JOURNAL OF VIROLOGY, Nov. 2011, p. 11016–11021 Vol. 85, No. 21
0022-538X/11/$12.00 doi:10.1128/JVI.05647-11

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

Nicolas Vodovar, Bertsy Goic, Hervé Blanc, and Maria-Carla Saleh*

Institut Pasteur, Viruses and RNA Interference Group and Centre National de la Recherche Scientifique, URA 3015, 75015 Paris, France

Received 11 July 2011/Accepted 18 August 2011

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 Flock 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.

RNA interference (RNAi) is an essential component of the antiviral immune response in invertebrates and plants. In insects, one of the landmarks of the antiviral RNAi response is the production of 21-nucleotide (nt)-long virus-derived small interfering RNAs (vsiRNAs). As vsiRNAs derive from cleavage of viral double-stranded RNA (dsRNA) molecules (e.g., intermediates of replication) by Dicer enzymes, they constitute a fragmented image of the viral genome sequence in the sample. This property can be used to discover new viruses, but genomes have been only partially reconstructed by assembling vsiRNA reads using next-generation sequencing (NGS) technologies (9, 22).

Besides its importance in virus discovery, the recent development of NGS technologies constituted a critical step in our understanding of the antiviral RNAi response. In particular, profiling vsiRNA along viral genomes is used extensively to determine the nature and structure of the viral dsRNA exposed to and processed by the RNAi machinery. Such profile analyses in *Drosophila melanogaster* (3, 5, 13, 20, 22) and other insects (2, 3, 14, 16, 17) led to the conclusion that some viruses are targeted by Dicer-2 across the entire genome, while others are targeted on limited regions (5, 20). Since profiling vsiRNAs involves their alignment against a viral reference genome, the accuracy of these profiles and therefore, the conclusions drawn from them, strongly depends on the reference sequence being used.

As RNA viruses accumulate mutations at high frequency (15), their consensus sequence evolves rapidly, entailing the resequencing by classical methods of the virus being used. Given the amount of data produced by NGS experiments, it is tempting to think that the full-length genome of the virus used in an experiment could be deduced from vsiRNAs. Here we show that it is possible to reconstruct the entire viral genome and thereby obtain an accurate reference sequence from the vsiRNAs present in the sample. We propose a Perl script, named Paparazzi, that accurately reconstitutes viral genomes through an iterative alignment/consensus call procedure using an initial reference sequence as a scaffold. The resulting full-length consensus sequence is then reused to profile vsiRNAs. Paparazzi constitutes an all-in-one tool for viral sequence determination and accurate vsiRNA profiling that fully exploits the depth of NGS data and precludes the use of other time-consuming sequencing technologies. We discuss further applications of Paparazzi in virus identification and virus diversity when resequencing viral genomes.

MATERIALS AND METHODS

Tissue culture, virus production, and infection. Naïve *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 S2R+ cell line was maintained in our

* Corresponding author. Mailing address: Institut Pasteur, Viruses and RNAi Group, CNRS URA 3015, 25 rue du Dr Roux, 75724 Paris Cedex 15, France. Phone: (33) 1 45 68 85 47. Fax: (33) 1 40 61 36 27. E-mail: carla.saleh@pasteur.fr.

† Supplemental material for this article may be found at http://jvi.asm.org/.

▽ Published ahead of print on 31 August 2011.

‡ The authors have paid a fee to allow immediate free access to this article.

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 IIx (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.saleh.free.fr/softwares.php. The SSAKE 3.2 software (21) (http://www.bcgsc.ca/platform/bioinfo/software/ssake/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 (ANV^{NCBI}) SW-2009a segments RNA1 (GQ342965) and RNA2 (GQ342966), Flock house virus (FHV^{NCBI}) segments RNA1 (NC_004146) and RNA2 (NC_004144), *Drosophila* C virus (DCV^{NCBI}; NC_001834), and *Drosophila* X virus (DXV^{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 SRP005903.

All informatics analyses were performed on a desktop Mac Pro computer with the following configuration: Quad-core Intel Xeon E5462 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.

RT-PCR, RACE-PCR, and Sanger dideoxy sequencing. Total RNA from viral inoculates and from culture supernatant from infected cells were extracted using TRIzol (Invitrogen) and used as template for reverse transcription-PCR (RT-PCR) and rapid amplification of cDNA ends and PCR (RACE-PCR). RT-PCR products were sequenced directly by Sanger dideoxy sequencing. 5′ RACE-PCR and 3′ RACE-PCR were performed using the 5/3 RACE kit (Roche) according to the manufacturer's instructions. For 3′ RACE, RNAs were polyadenylated using yeast poly(A)-polymerase (Affymetrix) prior to anchored-oligo(dT)-primed reverse transcription. Before Sanger dideoxy sequencing, 5′ RACE and 3′ RACE products were cloned in TOPO-T/A vector (Invitrogen). The DNAStar Lasergene package was used for virus genome assembly. Flock house virus isolate VIA-022011 segments RNA1 and RNA2 were submitted to the NCBI Entrez Nucleotide Database under accession numbers JF461541 and JF461542, respectively.

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 FHV^{S2R+}, DCV^{S2R+}, and DXV^{S2R+}. We further confirmed this result by Sanger sequencing of the RNA1 and RNA2 segments of FHV^{S2R+} that displayed 2.83% and 4.80% difference with ANV^{NCBI} RNA1 and RNA2, respectively. Further comparison of different FHV isolates showed that the first 30 nucleotides of RNA1 and the last ~50 nucleotides of RNA2 are highly divergent and share no similarity when FHV^{S2R+} is compared to ANV^{NCBI} (data not shown).

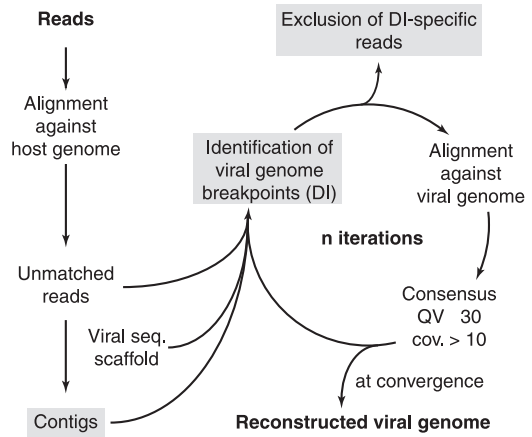

FIG. 1. Schematic representation of the Paparazzi script. Shaded boxes indicate the steps for filtering out the defective interfering particle (DI)-derived reads. QV, quality value; cov., coverage; seq., sequence.

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 ≥30 (probability that the base read is miscalled $\leq 10^{-3}$) 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; 8 iterations using 2 mismatches). Of particular note, Paparazzi also reconstructed the same sequence for FHV^{S2R+} RNA1 when using FHV^{NCBI} as the reference sequence, differing from ANV^{NCBI} 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^{NCBI} 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 sequence 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 artifactual 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. These data suggested that the reconstructed sequence did not represent the actual functional genome.

The RNA2 segment of FHV is prone to internal deletions that give rise to defective interfering particles (22, 23). These DIs cannot sustain infection by themselves but are efficiently replicated and encapsidated using proteins produced by wild-type viral genomes. The existence of DIs in the sample was confirmed by PCR (smaller amplification product than for the full-length genome) followed by Sanger sequencing (data not shown). The comparison of the sequence obtained for DIs and that of ANV^{NCBI} RNA2 allowed the identification of three breakpoints (Fig. 2A), which are similar to those previously described (11, 22, 23), suggesting that there are hot spots in FHV RNA2 for such deletions to occur. The sequences at

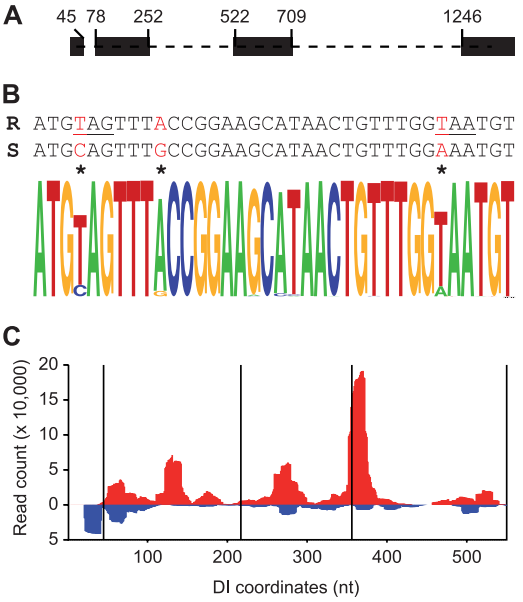

FIG. 2. FHV RNA2-derived DIs are heavily targeted by the RNAi machinery. (A) Structure of the major DI identified by Paparazzi and confirmed by Sanger sequencing. (B) Comparison of the sequence found by Sanger sequencing (S) and the reconstructed (R) sequence of FHV^{S2R+} RNA2 between positions 565 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 (×10,000).

these 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 (including genome-specific regions), and the artifactual duplications introduced by Paparazzi during reconstruction corresponded to DI breakpoints.

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-

resented by fewer reads (data not shown). This implementation significantly improved the quality of the reconstructed FHV^{S2R+} 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 FHV^{S2R+} 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 DCV^{S2R+} and DXV^{S2R+}. Reconstruction using 2 mismatches is faster and more effective when using a distant reference scaffold sequence. However, the number of mismatches must be set at 1 mismatch when dealing with DI-derived reads. To optimize reconstruction while taking DIs into account, Paparazzi was modified so that the number of mismatches allowed during reconstruction is automatically adjusted at each iteration. This number is set at 2 by default but is automatically set at 1 as soon as Paparazzi detects breakpoints within viral genomes.

Using this improved version of Paparazzi, we successfully reconstructed the genome of DCV^{S2R+} (9,264 nt) and DXV^{S2R+} (segment A, 3,360 nt; segment B, 3,243 nt). Paparazzi rebuilt the full genome sequence of DCV^{S2R+} 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 reconstituted 99.85% and 99.41% of DXV^{S2R+} segments A and B, respectively. The reconstructed genomes displayed 2.94%, 2.26%, and 2.1% differences with the sequences of DCV^{NCBI} and DXV^{NCBI} segments A and B, respectively.

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 ANV^{NCBI} RNA1 (Fig. S1B) and RNA2 (Fig. 3B) displayed regions uncovered by siRNAs. 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 FHV^{S2R+} RNA2 showed that RNA2 is fully covered by vsiRNA (Fig. 3A), contrary to previously published profiles (5, 20) and the profile obtained using ANV^{NCBI} reference sequence (Fig. 3B and C), and confirmed the profusion of DI-derived vsiRNAs. Similar improvement was obtained when profiling vsiRNAs against reconstructed DCV^{S2R+} (Fig. S2A to S2C) and DXV^{S2R+} (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 FHV^{S2R+} sequence, even under stringent alignment conditions (0 mismatch; Fig. 3A). In contrast, increasing the number of mismatches when aligning vsiRNAs against the ANV^{NCBI} 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 ANV^{NCBI} (Fig. 3B and C) as the reference sequence, while vsiRNAs mapped along the full-length FHV^{S2R+} 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

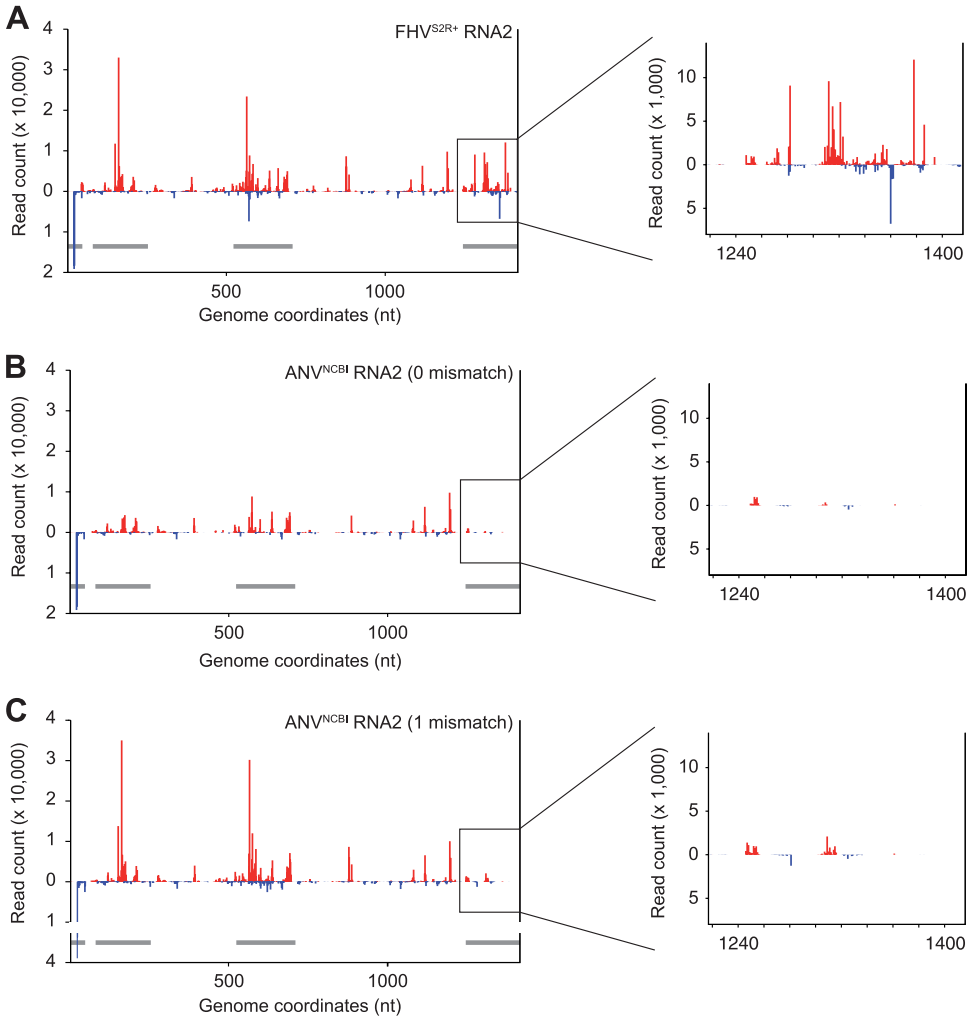

FIG. 3. Profiles of 21-nt FHV^{S2R+}-derived reads for the RNA2 segment. These profiles were obtained using either the reconstructed FHV^{S2R+} (A) or the ANV^{NCBI} genomes as reference allowing 0 mismatch (B) or 1 mismatch (C) during the alignment. The panels on the right display magnified profiles between positions 1217 and 1416. Each FHV^{S2R+}-derived small RNAs (vsiRNA) is represented by the position of its first nucleotide. The vsiRNAs matching the positive (red) and negative (blue) strands of RNA2 are shown. The horizontal axis represents the genome coordinates in nucleotides, while the vertical axis represents read coverage (×10,000 for panels A to C and ×1,000 for the right panels). The gray bars under the profiles represent regions corresponding to both genome and DI sequences.

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 artifactual 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^{S2R+} 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} that was determined using both *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^{S2R+}. 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^{S2R+}, we infer that the duplications observed in ANV^{NCBI} RNA2 result from artifacts in *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 vsiRNA 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 Schweisguth for critically reading the manuscript.

REFERENCES

1. **Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman.** 1990. Basic local alignment search tool. J. Mol. Biol. **215:**403–410.
2. **Brackney, D. E., J. E. Beane, and G. D. Ebel.** 2009. RNAi targeting of West Nile virus in mosquito midguts promotes virus diversification. PLoS Pathog. **5:**e1000502.
3. **Brackney, D. E., et al.** 2010. C6/36 Aedes albopictus cells have a dysfunctional antiviral RNA interference response. PLoS Negl. Trop. Dis. **4:**e856.
4. **Crooks, G. E., G. Hon, J. M. Chandonia, and S. E. Brenner.** 2004. WebLogo: a sequence logo generator. Genome Res. **14:**1188–1190.
5. **Flynt, A., N. Liu, R. Martin, and E. C. Lai.** 2009. Dicing of viral replication intermediates during silencing of latent Drosophila viruses. Proc. Natl. Acad. Sci. U. S. A. **106:**5270–5275.
6. **Gausson, V., and M. C. Saleh.** 2011. Viral small RNA cloning and sequencing. Methods Mol. Biol. **721:**107–122.
7. **Hafner, M., et al.** 2011. RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. RNA **17:**1697–1712.
8. **Kent, W. J.** 2002. BLAT–the BLAST-like alignment tool. Genome Res. **12:**656–664.
9. **Kreuze, J. F., et al.** 2009. Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: a generic method for diagnosis, discovery and sequencing of viruses. Virology **388:**1–7.
10. **Langmead, B., C. Trapnell, M. Pop, and S. L. Salzberg.** 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. **10:**R25.
11. **Li, Y., and L. A. Ball.** 1993. Nonhomologous RNA recombination during negative-strand synthesis of Flock house virus RNA. J. Virol. **67:**3854–3860.
12. **Linsen, S. E., et al.** 2009. Limitations and possibilities of small RNA digital gene expression profiling. Nat. Methods **6:**474–476.
13. **Mueller, S., et al.** 2010. RNAi-mediated immunity provides strong protection against the negative-strand RNA vesicular stomatitis virus in Drosophila. Proc. Natl. Acad. Sci. U. S. A. **107:**19390–19395.
14. **Myles, K. M., M. R. Wiley, E. M. Morazzani, and Z. N. Adelman.** 2008. Alphavirus-derived small RNAs modulate pathogenesis in disease vector mosquitoes. Proc. Natl. Acad. Sci. U. S. A. **105:**19938–19943.
15. **Sanjuan, R., M. R. Nebot, N. Chirico, L. M. Mansky, and R. Belshaw.** 2010. Viral mutation rates. J. Virol. **84:**9733–9748.
16. **Scott, J. C., et al.** 2010. Comparison of dengue virus type 2-specific small RNAs from RNA interference-competent and -incompetent mosquito cells. PLoS Negl. Trop. Dis. **4:**e848.
17. **Siu, R. W., et al.** 2011. Antiviral RNA interference responses induced by Semliki Forest virus infection of mosquito cells: characterization, origin and frequency-dependent functions of virus-derived small interfering RNAs. J. Virol. **85:**2907–2917.
18. **Smith, N. A., A. L. Eamens, and M. B. Wang.** 2010. The presence of high-molecular-weight viral RNAs interferes with the detection of viral small RNAs. RNA **16:**1062–1067.
19. **Szittya, G., et al.** 2010. Structural and functional analysis of viral siRNAs. PLoS Pathog. **6:**e1000838.
20. **van Rij, R. P., and E. Berezikov.** 2009. Small RNAs and the control of transposons and viruses in Drosophila. Trends Microbiol. **17:**163–171.
21. **Warren, R. L., G. G. Sutton, S. J. Jones, and R. A. Holt.** 2007. Assembling millions of short DNA sequences using SSAKE. Bioinformatics **23:**500–501.
22. **Wu, Q., et al.** 2010. Virus discovery by deep sequencing and assembly of virus-derived small silencing RNAs. Proc. Natl. Acad. Sci. U. S. A. **107:**1606–1611.
23. **Zhong, W., R. Dasgupta, and R. Rueckert.** 1992. Evidence that the packaging signal for nodaviral RNA2 is a bulged stem-loop. Proc. Natl. Acad. Sci. U. S. A. **89:**11146–11150.