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[1*]

[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:

so@tokai.ac.jp, tnakano@ompu.ac.jp

Keywords

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

Running headline: Bioinformatics pipeline 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 their sequences. Therefore, this study generated a dataset of RNA-dependent RNA polymerase (RdRp) domains essential for all RNA viruses belonging to Orthornavirae. Also, the collected 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. Additionally, using the UniProt dataset, we confirmed that almost all NeoRdRp HMM profiles could specifically detect RdRps in Orthornavirae. Furthermore, we compared the NeoRdRp dataset with two previously reported RNA virus detection methods to detect 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 RNA viruses

50 similar to human infectious viruses, such as coronaviruses, orthomyxoviruses, and
51 filoviruses, found in several non-mammalian/avian species, including fishes (Shi et al.,
52 2018). However, identifying RNA virus-derived sequences from metatranscriptome data is
53 challenging due to the low abundance of viral RNA in sequencing reads and high nucleotide
54 diversity among RNA viruses (Cobbin et al., 2021). Thus, we have developed a method of
55 double-strand RNA (dsRNA) sequencing named fragmented and primer ligated dsRNA
56 sequencing (FLDS) (Urayama et al., 2016, 2018; Hirai et al., 2021). In eukaryotic cells, long
57 dsRNA are rare; therefore, they are mainly derived from the genomes of dsRNA viruses or
58 replicative intermediates of single-strand RNA (ssRNA) viruses (Kumar and Carmichael,
59 1998). Additionally, FLDS enables us to obtain entire dsRNA sequences, including both
60 termi, and reconstruct multi-segmented genomes of RNA viruses based on terminal sequence
61 similarity (Urayama et al., 2016). Accordingly, FLDS can enrich potential RNA virus
62 genomes even though their sequences are not similar to known RNA viruses. However,
63 confirming RNA viral genomes based on the FLDS data remains a problem, mainly because
64 of the difficulty in identifying RNA virus-specific genes.
65 Viruses that utilize RNA as a genetic material form a monophyletic group named
66 Riboviria (ICTV) based on taxonomy. Orthornavirae is a major Riboviria kingdom
67 containing all the dsRNA and ssRNA viruses — but excluding retroviruses — that harbors
68 RNA-dependent RNA polymerase (RdRp). Therefore, RdRp genes can be used as markers
69 for most RNA viruses, whereas their nucleotide and amino acid sequences vary greatly, and
70 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

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). Thus, it is necessary to yield reliable multiple alignments of RdRp sequences by grouping them with high similarity and obtaining a reliable alignment.

This study collected and annotated RdRp domain sequences from 23,410 RNA viruses and constructed 2,234 hidden Markov model (HMM) profiles of RdRp motifs by clustering their similarities. HMM profiles are based on multiple sequence alignments of given homologous genes; therefore, they 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 (Goodacre et al., 2018; Bigot et al., 2019). Thus, this study analyzed RdRp sequences of RNA viruses obtained by FLDS with marine samples using our method and the two abovementioned approaches (i.e., VirSorter2 and RVDB). As a result, our method successfully detected RNA virus sequences, not by similarity search with significant reliability. The constructed RdRp dataset, including annotation, is publicly shared online (https://github.com/shoichisakaguchi/NeoRdRp). Furthermore, the RdRp dataset developed in this study could identify possible RdRp genes of various RNA viruses from metatranscriptomic sequencing data.

Materials and Methods

99 Data sets

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

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

117

118 Clustering based on amino acid sequence similarity of RdRp

119 Amino acid sequences of RdRp genes were clustered using the CD-HIT program v.4.7 (Li
120 and Godzik, 2006; Fu et al., 2012) with the following parameters: a similarity threshold of
121 40 % and word size of 2. For a cluster containing ≥5 genes, amino acid sequences were
122 aligned using the L-INS-i program in MAFFT version 7.450 (Katoh and Standley, 2013). We
123 selected regions with strong statistical evidence of shared homology for each multiple

124 sequence alignment and removed uncertainty and error regions using Divvier version 1.0 (Ali
125 et al., 2019). Furthermore, we split a multiple sequence alignment if more than one-quarter of
126 the sequences were gap at the position. Notably, a motif was defined by multiple sequence
127 alignment consisting of nine amino acid lengths. Multiple alignments of amino acid
128 sequences were visualized using weblogo version 3.7.8 (Crooks et al., 2004). Height
129 represents the probability at each site, and the width is proportional to the number of
130 sequences.
131
132 Creation of Hidden-Markov-Model (HMM) profiles
133 The HMM profiles were built for each multiple-sequence-alignment using hmmbuild, part of
134 the HMMER3 package version 3.1b2 (Eddy, 2011) with default parameters. Then, all HMM
135 profiles were compressed into an HMMER3 searchable database using hmmpress (from the
136 HMMER3 package).
137
138 Other RNA virus detection software and databases
139 VirSorter2 (Guo et al., 2021) and hmmsearch with HMM profiles of Reference Viral protein
140 DataBase (RVDB-prot) (Bigot et al., 2019) were used to identify the potential complete RNA
141 virus genome/genome segment sequences in FLDS-data. Additionally, VirSorter2 was used
142 with an option "the search group set to RNA viruses only", and other parameters were set to
143 default. Finally, RVDB-prot version 22.0 was used with hmmsearch with an E-value of 1E-
144 10.
145
146 **Results**
147 Constructing RdRp HMM profiles

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

157 Furthermore, we conducted comprehensive similarity searches using the 1,467 HMM
158 profiles with 18,790 amino acid sequences of 4,239 RefSeq RNA virus genomes downloaded
159 from NCBI Virus (https://www.ncbi.nlm.nih.gov/labs/virus/) on June 9, 2021. A total of
160 7,696 viral sequences were hit through the hmmsearch. If multiple hits in short intervals on a
161 single gene, a single domain may hit multiple regions. Therefore, if the detected sequences'
162 distance was less than 500 amino acid length, we regarded them as a single region and a
163 domain using BEDTools (Fig. S1) (Quinlan and Hall, 2010). The obtained RdRp motif
164 sequences were combined with Wolf-RdRp, resulting in 12,316 sequences named
165 "NeoRdRp-seq." The NeoRdRp-seq dataset was subjected to incremental clustering by CD-
166 HIT-2D, and as a result, 448 clusters consisting of more than five sequences (including 8,669
167 sequences) were identified. Notably, 1,476 clusters (including 3,123 sequences) containing
168 less than five sequences for each were excluded from the HMM profile construction.
169 Therefore, we performed multiple sequence alignment for 448 clusters using reliable regions,
170 and 2,234 motifs were extracted. Fig. 2 shows an example of the RdRp domain alignment.
171 The well-conserved motifs, DxxxxxD, GxxxTxxxN, and GDD (Koonin, 1991), were found.

172 For each motif, HMM profiles were created by hmmbuild in HMMER3. We named the 2,234
173 HMM profile set obtained from RdRp sequences as NeoRdRp-HMM version 1.0.
174
175 Evaluating RdRp HMM profiles
176 To annotate each NeoRdRp-HMM, we conducted a comprehensive HMM search using
177 UniProt amino acid sequences [UniProtKB Reviewed (Swiss-Prot) 2021-05-11]. As a result,
178 1,361/2,234 HMM profiles detected at least one amino acid sequence (including cases where
179 multiple HMM profiles detected the same amino acid sequence). Further, 963/995 database
180 hits (96.8 %) were derived from RNA viruses (i.e., annotated as "Riboviria"), indicating that
181 the RdRp HMM profiles can identify RNA viruses based on RdRp motifs. Furthermore, it
182 has been reported that, although retroviruses do not contain the RdRp, they have reverse
183 transcriptase partially similar to the RdRp (Wolf et al., 2018). However, in our UniProt
184 search, no retrovirus was detected.
185 We evaluated NeoRdRp-HMM profiles based on the UniProtKB annotation of Gene
186 Ontology (GO; molecular function). First, for each hit was scored it 1 if "RNA-directed 5'-3'
187 RNA polymerase activity" (GO:0003968) was included and 0 if excluded. Then, we
188 calculated a mean value of the scores for each NeoRdRp-HMM profile, summarized in
189 Supplementary Table 1. The average score of the profiles is 0.94, suggesting that most
190 NeoRdRp-HMM profiles annotated by the UniProtKB search were derived from the RdRp of
191 RNA viruses (1,145 HMMs, 84.1%). However, we also found that a few (25 HMMs, 1.8%)
192 had a score of 0; RNA binding [GO:0003723] was the most common with 11 HMM profiles,
193 followed by ATP binding [GO:0005524] with 10 cases, helicase activity [GO:0004386], and
194 hydrolase activity [GO:0016787] with 9, mRNA methyltransferase activity [GO:0008174]
195 and structural molecule activity [GO:0005198] with 8. Thus, these HMM profiles with a
196 score of 0 detect proteins that interact with RNA but do not contain RdRp domains.

197

198 Comparing RdRp identification

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

206 Using the 1,143 amino acid sequences including 228 RdRps in the FLDS-data, we
207 evaluated NeoRdRp-HMM and NeoRdRp-seq using hmmsearch and tBLASTn, respectively.
208 We merged the results of NeoRdRp-HMM and NeoRdRp-seq searches as NeoRdRp. We also
209 applied VirSorter2 (Guo et al., 2021) and hmmsearch with RVDB (Goodacre et al., 2018) for
210 comparison. First, we analyzed the 228 RdRp genes using the three methods and detected as
211 follows: 70 using all three methods, 112 by NeoRdRp and RVDB, two by NeoRdRp and
212 VirSorter2, 13 only by NeoRdRp, 11 only by RVDB, and 20 were not detected by all
213 methods (Fig. 3). Viral genomes detected only by NeoRdRp were classified into family
214 Partitiviridae, Reoviridae, Narnaviridae, Flaviviridae, and Endornaviridae based on HMM
215 pattern and genomic structure. However, viral genomes that the three methods could not
216 detect were recognized as members of Reoviridae and Megabirnaviridae in Urayama's report
217 (Urayama et al., 2018). Then, we further compared NeoRdRp-HMM and NeoRdRp-seq
218 (Supplementary Fig. 1). As a result, 161 RdRp genes were found by both methods, 14 only
219 by NeoRdRp-HMM and 22 only by NeoRdRp-seq. The fact that there were RdRp genes that
220 NeoRdRp-HMM could only detect means that the HMM search can detect diversified RdRp
221 genes in various RNA viruses. On the other hand, there were RdRp sequences that NeoRdRp-

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

247 incorporating newly detected RdRp sequences. Thus, in principle, it is possible to find more
248 RdRp sequences by repeating updates. Notably, the clusters that comprised less than five
249 sequences were excluded from HMM profile construction to maintain the accuracy of RdRp
250 domain identification. Additionally, as the number of RdRp sequences increases, the number
251 of clusters used for HMM construction increases. As a result, the detection power must
252 increase.

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

265 Therefore, the NeoRdRp dataset can be applied to the following research designs.
266 First, the metatranscriptomic data obtained by FLDS were assembled. Then, ORFs were
267 searched in each contig, and RdRp domains were searched for each ORF using NeoRdRp-
268 HMM. Undetected, remaining sequences were further searched with NeoRdRp-seq, a
269 relatively computationally intensive process (Table 1). From the results of these analyses and
270 their scores, we can identify contigs containing the RdRp domain, suggesting that they are
271 derived from RNA viruses. Furthermore, FLDS can identify multi-segmented genomes of

272 RNA viruses based on terminal sequence similarity (Urayama et al., 2016). Therefore, if one
273 contig is found to contain an RdRp domain, other contigs having the same sequence on their
274 terminals are strongly suggested to be derived from RNA viruses. Also, in those contigs,
275 other viral genes can be detected using other computational programs and datasets, including
276 VirSorter2 and RVDB.

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

288 **Acknowledgments**

289 This work was supported by KAKENHI Grant-in-Aid for Scientific Research (C) 20K06775
290 (to S.S. and S.N.), Early-Career Scientists 20K15685 (to S.S.), Scientific Research on
291 Innovative Areas 16H06429 (to S.U. and S.N.), 16K21723 (to S.U. and S.N.), 16H06437 (to
292 S.U.), 18H05368 (to T.N.), 19H04843 (to S.N.), and by AMED under Grant Numbers
293 JP19fm0208009 (to S.U.), JP20wm0325004 (to S.U.), and JP21wm0325004 (to S.U. and
294 S.N.). The computing resource was partially supported by the NIG supercomputer at ROIS
295 National Institute of Genetics, Japan, and SHIROKANE at the Human Genome Center, the
296 Institute of Medical Science, the University of Tokyo, Japan.

297 **References**

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

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

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

369 Urayama, S.-I., Takaki, Y., and Nunoura, T. (2016). FLDS: A Comprehensive dsRNA
370 Sequencing Method for Intracellular RNA Virus Surveillance. Microbes Environ. 31, 33–
371 40.
372 Wolf, Y. I., Kazlauskas, D., Iranzo, J., Lucía-Sanz, A., Kuhn, J. H., Krupovic, M., et al.
373 (2018). Origins and Evolution of the Global RNA Virome. MBio 9.
374 doi:10.1128/mBio.02329-18.
375 Wolf, Y. I., Kazlauskas, D., Iranzo, J., Lucía-Sanz, A., Kuhn, J. H., Krupovic, M., et al.
376 (2019). Reply to Holmes and Duchêne, "Can Sequence Phylogenies Safely Infer the
377 Origin of the Global Virome?": Deep Phylogenetic Analysis of RNA Viruses Is Highly
378 Challenging but Not Meaningless. MBio 10. doi:10.1128/mBio.00542-19.
379 Zayed, A. A., Lücking, D., Mohssen, M., Cronin, D., Bolduc, B., Gregory, A. C., et al.
380 (2021). efam: an expanded, metaproteome-supported HMM profile database of viral
381 protein families. Bioinformatics. doi:10.1093/bioinformatics/btab451.
382

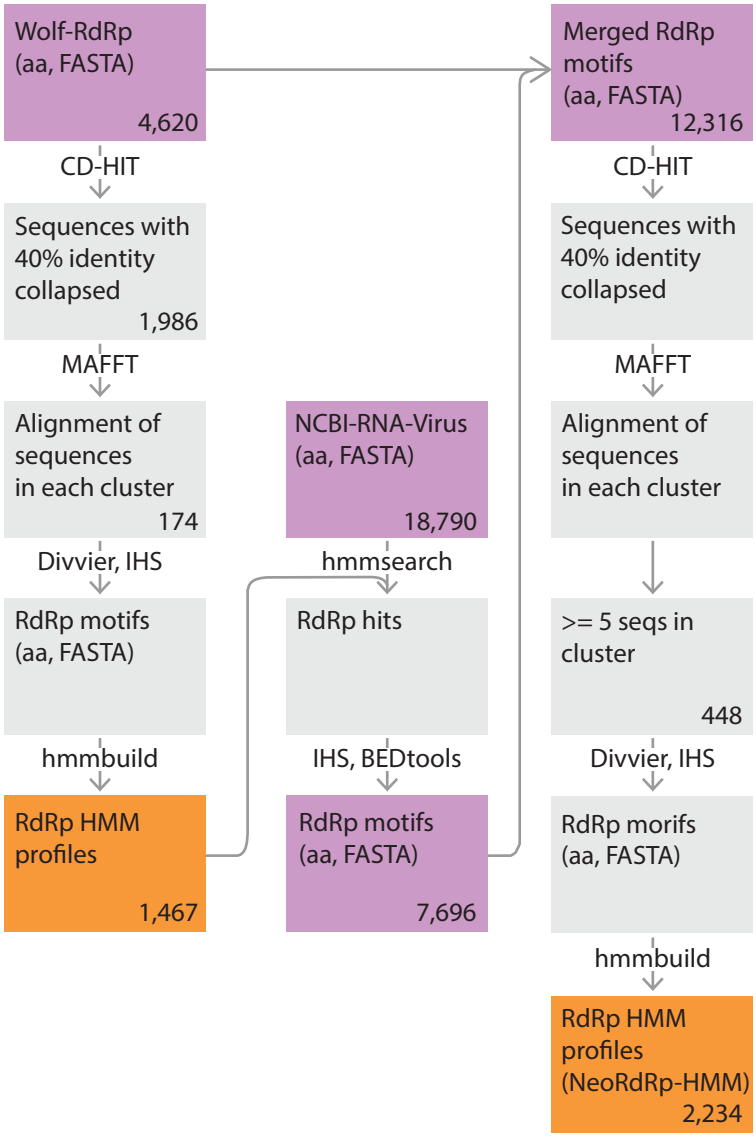


Fig. 1. A schematic workflow of NeoRdRp-HMM construction

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 sequence 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.

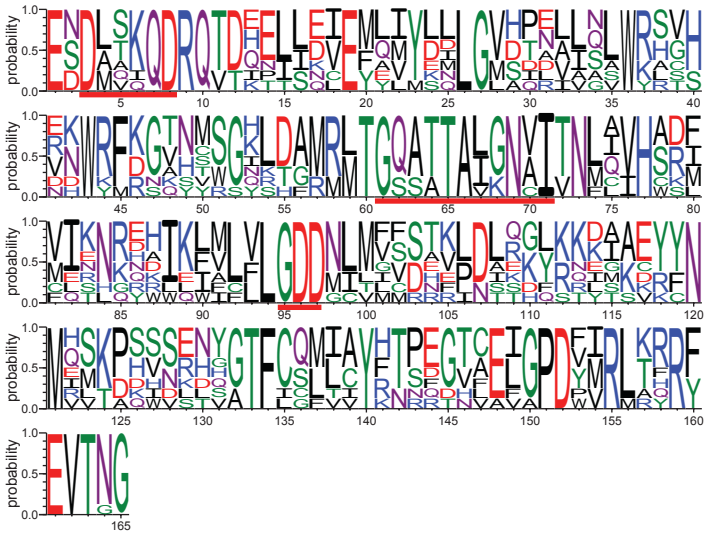


Fig. 2. An example of a multiple sequence alignment of amino acid sequences of the RdRp domain.

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.

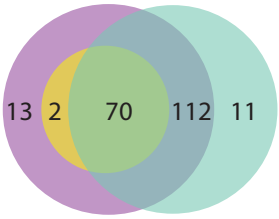
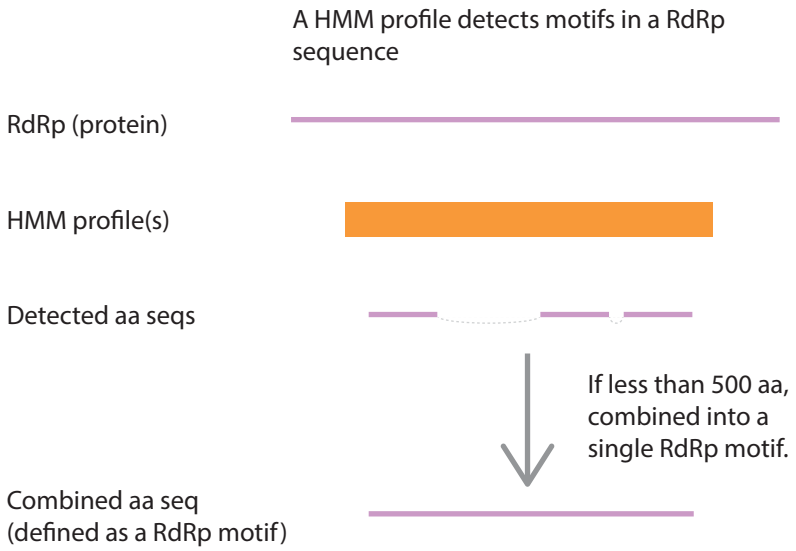


Fig. 3. Venn diagram of three different methods for identifying RNA viruses. Annotated marine metatranscriptome data were searched using NeoRdRp (hmmsearch and tBLASTn), RVDB (hmmsearch), and VirSorter2, respectively. Of all RdRp sequences detected, 13 were found only in NeoRdRp. Purple: NeoRdRp (HMM + tBLASTn), Green: RVDB, 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	hmmsearch	HMM	1.61	2 min 52 sec
VirSorter2	VirSorter2	VS2	11.13	11 min 46 sec



Supplementary Figure 1. Defining an RdRp motif.

The detected sequences' distance was less than 500 amino acid length in the hmmsearch result; we regarded them as a single region as a domain.

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

161 RdRp genes were found by both methods, 14 only by NeoRdRp-HMM, and 22 only by NeoRdRp-seq. Purple: NeoRdRp-HMM, green: NeoRdRp-seq

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

427 (provided as an Excel File)

428

429

430

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

432 **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

433