A RNA Virus Reference Database (RVRD) to Enhance Virus Detection in Metagenomic Data

Shaohua Lei

Thesis submitted to the faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
In
Computer Science and Application

Liqing Zhang, Chair
Yang Cao
Xiaowei Wu

September 5th, 2016
Blacksburg, Virginia

Keywords: RNA virus, Database, Virus detection, Metagenomics, Cluster

A RNA Virus Reference Database (RVRD) to Enhance Virus Detection in Metagenomic Data

Shaohua Lei

ABSTRACT

With the great promise that metagenomics holds in exploring virome composition and discovering novel virus species, there is a pressing demand for comprehensive and up-to-date reference databases to enhance the downstream bioinformatics analysis. In this study, a RNA virus reference database (RVRD) was developed by manual and computational curation of RNA virus genomes downloaded from the three major virus sequence databases including NCBI, ViralZone, and ViPR. To reduce viral sequence redundancy caused by multiple identical or nearly identical sequences, sequences were first clustered and all sequences except one in a cluster that have more than 98% identity to one another were removed. Other identity cutoffs were also examined, and Hepatitis C virus genomes were studied in detail as an example. Using the 98% identity cutoff, sequences obtained from ViPR were combined with the unique RNA virus references from NCBI and ViralZone to generate the final RVRD. The resulting RVRD contained 23,085 sequences, nearly 5 times the size of NCBI RNA virus reference, and had a broad coverage of RNA virus families, with significant expansion on circular ssRNA virus and pathogenic virus families. Compared to NCBI RNA virus reference in performance evaluation, using RVRD as reference database identified more RNA virus species in RNAseq data derived from wastewater samples. Moreover, using RVRD as reference database also led to the discovery of porcine

rotavirus as the etiology of unexplained diarrhea observed in pigs. RVRD is publicly available for enhancing RNA virus metagenomics.

A RNA Virus Reference Database (RVRD) to Enhance Virus Detection in Metagenomic Data

Shaohua Lei

GENERAL AUDIENCE ABSTRACT

Next-generation sequencing technology has demonstrated capability for the detection of viruses in various samples, but one challenge in bioinformatics analysis is the lack of well-curated reference databases, especially for RNA viruses. In this study, a RNA virus reference database (RVRD) was developed by manual and computational curation from the three commonly used resources: NCBI, ViralZone, and ViPR. While RVRD was managed to be comprehensive with broad coverage of RNA virus families, clustering was performed to reduce redundant sequences. The performance of RVRD was compared with NCBI RNA virus reference database using the pipeline FastViromeExplorer developed by our lab recently, the results showed that more RNA viruses were identified in several metagenomic datasets using RVRD, indicating improved performance in practice.

Dedicated to my parents and sister for all their unwavering love and support.

ACKNOWLEDGEMENTS

First of all, I owe unending honor and gratitude to my advisor, Dr. Liqing Zhang. Back to the summer in 2016, I decided to pursue this simultaneous Master's degree while working on my Ph.D. degree in Biomedical Science, which is a huge challenge for a graduate student, Dr. Zhang took a leap of faith and welcomed me into the lab in our first meeting. She has been an amazing advisor with her terrific mentoring, intelligence, dedication, and kindness, I truly appreciate her guidance and encouragement to keep me on track and to overcome the challenges. I would also like to thank Dr. Yang Cao and Dr. Xiaowei Wu for serving on my advisory committee, their time, comments, and suggestions are invaluable.

I am very grateful to have the group members Saima Sultana Tithi, Mohammad Shabbir Hasan, Gustavo Arango-Argoty, Dhoha Abid, Min Oh, Hong Tran, Jacob Porter, and Qinglai Bian. Your hardwork and productivity have motivated me so much, and your patience to answer my naïve coding and algorithm questions has carried me over those hard times. Thank you all for your assistance, advice, and friendship.

I also appreciate the Department of Computer Science at Virginia Tech for everything I have gained during my Master's study, and our graduate coordinator Sharon Kinder-Potter for helping me with paperwork promptly.

Last but not the least, I would like to thank my family from the bottom of my heart for their understanding, love, and support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1 Introduction and literature review

As the most abundant life entities and a primary reservoir of genetic diversity on Earth, viruses play important roles in shaping ecosystem dynamics and biogeochemical cycles of all habitats (1). Exploration of the virome in oceans, environmental niches, animals, and humans has revealed considerable genetic richness and complex involvement of viruses in host health and disease (1-3). Accumulating evidence has illuminated that the intestinal virome could be associated with the development of a large variety of human diseases, including type 1 diabetes, colorectal cancer, inflammatory bowel disease and Crohn's disease (4-6), presumably resulting from their immunomodulatory effects together with other components of the gut microbiome such as bacteria (3). Therefore, a broad view of virus genetic diversity and biogeographical distribution is in great demand for dissecting the roles that viruses play. Moreover, given the tremendous disease and economic burden associated with viral pathogens, it is critical to timely identify and annotate previously known viruses in environmental and clinical samples, as well as surveil evolving viruses and predict potential emerging viral outbreaks (7).

Recent advances in viral metagenomics have promoted the systematic and unbiased virus identification and quantification, and viral metagenomics has largely outperformed traditional techniques that are time-consuming and often inapplicable as they require virus isolation and culturing (8). However, unlike the ease of metagenomic data generation, there are ongoing challenges of upstream viral sequence enrichment and downstream bioinformatics analysis of a large amount of data (9). Within the enormous metagenomes generated from untargeted next-generation sequencing (NGS) technologies, short nucleotide sequences derived from known

viruses are typically far less than those of hosts or other microorganisms, limiting the efficiency and accuracy of data analysis after sequencing. Thus, prior to viral sample sequencing, virus particle enrichment and/or virus sequence amplification are widely performed to target on viral genomes. For instance, as viruses are the smallest life entities on Earth, size filtration and ultracentrifugation are two effective strategies to increase virus yield. However, one caveat is that such strategies might bias the composition of virus populations (9,10). An alternative option involves PCR amplification targeting specific viral species in a sample, also known as deep sequencing, which has enabled efficient recovery of virus genomes of interest (11,12).

Due to the lack of universal genomic marker genes and the limited reference databases, the downstream computational identification and annotation of virus species from metagenomic data are challenging (13). Currently, strategies for the development of integrated analytical pipelines could be generally classified into assembly-based and assembly-independent. The assembly-based approaches require the users to first assemble the metagenomic short reads into longer contigs using a sequence assembler, and then the assembled contigs are aligned to reference database(s) for taxonomic analysis. Pipelines in this category include VIROME (14), Metavir2 (15), VirSorter (16), and VirFinder (17). Compared with short reads, contigs with increased sequence length tend be more informative in generating accurate taxonomic analysis, and assembly is also necessary to generate whole genome sequences for viral variation analysis (18). However, the disadvantage of this approach is that read assembly is not only computationally intensive but also prone to genomic chimeras that can compromise downstream annotation (19). In contrast, the assembly-independent pipelines such as ViromeScan (20), HoloVir (21), and FastViromeExplorer (22) directly align metagenomic short reads to a virus

genome reference database for viral annotation. This strategy not only expedites the virus annotation process but also enables virus quantification as the abundance of aligned reads per virus species is associated with virus titers. Comparatively, the disadvantage of the assembly-free approach is that short reads are not as informative as assembled contigs about the entire viral genomes. Regardless, the common step of viral sequence annotation involves the comparison of either contigs or short reads to a reference sequence database, and the availability and reliability of reference databases are critical for the accuracy of the viral metagenomic annotation (8,9).

Currently, GenBank serves as the largest publicly available resource for nucleotide sequence retrieval and deposit (23), and is maintained by the National Center for Biotechnology Information (NCBI) at the National Institutes of Health (NIH, Bethesda, MD). The Reference Sequence (RefSeq) collection at NCBI provides a manually curated set of nonredundant reference sequences (24), and virus RefSeq data set included in the NCBI Viral Genomes has been broadly used by researchers in the viral metagenomics community (25). Technically, each viral species or significant subspecies contributes one RefSeq reference, which is intended to be nonredundant full-length virus genome, and new sequences are being manually added into the virus RefSeq collection. Fortunately, a higher rate of adding new viruses into RefSeq was observed, from 0.34 sequence/day in 2010 to 2.5 sequence/day in 2015 (9), but the rate has fallen far behind that of novel virus discovery in recent years. For example, only seven hepatitis C virus (HCV) Refseq sequences are included in NCBI Viral Genomes, whereas HCV has been classified into 67 sub-genotypes with 129 distinct sequences and the numbers keep increasing (26). Furthermore, there are currently fewer than 10,000 virus genomes in Refseq, but the number of distinct virus species was estimated to be more than 3 million (27). Therefore, the widely used NCBI virus reference

database is an under-representation of all virus species, and there is a great demand to expand virus reference database in a timely manner to facilitate virus detection in metagenomic data.

According to the Baltimore virus classification system (Fig. 1), viruses can be broadly classified into DNA viruses, RNA viruses, and retroviruses. While the reference databases for DNA viruses and retroviruses have been expanded and integrated previously (28), the lack of a well-established RNA virus reference database (RVRD) is much more pronounced. Although resources for RNA virus sequences exist, they are generally maintained separately with an emphasis on

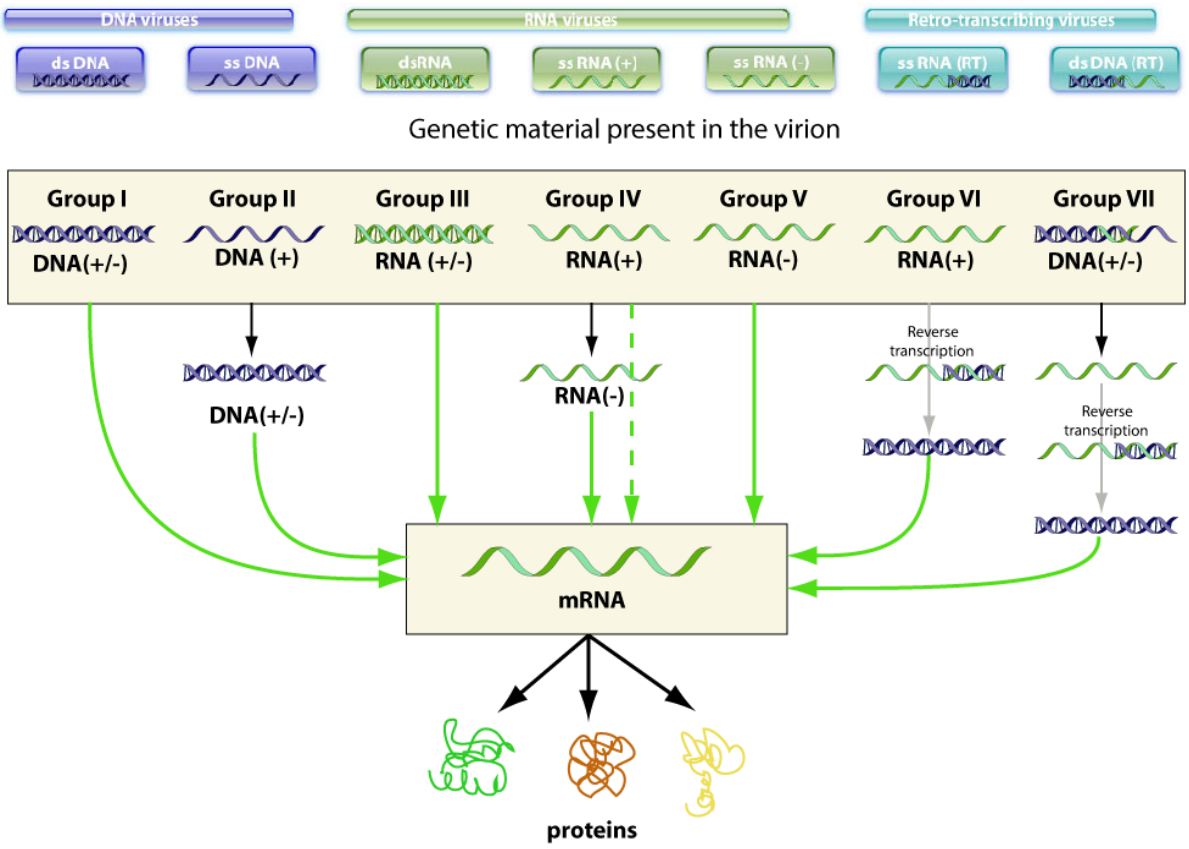

Figure 1. The Baltimore virus classification system. Viruses are clustered into families based on their type of genome, and RNA viruses have no DNA materials in the entire lifecycle. The figure was adapted from ViralZone (https://viralzone.expasy.org/254).

individual virus families, such as a specific pathogenic virus or family responsible for common infectious diseases, including HCV, hepatitis E virus, Dengue virus, influenza A virus, and West Nile virus (29), and some might have a high degree of sequence redundancy. Because of the lack of a well-established RVRD for RNA virus metagenomics, reads are often analyzed against a nonredundant viral protein database instead of a nucleotide RVRD (30-32). As shown in Fig. 1, RNA virus consists of double-stranded RNA virus (dsRNA virus), positive-sense single-stranded RNA virus (ssRNA(+) virus), and negative-sense single-stranded RNA virus (ssRNA(-) virus). Due to the lack of DNA material in their entire lifecycle, RNA virus could not be detected using sample DNA extraction and amplification, and thus RNA virus metagenomics requires extra efforts in library preparation, including elaborative sample RNA extraction and subsequent cDNA construction. Nevertheless, RNA virus metagenomics has gained increasing attention due to RNA virus diversity, the unique advantage of RNA sequencing, and the availability of a massive number of transcriptome data (33). In addition, most recent pandemic viral outbreaks were caused by emerging RNA viruses, such as Zika virus, Dengue virus, West Nile virus, Ebola virus, norovirus, Lassa virus, influenza A virus, SARS and MERS coronaviruses. Thus, more effort is required to create a comprehensive reference database for RNA viruses to explore the full potential of RNA virus metagenomics.

In this thesis project, aiming to develop a comprehensive and nonredundant RVRD, RNA virus genomes from the three best recognized resources: NCBI, ViralZone (34), and virus pathogen database and analysis resource (ViPR) (35) were pooled and further curated. The strategy to reduce sequence redundancy while retaining sufficient virus diversity was to cluster the sequences retrieved from ViPR using CD-HIT-EST (36) at 98% identity, and then the resulting

collection of sequences were combine with the unique RNA virus references from NCBI and ViralZone. Using the pipeline FastViromeExplorer together with RNAseq datasets derived from wastewater and animal tissue samples, the improved performance of RVRD in identifying RNA viral species was demonstrated as compared with NCBI RNA virus reference database. The developed RVRD and other resources are publicly available online at:

https://bench.cs.vt.edu/FastViromeExplorer/RNA_virus_database/

Chapter 2 Materials and methods

2.1 Retrieval of RNA virus sequences

Efforts to establish the RVRD were initiated by gathering RNA virus sequences from NCBI (February 2018), ViralZone (February 2018), and ViPR (March 2018). The RNA virus reference sequences in NCBI were retrieved from the NCBI Nucleotide database (https://www.ncbi.nlm.nih.gov/nucleotide/), using the species option “viruses” and source databases option “RefSeq”, and the search terms “ssRNA viruses”, “dsRNA viruses”, or “unclassified RNA viruses”. ViralZone has detailed viral taxonomy with entries for each family (https://viralzone.expasy.org/), but all the reference sequences are stored in other databases, and sequences for each RNA virus family had to be retrieved individually. By searching the virus family in ViralZone and then retrieving the corresponding sequences in NCBI, a total of 2410 RefSeq sequences were pooled manually from 13 dsRNA virus families, 40 ssRNA(+) virus families, 11 ssRNA(-) virus families, and two circular ssRNA virus families. ViPR is an integrated database focusing on virus pathogens with frequent updates on nucleotide sequences (https://www.viprbrc.org/). Complete genomes from six ssRNA(+) virus families, five ssRNA(-) virus families, and one dsRNA virus family were retrieved from ViPR. For each virus category mentioned above, genome sequences were stored separately in the first place for quality control analysis.

2.2 Nucleotide sequence clustering

To evaluate the quality and reduce the redundancy of pooled RNA virus genomes, nucleotide sequence clustering was performed using the tool CD-HIT-EST with multiple identity

cutoffs (36). The clustering algorithm in CD-HIT is a greedy incremental algorithm, it first sorts sequences in a decreasing length order, the longest one is the representative of the first sequence cluster, and then all remaining sequences are compared to the current representative(s). If the similarity with any representatives is greater than a given cutoff, the compared sequence is grouped into that cluster. Otherwise, it becomes the representative of a new cluster (Fig. 2). For each sequence comparison, CD-HIT applies k-mer filtering to speed up the clustering procedure. Briefly, an index table of k-mer is created for each sequence, if the index table similarity between a sequence and a representative is above the required value, a pairwise alignment is performed to confirm the sequence identity. Since k-mer can be easily indexed, k-mer filtering efficiently reduces the time-consuming pairwise alignments and thus the entire clustering (37). Although PSI-CD-HIT is optimized for clustering long sequences such as the whole genome of most species, program CD-HIT-EST was chosen in this study due to the relatively small sizes of RNA virus genomes, as well as its better performance in clustering at high sequence similarity. A series of identity cutoffs (100%, 98%, 95%, 92%, and 90%) were used to explore the quality of pooled sequences, and 99.9% identity cutoff was chosen to combine the NCBI + ViralZone and ViPR RNA virus sequences, because a lower cutoff could reduce the sequence diversity in NCBI + ViralZone (Supplementary Table 4). While all clustering was performed with eight CPUs, a k-mer size ten was used with identity cutoff ≥ 95% and a k-mer size eight with identities cutoffs 92% and 90%, respectively. As the default setting of CD-HIT-EST, the longest sequence of a cluster becomes the cluster representative.

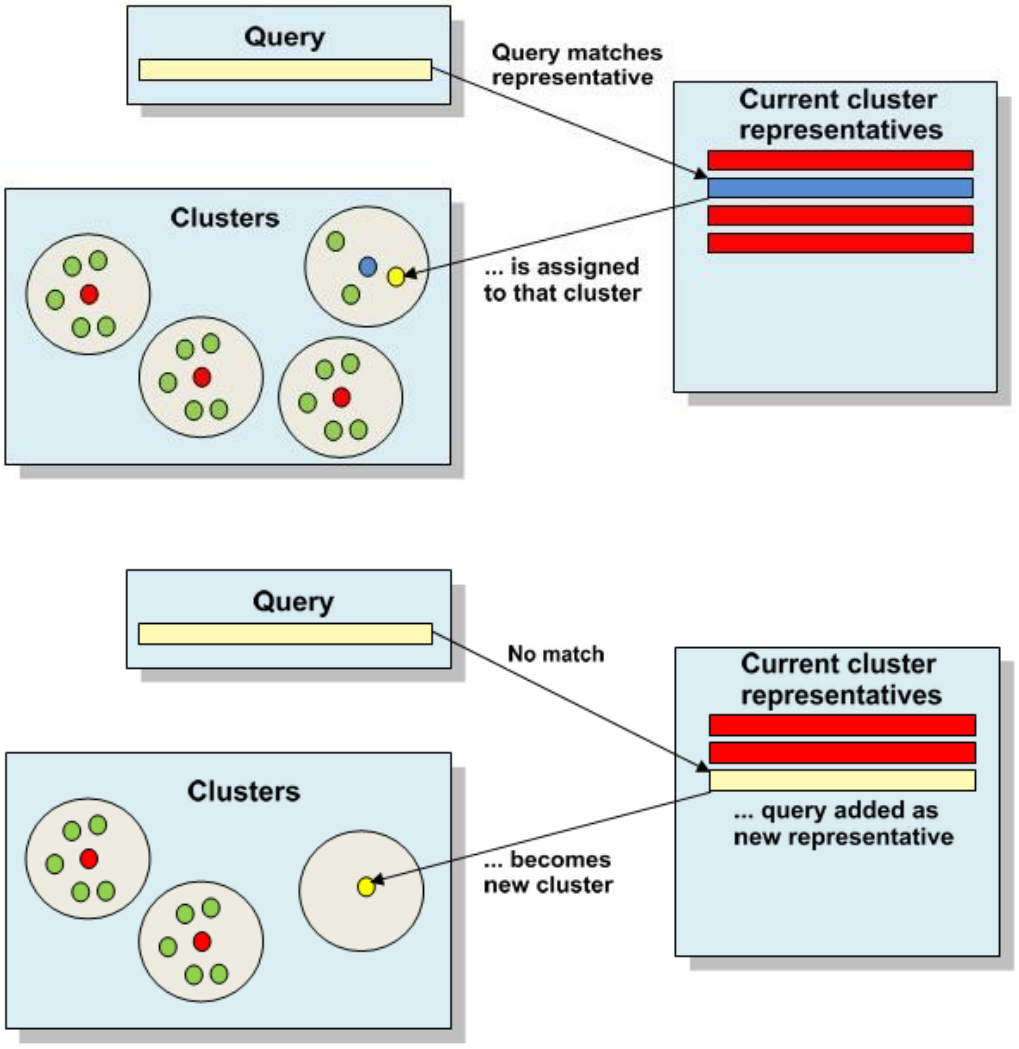

Figure 2. Schematic outline of nucleotide sequence clustering algorithm using CD-HIT. The figure was adapted from http://www.drive5.com/uclust/uclust_userguide_1_1_579.html

2.3 Distribution of virus genome sizes and families

RefSeq sequences of all viruses were obtained by searching “viruses” in NCBI Nucleotide with the options of “viruses” and “RefSeq” (March 2018). For both the developed RVRD and RefSeq of all viruses in NCBI, virus information containing the virus accession number and genome size was extracted from the corresponding fasta sequence files, using the bash script named *generateGenomeList.sh* provided in FastViromeExplorer (22). The frequency distribution

of genome sizes was analyzed and visualized in GraphPad Prism 6, and the last bin was arbitrarily set at 100,000 bps to gain an improved visualization of RNA virus genomes, which were less than 35,000 bps. The distribution of virus family in RVRD was analyzed by comparing and clustering sequences obtained from the three resources for each virus family.

2.4 Performance evaluation of RVRD

FastViromeExplorer was applied as the analytical pipeline to evaluate the performance of RVRD and compared with the NCBI RNA virus reference. FastViromeExplorer first uses kallisto for rapid alignment of input reads against the virus reference database, and then the alignment outputs are filtered using the ratio between the observed and expected genome coverage, which effectively filters a large amount of artifacts and thus improves the virus detection specificity (22). Before performing the evaluation with FastViromeExplorer, the kallisto index files for both databases were generated following the pipeline's instruction. For the datasets from the wastewater study, the RNAseq fastq files were retrieved from NCBI using SRA Toolkit fastq-dump. The RNAseq data from the germ-free pig's small intestinal tissue are available upon request.

Chapter 3 Results

3.1 RNA virus genomes curation and quality control.

In the last decade, RNA virus metagenomics has significantly advanced our understanding of virus diversity and provided a new tool for studying viruses in clinical and environmental settings (34). To enhance the capability of computational identification of RNA virus species in metagenomic data, a new RVRD was developed in this study by firstly pooling RNA virus genomes from the three best recognized resources, including NCBI, ViralZone, and ViPR. As shown in Table 1, by searching ssRNA viruses, dsRNA viruses, and unclassified RNA viruses in NCBI, a total of 2480, 1145, and 1104 sequences were retrieved, respectively (as of February 2018). ViralZone is a knowledge database with comprehensive viral taxonomy, and the reference sequences for each RNA virus family were downloaded and pooled (Table 1 and Supplementary Table 1), a total of 2034 sequences from 40 ssRNA(+) virus families and 11 ssRNA(-) virus families, 337 sequences from 13 dsRNA virus families, and 39 sequences from two circular ssRNA virus families were obtained (as of February 2018). Instead, there were only six ssRNA(+) virus families, five ssRNA(-) virus families, and one dsRNA virus family available in ViPR, but the corresponding numbers of complete genome sequences from each family were higher than those of NCBI and ViralZone (Table 1 and Supplementary Table 1).

To evaluate the quality of the RNA virus genomes, the sequence redundancy of each virus group in Table 1 was examined by performing CD-HIT-EST with 100% identity cutoff (36). Notably, the clustering of ssRNA viruses (altogether 4514, with 2480 from NCBI RefSeq and 2034 from

Table 1. RNA virus genomes curated from NCBI, ViralZone, and ViPR.

Virus group	Virus family	NCBI RefSeq	ViralZone	ViPR	CD-HIT-EST 100% identity
ssRNA viruses		2480	2034		2480
dsRNA viruses		1145	337		1145
unclassified RNA viruses		1104			1104
Circular ssRNA viruses	Avsunviroidae		4		4
	Pospiviroidae		35		35
ss(+)RNA viruses	Caliciviridae			1512	1442
	Coronaviridae			2519	2179
	Flaviviridae			12898	11513
	Hepeviridae			386	362
	Picornaviridae			4628	4330
	Togaviridae			1700	1545
ss(-)RNA viruses	Arenaviridae			1141	1030
	Bunyaviridae			6545	5547
	Filoviridae			594	396
	Paramyxoviridae			3226	2825
	Rhabdoviridae			1833	1609
dsRNA viruses	Reoviridae			51399	36065
Sum		4729	2410	88381	

ViralZone) resulted in 2480 sequences, and the clustering of dsRNA viruses (altogether 1482, with 1145 from NCBI RefSeq and 337 from ViralZone, Table 1) resulted in 1145 sequences. Further examination of the clustering results shows that sequences obtained from ViralZone were covered entirely by sequences in NCBI RefSeq and the number of viral genomes in each ssRNA virus and dsRNA virus family was the same between NCBI and ViralZone, except for the unclassified and environmental ones (Supplementary Table 1). However, sequence redundancy was observed in ViPR, as all virus families had reduced numbers of sequences after the clustering, and especially for Filoviridae and Reoviridae viral families, about 30% sequences were removed after clustering (Table 1). Therefore, additional sequence clustering was required to further reduce the degree of redundancy for RNA virus genome sequences obtained from ViPR.

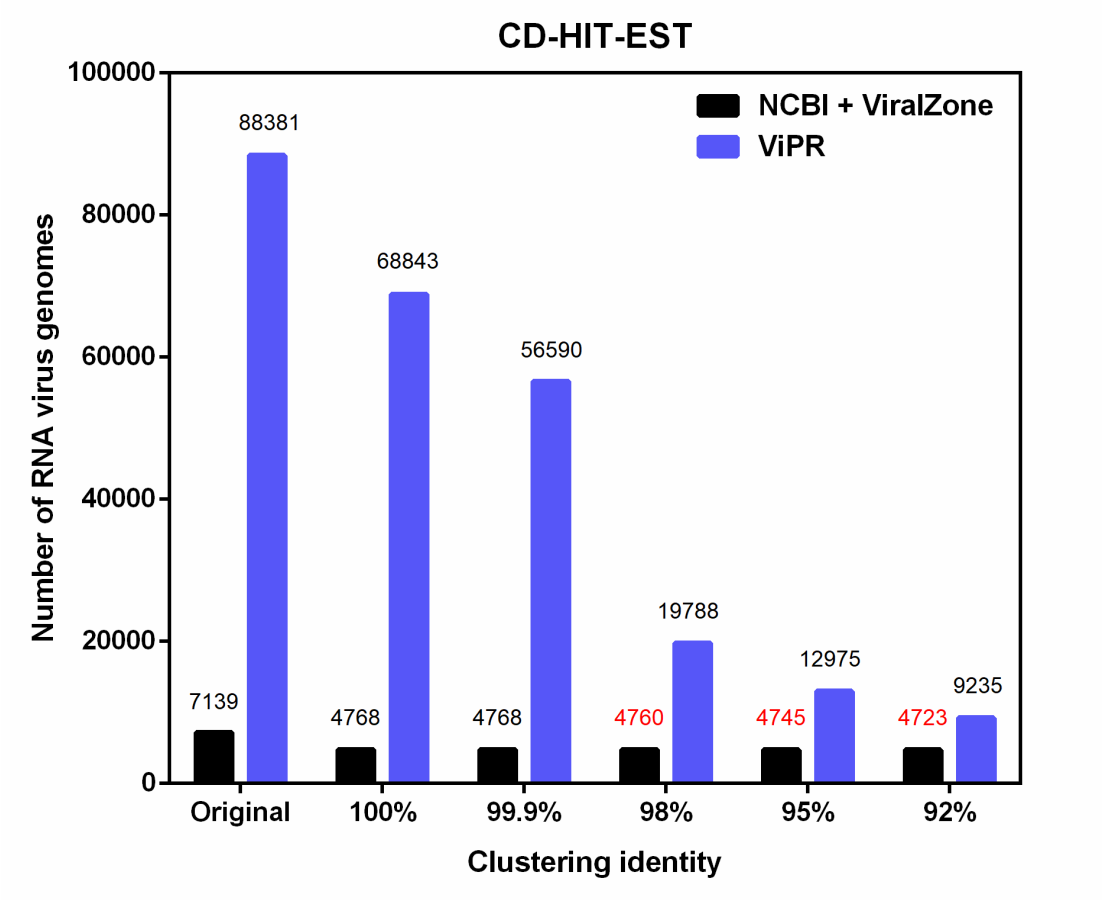

Figure 3. Clustering of the pooled RNA virus genome sequences. Sequences obtained from NCBI and ViralZone were combined to improve the analysis, and CD-HIT-EST with multiple identities was performed to evaluate the diversity and redundancy of pooled sequences.

3.2 Clustering of the pooled sequences.

Since the sequences obtained from both NCBI and ViralZone were RefSeq reference and sequence overlap was observed as discussed above, these sequences were combined and designated as NCBI + ViralZone. After applying CD-HIT-EST with 100% identity cutoff for clustering, the pooled NCBI + ViralZone had 4768 sequences, containing 4729 sequences from NCBI RefSeq and 39 sequences from the circular ssRNA virus families in ViralZone (Fig. 3). As expected, the number of sequences gets further reduced for NCBI + ViralZone with even lower clustering

identity cutoffs (e.g., from 98% to 92%), and the lower the identity cutoff, the higher the reduction. But overall the redundancy of the pooled sequences from NCBI + ViralZone is low. Comparatively, a much higher loss of sequences occurred in ViPR, for example, the total number of sequence was 68843 and 19788 when a clustering identity of 100% and 98% was applied, respectively (Fig. 3). Thus, to reduce the degree of redundancy while maintaining high virus sequence diversity, the optimal clustering identity and strategy need to be determined.

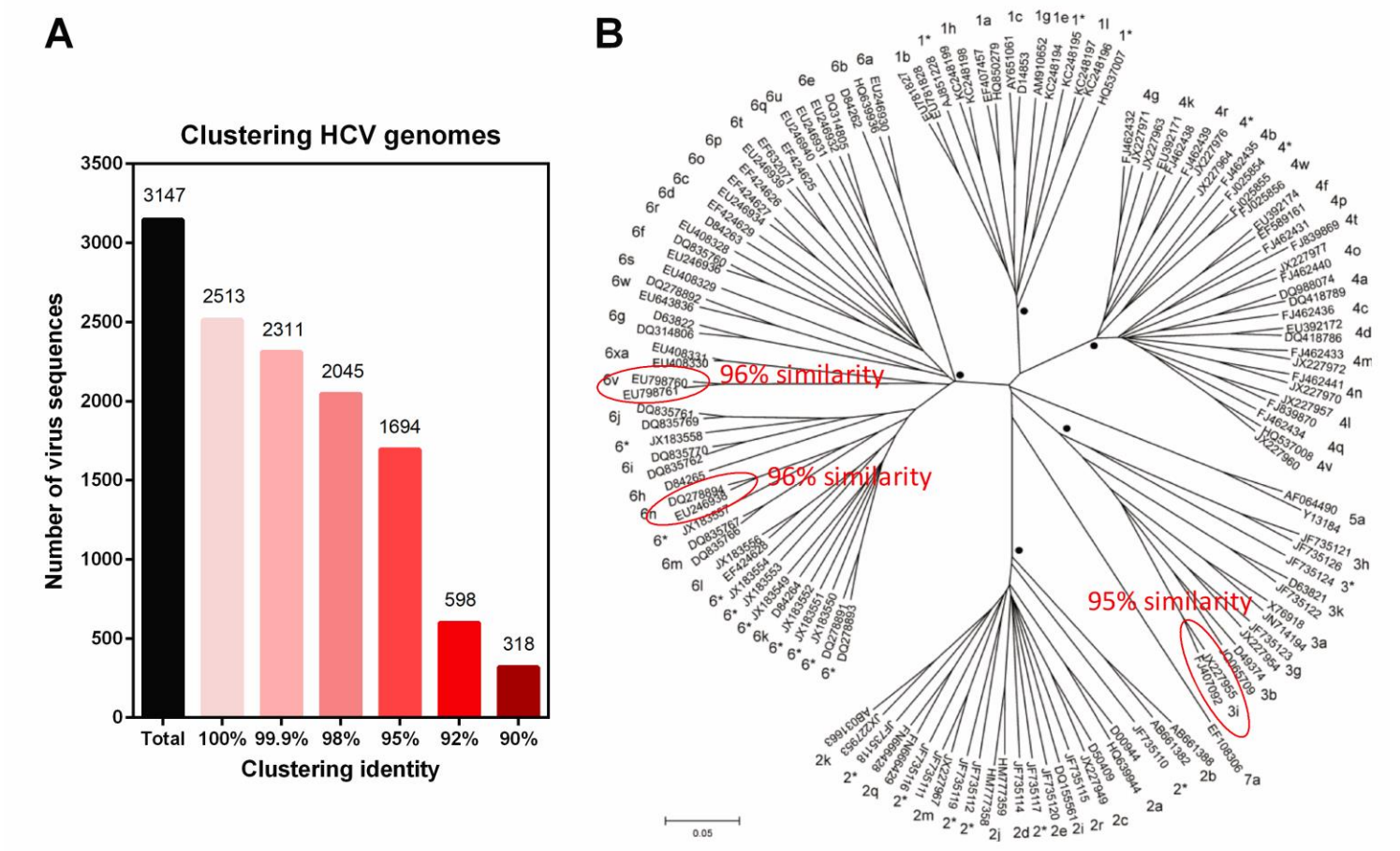

Figure 4. Optimization of cluster identity by analyzing HCV genomes. (**A**) Clustering HCV genomes obtained from ViPR using CD-HIT-EST at different identities. The number of sequences was indicated at the top of each bar (**B**) Phylogenetic tree of 129 distinct HCV genomes. The similarity of the two HCV genomes in red cycles was evaluated in BLAST, and the results were indicated around cycles in red.

To gain a more informative view about the influence of clustering identity on sequence redundancy and diversity, the complete HCV genomes were obtained from ViPR and examined in detail for the effect of different identity cutoffs on the number of sequence clusters. A significant reduction of sequences was observed from the identity cutoff 95% to 92% (Fig. 4A), most likely resulting in a dramatic loss of sequence diversity. Phylogenetic analysis indicated that there were 129 representative HCV sequences (26), and some have similarity to one another as high as 96% (Fig. 4B), which would be removed from the database if they were clustered at 95% identity. Taken together, the optimal identity for clustering HCV genomes was determined as 98%, which is also the cutoff used for virus database curation in a previous study (38). Thus, 98% identity cutoff was chosen for clustering RNA virus genomes retrieved from ViPR. In addition, NCBI virus RefSeq has been considered as the "gold standard" database due to its broad coverage

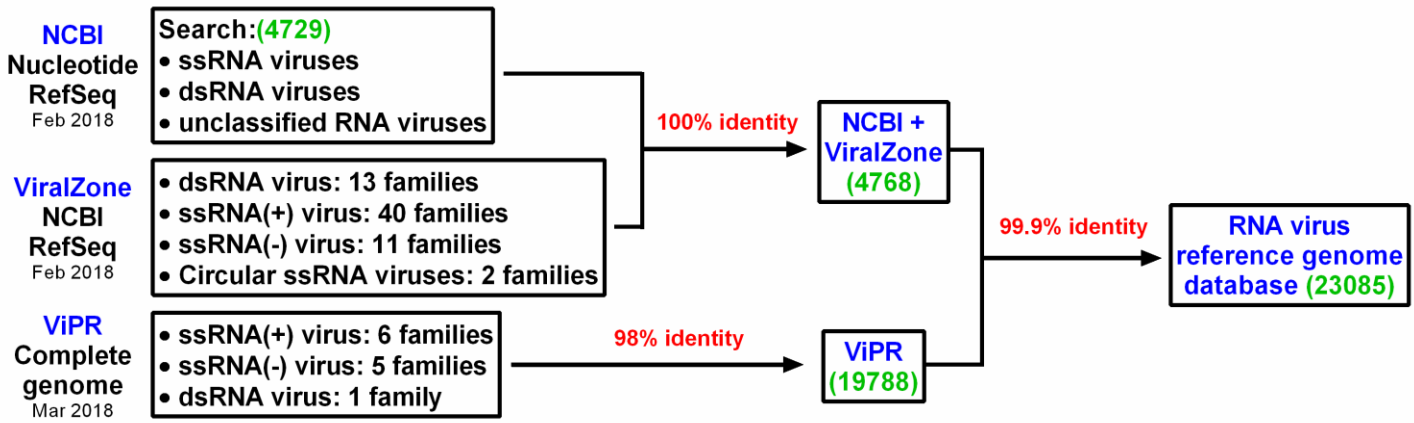

Figure 5. Schematic outline for the development of RVRD. RNA virus RefSeq sequences retrieved from NCBI and ViralZone were combined and clustered at 100% identity to remove the identical sequences, and RNA virus complete genomes from ViPR were clustered at 98% identity to reduce the redundancy. Then all sequences were clustered at 99.9% identity to avoid the sequence overlap in between and to maintain sequence diversity. Numbers in green indicate the total sequences in the corresponding groups.

on virus families, and sequences included in NCBI + ViralZone were all RefSeq in this study, so we managed to keep them in RVRD. The schematic strategy for the development of RVRD was outlined in Fig. 5, the NCBI + ViralZone and ViPR RNA virus sequences were clustered firstly at 100% and 98% identity, respectively, and then were combined together and clustered at 99.9% identity to filter the sequence overlap between them, as well as retaining the sequence diversity in NCBI + ViralZone. In all, RVRD had 23,085 sequences as compared with 4,729 sequences in the original NCBI RNA virus reference (Fig 4), and the increase was nearly 5-fold.

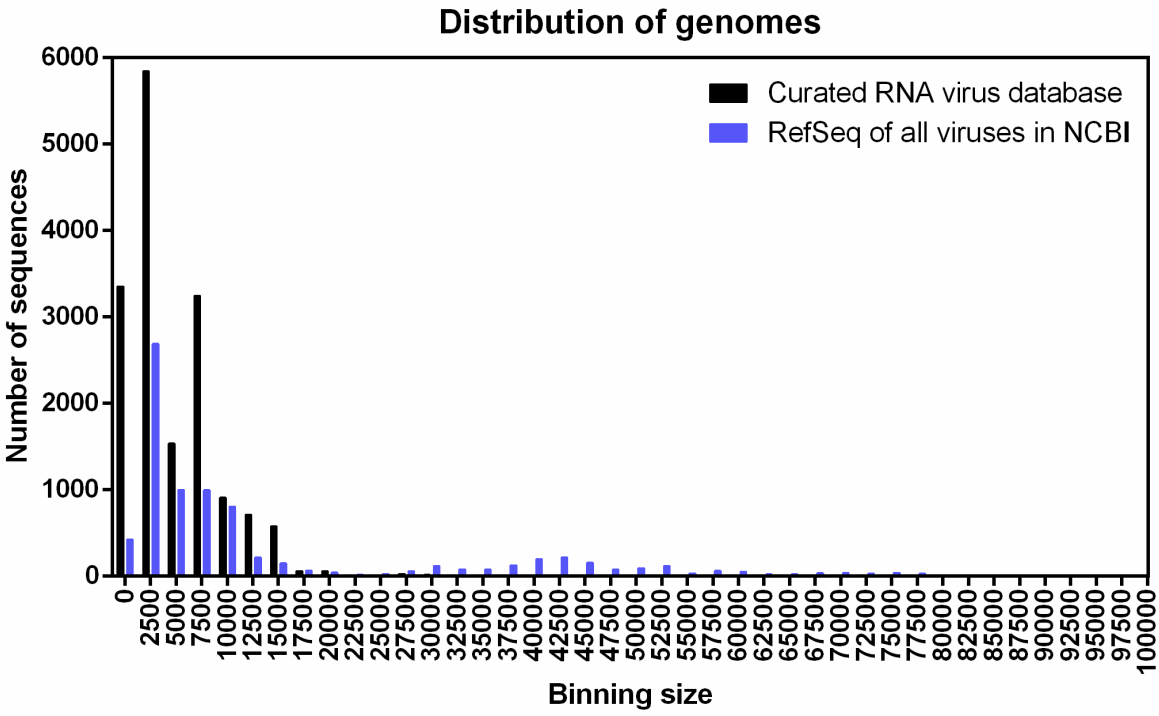

Figure 6. Distribution of virus genome sizes. Sequences in the curated RNA virus database RVRD and RefSeq of all viruses in NCBI were binned by size and plotted by number of sequences. Among each bin in the majority of 0 to 15,000 bp, there were more sequences from the RNA virus database.

3.3 Characterization of the developed RVRD.

The distribution of genomes on the basis of the sequence length/size was determined for the curated RVRD and for RefSeq of all viruses in NCBI to compare their differences (Fig. 6). The genome size of RNA viruses is generally smaller than that of other viruses such as DNA viruses and retroviruses, and the results indicated that the size of the largest genome was 33,452 bps in RVRD, as compared to 2,473,870 bps in RefSeq of all viruses. In addition, the majority of sequences in both were less than 17,500 bps, among which there were more sequences in RVRD for each binning size than that of RefSeq of all viruses, and RefSeq of all viruses had a considerable distribution peak at 42,500 bps (Fig. 6). The distribution of genomes on the basis of virus family was analyzed as well, and the detailed information was shown in Supplementary Table 1. Considering NCBI RNA virus reference as the framework of the developed RVRD, ViralZone only contributed 39 sequences as circular ssRNA viruses, and ViPR contributed sequences significantly from its 12 virus families consisting of pathogenic RNA viruses.

3.4 Performance evaluation of RVRD.

The NCBI virus reference is the most widely used public database in virus metagenomics, and thus the performance of RVRD was compared with the NCBI RNA virus reference. The two databases were used as custom databases in a recently developed pipeline FastViromeExplorer, which outperformed many other tools in virus detection and quantification in metagenomic data in terms of efficiency and accuracy (22), and several RNAseq datasets generated from environmental samples and animal tissues were analyzed regarding the performance of the two databases in the detection of RNA viruses.

Table 2. Performance evaluation of RVRD using RNAseq data of wastewater

Sample ID	Accession	Fastq file size	Number (abundance) of identified viruses	
			NCBI RNA virus RefSeq	RVRD
LI_13-9	SRX3606320	436 MB	14 (16582)	31 (17967)
LE_11-10	SRX3606321	938 MB	10 (33800)	13 (41763)

In a recent study determining the presence of viruses in a river, the virus particles in wastewater samples were concentrated first by precipitation and filtration to increase virus titers and reduce the noise of other organisms, and a diverse group of RNA viruses were identified in the shotgun RNAseq data against the NCBI virus protein database (30), and therefore such metagenomic datasets are suitable for the performance evaluation of RVRD. As shown in Table 2, for both RNAseq data sets (LI_13-9 and LE_11-10) analyzed in this study, higher numbers of RNA viruses were obtained using FastViromeExplorer with RVRD as the reference database than with the NCBI RNA virus RefSeq (31 versus 14, and 13 versus 10, respectively), and the higher numbers of virus outputs were supported by higher abundances of reads (17967 versus 16582 reads, and 41763 versus 33800 reads, respectively). In addition to the virus species identified using NCBI RNA virus RefSeq reference database, the output of using RVRD as the reference database contained additional RNA viruses, including picobirnavirus, porcine rotavirus C, and different human rotavirus strains and segments (see Supplementary Table 2 and 3). Taken together, RVRD enabled the identification of more RNA virus species and subspecies than NCBI RNA virus RefSeq in this environmental wastewater metagenomic data.

Viral metagenomics has been applied to determine the etiology of unexplained human and animal diseases such as diarrhea and pneumonia (39,40), and it is more rewarding when

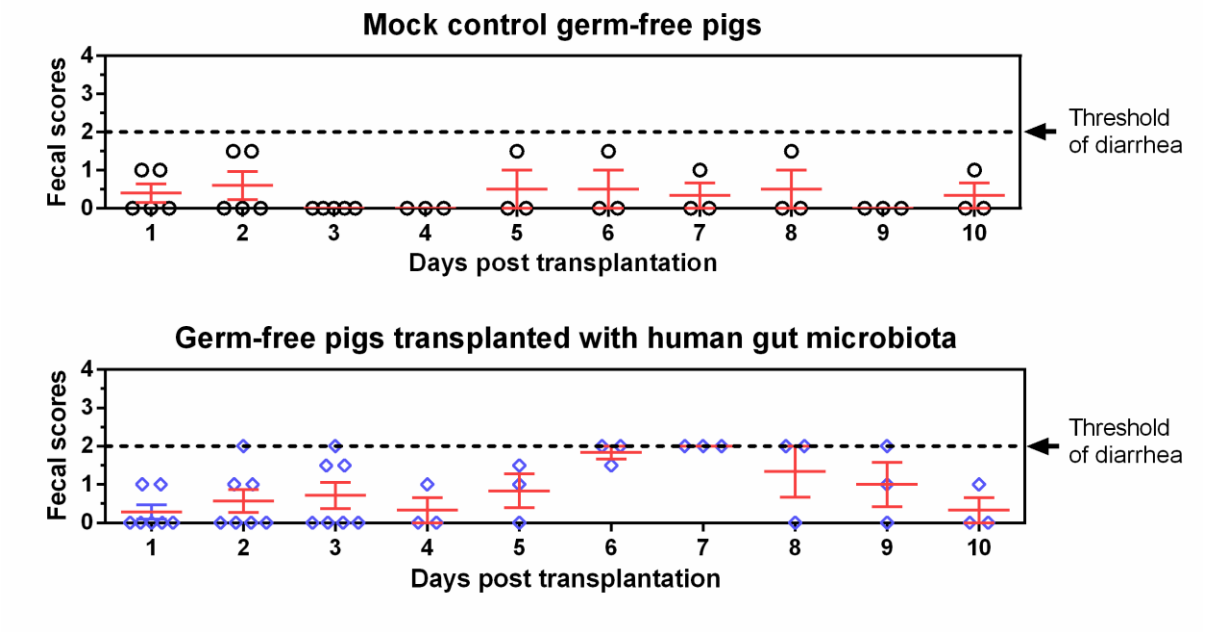

Figure 7. Porcine rotavirus was identified for the diarrhea in germ-free pigs transplanted with human gut microbiota. Fecal consistency scores were evaluated daily after the transplantation of human gut microbiota, the score scaling system was 0, solid; 1, semisolid; 2, pasty; 3, semiliquid; and 4, liquid. Pigs with daily fecal scores of ≥ 2 were considered diarrheic.

traditional techniques have excluded all potential pathogens known to be responsible to the disease symptoms. In a study evaluating the effect of human gut microbiota (HGM) on host health and immune function, an infant stool sample containing intact HGM was transplanted to seven germ-free pigs. The selected infant stool was screened and confirmed negative for human rotavirus, norovirus, astrovirus, sapovirus, adenovirus, and *Klebsiella* spp via PCR (41), but all the germ-free pigs experienced diarrhea after HGM transplantation (Fig. 7). To investigate the potential viral pathogens that might have caused the diarrhea in these pigs, shotgun RNAseq was performed for pig small intestinal tissues, and porcine rotavirus C (GenBank accession no.

KP776735) was found in all the seven HGM colonized pigs using RVRD as the reference database in FastViromeExplorer, and the average abundance was 146. However, no viruses were identified with NCBI RNA virus reference. Therefore, the expanded RNA virus genome database RVRD exhibited improved capacity for RNA virus detection in metagenomic data.

Chapter 4 Conclusion and discussion

Viruses are the most abundant biological entities on the planet, and it is believed that more than 40% of metagenomic sequences are derived from viruses (42). However, current bioinformatics analysis relies on alignment based classification, which is largely limited by the under-represented viruses in the reference database, and thus the majority of metagenomic reads belonging to viruses were unrecognized and designated as the “viral dark matter” (43). Despite extra challenge it might have as compared to DNA viruses and retroviruses, RNA virus metagenomics has been advancing our knowledge in virus diversity and addressing healthcare issues for humans and animals (33). In this study, to enhance the detection of RNA viruses in metagenomic data, an expanded database RVRD was developed by RNA virus genome curation from NCBI, ViralZone, and ViPR. RVRD has a comprehensive coverage of both eukaryotic and prokaryotic RNA viruses, and emphasizes on pathogenic RNA viruses. Additionally, our preliminary results using metagenomic data obtained from environmental and animal tissue samples demonstrated the higher efficiency of RVRD in identifying RNA viruses than NCBI RNA virus reference database.

RVRD contains sequences of two circular ssRNA virus families, *Avsunviroidae* and *Pospiviroidae*, and both of them were obtained from ViralZone. The clustering of RNA virus sequences from NCBI and ViralZone at 100% identity resulted in 4768 sequences, which was 39 more than the 4729 sequences observed from NCBI, indicating that the 39 circular ssRNA viruses were not included in our pooled sequences from NCBI (Fig. 3 and Table 1). The covalently closed circular RNAs are infectious long non-coding subviral agents also known as viroids, and they are

able to autonomously infect plant hosts and hijack host proteins for most biological functions including replication and pathogenesis (44). Circular ssRNA viruses cause diseases in several important crop plants, and the latent infections are also a potential threat to susceptible hosts that might lead to devastating outcomes (45). While the identification of circular ssRNA viruses is becoming an area of active study, it merits our attention that they are RNA present in host cells and freely in the environment, and thus sample pre-treatments such as virus particle enrichment might filter them out in many cases.

Using HCV genomes as an example, the optimal clustering identity was determined as 98% (Fig. 4), which aimed to reduce the database redundancy and retain diversity. However, the degree of redundancy varies in different virus species, and the uniqueness of two genomes are generally determined by the corresponding protein sequences/functions instead of nucleotide sequence similarities, thus it is unlikely to come up with a universal clustering identity optimal for all viruses, and it is also labor-intensive and computational expensive to cluster each virus species separately. Nevertheless, clustering is required in the development of reference database, since it not only reduces the redundancy and thus the database size but also saves time for the database users, especially when it comes to large metagenomic data. In addition, a more stringent identity such as 95% or even lower might be used, so that a database with more divergent and unique sequences could be generated, although the cost would be the loss of the degree of diversity. The version of RVRD generated by clustering sequences from ViPR at 95% identity is also available online.

FastViromeExplorer aligns metagenomic short reads to the reference database directly, which accelerates the classification process and enables the quantification of virus titers (22).

However, this type of algorithms might be somewhat blind to highly similar reference sequences, including subspecies and different virus strains. For example, there were 4 different strains of porcine rotavirus C VP3 gene in the output of performance evaluation using RVRD (Supplementary Table 3), and it was hard to tell how many strains there were in the original wastewater sample. In addition, similar confusion might occur when using NCBI RNA virus reference, because it contains highly similar sequences as well (Supplementary Table 4). In case of the user needs to know exactly what strain it is in such situations, assembly-based pipelines could be used instead, as it generates longer contigs first and thus might gain more accurate alignment among subspecies or different virus strains. Also in such cases, it could be more beneficial to use a database developed by clustering pooled sequences at 100% identity, because all potential virus subspecies or strains will be maintained in the database and considered for alignment.

The presence of porcine rotavirus C has been reported worldwide, and diarrhea outbreaks caused by porcine rotavirus C infection could occur in nursing, weaning, and post-weaning pigs (46). In this study, porcine rotavirus C was detected in the HGM transplanted germ-free pigs, suggesting that the original infant stool used for fecal transplantation might be contaminated with porcine rotavirus C. The origin of the virus might be the consumption of undercooked pork containing the virus or contamination during sample handling and processing. The nucleotide sequences of the observed porcine rotavirus C (GenBank accession no. KP776735) and the corresponding human rotavirus C (GenBank accession no. NC_007571) were compared by BLAST, and the results showed that there is no significant similarity, thus it is very likely that human

rotavirus specific PCR was not able to detect porcine rotavirus C during the virus screening procedure.

Overall, RVRD achieves an enhanced RNA virus identification in metagenomic data when compared to the broadly used NCBI RNA virus reference. Additional work includes updates from the original databases and looking for new resources, as well as a Web application for user feedback. The developed RVRD and other resources are available at:

https://bench.cs.vt.edu/FastViromeExplorer/RNA_virus_database/

Reference

1. Paez-Espino, D., Eloe-Fadrosh, E. A., Pavlopoulos, G. A., Thomas, A. D., Huntemann, M., Mikhailova, N., Rubin, E., Ivanova, N. N., and Kyrpides, N. C. (2016) Uncovering Earth's virome. *Nature* **536**, 425-430
2. Brum, J. R., Ignacio-Espinoza, J. C., Roux, S., Doulcier, G., Acinas, S. G., Alberti, A., Chaffron, S., Cruaud, C., de Vargas, C., Gasol, J. M., Gorsky, G., Gregory, A. C., Guidi, L., Hingamp, P., Iudicone, D., Not, F., Ogata, H., Pesant, S., Poulos, B. T., Schwenck, S. M., Speich, S., Dimier, C., Kandels-Lewis, S., Picheral, M., Searson, S., Tara Oceans, C., Bork, P., Bowler, C., Sunagawa, S., Wincker, P., Karsenti, E., and Sullivan, M. B. (2015) Ocean plankton. Patterns and ecological drivers of ocean viral communities. *Science* **348**, 1261498
3. Cadwell, K. (2015) The virome in host health and disease. *Immunity* **42**, 805-813
4. Zhao, G., Vatanen, T., Droit, L., Park, A., Kostic, A. D., Poon, T. W., Vlamakis, H., Siljander, H., Harkonen, T., Hamalainen, A. M., Peet, A., Tillmann, V., Ilonen, J., Wang, D., Knip, M., Xavier, R. J., and Virgin, H. W. (2017) Intestinal virome changes precede autoimmunity in type I diabetes-susceptible children. *Proc Natl Acad Sci U S A* **114**, E6166-e6175
5. Nakatsu, G., Zhou, H., Wu, W. K. K., Wong, S. H., Coker, O. O., Dai, Z., Li, X., Szeto, C. H., Sugimura, N., Yuen-Tung Lam, T., Chi-Shing Yu, A., Wang, X., Chen, Z., Chi-Sang Wong, M., Ng, S. C., Chan, M. T. V., Chan, P. K. S., Leung Chan, F. K., Jao-Yiu Sung, J., and Yu, J. (2018) Alterations in Enteric Virome Are Associated With Colorectal Cancer and Survival Outcomes. *Gastroenterology*
6. Carding, S. R., Davis, N., and Hoyles, L. (2017) Review article: the human intestinal virome in health and disease. *Alimentary pharmacology & therapeutics* **46**, 800-815
7. Geoghegan, J. L., and Holmes, E. C. (2017) Predicting virus emergence amid evolutionary noise. *Open biology* **7**
8. Nooij, S., Schmitz, D., Vennema, H., Kroneman, A., and Koopmans, M. P. G. (2018) Overview of Virus Metagenomic Classification Methods and Their Biological Applications. *Frontiers in microbiology* **9**, 749
9. Rose, R., Constantinides, B., Tapinos, A., Robertson, D. L., and Prosperi, M. (2016) Challenges in the analysis of viral metagenomes. *Virus evolution* **2**, vew022
10. Ruby, J. G., Bellare, P., and Derisi, J. L. (2013) PRICE: software for the targeted assembly of components of (Meta) genomic sequence data. *G3 (Bethesda, Md.)* **3**, 865-880
11. Cotten, M., Lam, T. T., Watson, S. J., Palser, A. L., Petrova, V., Grant, P., Pybus, O. G., Rambaut, A., Guan, Y., Pillay, D., Kellam, P., and Nastouli, E. (2013) Full-genome deep sequencing and phylogenetic analysis of novel human betacoronavirus. *Emerging infectious diseases* **19**, 736-742b
12. Manso, C. F., Bibby, D. F., and Mbisa, J. L. (2017) Efficient and unbiased metagenomic recovery of RNA virus genomes from human plasma samples. *Sci Rep* **7**, 4173
13. Khan, A. S., Vacante, D. A., Cassart, J. P., Ng, S. H., Lambert, C., Charlebois, R. L., and King, K. E. (2016) Advanced Virus Detection Technologies Interest Group (AVDTIG): Efforts on High Throughput Sequencing (HTS) for Virus Detection. *PDA journal of pharmaceutical science and technology* **70**, 591-595
14. Wommack, K. E., Bhavsar, J., Polson, S. W., Chen, J., Dumas, M., Srinivasiah, S., Furman, M., Jamindar, S., and Nasko, D. J. (2012) VIROME: a standard operating procedure for analysis of viral metagenome sequences. *Standards in genomic sciences* **6**, 427-439
15. Roux, S., Tournayre, J., Mahul, A., Debroas, D., and Enault, F. (2014) Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. *BMC Bioinformatics* **15**, 76

16. Roux, S., Enault, F., Hurwitz, B. L., and Sullivan, M. B. (2015) VirSorter: mining viral signal from microbial genomic data. *PeerJ* **3**, e985
17. Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A., and Sun, F. (2017) VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* **5**, 69
18. Roux, S., Emerson, J. B., Eloe-Fadrosh, E. A., and Sullivan, M. B. (2017) Benchmarking viromics: an in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. *PeerJ* **5**, e3817
19. Vazquez-Castellanos, J. F., Garcia-Lopez, R., Perez-Brocal, V., Pignatelli, M., and Moya, A. (2014) Comparison of different assembly and annotation tools on analysis of simulated viral metagenomic communities in the gut. *BMC Genomics* **15**, 37
20. Rampelli, S., Soverini, M., Turroni, S., Quercia, S., Biagi, E., Brigidi, P., and Candela, M. (2016) ViromeScan: a new tool for metagenomic viral community profiling. *BMC Genomics* **17**, 165
21. Laffy, P. W., Wood-Charlson, E. M., Turaev, D., Weynberg, K. D., Botte, E. S., van Oppen, M. J., Webster, N. S., and Rattei, T. (2016) HoloVir: A Workflow for Investigating the Diversity and Function of Viruses in Invertebrate Holobionts. *Frontiers in microbiology* **7**, 822
22. Tithi, S. S., Aylward, F. O., Jensen, R. V., and Zhang, L. (2018) FastViromeExplorer: a pipeline for virus and phage identification and abundance profiling in metagenomics data. *PeerJ* **6**, e4227
23. Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Ostell, J., Pruitt, K. D., and Sayers, E. W. (2018) GenBank. *Nucleic Acids Research* **46**, D41-D47
24. O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C. M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V. S., Kodali, V. K., Li, W., Maglott, D., Masterson, P., McGarvey, K. M., Murphy, M. R., O'Neill, K., Pujar, S., Rangwala, S. H., Rausch, D., Riddick, L. D., Schoch, C., Shkeda, A., Storz, S. S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R. E., Vatsan, A. R., Wallin, C., Webb, D., Wu, W., Landrum, M. J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T. D., and Pruitt, K. D. (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, D733-745
25. Brister, J. R., Ako-Adjei, D., Bao, Y., and Blinkova, O. (2015) NCBI viral genomes resource. *Nucleic Acids Res* **43**, D571-577
26. Smith, D. B., Bukh, J., Kuiken, C., Muerhoff, A. S., Rice, C. M., Stapleton, J. T., and Simmonds, P. (2014) Expanded classification of hepatitis C virus into 7 genotypes and 67 subtypes: updated criteria and genotype assignment web resource. *Hepatology (Baltimore, Md.)* **59**, 318-327
27. Anthony, S. J., Epstein, J. H., Murray, K. A., Navarrete-Macias, I., Zambrana-Torrelio, C. M., Solovyov, A., Ojeda-Flores, R., Arrigo, N. C., Islam, A., Ali Khan, S., Hosseini, P., Bogich, T. L., Olival, K. J., Sanchez-Leon, M. D., Karesh, W. B., Goldstein, T., Luby, S. P., Morse, S. S., Mazet, J. A., Daszak, P., and Lipkin, W. I. (2013) A strategy to estimate unknown viral diversity in mammals. *mBio* **4**, e00598-00513
28. Paez-Espino, D., Chen, I. A., Palaniappan, K., Ratner, A., Chu, K., Szeto, E., Pillay, M., Huang, J., Markowitz, V. M., Nielsen, T., Huntemann, M., TB, K. R., Pavlopoulos, G. A., Sullivan, M. B., Campbell, B. J., Chen, F., McMahon, K., Hallam, S. J., Denef, V., Cavicchioli, R., Caffrey, S. M., Streit, W. R., Webster, J., Handley, K. M., Salekdeh, G. H., Tsesmetzis, N., Setubal, J. C., Pope, P. B., Liu, W. T., Rivers, A. R., Ivanova, N. N., and Kyrpides, N. C. (2017) IMG/VR: a database of cultured and uncultured DNA Viruses and retroviruses. *Nucleic Acids Res* **45**, D457-D465
29. Sharma, D., Priyadarshini, P., and Vrati, S. (2015) Unraveling the web of viroinformatics: computational tools and databases in virus research. *J Virol* **89**, 1489-1501

30. Adriaenssens, E. M., Farkas, K., Harrison, C., Jones, D. L., Allison, H. E., and McCarthy, A. J. (2018) Viromic Analysis of Wastewater Input to a River Catchment Reveals a Diverse Assemblage of RNA Viruses. *mSystems* **3**
31. Bolduc, B., Shaughnessy, D. P., Wolf, Y. I., Koonin, E. V., Roberto, F. F., and Young, M. (2012) Identification of novel positive-strand RNA viruses by metagenomic analysis of archaea-dominated Yellowstone hot springs. *J Virol* **86**, 5562-5573
32. Buchfink, B., Xie, C., and Huson, D. H. (2015) Fast and sensitive protein alignment using DIAMOND. *Nature methods* **12**, 59-60
33. Greninger, A. L. (2018) A decade of RNA virus metagenomics is (not) enough. *Virus Res* **244**, 218-229
34. Masson, P., Hulo, C., De Castro, E., Bitter, H., Gruenbaum, L., Essioux, L., Bougueleret, L., Xenarios, I., and Le Mercier, P. (2013) ViralZone: recent updates to the virus knowledge resource. *Nucleic Acids Res* **41**, D579-583
35. Pickett, B. E., Greer, D. S., Zhang, Y., Stewart, L., Zhou, L., Sun, G., Gu, Z., Kumar, S., Zaremba, S., Larsen, C. N., Jen, W., Klem, E. B., and Scheuermann, R. H. (2012) Virus pathogen database and analysis resource (ViPR): a comprehensive bioinformatics database and analysis resource for the coronavirus research community. *Viruses* **4**, 3209-3226
36. Li, W., and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics (Oxford, England)* **22**, 1658-1659
37. Li, W., Jaroszewski, L., and Godzik, A. (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics (Oxford, England)* **17**, 282-283
38. Goodacre, N., Aljanahi, A., Nandakumar, S., Mikailov, M., and Khan, A. S. (2018) A Reference Viral Database (RVDB) To Enhance Bioinformatics Analysis of High-Throughput Sequencing for Novel Virus Detection. *mSphere* **3**
39. Delwart, E., Kapusinszky, B., Pesavento, P. A., Estrada, M., Seguin, M. A., and Leutenegger, C. M. (2017) Genome Sequence of Canine Polyomavirus in Respiratory Secretions of Dogs with Pneumonia of Unknown Etiology. *Genome announcements* **5**
40. Phan, T. G., Sdiri-Loulizi, K., Aouni, M., Ambert-Balay, K., Pothier, P., Deng, X., and Delwart, E. (2014) New parvovirus in child with unexplained diarrhea, Tunisia. *Emerging infectious diseases* **20**, 1911-1913
41. Twitchell, E. L., Tin, C., Wen, K., Zhang, H., Becker-Dreps, S., Azcarate-Peril, M. A., Vilchez, S., Li, G., Ramesh, A., Weiss, M., Lei, S., Bui, T., Yang, X., Schultz-Cherry, S., and Yuan, L. (2016) Modeling human enteric dysbiosis and rotavirus immunity in gnotobiotic pigs. *Gut Pathog* **8**, 51
42. Krishnamurthy, S. R., and Wang, D. (2017) Origins and challenges of viral dark matter. *Virus Res* **239**, 136-142
43. Zhang, Y. Z., Shi, M., and Holmes, E. C. (2018) Using Metagenomics to Characterize an Expanding Virosphere. *Cell* **172**, 1168-1172
44. Steger, G., and Perreault, J. P. (2016) Structure and Associated Biological Functions of Viroids. *Advances in virus research* **94**, 141-172
45. Kovalskaya, N., and Hammond, R. W. (2014) Molecular biology of viroid–host interactions and disease control strategies. *Plant Science* **228**, 48-60
46. Vlasova, A. N., Amimo, J. O., and Saif, L. J. (2017) Porcine Rotaviruses: Epidemiology, Immune Responses and Control Strategies. *Viruses* **9**

Supplementary Table 1. The taxonomy of RNA viruses in databases.

Virus group	Virus family/class	NCBI Viral Genomes	Viralzone	ViPR	Final
dsRNA viruses	Amalgaviridae	11	11		11
	Birnaviridae	9	9		9
	Chrysoviridae	13	13		13
	Cystoviridae	7	7		7
	Endornaviridae	32	32		32
	Hypoviridae	15	15		15
	Megabirnaviridae	3	3		3
	Partitiviridae	75	75		75
	Picobirnaviridae	6	6		6
	Quadriviridae	1	1		1
	Reoviridae	89	89	51300	10795
	Totiviridae	74	74		74
	unclassified dsRNA viruses	27	2		27
ssRNA(+) virus	Albetovirus	2	2		2
	Alphatetraviridae	4	4		4
	Alvernaviridae	1	1		1
	Astroviridae	55	55		55
	Aumaivirus	1	1		1
	Barnaviridae	1	1		1
	Benyviridae	4	4		4
	Blunervirus	1	1		1
	Bromoviridae	37	37		37
	Caliciviridae	27	27	1473	376
	Carmotetraviridae	1	1		1
	Cilevirus	2	2		2
	Closteroviridae	52	52		52
	Flaviviridae	138	138	12882	3578
	Hepeviridae	7	7	386	109
	Higrevirus	1	1		1
	Idaeovirus	3	3		3
	Leviviridae	11	11		11
	Luteoviridae	57	57		57
	Narnaviridae	43	43		43
	Nidovirales (Coronaviridae)	60	60	2503	785
	Nidovirales (others)	30	30		30
	Nodaviridae	16	16		16
	Ourmiavirus	4	4		4
	Papanivirus	1	1		1

	Permutotetraviridae	2	2		2
	Picornavirales (Picornaviridae)	140	140	4432	1056
	Picornavirales (others)	172	172		172
	Potyviridae	210	210		210
	Sarthroviridae	1	1		1
	Sinaivirus	7	7		7
	Solemoviridae	20	20		20
	Solinviviridae	2	2		2
	Togaviridae	33	33	1699	496
	Tombusviridae	77	77		77
	Tymovirales	223	223		223
	Virgaviridae	59	59		59
	Virtovirus	1	1		1
	unclassified ssRNA positive-strand viruses	41	0		41
	environmental samples	1	0		1
ssRNA(-) virus	Arenaviridae	44	44	1133	389
	Aspiviridae	7	7		7
	Bunyaviridae	190	190	6545	2178
	Deltavirus	1	0		1
	Mononegavirales (Filoviridae)	7	7	586	175
	Mononegavirales (Paramyxoviridae)	64	64	3221	1067
	Mononegavirales (Rhabdoviridae)	174	174	1830	559
	Mononegavirales (Others)	32	32		32
	Orthomyxoviridae	10	10		10
	Tilapinevirus	1	0		1
	unclassified ssRNA negative-strand viruses	16	0		16
Circular ssRNA virus	Avsunviroidae	0	4	0	4
	Pospiviroidae	0	35	0	35

Supplementary Table 2. Evaluation performance output using NCBI RNA virus reference

#VirusIdentifier	VirusName	EstimatedAbundance
NC_001367.1	Tobacco mosaic virus, complete genome	4409
NC_003630.1	Pepper mild mottle virus, complete genome	4074
NC_002692.1	Tomato mosaic virus, complete genome	3193
NC_002618.2	Cocksfoot mottle virus, complete genome	2330
NC_011507.2	Rotavirus A segment 1, complete genome	952
NC_002792.2	Ribgrass mosaic virus, complete genome	576
NC_011503.2	Rotavirus A segment 9, complete genome	560
NC_004106.1	Paprika mild mottle virus, complete genome	303
NC_011501.2	Rotavirus A segment 7, complete genome	72
NC_011504.2	Rotavirus A segment 10, complete genome	45
NC_001873.1	Turnip vein-clearing virus, complete genome	20
NC_007547.1	Rotavirus C segment 1, complete sequence	17
NC_028478.1	Tomato brown rugose fruit virus isolate Tom1-Jo, complete genome	16
NC_007543.1	Rotavirus C segment 6, complete sequence	15
	Total	16582

Supplementary Table 3. Evaluation performance output using RVRD

#VirusIdentifier	VirusName	EstimatedAbundance
NC_001367.1	1. Tobacco mosaic virus, complete genome	4409
NC_003630.1	2. Pepper mild mottle virus, complete genome	4074
NC_002692.1	3. Tomato mosaic virus, complete genome	3193
NC_002618.2	4. Cocksfoot mottle virus, complete genome	2330
KU128895	5. Rotavirus A strain RVA/Human-wt/Bel/BEF06018/2014/G29P41 VP4 gene, complete cds	1170
EF554115	6. Rotavirus A strain RVA/Human-wt/BEL/B10925/1997/G6P[14] VP1 gene, complete cds	876.952
NC_002792.2	7. Ribgrass mosaic virus, complete genome	576
NC_004106.1	8. Paprika mild mottle virus, complete genome	303
KP776735.1	9. Porcine rotavirus C strain RVC/Pig-wt/CZE/P21/2013 NSP2 (NSP2) gene, complete cds	169
KU128898	10. Rotavirus A strain RVA/Human-wt/Bel/BEF06018/2014/G29P41 NSP1 gene, complete cds	152
KY910004	11. UNVERIFIED: Rotavirus C isolate RVC/Pig-wt/CAN/A11-152/2014/G6P5 structural protein 2-like gene, complete sequence	122
KY909972	12. Rotavirus C isolate RVC/Pig-wt/CAN/A5-36/2014/G6P4 non structural protein 4 gene, complete cds	96
HQ185673	13. Human rotavirus C strain BK0830 VP1 gene, complete cds	87.9775
KP982879	14. Rotavirus C VP1 (VP1) gene, complete cds	52
LC122593	15. Porcine rotavirus C gene for VP2, complete cds, strain: RVC/Pig-wt/Tochigi-1-1/2015/G9P[4]	48
EF554123	16. Rotavirus A strain RVA/Human-wt/BEL/B10925/1997/G6P[14] NSP3 gene, complete cds	37.4599
KX442661	17. Rotavirus C strain RVC/SB-wt/IND/UP-SB37/2016 outer capsid protein (VP7) gene, complete cds	34.4472
LC307023	18. Porcine rotavirus C NSP3 gene for nonstructural protein 3, complete cds, strain: 86-H3	32
KY910009	19. UNVERIFIED: Rotavirus C isolate RVC/Pig-wt/CAN/A8-158/2014/G6P4 structural protein 3-like gene, complete sequence	26
LC307187	20. Porcine rotavirus C VP3 gene for viral structural protein 3, complete cds, strain: 87-I4	24
NC_001873.1	21. Turnip vein-clearing virus, complete genome	20
KT284778	22. Human rotavirus C isolate CAU14-1-242 segment 3, complete sequence	16.0952
NC_028478.1	23. Tomato brown rugose fruit virus isolate Tom1-Jo, complete genome	16
NC_034452.1	24. Picobirnavirus green monkey/KNA/2015 strain PBV/Simian/KNA/08984/2015 RNA-dependent RNA polymerase gene, complete cds	15
LC307194	25. Porcine rotavirus C VP3 gene for viral structural protein 3, complete cds, strain: 134-9	15
LC129108	26. Human rotavirus C NSP3 gene for NSP3 protein, complete cds, strain: RVC/Human-wt/JPN/HO-62/2005/G4P[2]	15

	27. Human rotavirus C isolate RVC/Human-wt/RUS/Novosibirsk/Nsk08-3414/2008 nonstructural protein NSP3	
JQ177071	(NSP3) gene, complete cds	13
	28. Porcine rotavirus C gene for VP3, complete cds, strain: RVC/Pig-wt/Ishi-1/2015/G13P[4]	
LC122626		12
	29. Rotavirus A strain dog-wt/GER/88977/2013/G8P1 nonstructural protein 3 gene, complete cds	
KJ940159		11.2826
	30. Porcine rotavirus C VP3 gene for viral structural protein 3, complete cds, strain: CJ3-6	
LC307175		11.0506
	31. Porcine rotavirus C isolate RVC/Pig-wt/KOR/2478/2012/GXPX outer capsid protein (VP4) gene, complete cds	
KJ814474		10
	Total	17967

Supplementary Table 4. Sequences from NCBI RNA virus reference with high similarity

ID	Name	Similarity
NC_034973.1	Aphis glycines virus 3 isolate S6-IA	99.40%
NC_033722.1	Wuhan insect virus 33 strain WHCCII11871	
NC_001411.2	Black beetle virus	99.07%
NC_004146.1	Flock house virus	
NC_024493.1	Senegalese sole Iberian betanodavirus	98.38%
NC_003449.1	Striped Jack nervous necrosis virus	
NC_008040.1	Redspotted grouper nervous necrosis virus	97.07%
NC_024492.1	Senegalese sole Iberian betanodavirus	
NC_027811.2	Fengkai orbivirus isolate D181/2008 segment 2	97.21%
NC_033783.1	Orbivirus SX-2017a VP2 gene	
NC_025810.1	Cangyuan orthoreovirus strain Cangyuan segment S4	96.28%
NC_020446.1	Melaka orthoreovirus segment S4	
NC_025344.1	Canine pneumovirus strain dog/Bari/100-12/ITA/2012	95.80%
NC_006579.1	Pneumonia virus of mice J3666	
NC_033710.1	Wuhan insect virus 9 strain WHCCII13328	95.55%
NC_033326.1	Hubei Wuhan insect virus 9 strain QTM27230	
NC_025807.1	Cangyuan orthoreovirus strain Cangyuan segment S3	95.25%
NC_020445.1	Melaka orthoreovirus segment S3	
NC_029054.2	Potiskum virus	94.94%
NC_033697.1	Saboya virus	
NC_007367.1	Influenza A virus (A/New York/392/2004(H3N2)) segment 7	94.94%
NC_007377.1	Influenza A virus (A/Korea/426/1968(H2N2)) segment 7	