NeoRdRp: A comprehensive dataset for identifying RNA-dependent RNA polymerase of various RNA viruses from metatranscriptomic data

Shoichi Sakaguchi¹, Syun-ichi Urayama², Yoshihiro Takaki³, Hong Wu¹, Youichi Suzuki¹, Takuro Nunoura⁴, Takashi Nakano^{1*}, So Nakagawa^{5*}

¹Faculty of Medicine, Department of Microbiology and Infection Control, Osaka Medical and Pharmaceutical University

²Laboratory of Fungal Interaction and Molecular Biology (donated by IFO), Department of Life and Environmental Sciences, University of Tsukuba

³Super-cuttingedge Grand and Advanced Research (SUGAR) Program, Japan Agency for Marine-Earth Science and Technology (JAMSTEC)

⁴Research Center for Bioscience and Nanoscience (CeBN), Japan Agency for Marine-Earth Science and Technology (JAMSTEC)

⁵Department of Molecular Life Science, Tokai University School of Medicine

*Address for correspondence:

tnakano@ompu.ac.jp (TNakano); so@tokai.ac.jp (SN)

Keywords

RNA virome, RNA-dependent RNA polymerase, hidden Markov model, metatranscriptome

Running headline: NeoRdRp for RdRp identification

Abstract

RNA viruses are distributed in various environments, and most RNA viruses have been recently identified by metatranscriptome sequencing. However, due to the high nucleotide diversity of RNA viruses, it is still challenging to identify novel RNA viruses from metatranscriptome data. Therefore, this study generated a dataset of RNA-dependent RNA polymerase (RdRp) domains essential for all RNA viruses belonging to *Orthornavirae*. Genes with RdRp domains from various RNA viruses were clustered by amino acid sequence similarity. For each cluster, a multiple sequence alignment was generated, and a hidden Markov model (HMM) profile was created if the number of sequences was greater than five. Using the 1,467 HMM profiles, we detected RdRp domains in the RefSeq RNA virus sequences, combined the hit sequences with the RdRp domains, and reconstructed the HMM profiles. As a result, 2,234 HMM profiles were generated from 12,316 RdRp domain sequences, and the dataset was named NeoRdRp. Almost all NeoRdRp HMM profiles were found to be able to detect specifically RdRps in the UniProt dataset. Furthermore, we compared the NeoRdRp dataset with two previously reported RNA virus detection methods to find RNA virus sequences from metatranscriptome sequencing data. Our methods can identify most of the RNA viruses in the datasets; however, some RNA viruses were not detected, similar to the other two methods. The NeoRdRp can be improved by repeatedly adding new RdRp sequences and can be expected to be widely applied as a system for detecting various RNA viruses from metatranscriptome data.

Introduction

Currently, many RNA viruses are identified by deep sequencing of RNA from various environmental samples (i.e., metatranscriptome) (Shi et al., 2018; Sakaguchi et al., 2020; Orba et al., 2021). Indeed, various metatranscriptomic analyses have shown that RNA viruses that were similar to human infectious viruses, such as coronaviruses, orthomyxoviruses, and filoviruses, were found in several non-mammalian/avian species, including fishes (Shi et al., 2018). However, identifying RNA virus-derived sequences from metatranscriptome data is challenging due to the low abundance of viral RNA in sequencing reads and high nucleotide diversity among RNA viruses (Cobbin et al., 2021). Thus, we have developed a method of double-strand RNA (dsRNA) sequencing named fragmented and primer ligated dsRNA sequencing (FLDS) (Urayama et al., 2016, 2018; Hirai et al., 2021). In eukaryotic cells, long dsRNAs are rare; therefore, they are mainly derived from the genomes of dsRNA viruses or replicative intermediates of single-strand RNA (ssRNA) viruses (Kumar and Carmichael, 1998). Additionally, FLDS enables us to obtain entire dsRNA sequences, including both termini, and reconstruct multi-segmented genomes of RNA viruses based on terminal sequence similarity (Urayama et al., 2016). Accordingly, FLDS can enrich potential RNA virus genomes even though their sequences are not similar to known RNA viruses. However, confirming RNA viral genomes based on the FLDS data remains a problem, mainly because of the difficulty in identifying RNA virus-specific genes.

Viruses that utilize RNA as a genetic material form a monophyletic group named *Riboviria* (ICTV) based on taxonomy. *Orthornavirae* is a major kingdom of *Riboviria* containing all the dsRNA and ssRNA viruses — but excluding retroviruses — that harbors RNA-dependent RNA polymerase (RdRp). Therefore, RdRp genes can be used as markers for most RNA viruses, whereas their nucleotide and amino acid sequences vary greatly, and some of them are identified as fusion genes with other protein motifs (Černý et al., 2014;

Wolf et al., 2018; Koonin et al., 2020, 2021). It is often challenging to identify RNA virus genes, including the RdRp gene, from the potential complete viral genome sequences obtained by FLDS based on similarity search using BLAST (Basic Local Alignment Search Tool) with major amino acid databases (e.g., nr database or RefSeq database) (Jia and Gong, 2019; Urayama et al., 2020).

In RdRp domain sequences, there are eight conserved amino acid motifs; notably, the sixth motif containing three amino acid residues, GDD, was widely conserved (Koonin, 1991). Based on these RdRp sequence motifs, Wolf and his colleagues inferred the phylogeny of 4,617 RNA viruses (Wolf et al., 2018). However, the accuracy of the multiple alignment of all RdRp amino acid sequences was disputed due to their diversity (Holmes and Duchêne, 2019; Wolf et al., 2019). Therefore, in order to obtain reliable multiple alignments of RdRp sequences, it is necessary to first classify them into groups with high similarity and then create multiple alignments for each group.

This study collected RdRp domain sequences from 23,410 RNA viruses and clustered them based on their similarity to build 2,234 hidden Markov model (HMM) profiles RdRp domains. Indeed, an HMM profile is based on a multiple sequence alignment; therefore, it can identify distantly related sequences more efficiently than a pairwise-based search (Mistry et al., 2021). In fact, several bioinformatics approaches using HMM have been devised to find more distantly related viruses, such as VirSorter2 (Guo et al., 2021) and RVDB-prot (Goodacre et al., 2018; Bigot et al., 2019). Thus, this study analyzed RdRp sequences of RNA viruses obtained by FLDS with marine samples and revealed that our method successfully detected RdRp domains efficiently rather than the two abovementioned approaches (i.e., VirSorter2 and RVDB-prot). The constructed RdRp dataset, including annotation, is publicly shared online (<https://github.com/shoichisakaguchi/NeoRdRp>).

95 Furthermore, the RdRp dataset developed in this study could identify possible RdRp genes of
96 various RNA viruses from metatranscriptomic sequencing data.

97

98

Materials and Methods

Data sets

A total of 4,620 amino acid sequences containing RdRp domains were downloaded from the NCBI (National Center for Biotechnology Information) GenBank database based on Wolf and his colleagues (Wolf et al., 2018). Additionally, we obtained 18,790 amino acid sequences of 4,239 RNA viruses from the NCBI Virus database (<https://www.ncbi.nlm.nih.gov/labs/virus/vssi/>) on June 9, 2021. These datasets will be referred to as Wolf-RdRps and NCBI-RNA-virus in this study, respectively. We also used the curated UniProt data, including 565,928 amino acid sequences with gene annotation [UniProtKB Reviewed (Swiss-Prot) 2021-05-11] (Boutet et al., 2016), as a dataset to annotate our HMM profiles.

Potential RNA virus genome sequences obtained from the marine metatranscriptomic analysis were used as a benchmark dataset (Urayama et al., 2018). This dataset contains 228 RdRp domains found in 1,143 potential complete RNA virus genomes/genome segments. These RNA virus genomes/genome segments were obtained by sequence read mapping in FLDS and grouping based on terminal sequence similarity. For each genome or genome segment, open reading frames (ORFs) were predicted using TransDecoder version 5.5.0 with default parameters (Haas et al., 2016), and 1,366 amino acid sequences were obtained. This amino acid dataset is referred to as FLDS-data in this study.

Clustering based on amino acid sequence similarity of RdRp

Amino acid sequences of RdRp genes were clustered using the CD-HIT program version 4.7 (Li and Godzik, 2006; Fu et al., 2012) with the following parameters: a similarity threshold of 40 % and word size of 2. For a cluster containing ≥ 5 genes, amino acid sequences were aligned using the L-INS-i program in MAFFT version 7.450 (Katoh and Standley, 2013). We

selected regions with solid statistical evidence of shared homology for each multiple sequence alignment and removed uncertainty and error regions using Divvier version 1.0 (Ali et al., 2019). Furthermore, we split a multiple sequence alignment if more than one-quarter of the sequences were gap at the position. Notably, a motif was defined by multiple sequence alignment consisting of nine amino acid lengths. Multiple alignments of amino acid sequences were visualized using WebLogo version 3.7.8 (Crooks et al., 2004). Height represents the probability at each site, and the width is proportional to the number of sequences.

Creation of Hidden-Markov-Model (HMM) profiles

The HMM profiles were built for each multiple-sequence-alignment using hmmbuild, part of the HMMER3 package version 3.1b2 (Eddy, 2011) with default parameters. Then, all HMM profiles were compressed into an HMMER3 searchable database using hmmcompress (from the HMMER3 package).

Other RNA virus detection software and databases

VirSorter2 (Guo et al., 2021) and hmmsearch with HMM profiles of Reference Viral protein DataBase (RVDB-prot) (Bigot et al., 2019) were used to identify the potential complete RNA virus genome/genome segment sequences in FLDS-data. Additionally, VirSorter2 was used with an option “the search group set to RNA viruses only,” and other parameters were set to default. Finally, RVDB-prot version 22.0 was used with hmmsearch with an E-value of 1E-10.

Results

Constructing RdRp HMM profiles

The schematic procedure for constructing RdRp HMM profile is illustrated in Fig. 1. First, we obtained 4,620 RdRp genes and genes with the RdRp domain selected by Wolf and colleagues from the GenBank database (named Wolf-RdRps) (Wolf et al., 2018). Next, we clustered 4,620 amino acid sequences containing RdRp motifs using CD-HIT-2D. From the resultant 1,986 clusters, those more than five sequences (i.e., 174 clusters in total) were used for the HMM profile construction. For each cluster, amino acid sequences were aligned using MAFFT, and the multiple sequence alignment was split using Divvier and an in-house script to remove unreliable regions. Then, for each split multiple sequence alignment, an HMM profile was created using the hmmbuild program in the HMMER3 suite.

Furthermore, we conducted comprehensive similarity searches using the 1,467 HMM profiles with 18,790 amino acid sequences of 4,239 RefSeq RNA virus genomes downloaded from NCBI Virus (<https://www.ncbi.nlm.nih.gov/labs/virus/>) on June 9, 2021. A total of 7,696 viral sequences were hit through the hmmsearch. If multiple hits in short intervals on a single gene, a single domain may hit multiple regions. Therefore, if the detected sequences' distance was less than 500 amino acid length, we regarded them as a single region and a domain using BEDTools (Fig. S1) (Quinlan and Hall, 2010). The obtained RdRp motif sequences were combined with Wolf-RdRp, resulting in 12,316 sequences named "NeoRdRp-seq." The NeoRdRp-seq dataset was subjected to incremental clustering by CD-HIT-2D, and as a result, 448 clusters consisting of more than five sequences (including 8,669 sequences) were identified. Notably, 1,476 clusters (including 3,123 sequences) containing less than five sequences for each were excluded from the HMM profile construction. Therefore, we performed multiple sequence alignment for 448 clusters using reliable regions, and 2,234 domains were extracted. Fig. 2 shows an example of the alignment clearly

indicating the well-conserved motifs, DxxxxxD, GxxxTxxxN, and GDD (Koonin, 1991). For each motif, an HMM profile was created by hmmbuild in HMMER3. We named the 2,234 HMM profile set obtained from RdRp sequences as NeoRdRp-HMM version 1.0.

Evaluating RdRp HMM profiles

To annotate each NeoRdRp-HMM, we conducted a comprehensive HMM search using UniProt amino acid sequences [UniProtKB Reviewed (Swiss-Prot) 2021-05-11]. As a result, 1,361/2,234 HMM profiles detected at least one amino acid sequence (including cases where multiple HMM profiles detected the same amino acid sequence). Further, 963/995 database hits (96.8 %) were derived from RNA viruses (i.e., annotated as “*Riboviria*”), indicating that the RdRp HMM profiles can identify RNA viruses based on RdRp motifs. Furthermore, it has been reported that, although retroviruses do not contain the RdRp, they have reverse transcriptase partially similar to the RdRp (Wolf et al., 2018). However, in our UniProt search, no retrovirus was detected.

We evaluated NeoRdRp-HMM profiles based on the UniProtKB annotation of Gene Ontology (GO; molecular function). First, for each hit was scored it 1 if “RNA-directed 5’-3’ RNA polymerase activity” (GO:0003968) was included and 0 if excluded. Then, we calculated a mean value of the scores for each NeoRdRp-HMM profile, summarized in Supplementary Table 1. The average score of the profiles is 0.94, suggesting that most NeoRdRp-HMM profiles annotated by the UniProtKB search were derived from the RdRp of RNA viruses (1,145 HMMs, 84.1%). However, we also found that a few (25 HMMs, 1.8%) had a score of 0; RNA binding [GO:0003723] was the most common with 11 HMM profiles, followed by ATP binding [GO:0005524] with 10 cases, helicase activity [GO:0004386], and hydrolase activity [GO:0016787] with 9, mRNA methyltransferase activity [GO:0008174]

and structural molecule activity [GO:0005198] with 8. Thus, these HMM profiles with a score of 0 detect proteins that interact with RNA but do not contain RdRp domains.

Comparing RdRp identification

We benchmarked the ability of NeoRdRp-HMM and NeoRdRp-seq to identify RNA viruses using the potential complete RNA virus genome/genome segment sequences obtained by FLDS associated with marine RNA viromes (Urayama et al., 2018). As a result, 1,143 amino acid sequences were detected in FLDS-data. In addition, Urayama and his colleagues conducted a BLAST search with the nr database and manually annotated them based on the known conserved amino acid motifs (Urayama et al., 2018). As a result, 228 amino acid sequences were annotated as RdRp in the FLDS-data.

Using the 1,143 amino acid sequences including 228 RdRps in the FLDS-data, we evaluated NeoRdRp-HMM and NeoRdRp-seq using hmmsearch and tBLASTn, respectively. We merged the results of NeoRdRp-HMM and NeoRdRp-seq searches as NeoRdRp. We also applied VirSorter2 (Guo et al., 2021) and hmmsearch with RVDB-prot (Goodacre et al., 2018) for comparison. First, we analyzed the 228 RdRp genes using the three methods and detected as follows: 70 using all three methods, 112 by NeoRdRp and RVDB-prot, two by NeoRdRp and VirSorter2, 13 only by NeoRdRp, 11 only by RVDB-prot, and 20 were not detected by all methods (Fig. 3). Viral genomes detected only by NeoRdRp were classified into family *Partitiviridae*, *Reoviridae*, *Narnaviridae*, *Flaviviridae*, and *Endornaviridae* based on HMM pattern and genomic structure. However, viral genomes that the three methods could not detect were recognized as members of *Reoviridae* and *Megabirnaviridae* in Urayama's report (Urayama et al., 2018). Then, we further compared NeoRdRp-HMM and NeoRdRp-seq (Supplementary Fig. 1). As a result, 161 RdRp genes were found by both methods, 14 only by NeoRdRp-HMM and 22 only by NeoRdRp-seq. The fact that there were

RdRp genes that NeoRdRp-HMM could only detect means that the HMM search can detect diversified RdRp genes in various RNA viruses. On the other hand, there were RdRp sequences that NeoRdRp-seq could only detect because the sequences that were less than 5 sequences as a clustering result were not in the HMM profiles. Also, we examined amino acid sequences of FLDS-data that were not annotated as RdRp. As a result, NeoRdRp-HMM detected 15 sequences as RdRps, indicating that our methods can identify unannotated RdRps in metatranscriptome data (Supplementary Table 2).

We also compared the computational time required to search for RNA viruses in FLDS-data using the three methods (Table 1): hmmsearch using NeoRdRp-HMM and tBLASTn using NeoRdRp-seq took 25 sec and 47 sec, respectively (i.e. 1 min and 12 sec in total for the NeoRdRp search); hmmsearch using RVDB-prot took 2 min 52 sec; and search using VirSorter took 11 min 46 sec. Furthermore, the database size required for each analysis was 0.15 GB for NeoRdRp, 1.61 GB for RVDB-prot, and 11.3 GB for VirSorter2 (Table 1). Therefore, the NeoRdRp dataset is small and easy to use for RNA virus detection.

Discussion

Since amino acid sequences of RdRp varied greatly among all RNA viruses, various bioinformatics attempts have been made to find RdRp genes from metatranscriptome data (Wolf et al., 2018; Bigot et al., 2019; Babaian and Edgar, 2021; Zayed et al., 2021). In this study, RdRp genes were comprehensively obtained (named NeoRdRp-seq), clustered based on amino acid similarity, and HMM profiles were constructed (named NeoRdRp-HMM) (Fig. 1). Using the dataset, we have obtained various RdRp genes from FLDS-data (Fig. 3). However, our dataset cannot identify some RNA viruses belonging to *Reoviridae* and *Megabirnaviridae*. Furthermore, 11/228 RdRp genes were found only using the RVDB-prot dataset, corresponding to RNA viruses belonging to *Narnaviridae*, *Picobirnaviridae*,

Partitiviridae, and unclassified dsRNA viruses. Therefore, the current NeoRdRp still has room for improvement. Indeed, our pipeline used for the NeoRdRp-HMM can create enhanced HMM profiles by incorporating newly detected RdRp sequences. Thus, in principle, it is possible to find more RdRp sequences by repeating updates. Notably, the clusters that comprised less than five sequences were excluded from HMM profile construction to maintain the accuracy of RdRp domain identification. Additionally, as the number of RdRp sequences increases, the number of clusters used for HMM construction increases. As a result, the detection power must increase.

Nucleic acid contamination from DNA viruses, hosts, and environments is a problem when analyzing RNA viromes. Therefore, techniques to selectively extract RNA virus-derived nucleic acids, such as the FLDS method (Urayama et al., 2016; Hirai et al., 2021), are used to solve this problem partially. However, it is still challenging to eliminate all contamination from the sequencing data. Also, since our data are based on the RdRp domain uniquely found in RNA viruses, the sequences identified in our methods strongly show that those are derived from RNA viruses. Indeed, other RNA virus identification bioinformatics tools are based not only on RdRp domains; therefore, they could show similarity to the sequences derived from those other than RNA viruses. Thus, our approach has advantages in the accuracy of RNA virus identification. However, our approach cannot identify RNA virus genes, not RdRp genes. Therefore, other methods should be applied to annotate non-RdRp genes of RNA viruses.

Therefore, the NeoRdRp dataset can be applied to the following research designs. First, the metatranscriptomic data obtained by FLDS were assembled. Then, ORFs were searched in each contig, and RdRp domains were searched for each ORF using NeoRdRp-HMM. Undetected, remaining sequences were further searched with NeoRdRp-seq, a relatively computationally intensive process (Table 1). From the results of these analyses and

their scores, we can identify contigs containing the RdRp domain, suggesting that they are derived from RNA viruses. Furthermore, FLDS can identify multi-segmented genomes of RNA viruses based on terminal sequence similarity (Urayama et al., 2016). Therefore, if one contig is found to contain an RdRp domain, other contigs having the same sequence on their terminals are strongly suggested to be derived from RNA viruses. Also, in those contigs, other viral genes can be detected using other computational programs and datasets, including VirSorter2 and RVDB-prot.

Furthermore, our method can detect various RdRps with high accuracy by creating an RdRp dataset, clustering the RdRp based on similarity, and creating HMM profiles for each RdRp. All sequence data and annotation developed in this study are available on the website (<https://github.com/shoichisakaguchi/NeoRdRp>). Therefore, RNA virome searches will be more accurate with RNA virus-selective metatranscriptome sequencing data using the dataset. Additionally, the NeoRdRp dataset is very compact compared with other methods for RNA virome search (Table 1), and the results are clear-cut due to the targeted approach. Furthermore, by adding new RdRp sequences in the future, it will be possible to find RdRp sequences with high accuracy from metatranscriptome data, which should become an essential technique for discovering unknown viruses.

Acknowledgments

This work was supported by KAKENHI Grant-in-Aid for Scientific Research (C) 20K06775 (to S.S. and S.N.), Early-Career Scientists 20K15685 (to S.S.), Challenging Research (Pioneering) 18H05368 (to S.U., Y.T., and T.Nunoura), and Scientific Research on Innovative Areas 16H06429 (to S.U. and S.N.), 16K21723 (to S.U. and S.N.), 16H06437 (to S.U.), 19H04843 (to S.N.), and by AMED under Grant Numbers JP19fm0208009 (to S.U.), JP20wm0325004 (to S.U.), and JP21wm0325004 (to S.U. and S.N.). The computing

297 resource was partially supported by the NIG supercomputer at ROIS National Institute of
298 Genetics, Japan, and SHIROKANE at the Human Genome Center, the Institute of Medical
299 Science, the University of Tokyo, Japan.

300

References

- Ali, R. H., Bogusz, M., and Whelan, S. (2019). Identifying Clusters of High Confidence Homologies in Multiple Sequence Alignments. *Mol. Biol. Evol.* 36, 2340–2351.
- Babaian, A., and Edgar, R. C. (2021). Ribovirus classification by a polymerase barcode sequence. *bioRxiv*, 2021.03.02.433648. doi:10.1101/2021.03.02.433648.
- Bigot, T., Temmam, S., Pérot, P., and Eloit, M. (2019). RVDB-prot, a reference viral protein database and its HMM profiles. *F1000Res.* 8, 530.
- Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., Bansal, P., Bridge, A. J., et al. (2016). UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View. *Methods Mol. Biol.* 1374, 23–54.
- Černý, J., Černá Bolfíková, B., Valdés, J. J., Grubhoffer, L., and Růžek, D. (2014). Evolution of tertiary structure of viral RNA dependent polymerases. *PLoS One* 9, e96070.
- Cobbin, J. C., Charon, J., Harvey, E., Holmes, E. C., and Mahar, J. E. (2021). Current challenges to virus discovery by meta-transcriptomics. *Curr. Opin. Virol.* 51, 48–55.
- Crooks, G. E., Hon, G., Chandonia, J.-M., and Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome Res.* 14, 1188–1190.
- Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Comput. Biol.* 7, e1002195.
- Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28, 3150–3152.
- Goodacre, N., Aljanahi, A., Nandakumar, S., Mikailov, M., and Khan, A. S. (2018). A Reference Viral Database (RVDB) To Enhance Bioinformatics Analysis of High-Throughput Sequencing for Novel Virus Detection. *mSphere* 3. doi:10.1128/mSphereDirect.00069-18.

Guo, J., Bolduc, B., Zayed, A. A., Varsani, A., Dominguez-Huerta, G., Delmont, T. O., et al. (2021). VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* 9, 37.

Haas, B., Papanicolaou, A., and Others (2016). TransDecoder (find coding regions within transcripts). URL <http://transdecoder.github.io>.

Hirai, M., Takaki, Y., Kondo, F., Horie, M., Urayama, S.-I., and Nunoura, T. (2021). RNA Viral Metagenome Analysis of Subnanogram dsRNA Using Fragmented and Primer Ligated dsRNA Sequencing (FLDS). *Microbes Environ.* 36. doi:10.1264/jsme2.ME20152.

Holmes, E. C., and Duchêne, S. (2019). Can Sequence Phylogenies Safely Infer the Origin of the Global Virome? *MBio* 10. doi:10.1128/mBio.00289-19.

Jia, H., and Gong, P. (2019). A Structure-Function Diversity Survey of the RNA-Dependent RNA Polymerases From the Positive-Strand RNA Viruses. *Front. Microbiol.* 10, 1945.

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780.

Koonin, E. V. (1991). The phylogeny of RNA-dependent RNA polymerases of positive-strand RNA viruses. *J. Gen. Virol.* 72 (Pt 9), 2197–2206.

Koonin, E. V., Dolja, V. V., Krupovic, M., and Kuhn, J. H. (2021). Viruses defined by the position of the virosphere within the replicator space. *Microbiol. Mol. Biol. Rev.*, e0019320.

Koonin, E. V., Dolja, V. V., Krupovic, M., Varsani, A., Wolf, Y. I., Yutin, N., et al. (2020). Global Organization and Proposed Megataxonomy of the Virus World. *Microbiol. Mol. Biol. Rev.* 84. doi:10.1128/MMBR.00061-19.

Kumar, M., and Carmichael, G. G. (1998). Antisense RNA: function and fate of duplex RNA in cells of higher eukaryotes. *Microbiol. Mol. Biol. Rev.* 62, 1415–1434.

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

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., et al. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Res.* 49, D412–D419.

Orba, Y., Matsuno, K., Nakao, R., Kryukov, K., Saito, Y., Kawamori, F., et al. (2021). Diverse mosquito-specific flaviviruses in the Bolivian Amazon basin. *J. Gen. Virol.* 102. doi:10.1099/jgv.0.001518.

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842.

Sakaguchi, S., Nakagawa, S., Mitsuhashi, S., Ogawa, M., Sugiyama, K., Tamukai, K., et al. (2020). Molecular characterization of feline paramyxovirus in Japanese cat populations. *Arch. Virol.* 165, 413–418.

Shi, M., Zhang, Y.-Z., and Holmes, E. C. (2018). Meta-transcriptomics and the evolutionary biology of RNA viruses. *Virus Res.* 243, 83–90.

Urayama, S.-I., Takaki, Y., Hagiwara, D., and Nunoura, T. (2020). dsRNA-seq Reveals Novel RNA Virus and Virus-Like Putative Complete Genome Sequences from *Hymeniacidon* sp. Sponge. *Microbes Environ.* 35. doi:10.1264/jsme2.ME19132.

Urayama, S.-I., Takaki, Y., Nishi, S., Yoshida-Takashima, Y., Deguchi, S., Takai, K., et al. (2018). Unveiling the RNA virosphere associated with marine microorganisms. *Mol. Ecol. Resour.* 18, 1444–1455.

Urayama, S.-I., Takaki, Y., and Nunoura, T. (2016). FLDS: A Comprehensive dsRNA Sequencing Method for Intracellular RNA Virus Surveillance. *Microbes Environ.* 31, 33–40.

Wolf, Y. I., Kazlauskas, D., Iranzo, J., Lucía-Sanz, A., Kuhn, J. H., Krupovic, M., et al. (2018). Origins and Evolution of the Global RNA Virome. *MBio* 9. doi:10.1128/mBio.02329-18.

Wolf, Y. I., Kazlauskas, D., Iranzo, J., Lucía-Sanz, A., Kuhn, J. H., Krupovic, M., et al. (2019). Reply to Holmes and Duchêne, “Can Sequence Phylogenies Safely Infer the Origin of the Global Virome?”: Deep Phylogenetic Analysis of RNA Viruses Is Highly Challenging but Not Meaningless. *MBio* 10. doi:10.1128/mBio.00542-19.

Zayed, A. A., Lücking, D., Mohssen, M., Cronin, D., Bolduc, B., Gregory, A. C., et al. (2021). efam: an expanded, metaproteome-supported HMM profile database of viral protein families. *Bioinformatics*. doi:10.1093/bioinformatics/btab451.

Fig. 1. A schematic workflow of NeoRdRp-HMM construction

Amino acid sequences of RdRp were clustered based on similarity to create HMM profiles. Then, these HMM profiles were used to detect the RdRp domain of the RefSeq sequences of RNA viruses. Finally, these RdRp sequences were merged and clustered again to create HMM profiles for RdRp detection. Pink: amino acid sequences, orange: HMM profiles.

Fig. 2. An example of a multiple sequence alignment of amino acid sequences of the RdRp domain.

The multiple sequence alignment of the HMM profile of 24_RDRP_0.25_120-285 was visualized using Weblogo software. Underlines in red indicate well-conserved DxxxxxD, GxxxTxxxN, and GDD motifs in RdRp at positions 3–8, 65–71, and 95–97, respectively.

Fig. 3. Venn diagram of three different methods for identifying RNA viruses.

Annotated marine metatranscriptome data were searched using NeoRdRp (hmmsearch and tBLASTn), RVDB-prot (hmmsearch), and VirSorter2, respectively. Of all RdRp sequences detected, 13 were found only in NeoRdRp. Purple: NeoRdRp (HMM + tBLASTn), Green: RVDB-prot, Yellow: VirSorter2.

Table 1. Database size and processing time of three methods for searching for RNA viruses in FLDS data.

Data	Tool	Database type	Database Size (GB)	Speed
NeoRdRp-HMM/	hmmsearch/	HMM/	0.15/	25 sec/
NeoRdRp-Seq	tBLASTn	blast (FASTA)	0.02	47 sec
RVDB-prot	hmmsearch	HMM	1.61	2 min 52 sec
VirSorter2	VirSorter2	VS2	11.13	11 min 46 sec

414

415

416

417 **Supplementary Figure 1. Defining an RdRp motif.**

418 The detected sequences' distance was less than 500 amino acid length in the hmmsearch

419 result; we regarded them as a single region as a domain.

420

421

422

423

424

425

426 **Supplementary Figure 2. Comparison of NeoRdRp-HMM and NeoRdRp-seq.**

427 161 RdRp genes were found by both methods, 14 only by NeoRdRp-HMM, and 22 only by

428 NeoRdRp-seq. Purple: NeoRdRp-HMM, green: NeoRdRp-seq

429

Supplementary Table 1. Summary of the scores for each NeoRdRp-HMM profile.

(provided as an Excel File)

Supplementary Table 2. Unannotated RdRps in metatranscriptome data detected by

NeoRdRp.

acc	virus
BDQA01000402	unannotated
BDQA01000334	unannotated
BDQA01000404	unannotated
BDQA01000783	<i>Picobirnaviridae</i>
BDQA01000784	<i>Picobirnaviridae</i>
BDQA01000789	<i>Picobirnaviridae</i>
BDQA01000749	unannotated
BDQA01000762	<i>Picobirnaviridae</i>
BDQA01000786	<i>Picobirnaviridae</i>
BDQA01000281	unclassified dsRNA viruses
BDQB01000243	unclassified viruses
BDQC01000150	unclassified viruses
BDQD01000090	unannotated
BDQD01000173	unclassified ssRNA positive-strand viruses
BDQD01000070	unannotated