vAMPirus: A versatile amplicon processing and analysis program for studying viruses

Alex Veglia[1], Ramon E Rivera-Vicens[2], Carsten Grupstra[3], Lauren Howe-Kerr[1], and Adrienne Correa[1]

[1]Rice University
[2]Inter American University of Puerto Rico - Barranquitas Campus
[3]Boston University

February 9, 2023

Abstract

Amplicon sequencing is an effective and increasingly applied method for studying viral communities in the environment. Here, we present vAMPirus, a user-friendly, comprehensive, and versatile DNA and RNA virus amplicon sequence analysis program, designed to support investigators in exploring virus amplicon sequencing data and running informed, reproducible analyses. vAMPirus intakes raw virus amplicon libraries and, by default, performs nucleotide- and protein-based analyses to produce results such as sequence abundance information, taxonomic classifications, phylogenies, and community diversity metrics. The vAMPirus pipelines additionally include optional approaches that can increase the biological signal-to-noise ratio in results by leveraging tools not yet commonly applied to virus amplicon data analyses. In this paper, we validate the vAMPirus analytical framework and illustrate its implementation into the general virus amplicon sequencing workflow by recapitulating findings from two previously published double-stranded DNA virus datasets. As a case study, we also apply the program to explore the diversity and distribution of a coral reef-associated RNA virus. vAMPirus is incorporated with the Nextflow workflow manager, offering straightforward scalability, standardization, and communication of virus lineage-specific analyses. The vAMPirus framework itself is also designed to be adaptable; community-driven analytical standards will continue to be incorporated as the field advances. vAMPirus supports researchers in revealing patterns of virus diversity and population dynamics in nature, while promoting study reproducibility and comparability.

Running head: Viruses, Amplicon sequencing, Bioinformatics

vAMPirus: A versatile amplicon processing and analysis program for studying viruses

Alex J. Veglia[1], Ramón E. Rivera-Vicéns[2,3], Carsten G.B. Grupstra[1,4], Lauren I. Howe-Kerr[1], and Adrienne M.S. Correa[1]

[1]BioSciences Department, Rice University, 6100 Main St, Houston, TX 77005, USA

[2]Department of Earth and Environmental Sciences, Paleontology & Geobiology, Ludwig-Maximilians-Universität München, Richard-Wagner-Str. 10, 80333 München, Germany

[3]Department of Sciences and Technology, Interamerican University of Puerto Rico at Barranquitas, PO Box 517, Barranquitas, PR 00794, USA

[4] Department of Biology, Boston University, Boston, Massachusetts, United States of America

KEYWORDS – Virus, bioinformatics, diversity, DNA, RNA, amplicon sequencing, amino acid

[1]Corresponding Author: [1] Alex J. Veglia *alex.veglia@gmail.com*

Abstract

Amplicon sequencing is an effective and increasingly applied method for studying viral communities in the environment. Here, we present vAMPirus, a user-friendly, comprehensive, and versatile DNA and RNA virus amplicon sequence analysis program, designed to support investigators in exploring virus amplicon sequencing data and running informed, reproducible analyses. vAMPirus intakes raw virus amplicon libraries and, by default, performs nucleotide- and protein-based analyses to produce results such as sequence abundance information, taxonomic classifications, phylogenies, and community diversity metrics. The vAMPirus pipelines additionally include optional approaches that can increase the biological signal-to-noise ratio in results by leveraging tools not yet commonly applied to virus amplicon data analyses. In this paper, we validate the vAMPirus analytical framework and illustrate its implementation into the general virus amplicon sequencing workflow by recapitulating findings from two previously published double-stranded DNA virus datasets. As a case study, we also apply the program to explore the diversity and distribution of a coral reef-associated RNA virus. vAMPirus is incorporated with the Nextflow workflow manager, offering straightforward scalability, standardization, and communication of virus lineage-specific analyses. The vAMPirus framework itself is also designed to be adaptable; community-driven analytical standards will continue to be incorporated as the field advances. vAMPirus supports researchers in revealing patterns of virus diversity and population dynamics in nature, while promoting study reproducibility and comparability.

Introduction

From the human gut to sediments in the deep ocean, viruses are abundant, diverse, and shape the systems they inhabit (Breitbart et al., 2018; Correa et al., 2021; Suttle, 2007). The advent of high-throughput sequencing (HTS) techniques like amplicon sequencing has transformed the field of virology, illuminating the currently unculturable virosphere (Labadie et al., 2020; Metcalf et al., 1995; Paez-Espino et al., 2017; Zayed et al., 2022) and helping identify the impacts of viruses on ecosystem and host function (Braga et al., 2020; Breitbart et al., 2018; Thurber et al., 2017; Uyaguari-Diaz et al., 2016). Amplicon sequencing is a targeted, polymerase chain reaction (PCR)-based HTS approach that allows deep characterization of genetic variants within virus populations (Short et al. 2010). The targeted nature of amplicon sequencing reduces the economic and computational investment required for spatiotemporal investigations of virus communities at ecologically relevant scales (see Finke & Suttle, 2019; Frantzen & Holo, 2019; Grupstra et al., 2022; Gustavsen & Suttle, 2021; Howe-Kerr et al., 2022; Montalvo-Proaño et al., 2017). The number of studies leveraging virus amplicon sequencing has increased rapidly over the last two decades (e.g., 16 peer-reviewed publications in 1998 compared to 127 in 2021 based on a Web of Science search of 'virus amplicon sequencing', November 2022).

The general virus amplicon sequencing workflow includes: 1. Extraction of virus nucleic acid (DNA or RNA), 2. PCR amplification of virus marker gene or transcript, 3. HTS of virus marker gene amplicons, and 4. Bioinformatic analysis of sequencing data (Short et al., 2010; Figure 1). The effective analysis and interpretation of amplicon sequencing data relies on biologically accurate binning of marker gene sequences into taxonomically or ecologically distinct units. Recognizing viral taxa or ecotypes, however, can be challenging. For example, non-model viruses have limited baseline information available to inform the selection of clustering thresholds. Other viruses, such as RNA viruses, have error-prone polymerases and produce quasispecies, a population structure consisting of large numbers of variant genomes (Domingo & Perales, 2019) that may not be easily resolved by the same clustering percentage. Amplicon sequence variants (ASVs) are a promising non-clustering-based approach for virus amplicon analyses that offers high precision and biological accuracy as error-derived sequence variants are removed during ASV generation (Callahan et al., 2017; Edgar, 2016b). In addition, since the identity of an ASV is not specific to a given dataset (as identity can be in clustering of marker gene sequences into *de novo* OTUs based on a percent identity value, Callahan et al., 2017), ASVs and their unique translations ('aminotypes', see Grupstra et al., 2022) can be compared directly among studies (Callahan et al., 2017).

To promote the standardization, reproducibility and cross-comparison of DNA and RNA virus amplicon sequence analyses, we developed the automated bioinformatics tool, vAMPirus

(github.com/Aveglia/vAMPirus). vAMPirus intakes raw (unprocessed) virus amplicon libraries, performs all read processing and diversity analysis steps, and produces reports detailing results (e.g., relative abundance plots, community diversity metrics) with interactive figures and tables. vAMPirus supports initial explorations of viral amplicon sequence datasets via a 'DataCheck' pipeline, which generates an HTML report with information on data quality and sequence diversity. Results from the exploratory DataCheck pipeline can then be used to optimize parameters in the read processing or ASV generation steps within the vAMPirus 'Analyze' pipeline; this can improve the signal-to-noise ratio in downstream analyses. vAMPirus is integrated with the Nextflow workflow manager, which uses a configuration file that can be shared among investigators, facilitating the standardization and dissemination of virus amplicon sequence analyses across projects and research groups. To that end, we also created the vAMPirus Analysis Repository (https://zenodo.org/communities/vampirusrepo/) to act as a central location for all published vAMPirus analyses. vAMPirus is intended to be accessible to researchers with a range of bioinformatics experience levels, and includes substantial help documentation with step-by-step instructions for running the tool (https://github.com/Aveglia/vAMPirus/blob/master/docs/). By facilitating the standardization of viral lineage-specific analyses and increasing the signal-to-noise ratio in community diversity analyses, vAMPirus will enhance the effectiveness of virus amplicon studies and lead to a more developed understanding the global virosphere.

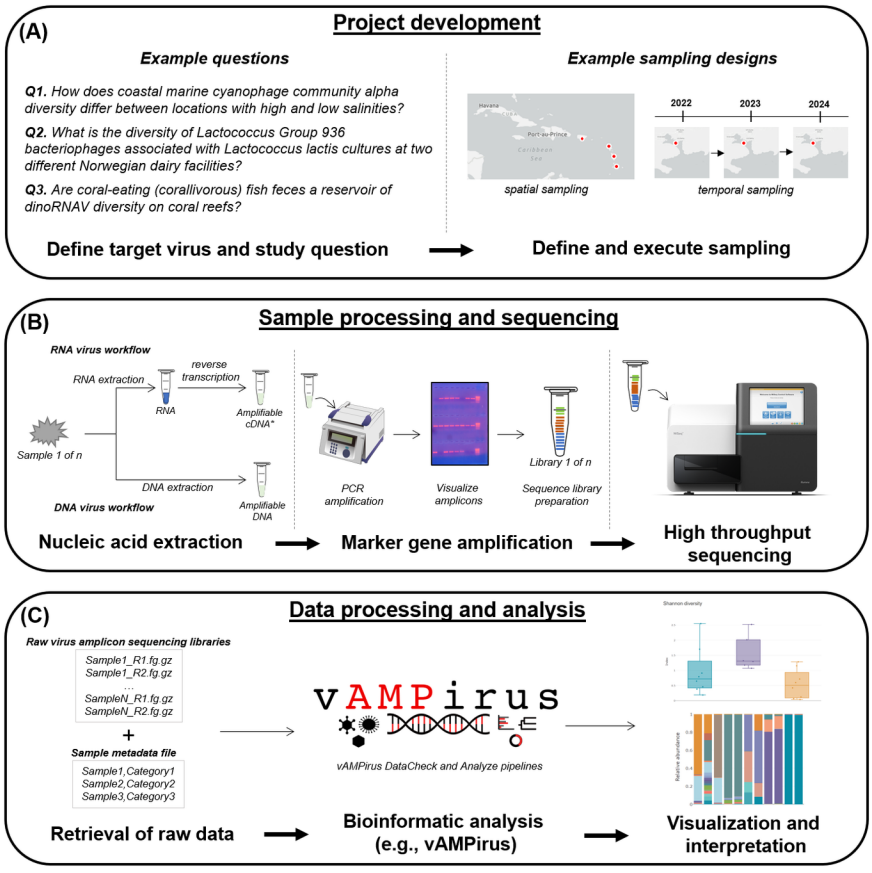

Figure 1. General workflow of virus amplicon sequencing projects (bolded text). *cDNA = complementary DNA.

Materials and methods

This section outlines the pipelines and processes therein that comprise vAMPirus v2.1.0 (Figure 2). For

a more detailed explanation of vAMPirus processes and output, see the manual stored in the vAMPirus GitHub repository (github.com/Aveglia/vAMPirus/).

2.1 vAMPirus implementation and configuration

2.1.1 Overview of vAMPirus execution

vAMPirus is composed of three main components that are recommended to be deployed sequentially: 1. A startup script to install dependencies and databases for taxonomy processes, 2. A 'DataCheck' pipeline that provides users with detailed information on data quality and diversity to inform subsequent analysis, and 3. An 'Analyze' pipeline that runs a comprehensive biology-focused analysis of the data using specified parameters and program options. vAMPirus is incorporated with Nextflow, a scientific workflow manager that allows easy configuration and deployment of the program using Conda, Docker, Singularity or cloud systems like Amazon Web Services (Di Tommaso et al., 2017). Nextflow natively communicates with scheduling managers like SLURM, PBS, or Torque, making it easy to run vAMPirus on high-performance computing clusters or on a local laptop or workstation. vAMPirus analyses can be configured using the Nextflow configuration file to promote efficient utilization of computing resources and reduce run times. Real-time monitoring and remote launching of vAMPirus analyses can be done using Nextflow Tower with no alterations to the vAMPirus script.

2.1.1 vAMPirus startup script

A startup script written in BASH is provided within the vAMPirus installation directory that will automatically install dependencies and prepare the vampirus.config file for use. Users can deploy this script to download the Nextflow workflow manager and Conda package management system if these programs are not already installed/accessible on the computer system. The script can also be directed to download one or more protein/taxonomy databases to be used in vAMPirus taxonomy processes. Available databases include: 1. The proteic version of the Reference Virus DataBase (RVDB, Bigot et al., 2020), 2. NCBI virus protein RefSeq database (Brister et al., 2015), and 3. Complete NCBI NR protein database (O'Leary et al., 2016). If directed to do so, the startup script will also download the NCBI Taxonomy Database (Schoch et al., 2020) and last common ancestor (LCA) information for sequences curated within the RVDB (Bigot et al., 2020). The script then edits the vAMPirus configuration file with the updated paths to any downloaded databases and to the vAMPirus installation directory. Lastly, text documents that include general next steps for the user and commands to test the installation are printed in the vAMPirus directory. If test analyses complete successfully, the user then updates the configuration file with project-specific parameters (e.g., project name, database for taxonomy inference, primer sequence information, number of allocated threads, working memory, scheduling manager) prior to running vAMPirus on a dataset.

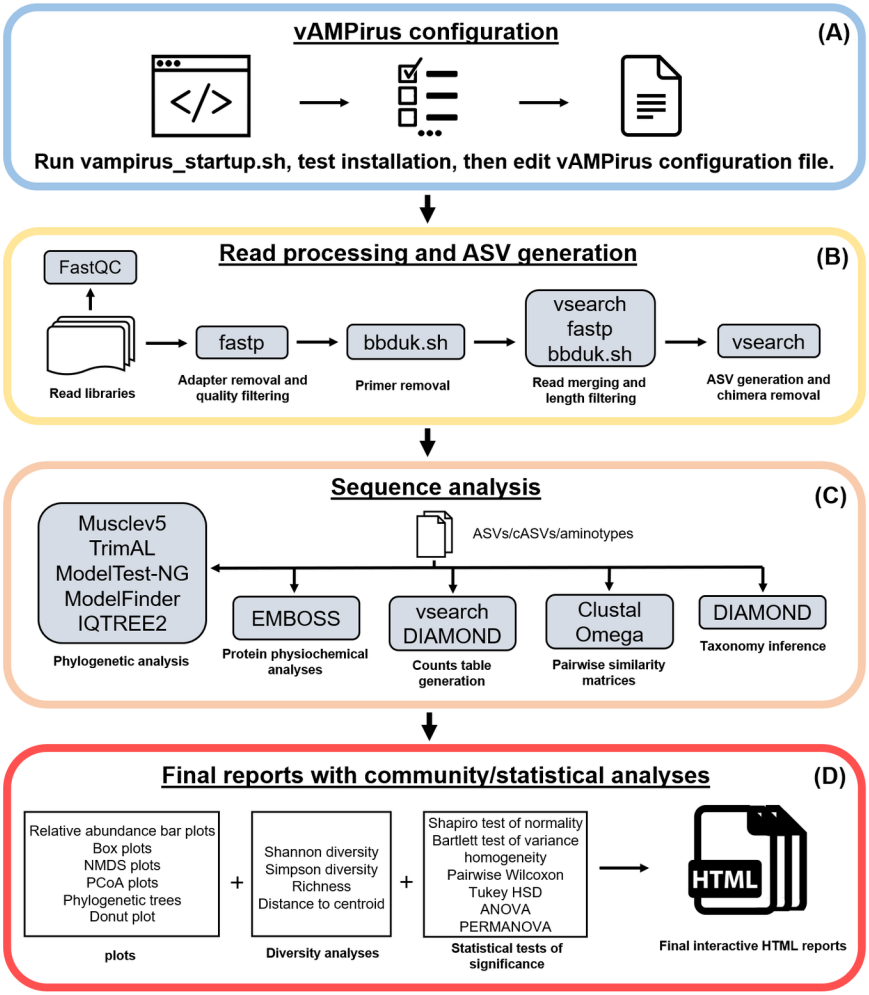

Figure 2. Generalized flowchart of vAMPirus v2.1.0, illustrating its configuration (box A), default analyses and programs used within the read processing (box B) and Analyze (boxes C and D) pipelines. For simplicity, only selected connections between processes are highlighted; processes generating the unique amino acid sequences ('aminotypes') and clustered ASVs ('cASVs'), as well as those specific to the DataCheck pipeline (see Supplemental Figure S2) are omitted. See Supplemental Figures S1, S2, and S3 for a more comprehensive illustration of vAMPirus pipelines and the processes therein.

2.2 Overview of the processes performed within vAMPirus

2.2.1 Read processing and generation of amplicon sequence variants

vAMPirus supports single- and paired-end raw Illumina read libraries as input. By default, read processing and ASV generation processes are performed prior to entering the DataCheck or Analyze pipelines (Figure 2, yellow box; Supplemental Figure S1). The read processing pipeline begins with a check of raw libraries using FastQC (v0.11.9, Andrews 2010), which creates and stores reports for review by the user. As FastQC is running, the program fastp (v0.20.1, Chen et al., 2018) automatically detects and removes adapter contamination, and performs quality/length filtering based on user-set parameters in the configuration file. fastp also performs over-representation analysis and (for paired-end input) base error correction during this step. Next, primers are removed from adapter-less reads using the bbduk.sh program within the BBTools software package (Bushnell 2014), and then another FastQC report is generated and stored. Cleaned reads

are then merged using the program VSEARCH (v2.21.1, Rognes et al., 2016) and merged reads (from all libraries) are then concatenated into a single fastq file. For accurate ASV generation, it is imperative that the merged reads be the same length (Edgar, 2016b). To ensure this, merged reads are globally trimmed to a user-specified maximum read length using fastp. Merged reads with the set length are then extracted from the total merged read file using the program bbduk.sh and dereplicated using the program VSEARCH (v2.21.1, Rognes et al., 2016), producing a unique read file containing read representation information. Amplicon sequence variants are then generated from this unique read file with VSEARCH and the UNOISE3 algorithm (Edgar, 2016b; Rognes et al., 2016). Chimeric ASVs are detected and removed using VSEARCH and the UCHIME3 algorithm (Edgar, 2016a). Prior to entering downstream pipelines, vAMPirus provides users the option to filter ASVs with DIAMOND blastx (v2.0.15, Buchfink et al., 2015) to remove non-target sequences or to focus their analyses on a subset of ASVs/aminotypes. These steps produce a final ASV fasta file that is then used as input for the DataCheck and Analyze pipelines.

2.2.2 Amplicon sequence variants and 'aminotypes'

By default, vAMPirus generates nucleotide-based (ASV) and protein-based (aminotype) results. ASVs support cross-study comparisons and offer a statistically supported view of virus sequence diversity, as biologically inaccurate sequences are removed during denoising (Callahan et al., 2017; Edgar, 2016b). However, ASV results for virus lineages with high mutation rates (e.g., RNA viruses with quasispecies heterogeneity) may still contain high levels of noise that mask biological patterns. It may be beneficial to group ASVs into distinct clusters based on genetic or ecological similarities in such use cases. In vAMPirus, 'aminotypes' (unique amino acid sequences, Grupstra et al. 2022) are generated by translating ASVs with VirtualRibosome (v2.0, Wernersson, 2006) and subsequently dereplicating these translations using the program CD-HIT (v4.8.1, Fu et al., 2012; Li & Godzik, 2006). As direct products of specific ASVs, aminotypes maintain sequence tractability, reproducibility, and comparability, and therefore differ from *de novo* OTUs or cASVs (see Section 2.2.3). The 'aminotyping' approach not only reduces noise; it also removes sequences with internal stop codons (a deleterious mutation) and reveals nonsynonymous mutations that may indicate differences in virus functionality (e.g., infection efficiency, host range; DeFilippis & Villarreal, 2000).

vAMPirus provides two additional (optional) ASV or aminotype "grouping" approaches that are alternatives to *de novo*clustering: Minimum Entropy Decomposition (MED) and phylogeny-based clustering or 'phylogrouping'. MED is a method of sequence clustering that utilizes Shannon entropy (Shannon, 1948) to partition marker gene datasets into 'MED nodes' (Eren et al., 2015). With this approach, users identify sequence positions in a set of ASVs or aminotypes that are information-rich (positions of high variability) or information-poor (positions of high conservation) and use these positions to assign ASVs/aminotypes to 'MED groups' (sequences with identical bases at specified positions) (Eren et al., 2015). Users can also specify and assign sequences to MED groups based on sequence positions of interest (e.g. positions of a protein sequence known to influence a viral characteristic such as host cell attachment; see Harvey et al., 2021). Phylogrouping is performed with the TreeCluster program (v1.0.3, Balaban et al., 2019). With this approach, ASV or aminotype sequences are assigned to "phylogroups" based on user specified TreeCluster parameters and the phylogenetic tree produced during analysis (see Figure 4-V, VI). All grouping methods can be applied at the same time; coupled with the use of the Nextflow '–resume' feature, adjusting specific parameters and generating new results to review and compare is straightforward and does not require re-running the entire DataCheck or Analyze pipelines.

2.2.3 Optional de novo *sequence clustering*

vAMPirus provides the option to perform *de novo* clustering of ASVs into 'clustered ASVs' or 'cASVs' based on pairwise nucleotide (ncASV) and/or protein (pcASV) sequence similarity using the programs VSEARCH (Rognes et al., 2016) and CD-HIT (Fu et al., 2012; Li & Godzik, 2006), respectively. cASVs differ from traditional *de novo* OTUs because for cASVs, denoising of sequences is done prior to clustering. The *de novo* clustering of ASVs is most useful for more developed virus systems where the degree of sequence divergence between taxonomically or ecologically distinct groups is known. Note that, from a methodological standpoint, representative sequences generated by a cASV approach exhibit the same issues as *de novo* OTUs

(e.g., dataset dependence; see Callahan et al., 2017).

2.2.4 vAMPirus DataCheck pipeline and report

The vAMPirus DataCheck pipeline can help investigators determine the optimal parameters for read processing, ASV generation, and other downstream analyses conducted in the Analyze pipeline. The DataCheck pipeline is particularly beneficial for investigators working on nascent virus systems because it facilitates the informed establishment of gene-, lineage- or system-specific analysis standards. The pipeline produces an HTML report that displays information such as sequencing success per sample, read characteristics (e.g., read length, GC content), and ASV/aminotype sequence properties. The DataCheck pipeline also provides insight into the ASV sequences by clustering them across a range of nucleotide and amino acid similarities and plotting the resultant number of cASVs per similarity value. Briefly, nucleotide-based *de novo* cASVs are produced by clustering ASV sequences using 24 different percent identity values (55%, 65%, 75%, 80-100%) with VSEARCH. To generate *de novo* pcASVs, ASVs are first translated using the program VirtualRibosome (v2.0, Wernersson, 2006), then clustered into *de novo* pcASVs using the same 24 percent identities with the program CD-HIT (v.4.8.1, Fu et al., 2012; Li & Godzik, 2006). For each percent identity value, the number of ncASVs and pcASVs is quantified and visualized as a scatter plot in the DataCheck report. This is a common approach used to determine the clustering percentage (e.g., Gustavsen and Suttle 2021): the percent similarity at which there is no longer a linear rise in the number of cASVs (the inflection point) is selected for sequence clustering. Optionally, users can also apply the program oligotyping (Eren et al., 2015) to calculate Shannon entropy values per sequence position for both ASV and aminotypes, which is then displayed in the report. An example vAMPirus DataCheck report is available at github.com/Aveglia/vAMPirusExamples.

2.2.5 vAMPirus Analyze pipeline and report

The Analyze pipeline includes multiple analyses (e.g., phylogenetics, taxonomy inference); results are summarized in a final HTML report with tables and interactive figures (Supplemental Figure S3). The pipeline also organizes and stores output (e.g., counts tables, similarity matrices, taxonomy files) from these analyses within a user-specified results directory. Integration into Nextflow allows different analytical approaches (e.g., ASV grouping approach) to be run in parallel. Primary processes and analyses are summarized below; additional processes, such as percent similarity matrix generation and protein physiochemical property analyses are reviewed in the supplemental materials (Section S1).

2.2.5.1 Counts table generation

Nucleotide- and amino acid-based counts tables are generated within the Analyze pipeline. Counts tables for ASVs and ncASVs are produced using the VSEARCH program and the USEARCH algorithm (Edgar, 2010). Optionally, users can replace the use of the USEARCH algorithm with the use of the "–search_exact" feature provided by VSEARCH for exact ASV counts tables (Rognes et al., 2016). The aminotype and pcASV counts tables are generated by aligning translated merged reads to reference amino acid sequences with DIAMOND blastx (Buchfink et al., 2015). Sequence abundance information (in a comma delimited counts table) is then generated with a custom BASH script, which quantifies the number of alignments to each reference aminotype or pcASV from the DIAMOND output file. Users have the option to edit and adjust option parameters for the VSEARCH and DIAMOND counts table generation processes within the configuration file. All count files are stored in the results directory and are processed and visualized within Analyze reports as relative abundance bar plots.

2.2.5.2 Taxonomy inference

Sequence taxonomy is inferred using DIAMOND blastx or blastp via the user-specified protein database and the option parameters specified in the vAMPirus configuration file. The taxonomy process produces several files that are stored within a DIAMOND specific directory: 1. Unmodified DIAMOND output file, 2. fasta file of taxonomy assignments within the sequence headers, and 3. Results summary files (phyloseq taxonomy file, tab-separated summary file, and summary table of the abundance of specific hits). Taxonomy inference results are visualized in the Analyze report as a donut plot.

2.2.5.3 Phylogenetic analysis

Phylogenetic analyses of ASVs, cASVs and aminotypes are conducted within the Analyze pipeline and all output files are stored in a dedicated directory within the results directory. First, sequences are aligned using the program muscle (v5.1; Edgar, 2021) and then trimmed automatically using the program trimAl (v1.4.1, Capella-Gutiérrez et al., 2009) using a heuristics-based approach. By default, substitution model testing is done with Modeltest-NG (v0.1.7, Darriba et al., 2020). The program IQTREE (v2.2.0.3; Minh et al., 2020) is used to generate a maximum-likelihood tree. The substitution model used to generate the tree can be set by the user, sourced from Modeltest-NG results, or automatically selected with ModelFinder (Kalyaanamoorthy et al., 2017). The tree produced is then used for phylogrouping with TreeCluster and is visualized in the Analyze report. Within the report, the user has the option to color code nodes based on sequence identity, taxonomy hit, MED group, or phylogroup assignment.

2.2.5.4 vAMPirus Analyze reports

The final process within the Analyze pipeline generates HTML reports using R Markdown (v2.3; (Xie et al., 2018). By default, an individual summary report containing community composition/diversity, taxonomy and phylogeny results is generated for ASVs, aminotypes, and cASVs. Users provide a metadata file that includes the sample name and category used to group the samples (i.e., sample treatment, location) for alpha and beta diversity analyses. vAMPirus summary reports are interactive and include: 1. Pre- and post-processing read statistics in tables and plots, 2. Rarefaction curves, 3. Shannon's diversity (H), Simpson's diversity (reported as 1-D), richness and distance to centroid box plots with statistical tests (Shapiro-Wilk normality test, Bartlett test of variance homogeneity, Kruskal-Wallis rank sum test, Wilcox test, ANOVA and Tukey HSD, as appropriate), 4. Two- and three-dimensional NMDS plots (if no convergence, then PCoA plots), 5. Relative abundance bar plots, and 6. Taxonomy and phylogenetic results. An example vAMPirus Analyze summary report can be downloaded and reviewed at github.com/Aveglia/vAMPirusExamples.

2.3 vAMPirus Analysis Repository

To encourage and simplify the dissemination of parameters and non-read files needed to reproduce vAMPirus analyses, we created the 'vAMPirus Analysis Repository' (zenodo.org/communities/vampirusrepo/). The vAMPirus Analysis Repository is a Zenodo Community intended as a central location where investigators can deposit vAMPirus configuration files, metadata files, databases used for taxonomy assignment or ASV filtering, and any other files required to reproduce an analysis. Instructions and recommendations for submission are available in the vAMPirus manual (shorturl.at/uCO28). Once uploaded, submissions to the vAMPirus Analysis Repository are given a DOI.

Validating the vAMPirus workflow with published double-stranded DNA (dsDNA) virus datasets

We assessed the functionality and performance of vAMPirus' analytical workflow using amplicon sequencing datasets from two previously published dsDNA virus studies (Table 1). Research questions associated with each study are used as examples in Figure 1A (Finke & Suttle 2019; Figure 1A, Q1; Frantzen & Holo 2019; Figure 1A, Q2). For each dataset, we ran a vAMPirus analysis that reproduced the analysis from the associated published paper as closely as possible. For example, if a study generated *de novo* OTUs based on 97% nucleotide identity, the vAMPirus equivalent was ncASVs generated at 97% nucleotide identity with similar data quality control constraints. We then compared the results of the vAMPirus-based analyses to the findings described in each source manuscript. In brief, vAMPirus identified the same biological patterns as those published by Finke & Suttle (2019, Figure 3) and Frantzen & Holo (2019, Figure 4) from their respective sequence datasets, and detected additional (previously unreported) virus diversity (Table 1). For example, Finke and Suttle (2019) reported increased cyanophage community alpha diversity in samples collected from sites with higher salinity (>27.5 practical salinity units, Figure 3-I, II); this pattern was present in the corresponding vAMPirus results (Figure 3-III, IV, V, VI), which included 86% more cyanophage pcASVs relative to the number of OTUs reported in Finke and Suttle (2019; Table 1). Similarly, the patterns of lactococcal phage OTU richness and relative abundances per sample reported by Franzten and Holo

(2019; Figure 4-I) were also present in the vAMPirus results (Table 2; Figure 4-II). vAMPirus reported 43% more lactococcal phage ncASVs, relative to the OTUs reported by Frantzen and Holo (2019; Table 1, Figure 4). In addition, vAMPirus ASV-level analysis (Figure 4-III) revealed high lactococcal phage nucleotide-level diversity (n=531), yet aminotyping results (Figure 4-IV) suggest that the mutations underlying this richness mostly result in synonymous mutations: ASV sequences translated to only 29 aminotypes. Aminotype phylogrouping (see Section 2.2.2) of these data with TreeCluster highlighted a previously hidden overlap of lactococcal phage diversity across samples and dairy plants (Figure 4-VI).

Some variation between results obtained from vAMPirus and previous publications was expected, as the pipelines used in these comparisons were not identical. The only striking difference between the original results (in Finke and Suttle 2019 and Frantzen and Holo 2019) and those produced by vAMPirus is the higher number of pcASVs and ncASVs (respectively) identified via the latter analytic pipeline. Taxonomy results generated with vAMPirus by DIAMOND blastx aligning sequences to the NCBI virus RefSeq database verified that the pcASVs and ncASVs are of cyanophage and lactococcal phage origin, respectively (Supplemental Figures S4 and S5). The higher diversity identified by vAMPirus may be attributable to differences in reference database used (boutique versus NCBI-curated), handling of singletons, and other factors.

Table 1. Breakdown of test datasets used during vAMPirus development, including the methods and results from the original (published) analysis, as well as results from vAMPirus analysis. vAMPirus results were generated using *de novo* clustering of ASVs into 'clustered ASVs' (cASVs) based on pairwise nucleotide (ncASV) and protein (pcASV) sequence similarity. dsDNA = double-stranded DNA.

Study	Target dsDNA virus group	Target gene	Original Methods
Finke & Suttle 2019	Myoviridae (T4-like cyanophage)	DNA polymerase	OTU clustering at 97% protein iden
Frantzen & Holo 2019	Siphoviridae (lactococcal bacteriophage)	portal protein	OTU clustering at 99.5% nucleotide

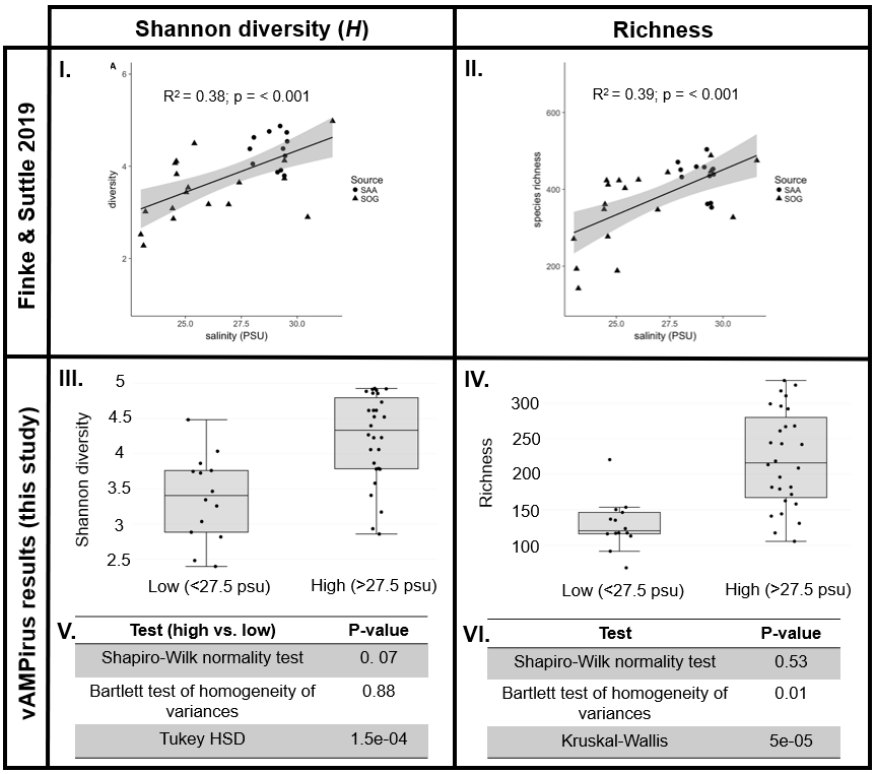

Figure 3. Comparison of results published by Finke & Suttle (2019) to those generated by the vAMPirus workflow for the same virus amplicon sequence dataset. The significant positive relationship between salinity and viral diversity (Shannon diversity (H) and richness) reported by Finke and Suttle (2019, panels I and II) was reproduced by vAMPirus (panels III-VI). Panels I and II display original results from Finke and Suttle (2019), based on 97% amino acid similarity OTUs. Panels III, IV, V, and VI include vAMPirus 97% pcASV results. Figures from Finke and Suttle (2019) and vAMPirus analyses were slightly modified for readability. For the vAMPirus analysis, which performs comparisons among categorical sample groups set by the user, all samples were assigned to either 'high' (>27.5) or 'low' (<27.5) salinity (practical salinity units; psu) groups. Figures from Finke and Suttle (2019) reprinted with permission from authors.

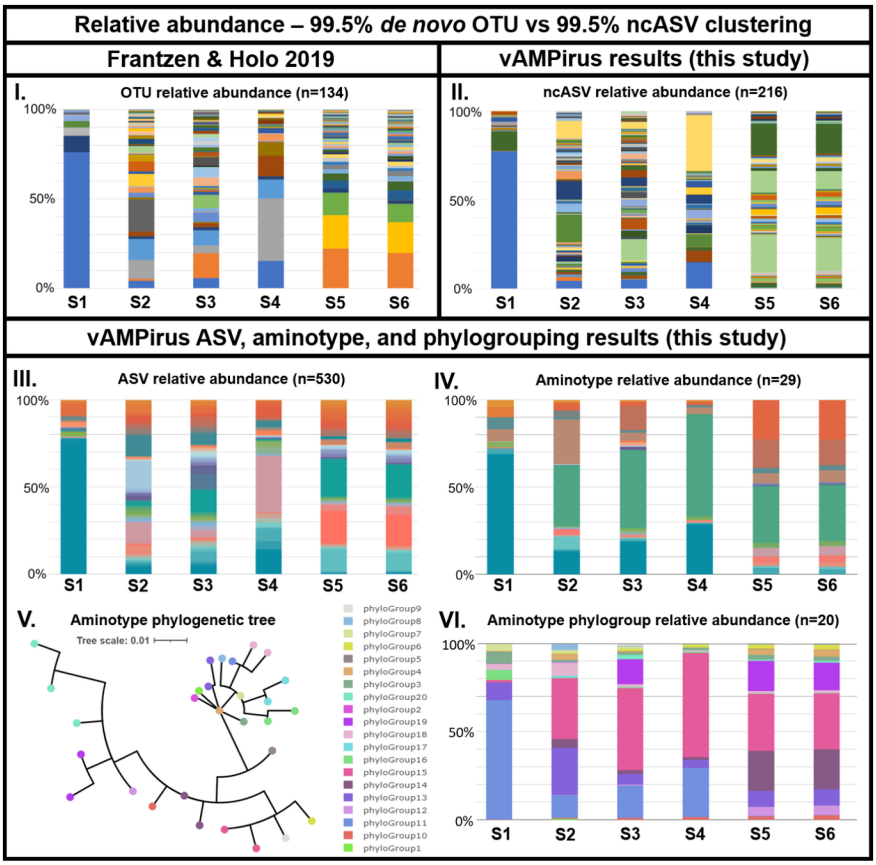

Figure 4. Comparison of results published by Frantzen and Holo (2019) to those generated by vAMPirus for the lactococcal phage amplicon sequence dataset. Patterns of lactococcal phage OTU relative abundance reported by Frantzen and Holo (2019; panel I) were recapitulated in vAMPirus ncASV results (panel II). Panel I displays original relative abundances of *de novo* OTUs based on 99.5% nucleotide similarity from Frantzen and Holo (2019). Panel II displays vAMPirus-generated 99.5% similarity ncASV relative abundance results. Panels III-VI demonstrate additional analyses output by vAMPirus, including ASV relative abundances, aminotype relative abundances, aminotype phylogeny with nodes colored by sequence phylogroup assignment, and aminotype phylogroup relative abundances. X-axis labels for all relative abundance graphs refer to dairy sample identity listed in Table 2. Colors in panels I-IV are plot-specific; similar colors across these panels do*not* indicate same sequences/clusters. However, colors assigned to phylogroups are consistent across panels V and VI. Figure from Frantzen and Holo (2019) reprinted with permission from authors.

Table 2. Comparison of Lactococcus phage sequence richness per sample between Frantzen and Holo (2019) results and vAMPirus (this study) results.

	Dairy Sample	S1	S2	S3	S4	S5	S6
	Dairy Plant	**1**	**1**	**1**	**2**	**2**	**2**
Frantzen and Holo 2019	# of OTUs with >100 reads	11	46	55	26	54	59
vAMPirus results	# of ncASVs with >100 reads	25	121	117	47	66	63

Applying vAMPirus to study a novel environmental RNA virus dataset

4.1 RNA virus study description

Dinoflagellate-infecting RNA viruses (dinoRNAVs) are positive-sense, single-stranded RNA viruses hypothesized to infect the dinoflagellate symbionts (Family Symbiodiniaceae) that live in the tissues of reef-building stony corals (Correa et al., 2013; Grupstra et al., 2022; Veglia et al., 2022). Although dinoRNAVs can be prevalent in coral colonies (Grupstra et al., 2022; Howe-Kerr et al., 2022; Montalvo-Proaño et al., 2017; Veglia et al., 2022), it is unclear how dinoRNAVs (or Symbiodiniaceae infected by dinoRNAVs) are transmitted among colonies. Recent work has shown that corallivorous (coral-eating) fishes disperse 100s of millions of live Symbiodiniaceae cells across reefscapes in their feces each day (Grupstra et al., 2021). To assess the extent to which corallivorous fish feces disperse dinoRNAVs in their feces (as in Grupstra et al., 2022), we characterized the presence and diversity of dinoRNAVs in various environmental reservoirs using amplicon sequencing of the dinoRNAV major capsid protein (*mcp*) gene. Given that dinoRNAV communities can vary across coral species and colonies (Grupstra et al., 2022; Howe-Kerr, 2022; Montalvo-Proaño et al., 2017) and that corallivorous fish actively 'sample' corals while feeding (Grupstra et al., 2021), we hypothesized that corallivorous fish feces are a reservoir of dinoRNAVs and that fish feces-associated dinoRNAV communities exhibit higher alpha diversity than coral colony-associated dinoRNAV communities. We generated 19 dinoRNAV *mcp* amplicon sequencing libraries from coral colony biopsies (*Acroporahyacinthus* , n=8; *Pocillopora* species complex, n=5), as well as the feces of corallivorous fishes (*Chaetodon reticulatus* , n=4; *Chaetodon ornatissimus,* n=2). All samples were collected from reefs off the north shore of Moorea, French Polynesia (South Pacific). Methods for sampling and sample processing to generate virus amplicon sequencing libraries are described in Grupstra et al. (2021,2022) and Howe-Kerr et al. (2022). DinoRNAV *mcp* amplicon libraries were processed and analyzed using vAMPirus (see doi.org/10.5281/zenodo.7574173).

4.2 RNA virus study results and discussion

Amplicon sequencing of the dinoRNAV *mcp* gene produced a total of 7.4 million raw reads across 19 samples representing three potential reservoirs of dinoRNAV diversity across the reef. The 7.4 million raw reads were processed and reduced to 2.8 million merged reads at the expected amplicon length of 420 bases. Merged *mcp* amplicons dereplicated into 1.1 million unique sequences from which 481 ASVs and 191 aminotypes were identified. The ASV-level results indicated a potential trend of higher dinoRNAV richness in corallivore feces relative to coral colonies (Kruskal-Wallis H test: p-value = 0.14, Figure 5-I). Aminotype results, however, revealed that dinoRNAV richness is significantly higher in corallivore feces, relative to*Pocillopora* coral colonies (Figure 5-II; Kruskal-Wallis H test: p-value = 0.005; Wilcoxon signed-rank test: *Pocillopora* vs. corallivore, p = 0.01, *Pocillopora* vs. *Acropora* , p = 0.04). We interpret that a biological difference in richness likely does exist between dinoRNAV communities in corallivore feces versus those in at least some species of coral holobionts, and this difference may be more readily detected with aminotype-based analyses (as ASV-based analyses may contain more "noise" due to errors arising during RNA virus replication). This use case illustrates the potential benefits of running nucleotide and protein-based amplicon analyses in tandem when testing hypotheses regarding virus community diversity and dynamics. Furthermore, both ASV and aminotypes differed significantly in composition according to dinoRNAV reservoir (anosim with Bray Curtis distances, R=0.99, p<0.01; Figure 5-III, IV), although some overlap (14%, 26 of 190 aminotypes) among dinoRNAV communities was observed (Supplemental Figure S6). Overall, this vAMPirus-based analysis of RNA virus amplicon sequencing data further corroborates that dinoRNAV communities differ across reef reservoirs (Grupstra et al. 2022, Montalvo-Proano et al. 2017, Howe-Kerr et al. 2022, Figure 5) and generates

a new hypothesis to be tested in future studies: corallivorous fishes are environmental hotspots of dinoRNAV diversity on reefscapes.

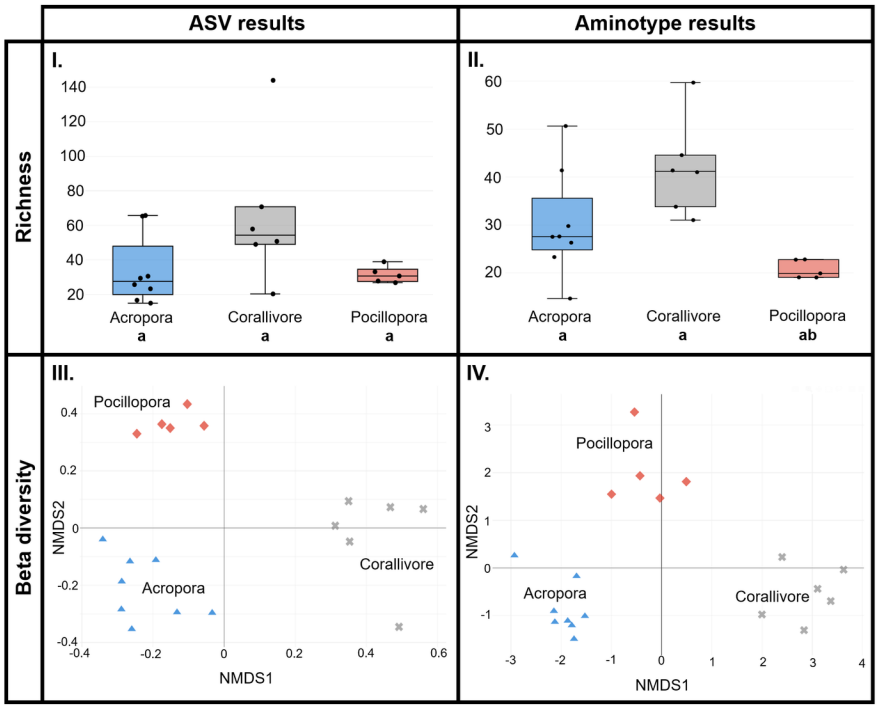

Figure 5. vAMPirus-generated dinoRNAV major capsid protein gene ASV and aminotype alpha (I, II) and beta (III, IV) diversity results from stony coral colonies *(Acropora* sp., *Pocillopora*species complex) and corallivorous (coral-eating) fish feces.Plots include three sample types: 1. *Acropora* biopsies (blue, triangle), 2. Corallivore feces (gray, x), and 3. *Pocillopora*biopsies (red, diamond). Letters beneath x axis labels on richness box plots (I, II) indicate statistically different groups. ASV and aminotype based NMDS plots (III, IV) were generated with Bray Curtis distances (stress values of 0.04 and 0.03, respectively).

Discussion

Targeted gene sequencing is increasingly being applied to explore spatiotemporal patterns of viral diversity (Adriaenssens & Cowan, 2014; Finke & Suttle, 2019; Frantzen & Holo, 2019; Grupstra et al., 2022; Gustavsen & Suttle, 2021; Howe-Kerr et al., 2022; Y. Li et al., 2018; Montalvo-Proaño et al., 2017; Prodinger et al., 2020; Short et al., 2010; Tong et al., 2016). The field of virology can now greatly benefit from the development of readily standardizable and reproducible pipelines for analyzing amplicon sequence datasets. Here, we present vAMPirus; a freely available, powerful, and flexible bioinformatics tool that streamlines the processing, analysis, and visualization of virus gene amplicon data. The availability of diverse bioinformatics approaches and tools within the vAMPirus program (*e.g.,* ASV calling, clustering, translation, phylogenetic clustering) empowers the user to adapt and set informed standards for their study system and easily share these standards with colleagues. With a user-friendly design and robust documentation, vAMPirus democratizes comprehensive virus amplicon sequencing analyses, making it a timely and valuable tool for virologists.

To inform virus amplicon data analyses, virologists have primarily relied on pipelines and tutorials geared towards bacterial or microeukaryote amplicon data (e.g., mothur (Schloss, 2020) and QIIME2 (Bolyen et al., 2019)). Although valuable insights have been made using these resources, an accessible virus-focused amplicon analysis pipeline will advance the field by offering via (1) automated pipelines that standardize

approaches for viral amplicon analyses (e.g., ASV and aminotype calling); (2) non-cluster-based alternatives to partitioning virus gene sequences (e.g., MED and phylogrouping); and (3) virus-focused taxonomy databases. Virus amplicon analyses have traditionally applied *de novo* clustering of marker gene sequences into *de novo* OTUs based on a percent identity value (i.e., 97% nucleotide identity, Callahan et al., 2017). However, clustering virus amplicons into biologically accurate *de novo* OTUs is challenging as the optimal clustering percentage is often unknown. vAMPirus provides users with the opportunity to transition from traditional *de novo*OTUs in virus amplicon sequencing analyses to using ASVs and aminotypes. We have illustrated here that ASV and aminotype-based analyses generally recapitulate findings generated via *de novo* OTU-based analyses (Figures 3, 4; Supplemental Table S1), while enabling reproducibility and cross-study comparisons (Callahan et al., 2017). Running analyses of amino acid and nucleotide sequence data in tandem, which is possible in vAMPirus, can aid in resolving virus phylogenies and reveal non-synonymous mutations that indicate virus protein property variability within a community (DeFilippis & Villarreal, 2000). This synergistic approach has been effective in developing dinoRNAVs and their dinoflagellate hosts (family Symbiodiniaceae, endosymbionts of stony corals) as a nascent study system. To characterize dinoRNAVs, studies have used the *mcp* gene, which has a high mutation rate and is hypothesized to be important in host cell attachment (Tomaru et al., 2004). vAMPirus aminotyping uncovered non-synonymous mutations in dinoRNAV *mcp* sequences, which may represent phenotypic differences that correlate with the distribution of host lineages across reefs (Grupstra et al., 2022; Howe-Kerr et al., 2022, this study). Aminotyping also effectively reduced noise from high mutation rates in ASV results, revealing temperature-driven increases in dinoRNAV infection productivity and community diversity across time and space (Grupstra et al. 2022: time only, Howe-Kerr et al., 2022). By making viral protein sequence analyses readily accessible in an amplicon sequence analysis workflow, vAMPirus helps reveal biological patterns in DNA and highly mutable RNA virus lineages by increasing signal-to-noise ratio in results (through collapse of synonymous nucleotide mutations, (Wernersson & Pedersen, 2003).

The increasing application of amplicon sequencing to the study of microbial diversity and dynamics has spurred efforts to improve the proficiency of tools that parse marker gene data. Such tools include the programs TreeCluster (Balaban et al., 2019) and oligotyping (Eren et al., 2015), which were developed as *de novo* clustering alternatives for partitioning genetic sequences into distinct units. In vAMPirus, these programs are utilized to assign ASVs and aminotypes to phylo- or MED groups based on user-set criteria (see Section 2.2.2). Assignment of ASV/aminotype sequences to groups rather than use of cluster representative sequences in analyses (such as, in the case of*de novo* OTUs and cASVs, Callahan et al., 2017) is done by vAMPirus to maintain reproducibility and comparability of results, while still permitting virus sequence classification into phylogenetically or ecologically distinct groups. These grouping approaches are instrumental for investigators because they can expose underlying patterns obscured by high sequence diversity (e.g., lactococcal phage phylogrouping results, Section 3, Figure 4). Phylogeny-based sequence clustering with TreeCluster has been applied to assess the diversity of microorganisms (and barley, Chen et al., 2022) and has been used to resolved virus transmission dynamics (HIV, Balaban et al., 2019; SARS-CoV-2, Plyusnin et al., 2022) and phylogenies (Ni et al., 2023). However, TreeCluster's potential utility for virus amplicon analyses is, for the most part, untapped. The inclusion of TreeCluster in the vAMPirus pipeline also opens the door to epidemiological insights, such as virus genetic linkage, transmission dynamics, and subpopulation mixing, from viral datasets (Balaban et al., 2019; Bezemer et al., 2015; Eshleman et al., 2011; Hué et al., 2014). Similarly, the program oligotyping developed by Eren et al. (2015) has been applied extensively to investigate microorganism diversity from marker gene data (cited 332 times as of February 2, 2023, Web of Science). However, only one published study has applied Minimum Entropy Decomposition sequence clustering with oligotyping to virus amplicon data (Needham et al., 2017). The MED grouping with oligotyping option provided by vAMPirus is a powerful approach for deciphering virus community diversity because it enables the grouping of sequences based on potential physiologically and/or ecologically relevant similarities. For example, users can identify gene sequence positions with non-synonymous mutations via aminotyping and then specify these positions in MED grouping to partition sequences into units of similar protein phenotypes (i.e., host cell attachment; see Harvey et al., 2021). The option to incorporate cutting-edge bioinformatic approaches, such as phylogrouping and MED grouping, into analyses of virus amplicon data makes vAMPirus a highly useful

"raw-reads-to-results" environmental virology workflow.

vAMPirus is an easy-to-use, open-source, and flexible tool that streamlines and simplifies the process of analyzing viral amplicon data. vAMPirus is designed to be community-driven; new features and programs (e.g., built-in lineage specific configuration files or databases, new bioinformatic tools) can easily be implemented at the request of investigators or when advances in best practices are made. vAMPirus advances studies of viral community diversity by facilitating informed analyses of amplicon sequence data with its DataCheck and Analyze pipelines in a standardized and reproducible manner.

Acknowledgments

The authors would like to thank Jan F. Finke, Curtis A. Suttle, Julia A. Gustavsen, Cyril Frantzen, Hedge Holo, Florian Prodinger and Hiroyuki Ogata (and their respective co-authors) for providing raw virus amplicon sequence data for vAMPirus testing and for permission to reprint original figures. We thank Samantha R. Coy for feedback and insight during early stages of vAMPirus development, and Nikolaos Schizas for access to computational resources during vAMPirus development. This work represents a contribution of the Moorea Coral Reef (MCR) LTER Site (NSF OCE 16–37396). This research was funded by a U.S. National Science Foundation Grant OCE 22-24354 (and earlier awards) to the Moorea Coral Reef LTER as well as a generous gift from the Gordon and Betty Moore Foundation. Research was completed under permits issued by the Territorial Government of French Polynesia (Delegation a la Recherche) and the Haut-Commissariat de la Republique en Polynesie Francaise (DTRT) (MCR LTER Protocole d'Accueil 2005–2022; Adrienne Correa Protocole d'Accueil 2013-2019), and we thank the Delegation a la Recherche and DTRT for their continued support. Start-up funds from Rice University, a NSF CAREER Award (OCE-2145472) and an Early-Career Research Fellowship (#2000009651) from the Gulf Research Program of the National Academies of Sciences to AMSC also contributed to this work. Additional funding was provided by Lewis and Clark Grants for Exploration to LHK and CGG, and Wagoner Foreign Study awards and to LHK, CGG, and AJV. The Kirk W. Dotson Endowed Graduate Fellowship in Ecology and Evolutionary Biology also helped support AJV as this work was conducted. RERV acknowledges funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 764840 (ITN IGNITE).

References

Adriaenssens, E. M., & Cowan, D. A. (2014). Using signature genes as tools to assess environmental viral ecology and diversity. *Applied and Environmental Microbiology* , *80* (15), 4470–4480. https://doi.org/10.1128/AEM.00878-14

Balaban, M., Moshiri, N., Mai, U., Jia, X., & Mirarab, S. (2019). TreeCluster: Clustering biological sequences using phylogenetic trees.*PLOS ONE* , *14* (8), e0221068. https://doi.org/10.1371/journal.pone.0221068

Bezemer, D., Cori, A., Ratmann, O., Sighem, A. van, Hermanides, H. S., Dutilh, B. E., Gras, L., Faria, N. R., Hengel, R. van den, Duits, A. J., Reiss, P., Wolf, F. de, Fraser, C., & Cohort, A. observational. (2015). Dispersion of the HIV-1 Epidemic in Men Who Have Sex with Men in the Netherlands: A Combined Mathematical Model and Phylogenetic Analysis.*PLOS Medicine* , *12* (11), e1001898. https://doi.org/10.1371/journal.pmed.1001898

Bigot, T., Temmam, S., Pérot, P., & Eloit, M. (2020). *RVDB-prot, a reference viral protein database and its HMM profiles* (8:530). F1000Research. https://doi.org/10.12688/f1000research.18776.2

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J. E., Bittinger, K., Brejnrod, A., Brislawn, C. J., Brown, C. T., Callahan, B. J., Caraballo-Rodríguez, A. M., Chase, J., … Caporaso, J. G. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology* , *37* (8), 852–857. https://doi.org/10.1038/s41587-019-0209-9

Braga, L. P. P., Spor, A., Kot, W., Breuil, M.-C., Hansen, L. H., Setubal, J. C., & Philippot, L. (2020).

Impact of phages on soil bacterial communities and nitrogen availability under different assembly scenarios. *Microbiome* , *8* (1), 52. https://doi.org/10.1186/s40168-020-00822-z

Breitbart, M., Bonnain, C., Malki, K., & Sawaya, N. A. (2018). Phage puppet masters of the marine microbial realm. *Nature Microbiology* , *3* (7), 754–766.

Brister, J. R., Ako-Adjei, D., Bao, Y., & Blinkova, O. (2015). NCBI viral genomes resource. *Nucleic Acids Research* ,*43* (Database issue), D571-577. https://doi.org/10.1093/nar/gku1207

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods* , *12* (1), Article 1. https://doi.org/10.1038/nmeth.3176

Callahan, B. J., McMurdie, P. J., & Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal* , *11* (12), Article 12. https://doi.org/10.1038/ismej.2017.119

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* , *25* (15), 1972–1973. https://doi.org/10.1093/bioinformatics/btp348

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* , *34* (17), i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Chen, Y.-Y., Schreiber, M., Bayer, M. M., Dawson, I. K., Hedley, P. E., Lei, L., Akhunova, A., Liu, C., Smith, K. P., Fay, J. C., Muehlbauer, G. J., Steffenson, B. J., Morrell, P. L., Waugh, R., & Russell, J. R. (2022). The evolutionary patterns of barley pericentromeric chromosome regions, as shaped by linkage disequilibrium and domestication.*The Plant Journal* , *111* (6), 1580–1594. https://doi.org/10.1111/tpj.15908

Correa, A. M. S., Howard-Varona, C., Coy, S. R., Buchan, A., Sullivan, M. B., & Weitz, J. S. (2021). Revisiting the rules of life for viruses of microorganisms. *Nature Reviews Microbiology* , *19* (8), 501–513. https://doi.org/10.1038/s41579-021-00530-x

Correa, A. M. S., Welsh, R. M., & Vega Thurber, R. L. (2013). Unique nucleocytoplasmic dsDNA and +ssRNA viruses are associated with the dinoflagellate endosymbionts of corals. *The ISME Journal* ,*7* (1), Article 1. https://doi.org/10.1038/ismej.2012.75

Darriba, D., Posada, D., Kozlov, A. M., Stamatakis, A., Morel, B., & Flouri, T. (2020). ModelTest-NG: A New and Scalable Tool for the Selection of DNA and Protein Evolutionary Models. *Molecular Biology and Evolution* , *37* (1), 291–294. https://doi.org/10.1093/molbev/msz189

DeFilippis, V. R., & VillarreaI, L. P. (2000). An Introduction to the Evolutionary Ecology of Viruses. *Viral Ecology* , 125–208. https://doi.org/10.1016/B978-012362675-2/50005-7

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology* , *35* (4), Article 4. https://doi.org/10.1038/nbt.3820

Domingo, E., & Perales, C. (2019). Viral quasispecies. *PLOS Genetics* , *15* (10), e1008271. https://doi.org/10.1371/journal.pgen.1008271

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* , *26* (19), 2460–2461. https://doi.org/10.1093/bioinformatics/btq461

Edgar, R. C. (2016a). *UCHIME2: Improved chimera prediction for amplicon sequencing* (p. 074252). bioRxiv. https://doi.org/10.1101/074252

Edgar, R. C. (2016b). *UNOISE2: Improved error-correction for Illumina 16S and ITS amplicon sequencing* (p. 081257). bioRxiv. https://doi.org/10.1101/081257

Edgar, R. C. (2021). *MUSCLE v5 enables improved estimates of phylogenetic tree confidence by ensemble bootstrapping* (p. 2021.06.20.449169). bioRxiv. https://doi.org/10.1101/2021.06.20.449169

Eren, A. M., Morrison, H. G., Lescault, P. J., Reveillaud, J., Vineis, J. H., & Sogin, M. L. (2015). Minimum entropy decomposition: Unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *The ISME Journal* , *9* (4), Article 4. https://doi.org/10.1038/ismej.2014.195

Eshleman, S. H., Hudelson, S. E., Redd, A. D., Wang, L., Debes, R., Chen, Y. Q., Martens, C. A., Ricklefs, S. M., Selig, E. J., Porcella, S. F., Munshaw, S., Ray, S. C., Piwowar-Manning, E., McCauley, M., Hosseinipour, M. C., Kumwenda, J., Hakim, J. G., Chariyalertsak, S., de Bruyn, G., … Hughes, J. P. (2011). Analysis of Genetic Linkage of HIV From Couples Enrolled in the HIV Prevention Trials Network 052 Trial. *The Journal of Infectious Diseases* , *204* (12), 1918–1926. https://doi.org/10.1093/infdis/jir651

Finke, J. F., & Suttle, C. A. (2019). The Environment and Cyanophage Diversity: Insights From Environmental Sequencing of DNA Polymerase.*Frontiers in Microbiology* , *10* . https://www.frontiersin.org/article/10.3389/fmicb.2019.00167

Frantzen, C. A., & Holo, H. (2019). Unprecedented Diversity of Lactococcal Group 936 Bacteriophages Revealed by Amplicon Sequencing of the Portal Protein Gene. *Viruses* , *11* (5), Article 5. https://doi.org/10.3390/v11050443

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data.*Bioinformatics* , *28* (23), 3150–3152. https://doi.org/10.1093/bioinformatics/bts565

Grupstra, C. G. B., Lemoine, N. P., Cook, C., & Correa, A. M. S. (2022). Thank you for biting: Dispersal of beneficial microbiota through "antagonistic" interactions. *Trends in Microbiology* ,*30* (10), 930–939. https://doi.org/10.1016/j.tim.2022.03.006

Grupstra, C. G. B., Rabbitt, K. M., Howe-Kerr, L. I., & Correa, A. M. S. (2021). Fish predation on corals promotes the dispersal of coral symbionts. *Animal Microbiome* , *3* (1), 25. https://doi.org/10.1186/s42523-021-00086-4

Grupstra, C. G., Howe-Kerr, L. I., Veglia, A. J., Bryant, R. L., Coy, S. R., Blackwelder, P. L., & Correa, A. (2022). Thermal stress triggers productive viral infection of a key coral reef symbiont. *The ISME Journal* , 1–12.

Gustavsen, J. A., & Suttle, C. A. (2021). Role of Phylogenetic Structure in the Dynamics of Coastal Viral Assemblages. *Applied and Environmental Microbiology* , *87* (11), e02704-20. https://doi.org/10.1128/AEM.02704-20

Harvey, W. T., Carabelli, A. M., Jackson, B., Gupta, R. K., Thomson, E. C., Harrison, E. M., Ludden, C., Reeve, R., Rambaut, A., Peacock, S. J., & Robertson, D. L. (2021). SARS-CoV-2 variants, spike mutations and immune escape. *Nature Reviews Microbiology* , *19* (7), Article 7. https://doi.org/10.1038/s41579-021-00573-0

Howe-Kerr, L. I. (2022). *Viruses of a key coral symbiont exhibit temperature-driven productivity across a reefscape* . https://doi.org/10.21203/rs.3.rs-1899377/v1

Hué, S., Brown, A. E., Ragonnet-Cronin, M., Lycett, S. J., Dunn, D. T., Fearnhill, E., Dolling, D. I., Pozniak, A., Pillay, D., Delpech, V. C., & Leigh Brown, A. J. (2014). Phylogenetic analyses reveal HIV-1 infections between men misclassified as heterosexual transmissions.*AIDS* , *28* (13), 1967. https://doi.org/10.1097/QAD.0000000000000383

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nature Methods* , *14* (6), Article 6. https://doi.org/10.1038/nmeth.4285

Labadie, T., Batéjat, C., Leclercq, I., & Manuguerra, J.-C. (2020). Historical Discoveries on Viruses in the Environment and Their Impact on Public Health. *Intervirology* , *63* (1–6), 17–32. https://doi.org/10.1159/000511575

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. https://doi.org/10.1093/bioinformatics/btl158

Li, Y., Hingamp, P., Watai, H., Endo, H., Yoshida, T., & Ogata, H. (2018). Degenerate PCR Primers to Reveal the Diversity of Giant Viruses in Coastal Waters. *Viruses* , *10* (9), 496. https://doi.org/10.3390/v10090496

Metcalf, T. G., Melnick, J. L., & Estes, M. K. (1995). ENVIRONMENTAL VIROLOGY: From Detection of Virus in Sewage and Water by Isolation to Identification by Molecular Biology—A Trip of Over 50 Years.*Annual Review of Microbiology* , *49* (1), 461–487. https://doi.org/10.1146/annurev.mi.49.100195.002333

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era.*Molecular Biology and Evolution* , *37* (5), 1530–1534. https://doi.org/10.1093/molbev/msaa015

Montalvo-Proaño, J., Buerger, P., Weynberg, K. D., & van Oppen, M. J. H. (2017). A PCR-Based Assay Targeting the Major Capsid Protein Gene of a Dinorna-Like ssRNA Virus That Infects Coral Photosymbionts.*Frontiers in Microbiology* , *8* . https://www.frontiersin.org/articles/10.3389/fmicb.2017.01665

Needham, D. M., Sachdeva, R., & Fuhrman, J. A. (2017). Ecological dynamics and co-occurrence among marine phytoplankton, bacteria and myoviruses shows microdiversity matters. *The ISME Journal* ,*11* (7), Article 7. https://doi.org/10.1038/ismej.2017.29

Ni, X.-B., Cui, X.-M., Liu, J.-Y., Ye, R.-Z., Wu, Y.-Q., Jiang, J.-F., Sun, Y., Wang, Q., Shum, M. H.-H., Chang, Q.-C., Zhao, L., Han, X.-H., Ma, K., Shen, S.-J., Zhang, M.-Z., Guo, W.-B., Zhu, J.-G., Zhan, L., Li, L.-J., … Cao, W.-C. (2023). Metavirome of 31 tick species provides a compendium of 1,801 RNA virus genomes. *Nature Microbiology* , *8* (1), Article 1. https://doi.org/10.1038/s41564-022-01275-w

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., … Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research* , *44* (D1), D733-745. https://doi.org/10.1093/nar/gkv1189

Paez-Espino, D., Chen, I.-M. A., Palaniappan, K., Ratner, A., Chu, K., Szeto, E., Pillay, M., Huang, J., Markowitz, V. M., Nielsen, T., Huntemann, M., K. Reddy, T. B., Pavlopoulos, G. A., Sullivan, M. B., Campbell, B. J., Chen, F., McMahon, K., Hallam, S. J., Denef, V., … Kyrpides, N. C. (2017). IMG/VR: A database of cultured and uncultured DNA Viruses and retroviruses. *Nucleic Acids Research* ,*45* (D1), gkw1030. https://doi.org/10.1093/nar/gkw1030

Plyusnin, I., Truong Nguyen, P. T., Sironen, T., Vapalahti, O., Smura, T., & Kant, R. (2022). ClusTRace, a bioinformatic pipeline for analyzing clusters in virus phylogenies. *BMC Bioinformatics* ,*23* (1), 196. https://doi.org/10.1186/s12859-022-04709-8

Prodinger, F., Endo, H., Gotoh, Y., Li, Y., Morimoto, D., Omae, K., Tominaga, K., Blanc-Mathieu, R., Takano, Y., Hayashi, T., Nagasaki, K., Yoshida, T., & Ogata, H. (2020). An Optimized Metabarcoding Method for Mimiviridae. *Microorganisms* , *8* (4), Article 4. https://doi.org/10.3390/microorganisms8040506

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: A versatile open source tool for metagenomics. *PeerJ* ,*4* , e2584. https://doi.org/10.7717/peerj.2584

Schloss, P. D. (2020). Reintroducing mothur: 10 Years Later.*Applied and Environmental Microbiology* , *86* (2). https://doi.org/10.1128/AEM.02343-19

Schoch, C. L., Ciufo, S., Domrachev, M., Hotton, C. L., Kannan, S., Khovanskaya, R., Leipe, D., Mcveigh, R., O'Neill, K., Robbertse, B., Sharma, S., Soussov, V., Sullivan, J. P., Sun, L., Turner, S., & Karsch-Mizrachi, I. (2020). NCBI Taxonomy: A comprehensive update on curation, resources and tools. *Database: The Journal of Biological Databases and Curation* , *2020* , baaa062. https://doi.org/10.1093/database/baaa062

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal* , *27* (3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Short, S. M., Chen, F., & Wilhelm, S. (2010). The construction and analysis of marker gene libraries. *Manual of Aquatic Viral Ecology* , 82–91.

Suttle, C. A. (2007). Marine viruses—Major players in the global ecosystem. *Nature Reviews Microbiology* , *5* (10), 801–812. https://doi.org/10.1038/nrmicro1750

Thurber, R. V., Payet, J. P., Thurber, A. R., & Correa, A. M. S. (2017). Virus–host interactions and their roles in coral reef health and disease. *Nature Reviews Microbiology* , *15* (4), 205–216. https://doi.org/10.1038/nrmicro.2016.176

Tomaru, Y., Katanozaka, N., Nishida, K., Shirai, Y., Tarutani, K., Yamaguchi, M., & Nagasaki, K. (2004). Isolation and characterization of two distinct types of HcRNAV, a single-stranded RNA virus infecting the bivalve-killing microalga Heterocapsa circularisquama. *Aquatic Microbial Ecology* , *34* (3), 207–218. https://doi.org/10.3354/ame034207

Tong, Y., Liu, B., Liu, H., Zheng, H., Gu, J., Liu, H., Lin, M., Ding, Y., Song, C., & Li, Y. (2016). New universal primers for genotyping and resistance detection of low HBV DNA levels. *Medicine* ,*95* (33), e4618. https://doi.org/10.1097/MD.0000000000004618

Uyaguari-Diaz, M. I., Chan, M., Chaban, B. L., Croxen, M. A., Finke, J. F., Hill, J. E., Peabody, M. A., Van Rossum, T., Suttle, C. A., Brinkman, F. S. L., Isaac-Renton, J., Prystajecky, N. A., & Tang, P. (2016). A comprehensive method for amplicon-based and metagenomic characterization of viruses, bacteria, and eukaryotes in freshwater samples. *Microbiome* , *4* (1), 20. https://doi.org/10.1186/s40168-016-0166-1

Veglia, A. J., Bistolas, K. S. I., Voolstra, C. R., Hume, B. C. C., Planes, S., Allemand, D., Boissin, E., Wincker, P., Poulain, J., Moulin, C., Bourdin, G., Iwankow, G., Romac, S., Agostini, S., Banaigs, B., Boss, E., Bowler, C., Vargas, C. de, Douville, E., … Thurber, R. L. V. (2022). *Endogenous viral elements reveal associations between a non-retroviral RNA virus and symbiotic dinoflagellate genomes*(p. 2022.04.11.487905). bioRxiv. https://doi.org/10.1101/2022.04.11.487905

Wernersson, R. (2006). Virtual Ribosome—A comprehensive DNA translation tool with support for integration of sequence feature annotation. *Nucleic Acids Research* , *34* (suppl_2), W385–W388. https://doi.org/10.1093/nar/gkl252

Wernersson, R., & Pedersen, A. G. (2003). RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences. *Nucleic Acids Research* , *31* (13), 3537–3539.

Xie, Y., Allaire, J. J., & Grolemund, G. (2018). *R Markdown: The Definitive Guide* . Chapman and Hall/CRC, Boca Raton, Florida. https://bookdown.org/yihui/rmarkdown/

Zayed, A. A., Wainaina, J. M., Dominguez-Huerta, G., Pelletier, E., Guo, J., Mohssen, M., Tian, F., Pratama, A. A., Bolduc, B., Zablocki, O., Cronin, D., Solden, L., Delage, E., Alberti, A., Aury, J.-M., Carradec, Q., da Silva, C., Labadie, K., Poulain, J., … Sullivan, M. B. (2022). Cryptic and abundant marine viruses at the evolutionary origins of Earth's RNA virome. *Science* , *376* (6589), 156–162. https://doi.org/10.1126/science.abm5847

Conflict of Interest

The authors declare that they have no financial conflict of interest with the content of this article.

Authors' Contributions

AJV, CBG, LHK conceived of the program with support from AMSC; CBG and LHK contributed R code used in the vAMPirus reports; RERV contributed R code and helped execute vAMPirus incorporation into Nextflow; CBG, LHK and AMSC processed samples and generated the RNA virus dataset; AJV designed the pipelines with input from CBG and LHK; AJV wrote bash and R code used in the program, analyzed data, and wrote the initial draft of the manuscript, with contributions by all authors.

Data availability Source code, scripts, and help documentation are available online at github.com/aveglia/vAMPirus. RNA virus sequencing libraries are available on NCBI SRA associated with the BioProject PRJNA923642 as well as in the vAMPirus Analysis Repository (doi.org/10.5281/zenodo.7574173). All non-read files required to reproduce all analyses and results described in this manuscript can be found on the vAMPirus Analysis Repository (doi.org/10.5281/zenodo.7574173).

ORCIDAJV - 0000-0003-3118-5127 RERV - 0000-0002-6229-3537 CGG - 0000-0001-5083-4570 LHK - 0000-0002-8086-5869 AMSC - 0000-0003-0137-5042