MISIS-2: a bioinformatics tool for in-depth analysis of small RNAs and representation of consensus master genome in viral quasispecies

Jonathan Seguin^{1,2}, Patricia Otten², Loïc Baerlocher², Laurent Farinelli², Mikhail M. Pooggin^{1*}

¹University of Basel, Department of Environmental Sciences, Botany, Hebelstrasse 1, 4056 Basel, Switzerland, ²Fasteris SA, Ch. du Pont-du-Centenaire 109, 1228 Plan-les-Ouates, Switzerland

*Corresponding author at: University of Basel, Institute of Botany, Hebelstrasse 1, 4056 Basel, Switzerland. Tel.: +41 61 2672977; fax: +41 61 2673504.

E-mail address: Mikhail.Pooggin@unibas.ch (M.M. Pooggin).

Abstract

In most eukaryotes, small RNA (sRNA) molecules such as miRNAs, siRNAs and piRNAs regulate gene expression and repress transposons and viruses. AGO/PIWI family proteins sort functional sRNAs based on size, 5'-nucleotide and other sequence features. In plants and some animals, viral sRNAs are extremely diverse and cover the entire viral genome sequences, which allows for *de novo* reconstruction of a complete viral genome by deep sequencing and bioinformatics analysis of viral sRNAs. Previously, we have developed a tool MISIS to view and analyze sRNA maps of viruses and cellular genome regions which spawn multiple sRNAs. Here we describe a new release of MISIS, MISIS-2, which enables to determine and visualize a consensus sequence and count sRNAs of any chosen sizes and 5'-terminal nucleotide identities. Furthermore we demonstrate the utility of MISIS-2 for identification of single nucleotide polymorphisms (SNPs) at each position of a reference sequence and reconstruction of a consensus master genome in evolving viral quasispecies. MISIS-2 is a Java standalone program. It is freely available along with the source code at the website <http://www.fasteris.com/apps>.

Keywords

Small RNA; siRNA; viral quasispecies; consensus sequence; single nucleotide polymorphism; bioinformatics tool.

1. Introduction

Small regulatory RNAs (sRNAs) are involved in various biological processes in both eukaryotes and prokaryotes (Ghildiyal and Zamore, 2009; Babski et al., 2014). In eukaryotes, sRNAs in a size-range of 20-30 nucleotides (nts) regulate gene expression and chromatin states as well as repress invasive nucleic acids such as transposons, transgenes and viruses. Based on the mechanisms of biogenesis and function, the eukaryotic sRNAs are classified in micro RNAs (miRNAs), short interfering RNAs (siRNAs) and PIWI-interacting RNAs (piRNAs) (Ghildiyal and Zamore, 2009; Meister, 2013). Functional sRNAs bind effector proteins from an AGO/PIWI family and guide the resulting RNA-induced silencing complexes (RISCs) to complementary target RNA or DNA. Sorting of sRNAs by the effector proteins is based mostly on sRNA size and 5'-nucleotide identity (Cenik and Zamore, 2011; Meister, 2013). Populations of sRNAs such as piRNAs and siRNAs which direct silencing of multiple genomic loci and viruses are extremely complex, with multiple sRNA species of different sizes, polarities (sense or antisense) and abundances across a given target sequence. In-depth analysis of such complex sRNA profiles can be indicative of a particular silencing mechanism for each target sequence.

Previously we have developed a bioinformatics tool MISIS to view and analyze maps of sRNA reads obtained by next generation sequencing and mapped to a given reference sequence (Seguin et al., 2014a). MISIS reads standard BAM/SAM files outputted by mapping tools such as Bowtie (Langmead et al., 2009) and BWA (Li and Durbin, 2009) and generates two text files that contain counts of the sRNA reads mapped to the reference sequence with zero mismatches, or with up to two mismatches, at each position of the reference's forward and reverse strand and for each of the chosen size-classes of sRNAs. These text files are then used by MISIS to display the sRNA maps as a histogram where the x-axis indicates positions of the 5'- or 3'-terminal nucleotide of sense and antisense sRNAs, respectively, along the reference sequence or its selected region and the y-axis the number of reads starting (for sense sRNA) or ending (for antisense sRNA) at each position. Size-classes of sRNAs can be visualized and compared separately or in combination. Thus, MISIS gives an overview of sRNA distribution along the reference sequence as well as detailed information on single sRNA species of different size-classes and abundances (Seguin et al., 2014a). Using MISIS, we have found out that viral siRNA species generated by the plant silencing machinery cover the entire genomes of RNA and DNA viruses without gaps in both sense and antisense orientation (Seguin et al., 2014b). This finding has confirmed and extended a pioneering work of Kreuze et al. (2009) showing that a complete genome of plant RNA virus can be reconstructed as multiple viral siRNA contigs using a bioinformatics tool Velvet (Zerbino and Birney, 2008). Furthermore, RNA viruses of invertebrate animals could also be detected and partially reconstructed by small RNA

sequencing and bioinformatics (Wu et al., 2010). Following these findings, we have developed an siRNA-omics (siRomics) approach based on next generation sequencing and bioinformatics which enabled *de novo* reconstruction of complete genomes of plant RNA and DNA viruses as single contigs of viral siRNAs. Moreover, the siRomics approach and MISIS-aided analysis of viral siRNAs allowed for reconstruction of a consensus master genome in viral quasispecies (Seguin et al., 2014b). Recently, it has been confirmed for potato virus Y that viral siRNAs faithfully represent the consensus master genome packaged in viral particles and can therefore be used for its reconstruction (Kutnjak et al., 2015).

Here we describe a second version of MISIS which enables a more comprehensive analysis of sRNA populations. In addition to all the above-mentioned features of MISIS, MISIS-2 can determine and visualize a consensus sequence supported by the majority of mapped sRNA reads, identify single nucleotide polymorphisms (SNPs) at any position of the reference sequence and count sRNA reads based on size and 5'-nucleotide identity. Furthermore, we demonstrate that MISIS-2 facilitates reconstruction of a consensus master genome in evolving viral quasispecies.

2. Results and Discussion

2.1. Features and applications of MISIS-2

MISIS-2 reads a BAM/SAM file containing sRNA mapping data for a given reference sequence and creates two text files (called table files) and two binary java files (called project files). To create these files the user should sequentially click "File", "Load BAM/SAM" and, in a popup window, select the BAM/SAM file, the reference sequence (in a fasta format) and the project folder. By default, the size range of sRNAs is from 20 to 25 nts, but the minimal and the maximal sizes can be changed in the popup window. The resulting table files are small-size intermediate files that can be visualised by MISIS-2 (click "File", "Load table" and select the table file from the project folder) and analyzed as described previously (Seguin et al., 2014a). They can also be loaded by a spreadsheet software such as Excel for further quantitative analysis. In each table, the columns contain the positions of 5'-terminal nucleotide of sense sRNAs and 3'-terminal nucleotide of antisense sRNAs along the reference sequence, the counts of the sense and antisense sRNAs of each size-class, and the total counts of sRNAs mapped on the forward and reverse strands and on both strands. By reading CIGAR and tag "MD" values, the MISIS-2 algorithm simultaneously creates two output tables: one containing the counts of sRNAs mapped to the reference sequence with zero mismatches ("_pm" in the file name stands for perfect match) and another the combined counts for sRNAs mapped with mismatches according to the mapping parameters. Side-by-side comparison of

the two tables is informative of nucleotide difference (SNPs or indels) between the sRNAs and the reference sequence. Likewise, the two binary project files created by MISIS-2 differ by the number of mismatches in mapped sRNA reads: the file "_project_pm.b2btf" contains the reads without mismatches and the file "_project.b2btf" contains the reads with and without mismatches.

To visualize and analyze the project files the user should click "File" and "Load Project". Note that only the file "_project.b2btf" can be used to determine and display a consensus sequence and to identify and count SNPs. MISIS-2 reads the project file and creates an interactive histogram where the x-axis indicates the nucleotide positions along the reference sequence and the y-axis the number of mapped sRNAs starting (sense) or ending (antisense) at each position of the forward and reverse strands, respectively (Fig. 1A). The user can select sRNA size-classes separately or in any chosen combination (click "View" and "Select the sizes"), which will then appear as single or multiple histograms, respectively. When two or more histograms are displayed, the user can change the scale of y-axis (check "Set y-axis"). This allows for quantitative comparison of sRNA profiles. Furthermore, the user can zoom into any region of the reference sequence and move along the sequence (see a tutorial on the MISIS web-site). If more than one size-class of sRNAs is selected and displayed in one histogram, the user can visualize each size using a colour code (click "Summary"). A panel below "Summary" will indicate size-colour association (Fig. 1A): the user can click in any colour in the panel and choose a new colour in a popup window, or return to the default histogram (click "Summary"). All the histograms displayed by MISIS-2 can be saved in jpeg, png or gif formats.

To determine the 5' nucleotide frequency for selected size-classes of sRNAs the user should click "View" and "5' nucleotide distribution". A popup window will display the bar graph representing the frequency of 5' nucleotides (A, C, G, U) per each size-class for the selected region of reference sequence (Fig. 1B). Within this window, the borders of the selected region (start and end nucleotide positions) and the polarities (forward, reverse, or total) are indicated, and these parameters can be changed by the user. The bar colours can also be changed by clicking directly on the corresponding colour. An image of the bar graph can be saved in a file.

To determine and display a consensus sequence supported by the majority of sRNA reads mapped to a given reference sequence with and without mismatches and to identify and count SNPs the user should load the file "_project.b2btf" and then check "Show reference" and "Show logo" on the right panel. Note that the reference sequence and the logo sequence will be displayed below the histogram only when the minimal width of bars is 10 pixels (Fig. 1A); the user can then slide along the reference using a cursor below the logo. The logo displays the nucleotides A, C, G, or T at each position of the reference sequence, with their size corresponding to their frequency in total sRNA

reads mapped to this position. If the most frequent nucleotide deviates from the reference sequence nucleotide and is supported by more than 50% reads, it will be highlighted by dark grey background (Fig. 1A). If the deviating nucleotide is supported by 10 to 50% reads it will be highlighted by light grey background. The nucleotide supported by the majority of the sRNA reads will be integrated into the consensus sequence which can be downloaded in a fasta format (click "File" and "Export consensus sequence"). The lower and upper thresholds (10 and 50% by default) for the colour codes of the deviating nucleotides can be adjusted (click "View" and "Change the thresholds for logos"). Furthermore, the reference sequence can be replaced (click "View" and "Load the reference"). The logo will then be altered accordingly to highlight nucleotides deviating from the new reference sequence. Furthermore, counts of the sRNA reads containing each of the four possible nucleotides per each position of the reference sequence can be displayed in a popup window by right-clicking on the corresponding position of the logo (Fig. 1A). This helps to verify whether or not the highlighted SNP or mismatch is supported by a substantial number of sRNA reads. The user can download a list of the SNPs with their frequencies as a tabulated text file (click "File" and "Export table mismatches/SNP"). Note that the presence of short indels (insertions/deletions) in a majority of mapped sRNA reads will result in a shift along the logo (several successive SNPs highlighted in dark grey) in comparison to the reference. The shift's characteristics allow the user to deduce the indel's position and size.

In addition to the above-described "all nucleotides" logo, the user can select a second "5' nucleotide logo" which will display the frequency of only 5'-terminal nucleotides of sRNAs per each position of the reference sequence. This logo can point at a particular mechanism of sRNA biogenesis for a given region of the reference sequence.

2.2. Reconstruction of a consensus master genome in viral quasispecies

To test if MISIS-2 can be applied to facilitate reconstruction of a consensus virus genome in viral quasispecies, we made use of sRNA sequencing datasets obtained from *Arabidopsis* plants infected with wild-type viral particles of *Oilseed rape mosaic virus*, for which a full-length genome clone was not infections (Seguin et al., 2014b). Mapping of viral sRNAs to the clone sequence and MISIS-aided reconstruction of the consensus sequence revealed single nucleotide substitutions at three positions of the reference sequence (G-to-A at position 12, A-to-T at position 231, and G-to-T at position 5612; Fig. 1A). These substitutions were supported by 96.5–97.7% (A12), 99.5–99.9% (T231) and 88.6–93.8% (T5612) of the viral sRNA reads at two time points (14 and 22 days) post-inoculation. These apparent mismatches between the reconstructed consensus virus sequence and the cloned virus sequence could stem from mutations introduced during the cloning process, which abolished infectivity of the virus, or from accidental cloning of one of the defective mutant genomes

generated through replication errors in the evolving viral quasispecies (Domingo et al., 2012). The latter hypothesis is supported by the presence of alternative nucleotides at all the three positions of the reconstructed viral quasispecies (Fig. 1A). To verify if the reconstructed consensus genome indeed represents the biologically-active wild-type virus genome, the three mismatches were repaired in the viral full-length clone and the resulting clone was tested for infectivity. Strikingly, the repaired virus clone was fully biologically active and caused disease symptoms indistinguishable from those of the wild-type viral particles (Seguin et al., 2014b). We also tested MISIS-2 for reconstruction of consensus genomes of other plant and animal viruses and found that viral small RNAs faithfully represent the consensus master genome in viral quasispecies (to be published elsewhere). This finding is supported by comparative sequencing analysis of the viral siRNA population and the viral particles from a *Potato virus Y*-infected plant (Kutnjak et al., 2015).

Using the script implemented in MISIS-2, analysis of the 5'-nucleotide identity and frequency was performed for the three major size-classes of plant and viral small RNAs (21-nt, 22-nt and 24-nt) from virus-infected rice, banana and tomato plants. The results confirmed and extended previous findings indicating that the plant and viral siRNAs are sorted by AGO proteins based on size and 5'-nucleotide identity (Rajeswaran et al. 2014a; Rajeswaran et al. 2014b; Fuentes et al., 2016). Interestingly, a novel class of abundant 20-nt small RNAs with 5'-terminal guanidine was identified in rice and banana plants (Rajeswaran et al. 2014a; Rajeswaran et al. 2014b).

2.3. Conclusions

In summary, MISIS-2 is a simple and user-friendly tool for visualizing and analyzing sRNA maps of the viruses and cellular genome regions generating multiple sRNAs. Moreover, this tool facilitates reconstruction of a consensus master sequence and identification of SNPs and indels in the evolving viral quasispecies. MISIS-2 complements other sRNA map-view and analysis tools such as Integrative Genomic Viewer (Thorvaldsdóttir et al., 2013) (<http://www.broadinstitute.org/igv>), Seqmonk (www.bioinformatics.bbsrc.ac.uk/projects/seqmonk/) and UEA Small RNA Workbench (Stocks et al., 2012) (srna-workbench.cmp.uea.ac.uk/).

3. Materials and methods

Like MISIS (Seguin et al., 2014a), MISIS-2 was written in Java language to be compatible with Windows, Mac and Linux. The following Java Packages were implemented in MISIS-2: Biojava version 3.0.7 and 1.7 (http://biojava.org/wiki/Main_Page) (Prlić et al., 2012) to read and write fasta files, and Sam version 1.77 from Picard Tools (<http://broadinstitute.github.io/picard/>) to read

Bam/Sam files.

Acknowledgements

The work was supported by Swiss National Science Foundation (grant 31003A_143882/1 to M.M.P.) and European Cooperation in Science and Technology (COST; grant SERI No. C09.0176 to L.F. and M.M.P.).

References

- Babski, J., Maier, L.K., Heyer, R., Jaschinski, K., Prasse, D., Jäger, D., Randau, L., Schmitz, R.A., Marchfelder, A., Soppa, J., 2014. Small regulatory RNAs in Archaea. *RNA Biol.* 11, 484-493.
- Cenik, E.S., Zamore, P.D., 2011. Argonaute proteins. *Curr. Biol.* 21, R446-449.
- Domingo, E., Sheldon, J., Perales, C., 2012. Viral quasispecies evolution. *Microbiol. Mol. Biol. Rev.* 76, 159–216.
- Fuentes, A., Carlos, N., Ruíz, Y., Callard, D., Sánchez, Y., Ochagavía, M.E., Seguin, J., Malpica-López, N., Hohn, T., Lecca, M.R., Pérez, R., Doreste, V., Rehrauer, H., Farinelli, L., Pujol, M., Pooggin, M.M., 2016. Field trial and molecular characterization of RNAi-transgenic tomato plants that exhibit resistance to tomato yellow leaf curl geminivirus. *Mol. Plant Microbe Interact.* Feb 11: MPMI08150181R. [Epub ahead of print]..
- Ghildiyal, M., Zamore, P.D., 2009. Small silencing RNAs: an expanding universe. *Nat. Rev. Genet.* 10, 94-108.
- Kreuze, J.F., Perez, A., Untiveros, M., Quispe, D., Fuentes, S., Barker, I., Simon, R., 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.
- Kutnjak, D., Rupa, M., Gutierrez-Aguirre, I., Curk, T., Kreuze, J.F., Ravnika, M., 2015. Deep sequencing of virus-derived small interfering RNAs and RNA from viral particles shows highly similar mutational landscapes of a plant virus population. *J. Virol.* 89, 4760-4769.
- Langmead, B., Trapnell, C., Pop, M., Salzberg, S.L., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10, R25.
- Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754-1760.
- Meister, G., 2013. Argonaute proteins: functional insights and emerging roles. *Nat. Rev. Genet.* 14, 447-459.
- Prlić, A., Yates, A., Bliven, S.E., Rose, P.W., Jacobsen, J., Troshin, P.V., Chapman, M., Gao, J., Koh, C.H., Foisy, S., Holland, R., Rimsa, G., Heuer, M.L., Brandstätter-Müller, H., Bourne, P.E., Willis, S., 2012. BioJava: an open-source framework for bioinformatics in 2012. *Bioinformatics* 28, 2693-2695.
- Rajeswaran, R., Golyaev, V., Seguin, J., Zvereva, A.S., Farinelli, L., Pooggin, M.M., 2014a. Interactions of Rice tungro bacilliform pararetrovirus and its protein P4 with plant RNA-silencing machinery. *Mol. Plant Microbe Interact.* 27, 1370-1378.
- Rajeswaran, R., Seguin, J., Chabannes, M., Duroy, P.O., Laboureau, N., Farinelli, L., Iskra-Caruana, M.L., Pooggin, M.M., 2014b. Evasion of short interfering RNA-directed antiviral silencing in *Musa acuminata* persistently infected with six distinct banana streak pararetroviruses. *J.*

- Virol. 88, 11516-11528. Seguin, J., Otten, P., Baerlocher, L., Farinelli, L., Pooggin, M.M., 2014a. MISIS: a bioinformatics tool to view and analyze maps of small RNAs derived from viruses and genomic loci generating multiple small RNAs. *J. Virol. Methods* 195, 120-122.
- Seguin, J., Rajeswaran, R., Malpica-López, N., Martin, R.R., Kasschau, K., Dolja, V.V., Otten, P., Farinelli, L., Pooggin, M.M., 2014b. De novo reconstruction of consensus master genomes of plant RNA and DNA viruses from siRNAs. *PLoS One* 9, e88513.
- Stocks, M.B., Moxon, S., Mapleson, D., Woolfenden, H.C., Mohorianu, I., Folkes, L., Schwach, F., Dalmay, T., Moulton, V., 2012. The UEA sRNA workbench: a suite of tools for analysing and visualizing next generation sequencing microRNA and small RNA datasets. *Bioinformatics* 28, 2059-2061.
- Thorvaldsdóttir, H., Robinson, J.T., Mesirov, J.P., 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* 14, 178-192.
- Wu, Q., Luo, Y., Lu, R., Lau, N., Lai, E.C., Li, W.X., Ding, S.W., 2010. Virus discovery by deep sequencing and assembly of virus-derived small silencing RNAs. *Proc. Natl. Acad. Sci. U S A* 107, 1606-1611.
- Zerbino, D.R., and Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 18, 821–829.

Legends to Figures

Fig. 1. An example of a viral sRNA map created, visualized and analyzed by MISIS-2. (A)

Single nucleotide resolution map of 20-25 nt viral sRNAs derived from *Oilseed rape mosaic virus* (ORMV) in *A. thaliana* plants at 14th day post inoculation with the wild-type viral particles (Seguin et al., 2014b). The user loaded a BWA-generated BAM file (deposited at the MISIS-2 web-site for tutorial) with viral sRNAs mapped to the ORMV reference sequence and let MISIS-2 create the table and the project files. Then the user loaded the project file "ORMV_project_20_25.b2btf" and checked "Summary" button. The resulting histogram depicts the map of viral sRNAs of six size-classes (20, 21, 22, 23, 24, and 25 nt, selected by default and color-coded as indicated on the right panel). The histogram's x-axis indicates the ORMV genome sequence positions (from 1 to 6303) against which the multiple viral sRNAs are mapped. The y-axis indicates the counts of each mapped sRNA (from 0 to 7000 reads) in the forward (bars above the x-axis) and reverse (bars below the x-axis) orientations. The size of each bar corresponds to the total count of 20-25 nt reads mapped to this position with their 5'-terminal nucleotide for the forward reads or their 3'-terminal nucleotide for the reverse reads. The colors within each bar indicate the relative proportion of each size-class in the total counts. To determine and visualize the consensus sequence and the SNP mismatches to the ORMV reference sequence, the user checked "Show reference" and "Show logo" buttons and zoomed in three regions of the viral genome as shown in three separate images below the histogram. The dark grey background indicates the nucleotides that do not match to the reference nucleotide at the positions 12, 231 and 5612 of the virus genome. The popup windows inside the images (created by right-clicks on the corresponding nucleotides of the logo) show the numbers of sRNAs reads containing the SNP at these positions. **(B)** The bar graph diagram showing the frequencies of A, G, C, and U at the 5'-terminal base position for each size-class of the 20-25 nt viral sRNAs mapped to both strands of the ORMV genome reference sequence. To visualize this bar graph in the popup window the user clicked on "View" and "5'-nucleotide distribution".

A. single nucleotide resolution map of viral sRNAs

B. sRNA 5' nucleotide frequency

SNP/Mismatch: G12A

SNP/Mismatch: A231T

SNP/Mismatch: G5612T