In Silico Reconstruction of Viral Genomes from Small RNAs Improves Virus-Derived Small Interfering RNA Profiling

Nicolas Vodovar, Bertsy Goic, Hervé Blanc, Maria-Carla P Saleh

To cite this version:

HAL Id: pasteur-01379334
https://hal-pasteur.archives-ouvertes.fr/pasteur-01379334
Submitted on 11 Oct 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives| 4.0 International License
RNA interference (RNAi) is the essential component of antiviral immunity in invertebrates and plants. One of the landmarks of the antiviral RNAi response is the production of virus-derived small interfering RNA (vsiRNA) from viral double-stranded RNA (dsRNA). vsiRNAs constitute a fragmented image of the viral genome sequence that results from Dicer cleavage. vsiRNA sequence profiling is used extensively as a surrogate to study the antiviral RNAi response by determining the nature of the viral dsRNA molecules exposed to and processed by the RNAi machinery. The accuracy of these profiles depends on the actual viral genome sequence used as a reference to align vsiRNA reads, and the interpretation of inaccurate profiles can be misleading. Using Flies house virus and Drosophila melanogaster as a model RNAi-competent organism, we show accurate reconstruction of full-length virus reference sequence from vsiRNAs and prediction of the structure of defective interfering particles (DIs). We developed a Perl script, named Paparazzi, that reconstitutes viral genomes through an iterative alignment/consensus call procedure using a related reference sequence as scaffold. As prevalent DI-derived reads introduce artifacts during reconstruction, Paparazzi eliminates DI-specific reads to improve the quality of the reconstructed genome. Paparazzi constitutes a promising alternative to Sanger sequencing in this context and an effective tool to study antiviral RNAi mechanisms by accurately quantifying vsiRNA along the replicating viral genome. We further discuss Paparazzi as a companion tool for virus discovery as it provides full-length genome sequences and corrects for potential artifacts of assembly.

Received 11 July 2011/Accepted 18 August 2011

Tissue culture, virus production, and infection. Naive Drosophila melanogaster S2 (Invitrogen) and infected S2R+ cells (Drosophila Genomics Resource Center) were cultured in Schneider’s medium (Invitrogen) supplemented with 10% fetal bovine serum (FBS) (Invitrogen). The S2+ cell line was maintained in our
laboratory for several passages. Titration by 50% tissue culture infective dose (TCID$_{50}$) of S2R+ culture supernatant was done on naïve S2 cells. For deep-sequencing experiments, naïve S2 cells were infected with virus present in S2R+ supernatant at a multiplicity of infection of 0.1 as determined by TCID$_{50}$. Cells and culture supernatant were recovered when the cytopathic effect was observed in 50% of the cells.

Small RNA library preparation and analysis. Total RNA was extracted for all samples using TRIzol reagent (Invitrogen) according to the manufacturer’s instructions. Small RNA libraries were prepared as previously described (6) and sequenced on a Genome Analyzer Ix (Illumina).

Reads (~25 millions) were clipped for adapters using the FASTX-Toolkit suite (http://hannonlab.cshl.edu/fastx_toolkit/) keeping reads between 19 and 30 nucleotides (nt) and discarding reads in which the adapter sequence could not be detected. Adapter-clipped reads were used as input for the Paparazzi script that is freely available at http://carla.salec.free.fr/soaires.php. The SSAKE 3.2 software (21) (http://www.bcgsc.ca/platform/bioinfo/software/ssa/ke/releases/3.2) was run using a seed length of 15 nucleotides (default parameters). Bowtie (10) (http://bowtie-bio.sourceforge.net/index.shtml) was run using the v alignment mode, the −all, −best and −strata options and using 4 parallel threads. Alignment against the Drosophila genome (ftp://ftp.flybase.net/genomes/Drosophila_melanogaster/current/fasta/) was performed allowing 0 mismatch; alignments against viral genomes were performed allowing either 1 or 2 mismatches (see Results). BLAT (8) (http://users.soe.ucsc.edu/−kent/src/) was run with a tile size of 8 and the fine option enabled. The viral reference sequences used were obtained from the National Center for Biotechnology Information (NCBI): Drosophila melanogaster American nodavirus (ANVNCBI) SW-2009a segments RNA1 (GG342965) and RNA2 (GG342966). Flock house virus (FHV NCBI) segments from the National Center for Biotechnology Information (NCBI): /H11001 was run with a tile size of 8 and nogaster/current/fasta/) was performed allowing 0 mismatch; alignments against Drosophila C virus (DCV NCBI) segments A (NC_004177) and B (NC_004169). The sequence logo in Fig. 2B was generated using weblogo (http://weblogo.berkeley.edu/) (4) from an average coverage of 98.831. The small RNA data set was submitted to the NCBI Small Read Archive under the accession number SRP009503.

All informatics analyses were performed on a desktop Mac Pro computer with the following configuration: Quad-core Intel Xeon E5426 processor running at 2.8 GHz, 20 Gb random-access memory (RAM) (4 Gb required) and Linux Ubuntu 11.04 operating system under 64 bits architecture. Under this configuration, filtering of host-derived reads and assembly of the unmatched reads into contigs takes approximately 25 s/million reads present in the input fastq file; reconstruction takes approximately 3 min per iteration.

RESULTS

Drosophila S2R+ cells are chronically infected by several viruses. Drosophila melanogaster S2 cells and their S2R+ derivative were shown to be chronically infected by several viruses (22). To identify the infectious viruses being produced by the S2R+ cell line from our laboratory and being processed by the RNAi machinery, we infected naïve Drosophila S2 cells with the culture supernatant of S2R+ cells. Small RNAs were recovered from infected samples and deep sequenced. The small RNA reads were then de novo assembled using SSAKE, and the contigs obtained were subjected to BLAST (1) alignment against the nucleotide database at NCBI. Using this approach, we obtained alignments against different regions of the Flock house virus (FHV) variant American nodavirus (ANV) (Nodaviridae, bisegmented single-stranded positive [ss (+)] genome; segments RNA1 and RNA2), the Drosophila C virus (DCV) [Dicistroviridae, ss (+)] genome, and the Drosophila X virus (DXV) (Birnaviridae, bisegmented dsRNA genome; segments A and B). However, the alignments obtained showed differences between the contigs and the sequences available at NCBI, suggesting that the genome sequences of the viruses present in these supernatants were different from those previously published. Therefore, we refer to these three viruses as FHVS2R-, DCV$_{S2R-}$, and DXV$_{S2R-}$. We further confirmed this result by Sanger sequencing of the RNA1 and RNA2 segments of FHVS2R- that displayed 2.83% and 4.80% difference with ANV$_{NCBI}$ RNA1 and RNA2, respectively. Further comparison of different FH isolates showed that the first 30 nucleotides of RNA1 and the last ~50 nucleotides of RNA2 are highly divergent and share no similarity when FHVS2R+ is compared to ANV$_{NCBI}$ (data not shown).

Full-length viral genome reconstruction from small RNA reads by iterative alignment/consensus call. During the alignment of small RNA reads against sequences of the three viruses derived by Sanger sequencing or the NCBI sequences of the three viruses, we observed a very high number of virus-derived reads (see Table S1 in the supplemental material). We therefore hypothesized that genomes could be fully reconstructed by aligning virus-derived small RNAs using the NCBI viral genome identified above as scaffold in order to bypass Sanger sequencing and directly obtain accurate sequence information about which genomic regions are being targeted by the antiviral RNAi response. To this end, we developed a Perl script, called Paparazzi (Fig. 1), that uses an iterative alignment/consensus call procedure and a reference sequence as a scaffold. Paparazzi aligns 19- to 30-nt-long reads against a reference genome using Bowtie, allowing one or two mismatches and no gaps between reads and the reference sequence. From this alignment, a $>50\%$ high-quality consensus sequence is calculated according to the following parameters: nucleotides with quality values of $\geq 30$ (probability that the base read is miscalled $\leq 10^{-2}$) and positions covered by $>10$ reads. The obtained consensus is then used as a new reference
to realign all the reads. This process is iterated until convergence is reached, i.e., when the number of reads aligned against the reference consensus sequence at cycle $n$ is equal to the number of reads aligned at cycle $n-1$.

Paparazzi reconstructed the 3,107-nucleotide-long genome sequence of FHV$^{S2R+}$ RNA1, including its 5' divergent extremity, with the exception of the four first nucleotides that remained unsolved (0.13%), i.e., the criteria to call a consensus were not met. The same sequence was obtained whether one or two mismatches were permitted during alignment; allowing 2 mismatches simply reduced the number of iterations required to reconstruct the genome sequence (15 iterations using 1 mismatch; 10 iterations using 2 mismatches). Of particular note, Paparazzi also reconstructed the same sequence for FHV$^{S2R+}$ RNA1 when using FHV$^{NCHI}$ as the reference sequence, differing from ANV$^{NCHI}$ by 9.14% and requiring only two additional iterations (17 iterations using 1 mismatch; 10 iterations using 2 mismatches). When using such a distant reference sequence as the scaffold, the 5' divergent extremity of FHV$^{S2R+}$ could be determined only by allowing 2 mismatches during reconstruction. These results show that Paparazzi can reconstruct full-length genome sequences from small RNA reads even when distantly related reference sequences are used.

To evaluate the accuracy of Paparazzi, the reconstructed and the Sanger sequencing-determined sequences of FHV$^{S2R+}$ RNA1 were compared. This analysis showed that these sequences differed by a single nucleotide (0.03%) while the sequences of ANV$^{NCHI}$ and FHV$^{S2R+}$ RNA1 differ by 2.83% over 3,107 nucleotides. This result indicates that Paparazzi can be used as a substitute to Sanger sequencing in this context. The presence of a single mismatch at position 5 between the reconstructed sequence and the Sanger sequencing-determined sequences strongly suggests that Paparazzi takes a snapshot of the viral genome processed by the RNAi machinery at a given time, rather than the input or output viral population.

**Genome breakpoint identification improves reconstruction accuracy in the presence of defective interfering particle (DI) genomes.** In contrast to RNA1, the reconstructed sequence of RNA2 displayed 64 differences with that determined by Sanger sequencing, introducing premature stop codons in the FHV$^{S2R+}$ RNA2 coding sequence. Moreover, Paparazzi introduced two artificial internal duplications (from nt 701 to 723 and nt 1237 to 1259 [701–723/1237–1259] and 248–267/517–536) that are absent from the Sanger sequencing-determined sequence of FHV$^{S2R+}$ RNA2. The differences between the two sequences are shown in red, and the stop codons introduced during the reconstruction process by Paparazzi are underlined. The sequence logo represents the nucleotide matrix calculated between the two sequences, with asterisks indicating the existence of minor variants, whose positions correspond to the difference observed between the S and R sequences. The differing sequence breakpoints constitute DI-specific signatures. When analyzing vsiRNAs that align against the sequence of these breakpoints, we observed a high proportion of DI-derived vsiRNAs that hampered genome reconstruction. Indeed, all differences observed between the Paparazzi-reconstructed and Sanger sequencing-determined FHV$^{S2R+}$ RNA2 sequences resided in the regions covered by DIs or around DI breakpoints. The vsiRNAs matching the positive and negative strands of the DI sequence are shown in red and blue, respectively. The vertical black lines indicate the sequence breakpoints within DI sequences compared to full-length FHV$^{S2R+}$ RNA2. The horizontal axis represents the DI coordinates in nucleotides, while the vertical axis represents read coverage ($\times 10,000$).

**FIG. 2.** FHV RNA2-derived DIs are heavily targeted by the RNAi machinery. (A) Structure of the major DI identified by Paparazzi and confirmed by Sanger sequencing. (B) Comparison of the sequence found by Sanger sequencing (S) and the reconstructed (R) sequence of FHV$^{S2R+}$ RNA2 between positions 565 and 600. The differences between the two sequences are shown in red, and the stop codons introduced during the reconstruction process by Paparazzi are underlined. The sequence logo represents the nucleotide matrix calculated between positions 565 and 600 for the FHV$^{S2R+}$ RNA2 sequence. Asterisks indicate the existence of minor variants, whose positions correspond to the difference observed between the S and R sequences. (C) Coverage of FHV$^{S2R+}$-derived siRNA along the sequence of the major RNA2 DI. The total length of each DI-derived vsiRNA was mapped on the DI sequence. The vsiRNAs matching the positive and negative strands of the DI sequence are shown in red and blue, respectively. The vertical black lines indicate the sequence breakpoints within DI sequences compared to full-length FHV$^{S2R+}$ RNA2. The horizontal axis represents the DI coordinates in nucleotides, while the vertical axis represents read coverage ($\times 10,000$).

To limit this effect, Paparazzi was instructed to identify breakpoints in viral genomes and to eliminate reads aligning against these breakpoints at each reconstruction step (Fig. 1, shaded boxes). Reads were first filtered against the host (*Drosophila*) genome using Bowtie, and the unmatched reads were *de novo* assembled using SSAKE. At each cycle, the contigs obtained were aligned against the virus consensus sequence generated using BLAT, allowing the identification of breakpoints within alignments. Using this feature, Paparazzi identified the three major breakpoints previously defined by Sanger sequencing (Fig. 2A), along with new minor breakpoints rep-
represented by fewer reads (data not shown). This implementation significantly improved the quality of the reconstructed FHV12/2+ RNA2 sequence, although eight nucleotides remained unsolved (nt 254 to 260 corresponding to one DI breakpoint and nt 1416). The reconstruction was effective only when 1 mismatch was used; allowing 2 mismatches did not correct for artifacts (e.g., internal duplication; see below). Of note, Paparazzi could not properly reconstruct the last ∼50 nucleotides of the RNA2 segment because this region (from nt 1373 to the end) is highly variable in length and sequence among FHV isolates. Nevertheless, the reconstructed genome no longer displayed duplication, and only 30 nucleotides were different from the sequences of the input and output virus determined by Sanger sequencing. All these differences fell into regions covered by DIs, confirming that DIs have diverged from the actual sequence of the genome. Moreover, analysis of the matrix generated throughout the reconstruction process revealed minor genomic variants. In regions shared by the actual genome and the DIs, these variants corresponded to the actual genome sequence (Fig. 2B).

Altogether, these results indicate that Paparazzi (i) is precise in reconstructing genomes in regions that do not overlap with DI sequences, (ii) can limit the deleterious effect of DI-derived reads during genome reconstruction, and (iii) provides information on viral genome diversity. In the context of DIs, this information may allow the prediction of the actual genome sequence. Interestingly, for FHV12/2+ RNA2, we not only observed vsiRNAs mapping in regions corresponding to the DI sequence, as previously observed (5, 22), we also observed abundant vsiRNAs aligning at the breakpoints of the DIs (Fig. 2C). These data suggest that DIs are heavily targeted by the RNAi machinery.

**Paparazzi successfully reconstructs the genomes of DCVS2R+ and DXVS2R+.** Reconstruction using 2 mismatches is faster and more effective when using a distant reference scaffold sequence. However, the number of mismatches must be corrected for artifacts (e.g., internal duplication; see below). Of note, Paparazzi could not properly reconstruct the last ∼50 nucleotides of the RNA2 segment because this region (from nt 1373 to the end) is highly variable in length and sequence among FHV isolates. Nevertheless, the reconstructed genome no longer displayed duplication, and only 30 nucleotides were different from the sequences of the input and output virus determined by Sanger sequencing. All these differences fell into regions covered by DIs, confirming that DIs have diverged from the actual sequence of the genome. Moreover, analysis of the matrix generated throughout the reconstruction process revealed minor genomic variants. In regions shared by the actual genome and the DIs, these variants corresponded to the actual genome sequence (Fig. 2B).

Altogether, these results indicate that Paparazzi (i) is precise in reconstructing genomes in regions that do not overlap with DI sequences, (ii) can limit the deleterious effect of DI-derived reads during genome reconstruction, and (iii) provides information on viral genome diversity. In the context of DIs, this information may allow the prediction of the actual genome sequence. Interestingly, for FHV12/2+ RNA2, we not only observed vsiRNAs mapping in regions corresponding to the DI sequence, as previously observed (5, 22), we also observed abundant vsiRNAs aligning at the breakpoints of the DIs (Fig. 2C). These data suggest that DIs are heavily targeted by the RNAi machinery.

Using this improved version of Paparazzi, we successfully reconstructed the genome of DCVS2R+ (9,264 nt) and DXVS2R+ (segment A, 3,360 nt; segment B, 3,243 nt). Paparazzi rebuilt the full genome sequence of DCVS2R+ except for 44 nucleotides (0.48% of the genome sequence), of which 35 (0.27%) corresponded to a previously unidentified putative DI breakpoint (nt 8115 to 8150) (data not shown). Paparazzi also reconstructed 99.85% and 99.41% of DXVS2R+ (nt 8115 to 8150) (data not shown). Paparazzi successfully reconstructed the full genome sequence of DCVS2R+ (Fig. S2A to S2C) and DXVS2R+ (Fig. S2D to S2I) sequences.

Paparazzi improves vsiRNA quantification along viral genome. The reconstructed sequences for the three viruses displayed less than 3% differences compared to their closest related NCBI sequences. To evaluate the impact of such a difference in vsiRNA profiling, we compared vsiRNA profiles obtained using either NCBI sequences or Paparazzi-reconstructed viral genomes as reference sequences.

The use of reconstructed sequence to perform vsiRNA profile improved the quality of vsiRNA profiles compared to those obtained using NCBI sequences for the three viruses tested (Fig. 3A and B; see Fig. S1A to S1C and Fig. S2 in the supplemental material). In particular, the profiles obtained using ANVNCBI RNA1 (Fig. S1B) and RNA2 (Fig. S3B) displayed regions uncovered by vsiRNAs. In contrast, these uncovered regions were no longer observed when the reconstructed sequence was used (Fig. 3A and Fig. S1A). In addition, the profile obtained for FHV12/2+ RNA2 showed that RNA2 is fully covered by vsiRNA (Fig. 3A), contrary to previously published profiles (5, 20) and the profile obtained using ANVNCBI reference sequence (Fig. 3B and C), and confirmed the profusion of DI-derived vsiRNAs. Similar improvement was obtained when profiling vsiRNAs against reconstructed DCVS2R+ (Fig. S2A to S2C) and DXVS2R+ (Fig. S2D to S2I) sequences.

Importantly, relaxing alignment stringency while using NCBI reference sequences only partially compensated the inaccuracy of the sequence and therefore the vsiRNA profiles obtained (Fig. 3; see Fig. S1 in the supplemental material). Indeed, the 3′ end of RNA2 is fully covered by vsiRNA when profiling is performed using the reconstructed FHV12/2+ sequence, even under stringent alignment conditions (0 mismatch; Fig. 3A). In contrast, increasing the number of mismatches when aligning vsiRNAs against the ANVNCBI RNA2 sequence did not significantly improve the very low vsiRNA coverage of this region (Fig. 3B and C). Similar results were obtained for the 30 first nucleotides of the RNA1 sequence (Fig. S1G).

These results showed (i) that the reconstructed sequence provided by Paparazzi improves the quality of vsiRNA profiles and (ii) that relaxing alignment stringency only partially compensates the inaccuracy of vsiRNA profile against approximate reference sequences. Altogether, these data reemphasize the need for an accurate reference genome when profiling vsiRNA that can be readily generated by Paparazzi.

**DISCUSSION**

In recent years, due to the development of NGS technologies, vsiRNA profiling has been extensively used as a surrogate to study the processing of viral genomes by the RNAi machinery. Although vsiRNAs align along full-length viral genomes, some regions are less covered than others (cold spots). Such cold spots were observed in the profile of the Semliki forest virus (SFV), and they correlated with predicted secondary structure that appears to be insensitive to Dicer processing (17). A similar interpretation was given for cold spots observed in the profile of FHV (5, 20). However, these profiles were incomplete due to an imprecise reference sequence. Indeed, we obtained similar incomplete profiles when we used ANVNCBI (Fig. 3B and C) as the reference sequence, while vsiRNAs mapped along the full-length FHV12/2+ genome after reconstruction (Fig. 3A). These results show that the identification of artifactual cold spots results from disparities between the sequence of the virus studied and the one used as reference. It is worth mentioning that artifactual cold spots...
may also be linked to technical factors, such as the small RNA library preparation protocol or the choice of the sequencing platform (7, 12, 18, 19).

Interestingly, in the case of SFV, synthetic siRNAs corresponding to cold spots were shown to inhibit viral replication more than synthetic siRNAs corresponding to hot spots. These results suggest that cold spots constitute a promising target for the design of effective antiviral siRNA-based therapeutics applied to agronomic or human health. To optimize such an approach, it is important to limit the identification of artificial cold spots, and therefore, the a priori knowledge of the actual virus reference sequence becomes critical.

Determining the actual reference sequence is particularly important when working with rapidly evolving RNA viruses (15). Although in some cases, the accumulation of changes in consensus sequence may be a slow process, viral genomes may vary in even a limited number of replication cycles, particularly if genetic bottlenecks and positive selection are occurring.

With more researchers turning to NGS technology to study viruses, we developed Paparazzi as a one-step tool that fully exploits the data generated by NGS to reconstruct viral genomes and therefore profile vsiRNAs accurately. Paparazzi successfully and rapidly reconstructed the sequences of the three viruses present in our infected sample, even if the initial reference sequence used as the scaffold differed by up to 10% with the sequence of the actual replicating virus. Altogether, our results show that Paparazzi can be used as a proxy to the Sanger method to resequence and potentially genotype viral strains from RNAi-competent organisms.

A key aspect of Paparazzi is its ability to detect variations in viral genomes without prior knowledge of their existence. First, single-nucleotide polymorphisms can be quantified using the nucleotide frequency matrix that is calculated during genome reconstruction. Second, both major and minor DI breakpoints can be identified by the DI discovery snippet, which improves the accuracy of the viral genome reconstruction when DI-
derived reads are overrepresented. The application of this last feature allowed us to show that DI-derived reads are highly targeted by the RNAi machinery for the FHV\textsuperscript{N2R} RNA2 genome. Two nonmutually exclusive hypotheses may account for these observations. (i) DIs replicate faster because of their smaller size, and the DI-genome-derived siRNA ratio follows the stoichiometry of these two species. (ii) DIs are less efficiently encapsidated than the full-length RNA2 sequence and are thus more exposed to the RNAi machinery. It was previously proposed that the prevalence of DI-derived siRNAs correlates with FHV persistence in S2 cells (5). However, we found a similar overrepresentation of DI-derived siRNAs under our experimental conditions, which is lethal for the naïve S2 cells. Therefore, the abundance of DIs is not sufficient to allow persistence of FHV infection, although it may contribute to this phenomenon.

While analyzing our data, we noticed that the overrepresentation of DI-derived reads affects the accuracy of virus genome reconstruction. Indeed, the genome sequence of ANV NCBI was determined using both \textit{de novo} assembly of virus-derived small RNAs and Sanger sequencing-based gap fill-in (22) displays two internal duplications (248–261/517–530 [100% identity] and 112–153/1349–1390 [89% identity]) in its RNA2 segment. None of these duplications were observed in any other FHV genome, including that of FHV\textsuperscript{N2R}. Interestingly, we obtained the same duplications when Paparazzi was instructed not to filter DIs (248–261/517–530), or if Paparazzi is applied allowing 2 mismatches for genome reconstruction (116–152/1353–1389), or if Paparazzi was instructed not to filter DIs (248–261/517–530), or if Paparazzi is applied allowing 2 mismatches for genome reconstruction (116–152/1353–1389). As these duplications are absent from the Sanger sequencing-determined sequence of FHVS2R (116–152/1353–1389). As these duplications are absent from our experimental conditions, which is lethal for the naïve S2 cells. Therefore, the abundance of DIs is not sufficient to allow persistence of FHV infection, although it may contribute to this phenomenon.

While analyzing our data, we noticed that the overrepresentation of DI-derived reads affects the accuracy of virus genome reconstruction. Indeed, the genome sequence of ANV NCBI was determined using both \textit{de novo} assembly of virus-derived small RNAs and Sanger sequencing-based gap fill-in (22) displays two internal duplications (248–261/517–530 [100% identity] and 112–153/1349–1390 [89% identity]) in its RNA2 segment. None of these duplications were observed in any other FHV genome, including that of FHV\textsuperscript{N2R}. Interestingly, we obtained the same duplications when Paparazzi was instructed not to filter DIs (248–261/517–530), or if Paparazzi is applied allowing 2 mismatches for genome reconstruction (116–152/1353–1389). As these duplications are absent from the Sanger sequencing-determined sequence of FHV\textsuperscript{N2R}, we infer that the duplications observed in ANV NCBI RNA2 result from artifacts in \textit{de novo} assembly, and therefore, the actual sequence of ANV RNA2 may differ from the one previously published (22). Although Paparazzi is not a viral discovery pipeline, its ability to reconstruct full-length viral genome sequences and spot artifacts of reconstruction make it a powerful companion tool to polish the results obtained from virus discovery pipelines. Of note, these features of Paparazzi can also be exploited when directly resequencing viral RNA by NGS technologies.

In conclusion, Paparazzi provides an effective tool for viral genome reconstruction, accurate siRNA profiling, and studying RNAi processing. The development of such a tool fits into a constant effort to limit the bottleneck between NGS data generation and analysis in a context where NGS technologies become more affordable and efficient by the day.

ACKNOWLEDGMENTS

This work was supported by the French Agence Nationale de la Recherche (ANR-09-JCJC-0045-01) and the European Research Council (FP7/2007-2013 ERC 242703) to M.-C.S.

N.V. conceived and designed experiments, developed the bioinformatics tools, and wrote the paper. B.G. and H.B. performed experiments. M.-C.S participated in project conception, experimental design, interpretation of results, and writing of the paper.

We thank Valérie Dorey for technical assistance, the members of the Saleh lab for fruitful discussions, and Marco Vignuzzi and François Schwegmann for critically reading the manuscript.