



HOKKAIDO UNIVERSITY

Title	Current trends in RNA virus detection through metatranscriptome sequencing data
Author(s)	Nakagawa, So; Sakaguchi, Shoichi; Ogura, Atsushi et al.
Citation	FEBS Open Bio, 13(6), 992-1000 https://doi.org/10.1002/2211-5463.13626
Issue Date	2023-05-10
Doc URL	https://hdl.handle.net/2115/96094
Rights(URL)	https://creativecommons.org/licenses/by/4.0/
Type	journal article
File Information	FEBS Open Bio 13-6_951-1133.pdf



REVIEW

Current trends in RNA virus detection through metatranscriptome sequencing data

So Nakagawa^{1,2,3,4} , Shoichi Sakaguchi⁵, Atsushi Ogura⁶ , Katsuhiko Mineta^{7,8,9} , Toshinori Endo¹⁰, Yoshiyuki Suzuki¹¹ and Takashi Gojobori^{7,9}

1 Department of Molecular Life Science, Tokai University School of Medicine, Kanagawa, Japan

2 Division of Genome Sciences, Institute of Medical Sciences, Tokai University, Kanagawa, Japan

3 Division of Interdisciplinary Merging of Health Research, Micro/Nano Technology Center, Tokai University, Kanagawa, Japan

4 Bioinformation and DDBJ Center, National Institute of Genetics, Mishima, Japan

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

6 Graduate School of Bioscience, Nagahama Institute of Bioscience and Technology, Japan

7 Computational Bioscience Research Center (CBRC), King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

8 Research Organization for Nano & Life Innovation, Waseda University, Tokyo, Japan

9 Marine Open Innovation Institute (MaOI), Shizuoka, Japan

10 Faculty of Information Science and Technology, Hokkaido University, Sapporo, Japan

11 Graduate School of Science, Nagoya City University, Japan

Keywords

endogenous viral elements;
metatranscriptome; RNA virome; RNA-
dependent RNA polymerase; SARS-CoV-2

Correspondence

S. Nakagawa, Department of Molecular Life
Science, Tokai University School of
Medicine, Kanagawa 259-1193, Japan
E-mail: so@tokai.ac.jp

(Received 1 November 2022, revised 1 May
2023, accepted 8 May 2023)

doi:10.1002/2211-5463.13626

With advances in sequencing technology, metatranscriptome sequencing from a variety of environmental and biological sources has revealed the existence of various previously unknown RNA viruses. This review presents recent major RNA virome studies sampled from invertebrate and vertebrate species as well as aquatic environments. In particular, we focus on severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) and related RNA virus identification through metatranscriptome sequencing analyses. Recently developed bioinformatics software and databases for RNA virus identification are introduced. A relationship between newly identified RNA viruses and endogenous viral elements in host genomes is also discussed.

Viruses are often classified as something between organisms and material. Although viruses own their genetic materials and successfully reproduce themselves in host cells, they cannot grow and replicate by themselves. Viruses that utilize RNA as a genetic material comprise realms named *Riboviria* and *Ribozyviria* (International Committee on Taxonomy of Viruses, <https://talk.ictvonline.org/>, access on April 26, 2023). *Orthornavirae* is a major kingdom of *Riboviria* that contains all the double-strand RNA (dsRNA) and

single-strand RNA (ssRNA) viruses, excluding retroviruses (belonging to kingdom *Pararnavirae*). For their genome replications, RNA viral genomes of *Orthornavirae* code RNA-dependent RNA polymerase (RdRp), while retroviruses code RNA-dependent DNA polymerase (i.e., reverse transcriptase). As for *Ribozyviria*, it consists of only a single family *Kolmioviridae* containing *Deltavirus* such as hepatitis delta virus [1]. As RNA genomes of *Ribozyviria* do not code RdRp, they utilize RdRp of other RNA viruses or host replication

Abbreviation

cDNA, complementary DNA; CNN, convolutional neural network; dsRNA, double-strand RNA; ERV, endogenous retroviruses; EVE, endogenous viral element; FLDS, fragmented and primer ligated dsRNA sequencing; HMM, hidden Markov model; MLP, multilayer perceptron; RdRp, RNA-dependent RNA polymerase; RNN, recurrent neural network; SARS-CoV-2, severe acute respiratory syndrome coronavirus-2; ssRNA, single-strand RNA; SVM, support vector machine.

enzymes for their replications [2]. Furthermore, it is known that some *Eptesicus* bats also code RdRp that was integrated into their genomes [3]. However, despite several exceptions, genes containing the RdRp domain are almost unique to all RNA viruses. Therefore, the presence of RdRp is known as the hallmark of RNA viruses [4]. Recently, many RNA viruses that used RNA as a genetic material were found to exist in various environments, including wild animals [5–9], oceans [10,11], and even in the genomes of various species [12–18]. The discovery of such a large number of viruses in recent years is largely due to advances in DNA sequencing technology. At the same time, the development of bioinformatics technology to analyze such large-scale nucleotide sequences has made significant contributions. In this review, to understand how metatranscriptome analyses have been used to discover RNA viruses, especially concerning severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2), we first briefly introduce recent major metatranscriptome RNA virome analyses. Then, we summarized bioinformatics tools for RNA virus detection. We also address the endogenous viral elements studies that could be related to novel RNA virus detection.

Metatranscriptome analysis for the RNA virome

RNA viruses have been comprehensively identified using metatranscriptome (RNA metagenome) data. The standard procedure of metatranscriptome analysis is summarized in Fig. 1. First, RNA is extracted from a given sample. After removing ribosomal RNA genes, reverse transcriptase is used to generate complementary DNA (cDNA) from the remaining RNA. Then, cDNAs are enriched by PCR using random primers. The enriched cDNAs are massively sequenced by so-called next-generation sequencers. The sequencing data are assembled, and contigs are generated. Note that this assembly step may be unnecessary if long-read sequencing is applied [19]. Each contig is searched using viral and host genomes and/or genes. Based on the sequence search results, viral genomes are thought to exist in a given sample. Recent advances in sequencing and computer technology have enabled metatranscriptome sequencing at a more affordable price, and many laboratories utilize metatranscriptome methods to identify RNA viruses from various samples.

Arthropod RNA virome

Since arthropod species transmit various RNA viruses that infect humans, various metatranscriptome

analyses of arthropod species have been conducted and revealed a vast diversity of RNA viruses not previously isolated. Li *et al.* [5] conducted metatranscriptome analysis of 70 arthropod species and identified 112 novel negative-sense RNA viruses. Shi *et al.* [6] also conducted a metatranscriptome analysis of invertebrates such as mosquitoes and ticks and reported thousands of previously undetected RNA viruses. Orba *et al.* [20] sampled mosquitos in the Bolivian Amazon basin and revealed flaviviruses with distinct features. Metatranscriptome analyses of human clinical samples have identified various arthropod-derived RNA viruses as the cause of the various symptoms. As examples, Alongshan virus belonging to *Jingmenvirus* was found to be associated with human febrile illness [21], while Yezo virus belonging to *Orthonairovirus* was identified in patients showing acute febrile illness with thrombocytopenia and leukopenia after a tick bite [22].

SARS-CoV-2 related RNA virome

One of the most prominent examples of metatranscriptome analyses for RNA virus detection was the discovery of SARS-CoV-2 and related coronaviruses. The genome sequence of SARS-CoV-2, a coronavirus causing severe respiratory illness named COVID-19, was revealed by metatranscriptome analyses [23–25]. Before the COVID-19 pandemic, the closest coronavirus to SARS-CoV-2 was RaTG13, which was isolated from a horseshoe bat (*Rhinolophus affinis*) [26]. In addition, coronaviruses detected in Sunda pangolins (*Manis javanica*) were also found to be similar to SARS-CoV-2, particularly with respect to the receptor-binding domain of the spike protein [27–30]. Since the start of the pandemic, many metatranscriptome studies have identified various coronaviruses that are closely related to SARS-CoV-2 [31–34]; however, ancestral viruses of SARS-CoV-2 have not been found to date (as of April 2023). Furthermore, although a vast metatranscriptome analysis of various wild animals sampled from multiple regions in China revealed a variety of novel RNA viruses, no coronaviruses that were related to SARS-CoV-2 were found [8,9]. Essentially, the evolution of RNA viruses that cause human disease is complex because various RNA viruses that infect humans are not limited to mammals, but closely related RNA viruses also exist in fish [7]. Indeed, a recent study revealed fish could be the first host of the genus influenza virus that also infects humans [35]. Further RNA virome studies will reveal the diversity and evolution of RNA viruses that cause human diseases, including SARS-CoV-2.

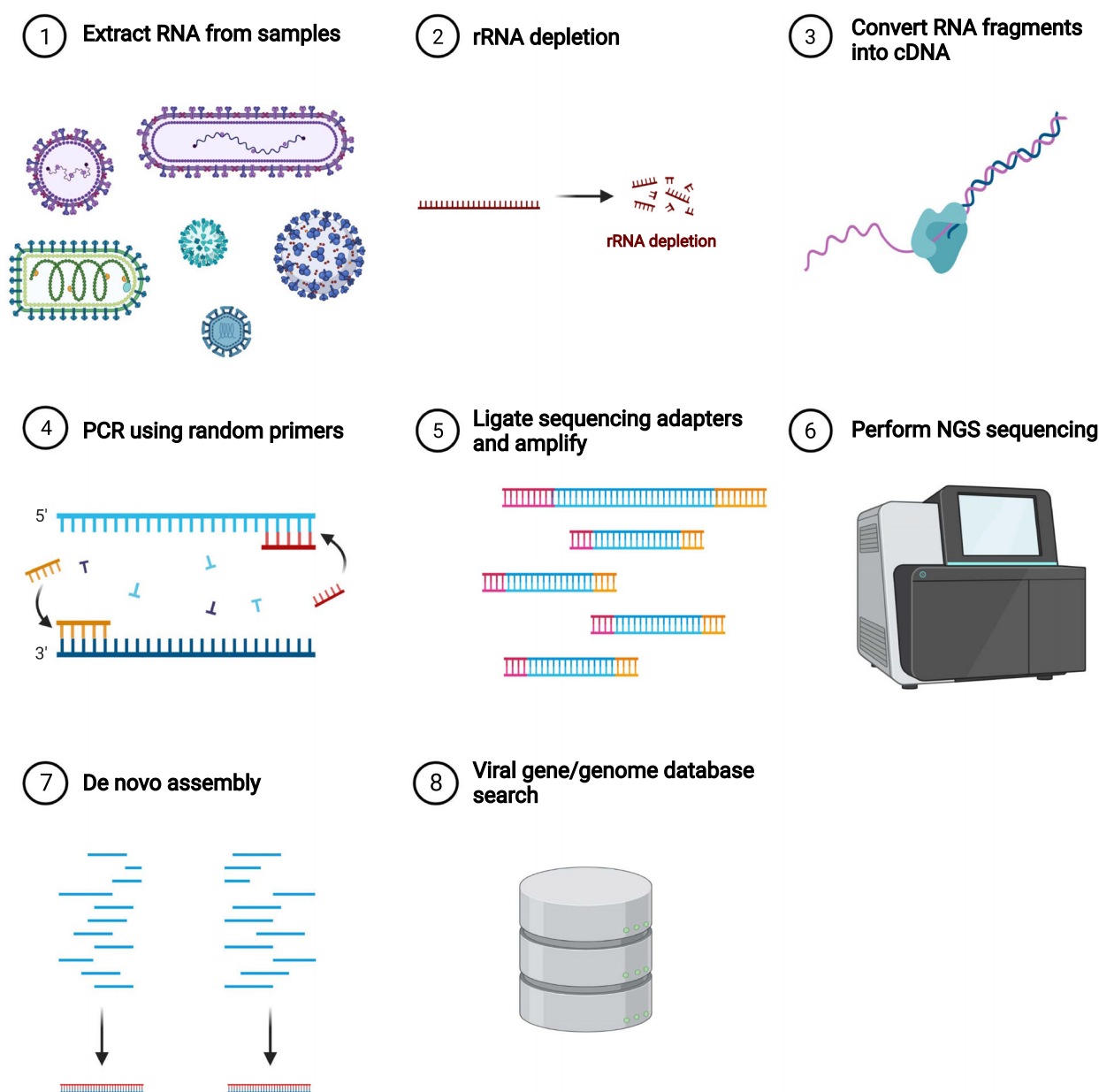


Fig. 1. Schematic procedure of a standard metatranscriptome analysis. (1) RNA is extracted from a given sample without enrichment of viruses. (2) Ribosomal RNA (rRNA) is removed. (3) Complementary DNA (cDNA) from the remaining RNA is generated using reverse transcriptase. (4) cDNAs are enriched by PCR using random primers. (5) Adapters are attached for terminals and the cDNAs are amplified. (6) The enriched cDNAs are sequenced. (7) Contigs are constructed by assembling reads. (8) Each contig is searched with viral genome databases. The figure was adapted from 'RNA sequencing', by BioRender.com (2022). Retrieved from <https://app.biorender.com/biorender-templates>.

Aquatic environment RNA virome

Metatranscriptome analyses have also revealed RNA virus communities in aquatic environments. In one of the oldest studies of the aquatic RNA virome, Culley *et al.* [36] sampled at Jericho Pier on English Bay and the Strait of Georgia, British Columbia, Canada. A

metatranscriptome analysis was conducted with Sanger sequencing technology, and RNA viral contigs belonging to *Picornavirales*, *Tombusviridae*, and *Umbravirus* were obtained. By using 454 sequencing technology which was the first commercially available next-generation sequencer, Djikeng *et al.* obtained RNA aquatic samples

at Lake Needwood, a freshwater lake in Maryland, USA, and identified various RNA virus-derived assemblies [37]. Subsequently, various studies have revealed the diversity of RNA viruses in oceans. In a major study, Wolf *et al.* [38] obtained 101 aquatic samples from Yangshan Deep-Water Harbor and yielded 4593 nearly full-length RNA viral RdRps that formed 2192 clusters with 75% amino acid identity. Recently, Zayed *et al.* [11] analyzed 771 metatranscriptome datasets from 121 locations throughout the world's five oceans and obtained 44 779 contigs, sequences of which 6686 encoded complete or near-complete RdRp domains. These aquatic metatranscriptome studies indicated the rich diversity of RNA viruses in the world's oceans.

RNA virome using public databases

Transcriptome data of various species and environments have been stored in public nucleotide sequencing databases, such as SRA [39], ERA [40], and DRA [41]. Kawasaki *et al.* [42] investigated more than 46 000 public sequencing datasets of mammalian and avian species and identified 882 RNA virus infections of 22 RNA virus families in 695 sequencing datasets from 53 host species. Moreover, Edgar *et al.* [43] comprehensively analyzed more than 3.8 million datasets of RNA-seq, metagenome, metatranscriptome, and metavirome stored in the SRA database. 131 957 contigs were found to have < 90% identity with known RdRps that are uniquely found in nonretroviral RNA viruses, suggesting that the analysis expanded the RNA virus diversity about 10-fold. The viral assembly information is available at their website: <https://serratus.io>. Neri *et al.* [44] also analyzed 5150 metatranscriptomes and revealed 124 873 distinct clusters of RNA virus genome segments at 90% identity, including two new candidate phyla and 74 new classes (<https://riboviria.org>). Those studies clearly suggest that there must be a huge number of RNA viruses that have not been identified to date. Such RNA viruses could also be revealed by metatranscriptome analyses in the future, which will be beneficial for understanding the nature of viruses and their impacts on human health. Meanwhile, no standards exist for the various current RNA virome studies; the quality and definition of RNA viruses reported for each study vary widely. In addition, the definition of RNA taxonomy is frequently revised to accommodate research on large-scale RNA virus discoveries.

Development of computational methods

Metatranscriptome analyses to detect RNA viruses largely depend on RNA viral genome sequences stored

in databases, as the assembled contigs are usually compared with existing RNA viral genomes to determine whether they are derived from RNA viruses. Owing to the advancement of DNA sequencing technology, the number of RNA viral gene and genome sequences stored in public databases has been increasing. Using such RNA viral gene and genome information, various bioinformatics tools and databases have been developed.

Table 1 summarizes the major software and databases for RNA virus identification available to date (as of April 2022). One of the major approaches utilizes a hidden Markov model (HMM). This is because RNA viral genes are highly divergent even for orthologous genes containing the RdRp domain [45–47]. An HMM profile based on a multiple sequence alignment can identify distantly related sequences more efficiently than a pairwise-based search [48]. RNA virus HMM profiles were constructed and distributed by vFams [49], RVDB-prot [50], efam [51], and VirSorter2 [52]; in particular, RVDB-prot has been updated frequently and has high potential applicability. NeoRdRp [53], RdRp-scan [54], and Olendraite *et al.* [55] also provide HMM-based databases; however, they target only the RdRp domain as the hallmark of nonretroviral RNA viruses (i.e., *Orthornavirae*). Palmscan with PALMdb also targets RdRp domains utilizing position-specific scoring matrix approaches [56].

Sequence similarity search programs, such as BLAST [57,58], BOWTIE2 [59], BWA [60], KRAKEN 2 [61], and DIAMOND [62], have been commonly used such as VIROMATCH [63], VIRMINE [64], VIRUSSEEKER [65], VIRFIND [66], VIRKRAKEN (<https://github.com/Strong-Lab/VirKraken>), SC > VIROMESCAN [67], and TAR-VIR [68]. Recently developed programs have applied machine-learning algorithms using *k*-mer (length of *k* nucleotides) frequencies of viral genes/genomes. VIRFINDER [69] is one of the pioneering virus identification programs that utilizes machine-learning algorithms. Additionally, other machine-learning algorithms such as convolutional neural network (CNN; DeepVirFinder [70], CHEER [71], CNN_Virus [72]), recurrent neural network (RNN; RNN-VirSeeker [73]), and multilayer perceptron (MLP) neural network (VIBRANT [74]), and attention-based long short-term memory (LSTM) network (Virtifier [75]) were reported. Since machine-learning-based virus identification programs are not based exclusively on sequence similarity to known viruses, they are capable of identifying a wide variety of unknown viruses. Some of those machine-learning-based virus identification programs were benchmarked by Glickman *et al.* [76]. However, with the development of machine-learning fields such as protein structure prediction and language model, various new algorithmic

Table 1. Bioinformatics tools for RNA virus identification.

Tool	Target	Search method	References	Web
VFAMS	Virus	HMM	[49]	https://derisilab.ucsf.edu/software/vFam/
RVDB-PROT	Virus	HMM	[50]	https://rvdb-prot.pasteur.fr/
EFAM	Virus	HMM	[51]	https://datacommons.cyverse.org/browse/iplant/home/shared/Virus/Zayed_efam_2020.1
VIRSORTER2	Virus	HMM	[52]	https://github.com/jiarong/VirSorter2
NEORDRP	RdRp	HMM	[53]	https://github.com/shoichisakaguchi/NeoRdRp
RDRP-SCAN	RdRp	HMM	[54]	https://github.com/JustineCharon/RdRp-scan
(Olendraite et al. 2023)	RdRp	HMM	[55]	https://github.com/ingridole/ViralRdRp_pHMMs
PALMSCAN/ PALMDB	RdRp	Position-specific scoring matrices (PSSMs)	[56]	https://github.com/rcedgar/palmscan
VIROMATCH	Virus	BWA-MEM, DIAMOND	[63]	https://twylie.github.io/viromatch/
VIRMI	Virus	BLAST	[64]	https://github.com/putonti/virMine
VIRUSSEEKER	Virus	BLAST	[65]	https://github.com/guoyanzhao/VirusSeeker-Virome
VIRFINDER	Virus	BLAST	[66]	http://virfinder.org/
VIRKRAKEN	Virus	Kraken2	–	https://github.com/Strong-Lab/VirKraken
VIROMESCAN	Virus	Bowtie2	[67]	https://sourceforge.net/projects/viromescan/
TAR-VIR	RNA virus	Bowtie2, BWA	[68]	https://github.com/chjiao/TAR-VIR
VIRFINDER	Virus	k-mer based SVM	[69]	https://github.com/jessieren/VirFinder
DEEPIRIFINDER	Virus	k-mer based CNN	[70]	https://github.com/jessieren/DeepVirFinder
CHEER	RNA virus	Hierarchically organized CNN	[71]	https://github.com/KennthShang/CHEER
CNN_VIRUS	Virus	Multitask learning model based on CNN	[72]	https://github.com/MaHaoran627/CNN_Virus
RNN-VIRSEEKER	Virus	k-mer based RNN	[73]	https://github.com/crazyinter/RNN-VirSeeker
VIBRANT	Virus	HMM and multilayer perceptron (MLP) neural network	[74]	https://github.com/AnantharamanLab/VIBRANT/
VIRTIFFIER	Virus	Attention-based long short-term memory (LSTM) network	[75]	https://github.com/crazyinter/Seq2Vec

bioinformatics programs continue to be proposed for virus search as well (please see the ‘Conclusion and future studies’ section as well). Since the purpose of each bioinformatics tool could differ and progress in bioinformatics methods is rapid, it will be necessary to constantly follow the latest papers for appropriate RNA virome analyses.

Endogenous viral elements and metatranscriptome analyses

Advances in genome research have also revealed many RNA virus-derived sequences in eukaryotic genomes. For example, approximately 8–10% of the human genome corresponds to retroviral origins [77–79], called endogenous retroviruses (ERVs). Since retroviruses code an RNA-dependent DNA polymerase (i.e., reverse transcriptase) that can replicate DNA based on an RNA template, the retroviral RNA genome is transcribed to DNA and can be integrated into the host genome using integrase. By chance, some of these are integrated into germline cells, which then become part of the host genome [80]. Since RNA viruses other than retroviruses do not encode reverse transcriptases, the

portion of nonretroviral RNA viruses is quite limited; however, some exist in the host genome, which are commonly called endogenous viral elements (EVEs) [11–17]. Metatranscriptome analyses have directly or indirectly identified EVEs in host genomes. Edgar et al. [43] estimated that around 1% of RNA viral contigs obtained through their metatranscriptome analyses correspond to EVEs. Furthermore, the recent discovery of many new RNA viruses facilitates the finding of previously unidentified EVEs. Kawasaki et al. employed carbovirus and cultervirus sequences that are newly identified bornavirus species as queries and identified various bornavirus-derived EVEs in a variety of mammalian species [16]. This observation clearly indicates that there are many unidentified EVEs in host genomes. Indeed, Kojima et al. [81] applied a *k*-mer-based support vector machine (SVM) classifier using EVE sequences as a training dataset and identified possible EVEs with low similarity to known RNA viruses in the human genome. Such EVEs could also be informative in the search for novel RNA viruses in nature. On the contrary, Kryukov et al. and Irwin et al. reported various viral contaminations in host

genome data [15,17], which may be incorrectly detected as EVEs. Further analyses and development of sequencing technology may be needed to understand the evolution of EVEs in detail.

Conclusion and future studies

Metatranscriptome analyses have revealed a variety of RNA viruses in various hosts and environments. However, it has also been shown that the RNA viruses found to date represent only a small subset of all RNA viruses. One of the bottlenecks in RNA virus identification from metatranscriptome data is its reliance on similarity to known RNA virus sequences. As noted, machine-learning-based methods could be useful; however, it is also difficult to confirm whether those detected sequences are derived from RNA viruses or not. To overcome such problems, a sequencing method named 'fragmented and primer ligated dsRNA sequencing' (FLDS) has been developed [10,82–84]. Since long dsRNAs are rare in eukaryotic cells, they are mainly derived from the genomes of dsRNA viruses or replication intermediates of ssRNA viruses [85]. Utilizing the biochemical properties of RNA viruses, sequences obtained by FLDS are likely to be RNA viruses even if they are not necessarily highly similar to known RNA viruses. Moreover, FLDS determines entire dsRNA sequences, including both terminals, and therefore, multisegmented genomes of RNA viruses can be reconstructed based on their terminal sequence similarities [82]. In addition to the FLDS methods, the three-dimensional structure of proteins is known to be more conserved than the underlying amino acid sequence [86]. Since AlphaFold2 can predict the protein structure of a given amino acid sequence with practical quality [87], RNA viral genes could also be predicted by their estimated three-dimensional structures. Therefore, more RNA viruses could be identified from metatranscriptome data combining various technologies in the near future.

Acknowledgments

This work was supported by JSPS KAKENHI Grant-in-Aid for Scientific Research (C) 20K06775 (to SN and SS); Early-Career Scientists 20K15685 and 22K14999 (to SS); and Scientific Research on Innovative Areas 16H06429 (to SN), 16K21723 (to SN), 17H05823 (to SN), 19H04843 (to SN); by JST CREST (JPMJCR20H6, to SN); and by AMED under grant numbers JP21wm0325004 (to SN) and JP22wm0325004 (to SN).

Conflict of interest

The authors declare no conflict of interest.

Author contributions

SN, AO, KM, TE, YS, and TG designed the concept of the manuscript. SN wrote the manuscript. SN and SS made a table. SS made a figure. All authors edited and approved the manuscript.

Data accessibility

Not applicable.

References

- Hepojoki J, Hetzel U, Paraskevopoulou S, Drosten C, Harrach B, Zerbini M, Koonin EV, Krupovic M, Dolja V and Kuhn JH (2020) Create one new realm (Ribozyviria) including one new family (Kolmioviridae) including genus *Deltavirus* and seven new genera for a total of 15 species. <https://ictv.global/ictv/proposals/2020.012D.R.Ribozyviria.zip>.
- Lai MMC (2005) RNA replication without RNA-dependent RNA polymerase: surprises from Hepatitis Delta virus. *J Virol* **79**, 7951–7958.
- Horie M, Kobayashi Y, Honda T, Fujino K, Akasaka T, Kohl C, Wibbelt G, Mühldorfer K, Kurth A, Müller MA *et al.* (2016) An RNA-dependent RNA polymerase gene in bat genomes derived from an ancient negative-strand RNA virus. *Sci Rep* **6**, 25873.
- Koonin EV and Dolja VV (2014) Virus world as an evolutionary network of viruses and Capsidless selfish elements. *Microbiol Mol Biol Rev* **78**, 278–303.
- Li CX, Shi M, Tian JH, Lin XD, Kang YJ, Chen LJ, Qin XC, Xu J, Holmes EC and Zhang YZ (2015) Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. *Elife* **4**, e05378.
- Shi M, Lin XD, Tian JH, Chen LJ, Chen X, Li CX, Qin XC, Li J, Cao JP, Eden JS *et al.* (2016) Redefining the invertebrate RNA virosphere. *Nature* **540**, 539–543.
- Shi M, Lin XD, Chen X, Tian JH, Chen LJ, Li K, Wang W, Eden JS, Shen JJ, Liu L *et al.* (2018) The evolutionary history of vertebrate RNA viruses. *Nature* **556**, 197–202.
- He WT, Hou X, Zhao J, Sun J, He H, Si W, Wang J, Jiang Z, Yan Z, Xing G *et al.* (2022) Virome characterization of game animals in China reveals a spectrum of emerging pathogens. *Cell* **185**, 1117–1129.e8.
- Chen YM, Sadiq S, Tian JH, Chen X, Lin XD, Shen JJ, Chen H, Hao ZY, Wille M, Zhou ZC *et al.* (2022) RNA viromes from terrestrial sites across China expand

- environmental viral diversity. *Nat Microbiol* **7**, 1312–1323.
- 10 Urayama S, 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.
 - 11 Zayed AA, Wainaina JM, Dominguez-Huerta G, Pelletier E, Guo J, Mohssen M, Tian F, Pratama AA, Bolduc B, Zablocki O *et al.* (2022) Cryptic and abundant marine viruses at the evolutionary origins of Earth's RNA virome. *Science* **376**, 156–162.
 - 12 Taylor DJ and Bruenn J (2009) The evolution of novel fungal genes from non-retroviral RNA viruses. *BMC Biol* **7**, 88.
 - 13 Horie M, Honda T, Suzuki Y, Kobayashi Y, Daito T, Oshida T, Ikuta K, Jern P, Gojobori T, Coffin JM *et al.* (2010) Endogenous non-retroviral RNA virus elements in mammalian genomes. *Nature* **463**, 84–87.
 - 14 Katzourakis A and Gifford RJ (2010) Endogenous viral elements in animal genomes. *PLoS Genet* **6**, e1001191.
 - 15 Kobayashi Y, Horie M, Nakano A, Murata K, Itou T and Suzuki Y (2016) Exaptation of Bornavirus-like nucleoprotein elements in Afrotherians. *PLoS Pathog* **12**, e1005785–21.
 - 16 Kryukov K, Ueda MT, Imanish T and Nakagawa S (2019) Systematic survey of non-retroviral virus-like elements in eukaryotic genomes. *Virus Res* **262**, 30–36.
 - 17 Kawasaki J, Kojima S, Mukai Y, Tomonaga K and Horie M (2021) 100-my history of bornavirus infections hidden in vertebrate genomes. *Proc National Acad Sci USA* **118**, e2026235118.
 - 18 Irwin NAT, Pittis AA, Richards TA and Keeling PJ (2021) Systematic evaluation of horizontal gene transfer between eukaryotes and viruses. *Nat Microbiol* **7**, 327–336.
 - 19 Boldogkői Z, Moldován N, Balázs Z, Snyder M and Tombácz D (2019) Long-read sequencing – a powerful tool in viral transcriptome research. *Trends Microbiol* **27**, 578–592.
 - 20 Orba Y, Matsuno K, Nakao R, Kryukov K, Saito Y, Kawamori F, Loza Vega A, Watanabe T, Maemura T, Sasaki M *et al.* (2021) Diverse mosquito-specific flaviviruses in the Bolivian Amazon basin. *J Gen Virol* **102**, 001518.
 - 21 Wang ZD, Wang B, Wei F, Han SZ, Zhang L, Yang ZT *et al.* (2019) A new segmented virus associated with human febrile illness in China. *N Engl J Med* **380**, 2116–2125.
 - 22 Kodama F, Yamaguchi H, Park E, Tatemoto K, Sashika M, Nakao R, Terauchi Y, Mizuma K, Orba Y, Kariwa H *et al.* (2021) A novel nairovirus associated with acute febrile illness in Hokkaido, Japan. *Nat Commun* **12**, 5539.
 - 23 Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, Wang W, Song H, Huang B, Zhu N *et al.* (2020) Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* **395**, 565–574.
 - 24 Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG *et al.* (2020) A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269.
 - 25 Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, Zhao X, Huang B, Shi W, Lu R *et al.* (2020) A novel coronavirus from patients with pneumonia in China, 2019. *N Engl J Med* **382**, 727–733.
 - 26 Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, Si HR, Zhu Y, Li B, Huang CL *et al.* (2020) A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273.
 - 27 Liu P, Chen W and Chen JP (2019) Viral metagenomics revealed Sendai virus and coronavirus infection of Malayan pangolins (*Manis javanica*). *Viruses* **11**, 979–915.
 - 28 Lam TTY, Jia N, Zhang YW, Shum MHH, Jiang JF, Zhu HC *et al.* (2020) Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. *Nature* **583**, 282–285.
 - 29 Xiao K, Zhai J, Feng Y, Zhou N, Zhang X, Zou JJ, Li N, Guo Y, Li X, Shen X *et al.* (2020) Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins. *Nature* **583**, 286–289.
 - 30 Zhang T, Wu Q and Zhang Z (2020) Probable pangolin origin of SARS-CoV-2 associated with the COVID-19 outbreak. *Curr Biol* **30**, 1578.
 - 31 Zhou H, Chen X, Hu T, Li J, Song H, Liu Y, Wang P, Liu D, Yang J, Holmes EC *et al.* (2020) A novel bat coronavirus closely related to SARS-CoV-2 contains natural insertions at the S1/S2 cleavage site of the spike protein. *Curr Biol* **30**, 2196–2203.e3.
 - 32 Delaune D, Hul V, Karlsson EA, Hassanin A, Ou TP, Baidaliuk A, Gámbaro F, Prot M, Tu VT, Chea S *et al.* (2021) A novel SARS-CoV-2 related coronavirus in bats from Cambodia. *Nat Commun* **12**, 6563.
 - 33 Wacharapluesadee S, Tan CW, Maneeorn P, Duengkae P, Zhu F, Joyjinda Y, Kaewpom T, Chia WN, Ampoot W, Lim BL *et al.* (2021) Evidence for SARS-CoV-2 related coronaviruses circulating in bats and pangolins in Southeast Asia. *Nat Commun* **12**, 972.
 - 34 Temmam S, Vongphayloth K, Baquero E, Munier S, Bonomi M, Regnault B, Douangboubpha B, Karami Y, Chrétien D, Sanamxay D *et al.* (2022) Bat coronaviruses related to SARS-CoV-2 and infectious for human cells. *Nature* **604**, 330–336.
 - 35 Petrone ME, Parry R, Mifsud JCO, Brussel KV, Vorhees I, Richards ZT *et al.* (2023) Evidence for an aquatic origin of influenza virus and the order *Articulavirales*. *bioRxiv* doi: [10.1101/2023.02.15.528772](https://doi.org/10.1101/2023.02.15.528772) [PREPRINT]
 - 36 Culley AI, Lang AS and Suttle CA (2006) Metagenomic analysis of coastal RNA virus communities. *Science* **312**, 1795–1798.

- 37 Djikeng A, Kuzmickas R, Anderson NG and Spiro DJ (2009) Metagenomic analysis of RNA viruses in a fresh water Lake. *PLoS One* **4**, e7264.
- 38 Wolf YI, Silas S, Wang Y, Wu S, Bocek M, Kazlauskas D, Krupovic M, Fire A, Dolja VV and Koonin EV (2020) Doubling of the known set of RNA viruses by metagenomic analysis of an aquatic virome. *Nat Microbiol* **5**, 1262–1270.
- 39 Katz K, Shutov O, Lapoint R, Kimelman M, Brister JR and O'Sullivan C (2022) The sequence read archive: a decade more of explosive growth. *Nucleic Acids Res* **50**, D387–D390.
- 40 Cummins C, Ahamed A, Aslam R, Burgin J, Devraj R, Edbali O, Gupta D, Harrison PW, Haseeb M, Holt S *et al.* (2022) The European nucleotide archive in 2021. *Nucleic Acids Res* **50**, D106–D110.
- 41 Okido T, Kodama Y, Mashima J, Kosuge T, Fujisawa T and Ogasawara O (2022) DNA data Bank of Japan (DDBJ) update report 2021. *Nucleic Acids Res* **50**, D102–D105.
- 42 Kawasaki J, Kojima S, Tomonaga K and Horie M (2021) Hidden viral sequences in public sequencing data and warning for future emerging diseases. *MBio* **12**, e01638-21.
- 43 Edgar RC, Taylor J, Lin V, Altman T, Barbera P, Meleshko D, Lohr D, Novakovsky G, Buchfink B, Al-Shayeb B *et al.* (2022) Petabase-scale sequence alignment catalyses viral discovery. *Nature* **602**, 142–147.
- 44 Neri U, Wolf YI, Roux S, Camargo AP, Lee B, Kazlauskas D, Chen IM, Ivanova N, Zeigler Allen L, Paez-Espino D *et al.* (2022) Expansion of the global RNA virome reveals diverse clades of bacteriophages. *Cell* **185**, 4023–4037.e18.
- 45 Wolf YI, Kazlauskas D, Iranzo J, Lucía-Sanz A, Kuhn JH, Krupovic M, Dolja VV and Koonin EV (2018) Origins and evolution of the global RNA Virome. *mBio* **9**, e02329-18.
- 46 Holmes EC and Sebastián D (2019) Can sequence phylogenies safely infer the origin of the global Virome? *mBio* **10**, e00289-19.
- 47 Wolf YI, Kazlauskas D, Iranzo J, Lucía-Sanz A, Kuhn JH, Krupovic M, Dolja VV and Koonin EV (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**, e00542-19.
- 48 Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ *et al.* (2021) Pfam: the protein families database in 2021. *Nucleic Acids Res* **49**, D412–D419.
- 49 Skewes-Cox P, Sharpton TJ, Pollard KS and DeRisi JL (2014) Profile hidden Markov models for the detection of viruses within metagenomic sequence data. *PLoS One* **9**, e105067.
- 50 Bigot T, Temmam S, Pérot P and Eloit M (2019) RVDB-prot, a reference viral protein database and its HMM profiles. *F1000research* **8**, 530.
- 51 Zayed AA, Lücking D, Mohssen M, Cronin D, Bolduc B, Gregory AC, Hargreaves KR, Piehowski PD, White Iii RA, Huang EL *et al.* (2021) Efam: an expanded, metaproteome-supported HMM profile database of viral protein families. *Bioinformatics* **37**, 4202–4208.
- 52 Guo J, Bolduc B, Zayed AA, Varsani A, Dominguez-Huerta G, Delmont TO, Pratama AA, Gazitúa MC, Vik D, Sullivan MB *et al.* (2021) VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* **9**, 37.
- 53 Sakaguchi S, Urayama S, Takaki Y, Hirotsuna K, Wu H, Suzuki Y, Nunoura T, Nakano T and Nakagawa S (2022) NeoRdRp: a comprehensive dataset for identifying RNA-dependent RNA polymerases of various RNA viruses from Metatranscriptomic data. *Microbes Environ* **37**, ME22001.
- 54 Charon J, Buchmann JP, Sadiq S and Holmes EC (2022) RdRp-scan: a bioinformatic resource to identify and annotate divergent RNA viruses in metagenomic sequence data. *Virus Evol* **8**, veac082.
- 55 Olendraite I, Brown K and Firth AE (2023) Identification of RNA virus-derived RdRp sequences in publicly available transcriptomic data sets. *Mol Biol Evol* **40**, msad060.
- 56 Babaian A and Edgar R (2022) Ribovirus classification by a polymerase barcode sequence. *PeerJ* **10**, e14055.
- 57 Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* **215**, 403–410.
- 58 Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K and Madden TL (2009) BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421.
- 59 Langmead B and Salzberg SL (2012) Fast gapped-read alignment with bowtie 2. *Nat Methods* **9**, 357–359.
- 60 Li H and Durbin R (2009) Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics* **25**, 1754–1760.
- 61 Wood DE, Lu J and Langmead B (2019) Improved metagenomic analysis with kraken 2. *Genome Biol* **20**, 257.
- 62 Buchfink B, Xie C and Huson DH (2015) Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**, 59–60.
- 63 Wylie TN and Wylie KM (2021) ViroMatch: a computational pipeline for the detection of viral sequences from complex metagenomic data. *Microbiol Resour Announc* **10**, e01468-20.
- 64 Garretto A, Hatzopoulos T and Putonti C (2019) virMine: automated detection of viral sequences from complex metagenomic samples. *PeerJ* **7**, e6695-21.

- 65 Zhao G, Wu G, Lim ES, Droit L, Krishnamurthy S, Barouch DH, Virgin HW and Wang D (2017) VirusSeeker, a computational pipeline for virus discovery and virome composition analysis. *Virology* **503**, 21–30.
- 66 Ho T and Tzanetakis IE (2014) Development of a virus detection and discovery pipeline using next generation sequencing. *Virology* **471**, 54–60.
- 67 Rampelli S, Soverini M, Turrone S, Quercia S, Biagi E, Brigidi P and Candela M (2016) ViromeScan: a new tool for metagenomic viral community profiling. *BMC Genomics* **17**, 1–9.
- 68 Chen J, Huang J and Sun Y (2019) TAR-VIR: a pipeline for TARgeted VIRal strain reconstruction from metagenomic data. *BMC Bioinformatics* **20**, 305.
- 69 Ren J, Ahlgren NA, Lu YY, Fuhrman JA and Sun F (2017) VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* **5**, 1–20.
- 70 Ren J, Song K, Deng C, Ahlgren NA, Fuhrman JA, Li Y, Xie X, Poplin R and Sun F (2020) Identifying viruses from metagenomic data using deep learning. *Quant Biol* **8**, 64–77.
- 71 Shang J and Sun Y (2020) CHEER: Hierarchical taxonomic classification for viral metagenomic data via deep learning. *Methods* **189**, 95–103.
- 72 Ma H, Tan TW and Ban KHK (2021) A multi-task CNN learning model for taxonomic assignment of human viruses. *BMC Bioinformatics* **22**, 194.
- 73 Liu F, Miao Y, Liu Y and Hou T (2020) RNN-VirusSeeker: a deep learning method for identification of short viral sequences from metagenomes. *IEEE/ACM Trans Comput Biol Bioinform* **19**, 1840–1849.
- 74 Kieft K, Zhou Z and Anantharaman K (2020) VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* **8**, 90.
- 75 Miao Y, Liu F, Hou T and Liu Y (2021) Virifier: a deep learning-based identifier for viral sequences from metagenomes. *Bioinformatics* **38**, 1216–1222.
- 76 Glickman C, Hendrix J and Strong M (2021) Simulation study and comparative evaluation of viral contiguous sequence identification tools. *BMC Bioinformatics* **22**, 329.
- 77 International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.
- 78 de Koning APJ, Gu W, Castoe TA, Batzer MA and Pollock DD (2011) Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet* **7**, e1002384–12.
- 79 Nurk S, Koren S, Rhie A, Rautiainen M, Bizikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A *et al.* (2022) The complete sequence of a human genome. *Science* **376**, 44–53.
- 80 Johnson WE (2014) Endogenous retroviruses in the genomics era. *Ann Rev Virol* **2**, 1–25.
- 81 Kojima S, Yoshikawa K, Ito J, Nakagawa S, Parrish NF, Horie M, Kawano S and Tomonaga K (2021) Virus-like insertions with sequence signatures similar to those of endogenous nonretroviral RNA viruses in the human genome. *Proc National Acad Sci USA* **118**, e2010758118.
- 82 Urayama S, Takaki Y and Nunoura T (2016) FLDS: a comprehensive dsRNA sequencing method for intracellular RNA virus surveillance. *Microbes Environ* **31**, 33–40.
- 83 Hirai M, Takaki Y, Kondo F, Horie M, Urayama S and Nunoura T (2021) RNA viral metagenome analysis of subnanogram dsRNA using fragmented and primer ligated dsRNA sequencing (FLDS). *Microbes Environ* **36**, ME20152.
- 84 Izumi T, Morioka Y, Urayama S, Motooka D, Tamura T, Kawagishi T *et al.* (2021) DsRNA sequencing for RNA virus surveillance using human clinical samples. *Viruses* **13**, 1310.
- 85 Kumar M and Carmichael GG (1998) Antisense RNA: function and fate of duplex RNA in cells of higher eukaryotes. *Microbiol Mol Biol Rev* **62**, 1415–1434.
- 86 Illergård K, Ardell DH and Elovsson A (2009) Structure is three to ten times more conserved than sequence—a study of structural response in protein cores. *Proteins* **77**, 499–508.
- 87 Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Židek A, Potapenko A *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589.