vsRNAfinder: a novel method for identifying high-confidence viral small RNAs from small RNA-Seq data

Zena Cai[1,#], Ping Fu[1,#], Ye Qiu[1], Aiping Wu[2,3], Gaihua Zhang[4], Yirong Wang[1], Taijiao Jiang[5], Xing-Yi Ge[1], Haizhen Zhu[1], Yousong Peng[1,*]

[1]Bioinformatics Center, College of Biology, Hunan Provincial Key Laboratory of Medical Virology, Hunan University, Changsha 410082, China

[2]Institute of Systems Medicine, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China

[3]Suzhou Institute of Systems Medicine, Suzhou, China

[4] The National and Local Joint Engineering Laboratory of Animal Peptide Drug Development, College of Life Sciences, Hunan Normal University, Changsha 410081, China

[5] Guangzhou Laboratory, Guangzhou, China

# Contributed equally;

*To whom correspondence should be addressed.

Abstract

Virus-encoded small RNAs (vsRNA) have been reported to play an important role in viral infection. Unfortunately, there is still a lack of an effective method for vsRNA identification. Besides, a systematic characterization of vsRNAs is also needed. Herein, we presented vsRNAfinder, a de novo method for identifying high-confidence vsRNAs from small RNA-Seq (sRNA-Seq) data based on peak calling and Poisson distribution and is public available at https://github.com/ZenaCai/vsRNAfinder.

vsRNAfinder outperformed two widely-used methods namely miRDeep2 and ShortStack in identifying viral miRNA, siRNA and piRNA. It can also be used to identify sRNAs in animals and plants with similar performance to miRDeep2 and ShortStack. Based on vsRNAfinder, a total of 19,734 high-confidence vsRNAs including 2,746 miRNAs were identified in 64 viral species from public sRNA-Seq data. The number of vsRNAs identified in viruses varied from 1 to 2,489 with a median of 170. The length distribution of vsRNAs peaked at 21 and 22 nt. Plant viruses were found to express larger number and higher levels of sRNAs than those of animal viruses. Besides, the number of vsRNAs identified increased as the viral infection persisted. Interestingly, the vsRNA showed strong expression specificity as little overlap was observed among vsRNAs identified in different strains of a virus, or in different hosts, cells or tissues infected by the same virus. Little conservation was observed among vsRNAs of different viruses. Overall, the study not only presents an effective method for vsRNA identification, but also provides a rich resource for further studies of vsRNAs.

Introduction

Small RNAs (sRNAs) are a kind of regulatory RNAs in cellular processes [1]. They are typically less than 200 nucleotides (nt) in length [1]. The sRNAs can be mainly divided into three classes including microRNA (miRNA) with a length of 18-24 nt, small interfering RNA (siRNA) with a length of 21-23 nt and piwi-interacting RNA (piRNA) with a length of 24-31 nt based on the mechanism of biogenesis and function

[2]. By the mechanism of biogenesis, the miRNAs are derived from single-stranded RNA precursors with a hairpin structure which are firstly processed by Drosha in the nucleus and then processed by Dicer after being transported to the cytoplasm; the siRNAs are derived from double-stranded RNA precursors in the cytoplasm and are processed by Dicer; the piRNAs are derived from the fragmentation of single-stranded RNA precursors via a Dicer-independent mechanism [3-5]. By the mechanism of function, although all three kinds of sRNAs can target the mRNAs to regulate their expressions by incorporating sRNAs into the RNA-induced silencing complex (RISC), their function mechanisms differ a lot: with the help of AGO protein, the miRNAs can repress mRNAs translation by binding imperfectly to their mRNA targets, while the siRNAs can degrade mRNAs by binding perfectly to their mRNA targets; the piRNAs can degrade mRNAs by binding perfectly their mRNA targets with the help of PIWI protein [5, 6]. The sRNAs play a fundamental role in the maintenance of normal physiology and serve as potential biomarkers for early diagnosis and therapeutic targets for a wide spectrum of diseases, particularly the cancer [1]. For example, miR-130 can down-regulate the expression of miRNA-221 and miRNA-222 by targeting the MET, thus reducing TNF-related apoptosis-induced ligand (TRAIL) resistance in non-small cell lung cancer [7].

Numerous sRNAs have been discovered in multiple organisms as the development of the next generation sequencing (NGS) technology and a growing number of sRNA databases are emerging, such as the miRBase [8], SEAweb [9] and ASRD [10]. The miRBase database is the most comprehensive database of miRNAs. It contains 48,860

mature miRNAs from 271 organisms, and provides rich information such as the expression, confidence, and method of identifying miRNAs. The SEAweb database integrates multiple classes of sRNAs from 10 organisms based on data-mining from over 4,200 small RNA sequencing (sRNA-Seq) samples. The ASRD database provides a platform for the comparison of sRNA abundances across over 2,000 sRNA-Seq samples in *Arabidopsis*. In recent years, lots of virus-encoded sRNAs (vsRNA) have been identified and proved to be functional in the regulation of viral life cycle and pathogenesis [11, 12]. For example, the DENV-vsRNA-5 was reported to suppress viral replication by targeting the coding region of viral non-structural protein 1 [13], and the svRNA-N of the Severe Acute Respiratory Syndrome coronavirus (SARS-CoV) was involved in the enhancement of lung inflammatory pathology [14]. However, most studies only investigated specific vsRNA class such as miRNA, or focused on vsRNAs from one or a few viruses. There is still a lack of systematic studies about vsRNAs.

Numerous methods have been developed to identify sRNAs from sRNA-Seq data. They can be mainly divided into two kinds. The first kind of method is to identify sRNAs based on a known sRNA library, such as the sRNAbench [15], Manatee [16] and sRNAnalyzer [17]. This kind of method can be used for accurate quantification of known sRNA. However, they may not be suitable for organisms with only a few known sRNAs such as viruses. The other kind of method is to identify sRNAs *de novo* by firstly finding the precursor of sRNAs from assembly of reads and then detecting the mature sRNA, such as miRDeep2 [18] and ShortStack [19]. Most

methods mentioned above have been mainly used in the identification of sRNAs in animals and plants. The viruses have much smaller genomes and stronger coding ability compared to animals and plants. Thus, the distribution of vsRNAs is much denser than that of other organisms, which may lead to assembly of adjacent vsRNAs with overlaps into long contigs and missed detection of many vsRNAs by current methods.

Herein, we presented vsRNAfinder, a *de novo* method for the identification of high-confidence vsRNAs from sRNA-Seq data based on peak calling and Poisson distribution. The method can also be used in the identification of sRNAs in animals and plants. Then, we used vsRNAfinder to systematically identify vsRNAs from over 60 viral species and further characterized these vsRNAs. The work would greatly facilitate further studies of vsRNAs.

Materials and methods

Data collection

The viral-infection-related sRNA-Seq data were obtained by firstly searching “virus” in the NCBI GEO [20] and SRA [21] databases. Then, the searching results were filtered by the study type and by the strategy, respectively, in the NCBI GEO and SRA database. Only those with the study type of “Expression profiling by high throughput sequencing” or “Non-coding RNA profiling by high throughput sequencing”, and those with the strategy of “RNASeq” or “other” were kept. Then, the sRNA-Seq data were further filtered manually and only those of viral infections with the genome

sequence of the virus used in the infection available were kept for further analysis. A total of 102 sRNA-Seq datasets were obtained for use in the study (Table S1).

To demonstrate the usage of our method in the identification of sRNAs from other organisms, some representative sRNA-Seq datasets of several widely-studied organisms including *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster*, *Arabidopsis thaliana* and *Oryza sativa* were also obtained from NCBI GEO [20] and SRA [21] database (Table S3).

To validate the performance of sRNA identification methods, the known miRNAs in viruses and other organisms were obtained from the database of miRBase (Release 22.1, http://mirbase.org/) [8]. The experimentally-verified siRNAs in Sugarcane mosaic virus and piRNAs in Dengue viruses were collected from the studies of Xia et al. [22] and Wang et al. [23], respectively.

The vsRNAfinder workflow

The vsRNAfinder workflow included four modules: Preprocessing, Identification, Filtering and annotation, and Quantification, which were described as follows.

Preprocessing module. Firstly, the adapter sequences are trimmed from the raw reads with cutadapt (version 2.6, parameter: '-m 15') [24] (Figure 1). Then, the clean reads are mapped to the viral genome directly (the default), or the clean reads are firstly mapped to the host genome and then the unmapped reads are mapped to the viral

genome, with bowtie (version 1.3.0, parameter: '-v 0 -m 2') [25]. The resulted SAM files are converted to BAM files which are sorted by samtools (version 1.9) [26]. The sorted BAM files are further converted to BED files by bedtools (version 2.29.2) [27].

Identification module. The coverage of mapped reads on the viral genome is calculated by bedtools [27] and then smoothed with the window-size of 11 nt. The peaks of the smoothed curve are identified using the function of *findPeaks* ("minpeakheight = 5") in the R package pracma (version 2.3.3) [28]. The start and end site of the candidate sRNA is restricted to the upstream and downstream 30 nt of the peak, respectively. Then, the number of reads starting at each position in the upstream region of the peak, and that ending at each position in the downstream of the peak are calculated. The position with the highest number of reads in the upstream and that in the downstream are considered as the start and end site of the candidate sRNA, respectively.

Filtering and annotation module. To obtain the high-confidence sRNAs, the abundances of the start and end sites (defined as boundary sites) of the candidate sRNA are normalized using RPM (Reads Per Million) which is calculated by dividing the number of reads starting at the start site or ending at the end site by the total number of clean reads, and then multiplied by 10^6. Only the candidate sRNAs with both of the boundary sites having at least 5 RPM and with lengths ranging from 15 to 34 nt are retained for further analysis.

Then, the Poisson distribution is used to evaluate the statistical significance of candidate sRNAs. Poisson distribution is a discrete probability distribution that

expresses the probability of a given number of events occurring in a fixed interval of time, which is widely applied in NGS data analysis [29]. We hypothesized that the reads should be enriched in the region where the sRNA is expressed compared to the background region. Thus, the significance of candidate sRNAs could be analyzed by the Poisson distribution of which the parameter λ can be calculated as the average coverage of the mapped reads on the background region. For each candidate sRNA, the background region is defined as extending the upstream and downstream 10 nt of the candidate sRNA. Therefore, the λ is local and is calculated as follows:

$$\lambda = \frac{C}{L}$$

where C is the sum of the coverage of mapped reads on each position of the background region, and L is the length of the background region.

Based on the local λ, for a given position X in the candidate sRNA with *k* mapped reads covering the position, the p-value of the position is calculated as follows:

$$P_{\lambda}\,(X = k) = \frac{\lambda^{k}}{k! * e^{-\lambda}}$$

The p-value of the candidate sRNA is calculated as the 80% quantile of p-values in all positions of the candidate sRNA. The candidate sRNAs with a p-value smaller than 0.05 are considered as statistically significant and highly confident.

The miRNAs are then annotated from the high-confidence sRNAs in two steps. Firstly, the hypothesized precursor of miRNA (pre-miRNA) is obtained. Since the miRNA may be located in either 5’-arm or 3’-arm of the hypothesized pre-miRNA, both cases are considered. If the miRNA is supposed to be in the 5’-arm of the hypothesized pre-miRNA, the start and end positions of hypothesized pre-miRNAs are determined

as follows according to Michael et al.'s study [30]:

$$start_{pre-miRNA} = start_{sRNA} - step * i$$

$$end_{pre-miRNA} = start_{sRNA} + 2 * L_{sRNA} - 1 + L_{loop} + step * i$$

otherwise,

$$start_{pre-miRNA} = start_{sRNA} - (L_{loop} + L_{sRNA} + step * i)$$

$$end_{pre-miRNA} = end_{sRNA} + step * i$$

where $start_{sRNA}$ and end_{sRNA} refer to the start and end position of the sRNA, respectively; L_{sRNA} refers to the length of sRNA; L_{loop} refers to the length of loop and is set to be 15; *step* and *i* are set to be 5 and 8 for animals and viruses, 7 and 10 for plants as plant pre-miRNA sequences are longer than those of animals and viruses, according to Michael et al.'s study [30].

Then, the secondary structure of the hypothesized pre-miRNA is predicted by RNAfold (version 2.4.18) [31] with default parameters. The hypothesized pre-miRNA which has the minimum free energy of less than -25 kcal/mol, and in which at least 70% of positions in the corresponding sRNA form base pairs, is considered to be pre-miRNA having the hairpin structure, and is further processed into the mature miRNA.

Quantification module. The abundances of sRNAs are normalized using RPM. The total number of reads mapping to the sRNA is calculated as the sum of reads that overlap with the sRNA by 80% and those that are utilized to estimate boundary sites of the sRNA. The RPM of each sRNA is calculated by dividing the total number of reads mapping to the sRNA by the total number of clean reads in the sample, and then

multiplied by 10^6.

Prediction of sRNAs with miRDeep2 and ShortStack

The miRDeep2 (version 0.1.2, https://github.com/rajewsky-lab/mirdeep2) [18] and ShortStack (version 3.8.5, https://github.com/MikeAxtell/ShortStack) [19] were used for prediction of sRNAs with default parameters.

Evaluating the performance of computational methods in identifying sRNAs

The known sRNAs (mentioned in Data collection) including miRNAs, siRNAs and piRNAs were used to evaluate the performance of computational methods in identifying sRNAs. For a given organism, the identified sRNAs whose sequences match perfectly to those of known sRNAs are considered as true positives, otherwise false positives; the known sRNAs which are missed by the method are considered as false negatives. Three metrics including Precision, Recall and F1-score were used to evaluate the performance of computational methods in identifying sRNAs, which are calculated as follows:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall}$$

while TP is the number of true positives, FP is the number of false positives, FN is the number of false negatives.

Statistical analysis

The Mann-Whitney U test was used to compare the expression differences between the conserved vsRNAs and non-conserved vsRNAs. A p-value less than 0.05 was considered as statistically significant. All statistical analyses were conducted in Python (version 3.8).

Results

Overview of vsRNAfinder

The vsRNAfinder consisted of four modules including Preprocessing, Identification, Filtering and annotation, and Quantification (Figure 1). The Preprocessing module processed raw reads to clean reads by trimming adapters. Then, the clean reads were firstly mapped to the host genome and then the unmapped reads were mapped to the viral genome, or the clean reads were mapped to the viral genome directly (the default) since the host reference genome was not always available.

The Identification module identified the candidate sRNAs with four steps: firstly, the coverage of mapped reads across the viral genome was calculated; secondly, the coverage was smoothed to reduce the noise of the peak detection and enhance the peak signal; thirdly, the peak was detected which had a higher expression than its adjacent sites; finally, the start and end sites of the candidate sRNA were determined at the upstream and downstream of the peak identified, respectively (see Materials and methods).

The Filtering and annotation module was designed for obtaining high-confidence

sRNAs and annotating miRNAs. The candidate sRNAs identified above were firstly filtered by the abundance of the start and end sites (defined as boundary sites) of sRNAs. Only the candidate sRNAs with both boundary sites having at least 5 RPM were kept. Then, they were further filtered by the length. Only the candidate sRNAs with 15-34 nt were kept which covered the length of most miRNAs, siRNAs and piRNAs [2]. Then, the statistical significance of candidate sRNAs was obtained based on the Poisson distribution. The remaining candidate sRNAs with a p-value smaller than 0.05 were considered as statistically significant and highly confident. Finally, the RNAfold [31] was used to annotate miRNAs from the high-confidence sRNAs (see Materials and methods).

The Quantification module quantified the abundance of sRNAs by RPM. The total number of reads mapping to the sRNA was calculated as the sum of reads that overlapped with the sRNA by 80% and those that were utilized to estimate boundary sites of the sRNA (see Materials and methods).

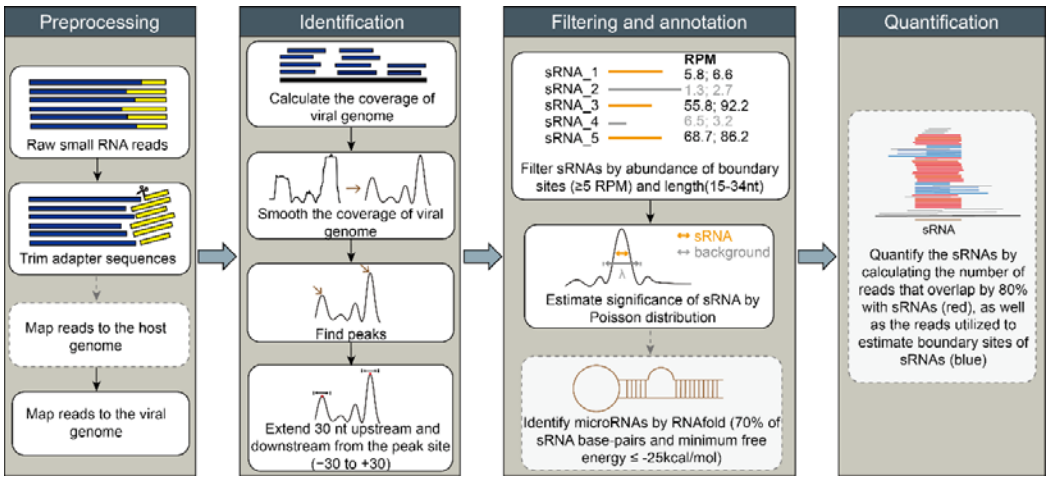

Figure 1. Overview of vsRNAfinder which consisted of four modules including Preprocessing, Identification, Filtering and annotation, and Quantification. Please see Materials and methods for details.

Performance analysis of vsRNAfinder in identifying viral miRNAs

To evaluate the performance of vsRNAfinder in identifying vsRNAs, vsRNAfinder was used to identify viral miRNAs since the miRNAs are the most studied vsRNAs. Four viruses including Murid gammaherpesvirus 4 (MuHV-4), Murid betaherpesvirus 1 (MuHV-1), Human betaherpesvirus 5 (HHV-5) and Human gammaherpesvirus 4 (HHV-4) were selected in the evaluation which had more than 20 viral miRNAs reported in public databases. For comparison, two *de novo* methods, namely the miRDeep2 [18] and ShortStack [19] which are widely used in identifying sRNAs, were selected in identifying viral miRNAs in four viruses.

As shown in Figure 2A, vsRNAfinder achieved the highest Recall ranging from 0.48 to 0.8 in all viruses, followed by miRDeep2 (0.21-0.64) and ShortStack (0.03-0.34). Taking the MuHV-4 as an example, vsRNAfinder had a Recall of 0.54, while miRDeep2 and ShortStack had a Recall of 0.21 and 0.07, respectively. In terms of Precision, the ShortStack performed best with Precision higher than 0.5 in all viruses. The vsRNAfinder had Precisions ranging from 0.28-0.36 which were similar to those of miRDeep2. Overall, when considering both the Recall and Precision using the F1-score, the vsRNAfinder achieved the F1-scores of 0.41, 0.41 and 0.47, respectively, in viruses of MuHV-4, MuHV-1 and HHV-5, which were much higher than those of

the miRDeep2 (0.26, 0.33 and 0.26) and ShortStack (0.13, 0.07 and 0.32). While in the virus of HHV-4, three methods had similar F1-scores.

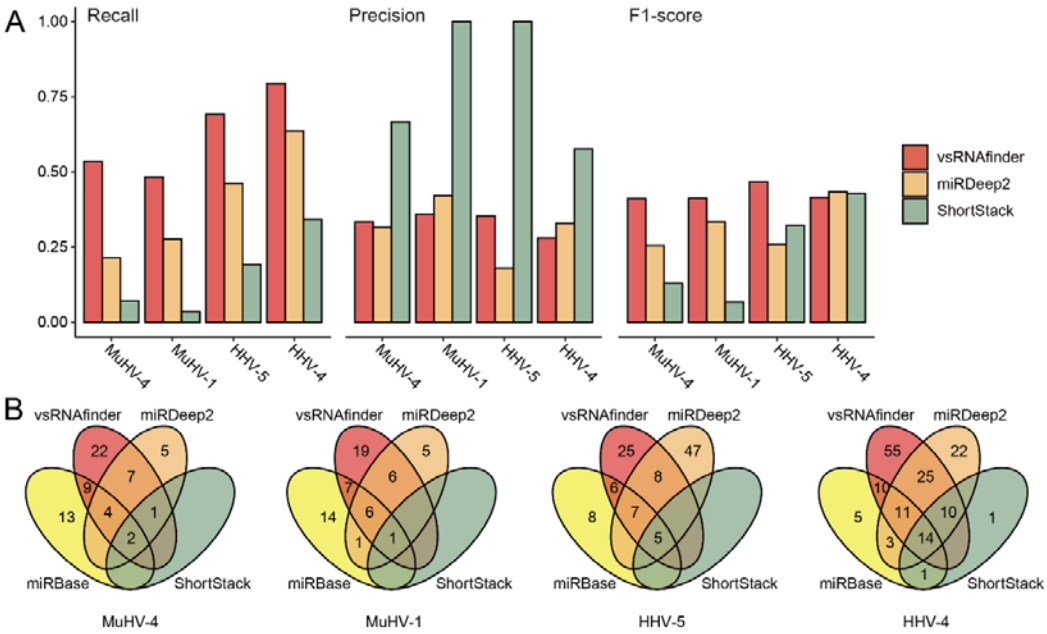

Figure 2. Performance analysis of vsRNAfinder, miRDeep2 and ShortStack in identifying viral miRNAs. (A) Comparison of Recall, Precision and F1-score of vsRNAfinder with miRDeep2 and ShortStack in identifying viral miRNAs in viruses of Murid gammaherpesvirus 4 (MuHV-4), Murid betaherpesvirus 1 (MuHV-1), Human betaherpesvirus 5 (HHV-5) and Human gammaherpesvirus 4 (HHV-4). (B) Venn diagram showing overlaps between viral miRNAs identified by vsRNAfinder, miRDeep2, ShortStack, and those in the miRBase database.

When comparing the number of viral miRNAs identified by three methods and those in miRBase, we found that most viral miRNAs identified by miRDeep2 and ShortStack were also identified by vsRNAfinder. For example, 19 and 3 viral

miRNAs were identified by miRDeep2 and ShortStack, respectively, in the MuHV-4, among which 14 and 3 were also identified by vsRNAfinder. Moreover, vsRNAfinder identified more novel viral miRNAs than the other two methods in viruses of MuHV-4, MuHV-1 and HHV-4 (Figure 2B).

Identification of other classes of sRNAs with vsRNAfinder

Besides viral miRNAs, viruses can also express siRNAs and piRNAs during viral infections [22, 23]. To evaluate the performance of vsRNAfinder in the identification of other vsRNA classes, 11 experimentally-verified siRNAs in Sugarcane mosaic virus (SCMV) [22] and 6 experimentally-verified piRNAs in Dengue virus (DENV) [23] were collected from literature. All SCMV siRNAs and four of six DENV piRNAs were identified correctly by vsRNAfinder (Table S4). The remaining two DENV piRNAs were missed by vsRNAfinder due to low abundance. For comparison, ShortStack only identified correctly one SCMV siRNA and two DENV piRNAs (Table S4), while miRDeep2 can only be used to identify the miRNA.

Systematic identification and characterization of vsRNAs

Then, the vsRNAfinder was used to systematically identify vsRNAs from viral-infection-related sRNA-Seq data in public databases. A total of 19,734 high-confidence vsRNAs were detected from 878 samples in 102 sRNA-Seq datasets by vsRNAfinder (Table S5). Among them, 2,746 vsRNAs were further annotated as miRNAs (Table S5&6). The vsRNAs belonged to 64 viral species from 33 viral

families of seven viral groups (Figure 3A). Over 50% (11,801) of vsRNAs were detected from the positive-sense single-stranded RNA (+ssRNA) viruses, and 3,524 (17.9%) vsRNAs were detected from double-stranded DNA (dsDNA) viruses. The number of vsRNAs identified in viruses varied much and ranged from 1 to 2,489 with a median of 170. The Invertebrate iridescent virus 6 (IIV-6, 2,489) contained most vsRNAs among all viruses, followed by the Orsay virus (OrV, 1,142), the Cucumber mosaic virus (CMV, 1,085), the Polygonum ringspot orthotospovirus (POLRSV, 972), the Sweet potato feathery mottle virus (SPFMV, 949), and so on.

Then, the length, expression and genomic locations of vsRNAs were analyzed. The lengths of vsRNAs showed a unimodal distribution which peaked in 21 and 22 nt (Figure 3B). The vsRNAs with 21 nt accounted for 43% of all vsRNAs, followed by those of 22 nt (17.5%) and 23 nt (7.0%). The expression level of vsRNAs followed an approximate normal distribution (Figure 3C). Over 80% of vsRNAs had more than 20 RPM. The viral miRNAs had similar expression levels with that of other vsRNAs. To analyze the location preference of vsRNAs along the viral genome, the number of vsRNAs located within each 10% of the viral genomes was calculated. As shown in Figure 3D, a slightly higher number of viral miRNAs and other vsRNAs were observed in both ends of the viral genome than those in other genomic regions. Taking the Zika virus as an example, there was 12-13 vsRNAs per Kb in the region around the 10 Kb of the viral genome (near the 3’-end) which were much larger than that observed in other genomic regions (less than 9 vsRNAs per Kb) (Figure 3E).

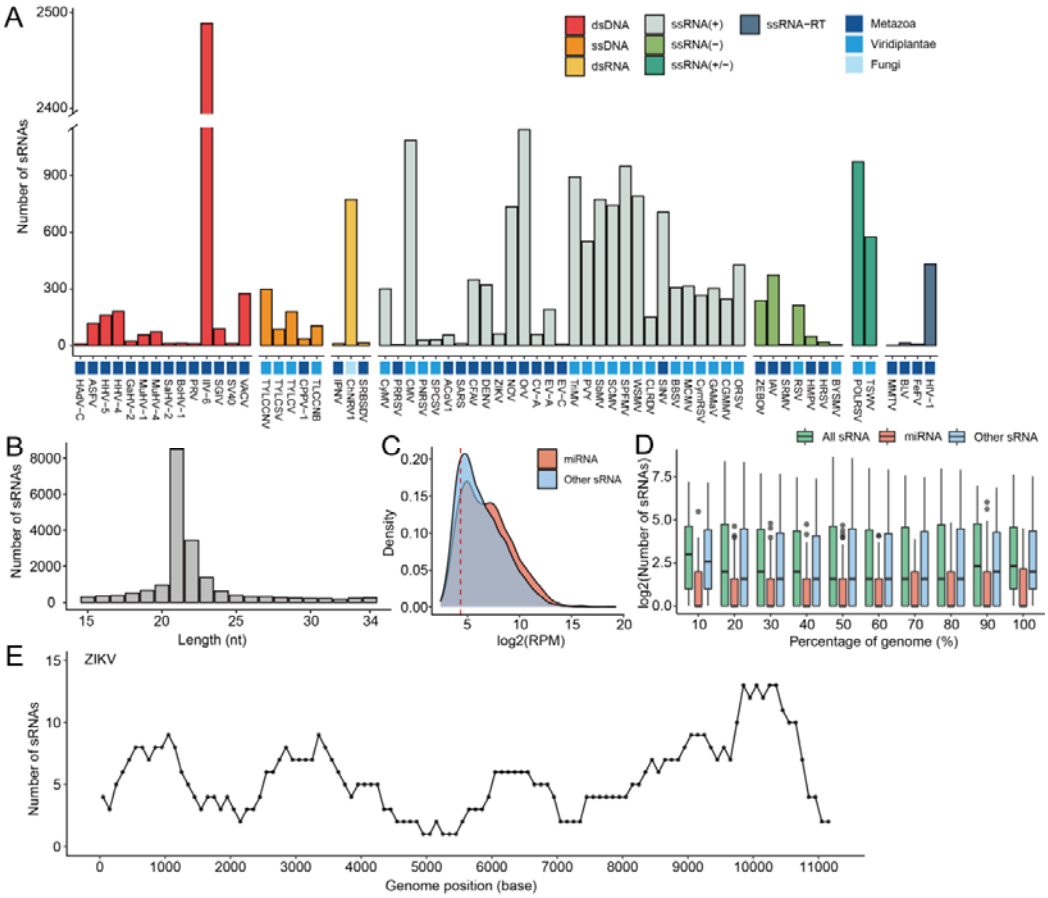

Figure 3. Data distribution of vsRNAs.

(A) The number of vsRNAs detected in each virus. The full name of each virus corresponding to the abbreviations of that was shown in Table S2.

(B) and (C) referred to the distribution of length and expression level of vsRNAs. RPM, Reads Per Million. The red dashed line in the (C) referred to 20 RPM.

(D) The number of vsRNAs located within each 10% of the viral genomes.

(E) The number of vsRNAs identified in a 1 Kb sliding-window with a step size of 100 bases along the genome of the Zika virus.

The vsRNAs were detected in viruses of 34 hosts from three kingdoms, including 20, 13 and 1 organism(s) in metazoa (animal), viridiplantae (plant) and fungi, respectively.

Therefore, the vsRNAs from viruses of different host kingdoms were analyzed. Comparing the number of vsRNAs identified in viruses of different hosts from three kingdoms showed that viruses from viridiplantae and fungi had much more vsRNAs than those from metazoa (Figure 4A). Attention to note, only one virus (Colletotrichum higginsianum non-segmented dsRNA virus 1, ChNRV1) which infected the fungi was used in the study, although it had 774 vsRNAs identified. Comparing the expression level of vsRNAs in viruses of different hosts from three kingdoms showed that the expression level of vsRNAs from viruses of viridiplantae was similar to that of vsRNAs from viruses of fungi; both of them were higher than that of vsRNAs from viruses of metazoa (Figure 4B). The length distribution of vsRNAs in viruses of metazoa was similar to that of viruses of viridiplantae (Figure 3C). Both of them had a sharp peak in 21 nt. While for viruses of fungi, the vsRNAs had gentle peaks ranging from 19 to 23 nt.

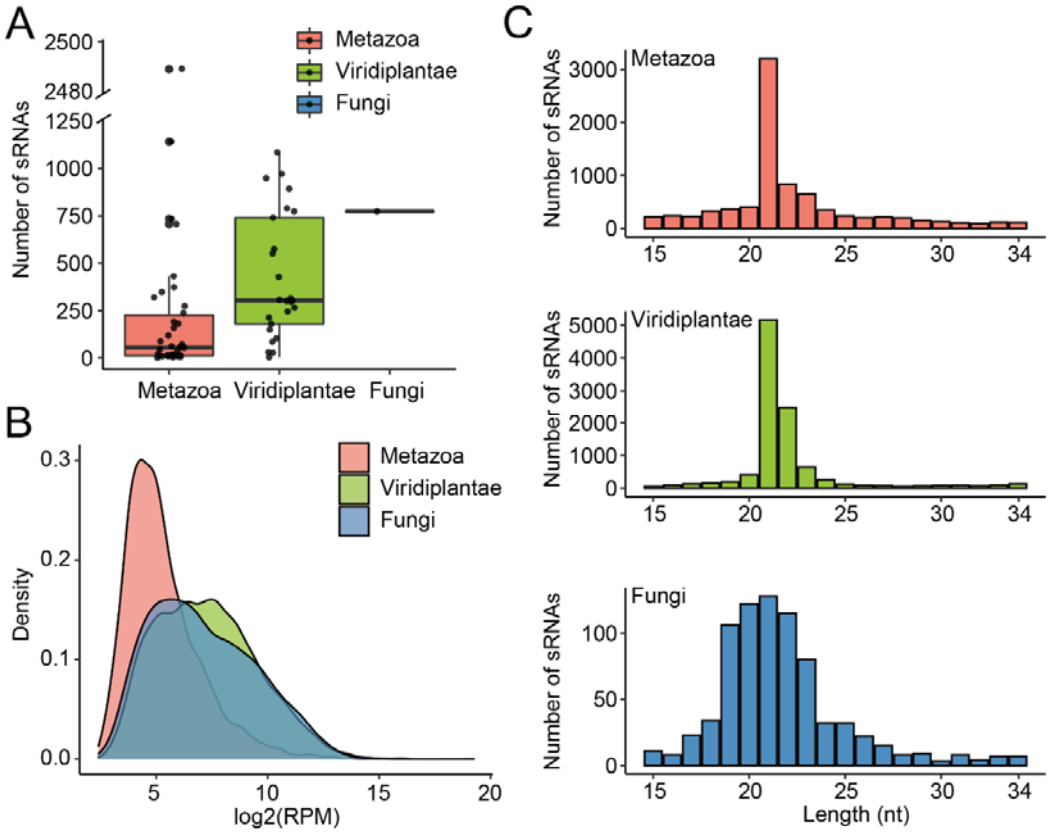

Figure 4. Analysis and comparison of vsRNAs by the viral host.

(A) The number of vsRNAs identified in viruses of different hosts from three kingdoms;

(B) and (C) referred to the distributions of the expression level and length of vsRNAs in viruses of different hosts from three kingdoms.

Dynamic expression of vsRNAs after viral infection

Then, the dynamic expression of vsRNAs after viral infection was analyzed in 17 viruses (Table S7). In most viruses, the number of vsRNAs increased as the viral infection persisted. Taking the African swine fever virus (ASFV) and influenza A virus (IAV) as examples, 0, 13 and 111 vsRNAs were detected at 6, 12 and 24 hours

post infection (hpi) for the ASFV, while for the IAV increased expression of vsRNAs was observed from 24 to 48 and 120 hpi (Figure 5A). Some viruses showed complex dynamics of vsRNA expressions after infections. For example, the number of vsRNAs from Sweet potato feathery mottle virus (SPFMV) were 133 and 157 at 3 and 6 weeks post□infection, respectively, while that decreased to 103 at 12 weeks post□infection.

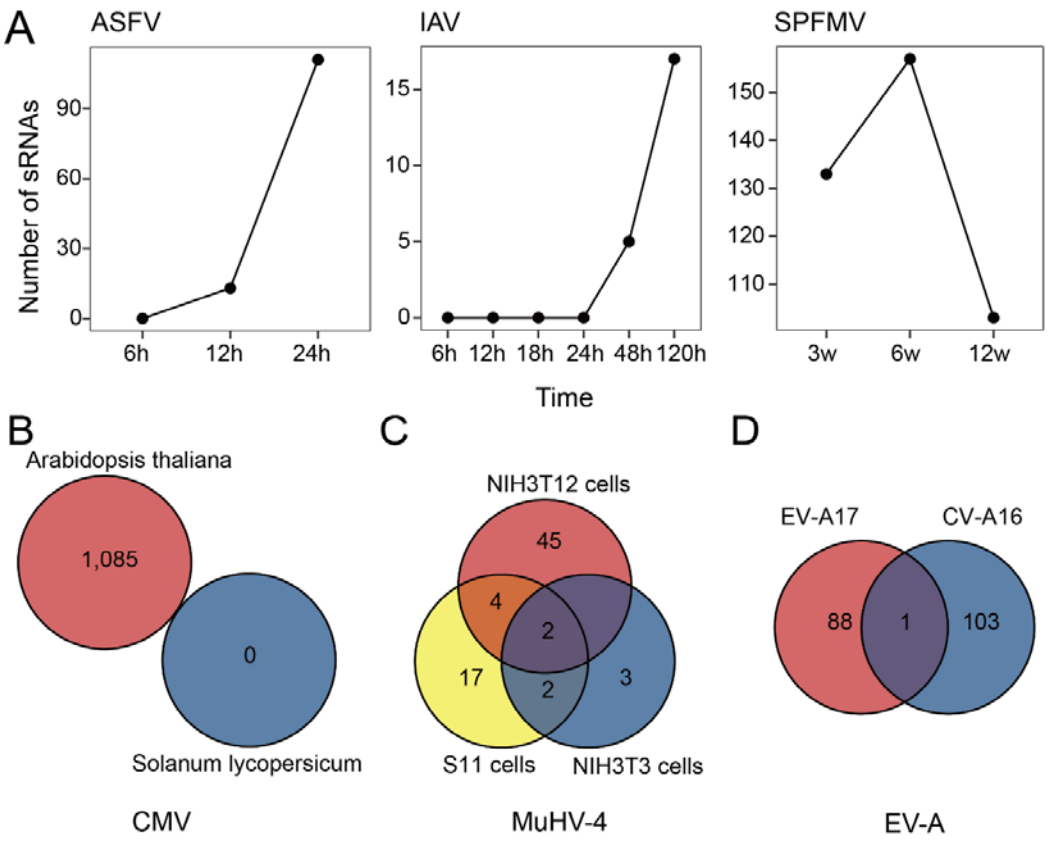

Figure 5. Analysis of the expression dynamics and specificity of vsRNAs.

(A) The number of vsRNAs detected in different time points after infection of African swine fever virus (ASFV), Influenza A virus (IAV) and Sweet potato feathery mottle virus (SPFMV).

(B)-(D) Venn diagram showing the overlap of vsRNAs detected in a given virus

infecting different hosts (B) or cell/tissue types (C), or of different strains (D).

Expression specificity of vsRNAs by the host, cell/tissue type and strain

The expression specificity of vsRNAs by the host was analyzed for four plant viruses that infected the same kind of tissue (leaves) of different hosts. For a given virus, the vsRNAs identified when infecting different hosts were analyzed and compared (Table S8). For all four viruses, only a small proportion of vsRNAs overlapped between viruses infecting different hosts. For example, 1,085 vsRNAs were detected when the Cucumber mosaic virus (CMV) infected *Arabidopsis thaliana*, while no vsRNA was detected when the virus infected *Solanum lycopersicum* (Figure 5B).

The expression specificity of vsRNAs by cell/tissue was analyzed for 11 viruses that infected two or more cells/tissues of the same host (Table S9). No or little overlap was observed between vsRNAs identified in different cells/tissues for most of these viruses. For example, the MuHV-4 had 7, 25 and 51 vsRNAs when infecting NIH3T3 cells, S11 cells, NIH3T12 cells, respectively, while only 2 vsRNAs were observed in all three cells (Figure 5C).

The expression specificity of vsRNAs by viral strain was analyzed for five viruses of which different viral strains infected the same cell or tissue (Table S10). For all five viruses, no or little overlap was observed between vsRNAs identified in different viral strains of the same virus. Notably, although the viral strains Enterovirus A71 (EV-A17) and Coxsackievirus A16 (CV-A16) of Enterovirus A (EV-A) shared approximately 80% identity in the genome, only one vsRNA overlapped between 89 and 104 vsRNAs

identified from EV-A17 and CV-A16, respectively (Figure 5D).

Conservation analysis of vsRNAs

Then, the sequence similarity of vsRNAs among different virus species was investigated using BLAST (version 2.11.0+) [32]. The vsRNAs of a given virus species were queried against those in other viruses. A pair of vsRNAs was considered to be similar if both the identity and coverage were greater than 90% in the reciprocal BLAST analysis. Only 37 pairs of similar vsRNAs with each vsRNA in viruses of the same or different viral families were observed in the analysis (Table S11). For example, 551 and 774 vsRNAs were identified in Potato virus Y and Soybean mosaic virus of the *Potyviridae* family, respectively, while only three similar vsRNA pairs were observed between the two viruses. Next, the conservation of vsRNAs in different viral strains of a given virus was analyzed. A vsRNA was considered as "conserved vsRNA" if it was encoded in at least two viral strains of the same virus; otherwise, the "non-conserved vsRNA". Unexpectedly, only 283 conserved vsRNAs were observed among 6,949 vsRNAs in 12 viruses with different viral strains (Figure S1). Only less than five conserved vsRNAs were observed in 8 of 12 viruses. For the other four viruses including Cucumber mosaic virus (CMV), Human gammaherpesvirus 4 (HHV-4), Invertebrate iridescent virus 6 (IIV-6) and Soybean mosaic virus (SbMV), the conserved vsRNAs had significantly higher expressions than non-conserved vsRNAs in viruses of HHV-4, IIV-6 and SbMV (Figure S1). For example, the conserved vsRNAs had a median of 259 RPM in the HHV-4, while the

non-conserved vsRNAs had a median of 89 RPM.

Performance analysis of vsRNAfinder in identifying sRNAs in plants and animals

To investigate the ability of vsRNAfinder in identifying sRNAs in other organisms besides viruses, the vsRNAfinder was used to identify sRNAs in *Homo sapiens* and four model organisms including *Mus musculus*, *Drosophila melanogaster*, *Arabidopsis thaliana* and *Oryza sativa*. The ability of vsRNAfinder was evaluated based on its performance in identifying miRNAs in these organisms. For comparison, miRDeep2 and ShortStack were also used to identify miRNAs in these organisms. As shown in Table 1, vsRNAfinder had Recalls ranging from 0.06 to 0.13 in humans and two animals which were a little lower than those of miRDeep2, while it had Recalls of 0.05 to 0.11 in two plants which were a little higher than those of miRDeep2. When compared to ShortStack, vsRNAfinder consistently had higher Recalls in five organisms. In terms of Precision, the ShortStack had the highest Precision among the three methods in five organisms. Although the Precision of vsRNAfinder was lower than that of ShortStack, it was consistently higher than that of miRDeep2 in five organisms. Overall, in terms of F1-score which combined both the Recall and Precision, the vsRNAfinder had an average F1-score of 0.122 which was similar to that of miRDeep2 (0.116) and was slightly higher than that of ShortStack (0.098).

Table 1. Performance comparison of three computational methods in identifying miRNAs in five organisms. Rec, Recall; Prec, Precision; F1, F1-score.

Organism	Method	Num of detected miRNA	Num of miRNAs in miRBase	Overlap of miRNA detected and those in miRBase	Rec	Prec	F1
Homo sapiens	vsRNAfinder	479	2632	157	0.06	0.33	0.1
	miRDeep2	924		213	0.08	0.23	0.12
	ShortStack	143		101	0.04	0.71	0.07
Mus musculus	vsRNAfinder	341	1963	145	0.07	0.43	0.13
	miRDeep2	524		177	0.09	0.34	0.14
	ShortStack	108		72	0.04	0.67	0.07
Drosophila melanogaster	vsRNAfinder	100	468	61	0.13	0.61	0.21
	miRDeep2	125		69	0.15	0.55	0.23
	ShortStack	51		37	0.08	0.73	0.14
Arabidopsis thaliana	vsRNAfinder	433	350	17	0.05	0.04	0.04
	miRDeep2	530		12	0.03	0.02	0.03
	ShortStack	14		9	0.03	0.64	0.05
Oryza sativa	vsRNAfinder	403	577	63	0.11	0.16	0.13
	miRDeep2	316		28	0.05	0.09	0.06
	ShortStack	116		55	0.1	0.47	0.16

Discussion

In this study, we presented vsRNAfinder, a novel method to identify and quantify vsRNAs *de novo* based on sRNA-Seq data. The vsRNAfinder was found to outperform two widely-used methods (miRDeep2 [18] and ShortStack [19]) in identifying vsRNAs, and it could also be used to identify sRNAs in multiple organisms of plants and animals with similar performances to current methods. Besides, 19,734 high-confidence vsRNAs in over 60 viral species were identified by vsRNAfinder and were further characterized, which could greatly facilitate further studies of vsRNAs.

Numerous methods have been developed to identify sRNAs from sRNA-Seq data. However, there is still a lack of an effective method for sRNA identification in viruses. The *de novo* methods such as miRDeep2 and ShortStack identified sRNAs based on reads assembly, which may merge reads from multiple adjacent sRNAs with overlaps. This became worse in viruses on whose genome the sRNAs were more densely distributed than those in other organisms. Hence, many vsRNAs may be missed by these methods. The vsRNAfinder presented in the study is also a *de novo* method and outperformed miRDeep2 and ShortStack in capturing viral miRNAs with improved Recalls. This may be explained by the following two reasons. Firstly, smoothing the coverage of mapped reads on the viral genome before the peak calling could reduce the noise of the peak detection and enhance the peak signal. Secondly, filtering

sRNAs by abundance and evaluating the statistical significance of vsRNAs based on the Poisson distribution could help identify high-confidence vsRNAs. Therefore, it is suggested to put vsRNAfinder on priority in the identification of vsRNAs.

Most methods have been developed for the identification of specific sRNA class such as miRNA, or the identification of sRNAs in a specific organism such as humans [18, 30]. Compared to previous methods, vsRNAfinder could be used to identify multiple classes of sRNAs including miRNA, piRNA and siRNA. Besides viruses, vsRNAfinder also achieved similar performances to current methods in the identification of sRNAs across several animal and plant organisms. This suggested that vsRNAfinder could be taken as a universal method for the identification of sRNAs, especially for the vsRNAs.

An increasing number of vsRNAs have been reported in recent years [8, 11]. However, most studies focused on vsRNAs from one or a few viruses [2, 13, 14, 23]. This work for the first time systematically identified and characterized vsRNAs from over 60 viral species. Most vsRNAs were detected from +ssRNA viruses, followed by dsDNA viruses. The ability of encoding sRNAs varied much among viruses. This was consistent with previous studies which showed large differences between viruses in their ability of encoding circRNAs [33]. It may reflect the large heterogeneity of viruses in producing RNA products. The plant viruses were found to express more vsRNAs than animal viruses. Besides, the abundance of vsRNAs in plant viruses was also higher than that of animal viruses, although the length distribution of vsRNAs was similar between plant and animal viruses. This may be because sRNAs in plants

are modified at their 3'-end by 2'-O-methylation which confers stability and protection from degradation, while the animals lack the mechanism [34]. The length of vsRNAs mainly focused on 21 and 22 nt which was the length of miRNAs and siRNAs. However, only 14% of vsRNAs were annotated as miRNAs. Unexpectedly, these miRNAs were extensively observed in RNA viruses that were reported to replicate in the cytoplasm and have no access to the miRNA biogenesis machinery located in the nucleus [35]. These miRNAs may be generated by the flexible or noncanonical pathways of miRNA biogenesis. For example, several viral miRNAs observed in infection of the Ebola virus (a negative-sense RNA virus) in humans were likely to be processed by the cytoplasmic Drosha [35]. Further studies are needed to verify the existence of viral miRNAs in RNA viruses and to clarify the mechanism of their biogenesis.

Strong expression specificity was observed for viral RNAs such as circRNAs [33]. Consistent with previous studies, we also found strong expression specificity of vsRNAs by the host, cell/tissue and viral strain [33]. For most viruses analyzed, no or little overlap was observed between vsRNAs identified in a given virus of different strains, or a virus infecting different hosts or cells/tissues. Besides, the dynamic expression of vsRNAs in different time points after viral infections was also observed. This suggested that the biogenesis of vsRNAs was highly specific and dynamic. It may be related to the complex and dynamic interactions between viruses and hosts. Viruses may use vsRNAs to fight against the host immunity and regulate the transcription and translation of both virus and host. For example, the Rice stripe virus

was reported to generate lots of host-specific viral siRNAs in rice and planthopper which potentially control defensive responses in hosts [36].

There are some limitations to the study. Firstly, only miRNAs were annotated in vsRNAfinder as there were few effective methods of annotating other vsRNA types such as piRNA and siRNA. For example, viral piRNAs were reported to lack the characteristic ping-pong signature that is common in piRNAs of other organisms, which makes it difficult to distinguish between viral piRNA and siRNA [37, 38]. More computational methods are needed to annotate vsRNAs. Secondly, the reads that mapped to more than two locations of the genome were excluded to reduce false positives in vsRNAfinder, which may lower the abundance of vsRNAs. Fortunately, only 0.007% of reads were mapped to more than two locations in the viral genomes as the viral genomes generally have a low level of sequence repeating. Thirdly, although lots of high-confidence vsRNAs were identified in over 60 viral species, the list of vsRNAs encoded by these viruses is far from complete considering the expression specificity and dynamics of vsRNAs. Further efforts are still needed to identify more vsRNAs, and to verify and characterize the vsRNAs identified.

Conclusion

Overall, this study developed a state-of-the-art method named vsRNAfinder which can not only be used for identifying high-confidence sRNAs in viruses with better performances than current methods, but also be used for sRNA identification in plants and animals. Besides, a total of 19,734 high-confidence vsRNAs identified in 64 viral

species provided a rich resource for further studies of vsRNAs.

Supplementary information

Additional file 1. Supplementary figure 1 and table S1-S4, S6-S11.

Additional file 2. Supplementary table S5.

Availability of data and materials

All data used in the study are available in supplementary materials. The vsRNAfinder is available at https://github.com/ZenaCai/vsRNAfinder.

Authors' contributions

Zena Cai and Ping Fu contributed equally to this work. Zena Cai developed the method, analyzed the data and wrote the draft; Ping Fu analyzed the data; Ye Qiu, Aiping Wu, Gaihua Zhang, Yirong Wang, Taijiao Jiang, Xing-Yi Ge, Haizhen Zhu and Yousong Peng supervised the project, reviewed and revised the manuscript; Yousong Peng designed the project, wrote and revised the manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (32170651) and Hunan Provincial Natural Science Foundation of China (2020JJ3006).

Acknowledgements

We thank people in the PengLab for helpful discussions on the manuscript.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

1. Romano G, Veneziano D, Acunzo M, Croce CM: **Small non-coding RNA and cancer.** *Carcinogenesis* 2017, **38:**485-491.
2. Liu C, Li F, Sun Y, Zhang X, Yuan J, Yang H, Xiang J: **Virus-derived small RNAs in the penaeid shrimp Fenneropenaeus chinensis during acute infection of the DNA virus WSSV.** *Sci Rep* 2016, **6:**28678.
3. Wahid F, Shehzad A, Khan T, Kim YY: **MicroRNAs: synthesis, mechanism, function, and recent clinical trials.** *Biochim Biophys Acta* 2010, **1803:**1231-1243.

4. Wilson RC, Doudna JA: **Molecular mechanisms of RNA interference.** *Annu Rev Biophys* 2013, **42:**217-239.

5. Neeb ZT, Zahler AM: **An expanding world of small RNAs.** *Dev Cell* 2014, **28:**111-112.

6. Jeang KT: **RNAi in the regulation of mammalian viral infections.** *BMC Biol* 2012, **10:**58.

7. Acunzo M, Visone R, Romano G, Veronese A, Lovat F, Palmieri D, Bottoni A, Garofalo M, Gasparini P, Condorelli G, et al: **miR-130a targets MET and induces TRAIL-sensitivity in NSCLC by downregulating miR-221 and 222.** *Oncogene* 2012, **31:**634-642.

8. Kozomara A, Birgaoanu M, Griffiths-Jones S: **miRBase: from microRNA sequences to function.** *Nucleic Acids Res* 2019, **47:**D155-d162.

9. Rahman RU, Liebhoff AM, Bansal V, Fiosins M, Rajput A, Sattar A, Magruder DS, Madan S, Sun T, Gautam A, et al: **SEAweb: the small RNA Expression Atlas web application.** *Nucleic Acids Res* 2020, **48:**D204-d219.

10. Feng L, Zhang F, Zhang H, Zhao Y, Meyers BC, Zhai J: **An Online Database for Exploring Over 2,000 Arabidopsis Small RNA Libraries.** *Plant Physiol* 2020, **182:**685-691.

11. Kincaid RP, Sullivan CS: **Virus-encoded microRNAs: an overview and a look to the future.** *PLoS Pathog* 2012, **8:**e1003018.

12. Weng KF, Hung CT, Hsieh PT, Li ML, Chen GW, Kung YA, Huang PN, Kuo RL, Chen LL, Lin JY, et al: **A cytoplasmic RNA virus generates functional**

viral small RNAs and regulates viral IRES activity in mammalian cells. *Nucleic Acids Res* 2014, **42:**12789-12805.

13. Hussain M, Asgari S: **MicroRNA-like viral small RNA from Dengue virus 2 autoregulates its replication in mosquito cells.** *Proc Natl Acad Sci U S A* 2014, **111:**2746-2751.

14. Morales L, Oliveros JC, Fernandez-Delgado R, tenOever BR, Enjuanes L, Sola I: **SARS-CoV-Encoded Small RNAs Contribute to Infection-Associated Lung Pathology.** *Cell Host Microbe* 2017, **21:**344-355.

15. Aparicio-Puerta E, Lebrón R, Rueda A, Gómez-Martín C, Giannoukakos S, Jaspez D, Medina JM, Zubkovic A, Jurak I, Fromm B, et al: **sRNAbench and sRNAtoolbox 2019: intuitive fast small RNA profiling and differential expression.** *Nucleic Acids Res* 2019, **47:**W530-w535.

16. Handzlik JE, Tastsoglou S, Vlachos IS, Hatzigeorgiou AG: **Manatee: detection and quantification of small non-coding RNAs from next-generation sequencing data.** *Sci Rep* 2020, **10:**705.

17. Wu X, Kim TK, Baxter D, Scherler K, Gordon A, Fong O, Etheridge A, Galas DJ, Wang K: **sRNAnalyzer-a flexible and customizable small RNA sequencing data analysis pipeline.** *Nucleic Acids Res* 2017, **45:**12140-12151.

18. Friedländer MR, Mackowiak SD, Li N, Chen W, Rajewsky N: **miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades.** *Nucleic Acids Res* 2012, **40:**37-52.

19. Shahid S, Axtell MJ: **Identification and annotation of small RNA genes**

using ShortStack. *Methods* 2014, **67:**20-27.

20. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, et al: **NCBI GEO: archive for functional genomics data sets--update.** *Nucleic Acids Res* 2013, **41:**D991-995.

21. Kodama Y, Shumway M, Leinonen R: **The Sequence Read Archive: explosive growth of sequencing data.** *Nucleic Acids Res* 2012, **40:**D54-56.

22. Xia Z, Peng J, Li Y, Chen L, Li S, Zhou T, Fan Z: **Characterization of small interfering RNAs derived from Sugarcane mosaic virus in infected maize plants by deep sequencing.** *PLoS One* 2014, **9:**e97013.

23. Wang Y, Jin B, Liu P, Li J, Chen X, Gu J: **piRNA Profiling of Dengue Virus Type 2-Infected Asian Tiger Mosquito and Midgut Tissues.** *Viruses* 2018, **10**.

24. Martin M: **Cutadapt removes adapter sequences from high-throughput sequencing reads.** *2011* 2011, **17:**3 %J EMBnet.journal.

25. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome Biol* 2009, **10:**R25.

26. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H: **Twelve years of SAMtools and BCFtools.** *Gigascience* 2021, **10**.

27. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing**

genomic features. *Bioinformatics* 2010, **26:**841-842.

28. Borchers HW: **pracma: Practical Numerical Math Functions.** 2022.

29. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szcześniak MW, Gaffney DJ, Elo LL, Zhang X, Mortazavi A: **A survey of best practices for RNA-seq data analysis.** *Genome Biol* 2016, **17:**13.

30. Hackenberg M, Rodríguez-Ezpeleta N, Aransay AM: **miRanalyzer: an update on the detection and analysis of microRNAs in high-throughput sequencing experiments.** *Nucleic Acids Res* 2011, **39:**W132-138.

31. Lorenz R, Bernhart SH, Höner Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL: **ViennaRNA Package 2.0.** *Algorithms Mol Biol* 2011, **6:**26.

32. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.

33. Cai Z, Fan Y, Zhang Z, Lu C, Zhu Z, Jiang T, Shan T, Peng Y: **VirusCircBase: a database of virus circular RNAs.** *Brief Bioinform* 2021, **22:**2182-2190.

34. Borges F, Martienssen RA: **The expanding world of small RNAs in plants.** *Nat Rev Mol Cell Biol* 2015, **16:**727-741.

35. Nanbo A, Furuyama W, Lin Z: **RNA Virus-Encoded miRNAs: Current Insights and Future Challenges.** *Front Microbiol* 2021, **12:**679210.

36. Yang M, Xu Z, Zhao W, Liu Q, Li Q, Lu L, Liu R, Zhang X, Cui F: **Rice stripe virus-derived siRNAs play different regulatory roles in rice and in the insect vector Laodelphax striatellus.** *BMC Plant Biol* 2018, **18:**219.

37. Varjak M, Donald CL, Mottram TJ, Sreenu VB, Merits A, Maringer K,

Schnettler E, Kohl A: **Characterization of the Zika virus induced small RNA response in Aedes aegypti cells.** *PLoS Negl Trop Dis* 2017, **11:**e0006010.

38. Öhlund P, Hayer J, Hesson JC, Blomström AL: **Small RNA Response to Infection of the Insect-Specific Lammi Virus and Hanko Virus in an Aedes albopictus Cell Line.** *Viruses* 2021, **13**.