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. 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 miRNAs with a significantly improved sensitivity. It can also be used to identify sRNAs in animals and plants with similar performance to miRDeep2 and ShortStack. The study would greatly facilitate effective identification 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

45 perfectly to their mRNA targets; the piRNAs can degrade mRNAs by binding
46 perfectly their mRNA targets with the help of PIWI protein[5, 6]. The sRNAs play a
47 fundamental role in the maintenance of normal physiology and serve as potential
48 biomarkers for early diagnosis and therapeutic targets for a wide spectrum of diseases,
49 particularly the cancer[1]. For example, miR-130 can down-regulate the expression of
50 miRNA-221 and miRNA-222 by targeting the MET, thus reducing TNF-related
51 apoptosis-induced ligand (TRAIL) resistance in non-small cell lung cancer[7].
52 Numerous methods have been developed to identify sRNAs from sRNA-Seq data.
53 They can be mainly divided into two kinds. The first kind of method is to identify
54 sRNAs based on a known sRNA library, such as the sRNAbench[8], Manatee[9] and
55 sRNAnalyzer[10]. This kind of method can be used for accurate quantification of
56 known sRNA. However, they may not be suitable for organisms with only a few
57 known sRNAs such as viruses. The other kind of method is to identify sRNAs *de*
58 *novo* by firstly finding the precursor of sRNAs from assembly of reads and then
59 detecting the mature sRNA, such as miRDeep2[11] and ShortStack[12]. Most
60 methods mentioned above have been mainly used in the identification of sRNAs in
61 animals and plants. The viruses have much smaller genomes and stronger coding
62 ability compared to animals and plants. Thus, the distribution of vsRNAs is much
63 denser than that of other organisms, which may lead to assembly of adjacent vsRNAs
64 with overlaps into long contigs and missed detection of many vsRNAs by current
65 methods.
66 Herein, we presented vsRNAfinder, a *de novo* method for the identification of

67 high-confidence vsRNAs from sRNA-Seq data based on peak calling and Poisson
68 distribution. The method can also be used in the identification of sRNAs in animals
69 and plants. The work would greatly facilitate the identification of vsRNAs.
70

71 **Materials and methods**

72 ***Data collection***

73 To evaluate the performance of computational methods in identifying viral miRNAs, a
74 total of 196 known miRNAs in nine viruses were obtained from the miRBase
75 database (Release 22.1, http://mirbase.org/) (Table S1). The viral-infection-related
76 sRNA-Seq data of these viruses were obtained from the NCBI GEO[13] and SRA[14]
77 databases. Only the sRNA-Seq data with the study type of "Expression profiling by
78 high throughput sequencing" or "Non-coding RNA profiling by high throughput
79 sequencing" in GEO database, and those with the strategy of "RNASeq" or "other" in
80 SRA database were kept. Then, the sRNA-Seq data were further filtered manually.
81 Only those of viral infections with the genome sequence of the virus used in the
82 infection available were kept for further analysis. A total of 14 sRNA-Seq datasets for
83 nine viruses were used in the study (Table S1).
84 To evaluate the performance of computational methods in identifying viral siRNAs
85 and piRNAs, the experimentally-verified siRNAs in Sugarcane mosaic virus and
86 piRNAs in Dengue viruses were collected from the studies of Xia et al.[15] and Wang
87 et al.[16], respectively. The sRNA-Seq datasets of these two viruses were obtained
88 from the NCBI GEO and SRA database as mentioned above (Table S2).

89 To demonstrate the usage of our method in the identification of sRNAs from other
90 organisms, the known miRNAs in several widely-studied organisms including *Homo*
91 *sapiens*, *Mus musculus*, *Drosophila melanogaster*, *Arabidopsis thaliana* and *Oryza*
92 *sativa* were obtained from the database of miRBase. Some representative sRNA-Seq
93 datasets of these organisms were also obtained from NCBI GEO and SRA databases
94 (Table S3).
95
96 ***The vsRNAfinder workflow***
97 The vsRNAfinder workflow included four modules: Preprocessing, Identification,
98 Filtering and annotation, and Quantification, which were described as follows.
99 ***Preprocessing module.*** Firstly, the adapter sequences are trimmed from the raw reads
100 with cutadapt (version 2.6, parameter: '-m 15')[17] (Figure 1). Then, the clean reads
101 are mapped to the viral genome directly (the default), or the clean reads are firstly
102 mapped to the host genome and then the unmapped reads are mapped to the viral
103 genome, with bowtie (version 1.3.0, parameter: '-v 0 -m 2')[18]. The resulted SAM
104 files are converted to BAM files which are sorted by samtools (version 1.9)[19]. The
105 sorted BAM files are further converted to BED files by bedtools (version 2.29.2)[20].
106 ***Identification module***. The coverage of mapped reads on the viral genome is
107 calculated by bedtools and then smoothed with the window-size of 11 nt. The peaks of
108 the smoothed curve are identified using the function of *findPeaks* ("minpeakheight =
109 5") in the R package pracma (version 2.3.3)[21]. The start and end site of the
110 candidate sRNA is restricted to the upstream and downstream 30 nt of the peak,

111 respectively. Then, the number of reads starting at each position in the upstream
112 region of the peak, and that ending at each position in the downstream of the peak are
113 calculated. The position with the highest number of reads in the upstream and that in
114 the downstream are considered as the start and end site of the candidate sRNA,
115 respectively.

116 ***Filtering and annotation module.*** To obtain the high-confidence sRNAs, the
117 abundances of the start and end sites (defined as boundary sites) of the candidate
118 sRNA are normalized using RPM (Reads Per Million) which is calculated by dividing
119 the number of reads starting at the start site or ending at the end site by the total
120 number of clean reads, and then multiplied by 10^6. Only the candidate sRNAs with
121 both of the boundary sites having at least 5 RPM and with lengths ranging from 15 to
122 34 nt are retained for further analysis.

123 Then, the Poisson distribution is used to evaluate the statistical significance of
124 candidate sRNAs. Poisson distribution is a discrete probability distribution that
125 expresses the probability of a given number of events occurring in a fixed interval of
126 time, which is widely applied in NGS data analysis[22]. We hypothesized that the
127 reads should be enriched in the region where the sRNA is expressed compared to the
128 background region. Thus, the significance of candidate sRNAs could be analyzed by
129 the Poisson distribution of which the parameter λ can be calculated as the average
130 coverage of the mapped reads on the background region. For each candidate sRNA,
131 the background region is defined as extending the upstream and downstream 10 nt of
132 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 Hackenberg et al.'s study[23]:

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

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

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

149 is set to be 15; *step* and *i* are set to be 5 and 8 for animals and viruses, 7 and 10 for
150 plants as plant pre-miRNA sequences are longer than those of animals and viruses,
151 according to Hackenberg et al.'s study[23].
152 Then, the secondary structure of the hypothesized pre-miRNA is predicted by
153 RNAfold (version 2.4.18)[24] with default parameters. The hypothesized pre-miRNA
154 which has the minimum free energy of less than -25 kcal/mol, and in which at least
155 70% of positions in the corresponding sRNA form base pairs, is considered to be
156 pre-miRNA having the hairpin structure, and is further processed into the mature
157 miRNA.
158 ***Quantification module.*** The abundances of sRNAs are normalized using RPM. The
159 total number of reads mapping to the sRNA is calculated as the sum of reads that
160 overlap with the sRNA by 80% and those that are utilized to estimate boundary sites
161 of the sRNA. The RPM of each sRNA is calculated by dividing the total number of
162 reads mapping to the sRNA by the total number of clean reads in the sample, and then
163 multiplied by 10^6.
164
165 ***Prediction of sRNAs with miRDeep2 and ShortStack***
166 The miRDeep2 (version 0.1.2, https://github.com/rajewsky-lab/mirdeep2)[11] and
167 ShortStack (version 3.8.5, https://github.com/MikeAxtell/ShortStack)[12] were used
168 for prediction of sRNAs with default parameters.
169
170 ***Evaluating the performance of computational methods in identifying sRNAs***

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

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

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

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

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

181

182 ***Statistical analysis***

183 The Wilcoxon signed-rank test was used to compare the F1-score between the
184 vsRNAfinder and the other methods. A p-value less than 0.05 was considered as
185 statistically significant. All statistical analyses were conducted in Python (version
186 3.8).

187

188 **Results**

189 ***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.

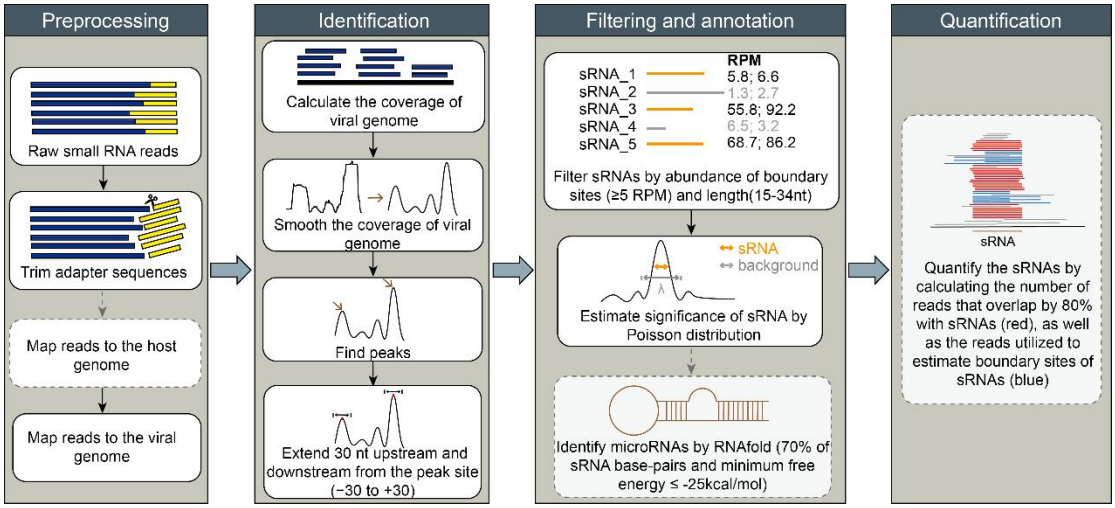


Figure 1. Overview of vsRNAfinder which consisted of four modules including Preprocessing, Identification, Filtering and annotation, and Quantification. Please see Materials and methods for details.

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

207 and methods).

208 The Filtering and annotation module was designed for obtaining high-confidence
209 sRNAs and annotating miRNAs. The candidate sRNAs identified above were firstly
210 filtered by the abundance of the start and end sites (defined as boundary sites) of
211 sRNAs. Only the candidate sRNAs with both boundary sites having at least 5 RPM
212 were kept. Then, they were further filtered by the length. Only the candidate sRNAs
213 with 15-34 nt were kept which covered the length of most miRNAs, siRNAs and
214 piRNAs[2]. Then, the statistical significance of candidate sRNAs was obtained based
215 on the Poisson distribution. The remaining candidate sRNAs with a p-value smaller
216 than 0.05 were considered as statistically significant and highly confident. Finally, the
217 RNAfold[24] was used to annotate miRNAs from the high-confidence sRNAs (see
218 Materials and methods).

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

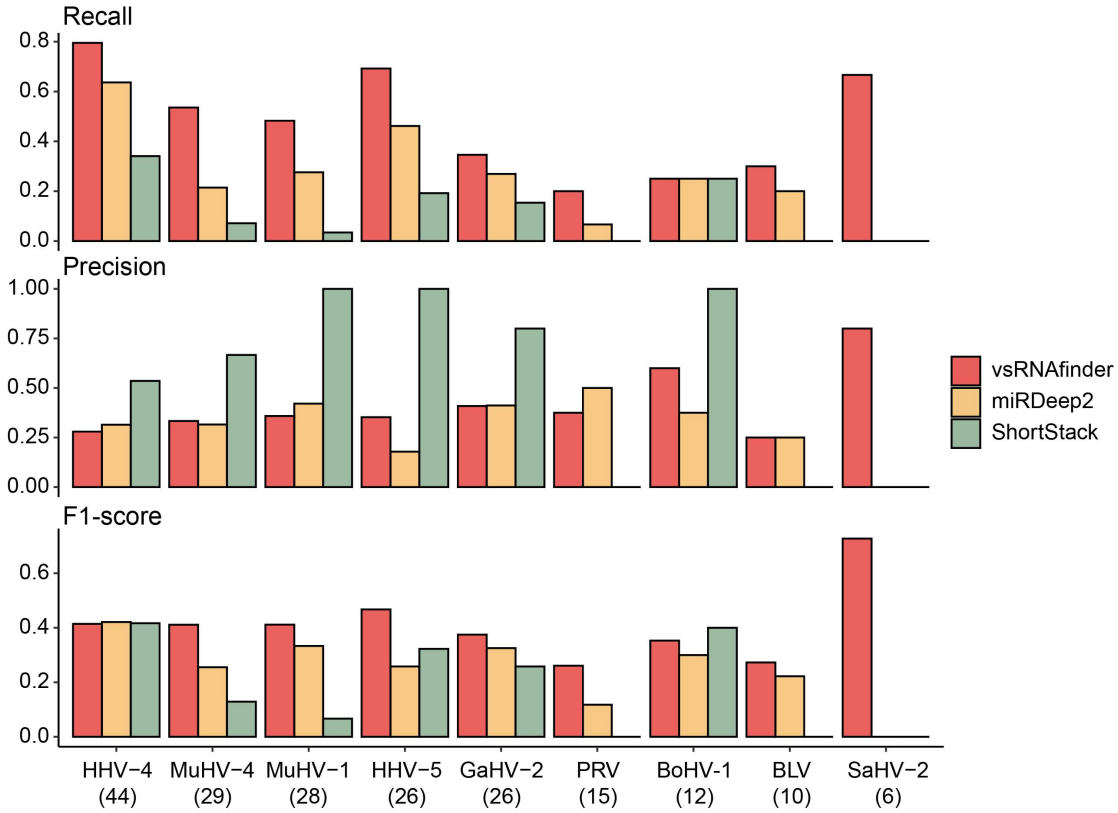
223

224 ***Performance analysis of vsRNAfinder in identifying viral miRNAs***

225 To evaluate the performance of vsRNAfinder in identifying vsRNAs, vsRNAfinder
226 was used to identify viral miRNAs since the miRNAs are the most studied vsRNAs.
227 Nine viruses which had a range of 6-44 miRNAs reported in miRBase were used in
228 the evaluation (Figure 2). For comparison, two *de novo* methods, namely the

229 miRDeep2[11] and ShortStack[12] which are widely used in identifying sRNAs, were

230 selected in identifying viral miRNAs in nine viruses.



231

232 **Figure 2.** Performance analysis of vsRNAfinder, miRDeep2 and ShortStack in

233 identifying viral miRNAs. Comparison of Recall, Precision and F1-score of

234 vsRNAfinder with miRDeep2 and ShortStack in identifying viral miRNAs in viruses

235 of Human gammaherpesvirus 4 (HHV-4), Murid gammaherpesvirus 4 (MuHV-4),

236 Murid betaherpesvirus 1 (MuHV-1), Human betaherpesvirus 5 (HHV-5), Gallid

237 alphaherpesvirus 2 (GaHV-2), Suid alphaherpesvirus 1 (PRV), Bovine

238 alphaherpesvirus 1 (BoHV-1), Bovine leukemia virus (BLV) and Saimiriine

239 gammaherpesvirus 2 (SaHV-2). The numbers in parentheses under the viral name

240 referred to the number of miRNAs reported in miRBase.

241

242 As shown in Figure 2, vsRNAfinder achieved the highest Recall ranging from 0.2 to
243 0.8 in all viruses, followed by miRDeep2 (0-0.64) and ShortStack (0-0.34). Taking the
244 MuHV-4 as an example, vsRNAfinder had a Recall of 0.54, while miRDeep2 and
245 ShortStack had a Recall of 0.21 and 0.07, respectively. In terms of Precision, the
246 ShortStack performed best with Precision higher than 0.5 in all viruses except three
247 viruses for which no miRNAs were identified by ShortStack. The vsRNAfinder had
248 Precisions ranging from 0.25-0.8 which were similar to those of miRDeep2. Overall,
249 when considering both the Recall and Precision using the F1-score, the vsRNAfinder
250 achieved the best F1-scores in all viruses except HHV-4 and BoHV-1. It had a median
251 F1-score of 0.41 which was significantly larger than that of miRDeep2 (0.26, p-value
252 = 0.0098) and ShortStack (0.13, p-value = 0.0039).
253 When comparing the number of viral miRNAs identified by three methods, we found
254 that most viral miRNAs identified by miRDeep2 and ShortStack were also identified
255 by vsRNAfinder (Table S4). For example, 19 and 3 viral miRNAs were identified by
256 miRDeep2 and ShortStack, respectively, in the MuHV-4, among which 14 and 3 were
257 also identified by vsRNAfinder. Moreover, vsRNAfinder identified more novel viral
258 miRNAs than the other two methods in most viruses (Table S4).
259
260 ***Identification of other classes of sRNAs with vsRNAfinder***
261 Besides viral miRNAs, viruses can also express siRNAs and piRNAs during viral
262 infections[15, 16]. Although vsRNAfinder could not annotate siRNAs and piRNAs
263 from vsRNAs, it could identify them from sRNA-Seq data. To evaluate the

264 performance of vsRNAfinder in the identification of other vsRNA classes, 11
265 experimentally-verified siRNAs in Sugarcane mosaic virus (SCMV)[15] and 6
266 experimentally-verified piRNAs in Dengue virus (DENV)[16] were collected from
267 literature. All SCMV siRNAs and four of six DENV piRNAs were identified correctly
268 by vsRNAfinder (Table S5). The remaining two DENV piRNAs were missed by
269 vsRNAfinder due to low abundance. For comparison, ShortStack only identified
270 correctly one SCMV siRNA and two DENV piRNAs (Table S5), while miRDeep2
271 can only be used to identify the miRNA.
272
273 ***Performance analysis of vsRNAfinder in identifying sRNAs in plants and animals***
274 To investigate the ability of vsRNAfinder in identifying sRNAs in other organisms
275 besides viruses, the vsRNAfinder was used to identify sRNAs in *Homo sapiens* and
276 four model organisms including *Mus musculus*, *Drosophila melanogaster*,
277 *Arabidopsis thaliana* and *Oryza sativa*. The ability of vsRNAfinder was evaluated
278 based on its performance in identifying miRNAs in these organisms. For comparison,
279 miRDeep2 and ShortStack were also used to identify miRNAs in these organisms. As
280 shown in Table 1, vsRNAfinder had Recalls ranging from 0.06 to 0.13 in humans and
281 two animals which were a little lower than those of miRDeep2, while it had Recalls of
282 0.05 to 0.11 in two plants which were a little higher than those of miRDeep2. When
283 compared to ShortStack, vsRNAfinder consistently had higher Recalls in five
284 organisms. In terms of Precision, the ShortStack had the highest Precision among the
285 three methods in five organisms. Although the Precision of vsRNAfinder was lower

286 than that of ShortStack, it was consistently higher than that of miRDeep2 in five
287 organisms. Overall, in terms of F1-score which combined both the Recall and
288 Precision, the vsRNAfinder had an average F1-score of 0.122 which was similar to
289 that of miRDeep2 (0.116) and was slightly higher than that of ShortStack (0.098).

290

291 **Table 1.** Performance comparison of three computational methods in identifying
292 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

293

294

295 **Discussion**

296 Numerous methods have been developed to identify sRNAs from sRNA-Seq data.
297 However, there is still a lack of an effective method for sRNA identification in viruses.

298 The *de novo* methods such as miRDeep2 and ShortStack identified sRNAs based on
299 reads assembly, which may merge reads from multiple adjacent sRNAs with overlaps.
300 This became worse in viruses on whose genome the sRNAs were more densely
301 distributed than those in other organisms. Hence, many vsRNAs may be missed by
302 these methods. The vsRNAfinder presented in the study is also a *de novo* method and
303 outperformed miRDeep2 and ShortStack in capturing viral miRNAs with improved
304 Recalls. This may be explained by the following two reasons. Firstly, smoothing the
305 coverage of mapped reads on the viral genome before the peak calling could reduce
306 the noise of the peak detection and enhance the peak signal. Secondly, filtering
307 sRNAs by abundance and evaluating the statistical significance of vsRNAs based on
308 the Poisson distribution could help identify high-confidence vsRNAs. Therefore, it is
309 suggested to put vsRNAfinder on priority in the identification of vsRNAs.
310 Most methods have been developed for the identification of specific sRNA class such
311 as miRNA, or the identification of sRNAs in a specific organism such as humans[11,
312 23]. Compared to previous methods, vsRNAfinder could be used to identify multiple
313 classes of sRNAs including miRNA, piRNA and siRNA. Besides viruses,
314 vsRNAfinder also achieved similar performances to current methods in the
315 identification of sRNAs across several animal and plant organisms. This suggested
316 that vsRNAfinder could be taken as a universal method for the identification of
317 sRNAs, especially for the vsRNAs.
318 There are some limitations to the study. Firstly, only miRNAs were annotated in
319 vsRNAfinder as there were few effective methods of annotating other vsRNA types

320 such as piRNA and siRNA. For example, viral piRNAs were reported to lack the
321 characteristic ping-pong signature that is common in piRNAs of other organisms,
322 which makes it difficult to distinguish between viral piRNA and siRNA[25, 26]. More
323 computational methods are needed to annotate vsRNAs. Secondly, the reads that
324 mapped to more than two locations of the genome were excluded to reduce false
325 positives in vsRNAfinder, which may lower the abundance of vsRNAs. Fortunately,
326 only 0.007% of reads were mapped to more than two locations in the viral genomes as
327 the viral genomes generally have a low level of sequence repeating.
328

329 **Conclusion**

330 Overall, this study developed a state-of-the-art method named vsRNAfinder which
331 can not only be used for identifying high-confidence sRNAs in viruses with better
332 performances than current methods, but also be used for sRNA identification in plants
333 and animals.
334

335 **Availability of data and materials**

336 All data used in the study are available in supplementary materials. The vsRNAfinder
337 is available at https://github.com/ZenaCai/vsRNAfinder.
338

339 **Authors' contributions**

340 Zena Cai and Ping Fu contributed equally to this work. Zena Cai developed the
341 method, analyzed the data and wrote the draft; Ping Fu analyzed the data; Ye Qiu,

342 Aiping Wu, Gaihua Zhang, Yirong Wang, Taijiao Jiang, Xing-Yi Ge, Haizhen Zhu and
343 Yousong Peng supervised the project, reviewed and revised the manuscript; Yousong
344 Peng designed the project, wrote and revised the manuscript.
345

346 **Funding**

347 This work was supported by the National Natural Science Foundation of China
348 (32170651), Hunan Provincial Natural Science Foundation of China (2020JJ3006),
349 and Double-First Class Construction Funds of Hunan university (521119400156).
350

351 **Acknowledgements**

352 We thank people in the PengLab for helpful discussions on the manuscript.
353

354 **References**

355 1. Romano G, Veneziano D, Acunzo M et al. Small non-coding RNA and cancer,
356 Carcinogenesis 2017;38:485-491.
357 2. Liu C, Li F, Sun Y et al. Virus-derived small RNAs in the penaeid shrimp
358 Fenneropenaeus chinensis during acute infection of the DNA virus WSSV, Sci Rep
359 2016;6:28678.
360 3. Wahid F, Shehzad A, Khan T et al. MicroRNAs: synthesis, mechanism, function,
361 and recent clinical trials, Biochim Biophys Acta 2010;1803:1231-1243.
362 4. Wilson RC, Doudna JA. Molecular mechanisms of RNA interference, Annu Rev
363 Biophys 2013;42:217-239.
364 5. Neeb ZT, Zahler AM. An expanding world of small RNAs, Dev Cell
365 2014;28:111-112.
366 6. Jeang KT. RNAi in the regulation of mammalian viral infections, BMC Biol
367 2012;10:58.
368 7. Acunzo M, Visone R, Romano G et al. miR-130a targets MET and induces
369 TRAIL-sensitivity in NSCLC by downregulating miR-221 and 222, Oncogene
370 2012;31:634-642.
371 8. Aparicio-Puerta E, Lebrón R, Rueda A et al. sRNAbench and sRNAtoolbox 2019:

372 intuitive fast small RNA profiling and differential expression, Nucleic Acids Res
373 2019;47:W530-w535.
374 9. Handzlik JE, Tastsoglou S, Vlachos IS et al. Manatee: detection and
375 quantification of small non-coding RNAs from next-generation sequencing data, Sci
376 Rep 2020;10:705.
377 10. Wu X, Kim TK, Baxter D et al. sRNAnalyzer-a flexible and customizable small
378 RNA sequencing data analysis pipeline, Nucleic Acids Res 2017;45:12140-12151.
379 11. Friedländer MR, Mackowiak SD, Li N et al. miRDeep2 accurately identifies
380 known and hundreds of novel microRNA genes in seven animal clades, Nucleic Acids
381 Res 2012;40:37-52.
382 12. Shahid S, Axtell MJ. Identification and annotation of small RNA genes using
383 ShortStack, Methods 2014;67:20-27.
384 13. Barrett T, Wilhite SE, Ledoux P et al. NCBI GEO: archive for functional
385 genomics data sets--update, Nucleic Acids Res 2013;41:D991-995.
386 14. Kodama Y, Shumway M, Leinonen R. The Sequence Read Archive: explosive
387 growth of sequencing data, Nucleic Acids Res 2012;40:D54-56.
388 15. Xia Z, Peng J, Li Y et al. Characterization of small interfering RNAs derived
389 from Sugarcane mosaic virus in infected maize plants by deep sequencing, PLoS One
390 2014;9:e97013.
391 16. Wang Y, Jin B, Liu P et al. piRNA Profiling of Dengue Virus Type 2-Infected
392 Asian Tiger Mosquito and Midgut Tissues, Viruses 2018;10.
393 17. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing
394 reads, 2011 2011;17:3 %J EMBnet.journal.
395 18. Langmead B, Trapnell C, Pop M et al. Ultrafast and memory-efficient alignment
396 of short DNA sequences to the human genome, Genome Biol 2009;10:R25.
397 19. Danecek P, Bonfield JK, Liddle J et al. Twelve years of SAMtools and BCFtools,
398 Gigascience 2021;10.
399 20. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing
400 genomic features, Bioinformatics 2010;26:841-842.
401 21. Borchers HW. pracma: Practical Numerical Math Functions 2022.
402 22. Conesa A, Madrigal P, Tarazona S et al. A survey of best practices for RNA-seq
403 data analysis, Genome Biol 2016;17:13.
404 23. Hackenberg M, Rodríguez-Ezpeleta N, Aransay AM. miRanalyzer: an update on
405 the detection and analysis of microRNAs in high-throughput sequencing experiments,
406 Nucleic Acids Res 2011;39:W132-138.
407 24. Lorenz R, Bernhart SH, Höner Zu Siederdissen C et al. ViennaRNA Package 2.0,
408 Algorithms Mol Biol 2011;6:26.
409 25. Varjak M, Donald CL, Mottram TJ et al. Characterization of the Zika virus
410 induced small RNA response in Aedes aegypti cells, PLoS Negl Trop Dis
411 2017;11:e0006010.
412 26. Öhlund P, Hayer J, Hesson JC et al. Small RNA Response to Infection of the
413 Insect-Specific Lammi Virus and Hanko Virus in an Aedes albopictus Cell Line,
414 Viruses 2021;13.

415