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

Shoichi Sakaguchi[1], Syun-ichi Urayama[2], Yoshihiro Takaki[3], Hong Wu[1], Youichi Suzuki[1], Takuro Nunoura[4], Takashi Nakano[1*], So Nakagawa[5*]

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

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

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

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

[5]Department 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 termi, 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;

71 Wolf et al., 2018; Koonin et al., 2020, 2021). It is often challenging to identify RNA virus
72 genes, including the RdRp gene, from the potential complete viral genome sequences
73 obtained by FLDS based on similarity search using BLAST (Basic Local Alignment Search
74 Tool) with major amino acid databases (e.g., nr database or RefSeq database) (Jia and Gong,
75 2019; Urayama et al., 2020).

76 In RdRp domain sequences, there are eight conserved amino acid motifs; notably, the
77 sixth motif containing three amino acid residues, GDD, was widely conserved (Koonin,
78 1991). Based on these RdRp sequence motifs, Wolf and his colleagues inferred the
79 phylogeny of 4,617 RNA viruses (Wolf et al., 2018). However, the accuracy of the multiple
80 alignment of all RdRp amino acid sequences was disputed due to their diversity (Holmes and
81 Duchêne, 2019; Wolf et al., 2019). Therefore, in order to obtain reliable multiple alignments
82 of RdRp sequences, it is necessary to first classify them into groups with high similarity and
83 then create multiple alignments for each group.

84 This study collected RdRp domain sequences from 23,410 RNA viruses and clustered
85 them based on their similarity to build 2,234 hidden Markov model (HMM) profiles RdRp
86 domains. Indeed, an HMM profile is based on a multiple sequence alignment; therefore, it
87 can identify distantly related sequences more efficiently than a pairwise-based search (Mistry
88 et al., 2021). In fact, several bioinformatics approaches using HMM have been devised to
89 find more distantly related viruses, such as VirSorter2 (Guo et al., 2021) and RVDB-prot
90 (Goodacre et al., 2018; Bigot et al., 2019). Thus, this study analyzed RdRp sequences of
91 RNA viruses obtained by FLDS with marine samples and revealed that our method
92 successfully detected RdRp domains efficiently rather than the two abovementioned
93 approaches (i.e., VirSorter2 and RVDB-prot). The constructed RdRp dataset, including
94 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

99 **Materials and Methods**

100 Data sets

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

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

118

119 Clustering based on amino acid sequence similarity of RdRp

120 Amino acid sequences of RdRp genes were clustered using the CD-HIT program version 4.7
121 (Li and Godzik, 2006; Fu et al., 2012) with the following parameters: a similarity threshold
122 of 40 % and word size of 2. For a cluster containing ≥5 genes, amino acid sequences were
123 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 hmmpress (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.

148 **Results**

149 Constructing RdRp HMM profiles

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

159 Furthermore, we conducted comprehensive similarity searches using the 1,467 HMM
160 profiles with 18,790 amino acid sequences of 4,239 RefSeq RNA virus genomes downloaded
161 from NCBI Virus (https://www.ncbi.nlm.nih.gov/labs/virus/) on June 9, 2021. A total of
162 7,696 viral sequences were hit through the hmmsearch. If multiple hits in short intervals on a
163 single gene, a single domain may hit multiple regions. Therefore, if the detected sequences'
164 distance was less than 500 amino acid length, we regarded them as a single region and a
165 domain using BEDTools (Fig. S1) (Quinlan and Hall, 2010). The obtained RdRp motif
166 sequences were combined with Wolf-RdRp, resulting in 12,316 sequences named
167 "NeoRdRp-seq." The NeoRdRp-seq dataset was subjected to incremental clustering by CD-
168 HIT-2D, and as a result, 448 clusters consisting of more than five sequences (including 8,669
169 sequences) were identified. Notably, 1,476 clusters (including 3,123 sequences) containing
170 less than five sequences for each were excluded from the HMM profile construction.
171 Therefore, we performed multiple sequence alignment for 448 clusters using reliable regions,
172 and 2,234 domains were extracted. Fig. 2 shows an example of the alignment clearly

173 indicating the well-conserved motifs, DxxxxxD, GxxxTxxxN, and GDD (Koonin, 1991). For
174 each motif, an HMM profile was created by hmmbuild in HMMER3. We named the 2,234
175 HMM profile set obtained from RdRp sequences as NeoRdRp-HMM version 1.0.
176
177 Evaluating RdRp HMM profiles
178 To annotate each NeoRdRp-HMM, we conducted a comprehensive HMM search using
179 UniProt amino acid sequences [UniProtKB Reviewed (Swiss-Prot) 2021-05-11]. As a result,
180 1,361/2,234 HMM profiles detected at least one amino acid sequence (including cases where
181 multiple HMM profiles detected the same amino acid sequence). Further, 963/995 database
182 hits (96.8 %) were derived from RNA viruses (i.e., annotated as "*Riboviria*"), indicating that
183 the RdRp HMM profiles can identify RNA viruses based on RdRp motifs. Furthermore, it
184 has been reported that, although retroviruses do not contain the RdRp, they have reverse
185 transcriptase partially similar to the RdRp (Wolf et al., 2018). However, in our UniProt
186 search, no retrovirus was detected.
187 We evaluated NeoRdRp-HMM profiles based on the UniProtKB annotation of Gene
188 Ontology (GO; molecular function). First, for each hit was scored it 1 if "RNA-directed 5'-3'
189 RNA polymerase activity" (GO:0003968) was included and 0 if excluded. Then, we
190 calculated a mean value of the scores for each NeoRdRp-HMM profile, summarized in
191 Supplementary Table 1. The average score of the profiles is 0.94, suggesting that most
192 NeoRdRp-HMM profiles annotated by the UniProtKB search were derived from the RdRp of
193 RNA viruses (1,145 HMMs, 84.1%). However, we also found that a few (25 HMMs, 1.8%)
194 had a score of 0; RNA binding [GO:0003723] was the most common with 11 HMM profiles,
195 followed by ATP binding [GO:0005524] with 10 cases, helicase activity [GO:0004386], and
196 hydrolase activity [GO:0016787] with 9, mRNA methyltransferase activity [GO:0008174]

197 and structural molecule activity [GO:0005198] with 8. Thus, these HMM profiles with a
198 score of 0 detect proteins that interact with RNA but do not contain RdRp domains.
199
200 Comparing RdRp identification
201 We benchmarked the ability of NeoRdRp-HMM and NeoRdRp-seq to identify RNA viruses
202 using the potential complete RNA virus genome/genome segment sequences obtained by
203 FLDS associated with marine RNA viromes (Urayama et al., 2018). As a result, 1,143 amino
204 acid sequences were detected in FLDS-data. In addition, Urayama and his colleagues
205 conducted a BLAST search with the nr database and manually annotated them based on the
206 known conserved amino acid motifs (Urayama et al., 2018). As a result, 228 amino acid
207 sequences were annotated as RdRp in the FLDS-data.
208 Using the 1,143 amino acid sequences including 228 RdRps in the FLDS-data, we
209 evaluated NeoRdRp-HMM and NeoRdRp-seq using hmmsearch and tBLASTn, respectively.
210 We merged the results of NeoRdRp-HMM and NeoRdRp-seq searches as NeoRdRp. We also
211 applied VirSorter2 (Guo et al., 2021) and hmmsearch with RVDB-prot (Goodacre et al.,
212 2018) for comparison. First, we analyzed the 228 RdRp genes using the three methods and
213 detected as follows: 70 using all three methods, 112 by NeoRdRp and RVDB-prot, two by
214 NeoRdRp and VirSorter2, 13 only by NeoRdRp, 11 only by RVDB-prot, and 20 were not
215 detected by all methods (Fig. 3). Viral genomes detected only by NeoRdRp were classified
216 into family *Partitiviridae*, *Reoviridae*, *Narnaviridae*, *Flaviviridae*, and *Endornaviridae* based
217 on HMM pattern and genomic structure. However, viral genomes that the three methods
218 could not detect were recognized as members of *Reoviridae* and *Megabirnaviridae* in
219 Urayama's report (Urayama et al., 2018). Then, we further compared NeoRdRp-HMM and
220 NeoRdRp-seq (Supplementary Fig. 1). As a result, 161 RdRp genes were found by both
221 methods, 14 only by NeoRdRp-HMM and 22 only by NeoRdRp-seq. The fact that there were

222 RdRp genes that NeoRdRp-HMM could only detect means that the HMM search can detect
223 diversified RdRp genes in various RNA viruses. On the other hand, there were RdRp
224 sequences that NeoRdRp-seq could only detect because the sequences that were less than 5
225 sequences as a clustering result were not in the HMM profiles. Also, we examined amino
226 acid sequences of FLDS-data that were not annotated as RdRp. As a result, NeoRdRp-HMM
227 detected 15 sequences as RdRps, indicating that our methods can identify unannotated RdRps
228 in metatranscriptome data (Supplementary Table 2).
229 We also compared the computational time required to search for RNA viruses in
230 FLDS-data using the three methods (Table 1): hmmsearch using NeoRdRp-HMM and
231 tBLASTn using NeoRdRp-seq took 25 sec and 47 sec, respectively (i.e. 1 min and 12 sec in
232 total for the NeoRdRp search); hmmsearch using RVDB-prot took 2 min 52 sec; and search
233 using VirSorter took 11 min 46 sec. Furthermore, the database size required for each analysis
234 was 0.15 GB for NeoRdRp, 1.61 GB for RVDB-prot, and 11.3 GB for VirSorter2 (Table 1).
235 Therefore, the NeoRdRp dataset is small and easy to use for RNA virus detection.
236
237 **Discussion**
238 Since amino acid sequences of RdRp varied greatly among all RNA viruses, various
239 bioinformatics attempts have been made to find RdRp genes from metatranscriptome data
240 (Wolf et al., 2018; Bigot et al., 2019; Babaian and Edgar, 2021; Zayed et al., 2021). In this
241 study, RdRp genes were comprehensively obtained (named NeoRdRp-seq), clustered based
242 on amino acid similarity, and HMM profiles were constructed (named NeoRdRp-HMM)
243 (Fig. 1). Using the dataset, we have obtained various RdRp genes from FLDS-data (Fig. 3).
244 However, our dataset cannot identify some RNA viruses belonging to *Reoviridae* and
245 *Megabirnaviridae*. Furthermore, 11/228 RdRp genes were found only using the RVDB-prot
246 dataset, corresponding to RNA viruses belonging to *Narnaviridae*, *Picobirnaviridae*,

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

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

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

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

279 Furthermore, our method can detect various RdRps with high accuracy by creating an
280 RdRp dataset, clustering the RdRp based on similarity, and creating HMM profiles for each
281 RdRp. All sequence data and annotation developed in this study are available on the website
282 (https://github.com/shoichisakaguchi/NeoRdRp). Therefore, RNA virome searches will be
283 more accurate with RNA virus-selective metatranscriptome sequencing data using the
284 dataset. Additionally, the NeoRdRp dataset is very compact compared with other methods for
285 RNA virome search (Table 1), and the results are clear-cut due to the targeted approach.
286 Furthermore, by adding new RdRp sequences in the future, it will be possible to find RdRp
287 sequences with high accuracy from metatranscriptome data, which should become an
288 essential technique for discovering unknown viruses.
289

290 **Acknowledgments**

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

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

301 **References**

302 Ali, R. H., Bogusz, M., and Whelan, S. (2019). Identifying Clusters of High Confidence
303 Homologies in Multiple Sequence Alignments. Mol. Biol. Evol. 36, 2340–2351.
304 Babaian, A., and Edgar, R. C. (2021). Ribovirus classification by a polymerase barcode
305 sequence. bioRxiv, 2021.03.02.433648. doi:10.1101/2021.03.02.433648.
306 Bigot, T., Temmam, S., Pérot, P., and Eloit, M. (2019). RVDB-prot, a reference viral
307 protein database and its HMM profiles. F1000Res. 8, 530.
308 Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., Bansal, P., Bridge, A. J., et al.
309 (2016). UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt
310 KnowledgeBase: How to Use the Entry View. Methods Mol. Biol. 1374, 23–54.
311 Černý, J., Černá Bolfíková, B., Valdés, J. J., Grubhoffer, L., and Růžek, D. (2014).
312 Evolution of tertiary structure of viral RNA dependent polymerases. PLoS One 9,
313 e96070.
314 Cobbin, J. C., Charon, J., Harvey, E., Holmes, E. C., and Mahar, J. E. (2021). Current
315 challenges to virus discovery by meta-transcriptomics. Curr. Opin. Virol. 51, 48–55.
316 Crooks, G. E., Hon, G., Chandonia, J.-M., and Brenner, S. E. (2004). WebLogo: a
317 sequence logo generator. Genome Res. 14, 1188–1190.
318 Eddy, S. R. (2011). Accelerated Profile HMM Searches. PLoS Comput. Biol. 7,
319 e1002195.
320 Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering
321 the next-generation sequencing data. Bioinformatics 28, 3150–3152.
322 Goodacre, N., Aljanahi, A., Nandakumar, S., Mikailov, M., and Khan, A. S. (2018). A
323 Reference Viral Database (RVDB) To Enhance Bioinformatics Analysis of High-
324 Throughput Sequencing for Novel Virus Detection. mSphere 3.
325 doi:10.1128/mSphereDirect.00069-18.

326 Guo, J., Bolduc, B., Zayed, A. A., Varsani, A., Dominguez-Huerta, G., Delmont, T. O., et
327 al. (2021). VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA
328 and RNA viruses. Microbiome 9, 37.
329 Haas, B., Papanicolaou, A., and Others (2016). TransDecoder (find coding regions within
330 transcripts). URL http://transdecoder.github.io.
331 Hirai, M., Takaki, Y., Kondo, F., Horie, M., Urayama, S.-I., and Nunoura, T. (2021).
332 RNA Viral Metagenome Analysis of Subnanogram dsRNA Using Fragmented and
333 Primer Ligated dsRNA Sequencing (FLDS). Microbes Environ. 36.
334 doi:10.1264/jsme2.ME20152.
335 Holmes, E. C., and Duchêne, S. (2019). Can Sequence Phylogenies Safely Infer the
336 Origin of the Global Virome? MBio 10. doi:10.1128/mBio.00289-19.
337 Jia, H., and Gong, P. (2019). A Structure-Function Diversity Survey of the RNA-
338 Dependent RNA Polymerases From the Positive-Strand RNA Viruses. Front. Microbiol.
339 10, 1945.
340 Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software
341 version 7: improvements in performance and usability. Mol. Biol. Evol. 30, 772–780.
342 Koonin, E. V. (1991). The phylogeny of RNA-dependent RNA polymerases of positive-
343 strand RNA viruses. J. Gen. Virol. 72 (Pt 9), 2197–2206.
344 Koonin, E. V., Dolja, V. V., Krupovic, M., and Kuhn, J. H. (2021). Viruses defined by
345 the position of the virosphere within the replicator space. Microbiol. Mol. Biol. Rev.,
346 e0019320.
347 Koonin, E. V., Dolja, V. V., Krupovic, M., Varsani, A., Wolf, Y. I., Yutin, N., et al.
348 (2020). Global Organization and Proposed Megataxonomy of the Virus World.
349 Microbiol. Mol. Biol. Rev. 84. doi:10.1128/MMBR.00061-19.

350 Kumar, M., and Carmichael, G. G. (1998). Antisense RNA: function and fate of duplex
351 RNA in cells of higher eukaryotes. Microbiol. Mol. Biol. Rev. 62, 1415–1434.
352 Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large
353 sets of protein or nucleotide sequences. Bioinformatics 22, 1658–1659.
354 Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E.
355 L. L., et al. (2021). Pfam: The protein families database in 2021. Nucleic Acids Res. 49,
356 D412–D419.
357 Orba, Y., Matsuno, K., Nakao, R., Kryukov, K., Saito, Y., Kawamori, F., et al. (2021).
358 Diverse mosquito-specific flaviviruses in the Bolivian Amazon basin. J. Gen. Virol. 102.
359 doi:10.1099/jgv.0.001518.
360 Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for
361 comparing genomic features. Bioinformatics 26, 841–842.
362 Sakaguchi, S., Nakagawa, S., Mitsuhashi, S., Ogawa, M., Sugiyama, K., Tamukai, K., et
363 al. (2020). Molecular characterization of feline paramyxovirus in Japanese cat
364 populations. Arch. Virol. 165, 413–418.
365 Shi, M., Zhang, Y.-Z., and Holmes, E. C. (2018). Meta-transcriptomics and the
366 evolutionary biology of RNA viruses. Virus Res. 243, 83–90.
367 Urayama, S.-I., Takaki, Y., Hagiwara, D., and Nunoura, T. (2020). dsRNA-seq Reveals
368 Novel RNA Virus and Virus-Like Putative Complete Genome Sequences from
369 Hymeniacidon sp. Sponge. Microbes Environ. 35. doi:10.1264/jsme2.ME19132.
370 Urayama, S.-I., Takaki, Y., Nishi, S., Yoshida-Takashima, Y., Deguchi, S., Takai, K., et
371 al. (2018). Unveiling the RNA virosphere associated with marine microorganisms. Mol.
372 Ecol. Resour. 18, 1444–1455.

373 Urayama, S.-I., Takaki, Y., and Nunoura, T. (2016). FLDS: A Comprehensive dsRNA
374 Sequencing Method for Intracellular RNA Virus Surveillance. Microbes Environ. 31, 33–
375 40.
376 Wolf, Y. I., Kazlauskas, D., Iranzo, J., Lucía-Sanz, A., Kuhn, J. H., Krupovic, M., et al.
377 (2018). Origins and Evolution of the Global RNA Virome. MBio 9.
378 doi:10.1128/mBio.02329-18.
379 Wolf, Y. I., Kazlauskas, D., Iranzo, J., Lucía-Sanz, A., Kuhn, J. H., Krupovic, M., et al.
380 (2019). Reply to Holmes and Duchêne, "Can Sequence Phylogenies Safely Infer the
381 Origin of the Global Virome?": Deep Phylogenetic Analysis of RNA Viruses Is Highly
382 Challenging but Not Meaningless. MBio 10. doi:10.1128/mBio.00542-19.
383 Zayed, A. A., Lücking, D., Mohssen, M., Cronin, D., Bolduc, B., Gregory, A. C., et al.
384 (2021). efam: an expanded, metaproteome-supported HMM profile database of viral
385 protein families. Bioinformatics. doi:10.1093/bioinformatics/btab451.
386

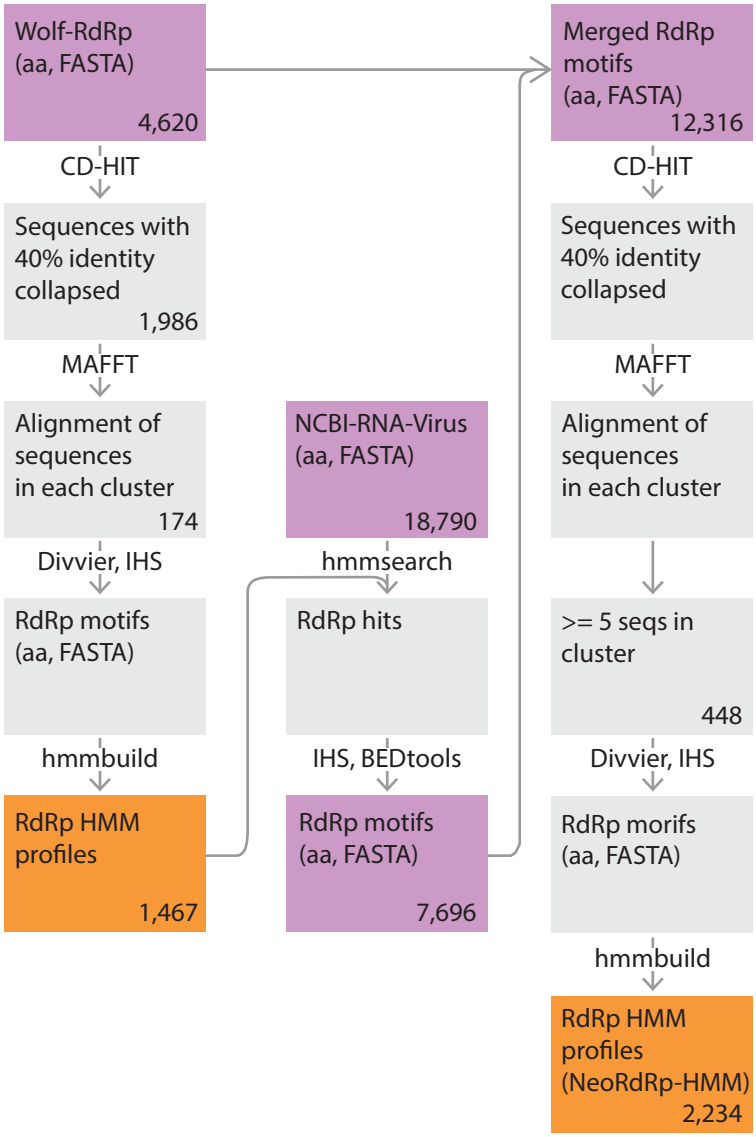


Fig. 1. A schematic workflow of NeoRdRp-HMM construction

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.

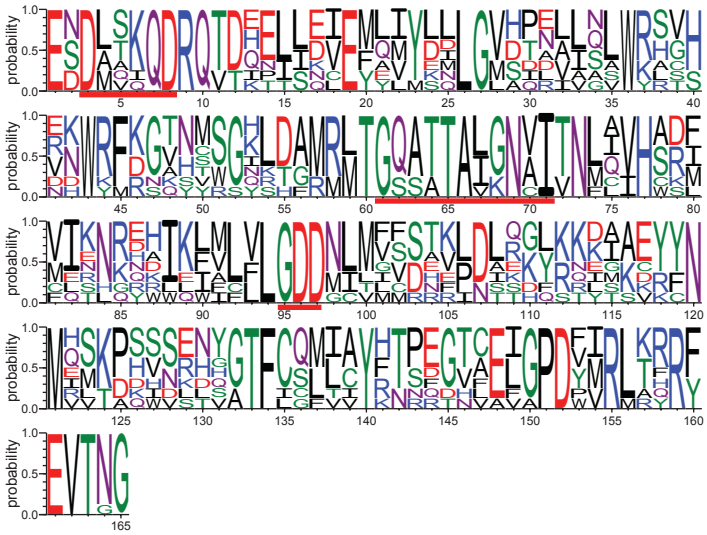


Fig. 2. An example of a multiple sequence alignment of amino acid sequences of the RdRp domain.

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.

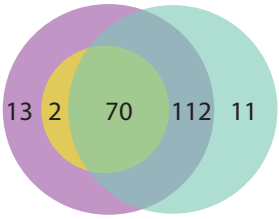
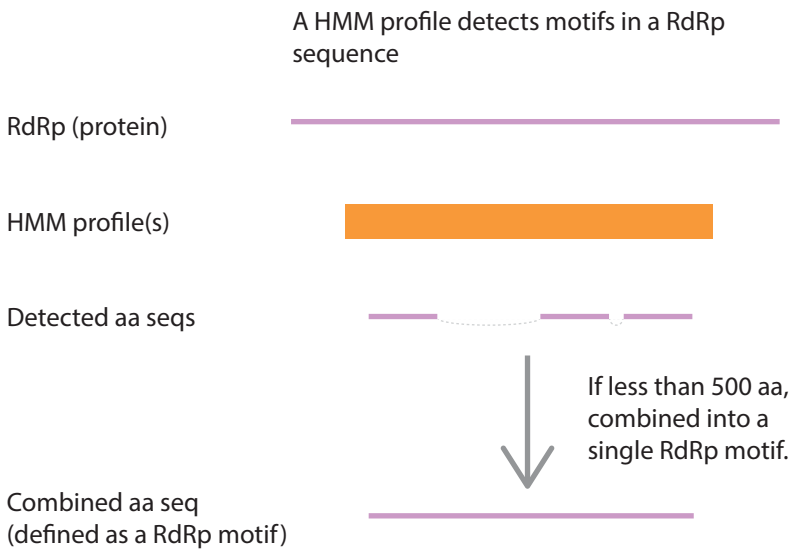


Fig. 3. Venn diagram of three different methods for identifying RNA viruses. Annotated marine metatranscriptome data were searched using NeoRdRp (hmmsearch and tBLASTn), RVDB-prot (hmmsearch), and VirSorter2, respectively. Of all RdRp sequences detected, 13 were found only in NeoRdRp. Purple: NeoRdRp (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/ NeoRdRp-Seq	hmmsearch/ tBLASTn	HMM/ blast (FASTA)	0.15/ 0.02	25 sec/ 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

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

431 (provided as an Excel File)

432

433

434

435 **Supplementary Table 2. Unannotated RdRps in metatranscriptome data detected by**

436 **NeoRdRp.**

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

437