellow, D., Mizrahi, I., & Shamir, R. (2020). Plasclass improves plasmid sequence classification. *PLoS computational biology*, *16*(4), e1007781.

Raoult, D., & Forterre, P. (2008). Redefining viruses: lessons from mimivirus. *Nature Reviews Microbiology*, *6*(4), 315–319.

Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A., & Sun, F. (2017). Virfinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome*, *5*, 69.

Ren, J., Song, K., Deng, C., Ahlgren, N. A., Fuhrman, J. A., Li, Y., Xie, X., Poplin, R., & Sun, F. (2020). Identifying viruses from metagenomic data using deep learning. *Quantitative Biology*.
URL `https://doi.org/10.1007/s40484-019-0187-4`

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. *Applied and Environmental Microbiology*, *66*(6), 2541–2547.

Roux, S., Enault, F., Hurwitz, B. L., & Sullivan, M. B. (2015). Virsorter: mining viral signal from microbial genomic data. *PeerJ*, *3*, e985.

Royer, G., Decousser, J. W., Branger, C., Dubois, M., Médigue, C., Denamur, E., & Vallenet, D. (2018). Plascope: a targeted approach to assess the plasmidome from genome assemblies at the species level. *Microbial Genomics*, *4*(9).

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.26.466018; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Schmidt, T. M., DeLong, E. F., & Pace, N. R. (1991). Analysis of a marine picoplankton community by 16s rrna gene cloning and sequencing. *Journal of Bacteriology*, *173*(14), 4371–4378.

Schulz, F., Roux, S., Paez-Espino, D., Jungbluth, S., Walsh, D., Denef, V. J., McMahon, K. D., Konstantinidis, K. T., Eloe-Fadrosh, E. A., Kyrpides, N., & et al. (2020). Giant virus diversity and host interactions through global metagenomics. *Nature*, (p. 1–7).

Sedlar, K., Kupkova, K., & Provaznik, I. (2017). Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics. *Computational and Structural Biotechnology Journal*, *15*, 48–55.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics (Oxford, England)*, *30*(14), 2068–2069.

Shen, W., & Ren, H. (2021). Taxonkit: A practical and efficient ncbi taxonomy toolkit. *Journal of Genetics and Genomics*.
URL `https://www.sciencedirect.com/science/article/pii/S1673852721000837`

Shi, Y., Zhang, H., Tian, Z., Yang, M., & Zhang, Y. (2018). Characteristics of arg-carrying plasmidome in the cultivable microbial community from wastewater treatment system under high oxytetracycline concentration. *Applied Microbiology and Biotechnology*, *102*(4), 1847–1858.

Sieracki, M. E., Poulton, N. J., Jaillon, O., Wincker, P., Vargas, C. d., Rubinat-Ripoll, L., Stepanauskas, R., Logares, R., & Massana, R. (2019). Single cell genomics yields a wide diversity of small planktonic protists across major ocean ecosystems. *Scientific Reports*, *9*(1), 1–11.

Simon, M., Grossart, H.-P., Schweitzer, B., & Ploug, H. (2002). Microbial ecology of organic aggregates in aquatic ecosystems. *Aquatic Microbial Ecology*, *28*(2), 175–211.

Slapeta, J., Moreira, D., & López-García, P. (2005). The extent of protist diversity: insights from molecular ecology of freshwater eukaryotes. *Proceedings. Biological Sciences*, *272*(1576), 2073–2081.

Stein, J. L., Marsh, T. L., Wu, K. Y., Shizuya, H., & DeLong, E. F. (1996). Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. *Journal of Bacteriology*, *178*(3), 591–599.

Suttle, C. A. (2005). Viruses in the sea. *Nature*, *437*(7057), 356–361.

Suttle, C. A. (2007). Marine viruses — major players in the global ecosystem. *Nature Reviews Microbiology*, *5*(1010), 801–812.

Treangen, T. J., Sommer, D. D., Angly, F. E., Koren, S., & Pop, M. (2011). Next generation sequence assembly with amos. *Current Protocols in Bioinformatics*, *Chapter 11*, Unit 11.8.

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., & Gordon, J. I. (2007). The human microbiome project. *Nature*, *449*(7164), 804–810.

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*(5667), 66–74.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.26.466018; this version posted October 27, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Vergin, K. L., Urbach, E., Stein, J. L., DeLong, E. F., Lanoil, B. D., & 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. *Applied and Environmental Microbiology*, *64*(8), 3075–3078.

Vieira-Silva, S., & Rocha, E. P. C. (2010). The systemic imprint of growth and its uses in ecological (meta)genomics. *PLOS Genetics*, *6*(1), e1000808.

Vorobev, A., Dupouy, M., Carradec, Q., Delmont, T. O., Annamalé, A., Wincker, P., & Pelletier, E. (2020). Transcriptome reconstruction and functional analysis of eukaryotic marine plankton communities via high-throughput metagenomics and metatranscriptomics. *Genome Research*, *30*(4), 647–659. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press publisher: Cold Spring Harbor Lab PMID: 32205368.

Wei, N., Quarterman, J., & Jin, Y.-S. (2013). Marine macroalgae: an untapped resource for producing fuels and chemicals. *Trends in Biotechnology*, *31*(2), 70–77.

Weissman, J. L., Hou, S., & Fuhrman, J. A. (2021). Estimating maximal microbial growth rates from cultures, metagenomes, and single cells via codon usage patterns. *Proceedings of the National Academy of Sciences*, *118*(12).
URL `https://www.pnas.org/content/118/12/e2016810118`

West, P. T., Probst, A. J., Grigoriev, I. V., Thomas, B. C., & Banfield, J. F. (2018). Genome-reconstruction for eukaryotes from complex natural microbial communities. *Genome Research*, *28*(4), 569–580.

Wilhelm, S. W., & 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*(10), 781–788.

Woese, C. R., & Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(11), 5088–5090.

Wolska, K. I. (2003). Horizontal dna transfer between bacteria in the environment. *Acta Microbiologica Polonica*, *52*(3), 233–243.

Yoon, G., Gaynanova, I., & Müller, C. L. (2019). Microbial networks in spring - semi-parametric rank-based correlation and partial correlation estimation for quantitative microbiome data. *Frontiers in Genetics*, *10*, 516.

Yoshimura, K., Ogawa, T., & Hama, T. (2009). Degradation and dissolution properties of photosynthetically-produced phytoplankton lipid materials in early diagenesis. *Marine Chemistry*, *114*(1), 11–18.

Zheng, J., Guan, Z., Cao, S., Peng, D., Ruan, L., Jiang, D., & Sun, M. (2015). Plasmids are vectors for redundant chromosomal genes in the bacillus cereus group. *BMC Genomics*, *16*(1), 6.

Zhou, F., & Xu, Y. (2010). cbar: a computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data. *Bioinformatics*, *26*(16), 2051–2052.