Next Generation Sequencing for studying viruses and RNA silencing-based antiviral defense in crop plants

Inauguraldissertation

zur

Erlangung der Würde eines Doktors der Philosophie

vorgelegt der

Philosophisch-Naturwissenschaftlichen Fakultät

der Universität Basel

von

Jonathan Seguin

von Frankreich

Basel, 2016

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät auf Antrag von Prof. Thomas Boller, PD Dr. Mikhail M. Pooggin und Prof. Mihaela Zavolan.

Basel, 9 dezember 2014

Prof. Dr. Jörg Schibler

General Preface

Financial support of this PhD work was provided by the COST action 'Food and Agriculture' (FA) 0806, which has for final objective to create a RNA-based vaccine to immunize crop plants against viral infection. This work was done within a collaboration between the Fasteris SA company, directed by Dr. Laurent Farinelli, and the team of Dr. Mikhail Pooggin from the Plant Physiology research group in Botany at the University of Basel. The expertise of Dr. Laurent Farinelli's company was requested in the objective to use Illumina-Solexa technology to sequence small RNA and perform bioinformatic analysis. The expertise of Dr. Mikhail Pooggin was requested in order to study the defense mechanisms based on sRNA within plant infected by Geminiviruses, Pararetroviruses and Tobamoviruses. Consequently, this work as part of the action FA0806 involved several different other collaborations with COST member european scientists.

Acknowledgments

I would like to thanks and prove my gratitude to all people who help me during my whole doctoral study and who make this thesis possible.

First, I would like to thank my supervisors, PD Dr. Mikhail Pooggin and Dr Laurent Farinelli for giving me the chance to do my thesis in collaboration with Fasteris and the University of Basel. I want to thank Dr. Pooggin for making me share his knowledge and guidance in data analysis and writing of the thesis. I want to thank Dr. Farinelli to have always welcomed me in his company, and have provided me all the technical and IT resources of Fasteris that I needed for the good progress of my thesis.

I thank my lab members : Rajeswaran Rajendran, Nachelli Malpica, Anna Zvereva, Victor Golyaev, Silvia Turco and Katya Ivanova for their warm hospitality and their invaluable assistance in understanding their respective project. I would especially like to thank Prof. Thomas Boller for accepting to be my representative in the faculty of botany and Prof. Mihaela Zavolan for being a member of my PhD committee. I also thank Prof. Thomas Hohn and all the other members of the Botanical Institute for their kindness.

I want to thank the Bioinformatic team of Fasteris SA: Patricia Otten especially who supervised me at Fasteris and teached me in the analysis of sequenced sRNA; Loic Baerlocher, William Baroni, Nicolas Gonzalez, and Julien Prados for their valuable advice. I would also like to thank Cécile Deluen and Magne Osteras for their explanation of the Solexa-Illumina NGS technologies, and to finish thanks to all Fasteris employees for their hospitality.

Special acknowledgements go to our collaborators: Matthew Chabannes, Pierre-Olivier Duroy and Marie-line Caruana from CIRAD for their collaboration on the analysis of Banana samples infected with BSV; Valerian Dolja and his team for the analysis of infected vines samples in Oregon; Madurai Kamaraj University for the analysis of cassava infected with ICMV and SLCMV, and Dr Basanta Borah and Dr. Basavaprabhu L. Patil; and finally a big thank to Alejandro Fuentes for providing the tomatoes samples infected with TYLCV. I would also like to thank Prof. Andreas Voloudakis the coordinator of COST Action FA0806 for supporting the Farinelli/Pooggin project application to the Swiss COST fund for financial support of my thesis work .

Finally, I would also like to thank my family for supporting me all these years in my personal and professional choices. I also want to thank my wife Hélène Méreau for her help and support, and my daughter Louise for the happiness and joy that inspires me daily.

Abstract

The main objectives of this work have been to use next generation sequencing (NGS) and develop bioinformatics tools for plant virus diagnostics and genome reconstruction as well as for investigation of RNA silencing-based antiviral defense. In virus-infected plants, the host Dicer-like (DCL) enzymes process viral double-stranded RNAs into 21-24 nucleotide (nt) short interfering RNAs (siRNAs) which can potentially associate with Argonaute (AGO) proteins and guide the resulting RNA-induce silencing complexes (RISCs) to target complementary viral RNA for post-transcriptional silencing and, in the case of DNA viruses, complementary viral DNA for transcriptional silencing. In the pioneering work, Kreuze et al. (2009) have demonstrated that an RNA virus genome can be reconstructed from multiple siRNA contigs generated by the short sequencing read assembler Velvet.

In this PhD study, we developed a bioinformatics pipeline to analyze viral siRNA populations in various model and crop plants experimentally infected with known viruses and naturally infected with unknown viruses. First, we developed a bioinformatics tool (MISIS) to view and analyze maps of small RNAs derived from viruses and genomic loci that generate multiple small RNAs (Seguin et al. 2014b). Using MISIS, we discovered that viral siRNAs cover the entire genomes of RNA and DNA viruses as well as viroids in both sense and antisense orientation without gaps (Aregger et al. 2012; Seguin et al. 2014a; Rajeswaran et al. 2014a, 2014b), thus allowing for de novo reconstruction of any plant virus or viroid from siRNAs. Then, we developed a de novo assembly pipeline to reconstruct complete viral genomes as single contigs of viral siRNAs, in which Velvet was used in combination with other assemblers: Metavelvet or Oases to generate contigs from viral redundant or non-redundant siRNA reads and Seqman to merge the contigs. Furthermore, we employed the mapping tool BWA and the map viewing tool IGV to verify the reconstructed genomes and identify a consensus master genome and its variants present in the virus quasispecies. The approach combining deep siRNA sequencing with the bioinformatics tools and algorithms, which enabled us to reconstruct consensus master genomes of RNA and DNA viruses, was named siRNA omics (siRomics) (Seguin et al. 2014a).

We utilized siRomics to reconstruct a DNA virus and two viroids associated with an emerging grapevine red leaf disease and generate an infectious wild type genome clone of oilseed rape mosaic virus (Seguin et al. 2014a). Furthermore, siRomics was used to investigate siRNA-based antiviral defense in banana plants persistently infected with six distinct banana streak pararetroviruses (Rajeswaran et al. 2014a) and rice plants infected with rice tungro bacilliform pararetrovirus (Rajeswaran et al. 2014b). Our results revealed that multiple host DCLs generate abundant and diverse populations of 21-, 22- and 24-nt viral siRNAs that can potentially associate with multiple AGO proteins to target viral genes for post-transcriptional and transcriptional

silencing. However, pararetroviruses appear to have evolved silencing evasion mechanisms such as overexpression of decoy dsRNA from a short non-coding region of the virus genome to engage the silencing machinery in massive siRNA production and thereby protect other regions of the virus genome from repressive action of viral siRNAs (Rajeswaran et al. 2014b). Furthermore, despite massive production of 24-nt siRNAs, the circular viral DNA remains unmethylated and therefore transcriptionally active, while the host genome is extensively methylated (Rajeswaran et al. 2014b). These findings shed new light at the siRNA generating machinery of economically-important crop plants. Our analysis of plant small RNAs in banana and rice revealed a novel class of highly abundant 20-nt small RNAs with 5'-terminal guanidine (5'G), which has not been identified in dicot plants. Interestingly, the 20-nt 5'G-RNA-generating pathway does not target the pararetroviruses, which correlates with silencing evasion (Rajeswaran et al. 2014a, 2014b).

This thesis work is a part of the European Cooperation in Science and Technology (COST) action that aims develop an RNA-based vaccine to immunize crop plants against viral infection. Our analysis of viral siRNA profiles in various virus-infected plants allowed to identify the regions in the viral genomes that generate low-abundance siRNAs, which are the candidate regions to be targeted by RNA interference (RNAi). Our analysis of RNAi transgenic tomato plants confirmed that targeting of the low-abundance siRNA region of Tomato yellow leaf curl virus (TYLCV) by transgene-derived siRNAs renders immunity to TYLCV disease, one of the major constraints for tomato cultivation worldwide.

Table des matières

List of abbreviations

AGO	Argonaute protein
BAM	Binary Alignment/Map
BSCAV	Banana Streak Cavendish Virus
BSGFV	Banana Streak Goldfinger Virus
BSIMV	Banana Streak Imove Virus
BSMYV	Banana Streak Mysore Virus
BSOLV	Banana Streak Obino l'Ewai Virus
BSV	Banana Streak Virus
BSVNV	Banana Streak Vietnam Virus
BWA	Burrows-Wheeler Alignment
BWT	Burrows-Wheeler Transform
CaLCuV	Cabbage Leaf Curl Virus
CaMV	Cauliflower Mosaic Virus
CP	Coat Protein
DCL	DiCer-Like protein
dsDNA	double-stranded DNA
dsRNA	double-stranded RNA
eBSV	endogenous BSV
ELISA	Enzyme-Linked ImmunoSorbent Assay
GRBaV	Grapevine Red Blotch-associated Virus
GRLaV	Grapevine RedLeaf-associated Virus
GUI	Graphical User Interface
GVGV	GrapeVine GeminiVirus
GYSVd	Grapevine Yellow Speckle Viroid
HSVd	Hop Stunt Viroid
ICMV	Indian Cassava Mosaic Virus
IGV	Integrative Genomics Viewer
miRNA	microRNA
MP	Movement Protein
MVC	Model-View-Controller
NGS	Next-Generation Sequencing
OLC	Overlap/Layout/Consensus

ORF	Open Reading Frame
ORMV	Oilseed Rape Mosaic Virus
PCR	Polymerase Chain Reaction
pri-miRNA	primary miRNA
pre-miRNA	precursor miRNA
PTGS	Post-Transcriptional Gene Silencing
RCA	Rolling Circle Amplification
RdDM	siRNA-directed DNA methylation
RDR	RNA-Dependant RNA polymerase
RISC	RNA-Induced Silencing Complex
RTBV	Rice Tungro Bacilliform Virus
RT-PCR	Real-Time Polymerase Chain Reaction
RTSV	Rice Tungro Spherical Virus
SAM	Sequence Alignment/Map
siRNA	short interfering RNA
SLCMV	Sri Lankan Cassava Mosaic Virus
SNP	Single Nucleotide Polymorphism
ssDNA	single-stranded DNA
ssRNA	single-stranded RNA
TGS	Transcriptional Gene Silencing
TMV	Tobacco Mosaic Virus
TYLCV	Tomato Yellow Leaf Curl Virus
vsiRNA	viral siRNA
WGS	Whole-Genome Shotgun

1. Introduction

Since the Neolithic age, humans domesticated wild plants to avoid hazards of the nature and to secure their food. The period of hunting and gathering gave way to the period of farming and breeding. Without need to move constantly and find food, the first great civilizations emerged: Sumerians, Babylonians, Egyptian, Chinese, etc. All these civilizations have left or continuously leave their marks in the history of the mankind. The control of agricultural crops is therefore in the basis of the human civilizations. Until today, 2500 wild plant species have been domesticated including 203 major and minor crop species (Meyer et al., 2012).

Big epidemics affecting crop plants had provoked famines which killed many thousands of humans. Moreover, human migration waves took place because of these epidemics. Until today, 14 crop plants supplied the majority of food for the human consumption. Nevertheless, more than 800 million humans do not have adequate food, and, in addition, it is considered that at least 10 % of worldwide food production is lost due to plant diseases. The sensitivity of crop plants to diseases is due to the intensity of agriculture strategy (e.g., use of monoculture, large fields size, etc.) and the genetic selection by humans. Major groups of pathogens that cause disease on crop plants are viruses, bacteria, oomycetes, fungi, nematodes and parasitic plants. More than 700 plant viruses are known to be involved in devastating diseases and many of those have wide host ranges (Strange and Scott, 2005).

Tobacco mosaic virus (TMV) was the first virus discovered more than 100 years ago, based on its ability to pass through the porcelane filter (which retains bacteria and fungi) and infect a new host plant. Since then TMV has provided the best model system for the virology research. Like all other viruses, TMV cannot reproduce itself without a host cell, which is the main feature distinguishing viruses from other domains of life. TMV is the type member of genus *Tobamovirus* of the family *Virgaviridae*. TMV and other tobamoviruses have provided the best model system for the virology research. TMV and other tobamoviruses have an RNA genome of ca. 6 kb encoding 4 proteins: the first two involved in RNA replication (p126 and p183), the third involved in viral movement (MP) and the fourth in RNA encapsidation (CP) (Creager et al., 1999). With the discovery of TMV, scientists tried to find viruses that cause plant disease in absence of other pathogenic factors such as bacteria, fungi or animals. In 2009, the International Committee for the Taxonomy of Viruses, ICTV, (King et al., 2011) listed up to 900 species of plant viruses. The majority of these plant viruses have been discovered and studied because of their impact on domesticated and cultivated plants and plant disease (Roossinck, 2011). However, viruses occur not only on cultivated but also on non-cultivated wild plants and can be neutral or even beneficial (Roossinck, 2011) The recent development of next generation sequencing technologies which allow to sequence a high number of genomes and metagenomes at low cost, have increased the

number of known plant virus species and this number is expected to raise in the future.

Given the growing number of plant viruses, ICTV has further improved their taxomonical classification, placing plant viruses in 3 orders, 21 families and 92 genera. The main taxonomic criteria are based on the viral particle structure, the mode of replication and the type of genome. The viral genome can be single or double-stranded DNA and RNA molecules. Moreover, the polarity of virus genome encapsidated into the viral particle can be positive, negative or ambisense. The genome can be contained in one molecule (monopartite) or divided in two or more molecules (bi-partite, multipartite), and can contain satellite molecules coding for optional genes. The mode of virus replication depends on the viral genome organisation and protein content. The viral particles differ in size, shape, presence/absence of envelope and other parameters (Hull, 2013) (Fig 1.1.1).

1.1 Descriptions of plant virus families

The majority of known plant viruses are single-stranded, positive-sense RNA viruses like TMV as they encode all the viral proteins on the genomic RNA encapsidated in the virion. There also exist RNA plant viruses with double-stranded genomes and single-stranded genomes of negative-sense or ambisense polarities as well as viruses with double-stranded or single-stranded DNA genomes (Scholthof et al., 2011, Hull, 2013) (Fig 1.1.1).

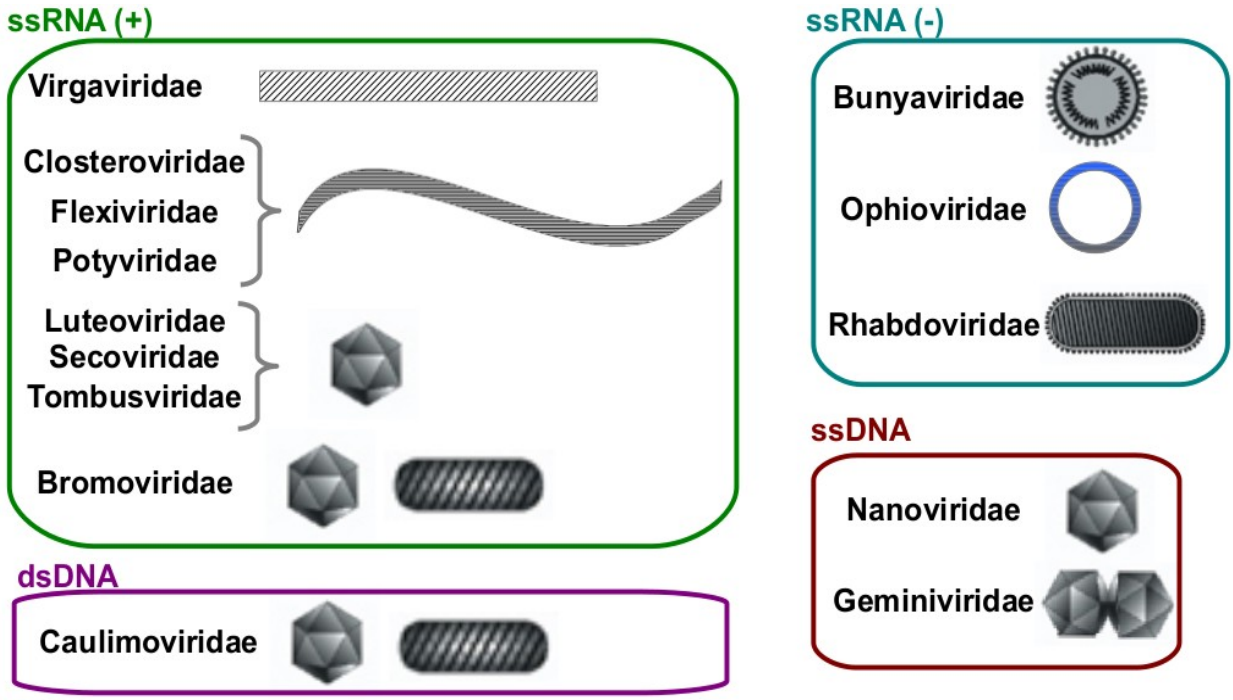


Figure 1.1.1: Representation of capsid structures in families of plant viruses.

Positive-sense single-stranded RNA viruses are classified into the families *Virgaviridae*, *Flexiviridae*, *Closteroviridae*, *Secoviridae*, *Tombusviridae*, *Bromoviridae*, *Luteoviridae* and *Potyviridae. Virgaviridae* including TMV possess alpha-like RNA-dependent RNA polymerase (RDR, or replicase), a single-stranded RNA genome with a 3'-tRNA-like structure and no polyA tail, rod-shaped virions of 20-25 nm in diameter, compounded by coat protein of 20-24 kDa, with a central « canal » (Adams et al., 2009). *Closteroviridae* share similar properties with the *Virgaviridae* family*,* but the virions and RNA genomes of closteroviruses are the longest among plant viruses. The length of *Closteroviridae* genomes varies between ~15,5 to ~19,5 kb which encode between 10 to 14 proteins. Moreover, the flexuous and filamentous virion incorporates at least five proteins that are assembled into a long helical body and a short segment tail (Dolja et al., 2006*).* The members of the *Flexiviridae* family have flexuous elongated helical virions. Their genomes are 3'-polyadenylated, monopartite, with a length between 6 to 9 kb. The filamentous virions vary from ~470 nm to ~1000 nm according to the viral genera, and have a diameter of 12-14 nm (Martelli et al., 2007). *Secoviridae* is a family characterized by a capsid which is always icosahedral with pseudo-T = 3 symmetry composed by 2 or 3 coat proteins. Their genomes can be mono- or bi-partite. This family was recently created by the fusion of *Sequiviridae* and *Comoviridae* families (Sanfaçon et al., 2009). *Tombusviridae* contains plant viruses which have their genome packaged into spherical particles with diameters of ~30 nm. The viral genome is polycistronic and is often not segmented. Moreover, they use the 3' cap-independent translational enhancers (3'CITEs) to mediate efficiently the translation of encoded viral proteins (Jiwan and White, 2011). *Bromoviridae* is the virus family responsible for major crop plant epidemics worldwide. Their genome is tripartite. Two subunits of the replicates are encoded by RNA1 and RNA2. The MP and the CP are encoded by RNA3. These three RNAs are encapsidated separately. The structure of the capsid in different genera of *Bromoviridae* can be bacilliform, spheroidal, or (quasi-)spherical (Bol, 1999). The genome of *Luteoviridae* is encapsided within non-enveloped icosahedral particles where the diameter is between 25 to 28 nm (Gray and Gildow, 2003). *Potyviridae* is the largest family of plant viruses. Their virions are flexuous, non rod-shaped and composed of around 2000 units of CP. The common feature of *Potyviridae* is the induction of characteristic pinwheel or scroll-shaped inclusion bodies in the cytoplasm during the infection. In the majority of *Potyviridae* viruses, the genome encodes up to 10 proteins which are multifunctional (Rohozková and Navrátil, 2011).

Negative-sense single-stranded RNA viruses, which encode major proteins such as RDR on the negative strand that is not encapsidated, are classified in the families *Bunyaviridae*, *Ophioviridae,* and *Rhabdoviridae.* The *Bunyaviridae* family is characterized by a tripartite RNA genome encapsided within spherical, membrane bound with a diameter of approximately 80-120 nm. The genome segments have either negative or ambisense polarities. The *Ophioviridae* family contains multipartite viruses. Their genomes are encapsided within linear virions. The virions are

composed by circular filamentous ribonucleoproteins about 3nm in diameter of different contour lengths. The genome is divided in three or four segments. The *Rhabdoviridae* family contains viruses which have monopartite negative-sense single-stranded RNA encapsided within bullet-shaped or bacilliform virions. This family is divided into 2 genera according to the localization of their replication within the infected cells: *Nucleorhabdoviruses* replicate in the nucleus, whereas *Cytorhabdoviruses* in the cytoplasm (Kormelink et al., 2011).

The *Endornaviridae* family contains viruses which have double-stranded RNA genome without evidence of encapsidation. The majority of *Endornaviridae* viruses have one very long ORF, a nick in the plus strand and a poly-C at the 3' end. Only the RDR domain is conserved among these viral species (Roossinck et al., 2011).

Plant DNA viruses are classified in three families: *Geminiviridae*, *Nanoviridae*, and *Caulimiviridae*. The *Geminiviridae* family includes viruses with a twinned icosahedral capsid and a circular single-stranded DNA genome of ca. 2.5 to 3.2 kb. They can be mono- or bi-partite. The monopartite geminiviruses can be accompanied by satellite circular ssDNA molecules of smaller size: alphasatellites and betasatellites (Nawaz-ul-Rehman and Fauquet, 2009). The *Nanoviridae* family contains viruses with multipartite circular single-stranded DNA genome. Each of the six to eight DNA components of about 1kb in length is packed separately within small isometric virions (Gronenborn, 2004). Both geminiviruses and nanoviruses replicate their DNA in the nucleus by a rolling circle-type mechanism. The members of the *Caulimoviridae* family package their double-stranded DNA genome within icosahedral or bacilliform virions. The *Caulimoviridae* are also known as pararetroviruses because they use a viral reverse transcriptase to replicate the genome from a pregenomic RNA (Hohn et al., 2001). Contrary to the mammalian retroviruses, their genomes are not obligatorily integrated into the host genome. Their circular DNA genomes accumulate as multiple minichromosomes (episomes) in the nucleus of infected plant cells and can have a length between 7,5 and 9,3 kb (Hohn and Rothnie, 2013).

In addition to viruses, plants can be infected with viroids which are circular and non-protein-coding RNA molecules of 246 to 401 nucleotides in length. They depend totally on the host proteins to replicate their genome by a rolling-circle mechanism. The 30 known species are classified into two families. The *Pospiviroidae* family contains viroids which replicate in the nucleus and have a rod-like secondary structure. The *Avsunviroidae* family contains viroids which replicate within plastids (Navarro et al., 2012).

1.2 Viruses investigated in this study

The Pooggin group at the University of Basel studies mostly DNA viruses with the emphasis on *Cauliflower mosaic virus* (CaMV) and *Cabbage leaf curl virus* (CaLCuV) which belong to the *Caulimoviridae* (pararetroviruses) and *Geminiviridae* families, respectively, and can infect the model plant *Arabidopsis thaliana*. For comparative studies in *Arabidopsis*, the RNA tobamovirus *Oilseed rape mosaic virus* (ORMV) of the *Virgaviridae* family is also used. In addition to these three viruses, in this study, we have investigated economically-important DNA viruses that cause severe diseases in crop plants including the *Caulimoviridae* family *Banana streak virus* (BSV) in banana and *Rice tungro bacilliform virus* (RTBV) in rice, and the *Geminiviridae* family *Indian/Sri Lankan cassava mosaic virus* (ICMV, SLCMV) and *Grapevine red blotch-associated virus* (GRBaV) in cassava and grapevine, respectively.

1.2.1 Cauliflower mosaic virus

Cauliflower mosaic virus (CaMV) is a DNA virus from genus *Caulimovirus* of the family *Caulimoviridae* (pararetroviruses). The CaMV genome is double-stranded, open-circular, and has around 8000 bp (in different virus strains/isolates). It is the first completely sequenced pararetrovirus (Franck et al., 1980). The main characteristic of pararetroviruses is that they replicate by reverse transcription using a pregenomic RNA intermediate as a template for the virus-encoded reverse transcriptase to produce the double-stranded DNA genome (Hoh et al., 2010). As a consequence of this mode of replication, there are three different forms of viral genome during the infection : the circular double-stranded DNA with discontinuities on both strands (left after reverse transcription) stored in virus particles, a covalently-closed double-stranded DNA accumulating as multiple mini-chromosome in nucleus and the pregenomic RNA which is produced in the nucleus by the host DNA-dependent RNA polymerase II transcription and present in the cytoplasm as a template for translation of viral proteins on ribosomes and for reverse transcription in the inclusion bodies producing viral particles (virions) (Khelifa et al., 2010; Haas et al., 2002).

The CaMV genome encodes seven open reading frames (ORF I to VII). All ORFs are located on one strand, and are separated or overlap by a few nucleotides, except for ORF VI. The ORF VI is separated by two intergenic regions of about 150 and 700 bp containing regulatory sequences. Two capped and polyadenylated RNA, named 35S and 19S, are transcribed by the cellular RNA polymerase II by using the CaMV mini-chromosome as template. The small intergenic regions contains the promoters of 19S and 35S RNAs. The 35S RNA covers the total genome plus about 180 nucleotides. It serves as a polycistronic messenger RNA for synthesis of proteins P1 to

P5 (respectively ORF I to ORF V) and as template for reverse transcription. The 19S RNA encodes the P6 protein, its promoter is located in the small intergenic region. P1 protein forms tubules through the plasmodesmata to allow a cell-to-cell movement (Haas et al., 2002). P2 is involved in the interaction between the virion and the stylet of aphid which serves as an insect vector of CaMV. P3 allows to link P2 and the virion (Drucker et al., 2002). P3 is also involved in capsid formation (Hoh et al., 2010). P4 is the precursor of the capsid proteins (Fig 1.2.1.1). P5 is a polyprotein essential for the replication of the viral genome. P6 is a multifunctional protein: it acts as a translational transactivator which promotes translation of ORFs, it plays a role in virus cell-to-cell movement, it is a suppressor of RNA-silencing, and it is the major genetic determinant of pathogenicity of infected plant (Love et al., 2012). The function of P7 is not known (Haas et al., 2002). The deletion of ORF VII does not have any incidence on viral infectivity, but mutagenesis of its initiation codon is reverted at a high frequency. P7 seems to be unstable because it is targeted by the protease function of P5. Moreover, yeast two-hybrid experiments show that P6 interacts with P7 (Lutz et al., 2012) (Fig 1.2.1.2).

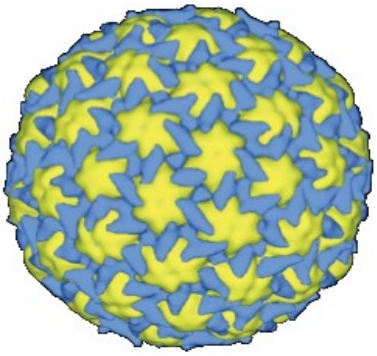


Figure 1.2.1.1: Structural model of the CaMV capsid.

The blue structures are formed by P3, and the yellow structures are formed by P4. Adapted from (Plisson et al., 2005).

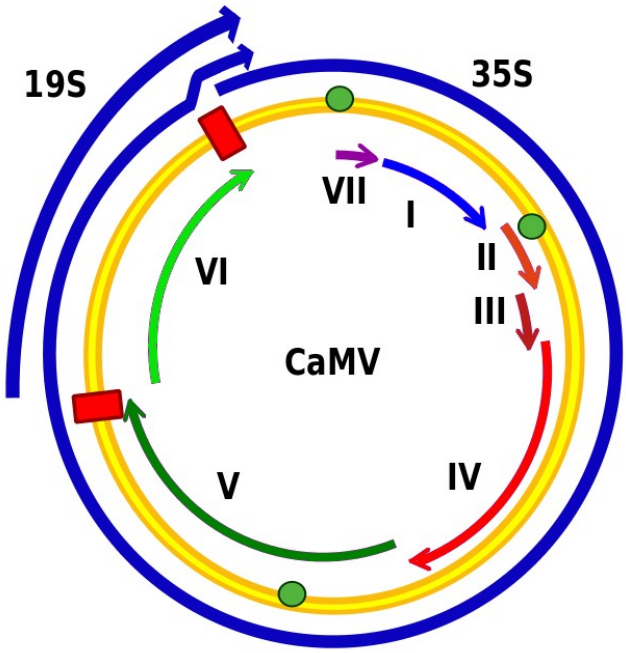


Figure 1.2.1.2: Schematic diagram of the CaMV genome.

Thin yellow lines represent the double-stranded circular DNA (8 kbp) with the green circles indicating the positions of the single-strand discontinuities. The red blocks show the positions of promoters. The coloured arrows represent the ORFs: the cell-to-cell movement protein (I), aphid transmission factors (II and III), the precursor of the capsid proteins (IV), the precursor of aspartic proteinase, reverse transcriptase and Rnase H (V), and an inclusion body protein/translational transactivator (VI). The two external arrowed lines correspond to the 35S and 19S RNA. Adapted from (Hull, 2001).

For the replication, the viral genome moves to the nucleus. The gaps of the viral genome are closed by host enzymes. Then, the double-stranded viral DNA forms one minichromosome with the host histones. The host Pol II transcribes the minichromosomes in two RNAs which move to the cytoplasm. The shorter RNA (19S) is the mRNA of P6 which produces the viroplasm protein. The 35S RNA is the RNA template for the reverse transcription. A host methionyl tRNA molecule is associated to the 35S RNA to form the primer for the reverse transcription. The RNAse H activity of P5 releases the RNA from the RNA:DNA duplex. The reverse transcription is finished when it degrades the tRNA. Two short regions of RNA are not degraded by RNAse H activity and are used as primers to replicate the complementary-sense of viral genome. The final degradation of these two primers and the tRNA are responsible of the presence of gaps within the DNA viral genome of CaMV (Matthews and Hull, 2002) (Fig 1.2.1.3).

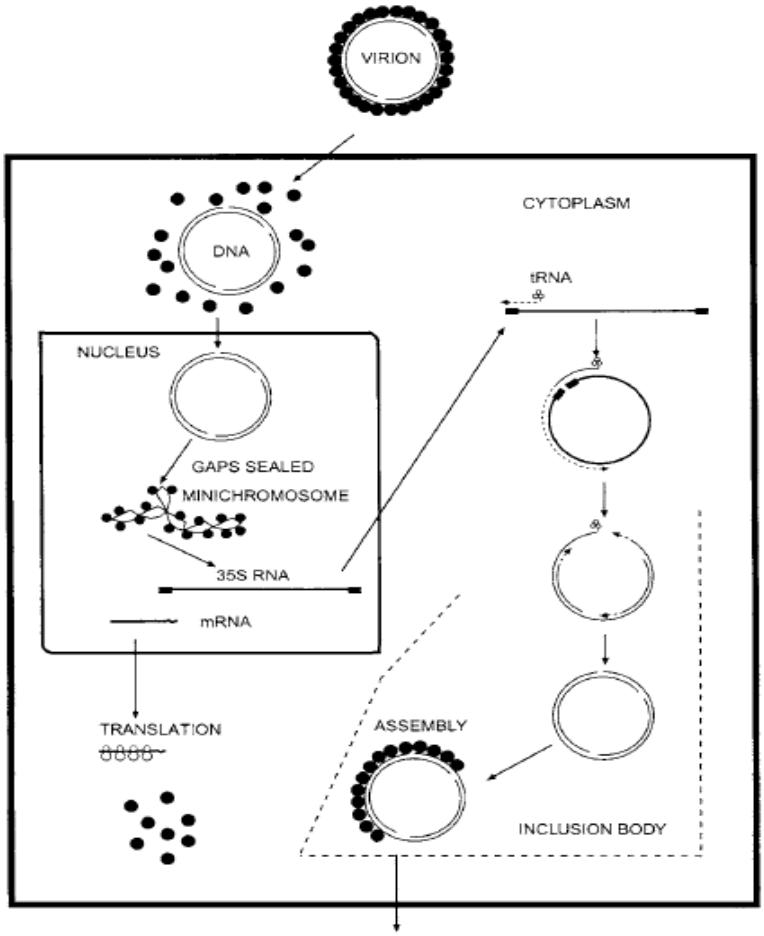


Figure 1.2.1.3: Diagram of the replication cycle of CaMV. Reprinted from (Matthews and Hull, 2002), Copyright (2002), with permission from Elsevier.

In the cell, the dsDNA genome is uncoated and moves to the nucleus. In the nucleus, the dsDNA forms a minichromosome with the host histones. Host RNA pol II transcribes the 35S and 19S RNA. These two RNAs move to the cytoplasm. The 19S RNA is translated into the viroplasm proteins. The 35S RNA is used as template to replicate the viral dsDNA genome by the viral reverse-transcriptase. When the genome is totally replicated, it is encapsided with the CP proteins and the resulting virion moves out the cell.

1.2.2 Banana streak virus

Banana streak virus (BSV) is a DNA virus from genus *Badnavirus* of the family *Caulimoviridae*. The BSV genome is circular and double-stranded. Its size is around 7,4 kb. It encodes three ORFs on one strand. The virions measure 120-150 x 27nm (Harper et al., 2002). The two first ORFs (I and II) encode two small proteins of unknown function. The last ORF encodes a polyprotein containing a cell-to-cell movement protein, the coat protein, and aspartic protease and the viral replicase with the reverse transcriptase and ribonuclease H functions (Fig 1.2.2.1). The genome is protected within bacilliform-shaped virions. Moreover, some BSV can provoke diseases in banana plants (*Musa spp*) due to infectious endogenous BSV (eBSV) sequences integrated in the plant genome (Iskra-Caruana et al., 2010).

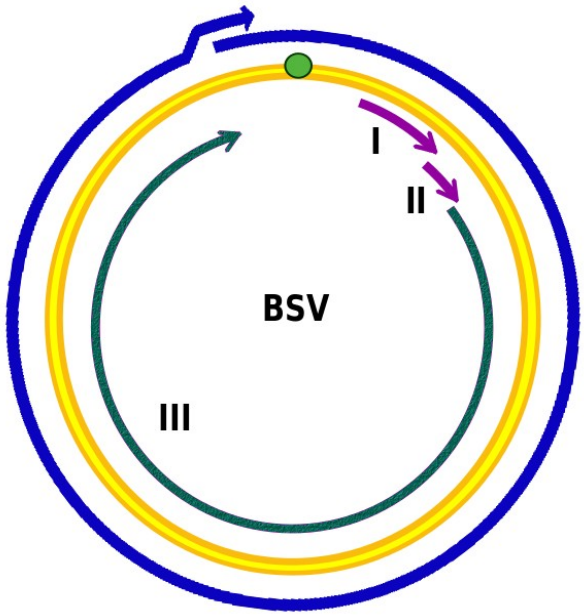


Figure 1.2.2.1: Schematic diagram of the BSV genome.

Thin yellow lines represent the double-stranded circular DNA (7.4 kbp) with the green circle indicating the positions of the single-strand discontinuities. The coloured arrows represent the ORFs: proteins of unknown function (I and II) and the polyprotein which encodes the movement protein, the coat protein, an aspartate protease and the reverse-transcriptase associated with the ribonuclease H (III). The external arrowed line corresponds to the 35S RNA. Adapted from (Hull, 2001) and (Harper et al., 1999).

1.2.3 Rice tungro bacilliform virus

Rice tungro bacilliform virus (RTBV) is the only known virus from genus *Tungrovirus* of the family *Caulimoviridae*. Like *Badnavirus*, *Tungrovirus* have bacilliform particles and no cytoplasmic inclusion bodies (Hull, 2001). The size of the RTBV genome is 8002 bp (Hay et al., 1991). To replicate its circular and double-stranded genome, it is transcribed into a pregenomic RNA (Pooggin et al., 2012). The RTBV genome has two site-specific discontinuities due to the replication process (Banerjee et al., 2011). The genome encodes 4 ORFs, three of which are closely packed. ORF IV is separated from the ORF III by a short non-coding region and from the

ORF I by a long intergenic region. ORF I encodes a protein of 24 kDa likely involved in particle assembly. ORF II encodes a protein of 12 kDa (Hull, 1996). The 12 kDa protein interacts with the coat protein to form the capsid (Herzog et al., 2000). ORF III encodes a polyprotein of 194 kDa (Hull, 1996). The cleavage of this polyprotein allows the release of the movement protein, the single coat protein (Marmey et al., 1999), an aspartate protease and the reverse transcriptase/ribonuclease H (Banerjee et al., 2011). ORF IV encodes a protein of 46 kDa. It is expressed from a spliced mRNA. It is likely involved in the control of expression of the RTBV genome (Hull, 1996) (Fig 1.2.3.1).

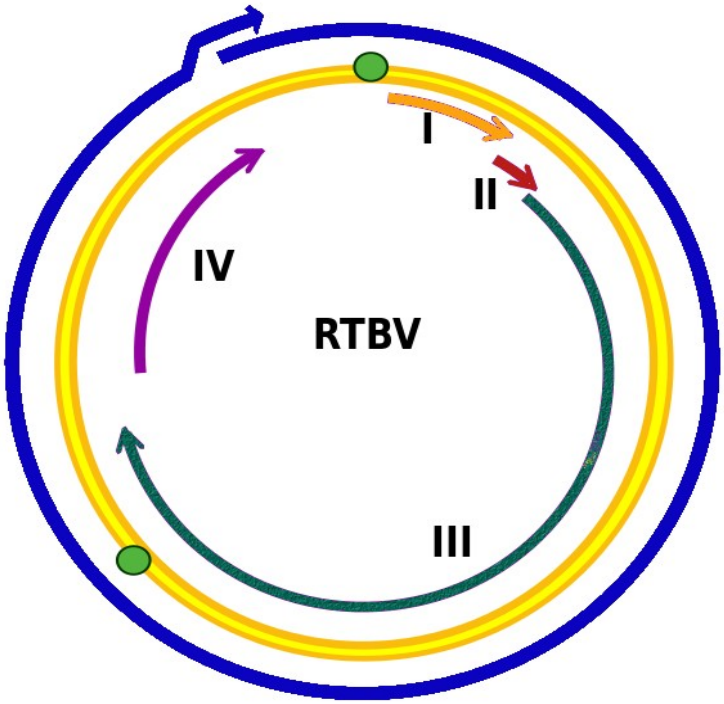


Figure 1.2.3.1: Schematic diagram of the RTBV genome.

Thin yellow lines represent the double-stranded circular DNA (8 kbp) with the green circles indicating the positions of the single-strand discontinuities. The coloured arrows represent the ORFs: P24 (I), P12 which interacts with the coat protein (II), the polyprotein P194 which encodes the movement protein, the coat protein, an aspartate protease and the reverse-transcriptase associated with the ribonuclease H (III), and P46 (IV). The external arrowed line corresponds to the 35S RNA. Adapted from (Hull, 2001).

The tungro disease of rice plants is due to the association of RTBV and *Rice tungro spherical virus* (RTSV). The tungro disease is characterized by stunting growth and yellowing or orange discolouration of leaves (Hibino, 1978). RTSV is a RNA virus which has a genome of polyadenylated single-stranded RNA of about 10 kb. RTSV is transmitted by the green leafhopper *Nephotettix virescens.* RTSV, alone, induces mild symptoms of the disease; only RTBV induces severe symptoms but it is not transmitted by leafhopper. When these two viruses infect the same plant, a transmission factor from RTSV allows for RTBV transmission by leafhopper vectors to spread in the rice fields; in these conditions, the disease provokes substantial losses of rice yields (Jones et al., 1991, Hull, 1996).

1.2.4 Cabbage leaf curl virus

Seven different genera of geminiviruses are described by ICTV according to the genome organisation and insect vectors. The genera are *Begomovirus, Mastrevirus, Curtovirus, Becurtovirus, Eragrovirus, Topocuvirus* and *Turncurtovirus* (Hanley-Bowdoin et al., 2013). *Cabbage leaf curl virus* (CaLCuV) is a bi-partite DNA virus which belongs to genus *Begomovirus*. The geminiviruses are characterized by circular single- stranded DNA genome and geminate (twinned) icosahedral virions which give the name of this family (Fig 1.2.4.1).

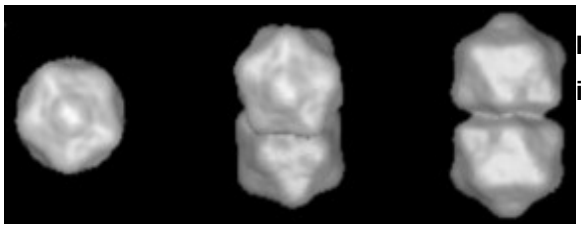


Fig 1.2.4.1: Surface representation of twinned icosahedral virions of geminivirus.

Copied from (Bottcher et al., 2004).

The size of the geminiviral genome varies between 2.5 and 2.7 kb (in the case of bipartite geminiviruses, for both DNA-A and DNA-B). The DNA-A and DNA-B genomes contain a common intergenic region of ca. 200 bp containing the origin of rolling circle replication and the bidirectional promoter driving Pol II transcription of the leftward and the rightward viral genes. The viral double-stranded DNA serving as a template for both transcription and replication, accumulates in the nucleus as multiple minichromosomes (Fig 1.2.4.2).

The CaLCuV DNA-A length is 2583 nucleotides-long and the DNA-B is 2513 nucleotides-long. The CaLCuV genome encodes seven ORFs, five in DNA-A and two in DNA-B. The DNA-A genome encodes for five proteins: two involved in replication (AC1 and AC3), transcription (AC2) and the coat protein (AV1). The DNA-B genome encodes proteins involved in movement functions: BV1 is the nuclear shuttle protein and BC1 is the movement protein (Trejo-Saavedra et al., 2009; Aregger et al., 2012) (Fig 1.2.4.3).

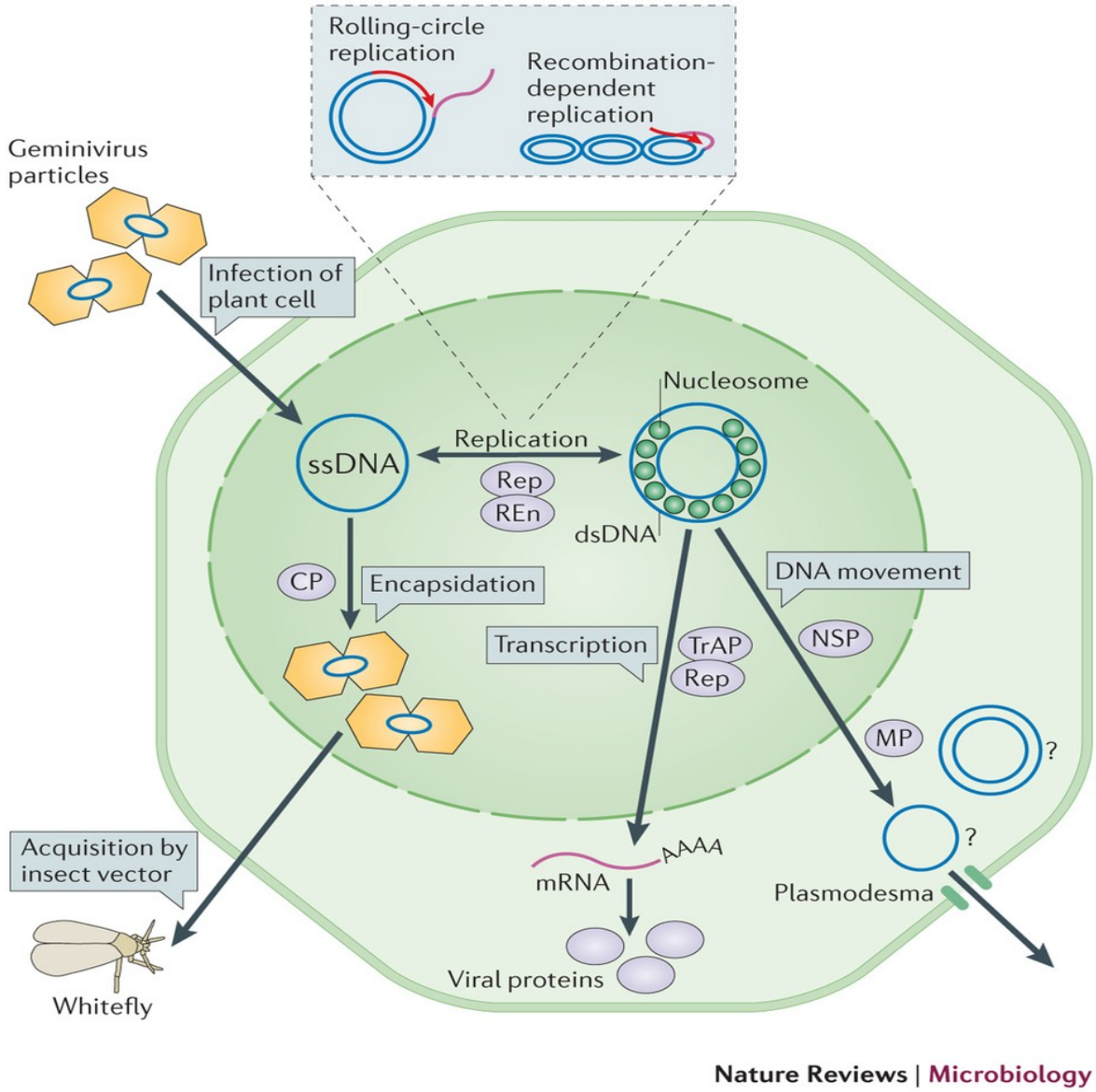


Figure 1.2.4.2: The begomovirus life cycle.

Infection begins in a plant cell when viral single-stranded DNA (ssDNA) is released from virions and copied to generate double-stranded DNA (dsDNA). The dsDNA, which assembles in nucleosomes, is transcribed by host RNA polymerase II, allowing production of replication initiator protein (Rep). Rep initiates rolling-circle replication by introducing a nick into a viral dsDNA molecule to generate a free 3'-hydroxyl end that primes ssDNA is converted to dsDNA to re-enter the replication cycle. Viral replication transitions to recombination-dependent replication, which is initiated by homologous recombination between a partially replicated ssDNA and a closed, circular dsDNA to form a looped molecule that serves as a template for both ssDNA and dsDNA synthesis (inset). Later in infection, Rep represses its own transcription, leading to activation of transcriptional activator protein (TrAP) expression, which in turn activates coat protein (CP) and nuclear shuttle protein (NSP) expression. Circular ssDNA can then be encapsidated by CP into virions, which are available for whitefly acquisition. NSP binds to viral DNA and moves it across the nuclear envelope, where movement protein (MP) traffics it across a plasmodesma. It is not known whether viral DNA moves as ssDNA versus dsDNA or as a linear versus a circular molecule. Reprinted by permission from Macmillan Publishers Ltd: Nature Review Microbiology (Hanley-Bowdoin et al., 2013), copyright (2013).

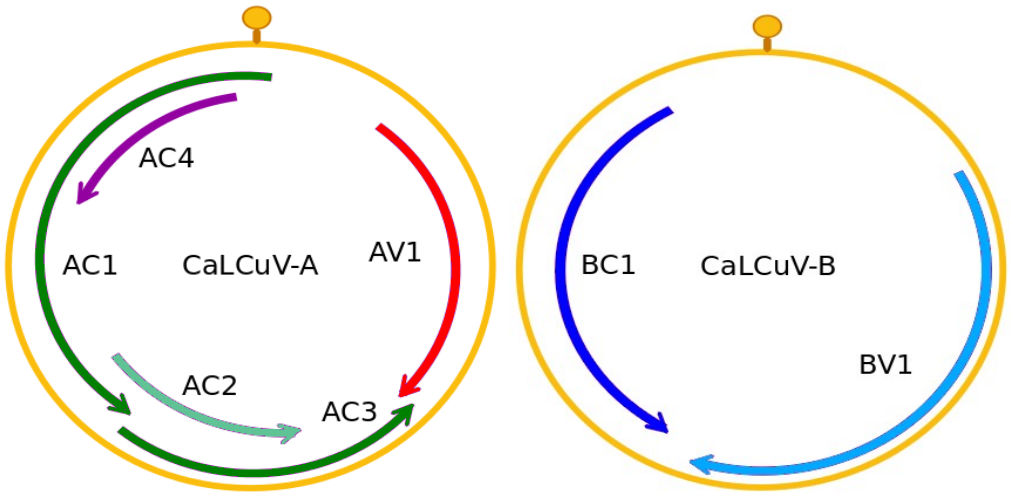


Figure 1.2.4.3: Schematic diagram of the CaLCuV bi-partite genome.

The yellow line represents the single-stranded circular DNA -A and -B. The coloured arrows represent the ORFs: the coat protein (AV1), the replicase associated proteins (AC1 and AC3), the transactivator protein (AC2), the nuclear shuttle protein (BV1) and the movement protein (BC1).

1.2.5 Sri Lankan cassava mosaic virus and Indian cassava mosaic virus

Sri Lankan cassava mosaic virus (SLCMV) and *Indian cassava mosaic virus* (ICMV) are bi-partite DNA viruses which belong to genus *Begomovirus* of the family *Geminiviridae*. SLCMV is likely derived from ICMV after recombination with a monopartite begomovirus located in Sri Lanka (Saunders et al., 2002). They are transmitted by the whiteflies *Bemisia tabaci* and provoke the cassava mosaic disease, which is one of the most important viral diseases in cassava plants (*Manihot esculenta*). Cassava is the third most important food crop in the tropics; consequently the cassava mosaic disease has a high impact during the cassava cultivation (Saunders et al., 2002).

The DNA-A genome encodes six genes (AC1-4 in the complementary sense and AV1-2 in the virion sense) involved in the replication, transcription enhancement and encapsidation functions. The DNA-B genome encodes two genes : BV1 which is required for intracellular movement and BC1 which is involved for intercellular movement (Fig 1.2.5.1). Their single-stranded genomes have a size between 2.7 to 2.8 kb (Patil and Fauquet, 2009). Their genomes are replicated by rolling-circle amplification: in first time, the single-stranded genome is used as template for the synthesis of the complementary strand (the “minus” strand) in order to create a circular double-stranded genome, and during the second step, the double-stranded genome is used as template to generate successive single-stranded DNA genomes (Hanley-Bowdoin et al., 1999).

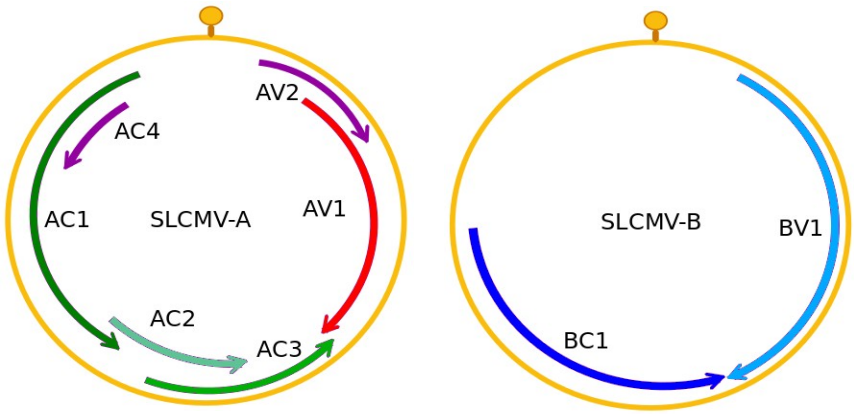


Figure 1.2.5.1: Schematic diagram of the SLCMV bi-partite genome.

The yellow line represents the single-stranded circular DNA -A and -B. The coloured arrows represent the ORFs: the coat protein (AV1), the replication associated protein (AC1), the replication enhancer protein (AC3), the transcriptional transactivator protein (AC2), the nuclear shuttle protein (BV1) and the movement protein (BC1).

1.2.6 *Grapevine red blotch-associated virus*

Grapevine red blotch-associated virus (GRBaV) belongs to the *Geminiviridae family*. It provokes red veins and red blotches symptoms on leaf in infected grapevine (*Vitis vinifera*). Its monopartite genome has a size of 3,2 kb and encodes 6 ORFs: 3 on the virion sense (V1, V2 and V3) and 3 on the complementary sense (C1, C2 and C3). V1 encodes the CP. C1 and C2 encode proteins involved in the initiation of replication (Fig 1.2.6.1) (Al Rwahnih et al., 2013).

It is transmitted by graft derived by infected grapevine, but can likely be transmitted by insect vectors because the *Grapevine redleaf-associated virus* (GRLaV) geminivirus, which provokes similar symptoms on infected grapevine, can also be transmitted by grafting and/or by the leafhopper *Erythroneura ziczac Walsh* (Poojari et al., 2013).

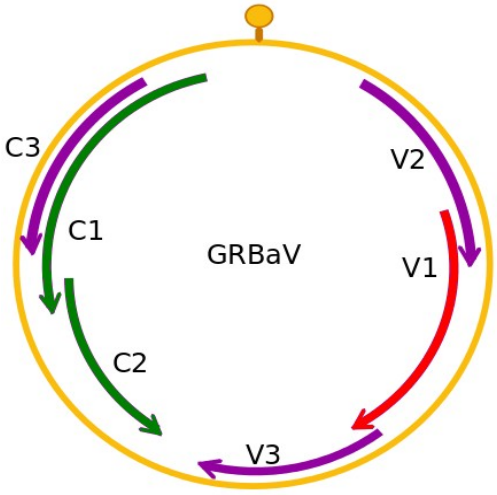


Figure 1.2.6.1: Schematic diagram of the GRBaV genome.

The yellow line represents the single-stranded circular DNA. The coloured arrows represent the ORFs: the coat protein (V1), and the proteins involved within the replication (C1 and C2).

1.2.7 Oilseed rape mosaic virus

Oilseed rape mosaic virus (ORMV) belongs to the genus *Tobamovirus* of the family *Virgaviridae*. This genus is characterized by a single-stranded RNA genome of positive polarity (encoding the proteins) and by rod-shaped particles (Fig 1.2.7.1) (Adams et al., 2009). The amino acid composition of movement protein (MP) determines phylogenetic subgroups of tobamoviruses: the first subgroup contains viruses isolated from solanaceous plants, the second group (including ORMV) isolated from cruciferous plants and the third subgroup from several other dicotyledonous plants. The common characteristic of the subgroup 2 is the overlap between MP and the coat protein (CP) ORFs (Mansilla et al., 2009). The length of ORMV genome is 6303 nucleotides. ORMV was originally isolated from an infected oilseed rape plant (*Brassica napus* L. var. *oleifera* DC). It can infect other plants belonging to the *Brassica ssp* or non-crucifer species such as tobacco (*Nicotiana tabacum*) (Aguilar et al., 1996).

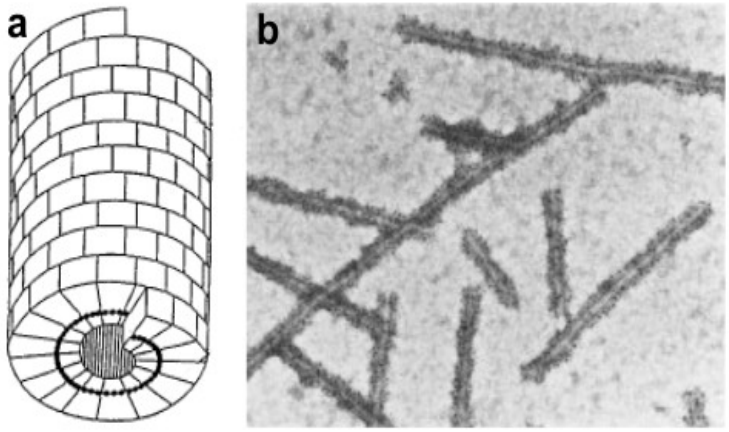


Figure 1.2.7.1: Capsid of TMV (tobamovirus).

(a) Schematic model of the capsid of TMV constituted by rod-shaped particles. (b) Electronic microscopy of TMV. Copied from (Van Regenmortel, 1999).

The genome contains three ORFs which encode four proteins : the components of RNA-dependent RNA polymerase (RDR) p126 and p180 (the latter translated from genomic RNA by readthrough of the translation stop codon of the former) and the MP and the CP translated from the two subgenomic RNAs (Fig 1.2.7.2). ORMV can infect both cruciferous and solanaceous plants (Mansilla et al., 2009).

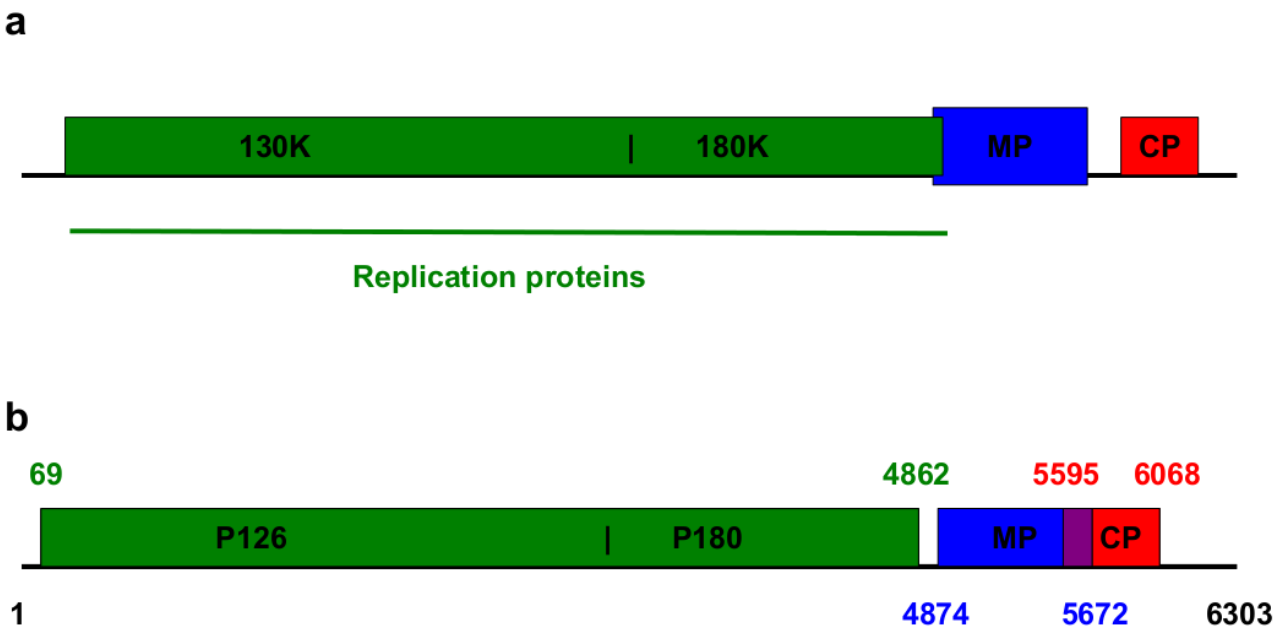


Figure 1.2.7.2: Genome organisation of tobamovirus.

(**a**) The 130K and 180K proteins are replication proteins, while MP and CP represent the movement and coat proteins, respectively. Adapted from (Ishibashi et al., 2012). (**b**) The genome organisation of ORMV: the coloured values indicate the position of the corresponding gene. The purple region corresponds to the overlapped region of MP and CP genes.

1.3 Role of small RNAs in plant antiviral defense

In 1998, experiments performed on *Caenorhabditis elegans* led to the discovery of RNA interference (RNAi), in which double-stranded RNA (dsRNA) could cause silencing of a cognate gene, Later it was shown that dsRNA is processed by Dicer or Dicer-like (DCL) enzyme into small RNAs that direct gene silencing in association with Argonaute (AGO) family proteins (Ghildiyal and Zamore, 2009). In 1999, the silencing system was also discovered in plants where it defends against invasive nucleic acids such as transposons, transgenes and viruses. Small RNAs of approximately 25-nts in length (later corrected to be 21, 22 or 24-nts) were found in virus-infected plants and in transgenic plants in which the transgene was silenced, and the term short interfering RNAs (siRNA) was coined (Hamilton and Baulcombe, 1999). The length of small RNAs involved in RNAi and silencing phenomena in various eukaryotes varies between 20 to 30-nts. Based on the mechanisms of biogenesis and function, the small RNAs are classified into siRNAs, micro RNAs (miRNAs), and Piwi-interacting RNAs (piRNAs). So far, only miRNAs and siRNAs have been discovered in plants.

1.3.1 microRNA

MicroRNAs (miRNA) are single-stranded RNA molecules which have a length of 21-22 nucleotides in plants (Allen and Howell, 2010; Bartel, 2004). They can be involved in regulation of plant development, signal transduction, protein degradation and response to environmental stress and pathogen invasion (Lu et al., 2008).

miRNAs are produced by transcription of miRNA genes within the plant genome. The RNA polymerase II transcribes miRNAs genes into primary miRNA transcripts (pri-miRNA) which form hairpin-like stem-loop secondary structures (Fig. 1.3.1.1). A ribonuclease III-like nuclease, named Dicer-like 1 (DCL1), cleaves this structure around 15 nucleotides from the base of the stem (Rogers and Chen, 2012). This process releases a precursor miRNA (pre-miRNA) in the nucleus. Then, the pre-miRNA is processed in mature miRNA:miRNA* duplex by DCL1. These miRNA:miRNA* duplexes have 2-nucleotide 3' overhangs and 19-20 complementary nucleotides (Allen and Howell, 2010). They are exported to the cytoplasm, and, through interaction with an AGO family protein, one of the duplex strands (miRNA) is incorporated into the RISC complex (Lu et al., 2008). The miRNA* strand is discarded during formation of RISC. The miRNA will be used to recognize target single-stranded mRNA by sequence complementarity. The association of the RISC with the mRNA induces the post-transcriptional gene silencing (PTGS) by cleaving the mRNA or inhibiting its translation. In *Arabidopsis thaliana,* the majority of miRNAs are bound to AGO1 due to their 5' terminal U (Allen and Howell, 2010). Some miRNAs are associated with other AGOs. For example, miR390 is recognized by AGO7 due to the presence of 5' A and some other sequence features (Montgomery et al., 2008). AGO10 is similar to AGO1. These two AGOs are involved in the temporal program of floral stem cells (Ji et al., 2011).

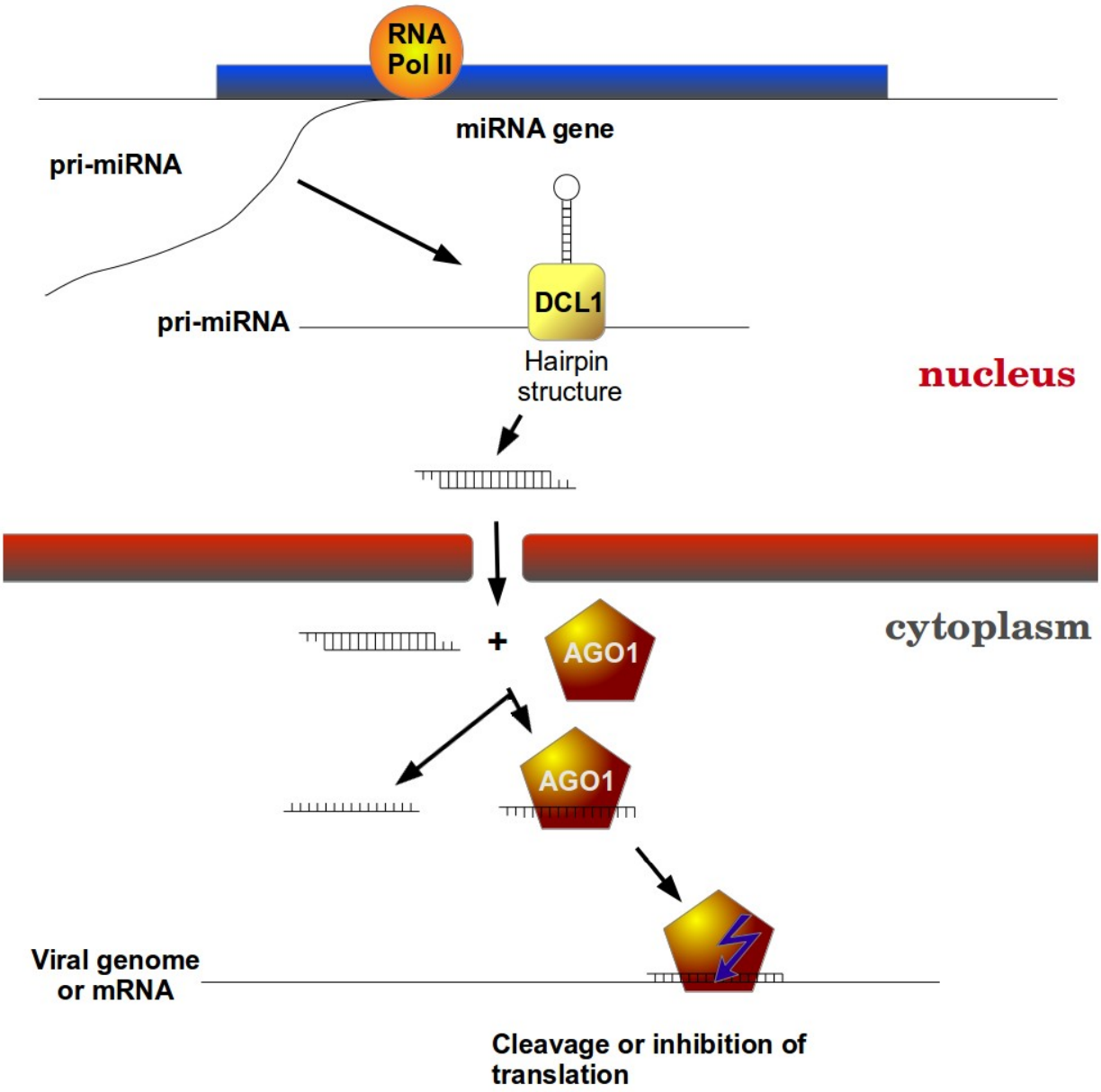


Figure 1.3.1.1: Summary of the miRNA pathway.

The miRNA gene is transcribed into primiRNA by the host RNA pol II. The primiRNA has a hairpin structure which is recognized by the DCL1 protein, is cleaved in miRNAs duplex and released in the cytoplasm. AGO1 protein recognizes the duplex, releases one miRNA and keeps the other to target viral genome or mRNA by nucleotidic complementarity. This association induces the cleavage or inhibits the translation of viral genome or mRNA.

1.3.2 short interfering RNA

Short interfering RNAs (siRNA) are singled-stranded RNA molecules which have a length from 21 to 24 nucleotides. These RNAs are produced from double-stranded RNA (dsRNA) precursors by DCL proteins. The dsRNA precursors are produced by DNA polymerase-mediated senses and antisense transcription or by RNA-dependent RNA-polymerase (RDR) from a single-stranded RNA (ssRNA) as template. The ssRNA template can be transcribed from the plant or viral genome. In RNA virus infections, viral siRNAs (vsiRNA) are processed from the dsRNA produced by plant RDR or viral RDR. The *Arabidopsis thaliana* genome encodes 4 DCL and 6 RDR proteins. Only three RDRs (RDR 1, RDR2 and RDR 6) are known to be functional in siRNA biogenesis (Voinnet, 2008; Pooggin, 2013). The length of siRNA depends on the DCL proteins. For example, in *Arabidopsis thaliana*, DCL2 produces 22-nt siRNAs, DCL3 - 24-nt siRNA and DCL4 21-nt siRNAs.

DCL4 and DCL2 generate 21-nt and 22-nt vsiRNAs involved in the antiviral immunity in RNA virus-infected plants (Deleris et al., 2006). In addition, DCL3 makes 24-nt vsiRNAs in DNA virus-infected plants as was discovered by Pooggin's team (Akbergenov et al., 2006; Blevins et al., 2006; Blevins et al., 2011; Aregger et al., 2012). Like miRNAs, 21 and 22-nt siRNAs are associated with AGO proteins and can direct cleavage of complementary target RNA (Fig 1.3.2.1).

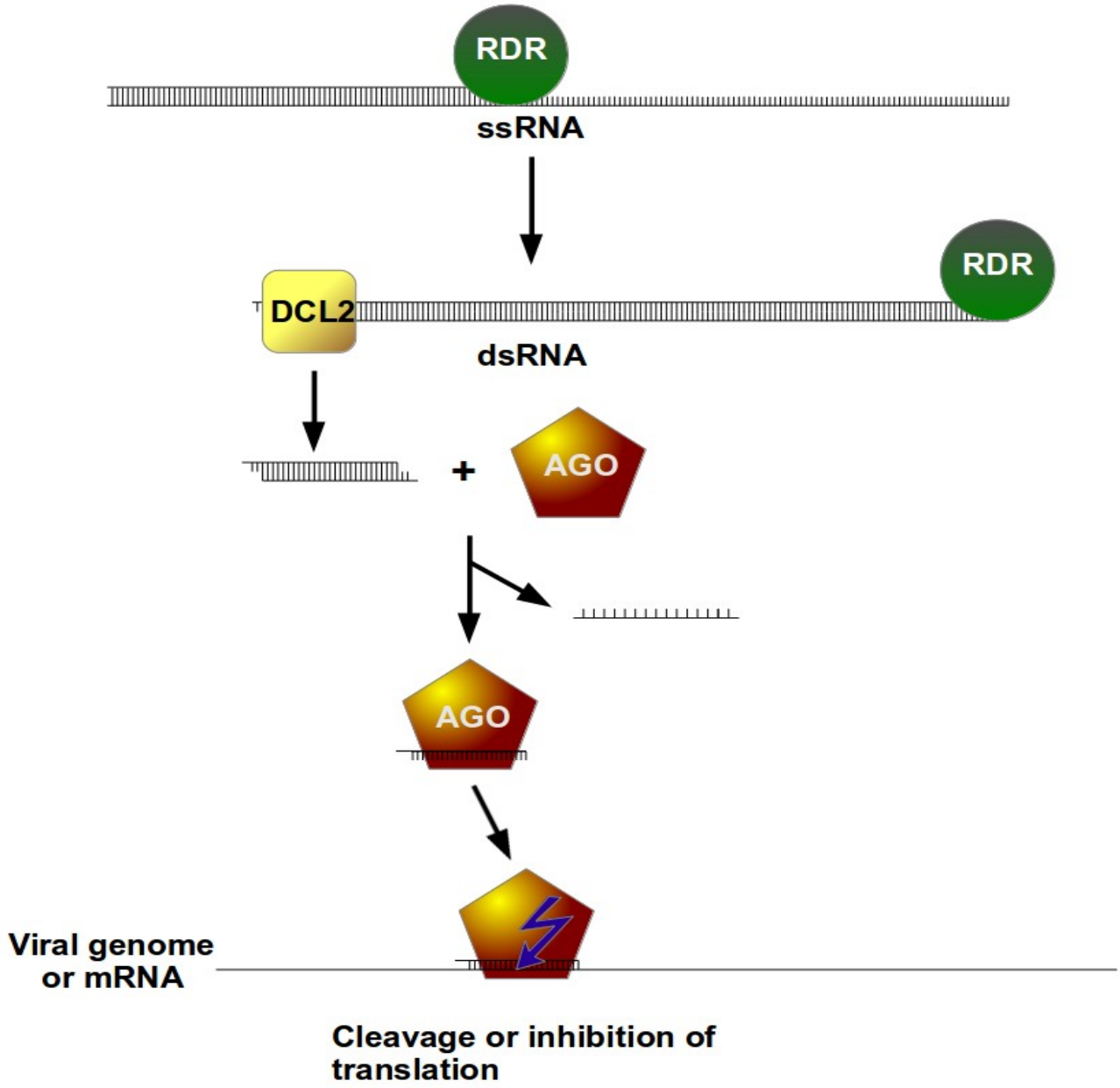


Figure 1.3.2.1: the siRNA pathway involved in the RNA regulation.

The complementary strand of a ssRNA is synthesized by a RDR protein. The dsRNA is cleaved into sRNA duplexes by DCL4 or DCL2. Then, AGO proteins recognize the duplex, release one sRNA and keep the other to target viral genome or mRNA by nucleotidic complementarity. This association induces the cleavage or inhibits the translation of viral genome or mRNA.

24-nt siRNA can target complementary DNA for cytosine methylation and thereby cause transcriptional gene silencing (TGS) (Fig 1.3.2.2). The siRNA-directed DNA methylation (RdDM) pathway involves DCL3, RDR2, AGO4, AGO6 and AGO9. RdDM regulates gene expression

through heterochromatization and defends against invasive nucleic acids such as transposons, transgenes and DNA viruses. Genomic sites for *de novo* DNA methylation are targeted by the 24-nt siRNAs produced by DCL3. AGO4 associated with 24-nt siRNAs target the RNA transcribed by the host RNA polymerase Pol V. The complementary interaction between the siRNA and the transcribed RNA allows the association of AGO4 with Pol V to recruit the de novo methyltransferase DRM2 and other RdDM factors. RDR2 produces the double-stranded RNA precursors of 24-nt siRNAs by using as template the RNA transcribed by the RNA polymerase Pol IV. The activity of RDR2 depends to its association with Pol IV. AGO4 is a main protein mediating RdDM in most tissues, while AGO6 and AGO9 have tissue specific expression. The DNA viruses have evolved various mechanisms to evade the silencing mechanism based on RdDM (Pooggin, 2013).

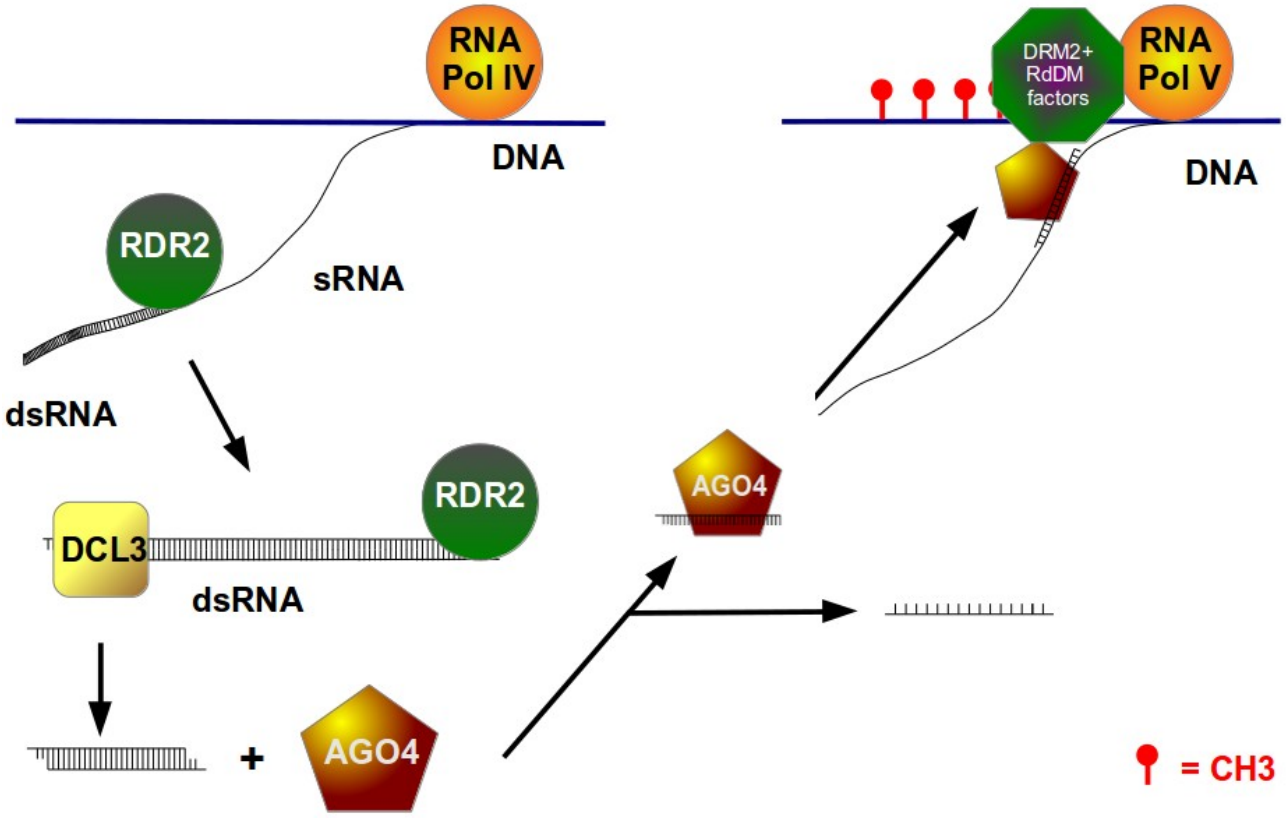


Figure 1.3.2.2: the siRNA pathway involved in the DNA methylation

DNA genome is transcribed into ssRNA by the host RNA pol IV. The complementary strand of this ssRNA is synthesized by the RDR2 protein. The dsRNA is cleaved into sRNA duplexes by DCL3. Then, AGO4 proteins recognizes the duplex, release one sRNA and keep the other to target ssRNA synthesized by the RNA pol V by nucleotidic complementarity. This association induces the recruiting of proteins involved in the DNA methylation, and methylated the DNA transcribed by the host RNA pol V.

1.4 Methods of viral diagnostics

The first method used for the viral diagnostics was an enzyme-linked immunosorbent assay (ELISA) (Clark and Adams, 1977). ELISA method simplified detection and reduced the time to obtain conclusive results. Moreover, before ELISA, only specialists with years of experience were able to recognize viruses according to the description of virus symptoms on hosts by using complex and cumbersome methods, and elaborated techniques like transmission electron microscopy. Nevertheless, ELISA was not adapted to correctly identify virus strains because viral coat proteins are often conserved in particular genus that antibodies cannot be used to discriminate. Also, the production of high-quality antisera required a lengthy and costly process.

To improve the virus diagnostics, new methods are developed based on the detection of viral RNA and DNA. In 1990, a virus was detected with a Polymerase Chain Reaction (PCR) method (Vunsh et al., 1990). This method has been improved through new diagnostics methods based on Real-Time Polymerase Chain Reaction (RT-PCR). These methods were improved to be used directly within field by using an isothermal amplification step during PCR, or to detect many viruses with a single test by multiplex methods. Nevertheless, ELISA and RT-PCR methods can detect only known viruses which have antibodies or primers already designed. To detect new viruses, the next-generation sequencing (NGS) technologies propose different interesting methods which must be completed by bioinformatics analysis (Boonham et al., 2014).

1.5 Next generation sequencing technologies for deep sequencing of viral siRNA populations

During the last couple of years the research efforts of many groups including the Pooggin group (this thesis work) have revealed that it is possible to identify a virus involved in a plant disease by sequencing the total sRNA population from the infected plant. With the development of next generation deep-sequencing technologies, large quantities of sRNAs can be sequenced per sample. The development of bioinformatics tools allows for a *de novo* assembly of viral genome using short sequencing reads. In the pioneering work of Kreuze et al. (2009) an RNA virus genome was assembled completely from vsiRNAs. The identification of viruses with the next generation deep-sequencing technologies is also possible for viruses with unstable particles or non-encapsidated agents such as viroids or certain virus strains, which are difficult to isolate (Kreuze et al., 2009).

The next generation sequencing (NGS) technologies are the second generation of sequencing technologies. The first generation is based on Sanger sequencing. In 2001, the

completion of human genome project stimulated the development of new sequencing technologies in order to reduce the cost and, increase the speed of sequencing and the number of sequenced data (Liu et al., 2012).

Since 2005, the three main commercial platforms are the Roche 454 Genome Sequencer, the Illumina Genome Analyzer, and the Life Technologies SOLiD System. All these platforms use two steps for the sequencing: the first is preparation and amplification of DNA, and the second is the sequencing step. The Roche 454 Genome Sequencer is based on sequencing-by-synthesis with pyrophosphate chemistry. The Illumina Genome Analyzer is based on sequencing-by-synthesis with Sanger chemistry. The Life Technologies SOLiD system is based on a sequencing-by-ligation technology (Zhou et al., 2010).

The Illumina technology is the best next-generation deep-sequencing technology because it can sequence up to $85x10^9$ reads per run where each sequenced-read has a length of 50 to 100-nts. This technology is actually the best deep-sequencing platform for deep sequencing of sRNAs because it has the biggest output (Liu et al., 2012).

2. Material and Methods

2.1 Biological materials

All the plant samples analyzed in this study are summarized in Table 2.1.1. The model plants were *Arabidopsis thaliana* and *Nicotiana benthamiana*. To study ORMV, CaMV and CaLCuV, wild-type and mutant *Arabidopsis thaliana* plants were used. The mutant plants carried mutations that inactivate Argonaute, DCL or RDR proteins. *A thaliana* seedlings were infected with DNA clones of CaMV and CaLCuV by biolistic inoculation. *N benthamiana* and *A thaliana* were infected with ORMV by mechanical inoculation with sap from ORMV infected plants. The wild-type ORMV infection originates from ORMV virions taken from a naturally infected plant, because the available ORMV clone was not infectious (for more details, see Seguin et al., 2014a in the Annex).

For grapevine and cassava viruses, samples came from leaves of naturally-infected plants displaying disease symptoms. The grapevine samples were taken from infected grapevine leaves in Oregon State (USA), and were provided by the team of Prof Valerian Dolja. The cassava samples were collected in South India and provided by Prof. Veluthambi of Madurai Kamaraj University, Madurai, India.

Banana plants *Musa acuminata* infected with different BSV species in 2000 were maintained at CIRAD (France) in a tropical greenhouse by vegetative propagation. *Musa*

balbisiana and non-infected *Musa acuminata* were used as control. Banana leaves were collected by the team of Dr Marie-Line Iskra-Caruana (for more details, see Rajeswaran et al., 2014a in the Annex). The BSV species studied include *Banana streak obino l'Ewai virus* (BSOLV) (Harper and Hull, 1998), *Banana streak goldfinger virus* (BSGFV), *Banana streak mysore virus* (BSMYV) (Geering et al., 2005), *Banana streak vietnam virus* (BSVNV) (Lheureux et al., 2007), *Banana streak cavendish virus* (BSCAV) and *Banana streak imove virus* (BSIMV) (James et al., 2011).

Seedlings of rice (Oryza sativa japonica) were grown in phytochambers and three weeks after germination, inoculated with the infectious clone of RTBV by agroinfiltration. Then, leaves displaying symptoms were collected (for more details, see Rajeswaran et al., 2014b in the Annex).

Total RNA from all the plant samples from infected and control tissue plants, was extracted using the Trizol method established in the lab (Aregger et al. 2006; Blevins et al., 2006) with certain modifications (for details, see Rajeswaran et al., 2014a, 2014b in the Annex) and then, sent to Fasteris (Fasteris SA, Chemin du Pont-du-Centenaire 109, 1228 Plan-Les-Ouates, https://www.fasteris.com) for Illumina deep sequencing. The small RNA molecules were separated according to their sizes on polyacrylamide gel. Then, Fasteris and Illumina protocols were used to reverse-transcribe the RNA, and the resulting cDNA library was sequenced on Illumina Genome Analyzer HiSeq 2000. Otherwise, for DNA circular viruses (e.g. *Geminiviridae* and *Caulimoviridae*), rolling circle amplification (RCA) was used to amplify circular viral genomic DNA (Hadfield et al., 2011; Haible et al., 2006). Amplification by RCA was used for BSV samples according to the protocol provided by the “illustra TempliPhi Amplification Kit” (GE Healthcare Life Sciences) (for more details, see Rajeswaran et al., 2014a in the Annex). The sequencing of these amplified samples was also performed by Fasteris.

Table 2.1.1 : small RNA datasets used for bioinformatics analysis

The first column indicates if the sample is infected or not. The second column gives the sample description. The third column contains the dataset number. The fourth column indicates the infected plant species. The last column contains the reference of paper which contains analysis of the sample.

Virus name	sample	Dataset number	Plant species	Reference
none	Col-Mock	BPO-13	*Arabidopsis thaliana*	
Cb_CaMVL	Col-Cb_CaMVL	BPO-14		
none	ago2-mock	BPO-16		
CaMV	Col-CaMV	BPO-20		(Seguin et al., 2014a)
	ago2-CaMV	BPO-21		

CaMV	ago3-CaMV	BPO-22	*Arabidopsis thaliana*	
ICMV	H226-1	BPO-31	*Manihot esculenta*	
	M4-3	BPO-32		
	SMAL-5	BPO-33		
	STVM-7	BPO-34		
none	Col-0 mock 14d	BPO-35	*Arabidopsis thaliana*	
	r126 mock 14d	BPO-36		
	d234 mock 14d	BPO-37		
ORMV	Col-0 14d	BPO-38		(Seguin et al., 2014b)
	r126 14d	BPO-39		
	d234 14d	BPO-40		
none	Col-0 mock 22d	BPO-41		
	r126 mock 22d	BPO-42		
	d234 mock 22d	BPO-43		
ORMV	Col-0 22d	BPO-44		(Seguin et al., 2014b)
	r126 22d	BPO-45		
	d234 22d	BPO-46		
ICMV	H226_s	BPO-47	*Manihot esculenta*	
	M4_s	BPO-48		
	STVM_s	BPO-49		
	H226_iv	BPO-51		
	M4_3b	BPO-52		
	M4_4a1	BPO-53		
	SMAL_6a	BPO-54		
none	Col-Mock	BPO-56	*Arabidopsis thaliana*	(Arreger et al. 2012)
CaLCuV	Col	BPO-57		(Arreger et al. 2012), (Seguin et al., 2014b)
	r126	BPO-58		(Arreger et al. 2012)
none	musa acuminata	BPO-59	*Musa acuminata*	(Rajeswaran et al., 2014a)
PKW	musa balbisiana	BPO-60	*Musa balbisiana*	(Rajeswaran et al., 2014a)
BSV	BSGFV	BPO-61	*Musa acuminata*	(Rajeswaran et al., 2014a)
	BSVNV	BPO-62		(Rajeswaran et al., 2014a)

BSV	BSCAV	BPO-63	*musa acuminata*	(Rajeswaran et al., 2014a)
	BSOLV	BPO-64		(Rajeswaran et al., 2014a)
	BSMYV	BPO-65		(Rajeswaran et al., 2014a)
	BSIMV	BPO-66		(Rajeswaran et al., 2014a)
RTBV	RTBV	BPO-67	*Oriza sativa*	(Rajeswaran et al., 2014b)
RTBV	RTBV	BPO-68		(Rajeswaran et al., 2014b)
control	control	BPO-69		(Rajeswaran et al., 2014b)
none	H226-mock-y	BPO-70	*Manihot esculenta*	
ICMV	H226-virus-y	BPO-71		
	VTP-virus-y	BPO-72		
	VTP-virus-o	BPO-73		
	S857-virus-y	BPO-74		
	S857-virus-o	BPO-75		
	MVD-virus-y	BPO-76		
	MVD-virus-o	BPO-77		
	STVM-virus-o	BPO-78		
unknown viruses	PN_green_N8	BPO-104	*Vitis vinifera*	(Seguin et al., 2014b)
unknown viruses	PN_red_N11	BPO-105		(Seguin et al., 2014b)
unknown viruses	PN_red_N12	BPO-106		(Seguin et al., 2014b)

2.2 Illumina-Solexa sequencing technology

Before purchased by Illumina in 2007 (Liu et al., 2012), Solexa has developed a sequencing platform based on sequencing-by-synthesis chemistry. Fragmented DNA (<800 bp) are prepared for the amplification step: adaptors are added to the 5' and 3' termini of each DNA fragment. After denaturation, they are attached on a flow cell which contains up to eight different channels where are run simultaneously different samples (Metzker, 2010). Included within the adaptors, a nucleotidic index allows to discriminate the different samples present in the same channel. The flow cell contains the complementary sequences of the adaptors. These complementary sequence are primers for the following amplification step. A “bridge PCR” amplifies

the DNA fragments (Mayer et al., 2013). Many PCR cycles are performed in order to form “DNA colonies” where each colony contains one original DNA fragment. Once “DNA colonies” are formed, the sequencing step starts, using a reaction mixture with primers and nucleotides labelled by a specific fluorescent dye. Each nucleotide has a specific fluorescent colour and a reversible chemical terminator. After the link of the primers on the 3' adapter of amplified DNA fragments, the sequencing is performed by the following repeated steps: incorporation of labelled nucleotide by complementarity, reading of the fluorescent dye by a CDD camera and, removal of the dye and the terminator. These steps are repeated for an user-defined number of cycles (Zhou et al., 2010) (Fig. 2.2.1).

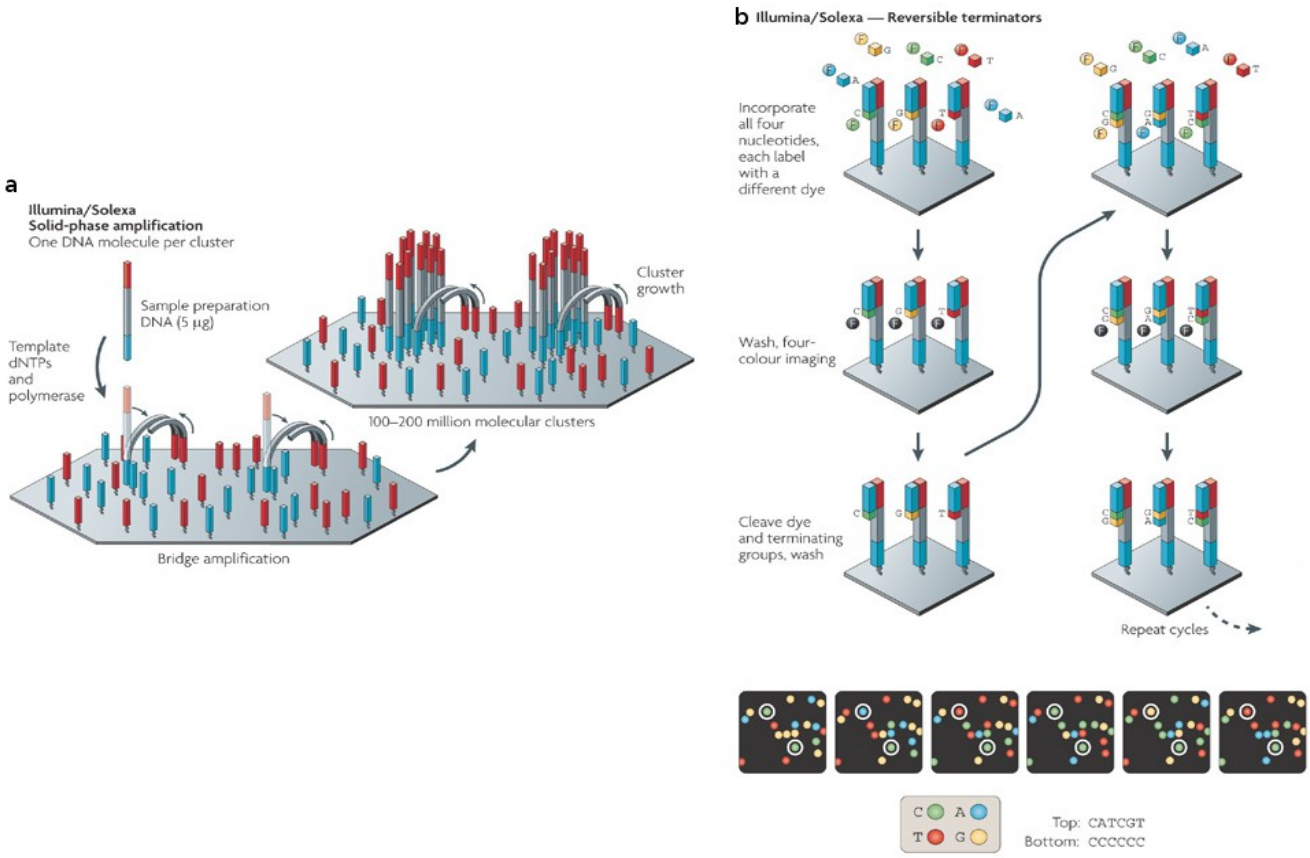


Figure 2.2.1: Summary of Illumina-Solexa deep-sequencing technology

(**a**) Illumina-Solexa amplification is composed of two basic steps: initial priming and extending of the single-stranded, single-molecule template, and bridge amplification of the immobilized template with immediately adjacent primers to form clusters. (**b**) The four-colour cyclic reversible termination (CRT) method uses Illumina/Solexa's 3'-O-azidomethyl reversible terminator chemistry. Following imaging, a cleavage step removes the fluorescent dyes and regenerates the 3'-OH group using the reducing agent tris(2-carboxyethyl)phosphine (TCEP). The four-colour images highlight the sequencing data from two clonally amplified templates. Reprinted by permission from Macmillan Publishers Ltd: Nature Review Genetics (Metzker, 2010), copyright (2010).

The standard output file provided by NGS technology is a fastq file. FASTQ stands for FASTA and Quality. In this file, each read is described by four informative lines. The first line contains the identifier of read starting with '@' symbol. The second line contains the sequence of read. This sequence is similar to FASTA format: it is restricted to IUPAC single letter codes for (ambiguous) DNA or RNA and the upper-case is conventional. The third line indicates the end of the sequence of read and begins with '+' symbol. The last line indicates the phred score encoded by an ASCII code. Each ASCII symbol corresponds to the nucleotide at the same position along the sequence. The Phred score represents the quality of sequencing for each nucleotide of read. This quality is an estimated probability to have a sequencing error at the corresponding nucleotide. To determine the quality, for each ASCII symbol corresponds one numeric value (Q) according to the kind of ASCII code (Table 2.2.2). The probability value is obtained according to the formula $P = 10^{(-Q/10)}$ (Cock et al., 2010).

Table 2.2.2: Summary of fastq file. Adapted from (Cock et al., 2010).

Line position	Description	example
1	@title and optional description	@HWIEAS210R_0001:1:1:1106:16572#NACTAT/1
2	sequence line(s)	AAGAGTGCTTGAAATTGTCG
3	+optional repeat of title line	+HWIEAS210R_0001:1:1:1106:16572#NACTAT/1
4	Quality line(s)	BCCAACCCCCCCCBCCC@CC

2.3 Bioinformatics analysis

2.3.1 Mapping

A mapping is a multiple alignment of sequencing reads along a reference sequence. In our case, it allows to localize precisely each sequenced short read along the viral genome or the host plant genome.

2.3.1.1 Mapping software : Burrows-Wheeler Alignment (BWA)

BWA is a tool which maps short reads produced by high-throughput sequencing like Illumina/Solexa (Li and Durbin, 2009, 2010). It uses the Burrows-Wheeler Transform (BWT) (Burrows and Wheeler, 1994) to improve the alignment of reads along the various reference sequences such as genes, transcripts, genomes (Li and Durbin, 2009). It is one of the main BWT programs including SOAPv2 (Li et al., 2009b) and Bowtie (Langmead et al., 2009). It takes a small memory footprint of computer, and can count the number of exact hits of a string (for example, a sequence of genome) of length *m* in O(*m*) time, independently of the genome size. First, BWA creates a prefix tree (also named prefix trie) with the sequences of genome. In this tree, each edge is labelled by one nucleotide, and the concatenation of nucleotides for each subsequence from a leaf until to the root gives one unique subsequence (Fig. 2.3.1.1.2). Then, values are given to each node according to the BWT provided by BWA. A “$” symbol is added to the end of the genome sequence. A suffix array is created by permutation of the first letter of the sequence to its end until the “$” is the first symbol of the sequence. Then the suffixes are sorted by alphabetical order during the BWT. The indexes, which indicate the position of suffixes within the BWT, are the values provided within each nodes (Fig. 2.3.1.1.1).

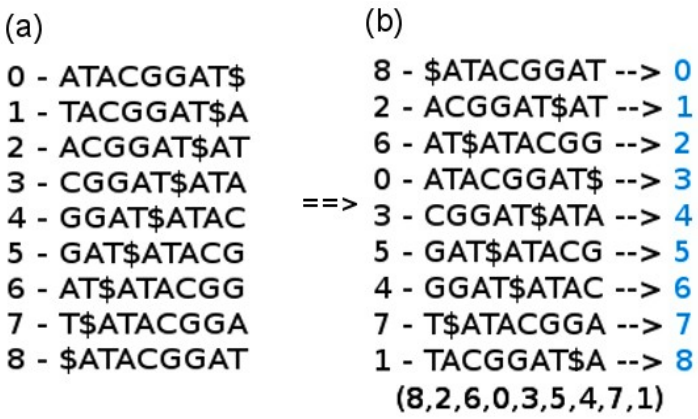


Figure 2.3.1.1.1: Creation of suffix array BWA

The sequence ATACGGAT are used to create the suffix array. (a) after addition of $ to the end, each letter of the sequence is permuted to the end. (b) After sorting, the obtained suffix array is (8, 2, 6, 0, 3, 5, 4, 7, 1). The index of suffix array is indicated in blue.

Each node from the prefix tree has two index values. They delimit an interval within the suffix array which contains the position of the corresponding subsequence (Fig. 2.3.1.1.2). These indexing and tree are saved within files, and consequently are used for different mappings against the corresponding genomic sequence.

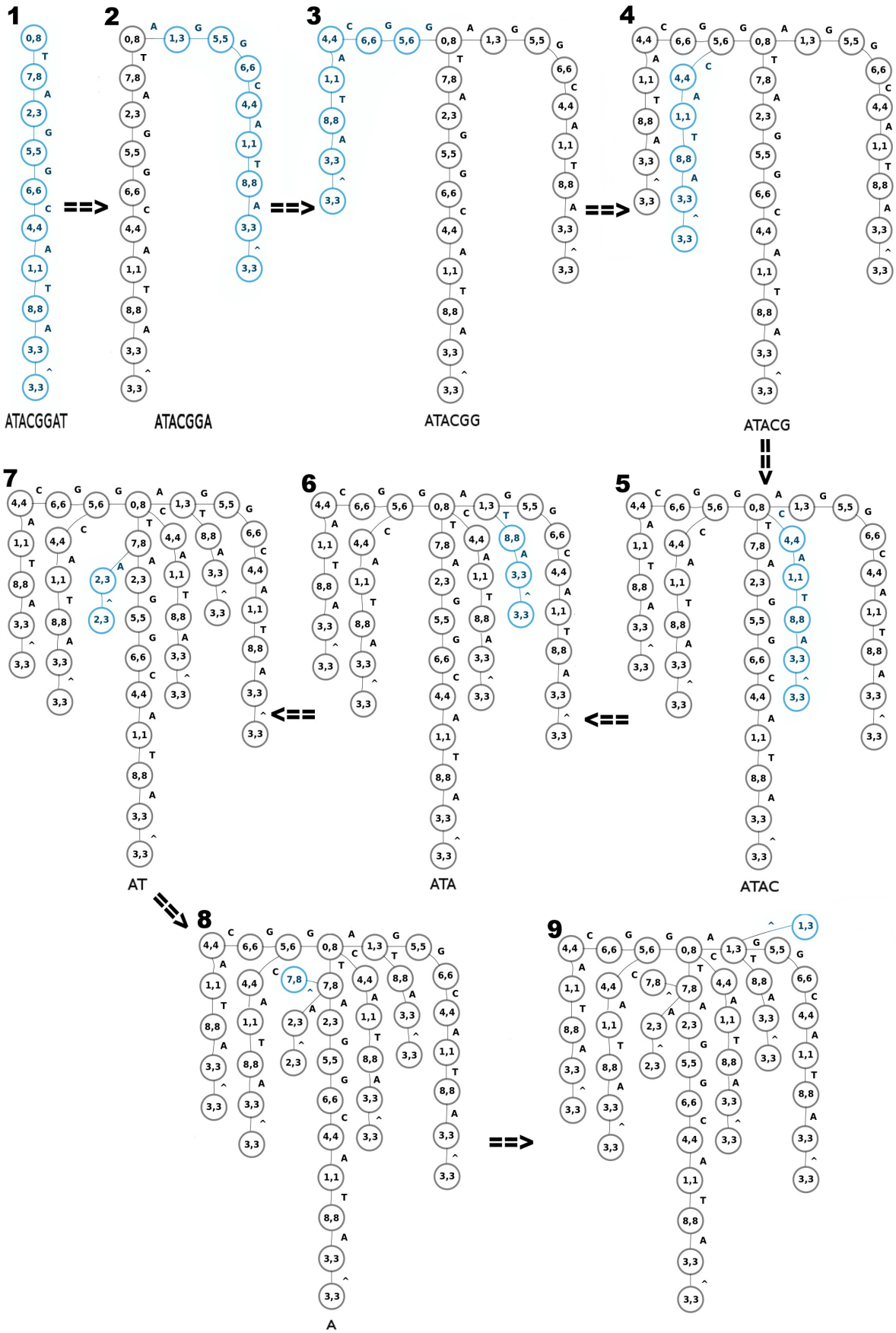


Figure 2.3.1.1.2: Creation of prefix tree

Each step allows to add one different subsequence to the tree.

The combination of the prefix tree and the suffix array allow to find quickly the exact positions of each subsequence along the genomic sequence. For each read, BWA searches for occurrences of the read sequence within the prefix tree by starting to the root which corresponds to the end of read sequence. The last node, corresponding to the first nucleotide of read sequence, contains the indexes of suffix array corresponding to the position of the read along the genomic sequence (Fig. 2.3.1.1.3).

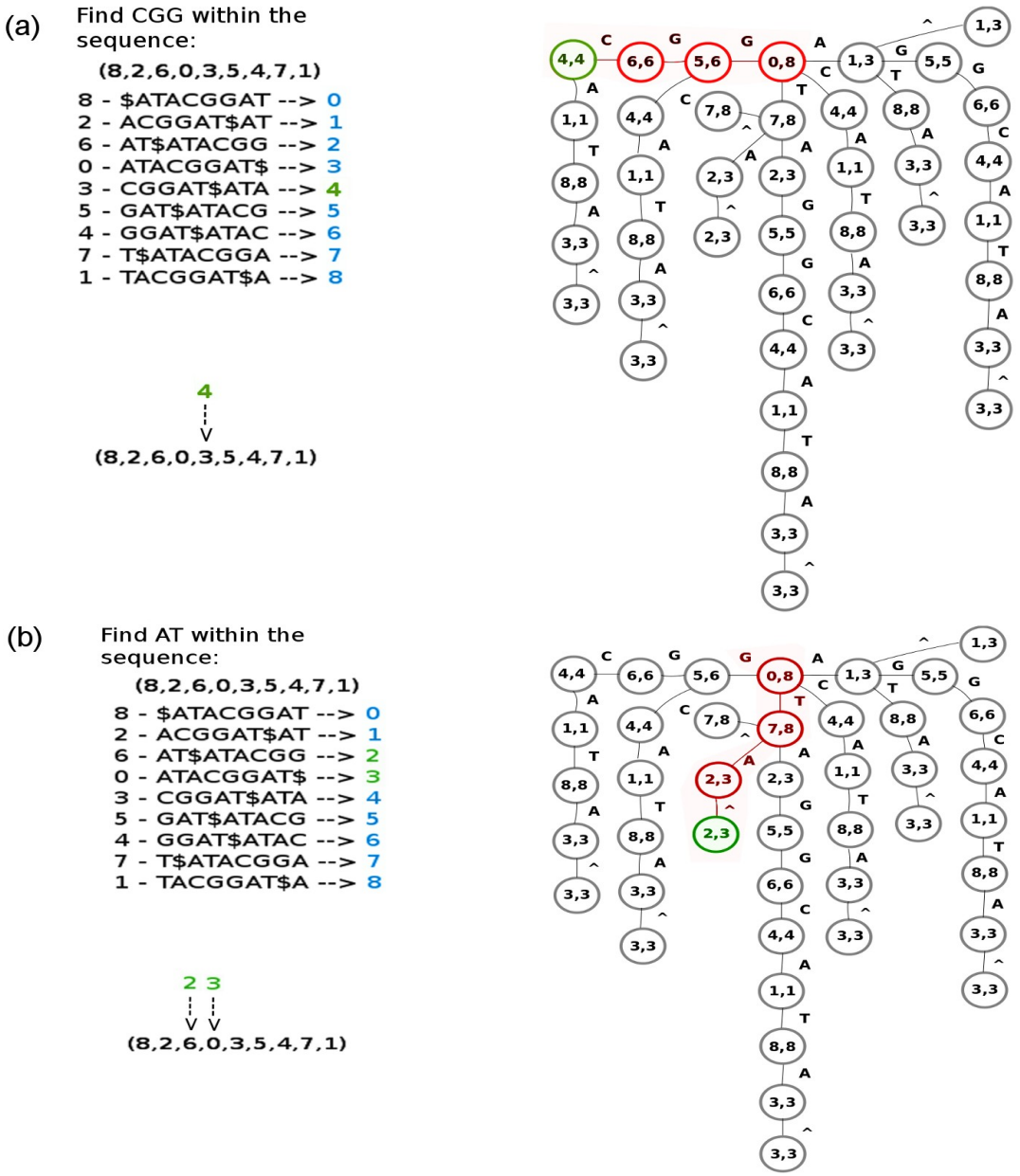


Figure 2.3.1.1.3: Research of positions for subsequences CGG and AT along the sequence ATACGGAT

Red nodes indicate the used path to find the subsequence along the prefix tree. Green nodes indicate the last node of path used to find the indexes which indicates the positions of the subsequences along the original sequence according to the suffix array. (**a**) The subsequence CGG is found at the position 3 along the sequence. (**b**) The subsequence AT is found at the position 6 and 0 along the sequence.

BWA takes as input a fasta file which contains the reference sequence and another file which contains all sequences to be aligned. This second input file can be a fasta file or a fastq files. As output, BWA generates a SAM/BAM file according to the user's parameters (Li and Durbin, 2009, 2010). For a circular reference genome (e.g. *Caulimoviridae, Geminiviridae* and *Viroids*), the first 50 nucleotides of reference sequence are added to the end in order to map reads overlapping the genome end in a linear form. When BWA finds different equally best positions for the same read along the genome, it selects randomly one position, saves it in the corresponding column within the SAM/BAM files, and saves the other positions within a tag column (see below for more information about SAM/BAM files). BWA can allow different number of mismatches during the mapping according to the length of reads. For reads between 20 to 25 nucleotides, BWA can include up to 2 mismatches during the alignment.

2.3.1.2 BAM/SAM files format

The Sequence alignment/Map (SAM) format is a generic format defined in 2009 for the output file provided by mapping analysis. The main goal was to have only one type of output files for different alignment tools. Each SAM file contains a header section and an alignment section. The header contains the name of reference sequence, found in the header of the fasta file (preceded by “>” symbol), and its length. The alignment section has 11 mandatory fields which can be followed by optional fields; all are TAB delimited. Each line corresponds to the mapping result for each read found within the fastq files (Li et al., 2009a).

The first column indicates the read identifier found within the fastq file. The second column uses a value, named flag, which provides information about the strand where the read is aligned. With short reads, this column can have 3 different values: 4 if the read is unmapped, 0 or 16 if the read is mapped on the forward or reverse strand of the reference sequence respectively. The third column indicates the name of the reference sequence. This column is important if the fasta file contains different sequences. The position where the read starts to map is indicated by the fourth column. If the read is mapped on the reverse strand, this position corresponds to the 3' mapped nucleotide of the read. The fifth column indicates the mapping quality by a Phred-scaled score. The sixth column is a CIGAR (a special letter/number code) which indicates the presence of insertion/deletion in the read aligned to the reference sequence. The tenth column contains the read sequence. The sequence always corresponds to the reference sequence found in the fasta file even if the read is mapped on the reverse strand (in this case, the tenth column contains the reverse-complemented sequence of the read). The last mandatory column contains the quality of sequenced read found in fastq file. This quality sequence is reversed if the read is mapped on the

reverse strand (Li et al., 2009a) (table 2.3.1.2.1).

The optional columns contain tags that bring complementary information according to the user's parameters. In our study, we used only the tag named 'MD:Z'. This tag allowed us to know the nucleotide which differs between the read and the reference sequence in order to find Single Nucleotide Polymorphisms (SNP). It uses a specific letter/number code similar to the CIGAR (Li et al., 2009a).

Binary Alignment/Map (BAM) file is the compressed SAM file. BAM file contains binary code understandable only by computer. This compression allows to reduce the size and the memory used to analyze the BAM file in comparison to the SAM file. A special software package named SAMtools allows to manipulate the SAM/BAM files. The conversion BAM to SAM and SAM to BAM can be performed by SAMtools (Li et al., 2009a).

Moreover SAMtools can display the mapping, sort the alignment and merge different BAM/SAM files (Li et al., 2009a). The statistical software named R (Ihaka and Gentleman, 1996) (http://www.r-project.org/) proposes a specific library named Rsamtools which includes all the functions to parse and analyze BAM/SAM files (http://www.bioconductor.org/).

Table 2.3.1.2.1: Mandatory fields in the SAM format.

Copied from (Li et al., 2009a)

No.	Name	Description
1	QNAME	Query NAME of the read or the read pair
2	FLAG	Bitwise FLAG (pairing, strand, mate strand, etc.)
3	RNAME	Reference sequence NAME
4	POS	1-Based leftmost POSition of clipped alignment
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIGAR	Extended CIGAR string (operations: MIDNSHP)
7	MRNM	Mate Reference NaMe ('=' if same as RNAME)
8	MPOS	1-Based leftmost Mate POSition
9	ISIZE	Inferred Insert SIZE
10	SEQ	Query SEQuence on the same strand as the reference
11	QUAL	Query QUALity (ASCII-33=Phred base quality)

2.3.1.3 Visualization software: IGV

Integrative Genomics Viewer (IGV) is a desktop application available for visualization of genomic data (http://www.broadinstitute.org/igv/) (Thorvaldsdóttir et al., 2012, 2013). It is written in the Java programming language, and consequently can be run on Windows, Mac and Linux computer environments. IGV can read the data provided by NGS technologies. It can load BAM/SAM files to analyze the mapping results.

IGV represents the mapping result by a histogram and a succession of arrow figures. The histogram represents the number of reads which cover each nucleotide of the reference genome. A colour code is used to indicate the presence of SNP or nucleotides which differ from the reference sequence. Each arrow represents one mapped read. The orientation of the arrow indicates the sense where the read maps. Each nucleotide differing from the reference is indicated within the arrow. Moreover, deletion and insertion of nucleotides are indicated by horizontal or vertical lines respectively, within the arrow. This kind of representation allows to analyze the observed mismatches during the mapping (Thorvaldsdóttir et al., 2012).

2.3.1.4 Visualization software: MISIS

MISIS is a new visualization software for the mapping results, which was developed as part of this PhD thesis. It displays a histogram of the coverage of mapped reads allowing to observe the hot and cold-spots of vsiRNA productions along a viral genome sequence (see section 3.1 of Results and the paper Seguin et al., 2014b in the Annexes).

2.3.1.5 Correction of the viral genome sequence

A viral reference sequence used for mapping may not correspond to the actual sequence of the virus in an infected sample. To identify the correct sequence of the virus, the analysis of mapping result is necessary. MISIS can show mismatches within the reference sequence by the presence of gaps in the coverage profile with the perfect matched reads. These gaps must have a length equal to the length of mapped reads. If this is the case, IGV is used to be sure that the gap is not due to an amplification error which occurred during the PCR amplification step, and to identify the nucleotide which differs from the reference. Moreover, IGV allows to identify single nucleotide polymorphisms (SNP) and consequently, deduce a consensus master sequence of the viral genome (Fig 2.3.1.5.1). This first step allows to correct the reference sequence and to redo the mapping with this new sequence before to starting the statistical analysis.

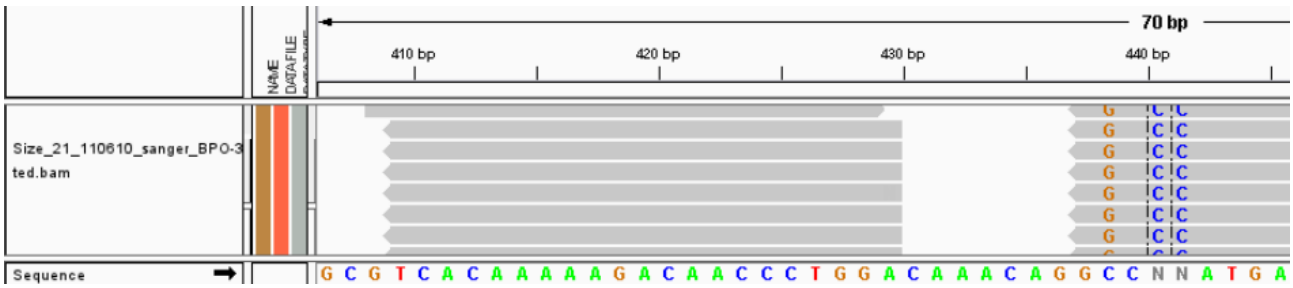


Figure 2.3.1.5.1: Example of mapping visualization with IGV.

Grey arrows represent the mapped reads. The reference sequence is indicated on the bottom with a colour code for each nucleotide and on the top with the positions of nucleotides. Two nucleotides (positions 440 and 441) are unknown within the reference sequence. This unknown region is covered by multiple reads which have CC nucleotides to these corresponding positions. Moreover, one SNP is visible at the position 438 bp.

2.3.1.6 Statistical analysis of mapping results

The count tables generated by MISIS can be used for specific statistical representation of the mapping result. R software (Ihaka and Gentleman, 1996) (http://www.r-project.org/) was used to represent the count of mapped reads between different samples in order to compare the difference between the sRNA antiviral mechanisms. The count tables were loaded in a R environment in order to draw different representation of counts of mapped reads according to the length and/or the strand of mapped reads, and the samples. These representations were saved in pdf files. The library Rsamtools was used to draw the logo of mapped reads and represents the proportion of each nucleotide at the 5'-end of reads (http://www.bioconductor.org/). To do this, specific functions were encoded at Fasteris and saved in an R library. These functions were called by the R script for statistical analysis.

2.3.2 *De novo* assembly algorithms

An assembly is a hierarchical data structure which reconstructs complete nucleotide sequences from shorter sequences. The *de novo* assembly is an assembly without a reference sequence. The sequencing reads provided by sequencers are concatenated by an assembler according to their overlapping regions into contigs. Each contig contains the consensus sequence determined by a multiple sequence alignment of reads. Each assembler provides consensus

sequences in an output fasta file. These sequences are often represented by strings of the characters A, C, G and T, for known nucleotides, and N or other letters if there is no consensus during the assembly steps. The quality of assembly is determined by the size and accuracy of their contigs and scaffolds. Quality values of an assembly include the maximum length of the contigs, their average length, the combined total length of all contigs and the N50. The N50 is the last contig of the group of largest contigs which represents at least 50% of the total assembly.

With the use of whole-genome shotgun (WGS) to sequence a complete genome, new *de novo* assembly algorithms and software were developed. These algorithms and software have been improved to face the emergencies of NGS which provides a big volume of data with short read lengths. All the NGS assemblers used algorithms based on graphs methods. Three main different algorithms are used for the assembly: the Overlap/Layout/Consensus (OLC) method based on an overlap graph, the *de Bruijn* graphs method based on K-mer graph, and the greedy-extension which is based on string comparison (Zhang et al., 2011).

For OLC, the overlapped regions are determined within read by pair-wise sequence alignments. The resulting graph contains nodes representing reads, and edges representing overlaps between the reads. The potential contigs are determined by paths through the graph. For the *de Bruijn* graphs method, the overlaps between subsequences of read are represented by a specific directed graph described by Nicholaas Govert de Bruijn. However, the assemblers use *de Bruijn* graphs based on K-mers. K corresponds to the length of subsequences created from the read sequence (Miller et al., 2010) (Fig. 2.3.2.1).

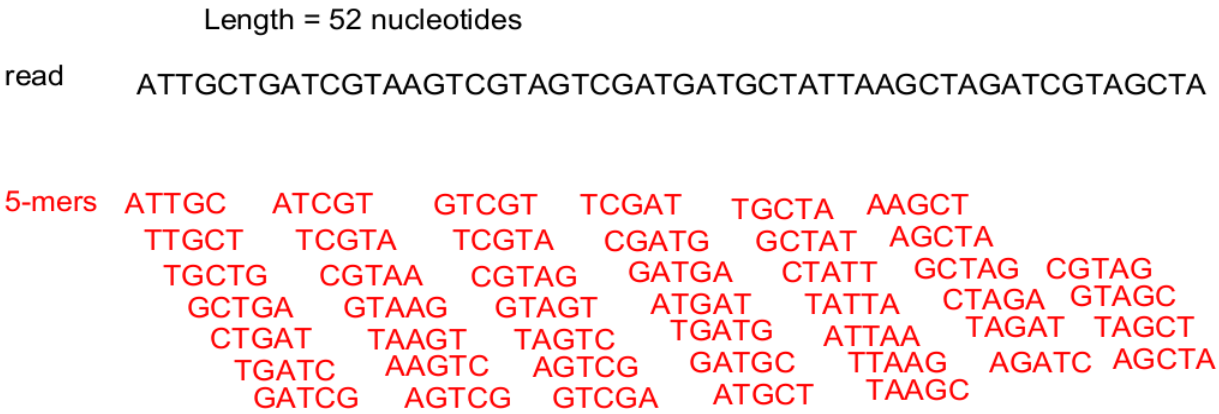


Figure 2.3.2.1: Example of creation of 5-mers from one read.

The different 5-mers are represented in red.

The *de Bruijn* graph is a specific k-mers graph where each edge is visited once to reconstruct the complete genome. In *de Bruijn* graph, each edge represents each k-mer and each node corresponds to (k-1)-mers (Compeau et al., 2011) (Fig. 2.3.2.2).

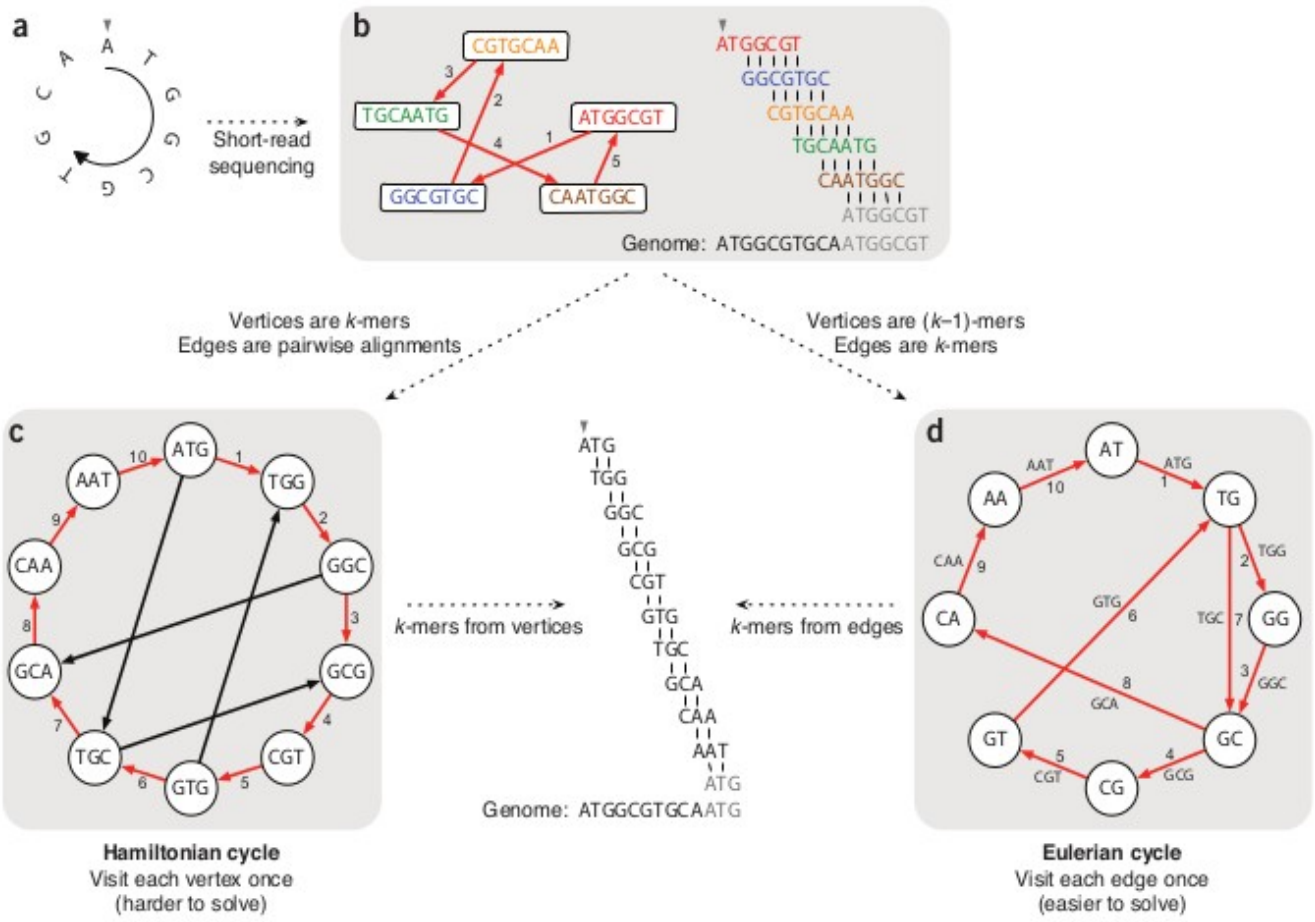


Figure 2.3.2.2: Comparison between a normal k-mers graph and a *de Bruijn* graph.

(**a**) An example small circular genome. (**b**) In traditional Sanger sequencing algorithms, reads were represented as nodes in a graph, and edges represented alignments between reads. Walking along a Hamiltonian cycle by following the edges in numerical order allows one to reconstruct the circular genome by combining alignments between successive reads. At the end of the cycle, the sequence wraps around to the start of the genome. The repeated part of the sequence is grayed out in the alignment diagram. (**c**) An alternative assembly technique first splits reads into all possible *k*-mers: with k =3, ATGGCGT comprises ATG, TGG, GGC, GCG and CGT. Following a Hamiltonian cycle (indicated by red edges) allows one to reconstruct the genome by forming an alignment in which each successive *k*-mer (from successive nodes) is shifted by one position. This procedure recovers the genome but does not scale well to large graphs. (**d**) Modern short-read assembly algorithms construct a ***de Bruijn*** graph by representing all *k*-mers prefixes and suffixes as nodes and then drawing edges that represent *k*-mers having a particular prefix and suffix. For example, the *k*-mer edge ATG has prefix AT and suffix TG. Finding an Eulerian cycle allows one to reconstruct the genome by forming an alignment in which each successive *k*-mer (from successive edges) is shifted by one position. This generates the same cyclic genome sequence without performing the computationally expensive task of finding a Hamiltonian cycle. Reprinted by permission from Macmillan Publishers Ltd: Nature Biotechnology (Compeau et al., 2011), copyright (2011).

Greedy-extension and OLC assemblers provide good results for small genomes with short and long reads when the computational power is limited. When the datasets contain more than hundred millions of short reads, the use of *de Bruijn* graph is preferable because of their short runtime and low RAM occupancy (Zhang et al., 2011). Consequently, d*e novo* assemblers based on *de Bruijn* graph were used during this study to assemble the viral genomes.

2.3.2.1 Velvet

Velvet is a group of algorithms which assemble a genome by building contigs from very short reads (length between 25 to 50 nucleotides) (Zerbino and Birney, 2008). To create contigs, Velvet builds a *de Bruijn* graph where each node corresponds to a series of overlapping *k*-mers.

The *de Bruijn* graph created by Velvet has nodes where only the last nucleotide is saved within the sequence of the node. Each node is linked to the node created by the reverse-complement sequence of *k*-mers. To create the graph, each read is hashed according to a *k*-mer length predefined by the user. The complementary sequences are automatically stored in a second database (Fig. 2.3.2.1.1).

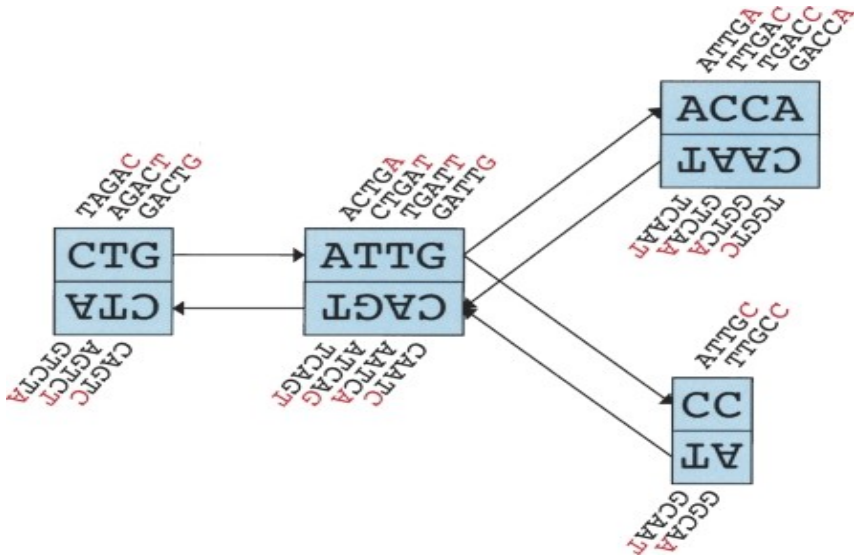


Figure 2.3.2.1.1: Schematic representation of the implementation of the *de Bruijn* graph.

Each node, represented by a single rectangle, represents a series of overlapping *k*-mers (in this case, *k* = 5), listed directly *above* or *below*. (Red) The last nucleotide of each *k*-mer. The sequence of those final nucleotides, copied in large letters in the rectangle, is the sequence of the node. The twin node, directly attached to the node, either *below* or *above*, represents the reverse series of reverse complement *k*-mers. Arcs are represented as arrows *between* nodes. The last *k*-mer of an arc's origin overlaps with the first of its destination. Each arc has a symmetric arc. Note that the two nodes on the *left* could be merged into one without loss of information, because they form a chain. Copied from (Zerbino and Birney, 2008).

When the graph creation is finished, it is simplified: the nodes which are linked only by one directed edge are merged. The non-merged edge must be due to the presence of errors. To remove the errors, Velvet proceeds in three steps: it removes ‘tips‘ which are a chain of nodes that is disconnected on one end, then it removes « bubbles» which are closely similar paths between similar nodes according to the path which have the highest coverage by the reads, and finally removes erroneous connections according to a threshold defined by the users. The created contigs are finally saved within fasta files according to the selected value of *k*-mers (Zerbino and Birney, 2008).

2.3.2.2 Oases

Oases is a *de novo* transcriptome assembler designed to reconstruct transcripts from short read sequencing (Schulz et al., 2012). It completes the Velvet assembly, and was adapted for transcriptome analysis. Oases allows to find alternative splicing and different isoforms within the final assembly. Oases uses a preliminary assembly produced by the Velvet assembler. The contig correction proposed by Oases differs from that of Velvet. To remove “bubbles” found in the graph, Oases can pass many times by the same nodes in order to have an exhaustive analysis. Edges are removed if their coverage is less than 10% of the sum of coverages of outgoing edges from that same node. Contigs are removed when their static coverage is less than 3 times (by default). A scaffold is created by a summary of distance information between the contigs. The connection between two contigs is direct when it is supported by at least one spanning read. A weight value is assigned for each connection. Then, the scaffold is filtered according to the weight or the support. The scaffold allows to separate the contigs according to their lengths: they are named 'long' if the contig length is bigger than a given threshold, otherwise they are named 'short'. After the filtering, the contigs are clustered within 'loci'. The long contigs are clustered into connected components. Then, the short contigs are added to the loci where they are connected to one of the long nodes. After the locus construction, the loci are reduced by removing the long distance connections when they can be replaced by short distance connections. Only connections between immediate neighbours are retained (Fig. 2.3.2.2.1). Each locus must be analyzed to find the alternative splicing events according to the abnormal topologies found within the *de Bruijn* Graph (Schulz et al., 2012). Here, Oases was used because it improves the *de novo* assembly provided by Velvet, and can be adapted for RNA viruses.

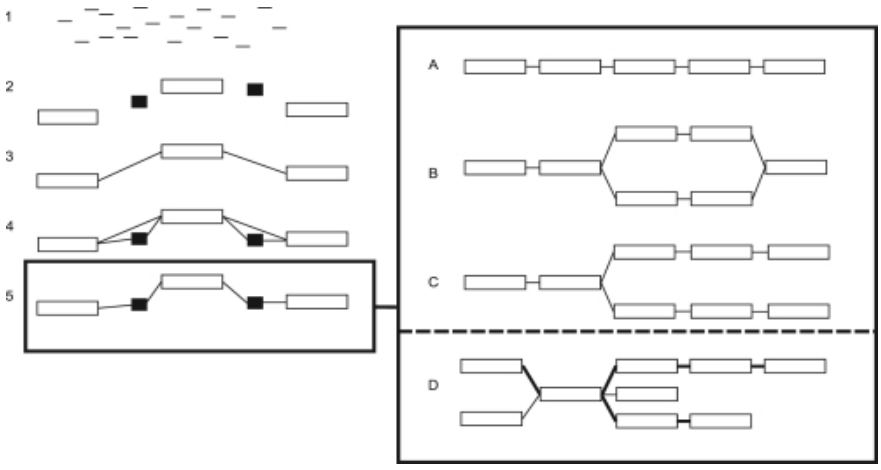


Figure 2.3.2.2.1: Schematic overview of the Oases pipeline.

(1) Individual reads are sequenced from an RNA sample, (2) contigs are built from those reads, some of them are labelled as *long* (clear), others *short* (dark), (3) long contigs, connected by single reads or read-pairs, are grouped into connected components, called *loci*, (4) short contigs are attached to the loci, (5) the loci are transitively reduced. Trans fragments are then extracted from the loci. The loci are divided into 4 categories: (A) chains, (B) bubbles, (C) forks and (D) complex (i.e. all the loci which did not fit into the previous categories). Copied from (Schulz et al., 2012).

2.3.2.3 Metavelvet

With the improvement of the deep-sequencing technologies and bioinformatic softwares, it is possible to sequence and analyse genomes of multiple organisms present in one sample, using a so-called metagenomics approach. In the case of infected plants, it is interesting to use *de novo* assemblers created especially for metagenome sequencing. Some reads can be shared between the host plants and the viral genome, and consequently, can be the origin of errors during the *de novo* assembly. Moreover, some begomoviruses studied here have two circular genomic components, and they must be reconstructed in two different contigs at minimum. Some plant diseases such as grapevine led leaf disease investigated here are caused by two or more viruses or viroids.

Metavelvet proposes to do a *de novo* assembly of metagenome by using and completing the velvet program. This program was developed to assemble multiple species by using mixed short reads. Metavelvet performs the *de novo* assembly of a metagenome in four steps. The first and last steps use Velvet's functions. The first step is the creation of *de Bruijn* graph with Velvet. The second step counts the *k*-mer frequency and represents it by one histogram. Each node is linked to one bar among the histogram according to the node coverage calculated by Velvet during

the *de Bruijn* graph creation. The histogram must have multiple peaks because the k-mer frequencies follow a Poisson distribution in a single-genome assembly, and consequently the histogram show a mixture of Poisson distribution due to the presence of different genomes (Wu and Ye, 2011) (Fig 2.3.2.3.1).

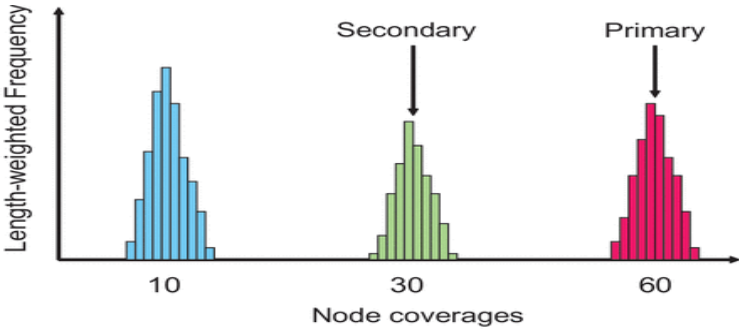


Figure 2.3.2.3.1: Detection of multiple peaks in the histogram of coverage values of the nodes.

Each peak corresponds to one genome. The primary peak indicates the nodes which must used to reconstruct the first genome. The secondary peak indicates the nodes to reconstruct the second genome . Copied from (Namiki et al., 2012).

The third step is to separate the *de Bruijn* graph in different subgraphs according to the highest peak observed within the histogram (Fig 2.3.2.3.2). Metavelvet finds the common nodes to delimitate the subgraphs according to the histogram. The last step uses the assembly methods provided by Velvet to create the contigs by reading the subgraph corresponding to the highest peak. Then, the corresponding nodes are removed from the initial *de Bruijn* graph, and the three last steps are repeated for each peak until the use of all nodes. The contigs and scaffold corresponding normally to each genome, are saved within fasta files (Namiki et al., 2012).

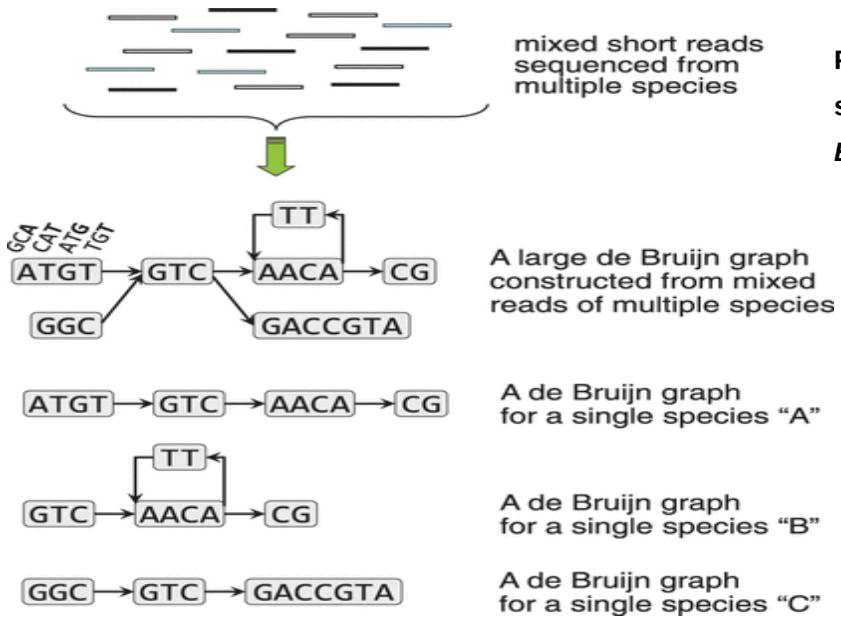


Figure 2.3.2.3.2: The MetaVelvet strategy to decompose a mixed *de Bruijn* graph.

Copied from (Namiki et al., 2012).

2.3.2.4 Seqman Pro (DNAstar)

Seqman Pro is an assembly tool provided by the DNAstar package. It uses an OLC method for the assembly. We used it for the scaffold step to assemble the contigs created with Oases or Metavelvet. In fact, the contigs created according different *k*-mers values can have overlapped regions. Seqman find these overlapped regions by aligning the sequences, and then merges and extends the corresponding contigs in order to obtain few longer contigs. The final objective of this step is to obtain the complete viral genomes.

3. Results

3.1 MISIS

3.1.1 Presentation of MISIS

In order to better visualize and analyze the diverse populations of vsiRNAs, we developed a new bioinformatics tool called MISIS. Biologists need to have a representation of mapping along the reference sequence per each position. The existing tools did not represent the strands of mapped read along the viral genome with different bars. All the counts are merged within only one bar with free softwares such as Seqmonk and IGV. So far, only those biologists who have informatics skills could draw the histogram of coverage of mapped reads on the forward or reverse strand with programming language such as R. We created a software, named MISIS, which allows to draw this kind of representation. The histogram is drawn directly from reading a file which contains a table of the mapped reads selected according to the size-classes and the position along the genome. Its positive bars correspond to the number of mapped reads on the forward strand, and its negative bars correspond to the mapped reads on the reverse strand. These two types of bars are in different colours to accentuate their visualization impact. This kind of histogram representation allows to observe hot- and cold-spots for the coverage of mapped reads. Moreover, MISIS allows to zoom in one particular region and to select only one size of mapped reads (Fig 3.3.1.1).

The main objective of this tool is to be easy to use for biologists and to allow to run one's

eye over the coverage in details. The other objective of MISIS is to parse the Bam/Sam files directly to extract the specific information of mapping. The two tables are generated after the analysis of a parsed file: one contains only the reads which map perfectly along the genome, and the second contains all mapped reads (number of mismatches is determined by the reads length). These tables are saved as text files in order to allow their use for other analyses such as statistical analysis for biologists.

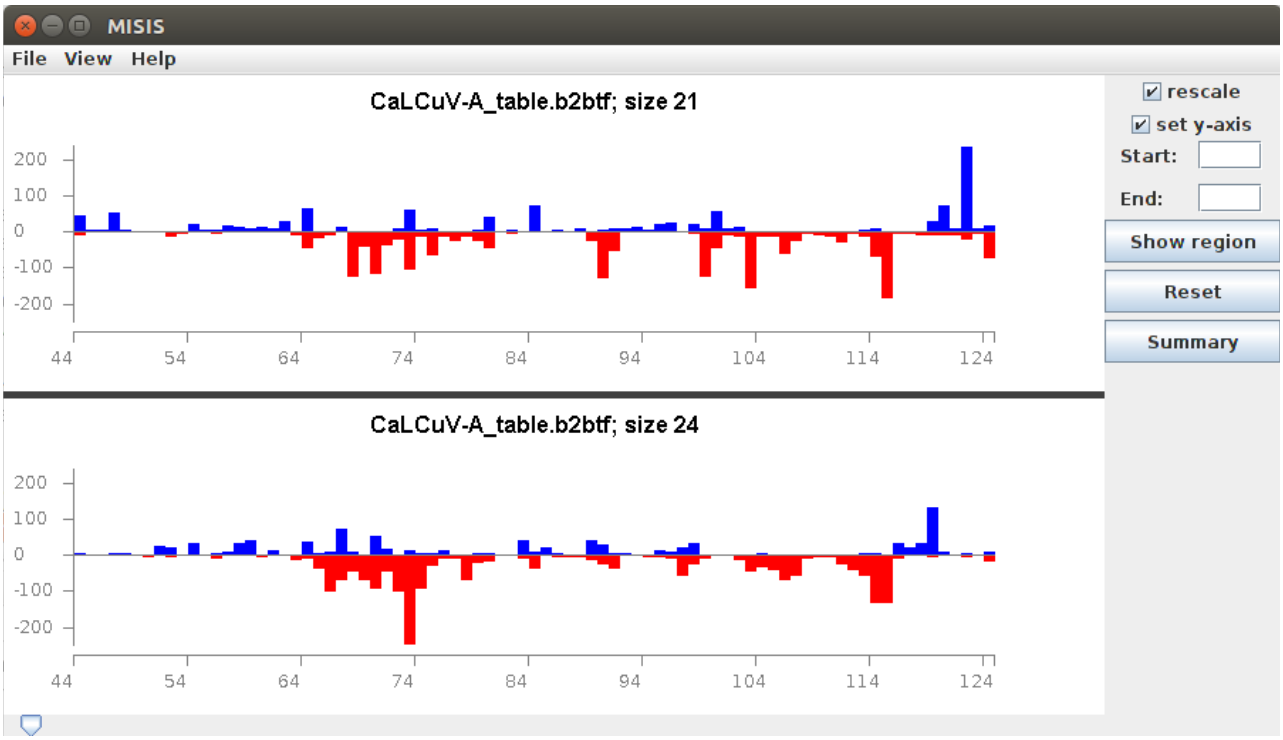


Figure 3.3.1.1: Visualization of mapping against CaLCuV-A genome with MISIS

Here, the mapped reads are represented with two histograms for sRNA of 21 and 24 nucleotides-long. The blue bars represent the mapped reads on the forward strand and the red bars on the reverse strand. The y-axis indicates the count of mapped reads and the x-axis indicates the positions along the CaLCuV-A viral genome.

3.1.2 Functioning of MISIS

When MISIS parses the Bam/Sam file, it reads firstly the header in order to extract the length of the mapped genome. Then, the tables are generated. The number of rows is determined by the length of the genome. The number of columns is determined by the range of size-classes of

sRNA chosen by the user: one column per size-class and per strand, and three additional columns for the merged count on forward and/or reverse strands. Each cell of tables is initialized to 0. After the header, each line of Bam/Sam files is read and the information about the position, the strand, the CIGAR, the sequence and the MD tag for each mapped read. The position allows to localize the row of table which must be incremented. The strand and the length of sequence allow to localize the columns of the row which must be incremented. The CIGAR and the MD tag allows to know if the table for perfect match mapped reads must be also incremented; the condition of this increment is that the CIGAR must indicate the read sequence added by the letter M and the MD tag must have a value corresponding to the read sequence. When the last line is read, the tables are saved as text files in the folder chosen by the user. If there are different genomes used for the mapping, the previous steps are repeated for each genome. The creation of tables is a step unlinked to the visualization of the tables. According to the capacity of the user's computer, the time of the table creation is variable and depends also to the length of reference genome and the number of mapped reads. All these variables prevent to estimate the time for the tables creation. To help the user to have an estimation of the time, a progress bar appears within a window to indicate the progression of the table's creation. When the text files are created, they are directly saved on the user's computer and can be used as long as they are not deleted by the user. These files can used to draw the corresponding histograms with MISIS.

The main objective for MISIS is to be easy to use for biologists. The use of MISIS must be clear and intuitive. MISIS generates a visual interface (GUI: Graphical User Interface) which allows the biologists to use his computer mouse and keyboard to draw the histogram. The user can zoom on specific region by indicating the position of this region or selecting with computer mouse. They can also select the colour of bars within the histogram, the size-classes which must be visualized and saved an image of the visualization as a file. All these functions of MISIS are explained in a tutorial provided on the website of MISIS (http://www.fasteris.com/apps/). MISIS is written in Java language because Java uses a virtual machine independent of the Operating System (OS). Consequently, it can run on windows, mac or linux environments.

3.1.3 Implementation of MISIS

The internal structure of MISIS is divided in three parts, based on informatics pattern named MVC for Model-View-Controller. The advantage of this pattern is that it allows to divide the source code of MISIS in three parts which have a specific role within MISIS: the View contains the code that manages the GUI, the Controller controls the request asked by the user via the GUI before to transmit it to the Model and, the Model contains the virtual model of the data and update

the View according to the last modification of the model. For MISIS, the View allows to draw the histogram, the Controller controls the data saved from the files and the Model contains the tables which are used to draw the corresponding histograms. Only the user's actions on GUI are analyzed by Controller when they have an incidence on the tables in the Model, otherwise all user's action on the histograms are managed by the View parts (Seguin et al., 2014b; see the paper in the Annexes). The three parts structures facilitate further developments of the tools by limiting the risk of bugs. Another advantage is to allow the development by different computer and biocomputer specialists which work on a specific part.

3.2 Mapping Results for ORMV, CaMV and CaLCuV

3.2.1 Mapping Result for ORMV

3.2.1.1 Correction of the viral genome sequence.

We mapped the small RNA populations from ORMV virion-infected Arabidopsis wild type (Col-0) and silencing mutant plants to the reference sequence of ORMV genomic RNA corresponding to the available ORMV full-length genome clone which did not exhibit wild type virus infectivity. The SNP profile showed the presence of 3 SNPs at positions 12, 231 and 5612 of the ORMV genome (Fig 3.2.1.1.1). The analyses for all the ORMV-infected plants resulted in the same profile where more than 80% of sRNA reads support these SNPs. The visualization of the mapping with IGV for the BPO-38 sample (ORMV-infected Col-0) showed that more than 98% of reads support the nucleotide changes: G12A, A231T, G5612T (the available 'non-infectious' ORMV clone sequence and the mapped read sequence, respectively). This result indicates the presence of sequencing or cloning errors in the ORMV clone. Based on these findings and *de novo* reconstruction results (see below) we corrected the ORMV reference sequence accordingly.

The A231T is a silent mutation for the viral replicase. The G12A mutation can potentially be responsible for the loss of infectivity for the available clone because it is expected to prevent ORMV replication. The G5612T mutation can also be responsible for the loss of infectivity because it introduces a premature stop codon in the MP ORF, without affecting the CP amino acid sequence. The experiments were undertaken to validate the mutation involved in the loss of infectivity (see below Fig 3.2.1.1.2).

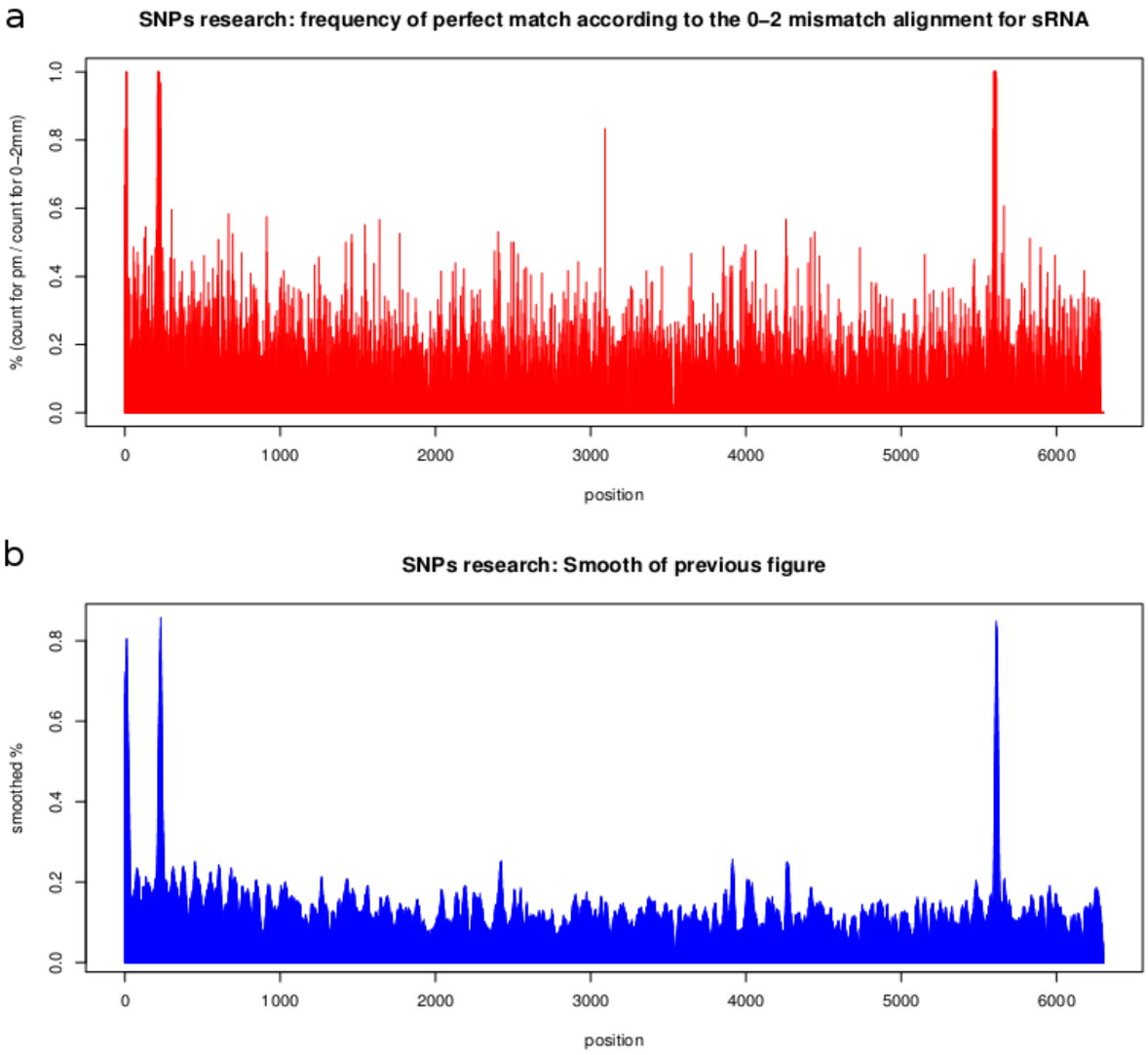


Figure 3.2.1.1.1: SNPs profile for ORMV genome

The x axis of these figures represent the position along the viral genome. **(**a) This SNP profile is generated by comparison of reads which map with perfect match and 0-2 mismatches. The y axis represents the percentage of count of perfect matched read divided by the count of mapped reads with mismatches. (b) This SNP profile corresponds to the smoothing of the previous profile. The smoothing has been performed by doing the average of percentage according to the size of mapped reads for each position.

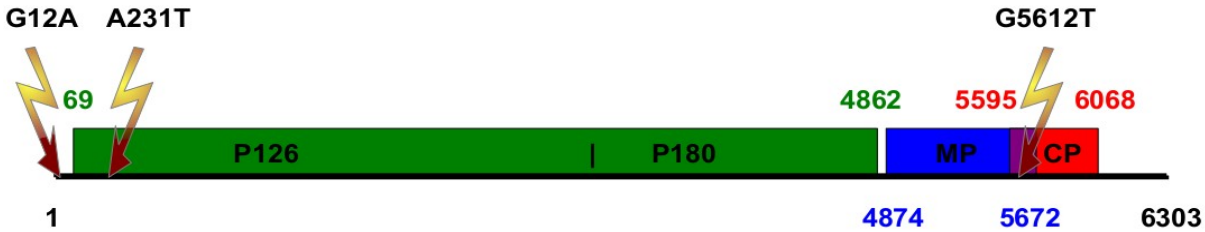


Figure 3.2.1.1.2: SNPs found on the ORMV genome

3.2.1.2 Analysis of ORMV-derived siRNAs.

The counts of ORMV genome-derived sRNAs showed that there is a dramatic decrease of viral sRNA production in the *dcl2/3/4* mutant plants infected by ORMV (BPO-40 and 46), compared to the Col-0 controls (BPO-38 and 44). No big difference was observed at two different time points post-inoculation (14 days post inoculation (dpi): BPO-38, 39 and 40; and 22 dpi: BPO-44, 45 and 46). In Col-0, the 21-nt viral reads constituted the majority of reads, followed by 20 and 22-nt reads, but this size profile was not observed for *dcl2/3/4*. (Fig 3.2.1.2.1). The counts did not differ between perfectly mapped reads and reads which map with 0-2 mismatches. There was no substantial difference between the wild type Col-0 and *rdr1/2/6* mutant plants infected by ORMV, contrary to *dcl2/3/4* plants which accumulated greatly reduced amounts of viral sRNAs. Thus, RDR1, RDR2 and RDR6 proteins do not appear to be involved in the biogenesis of ORMV-derived sRNAs, whereas DCL2, DCL3 and DCL4 are required for ORMV sRNA biogenesis.

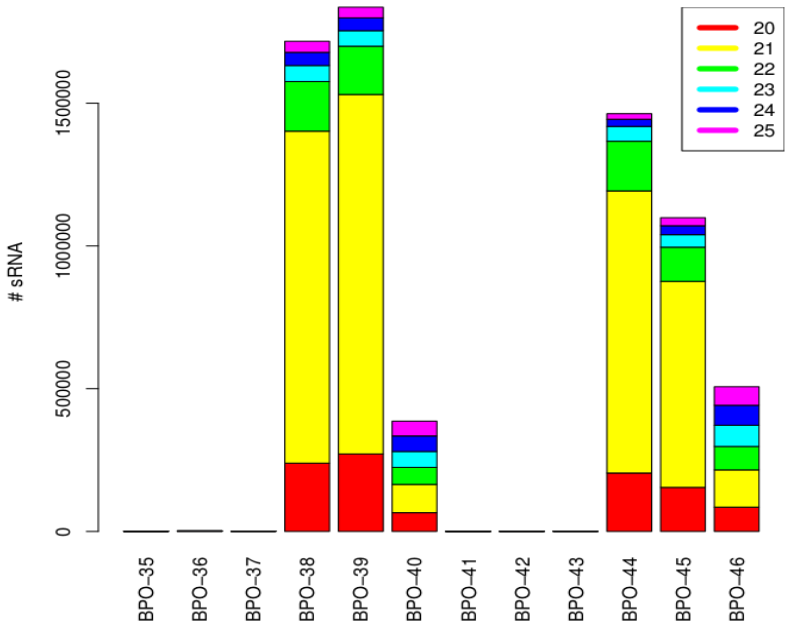


Figure 3.2.1.2.1: Distribution of mapped reads against the ORMV genome.

The histogram represents the count of mapped reads for the virus-infected (BPO-38-40 and BPO-44-46) and mock-inoculated (BPO-35-37 and BPO-41-43) samples. The y-axis represents the count of mapped reads. The bars have separated colours according to the size of mapped reads.

The ORMV genome coverage with viral sRNA reads was found to be homogeneous, without any major hot and cold-spot of sRNA production. Apparently the entire viral genome is targeted in both sense and antisense strands during the anti-viral defense. Furthermore, the coverage profile did not give any indication that the viral subgenomic RNAs for MP and CP are targeted preferentially, compared to the genomic RNA. However, the viral sRNA profiles in *dcl2/3/4* mutant plants lacking all DCLs but DCL1 (BPO-40 and 46) indicated that there is a preference for the forward strand for the sRNA biogenesis by the DCL1 pathway (Fig 3.2.1.2.2).

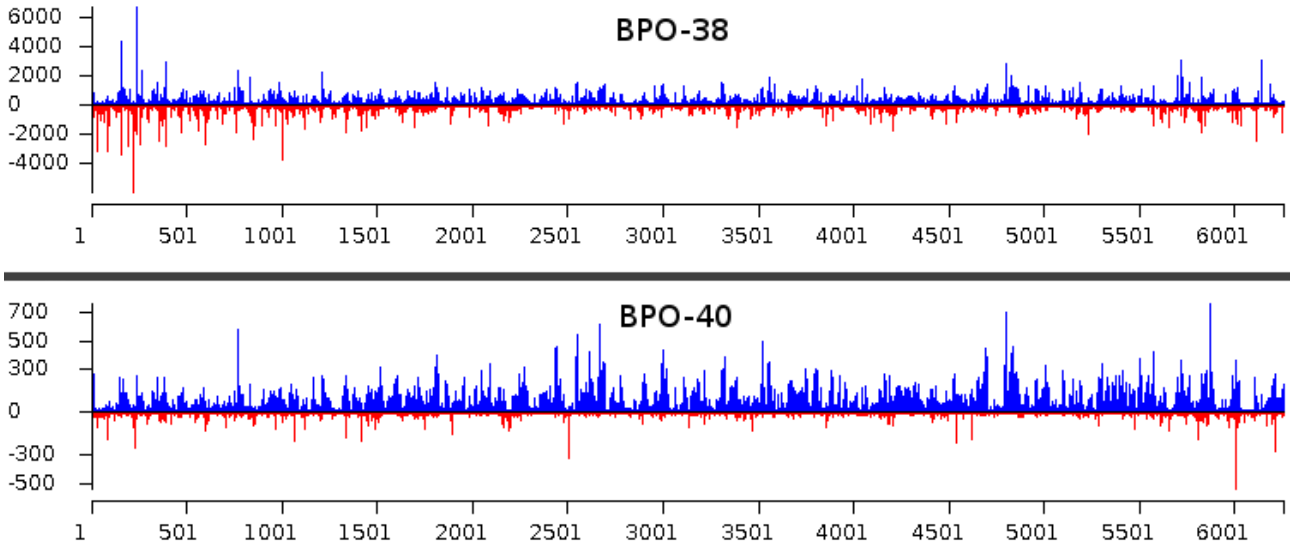


Figure 3.2.1.2.2: Coverage of mapped reads (20-25 nucleotides-long) along the ORMV genome for BPO-38 (wild type plant) and BPO-40 (*dcl-2/3/4* triple mutant plant).

The x-axis represents the ORMV genome. The y-axis represents the count of mapped reads. The blue bars represent the count of reads mapped on the forward strand, and the red bars represent the count of reads mapped on the reverse strand. These profiles include mapped reads with 0-2 mismatches.

The analysis of 5' nucleotides of ORMV sRNA reads showed that the majority of reads starts with U (~33 %), followed by A (~27 %). When the distribution of each nucleotide along the viral genome was compared (A = 28%, C = 19%, G = 26% and U = 27%), we observed an abnormal enrichment of 5' U sRNA, although the coverage showed relative homogeneity of mapped reads for all the size classes. The 22-nt reads had the highest proportion of 5' U. This bias was not observed for the *dcl2/3/4* mutant plants (BPO-40) (Fig 3.2.1.2.3). Based on these results, we hypothesize that AGO1 (and/or AGO10) may associate with the majority of viral sRNAs, followed by AGO2. DCL2, which produces 22-nt sRNA, may preferentially produce sRNA with 5' U, or 22-nt sRNAs are preferentially associated with and stabilized by AGO1 (and/or AGO10).

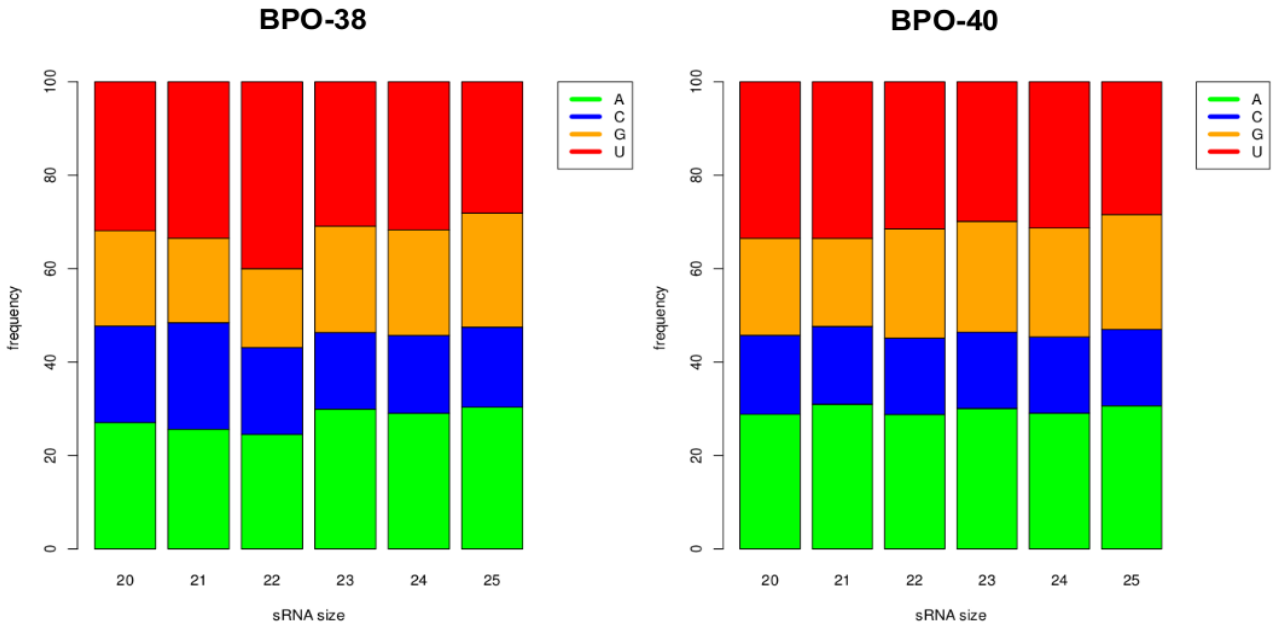


Figure 3.2.1.2.3: Distribution of 5' nucleotide among reads for BPO-38 (wild type plant) and BPO-40 (*dcl-2/3/4* triple mutant plant) along the ORMV genome.

Each bar corresponds to one size classes of reads. The y axis indicates the percentage values found for the distribution.

3.2.1.3 Analysis of endogenous sRNAs in A. thaliana.

The counts of reads mapped to the *Arabidopsis thaliana* genome in non-infected control plants (BPO-35 and BPO-41) show that the most abundant size class of plant sRNAs is 24-nt, followed by 23-nt and 21-nt. In contrast in ORMV-infected plants, the 21-nt class is the most abundant (BPO-38 and BPO-44), suggesting that ORMV infections have a major impact on the biogenesis of plant sRNAs. This virus impact on the host plant sRNA profile is similar to the impact of triple mutations of DCL2, DCL3 and DCL4 in dcl2/3/4 plants, and triple mutations of RDR1, RDR2 and RDR6 in *rdr1/2/6* plants in which plant siRNA biogenesis is diminished and only DCL1-dependent pathway generating predominantly 21-nt miRNAs is operating (Fig 3.2.1.3.1). Taken together, ORMV infection appears to boost DCL1-dependent miRNA production as was observed previously for selected miRNAs using RNA blot hybridization (Blevins et al., 2006).

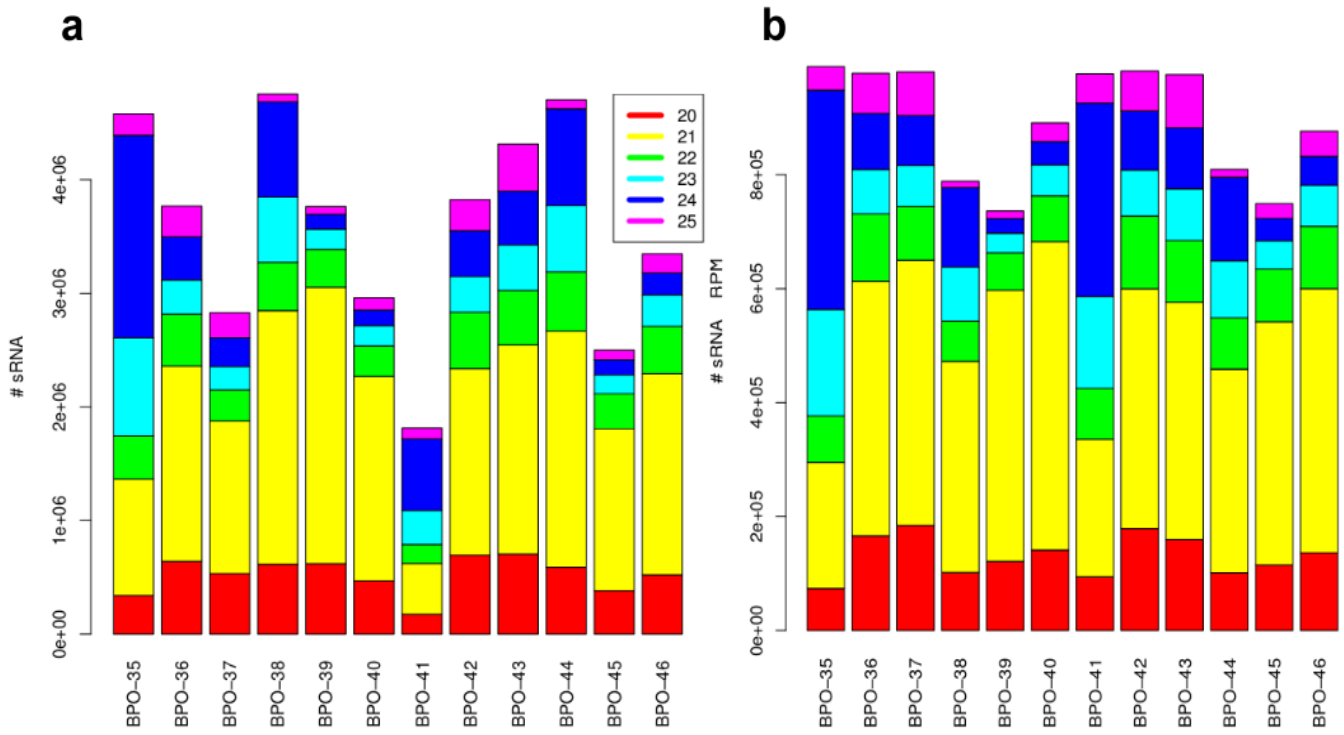


Figure 3.2.1.3.1: Distribution of mapped reads against *Arabidopsis thaliana* genome.

(a) The histogram represents the count of mapped reads for BPO-35 – 46 samples. ORMV-Infected samples are BPO-38,39,40 and BPO-44,45,46. The *rdr1/2/6* mutant plants are BPO-36,39,42,45. The *dcl2/3/*4 mutant plants are BPO-37,39,42,46. The y-axis represents the count of mapped reads. The bars have separated colours according to the size of mapped reads. (b) It is the same histogram after normalization of mapped reads per million of 20-25 nucleotides-long reads.

3.2.2 Mapping and counting of CaMV-derived siRNAs

The mapping of sRNAs from CaMV-infected wild type plants showed that the majority of vsiRNAs belong to 24-nt size class, followed by 21 and 22-nt size-classes. This result indicates that the 24-nt siRNA-directed DNA methylation pathway is involved in antiviral defence, in addition to the 21/21-nt siRNA-directed posttranscriptional silencing pathway. The size profile of vsiRNAs did not differ in CaMV-infected mutant plants for AGO2 and AGO3 (BPO-21 and BPO-22 respectively) (Fig 3.2.2.1), indicating that AGO2 and AGO3 are not required for vsiRNA biogenesis.

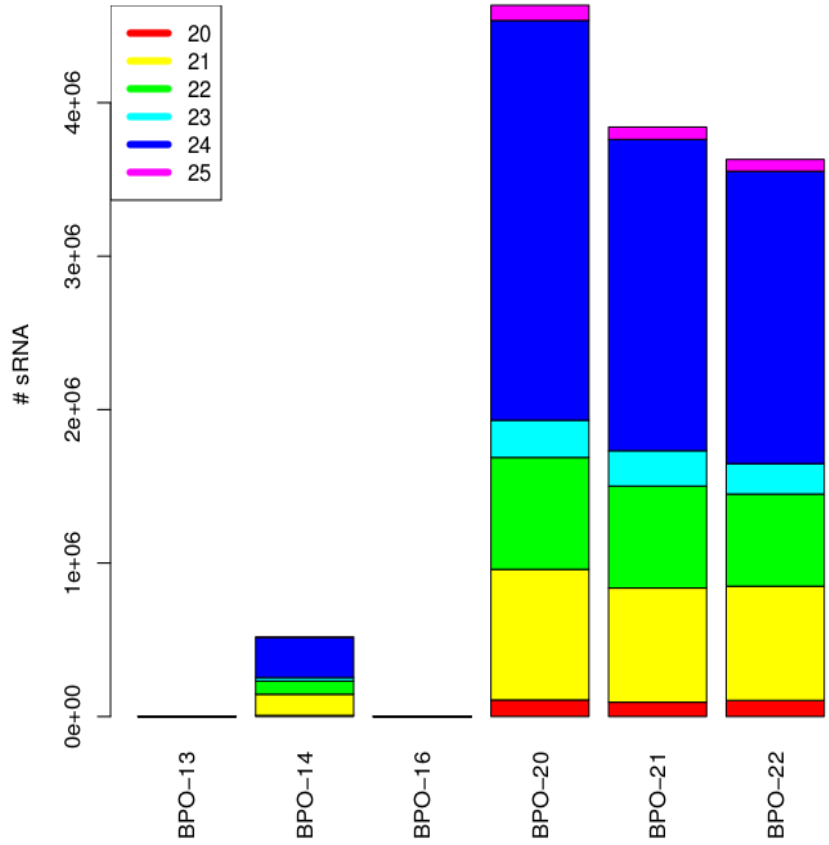


Figure 3.2.2.1: Distribution of mapped reads against CaMV genome.

The histogram represents the count of mapped reads for BPO-13, BPO-14, BPO-16 and BPO-20-22 samples. BPO-13 and BPO-16 are mock-inoculated samples. Other samples are infected with CaMV. BPO-16 and BPO-21 are *ago2* mutant plants. BPO-22 is an *ago3* mutant plant. The y-axis represents the count of mapped reads. The bars have separated colours according to the size of mapped reads.

The coverage of vsiRNAs along the CaMV genome is heterogeneous: the majority of vsiRNAs map to the 600 bp non-coding region of the genome in both strands (Fig 3.2.2.2). This region seems to be the main origin of the vsiRNA biogenesis involved in antiviral defense and viral counter-defense (Blevins et al. 2011). Nevertheless, the entire genome is covered by non-redundant sRNA reads, which allowed for a *de novo* assembly of the complete CaMV genome using these datasets (see below). This finding proved that, even if the coverage is heterogeneous, it is possible to reconstruct the complete viral genome from siRNAs generated by the NGS technology.

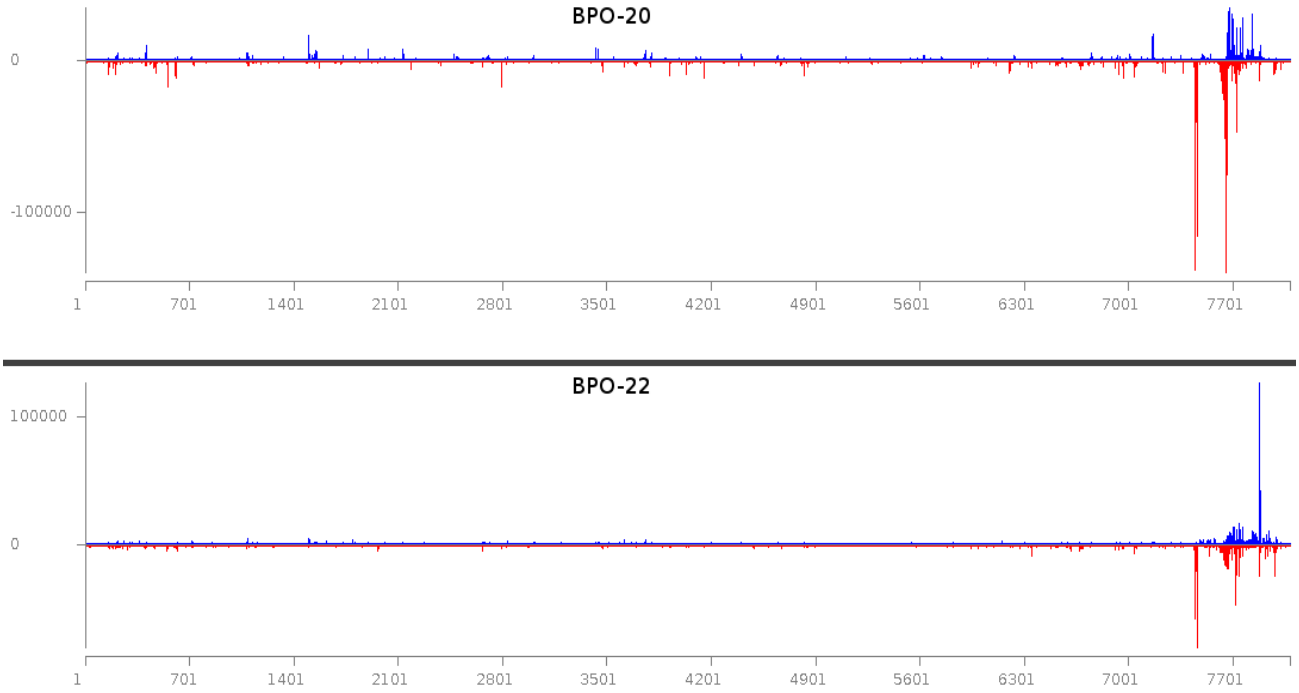


Figure 3.2.2.2: Coverage of mapped reads (20-25-nt) along the CaMV genome for BPO-20 (wild type plant) and BPO-22 (*ago3 mutant*).

The x-axis represents the ORMV genome. The y-axis represents the count of mapped reads. The blue bars represent the count of reads mapped on the forward strand, and the red bars represent the count of reads mapped on the reverse strand. These profiles include mapped reads with 0-2 mismatches.

The analysis of 5' nucleotides of the viral reads showed that the majority of 24 and 21-nt reads start with A followed by U. The majority of 22-nt vsiRNAs start with U. The bias for 5'-A indicates that AGO2 and AGO4, might be associated with viral 21-nt and 24-nt vsiRNAs, respectively. In *ago2* mutant plants, no substantial alteration in 5' nucleotide profiles was observed, except that 20-nt vsiRNAs showed a bigger proportion of 5'- C and a smaller proportion of 5' U. (Fig 3.2.2.3). This indicates that AGO2 may stabilize some of the 20-nt 5'A-sRNA, while, in the absence of AGO3, AGO5 takes over to stabilize 20-nt 5'C-sRNAs.

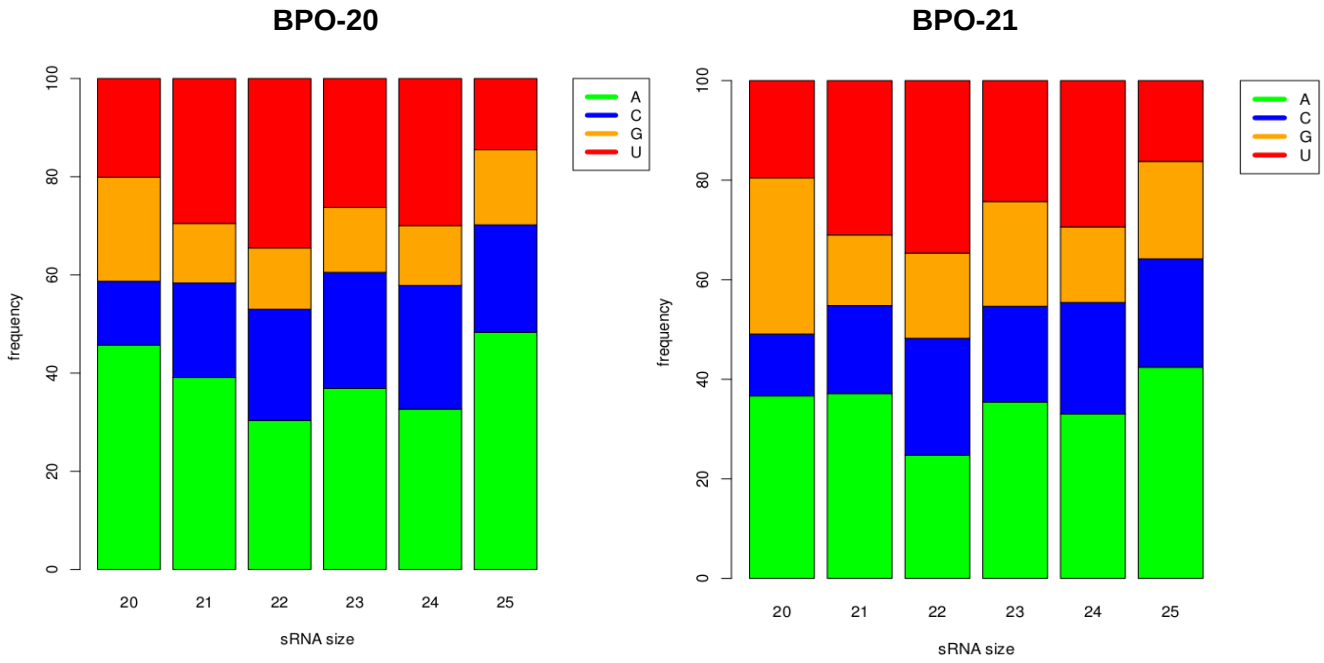


Figure 3.2.2.3: distribution of 5' nucleotide among reads for BPO-20 (wild type plant) and BPO-21 (*ago2* mutant plant) along CaMV genome.

Each bar corresponds to one size classes of reads. The y axis indicates the percentage values found for the distribution.

3.2.3 Mapping and counting of CaLCuV-derived siRNAs

The Mapping of sRNAs from CaLCuV-infected Arabidopsis plants showed that the majority of vsiRNA belong to 21-nt size-class, followed by 24 and 22-nt classes. The relative accumulation of 21-nt and 24-nt vsiRNA is comparable in wild type (BPO-57) and *rdr1/2/6* triple mutant infected plants (BPO-58). The total amount of DNA-B derived 20-25-nt siRNAs are increased in *rdr1/2/6*, while that of DNA-A-derived siRNAs is slightly decreased. (Fig 3.2.3.1). Moreover, there is no significant difference in the distribution of vsiRNA reads between the forward and reverse strands of the DNA-A or DNA-B. Taken together, these results suggest that RDR1, RDR2 and RDR6 are not required for vsiRNA biogenesis.

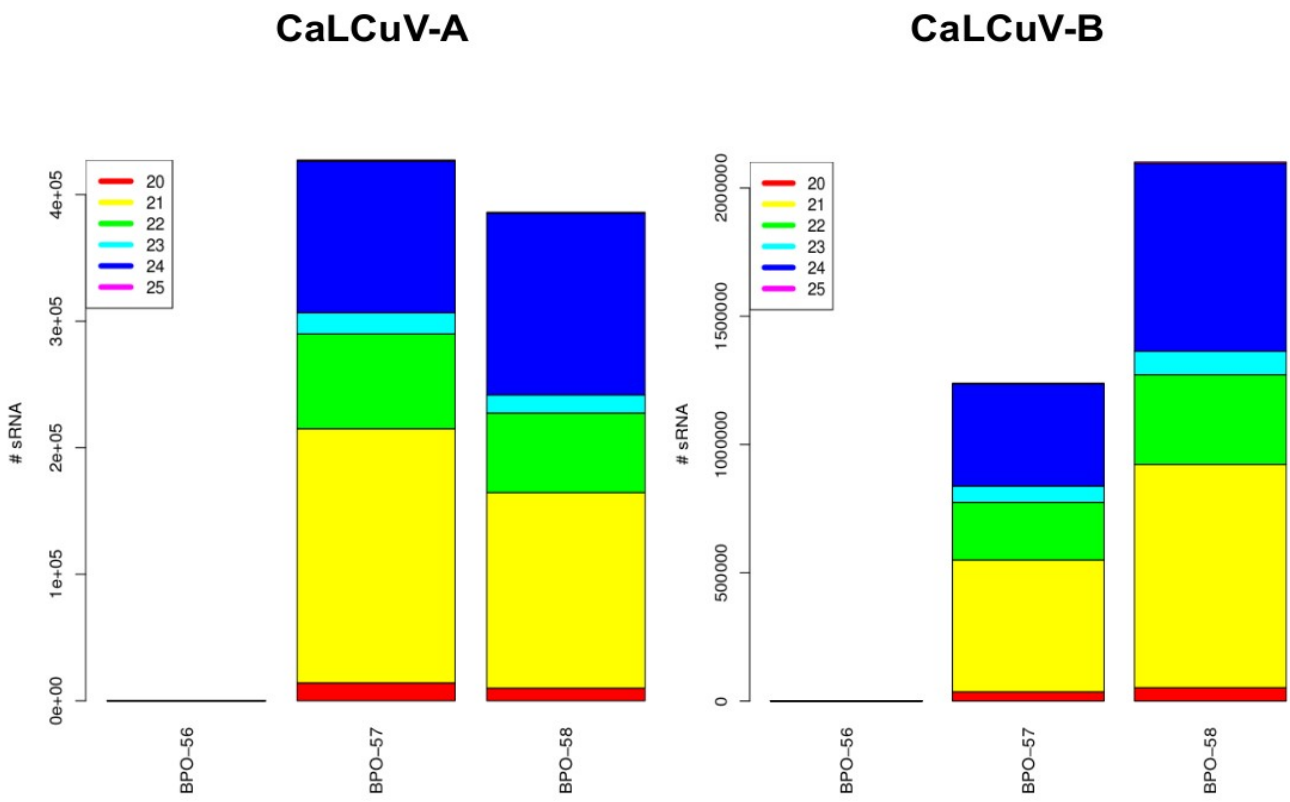


Figure 3.2.3.1: Distribution of mapped reads against CaLCuV-A & B genomes

The histograms represent the count of mapped reads for virus-infected samples BPO-57 and BPO-58. BPO-56 is the mock-inoculated wild-type plant sample. BPO-58 is a *rdr-1/2/6* triple mutant plant. The y-axis represents the count of mapped reads. The bars have separated colours according to the size of mapped reads.

The coverage of viral sRNA reads along the CaLCuV genome components A and B is heterogeneous. Two regions of the DNA-B are covered by the majority of reads. For the DNA-A, the coverage is somewhat more homogeneous. The siRNA hot-spot profiles are similar for each of the major size classes (Fig 3.2.3.2).

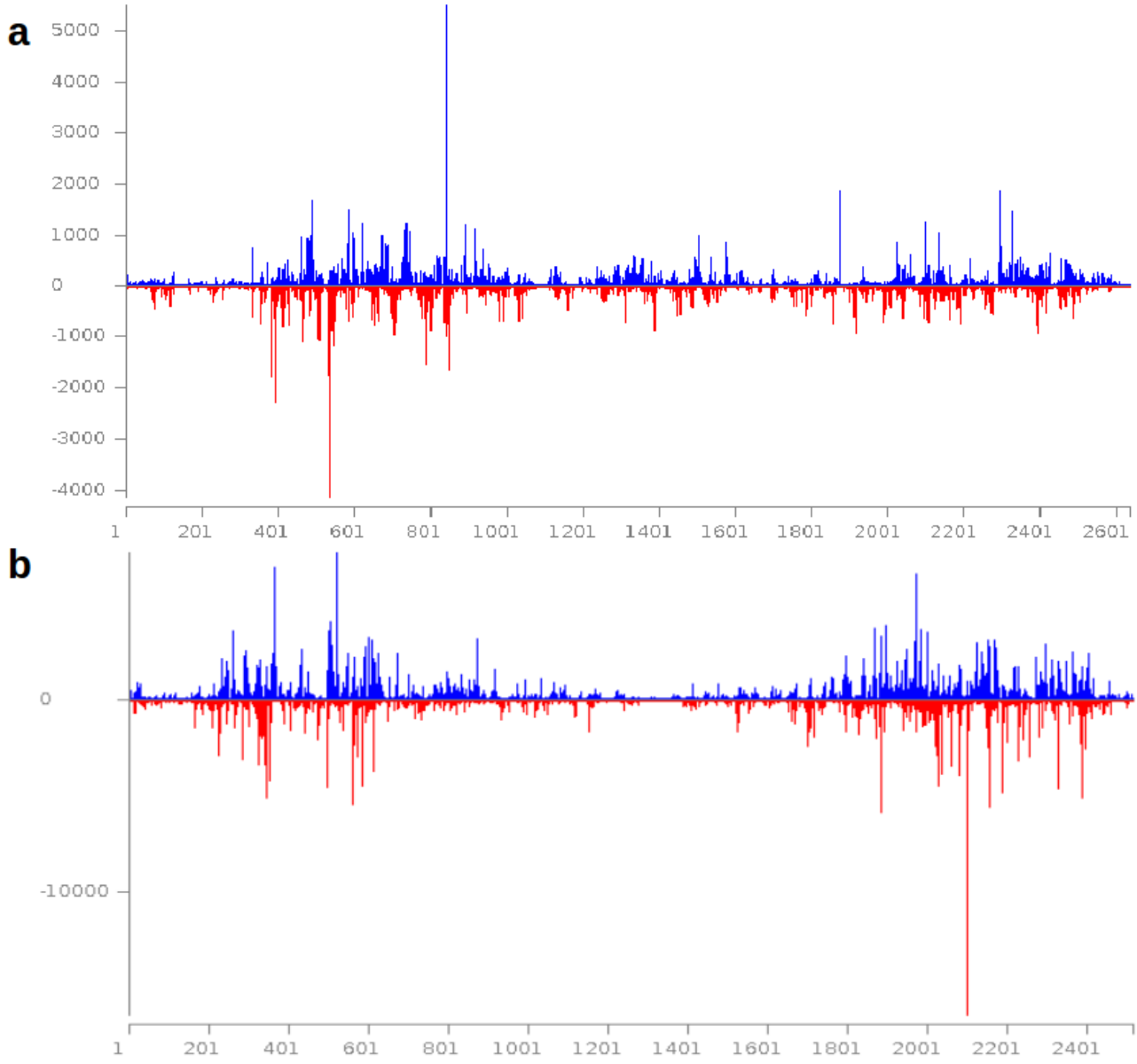


Figure 3.2.3.2: Profile of mapped reads along CaLCuV-A & B DNAs for BPO-57

The x-axis represents the DNA-A (**a**) and the DNA-B (**b**) of the CaLCuV genome. The y-axis represents the count of mapped reads. The blue bars represent the count of reads mapped on the forward strand, and the red bars represent the count of reads mapped on the reverse strand. These profiles include mapped reads with 0-2 mismatches.

The 5' nucleotide profile of vsiRNAs is similar for all size-classes and for both DNA-A and DNA-B components of the genome. There is no significant bias for any 5' nucleotide in any size class (Fig 3.2.3.3).

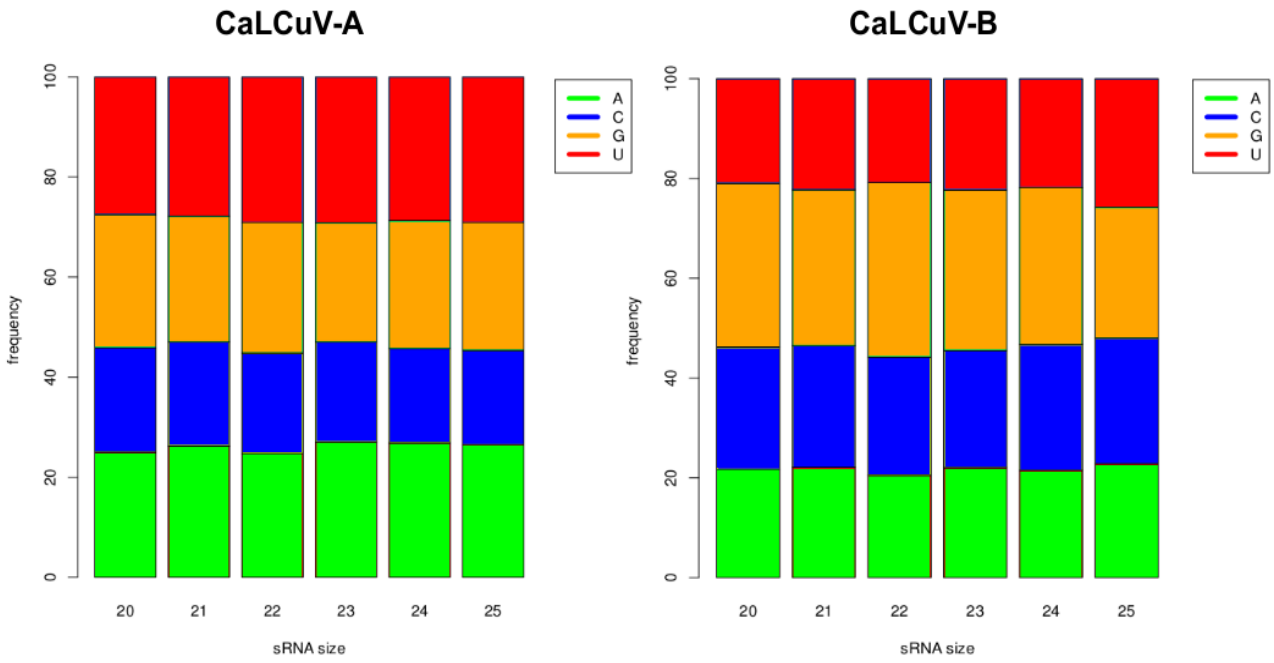


Figure 3.2.3.3: 5' nucleotide profile of viral sRNA reads for BPO-57 mapped along the CaLCuV-A and CaLCuV-B.

Each bar corresponds to one size classe of reads. The y axis indicates the percentage values found for the distribution.

3.2.4 Analysis of non-redundant viral reads

The analysis of the mapping analysis of non-redundant viral reads revealed that, for both RNA (ORMV) and DNA (CaMV and CaLCuV) viruses, the vsiRNA species cover the entire virus genome in both orientations without gaps (Fig 3.2.4.1). This suggested that the complete virus genomes can be reconstructed from vsiRNAs *de novo*, i.e. without reference genome sequences.

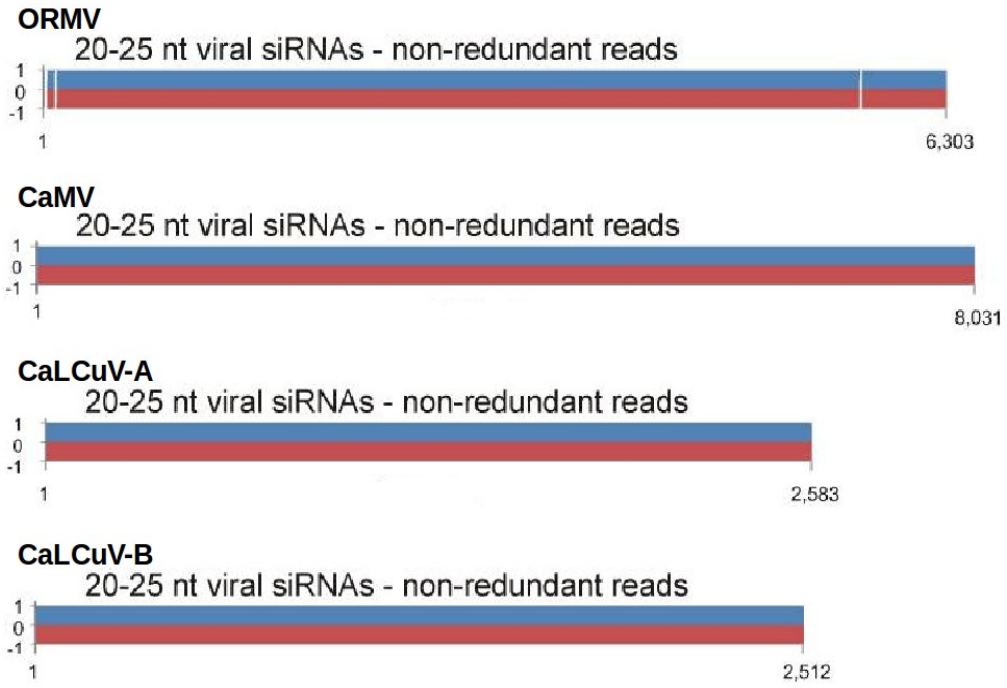


Figure 3.2.4.1 : Mapping of non-redundant reads along ORMV, CaMV and CaLCuV genomes.

Copied from (Seguin et al., 2014a).

3.3 De novo reconstruction of viral genomes from siRNAs

3.3.1 Strategies of virus genome reconstruction from short reads

Different strategies were developed and tested to reconstruct *de novo* the genomes of known viruses ORMV, CaMV and CaLCuV and then of unknown viruses (Seguin et al., 2014a). In all strategies, we used Velvet followed by Oases or Metavelvet to generate vsiRNA contigs and then Seqman to merge the contigs. The success of these strategies depended on the availability of the host genome sequence. In a proof-of-concept experiment, we used virus-infected *Arabidopsis thaliana* plants for which the complete genome is available since 2000 (The Arabidopsis Initiative, 2000).

The first strategy starts with the mapping of a short RNA library against the host genome. Then, the unmapped sRNAs, which contain mostly vsiRNAs, are used for *de novo* assembly of the virus genome. With this filtering step, the endogenous plant sRNA produced from microRNA, tasiRNA genes as well as heterochromatic (hc)siRNA from repetitive DNA such as transposons are

removed from the library. However, the removal of these reads can result in gaps within the contigs generated by Velvet/Oases or Metavelvet, because some of vsiRNAs may share identity with the host genome.

To verify the presence of shared read, a mapping was performed between the viral and the host genomes in order to determine the quantity of common sRNA. Then, we compared the quantity of common sRNA using the total reads (20 to 25-nucleotides long) and the non-redundant reads. For CaMV, ORMV and CaLCuV viruses, we observed a significant percentage of reads that could be mapped on both the virus and the plant genome. These percentages varied between ~0,5 % to ~2% for non-redundant reads, and between ~1,5% to ~8% for redundant reads (Fig 3.3.1.1). These common reads may correspond to vsiRNAs which can modify the infectivity by targeting host genes (Hanley-Bowdoin et al., 2013; Shimura et al., 2011).

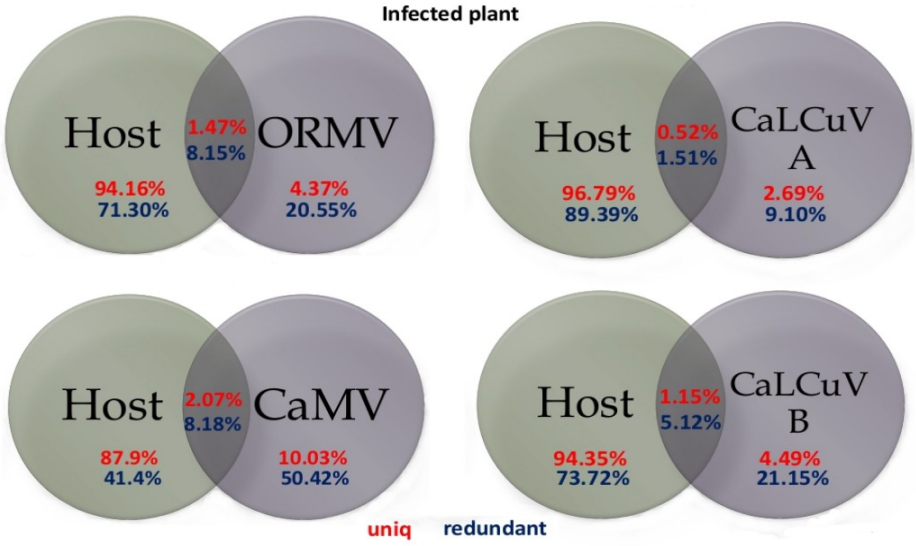


Figure 3.3.1.1: Percentage of mapped reads against viral and host genome.

Comparison of mapped reads between the host and the viral genome. The red percentage corresponds to the percentage of mapped non-redundant reads (uniq), and the redundant reads are represented in blue.

To reconstruct completely the viral genome, the shared reads must be kept. One of the possible solutions for this problem is to apply a filtering step after Velvet/Oases or Metavelvet. This ensures that long viral contigs that share only partial homology to the host genome are not removed. Consequently, the second strategy has two steps for the *de novo* assembly: the contigs generated by Velvet/Oases or Metavelvet are mapped against the host genome during the BWA filtering, before the final *de novo* assembly using Seqman (Fig 3.3.1.2). If the second strategy is not sufficient for reconstruction of the entire virus genome, a third Strategy can be used, in which the contigs created only with Velvet are taken like references for the filtering step. Then, the small RNAs that do not match the Velvet contigs are used for a new round of *de novo* assembly with the three assemblers.

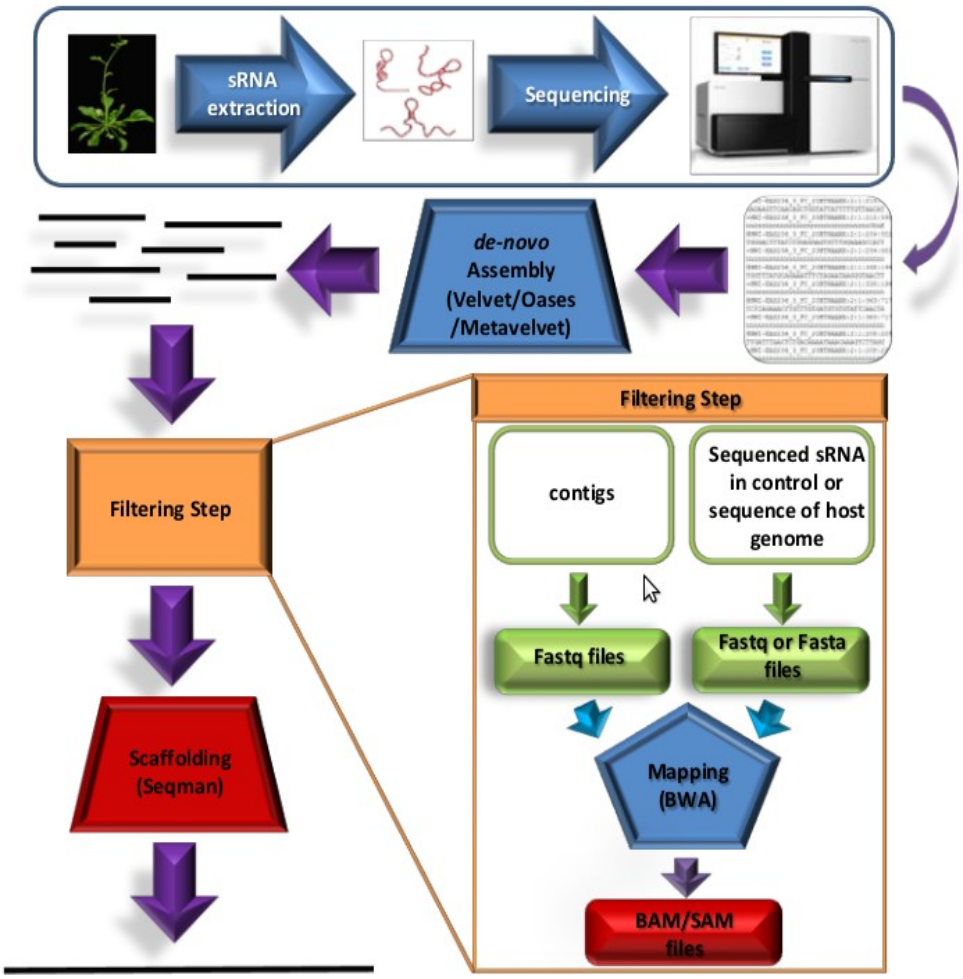


Figure 3.3.1.2: Summary of the second strategy.

After extraction of the sRNA from infected plant and their sequencing, the reads are used for a first *de novo* assembly with Velvet and, Oases or Metavelvet. The assembled contigs are filtered by a mapping step against the host genome or the control sample (if the sequence of genome is not available). The unmapped contigs are extracted from Bam/Sam files to be assembled by the scaffolding step with Seqman.

To eliminate potential sequencing errors, the non-redundant (unique) reads can be used. In fact, reads can contain errors introduced at different (enzymatic and chemical) steps during preparation of cDNA libraries and sequencing. Mapping of non-redundant reads of the *de novo* assembled virus genome allows for the identification and correction of the errors (by SNP/mismatch and INDEL call). Taking this into account, six strategies were used (table 3.3.1.3).

table 3.3.1.3 : Strategies for *de novo* reconstruction of a virus genome.

Strategy number	redundant reads	Filtering step
1	no	before velvet
2	yes	before velvet
3	no	after oases
4	yes	after oases
5	no	after velvet
6	yes	after velvet

For each sample, a selection of contigs created by Velvet/Oases according different *k*-mers can be done according to the N50 value and the maximal contig length. N50 corresponds to the contig size which includes 50% of all concatenated contigs. This concatenation of contigs is performed by a decrease sorting. According to these criteria, this first selection allows to select one or two fasta files containing contigs sequences generated by one or two different *k*-mers. A second selection can be done by the addition of previous and following *k*-mers or by the addition of intermediate *k*-mers if it is possible. In total, three different selections are done before the final assembly: one according to the N50 value and the maximal length, one by extension of the first selection and the last one with all contigs.

3.3.2 Reconstruction of the RNA tomabovirus genome (ORMV) from viral siRNAs and analysis of viral siRNAs

The datasets BPO-35 to BPO-46 represent sRNA libraries from ORMV-infected and mock-inoculated *Arabidopsis thaliana* wild-type (Col-0) and silencing-deficient mutant (*dcl2/3/4* and *rdr1/2/6*) plants with ORMV at 14 and 22 dpi (see the above mapping result with ORMV).

Our above-described analysis revealed that the majority of plant sRNAs in mock-inoculated and virus-infected plants belong to 21-22-nt and 24-nt classes. These sizes correspond to the production of plant miRNAs (21-22-nt) and heterochromatic siRNAs (24-nt).

The mapping with BWA was performed to evaluate the coverage of sRNAs along the viral genome of ORMV. In the case of virus-infected plants, the entire ORMV genome sequence was found to be covered by the sRNA species without gaps in sense or antisense polarity with or without redundant reads (Fig 3.3.2.1). The highest numbers of vsiRNA reads and the best genome coverage were obtained for ORMV-infected wild type (Col-0) plants. In the case of ORMV-infected *rdr-1/2/6* plants the ORMV genome coverage was similar to that of Col-0 but the number of 24-nt endogenous sRNAs was lower.

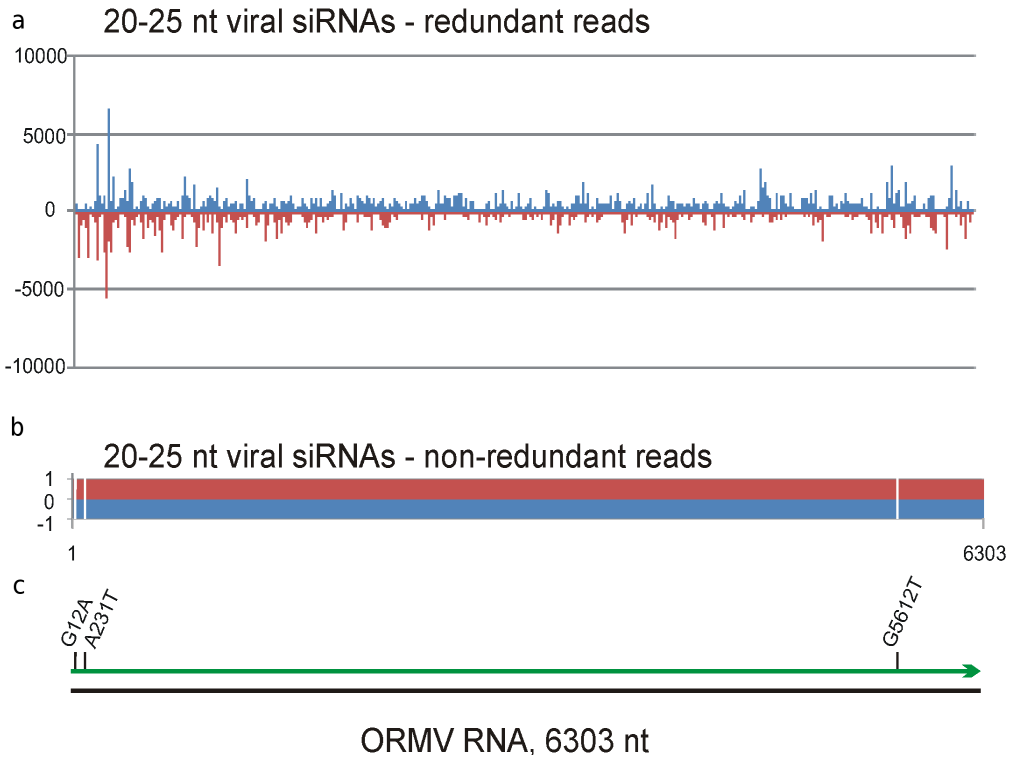


Figure 3.3.2.1: coverage of mapped redundant and non-redundant reads along the ORMV genome.

This figure represents the coverage of mapped vsiRNA with redundant (a) and non-redundant reads (b). The figure c represents the ORMV genome with the detected SNPs during the mapping. Copied from (Seguin et al., 2014a).

The datasets BPO-38 and BPO-44 that correspond to ORMV-infected wild-type plants at 14 and 22 dpi were taken for *de novo* reconstruction of the ORMV genome from vsiRNAs.

To test all the above described strategies, the *de novo* assembly was first performed without the filtering step. In all the cases, following Velvet/Oases the contigs were chosen for the filtering step according to the N50 and the contig maximum length: the best values were achieved using the *k*-mers 17 and 19 (table 3.3.2.2). The contigs obtained with these *k*-mers were then filtered through the *Arabidopsis thaliana* genome (TAIR9) and then Seqman was used to finalize *de novo* assembly. To complete the analysis, we also tested the extension of the selected contigs adding the previous and the next *k*-mers (15 and 21, respectively) and the full complement of contigs using all the possible *k*-mers from 13 to 21.

table 3.3.2.2: Summary of statistical values for contigs created by Velvet/Oases with Strategies 3 and 4.

BPO-38 and BPO-44 are datasets of reads from wild-type plant at 14 day or 22 days after infection with ORMV respectively. N50 and the maximum length of contigs are indicated in red because the *k*-mers are selected according to these values.

	BPO-38						BPO-44						
K-mers	**13**	**15**	**17**	**19**	**21**	**23**	**13**	**15**	**17**	**19**	**21**	**23**	
Total length raw	7869	144357	95759	30839	10550	1045	4551	149715	103080	34087	11021	228	**Strategy 3**
Sum of contig length	7869	144357	95759	30839	10550	1045	4551	149715	103080	34087	11021	228	
Number of contigs	60	908	518	147	59	9	36	967	524	146	63	2	
Undetermined Base	0	8	1	0	0	0	2	22	0	0	0	0	
Breaks	60	910	519	147	59	9	37	976	524	146	63	2	
Ambig Bases	0	0	0	0	0	0	0	0	0	0	0	0	
N50 of contigs	127	147	161	**243**	190	113	122	145	170	**268**	181	126	
Contigs for N50	23	286	118	25	18	4	15	329	104	19	20	0	
Average length of contigs	131	158	184	209	178	116	126	154	196	233	174	114	
Contig maximum length	289	2710	**3379**	3055	480	135	188	1581	**3326**	2869	616	126	
K-mers	**13**	**15**	**17**	**19**	**21**	**23**	**13**	**15**	**17**	**19**	**21**	**23**	
Total length raw	12682	180295	97102	31622	9923	1153	2635	186224	113804	37131	10016	368	**Strategy 4**
Sum of contig length	12682	180295	97102	31622	9923	1153	2635	186224	113804	37131	10016	368	
Number of contigs	101	933	501	148	58	10	21	1061	536	141	58	3	
Undetermined Base	9	17	3	0	0	0	0	20	0	0	0	0	
Breaks	104	937	502	148	58	10	21	1069	536	141	58	3	
Ambig Bases	0	0	0	0	0	0	0	0	0	0	0	0	
N50 of contigs	122	178	174	**451**	172	114	128	161	180	**451**	181	126	
Contigs for N50	43	214	104	26	18	4	9	296	84	15	18	1	
Average length of contigs	125	193	193	213	171	115	125	175	212	263	172	122	
Contig maximum length	208	2906	**5577**	3536	453	135	163	2999	**5577**	3536	634	140	

The results obtained without the filtering step show that a large number of contigs could be generated with Seqman using either redundant or non-redundant reads. When we used only *k*-mers 17 and 19, around 120 contigs were obtained. Nevertheless, the longest contigs had a length close to the viral genome size (6303-nt), but around 300 nucleotides of the ORMV genome were not included in these contigs. When the *k*-mers 15 and 21 were added, we obtained more than 300 contigs in each case. With the non-redundant reads, the size of the longest contig could be increased to 7 kb in the case of BPO-44. This increased size is due to the incorporation of polyA to the end and also other non-viral genomic sequences.

For the contigs obtained using the filtering step, we observed similar results. However, the maximum contig length was bigger for the viral genome with redundant reads for the extended *k*-mers selection, while it was not the case only with BPO-38 for non-redundant reads.

For the two last strategies, two contigs were only obtained for BPO-44 with the strategy 3 and the selection of *k*-mers 19 and 21, while one contig was observed for all the other parameters. Moreover, 7 contigs were longer than the viral genome, with, the longest contig having an

extension of 63 nucleotides at 3'-terminus (BPO-38, the strategy 4 and all contigs; Table 3.3.2.3).

table 3.3.2.3: Summary of contigs created with Seqman according to the selected k-mers and strategies for BPO-38 and BPO-44.

	BPO-38		BPO-44		BPO-38		BPO-44	
K-mers	size of longest contig	number of contigs	size of longest contig	number of contigs	size of longest contig	number of contigs	size of longest contig	number of contigs
	without filter steps							
17,19	6388	121	5594	125	6292	115	6273	126
15, 17, 19, 21	6396	346	7068	363	6294	323	6286	371
all	6396	348	7068	367	6295	324	6286	373
	Strategy 1				Strategy 2			
17,19	6310	177	5592	121	6302	123	6286	117
15, 17, 19, 21	6317	335	6783	359	6602	325	7393	366
all	6317	339	6783	362	6602	334	7392	370
	Strategy 3				Strategy 4			
17,19	6310	1	5597	2	6295	1	6281	1
15, 17, 19, 21	6321	1	6294	1	6325	1	6312	1
all	6321	1	6294	1	6366	1	6312	1

For multiple alignment analysis, we took the contigs obtained using the strategies 3 and 4, because it was difficult to identify the ORMV contigs among the contigs obtained using the other strategies. Even if the longest viral contig had a size close to the ORMV genome, there were different contigs longer than 1 kb which could be viral according to their size.

By multiple alignment of contigs with the viral genome, the best alignment is observed with the contigs created using the *k*-mers 17 and 19 for BPO-38. Moreover, there were lower numbers of indels or SNPs using the strategy 3 in comparison to the strategy 4. This difference can be attributed to the errors introduced during preparation of cDNA libraries and deep-sequencing as described above. In fact, with the strategy 3 using redundant sequences, much more abundant reads for a given small RNA species will be used for *de novo* assembly, and consequently, there will be a much lower probability of incorporation into the growing contig of the error-containing reads corresponding to the given sRNA species which are present in much smaller numbers. We also observed more indels and SNPs in BPO-44 compared to the BPO-38. This is likely due to the natural appearance of ORMV genome variants (containing SNPs and indels) during the course of infection, especially at the late stage (i.e.. at 22 dpi vs. 14 dpi). The increased number of SNPs and indels in particular regions of the viral genome can be used as a viral strategy to mutate itself in order to adapt to new host and/or conditions. We also noted an addition of oligoT-rich stretches at both ends of ORMV contigs in comparison to the reference genome. This might represent

oligourydilation of vsiRNA species which are not protected by HEN1-mediated methylation which was previously observed in ORMV-infected *N. benthamiana* and Arabidopsis plants as ORMV infection blocks HEN1 activity (Akbergenov et al., 2006; Blevins et al., 2006).

In conclusion, for ORMV (and perhaps RNA tobamoviruses) the strategies 3 and 4 appear to be the best strategies for a *de novo* reconstruction of the viral genome without the use of reference sequence In the case of ORMV, a single contig was created with selected *k*-mers of redundant and non-redundant reads; this contig corresponded to the complete RNA genome extended by oligoA.

3.3.3 Reconstruction of the pararetrovirus genome (CaMV) from viral siRNAs

Four samples from CaMV-infected plants were taken (BPO-20, wild type line Col-0; BPO-21, *ago2* mutant line; and BPO-22, *ago3* mutant line) and one: (BPO-13) from the mock-inoculated Col-0 plants.

Bioinformatics analysis of sRNA populations revealed that, the majority of sRNAs are 24-nt long, with 21-nt sRNAs being the second most abundant, followed by 22-nt sRNAs. Consequently, the sRNAs ranging in size between 20 to 25 nucleotides were selected for *de novo* assembly.

BWA mapping showed that the sRNA coverage of the CaMV genome is not uniform for the three samples of CaMV infected plant. There was no dramatic difference between the CaMV-infected samples (see the above mapping result with CaMV). Consequently, these three samples were studied for the final *de novo* assembly with Seqman.

The strategies 3 and 4 were used for CaMV-infected samples. According to the values of N50 and the length of longest contig, different *k*-mers can be selected for different samples. With the strategy 3, the 17-mers contigs were selected for BPO-20, the 15-mers and 19-mers had the best values for BPO-21 and only 21-mers for BPO-22 (« chosen » row in Table 3.3.2.1). To extend the selected contigs, the 15-mers, 17-mers and 19-mers were included for BPO-20 and BPO-21. For BPO-22, 19-mers and 23-mers were included too («extended *k*-mers » row in the table). With the strategy 4, the 15-mers is sufficient for BPO-20, the 15-mers and the 17-mers are selected for BPO-21 and, 19-mers and 21-mers for BPO-22 (« chosen » row in Table 3.3.2.1). 13-mers and 15-mers were used to extend the contigs selection for BPO-20. 13-mers and 19-mers could extend the BPO-21 selection. 17-mers and 23-mers were also selected to extend the BPO-22 selection (« extended *k*-mers » row in the table) (Table 3.3.3.1).

table 3.3.3.1: Summary of statistical values for contigs created by Velvet/Oases with Strategies 3 and 4.

BPO-20 to BPO-22 are dataset of reads from, respectively, wild-type, *ago2* mutant and *ago3* mutant, plants infected with CaMV. N50 and the maximum length of contigs are indicated in red because the *k*-mers are selected according to these values.

	BPO-20						BPO-21						BPO-22						
K-mers	**13**	**15**	**17**	**19**	**21**	**23**	**13**	**15**	**17**	**19**	**21**	**23**	**13**	**15**	**17**	**19**	**21**	**23**	
Total length raw	8778	51966	39287	17523	7982	394	13863	56124	40164	24506	9626	724	6697	134649	94905	58435	15448	6100	
Sum of contig length	8778	51966	39287	17523	7982	394	13863	56124	40164	24506	9626	724	6697	134649	94905	58435	15448	6100	
Number of contigs	68	270	126	65	45	1	98	298	162	76	50	3	56	843	460	144	45	30	
Undetermined Base	14	5	0	0	0	0	0	12	0	0	0	0	5	10	0	0	0	0	
Breaks	72	271	126	65	45	1	98	301	162	76	50	3	57	847	460	144	45	30	
Ambig Bases	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
N50 of contigs	125	206	**409**	304	168	394	132	173	304	**607**	202	500	116	154	216	1518	**8039**	208	
Contigs for N50	27	72	20	13	12	0	37	79	35	10	12	0	24	278	96	10	0	6	
Average length of contigs	129	192	311	269	177	394	141	188	247	322	192	241	119	159	206	405	343	203	Strategy 3
Contig maximum length	207	1284	**1717**	1396	604	394	334	**2231**	1206	1784	1013	500	179	2159	2059	5302	**8039**	1162	
K-mers	**13**	**15**	**17**	**19**	**21**	**23**	**13**	**15**	**17**	**19**	**21**	**23**	**13**	**15**	**17**	**19**	**21**	**23**	
Total length raw	41975	79934	32700	15961	8133	745	44003	92652	54967	17565	9331	828	24648	166604	107912	43532	15864	5829	
Sum of contig length	41975	79934	32700	15961	8133	745	44003	92652	54967	17565	9331	828	24648	166604	107912	43532	15864	5829	
Number of contigs	233	244	104	62	46	3	257	287	125	66	47	4	178	893	437	133	47	26	
Undetermined Base	14	0	0	0	0	0	0	6	0	0	0	0	7	10	3	0	0	0	
Breaks	240	244	104	62	46	3	257	289	125	66	47	4	179	897	438	133	47	26	
Ambig Bases	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
N50 of contigs	181	**643**	579	287	152	504	175	598	**1206**	317	219	504	134	174	297	1002	**8039**	220	
Contigs for N50	67	23	12	14	13	0	81	15	10	10	11	0	69	223	53	9	0	6	
Average length of contigs	180	327	314	257	176	248	171	322	439	266	198	207	138	186	246	327	337	224	Strategy 4
Contig maximum length	527	**5510**	3439	1490	650	504	685	**7083**	2984	1784	1016	504	335	2894	4767	**8069**	8039	1143	

We observed that the sRNAs from CaMV-infected ago2 mutant plant (BPO-21) could not be assembled in a complete CaMV genome (8031-nt) as a single Seqman contig. The minimal number of CaMV contigs was 12 for the first strategy and 22 for the second strategy. One Seqman contig covering the entire CaMV genome was generated with the strategy 3 for the wild-type and *ago3* mutant plants, and also with the strategy 4 for the *ago3* mutant plants. The common approach for all these samples was to use the extended selected *k*-mers. Otherwise, the first chosen *k*-mers for *ago3* mutant plant was sufficient to assemble the complete genome in a single contig, but not with all *k*-mers. We observed the opposite with the wild-type infected plant sample (BPO-20). The size of these contigs was from 8039 to 10754 nucleotides (Table 3.3.3.2).

table 3.3.3.2: Summary of contigs created with Seqman according to the selected *k*-mers and strategies for BPO-20 (wild type), BPO-21 (*Ago2* mutant) and BPO-22 (*Ago3* mutant).

	Strategy 3						Strategy 4					
	BPO-20		BPO-21		BPO-22		BPO-20		BPO-21		BPO-22	
K-mers	**size of longest contig**	**number of contigs**	**size of longest contig**	**number of contigs**	**size of longest contig**	**number of contigs**	**size of longest contig**	**number of contigs**	**size of longest contig**	**number of contigs**	**size of longest contig**	**number of contigs**
chosen	2958	8	3992	12	8039	1	6174	3	8438	22	10730	1
different k-mers	8075	1	7454	23	9341	1	6174	3	8332	32	10748	1
all	8078	1	4079	26	9339	2	7161	5	8083	34	10754	3

Multiple alignment of these contigs and the CaMV reference genome was performed with ClustalW2 or Blast. All BPO-20 contigs created with the strategy 3, showed precise match with the CaMV reference genome with only a few gaps or misalignment. The BPO-22 contigs created with only the 21-mers had a sequence matching to the reference genome. However, the blast alignment of the contig created with 19, 21 and 23-mers showed a repeated region that is why the contig is longer than the reference genome (1310-nt longer). The same phenomenon was observed with the contigs created with BPO-22 and the strategy 4, where the repeated region has a length superior to 3000 nucleotides (Fig 3.3.3.3). This is obviously due to a circular nature of the CaMV genome.

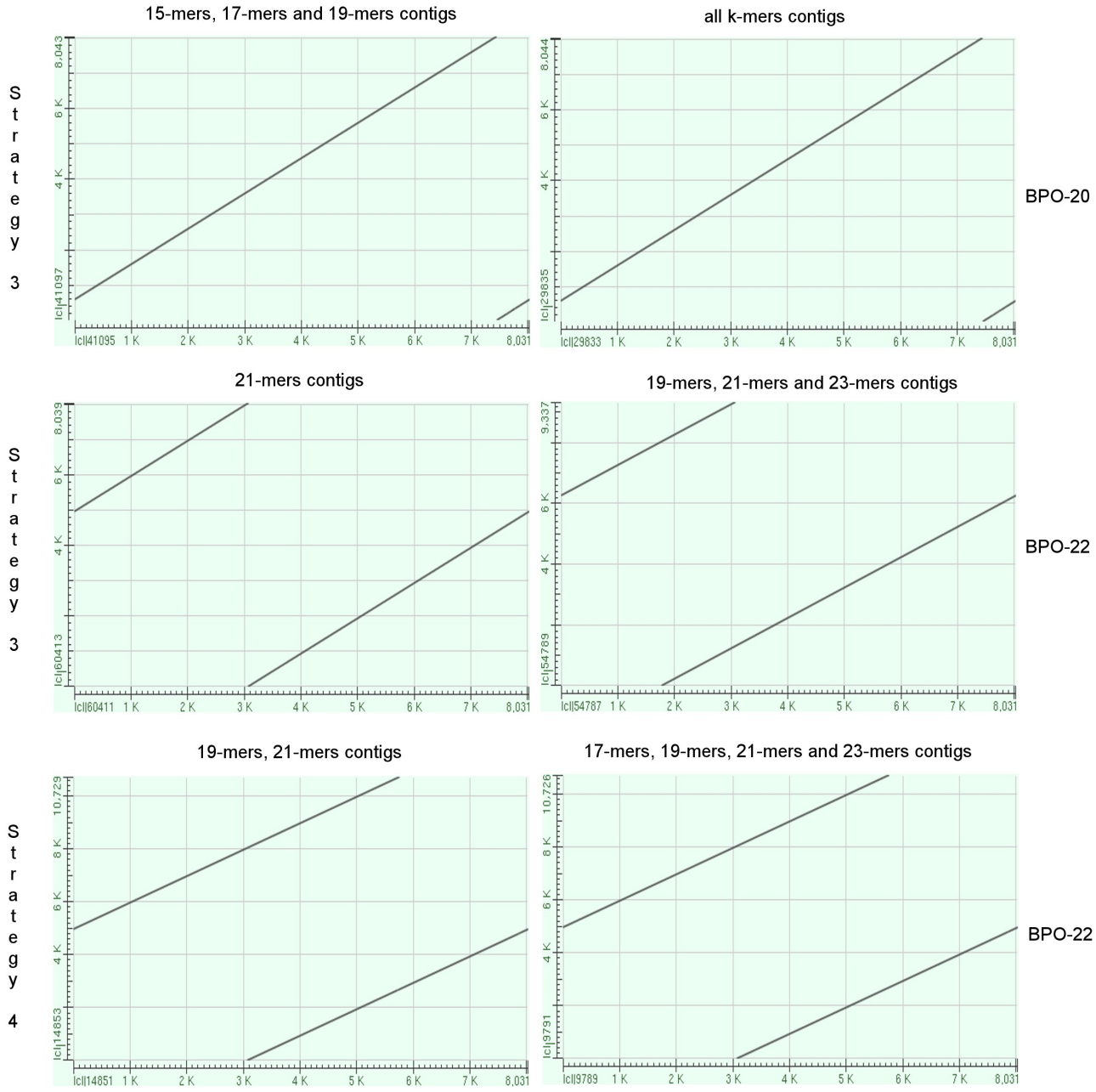


Figure 3.3.3.3: Result of blast alignments of Seqman contigs against the CaMV genome.

The x-axis represents the CaMV genome, the y-axis represents the contigs created with Seqman, according to the *k*-mers selection. BPO-20 contains small RNAs from a wild-type infected plant. BPO-22 contains small RNAs from *Ago3* mutant plant.

These results show that the use of non-redundant reads is likely preferable for the reconstruction of a circular genome of viruses because we finally obtain only one contig which covered the entire viral genome with more selections of non-redundant reads than with the redundant reads.

3.3.4 Reconstruction of the geminivirus genome (CaLCuV) from viral siRNAs

Three sRNA datasets were used for the CaLCuV study. BPO-56 is from Col-0 mock-inoculated plants. BPO-57 and BPO-58 are from CaLCuV-infected Col-0 (wild type) and *rdr1/2/6* mutant plants, respectively. The analysis of sRNA populations showed that the majority of viral sRNAs belong to 21-nt class, followed by 24-nt class. 20-25-nt sRNAs were used for the de novo assembly.

We started the reconstruction with the strategies 1 and 3 and found that an big number of sRNAs map to both the Arabidopsis genome, and the CaLCuV genome. So the filter step after the first *de novo* assembly steps was required in order to keep all the vsiRNAs.

Using strategy-3 with redundant reads, the 17-mers and 21-mers had the best values for the N50 and the longest contig for the CaLCuV-infected Col-0 plant (BPO-57). With the CaLCuV-infected *rdr1/2/6* mutant plant (BPO-58), the 15-mers and the 17-mers were selected. Before assembly with all *k*-mers, we extended the selection by adding 19-mers Velvet/Oases contigs for BPO-57 and, 13-mers and 21-mers for BPO-58. After Seqman for CaLCuV- infected Col-0 plant, we always observed the smallest number of contigs when we took all Velvet/Oases contigs. Although the longest contig had the closest size to the reference viral genome, three contigs were created with the final assembly, one for DNA-B and two for DNA-A (Fig 3.3.4.1).

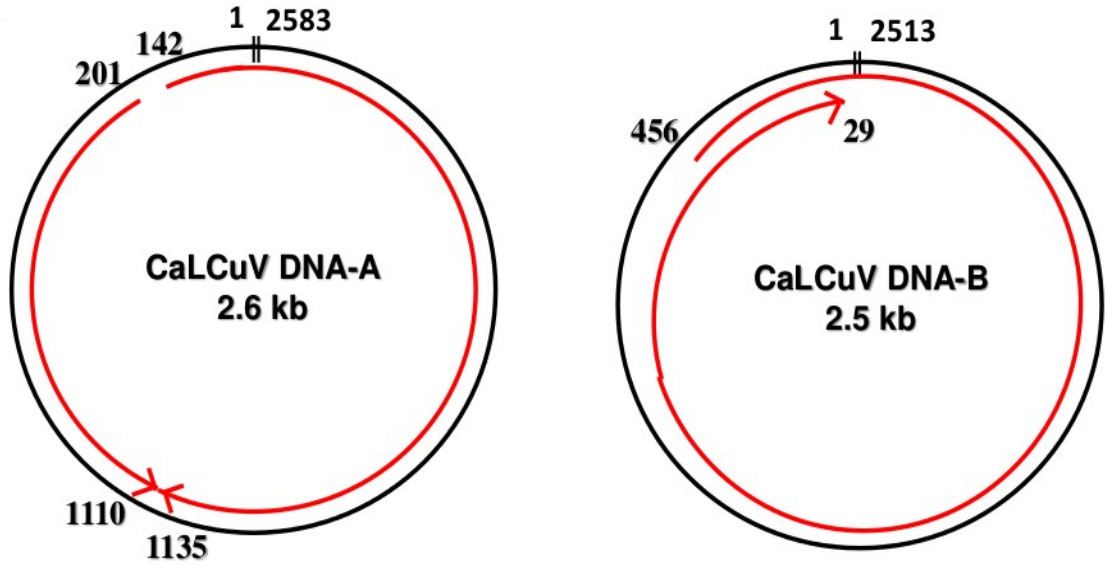


Figure 3.3.4.1: Contigs created by the Strategy 3

To obtain single contig for DNA-A, we combined the two libraries, but this did not improve the result. We also used all the values between 13 to 23 for *k*-mers and many combinations thereof. However, DNA-A was still assembled in two non-overlapping contigs, leaving two small gaps, one in the common region (shared with DNA-B) and another in the poly(A) site region (will the lowest number of reads). Using IGV with the DNA-A, the two contigs could be extended by the overlapping sRNAs covering the two gap regions and thereby the complete DNA-A could be reconstructed as a single contig.

To complete and improve these results, we used the strategies 5 and 6, in which one more step was added to the released host contigs. For this, a *de novo* assembly was performed only with velvet before the filter step. Then the contigs that were not mapped on the host genome were used to sort the sRNAs which map only to these contigs. Finally, the selected sRNAs were assembled with velvet, oases and Seqman. Like previously, this strategy was performed both with the non-redundant and redundant reads. The result of this *de novo* assembly shows that the strategies 5 and 6 do not improve the results obtained with strategies 3 and 4. The implementation of the strategies 5 and 6 is the most complicated that is why the strategies 3 and 4 seems to be the best strategies for our *de novo* assembly method.

To conclude with *de novo* assembly of known viruses, the complete genomes of ORMV and CaMV were assembled as single contigs of vsiRNAs. With CaLCuV, such de novo assembly without a reference genome gave the complete single-contig for DNA-B but two incomplete contigs for DNA-A. We assume that because the two DNAs share a near identical common region of 195-nts (with 7 SNPs), during *de novo* assembly the DNA-A siRNA contig gets split in two contigs within this region. In addition the low coverage of the DNA-A poly(A) site resulted in a gap. Both problems could be solved using the DNA-A reference sequence as a guide to extend the contigs by overlapping siRNAs. Thus, the use of existing bioinformatics tools with additional steps allowed for the reconstruction of complete viral genome only with sRNAs extracted from virus-infected plants. For the assembler Velvet, different *k*-mers can be used for the assembly, but all *k*-mers are not necessary to obtain the complete viral genome. Statistical value like N50 or the maximal length of created contig during the first *de novo* assembly step seems to be necessary to define the contigs which will be use for the last *de novo* assembly step with Seqman. Nevertheless, the selected *k*-mers are not always sufficient to obtain the viral genome. Consequently, it is important to extend the selection of *k*-mers to help the *de novo* assembly with Seqman.

A key finding of our study is that a homogeneous coverage of the viral genome with sRNA with redundant reads does not appear to be important for a success of complete genome reconstruction. For example, we obtained the complete genomes for both CaMV and ORMV, although the coverage of CaMV genome is not homogeneous. This finding is in line with the fact

that we obtained the best reconstruction results with non-redundant reads. Furthermore, our study shows that the origins of sRNAs in virus-infected plants can not be a priori separated between the viral genome and the host genome and a fraction of sRNAs can be shared between the two genomes. Consequently, the filter steps are necessary for complete reconstruction of the viral genome. These common sRNAs can be due to the presence of host genome-integrated viral sequences with a homology to the viral genome. These integrated sequences are expected to be silenced by the host siRNA-generating machinery and the siRNAs derived from these sequences are expected to be assembled into contigs. Therefore, filtering all the contigs generated by Velvet/Oases or Metavelvet should help to omit the interacted virus contigs from non-integrated viral contigs.

Although this study was performed with Velvet/Oases, Metavelvet and Seqman, it is advisable to use other bioinformatics tools in case of incomplete reconstruction of a viral genome. We named a “siRNA omics” (siRomics) approach the use of mapping and the *de* novo assembly strategies 3 and 4 with an additional step using IGV to reconstruct a consensus master genome in viral 'quasispecies cloud' (Seguin et al., 2014a).

3.3.5 Analysis of the viral quasispecies

Using siRomics we reconstructed the consensus master genomes of ORMV, CaMV and CaLCuV from infected *Arabidopsis thaliana,* which were supported by the majority of the sRNA reads at each nucleotide position. In the case of ORMV, the reconstructed master genome differed from the Genbank sequence and from the sequence of the available ORMV full-length clone which did not exhibit wild type infectivity. Correction of the ORMV clone to replace the mutated positions (cloning errors) and thereby create the consensus master genome clone restored the wild type infectivity. Additionally, the SNP analysis showed that 6 positions of the ORMV genome sequence have SNPs, but this nucleotide variation was supported by less than 20% of redundant sRNA and therefore represented the natural evolution of the ORMV quasispecies. For CaMV and CaLCuV DNA viruses, a few SNPs were also observed in their master genome: 5 SNPs in the CaMV master genome supported by less than 20% of redundant reads. 12 SNPs were found in the CaLCuV DNA-A and 5 SNPs in DNA-B. However, these two DNA components have the near identical common sequences of ~ 200 nucleotides, which were almost equally represented in the vsiRNA population. After removal of the respective SNPs in the common region, only 2 SNPs and 1 SNPs remained in the DNA-A and DNA-B, respectively, which were supported by a small proportion of sRNA reads (for details of SNP analysis, see Seguin et al. 2014a).

3.3.6 Reconstruction of a DNA virus and two viroids associated with emerging red blotch disease of grapevine

Over the last couple of years, grapevine plants in the vineyards in Oregon State, USA were damaged with a severe disease associated red leaf blotch symptoms, which did not resemble the symptoms of known grapevine viruses. In collaboration with Prof. Valerian Dolja from Oregon State University, we used the siRomics approach to reconstruct the genome of this unknown virus.

Total RNA were extracted from different samples of leaves taken from virus-infected and healthy grapevines plants in Oregon. Small RNA were sequenced using Illumina technology at Fasteris SA. The strategies 3 and 4 were used to generate the contigs. For the filtering step, we used the sequence of *Vitis vinifera* genome from NCBI Genome bank (PRJNA33471, (Jaillon et al., 2007)).

After the filtering step and the scaffolding step with Seqman, many contigs were kept : more than 20 contigs were generated from metavelvet contigs and more than 40 contigs were generated from oases contigs. Even if the average length of contigs was similar between metavelvet and oases, the maximum length of the contigs was twice longer for oases with redundant reads (Table 3.3.6.1). To detect the viruses present within the samples, we blasted the Seqman contigs against one plant viral database of DNA sequences generated thanks to the fasta files provided by the NCBI viral database (http://www.ncbi.nlm.nih.gov/genomes/VIRUSES/viruses.html, (Bao et al., 2004)). The blast results showed that the highest scores and total length are obtained for 5 different viruses and viroids: the *Grapevine red blotch-associated virus* isolate CF214-1 (GRBaV), the *Grapevine geminivirus* (GVGV) (Krenz et al., 2012), the *Grapevine yellow speckle viroid 1* (GYSVd-1), the *Grapevine yellow speckle viroid 2* (GYSVd-2) and the *Hop stunt viroid* (HSVd) (Table 3.3.6.2). The *Grapevine red blotch-associated virus* isolate CF214-1 (GRBaV) and the *Grapevine geminivirus* (GVGV) have the same genome which differs only by 9 nucleotides. The size of the contigs aligned against the viroids are longer than the viroids genome; these contigs show a concatenation of viroids genome likely due to the RCA amplifications. The grapevine geminiviruses (GVGV and GRBaV) were covered by 4 contigs when we merged the blasted contigs created by Metavelvet and Oases (Fig 3.3.6.3 a).

table 3.3.6.1: Summary of contigs created with Seqman according to the selected *k*-mers , the strategies and the software used before the filter step.

	Strategy 3					
K-mers	**all metavelvet**	**k-17-19-21-23 metavelvet**	**k-19-21 metavelvet**	**all oases**	**K-17-19-21 oases**	**k-19 oases**
Total length raw:	6178	3501	1502	10213	4590	1050
Sum of contig length:	6178	3501	1502	10213	4590	1050
Number of contigs:	26	13	5	41	16	3
Undetermined Base:	0	0	0	0	0	0
Breaks:	26	13	5	41	16	3
Ambig Bases:	28	13	14	72	2	12
N50 of contigs:	306	315	315	302	319	291
Contigs for N50:	8	4	2	12	5	1
Average length of contigs:	237	269	300	249	286	350
Contig maximum length:	**428**	428	319	**543**	495	498
	Strategy 4					
K-mers	**all metavelvet**	**k-17-19-21-23 metavelvet**	**k-19-21 metavelvet**	**all oases**	**k-15-17-19 oases**	**k-17 oases**
Total length raw:	6283	3991	1290	11246	11001	1144
Sum of contig length:	6283	3991	1290	11246	11001	1144
Number of contigs:	26	16	4	44	42	4
Undetermined Base:	0	0	0	0	0	0
Breaks:	26	16	4	44	42	4
Ambig Bases:	46	16	13	161	122	23
N50 of contigs:	306	306	315	324	327	324
Contigs for N50:	8	5	1	11	10	1
Average length of contigs:	241	249	322	255	261	286
Contig maximum length:	441	441	441	**889**	907	402

table 3.3.6.2: Summary of blasted contigs with highest score.

The contigs are generated by oases with different *k*-mers (13, 15, 17, 19, 21 and 23), merged and finally scaffolded by Seqman.

Identifiant	**score sum**	**e-value average**	**identity average**	**similarity average**	**total length**
Gi|529296147 – Grapevine red blotch-associated virus isolate CF214-1	2369	0	99,41745	99,125	2420
Gi|387600900 - Grapevine geminivirus	2353	0	99,0996	98,75	2420
Gi|20153376 - Grapevine yellow speckle viroid 2	289	1,43E-028	93,783394	93,5	385
Gi|11497495 - Hop stunt viroid	240	4,68E-018	96,09074	95,5	315
Gi|11496576 - Grapevine yellow speckle viroid 1	247	5,07E-015	90,72648	90,333336	427

To confirm the blast result, we performed mappings against GVGV, GYSVd-1, GYSVd-2 and HSVd. The mapping results validated the presence of the viroids in all the infected and healthy samples, while the GVGV virus was found only in infected sample with red leaves symptoms. All their genomes were covered entirely with non-redundant reads. In opposite, the coverage with redundant reads was heterogeneous (Fig 3.3.6.3). The analysis of SNP and In/Dels polymorphism with IGV allows to deduce that one of the viroids is close to the genome sequence of GYSVd-1 and the second corresponds to the HSVd genome. Moreover, the DNA virus had the genome which differs of the GVGV viral genome (Krenz et al., 2012) by only 11 nucleotides.

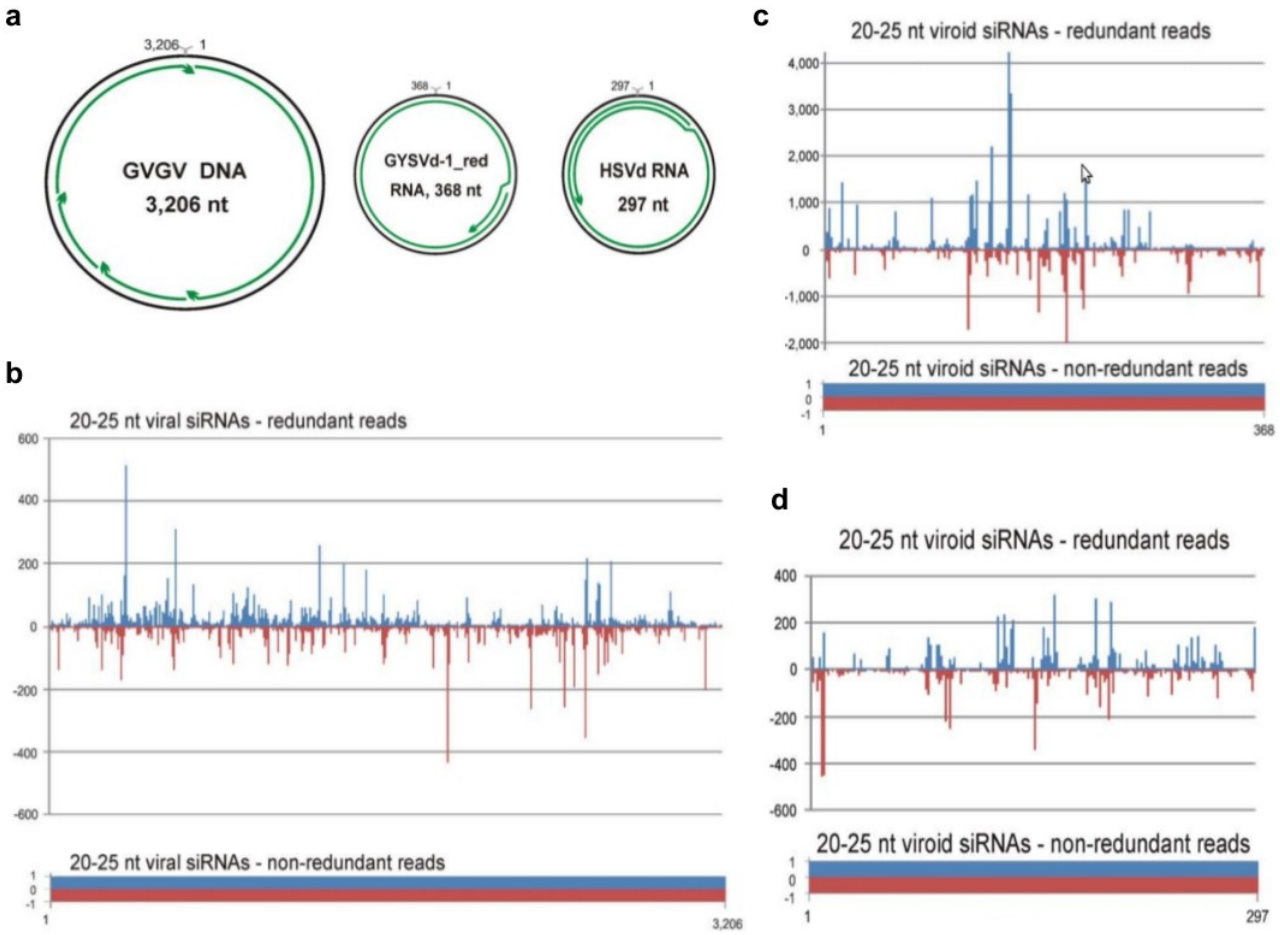


Figure 3.3.6.3: Blast and mapping results against GVGV, GYSVd-1 and HSVd.

(a) The figures represent the contigs aligned on the GVGV, GYSVd-1 and HSVd. (b) The figure represents the coverage of mapped redundant and non-redundant reads along the GVGV genome. (c) The figure represents the coverage of mapped redundant and non-redundant reads along the GYSVd-1 genome. (d) The figure represents the coverage of mapped redundant and non-redundant reads along the HSVd genome. Adapted from (Seguin et al., 2014a).

The majority of sRNA mapped to the GVGV genome are of 21-nt class, followed by 22 and 24-nt size-classes s (Fig 3.3.6.4). Similar size profile was observed for other geminiviruses too. The proportion of mapped sRNA along reverse and forward strands was comparable for all the size classes (Fig 3.3.6.5). This observation and other lines of evidence suggest that bi-directional readthrough transcription of circular viral DNA may generate dsRNA precursors (see Pooggin 2013).

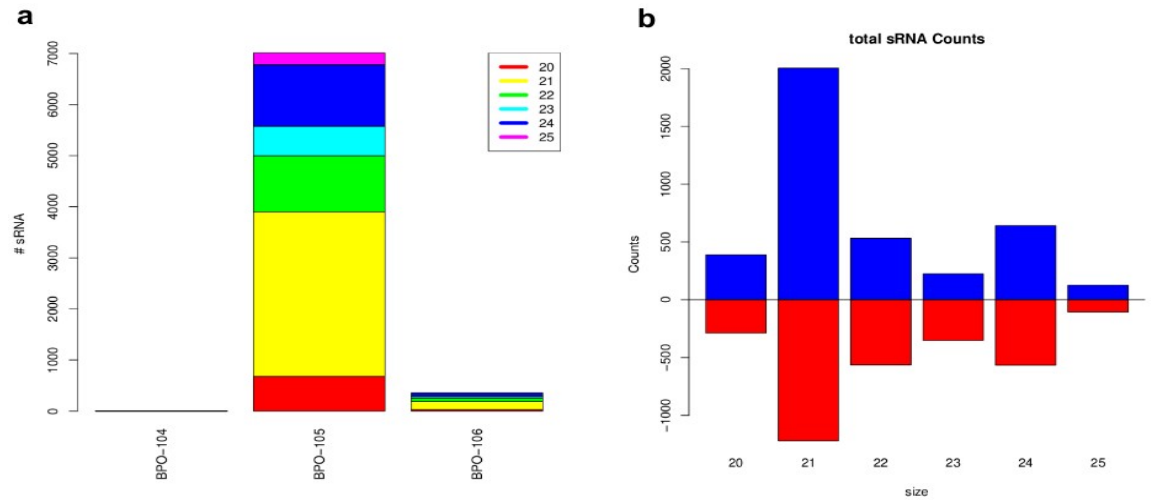


Figure 3.3.6.4: Distribution of mapped reads against GVGV genome.

(a) The histogram represents the count of mapped reads for BPO-104 (infected grapevine plant with green leaves), BPO-105 and BPO-106 samples (infected grapevine plant with red leaves). The y-axis represents the count of mapped reads. The bars have separated colours according to the size of mapped reads. (b) The histogram represents the count of mapped reads for BPO-105 according to the size-classes and the targeted strand.

5' nucleotide profiling revealed that the majority of 24-nt vsiRNAs have 5'A, suggesting their association with the AGO4 clade proteins. Other major size classes have less pronounced 5'-nucleotide biases. Taken together, multiple AGOs may bind GVGV-derived siRNAs.

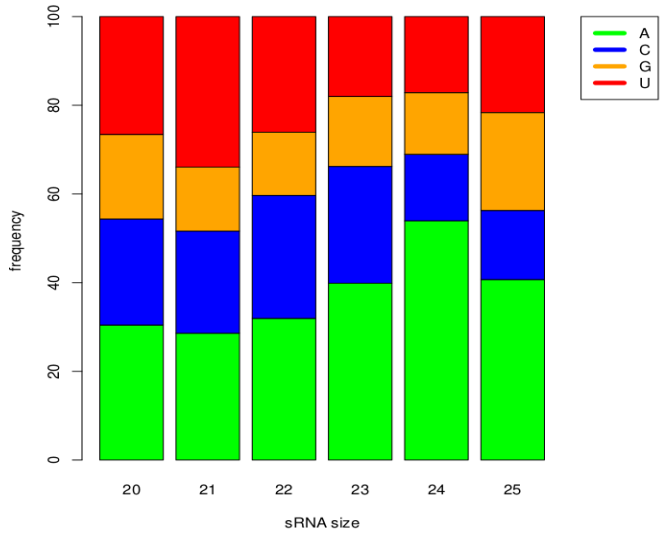


Figure 3.3.6.5: distribution of 5' nucleotide among reads for BPO-105 mapped along GVGV genome.

Each bar corresponds to one size classes of reads. The y axis indicates the percentage values found for the distribution.

To conclude, the siRNA omics approach can be used to detect unknown plant viruses and viroids. The use of metavelvet or oases depends on the nature of the plant and viral genomes. Even though Metavelvet seems to be the most adapted *de novo* assembler when the disease is

caused by a combination of viruses and/or viroids, Oases generates longer contigs. Moreover, the siRNA omics approach allows to investigate the plant antiviral defense mechanisms. Thus, the siRNA omics approach can be used as a universal diagnostics tool and a method to analyze the sRNA-based antiviral mechanisms. However, other methods such as small RNA blot hybridization are necessary to validate the results obtained by siRNA omics.

3.3.7 Reconstruction of consensus master genome and the infectious clone of Oilseed rape mosaic virus

As described above (section 3.2.1), the sRNA mapping results for ORMV revealed three single nucleotide mismatches compared to the sequence of genome in the NCBI Genbank as well as the corresponding ORMV cDNA clone. These mutant nucleotides can explain why the cDNA clone displayed dramatically reduced infectivity. Using the siRomics approach and two independent datasets of sRNAs from Arabidopsis plants infected with the wild type ORMV virions, we reconstructed the complete ORMV genome as a single contig. In both cases, the contig sequence had the three nucleotide mismatches, thus confirming the sRNA mapping results. Using synthetic gene fragments, we corrected the cDNA clone by incorporating the wild type nucleotides at the corresponding positions and tested the resulting clone for infectivity in *Nicotiana benthamiana* plants. Seven days post-inoculation, all plants inoculated by the corrected clone exhibited severe disease symptoms, similar to those on plants infected by the wild-type ORMV virions (Fig 3.3.7.1) (Seguin et al., 2014a). These results validated the quality of the de novo results obtained using the siRomics approach.

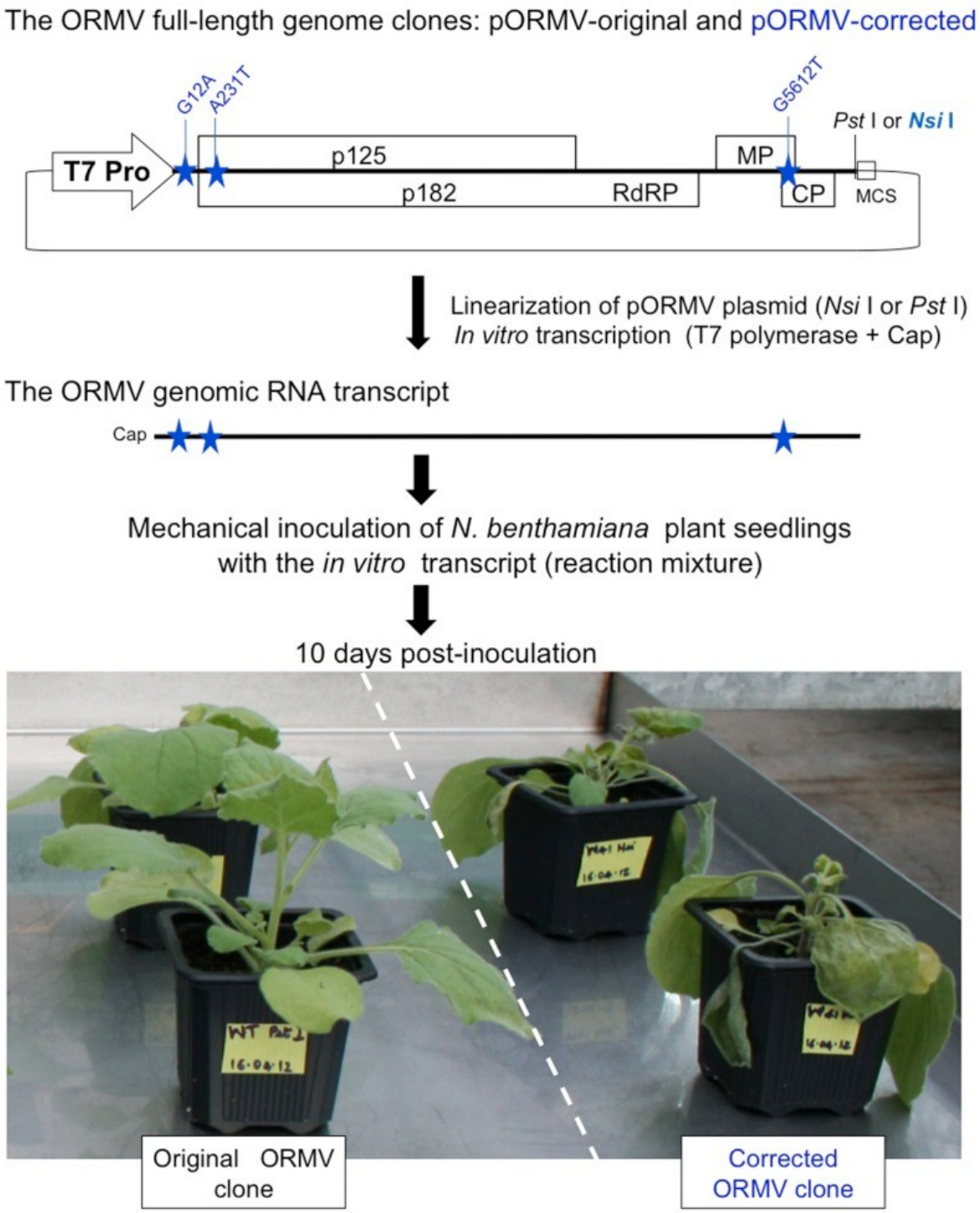


Figure 3.3.7.1: Test of the original and the corrected ORMV clones for infectivity.

The plasmid containing the full-length ORMV genome sequence (original or corrected) behind the T7 promoter is depicted schematically: the restriction site Pst I or Nsi I, respectively, just downstream of the genome (located in multiple cloning site) was used for linearization of the plasmid, followed by run-off transcription by the T7 polymerase in the presence of a cap analogue. The resulting in vitro transcript (ORMV genomic RNA) was taken for mechanical inoculation of *N. benthamiana* plants. The picture shows the inoculated plants at 10 days post-inoculation. Copied from (Seguin et al., 2014a).

3.4 Analysis of sRNA-based antiviral mechanisms in banana plants infected with Banana streak virus

3.4.1 RCA-based deep-sequencing approach to reconstruct episomal BSV species

We used the Rolling circle amplification (RCA) approach to reconstruct the genomes of six BSV species which persistently infect *Musa acuminata* plants after many years of vegetative propagation (Rajeswaran et al., 2014a). The circular BSV genomes were amplified by RCA and the resulting RCA products (concatemeric linear dsDNA) were sequenced using Illumina. The datasets of 50-nt sequencing reads were assembled de novo without any filtering step, because the viral DNA was greatly amplified by RCA and exceeded the residual amount of banana genomic DNA in the RCA product. The viral contigs were mapped against a fasta file containing the reference sequence for different BSV species to validate their presence in each samples. The use of only one fasta file for each viral sequence is mandatory to prevent errors of analysis: as we observed in the table 3.4.1.1, when the mapping is performed against different fasta files containing only one closely related sequence, false positive results can be obtained. For example, the sample BPO-101 had more contigs which mapped to the BSOLV than BSCAV although this sample is infected by BSCAV. Another solution to prevent these errors is to analyze the CIGAR value and MD:Z tag for each mapped contigs or to do a blast instead of a mapping with an analysis of score and/or alignment. But, the most efficient solution consists to use only one fasta file.

Table 3.4.1.1: mapped Seqman contigs against BSV species

The black and red values indicate the number of mapped contigs when the mapping was performed against one fasta file per BSV or one fasta file containing all BSV sequences respectively. The bold red value indicates that the count of mapped contigs is the same between the two different conditions.

BSV viruses	BPO-96	BPO-97	BPO-98	BPO-99	BPO-100	BPO-101
BSV-Cav	175/0	1/0	11/0	6/0	0/0	30/31
BSV-GF	0/0	**33**	0/0	1/0	0/0	0/0
BSV-Im	10/0	0/0	**76**	**26**	0/0	8/0
BSV-Mys	2/0	0/0	1/0	1/0	**40**	0/0
BSV-OL	**286**	4/0	11/0	1/0	0/0	39/0

IGV was used to identify SNPs and/or In/Dels in the contigs and between the contigs and the corresponding reference genomes in order to reconstruct a master genome for each BSV species. We used only non-redundant reads to determine the master genomes. 27 SNPs and 4 indels were identified in the BSOLV reference genome; 59 SNPs in BSGFV; 18 SNPs and 2 indels

for BSIMV; 20 SNPs and 6 indels in BSVNV; and 81 SNPs and 2 indels in BSMYV. For BSCAV, there was a much bigger difference between the contigs and the Genbank BSCAV sequence: 426 SNPs and 14 indels. Moreover, in order to close small observed gaps, we repeated the mapping against the corrected BSCAV genome five times with the deep sequencing reads; each mapping allowed us to replace one or two N by the 5' or 3' nucleotides of mapped reads.

To validate and complete the reconstruction of the BSV master genomes, the mapping was performed using redundant deep-sequencing reads of sRNAs from the same BSV-infected plants. The majority of non-redundant reads covering each nucleotide position of the genome confirmed the consensus virus genome sequences reconstructed from RCA DNA reads The analysis of sRNA sequence variants that deviated from the master genomes by in less than 50% but more than 10% of redundant reads revealed several SNPs in each genome. The majority of these SNPs were in the intergenic regions without any influence on the conserved cis-elements or in the coding regions without any change in the encoded amino acids. Just a few SNPs changed the amino acid in BSV ORFs. This observation suggests that the majority of SNPs constitute viable variants of the viral genomes present in the BSV quasispecies.

The comparison of the reconstructed master genomes with the corresponding Genbank reference sequences for six BSVs revealed that these sequences share between 98.9 and 99.7% of identity. Apparently, the nucleotide differences must be due to the evolution of BSV quasispecies in banana plants during their long vegetative persistence. Concerning BSVNV, our master genome had a longer ORFI produced by a frame shift due to 2 nucleotide indels. We hypothesize that the Genbank BSVNV contains sequencing errors, but experimental validations are needed to validate this hypothesis. For BSCAV, the difference between the sequences is the most dramatic (93.8% of identity with the BSCAV Genbank sequence) and the amino acids content of the three ORF are changed. For this reason, the reconstructed master BSCAV genome corresponds to a new strain of BSCAV. Next, we used these master BSV genomes as references to study the antiviral mechanisms based on sRNAs (Rajeswaran et al., 2014a).

3.4.2 Analysis of BSV-derived siRNAs

We mapped 20-25-nt sRNA populations from BSV-infected banana plants to the reconstructed master genomes. A significant proportion of 24-nt sRNAs from the control sample of *M. balbisiana* (BPO-59) mapped to four of the six BSV genomes, except for BSCAV and BSVNV (Fig 3.4.2.1). These sRNAs are likely produced from the integrated BSV genomes within the *M. balbisiana* (B) genome. In contrast, only negligible numbers of sRNAs from healthy *M. acuminata* banana plant mapped to the BSV genomes, consistent with the fact that the *M. acuminata* (A)

genome does not contain any integrated BSV. In BSV-infected *M. acuminata* plants, the majority of viral sRNAs belongs to 21-nt class; followed by the 22-nt class for BSCAV, BSGFV, BSIMV and BSVNV; or by the 24-nt class for BSMYV and BSOLV. In comparison, *A. thaliana* plants infected with the distantly-related pararetrovirus CaMV accumulate predominantly 24-nt viral sRNA. Thus, the PTGS pathway seems to be favoured for the antiviral mechanisms based on sRNA for BSGFV, BSCAV, BSIMV and BSVNV, while episomal BSOLV and BSMYV seems to be additionally targeted by TGS pathway in banana, like in the case of CaMV in Arabidopsis. In *M. balbisiana* plant, 24-nt viral sRNAs derived from the integrated BSGFV, BSIMV, BSOLV and BSMYV (Chabannes et al., 2013; Geering et al., 2005) are predominant, suggesting that TGS pathway is responsible for silencing of the integrated pararetroviruses in banana plants.

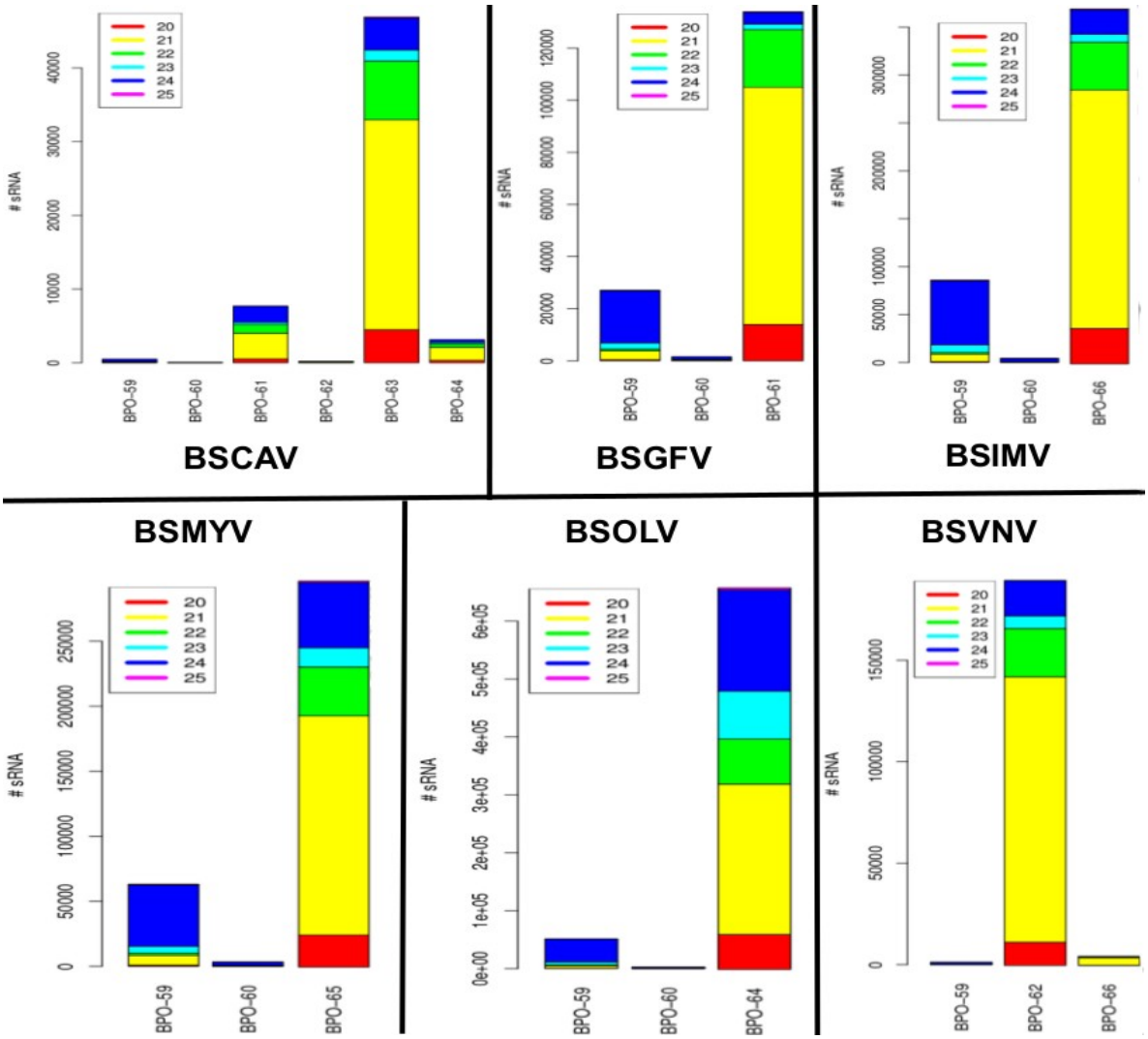


Figure 3.4.2.1: Distribution of mapped reads against BSV genomes.

The histogram represents the count of mapped reads for BPO-59 – 66 samples. The y-axis represents the count of perfect mapped reads. The bars have separated colours according to the size of mapped reads. The BPO-60 sample is the *Musa acuminata* control plant and the BPO-59 sample is the *Musa balbisiana* control plant.

The majority of 21 and 22-nt viral sRNAs starts with a 5' U for all BSV species from BSV infected *M. acuminata* plants. In *M. balbisiana* (BPO-59), 24-nt viral sRNAs start predominantly with 5' A or 5'U for all (BSGFV, BSIMV, BSMYV and BSOLV) integrants (Fig 3.4.2.2). These results suggest that banana Argonaute proteins are associated with vsiRNAs in both cases, similar to Arabidopsis AGO that sorts sRNAs based on the 5' nucleotide identity (Rajeswaran et al., 2014a).

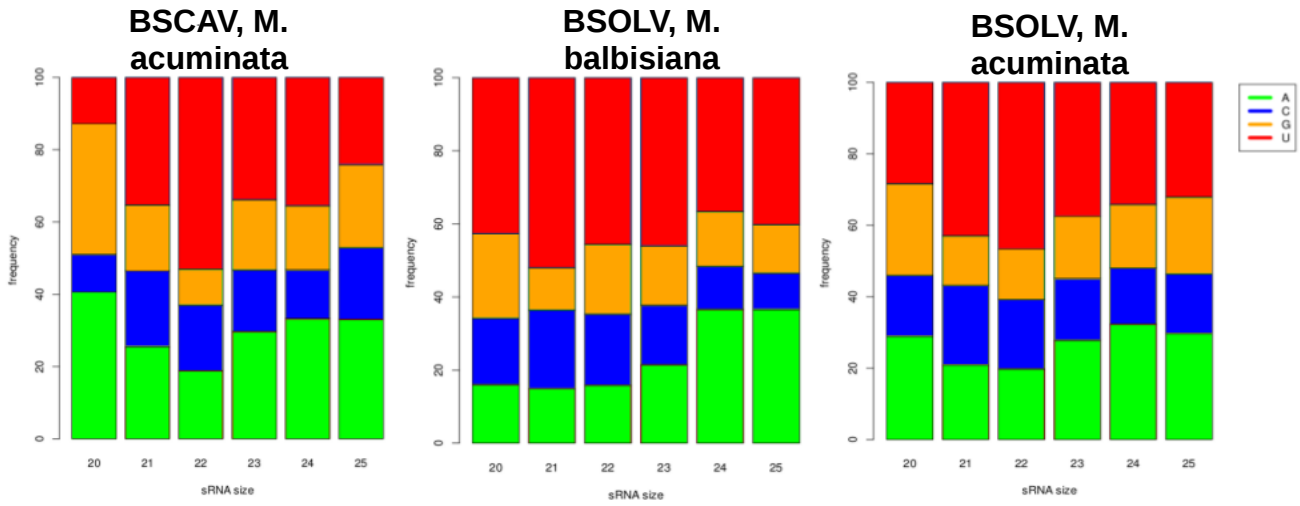


Figure 3.4.2.2: distribution of 5' nucleotide among reads mapped along BSCAV and BSOLV genome.

Each bar corresponds to one size class of reads. The y axis indicates the percentage values found for the distribution.

The coverage of redundant mapped reads along each BSV species shows that both strands of the viral genome are covered without gaps. The hot-spot of 21 and 22-nt viral sRNAs are located within the ORF I, the ORF II and the 5' region of ORF III along all the BSV genomes except for BSOLV (Fig 3.4.2.3). This result suggests that these sRNAs are produced from dsRNA precursors which may be preferentially amplified by a RDR activity. The 24-nt viral sRNAs are distributed along all the BSV genomes without major hot-spots, excepted for BSOLV. For BSOLV, the ORF III is covered by abundant 24-nt viral sRNAs, while although the 21 and 22-nt viral sRNAs are mainly generated from the ORF I and ORF II regions (Fig 3.4.2.4). Apparently, all these different size-classes of siRNAs are processed from different dsRNA precursors (Rajeswaran et al., 2014a).

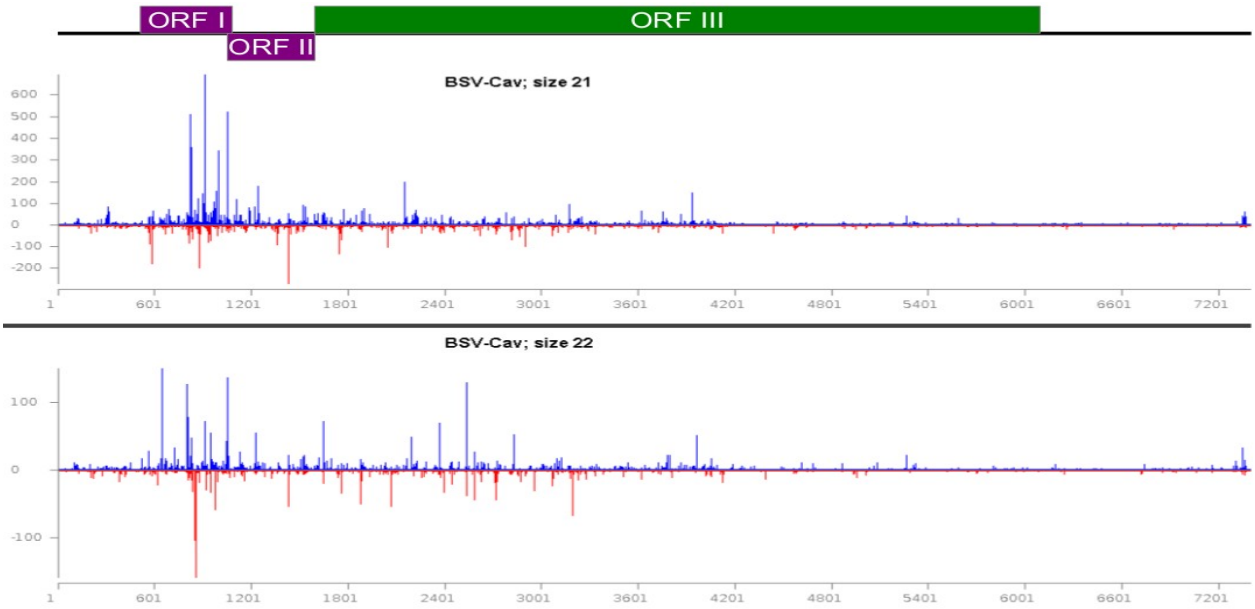


Figure 3.4.2.3: Profile of mapped reads along the BSCAV genome for BPO-63

The x-axis represents the BSCAV genome. The y-axis represents the count of mapped reads. The blue bars represent the count of reads mapped on the forward strand, and the red bars represent the count of reads mapped on the reverse strand. These profiles include mapped reads with 0-2 mismatches.

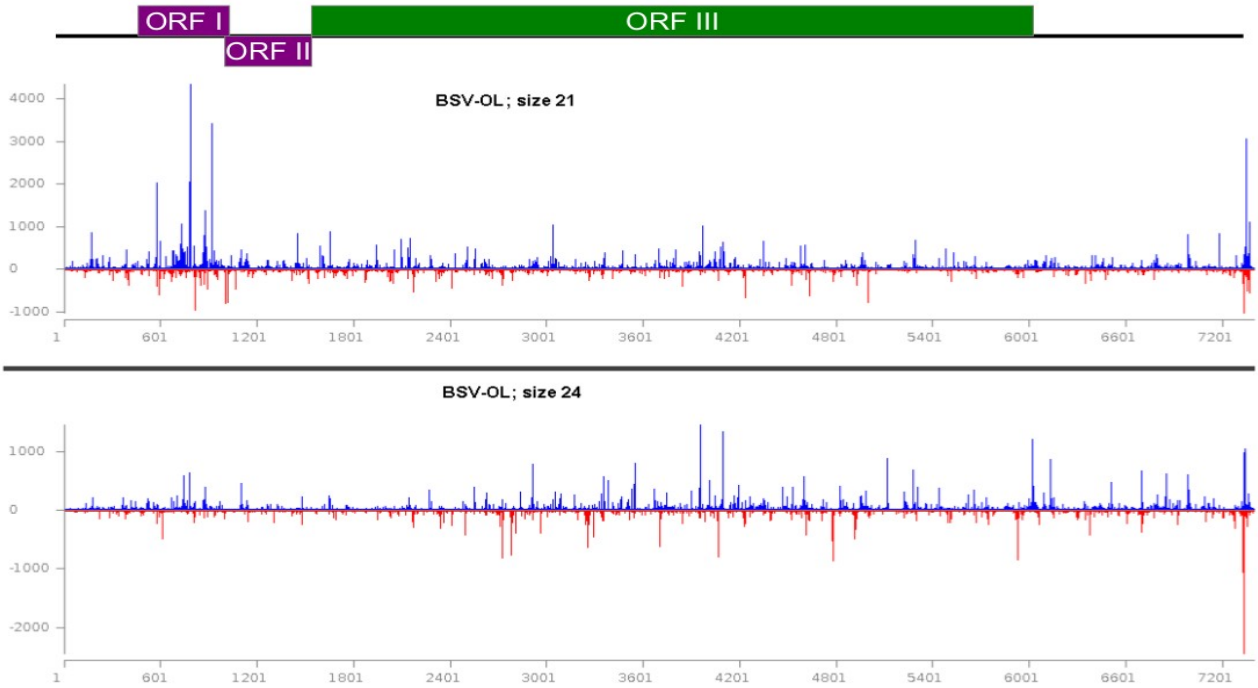


Figure 3.4.2.4: Profile of mapped reads along the BSOLV genome for BPO-64

The x-axis represents the BSOLV genome. The y-axis represents the count of mapped reads. The blue bars represent the count of reads mapped on the forward strand, and the red bars represent the count of reads mapped on the reverse strand. These profiles include mapped reads with 0-2 mismatches.

The SNP calling for the master genomes in six BSV pararetroviruses in banana plants revealed up to 21 SNPs. More than half of these SNPs were supported by more than 20 % (but less than 50%) of reads (cf table 3.4.2.5). This higher proportion of diversity in the BSV quasispecies, compared to the pararetrovirus CaMV in Arabidopsis, can be explained by the long persistence of BSV in vegetatively-propagated banana plants, while Arabidopsis had undergone only one infection cycle following inoculation with the CaMV infectious clone.

Table 3.4.2.5: Count of SNPs found in BSV master genomes

The first column indicates the name of BSV and the sample identifier. The second column indicates the number of found SNPs. The third column indicates the number of SNPs which are represented by more than 20% of reads. The last column is the percentage of SNPs which are represented by more than 20% of reads.

BSV viruses	# total SNPs	# SNPs > 20 % reads	% SNPs > 20 % reads
BSGFV (BPO-61)	8	5	62,50%
BSVNV(BPO-62)	18	13	72,22%
BSCAV (BPO-63)	21	13	61,90%
BSOLV (BPO-64)	8	5	62,50%
BSMYV (BPO-65)	17	14	82,35%
BSIMV (BPO-66)	21	12	57,14%

3.4.3 Analysis of endogenous sRNAs in M. acuminata

The sRNA populations were performed against the *M. acuminata* genome sequenced by CIRAD (D'Hont et al., 2012). This sequenced genome contains 11 chromosomes and ca 30% of the sequences that could not be anchored to the chromosomes. The analysis of the sRNA profile showed that there is a high proportion of 21 and 22-nt sRNA classes. This profile varied for the chromosomal sequences versus the unanchored sequences: the majority of 20-nt sRNAs mapped to the unanchored sequences, followed by the 24-nt sRNAs (Fig 3.4.3.1).

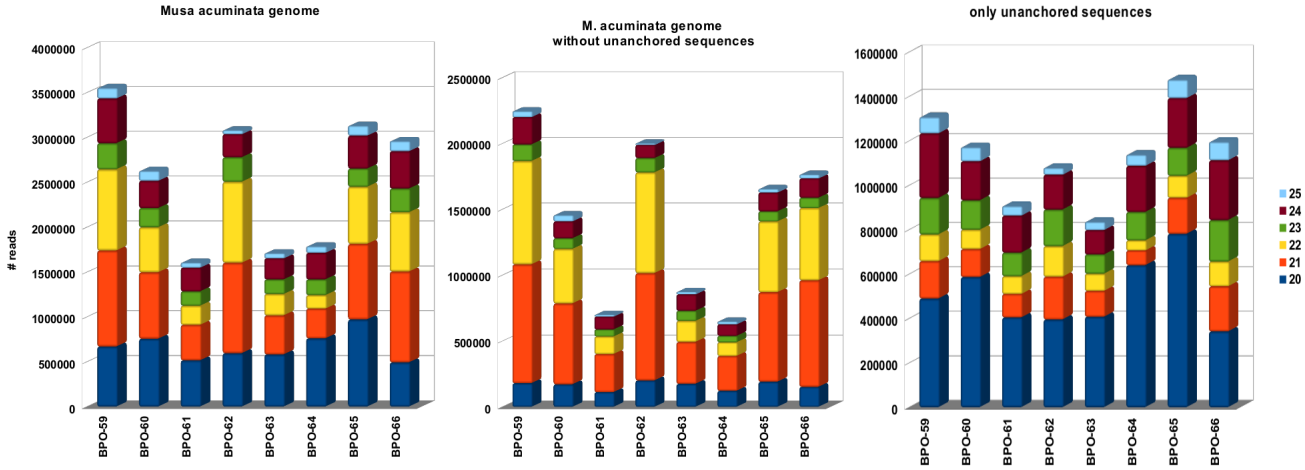


Figure 3.4.3.1: Distribution of mapped reads against *Musa acuminata* genome.

The histogram represents the count of mapped reads for BPO-59-66 samples. All samples are extracted from *Musa acuminata* plant excepted BPO-60 which is a *Musa balbisiana* sample. BPO-61-66 are infected by BSV. The y-axis represents the count of mapped reads. The bars have separated colours according to the size of mapped reads.

The analysis of 5' nucleotide of endogenous banana sRNAs showed that the majority of 21 and 22-nt sRNAs starts with 5'U. A high proportion of 24-nt sRNA starts with 5' A and 5'U. This 5'-nt distribution suggests that Argonaute proteins are involved in sRNA sorting and stabilization according to their 5' specificity as observed in *A.* thaliana. Moreover, 20-nt sRNA, have an increase of the distribution of 5'G when the unanchored sequences are included within the *M. acuminata* genome during the mapping (Fig 3.4.3.2). The abundance of 20-nt 5'G sRNA in banana suggests that they are a novel class of plant sRNAs (Rajeswaran et al., 2014a).

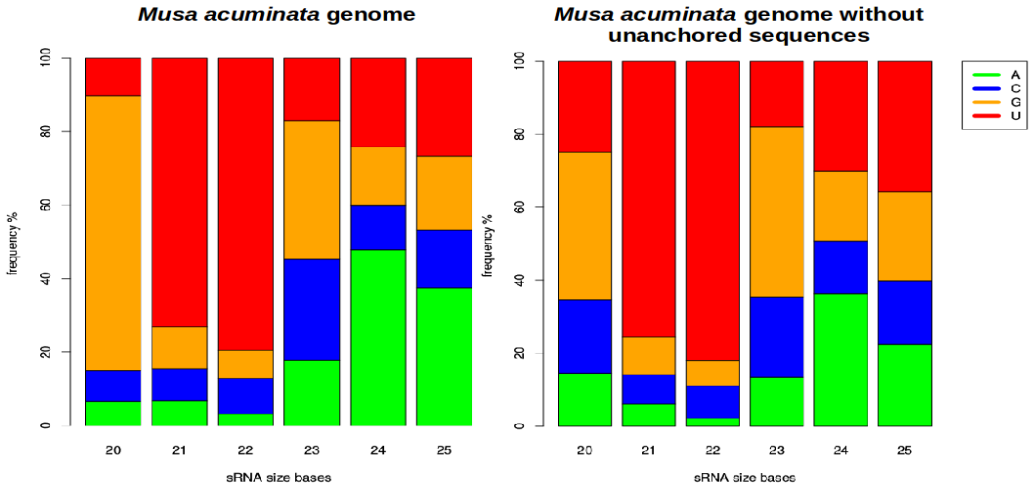


Figure 3.4.3.2: 5' nucleotide profile of sRNAs mapped to the *Musa acuminata* genome.

Each bar corresponds to one size class of reads. The y axis indicates the percentage values found for the distribution.

3.5 Analysis of sRNA-based antiviral mechanisms in rice plants infected with RTBV

3.5.1 Analysis of RTBV-derived viral siRNAs

20-25-nt sRNA populations from two RTBV infected rice plants were mapped against the RTBV genome. 21-nt sRNAs were found to be the major size-class of viral sRNAs, followed by the 22-nt- and 24-nt sRNAs (Fig 3.5.1.1). Both strands of the circular double stranded RTBV genome were covered by sRNAs.

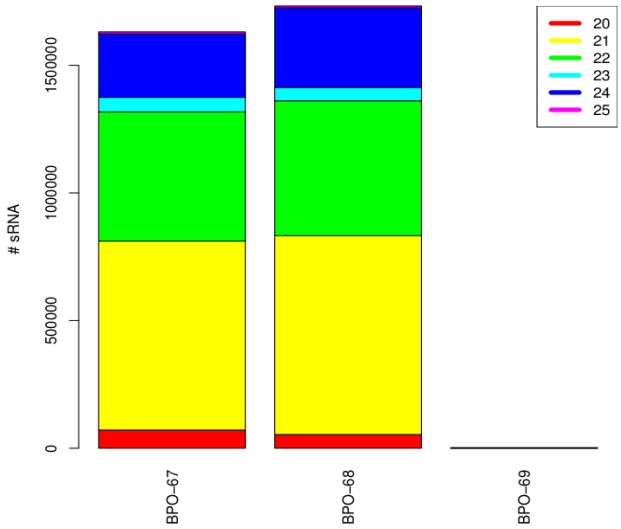


Figure 3.5.1.1: Distribution of mapped reads against RTBV genome

The histograms represent the count of mapped reads for BPO-67 - 69 samples (RTBV infected and control plants, respectively). The y-axis represents the count of mapped reads. The bars have separated colours according to the size of mapped reads.

The 5' nucleotide profile was found to be heterogeneous for the three major size-classes of viral sRNAs: the majority of 21 and 22-nt sRNAs starts with a 5' U, while the majority of 24-nt long starts with a 5' A (Fig 3.5.1.2). The 5' specificity is similar to the specificity of Argonaute proteins in *A. thaliana*.

The rice genome encodes 8 DCLs, 5 RDR and 19 AGO proteins (Kapoor et al., 2008). Three OsAGO1 clade proteins are involved in the 21-nt miRNA pathway and bind specifically the miRNA starting with 5' U (Wu et al., 2009). The 5' nucleotide identity of viral sRNA indicates that OsAGO1 clade proteins are likely associated with vsiRNAs. OsAGO4 clade proteins are involved in the rice TGS pathway by binding 24-nt miRNA and siRNAs. However, these AGOs are less specific to the 5' nucleotide, excepted for osAGO4a (Wu et al., 2010). OsAGO4a binds specifically 24-nt miRNA starting with 5'A. Thus, the TGS pathway is also involved in silencing the RTBV genome during infection in rice, even though the major antiviral mechanism against RTBV uses the PTGS pathway with the 21 and 22-nt viral sRNAs.

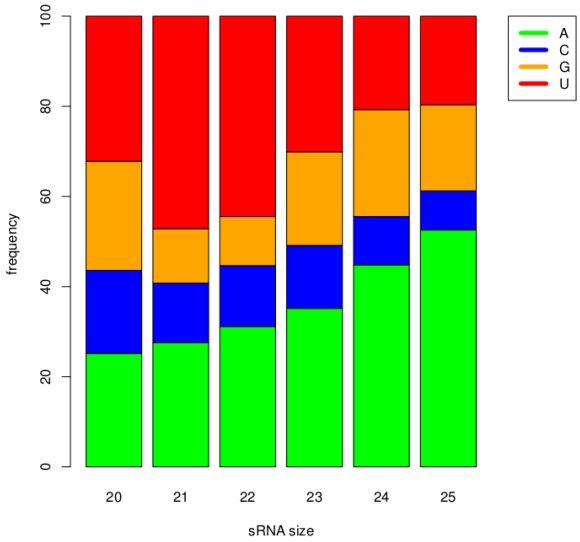


Figure 3.5.1.2: distribution of 5' nucleotide among reads mapped along RTBV genome for BPO-67.

Each bar corresponds to one size class of reads. The y axis indicates the percentage values found for the distribution.

The vsiRNA profile for RTBV in rice is similar to the profile observed for CaMV in Arabidopsis (Fig 3.5.1.3). The entire RTBV genome is covered by non-redundant vsiRNAs along both strands without gaps. This allows to reconstruct the complete RTBV genome. Comparison of the genome coverage with redundant sense and antisense vsiRNAs showed that the sRNAs was likely produced from perfect dsRNA precursors, rather than secondary structures of viral pgRNA of sense polarity.

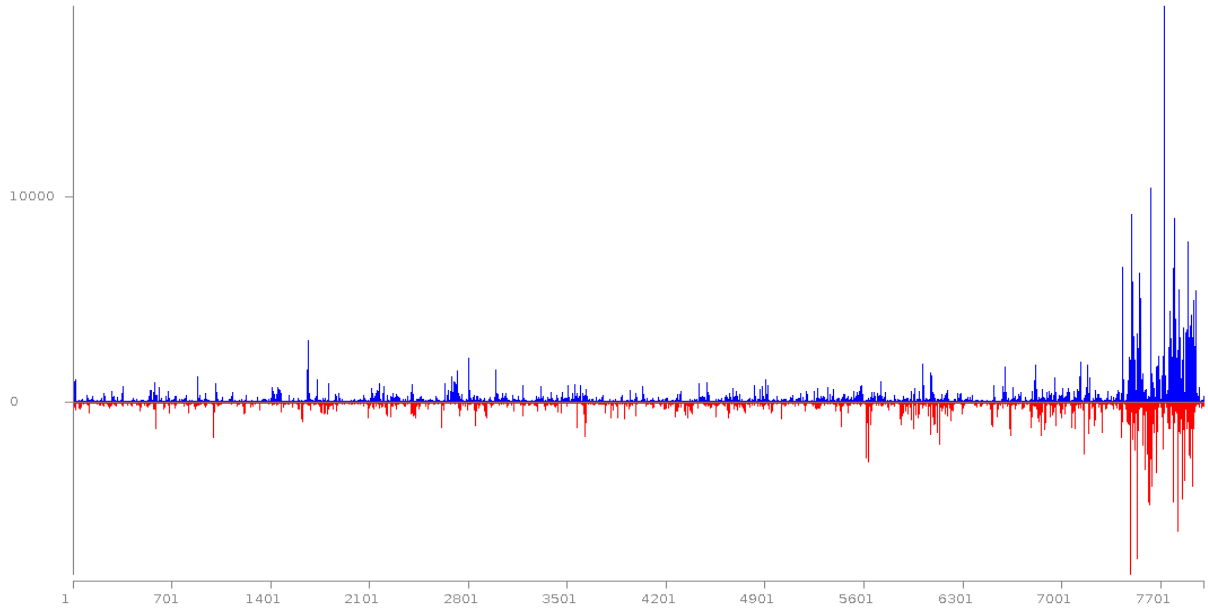


Figure 3.5.1.3: Profile of mapped reads along the RTBV genome for BPO-67

The x-axis represents the RTBV genome. The y-axis represents the count of mapped reads. The blue bars represent the count of reads mapped on the forward strand, and the red bars represent the count of reads mapped on the reverse strand. These profiles include mapped reads with 0-2 mismatches.

3.5.2 Analysis of endogenous sRNAs in Oryza sativa japonica

20-25-nt sRNAs form healthy and RTBV infected rice plants were mapped against the *Oryza sativa* genome MSU6 (http://rice.plantbiology.msu.edu/, (Ouyang et al., 2007; Project, 2005)). The mapping revealed that 24-nt reads are most abundant, followed by the 21 and 23-nt reads. The high proportion of 24-nt endogenous sRNAs indicates that the 24-nt sRNA-directed DNA methylation TGS pathway is a predominant silencing pathway in rice leaves. Similar predominance of 24-nt sRNAs was already observed in *A. thaliana,* whereas 21 and 22-nt sRNAs are the major size-classes in *M. acuminata*. Furthermore, there is a significant proportion of 20-nt sRNAs in rice, similar to banana but not Arabidopsis (Fig 3.5.2.1). The size class profile of endogenous sRNAs is slightly affected by RTBV infection (Rajeswaran et al., 2014b; Song et al., 2012).

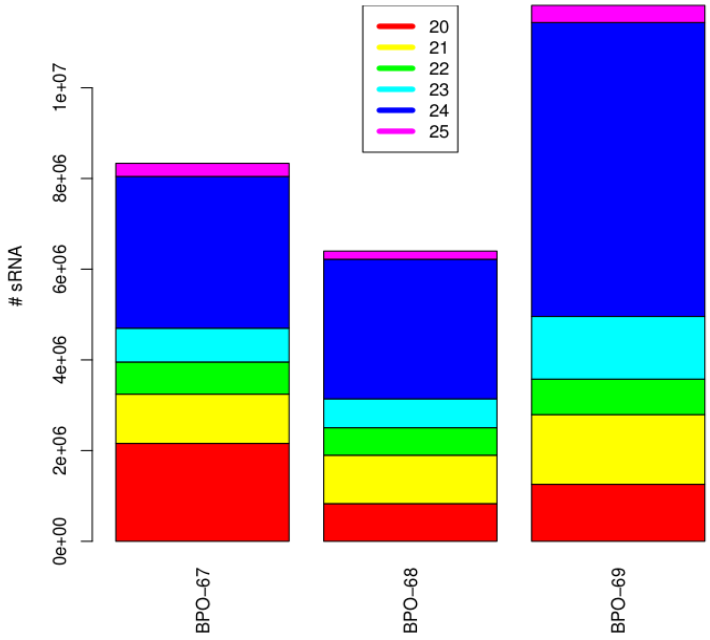


Figure 3.5.2.1: Distribution of mapped reads against *Oryza sativa japonica* genome

The histograms represent the count of mapped reads for BPO-67 - 69 samples (RTBV infected and control plants, respectively). The y-axis represents the count of mapped reads. The bars have separated colours according to the size of mapped reads.

Like with the RTBV genome, the profile of 5' nucleotides is relatively heterogeneous for each of the size-classes of mapped reads, except for 20-nt sRNAs: the overwhelming majority of 20-nt sRNAs starts with 5' G. (Fig 3.5.2.2). The major 5' G 20-nt sRNA class was already observed in *M. acuminata*; this suggest that 20-nt 5'G-sRNAs may be a common feature of monocotyledons plants (Rajeswaran et al., 2014b). the 5'-nt profile of endogenous rice sRNAs is not affected by RTBV infection, likely because RTBV replication is limited to the phloem tissues.

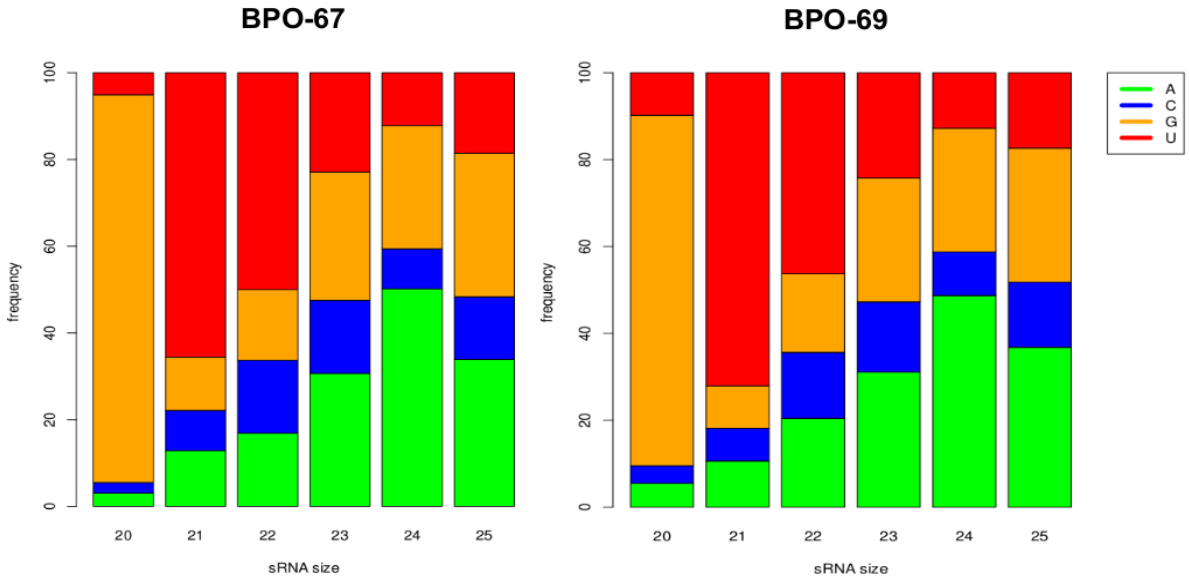


Figure 3.5.2.2: Distribution of the 5' nucleotide among reads mapped on the *O. sativa japonica* genome for BPO-67 (RTBV infected) and BPO-69 (control).

Each bar corresponds to one size class of reads. The y axis indicates the percentage values found for the distribution.

3.6 Analysis of sRNA-based antiviral mechanisms in cassava plants infected with ICMV/SLCMV

3.6.1 Analysis of ICMV/SLCMV-derived viral siRNAs

20-25-nt sRNA populations from healthy and I/SLCMV-infected plants were mapped to the corresponding ICMV or SLCMV genomes. The size profile of viral sRNAs dominated by 21-nt and 22-nt sRNAs, followed by the 24-nt sRNAs (Fig 3.6.1.1). There was no dramatic difference in the size-class profiles of sRNAs derived from different strains ICMV or SLCMV, although overall accumulation of vsiRNAs differed dramatically.

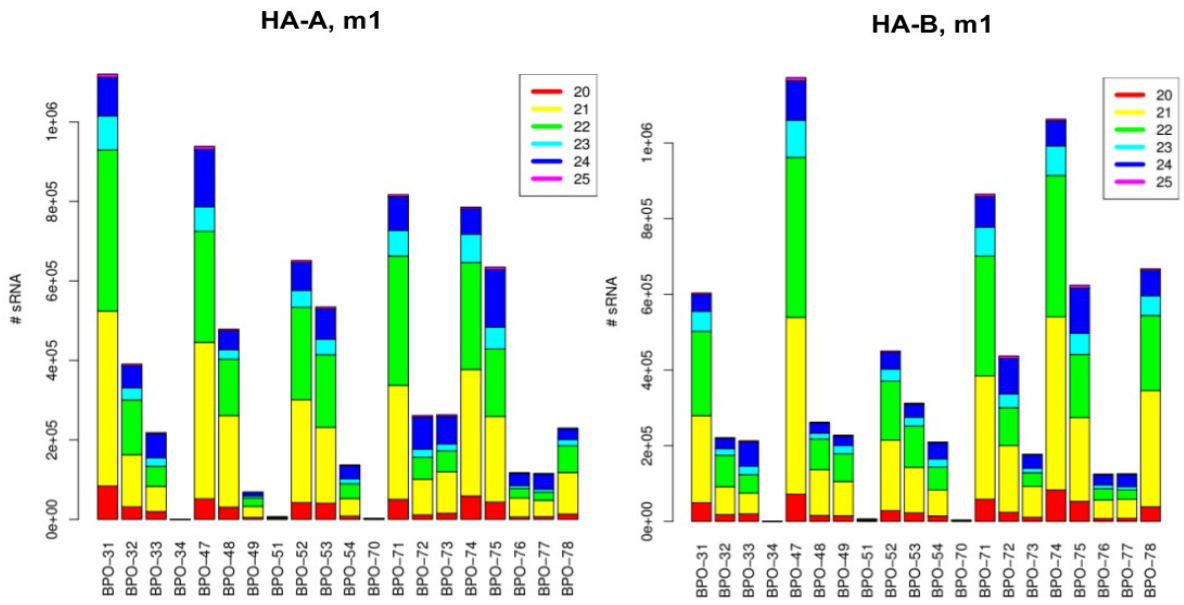


Figure 3.6.1.1: Distribution of mapped reads against strain HA of ICMV/SLCMV genome

The histograms represent the count of mapped reads for BPO-31-34, BPO-47-49, BPO-51-54 and BPO-70-78 samples. BPO-31, -47, -51 and -71 are virus- infected samples for cassava genotype H226; BPO-32, -48, -52 and -53 genotype M4; BPO-33 and -54 genotype SMAL; BPO-34, -49 genotype STVM; BPO-72-73 genotype VTP; BPO-74-75 genotype S857; BPO-76-77 genotype MVD; BPO-70 is a control sample. The y-axis represents the count of mapped reads. The bars have separated colours according to the size of mapped reads.

The majority of 24-nt viral sRNAs starts with 5' A, while high proportions of 21 and 22-nt sRNAs start with 5' U or 5' G. This 5'-nt profile slightly varies for different strains of ICMV/SLCMV (Fig 3.6.1.2). Nevertheless, the distribution of 5' nucleotide seems to be homogeneous for the two size-classes. If the AGO and DCL proteins of *Manihot esculenta* have the same specificity as their homologous proteins of *A thaliana*, these results suggest that the PTGS and TGS pathways are involved to defend the cassava plant against these viruses. The main pathway is PTGS directed by 21 or 22-nt vsiRNAs.

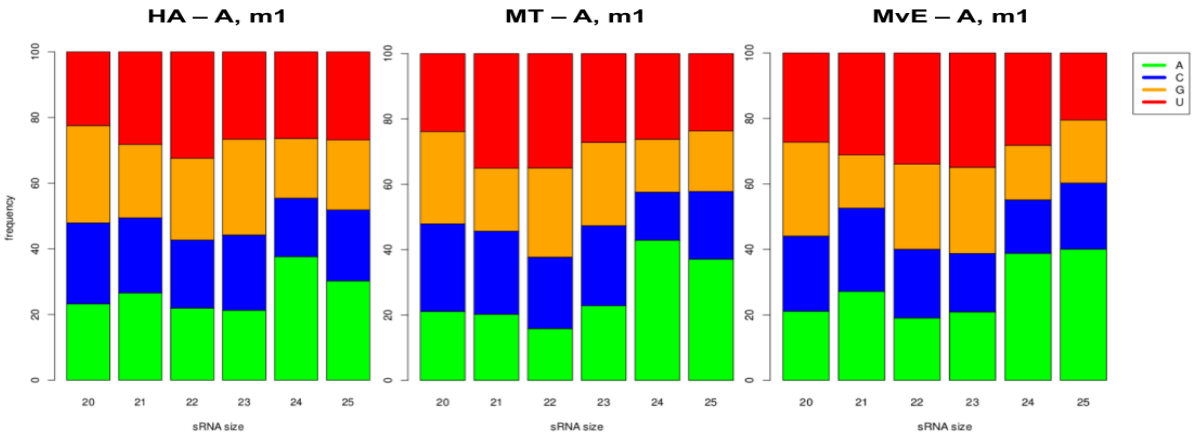


Figure 3.6.1.2: distribution of 5' nucleotide among reads mapped along different strains of ICMV/SLCMV

Each bar corresponds to one size classes of reads. The y axis indicates the percentage values found for the distribution.

The coverage of viral sRNAs along the viral genome was found to be heterogeneous. In the viral sRNA hotspot regions, no substantial bias to the forward or reverse strand was observed and seems to be equally covered in the same hotspot (Fig 3.6.1.3).

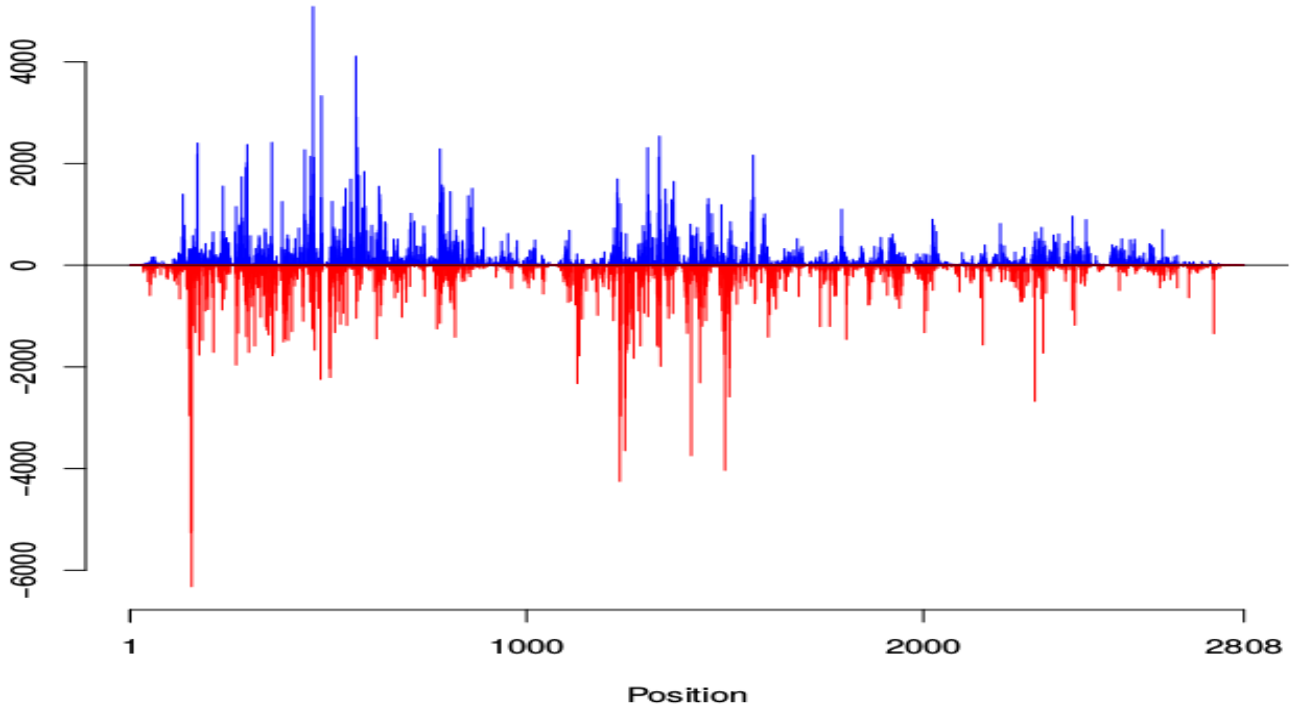


Figure 3.6.1.3: Profile of mapped reads along the strain HA of ICMV/SLCMV genome

The x-axis represents the HA strain genome. The y-axis represents the count of mapped reads. The blue bars represent the count of reads mapped on the forward strand, and the red bars represent the count of reads mapped on the reverse strand. This profiles includes mapped reads which perfectly matched.

3.6.2 Analysis of sRNAs derived from Manihot esculenta genome

Among 20-25-nt sRNAs mapped to the *Manihot esculenta* genome, 21-nt class was predominant, followed by the 24 and 22-nt classes (Fig 3.6.2.1).

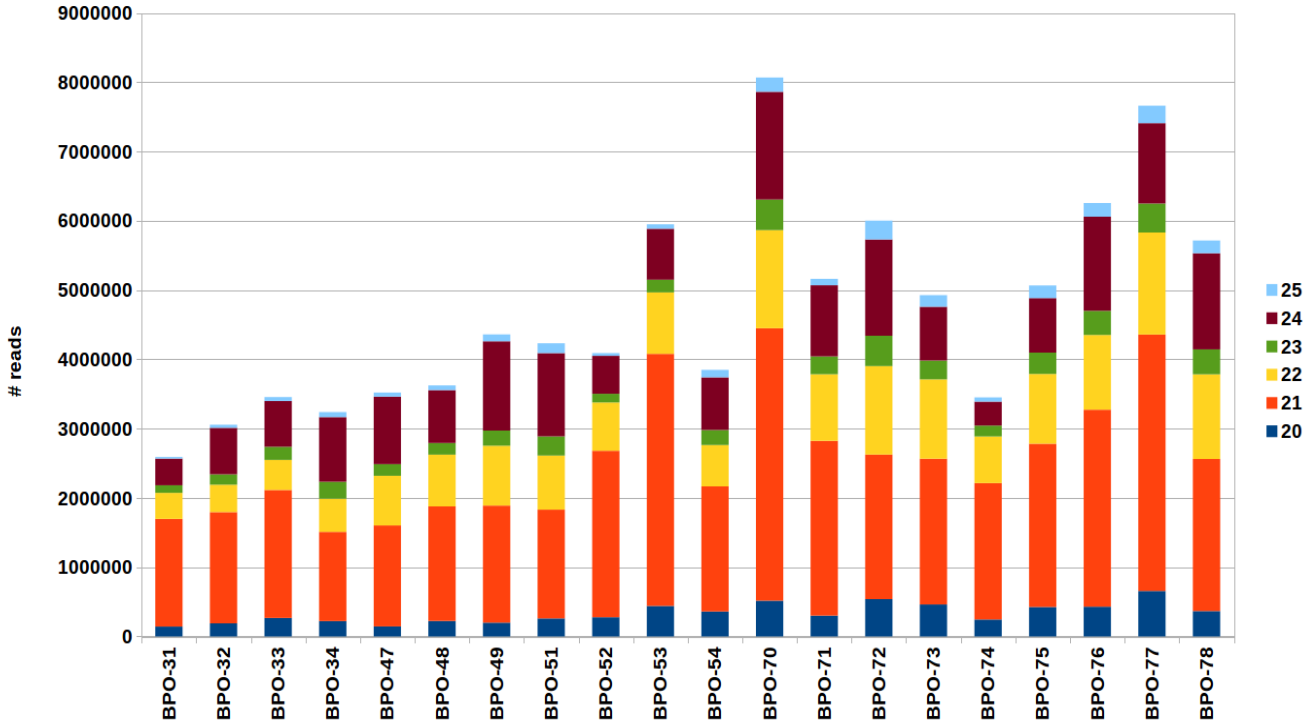


Figure 3.6.2.1: Distribution of mapped reads against *Manihot esculenta* genome

The histograms represent the count of mapped reads for BPO-31-34, BPO-47-49, BPO-51-54 and BPO-70-78 samples. BPO-31, -47, -51 and -71 are virus- infected samples for cassava genotype H226; BPO-32, -48, -52 and -53 genotype M4; BPO-33 and -54 genotype SMAL; BPO-34, -49 genotype STVM; BPO-72-73 genotype VTP; BPO-74-75 genotype S857; BPO-76-77 genotype MVD; BPO-70 is a control sample. The y-axis represents the count of mapped reads. The bars have separated colours according to the size of mapped reads.

The majority of 21 and 22-nt sRNA starts with 5' U. In contrast, the 24-nt sRNAs did not exhibit any dramatic bias in the 5'-nt profile. The results suggest the involvement of several distinct AGOs in sorting endogenous sRNAs in cassava.

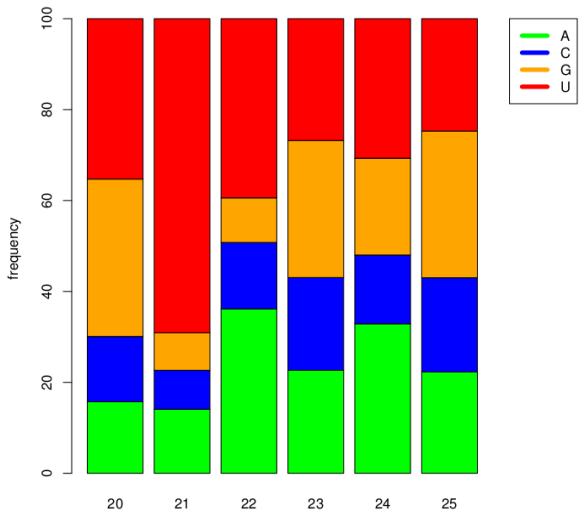


Figure 3.6.2.2: Distribution of 5' nucleotide among reads mapped along *Manihot esculenta* genome for BPO-70 (control plant).

Each bar corresponds to one size class of reads. The y axis indicates the percentage values found for the distribution.

4. Discussion

In this work, we used small RNA populations extracted from healthy and virus-infected *A thaliana*, *M. esculenta*, *O. sativa,* and *M. acuminata* plants to identify and reconstruct the viral genome and analyze the biogenesis and diversity of viral and endogenous sRNAs. The methods developed and applied in this thesis work for deep sequencing and bioinformatics analysis of vsiRNAs are summarized by the term 'siROmics'.

4.1 siROmics Approach

Currently, the Illumina technology allows for sequencing of many millions of small RNAs per sample and several samples can be multiplexed using indexes and sequenced in one lane of Illumina Genome Analyzer HiSeq 2000 which generates 150-200 million reads per lane. Our siROmics approach uses two main bioinformatics approaches for the analysis of Illumina sRNA datasets: the mapping and the *de novo* assembly.

4.1.1 Reconstruction *de novo*

In 2009, the group of Dr. Kreuze reported that it is possible to identify viruses in infected

plants by using *de novo* assembly bioinformatics tools such as Velvet with small RNA populations (Kreuze et al., 2009). Based on their work, we attempted to reconstruct complete genomes of RNA and DNA viruses from model (Arabidopsis) and crop plants. The idea was to develop an analysis pipeline which can be used to universal diagnostics of plant viral disease by reconstruction of the genome of viruses present in a plant. We selected bioinformatics tools known to be able to reconstruct a relatively small viral genome from siRNAs. We compared different strategies for the *de novo* assembly and found that using non-redundant and redundant sRNA reads give similar results although non-redundant reads generate longer contigs and allow to reduce the time of computing for the *de novo* assembly. Furthermore, our analysis revealed that to succeed with *de novo* reconstruction of a complete viral genome as a single contig of vsiRNAs, it is necessary to have a filter step in which host RNA sequences are removed by mapping a set of the sRNAs or the sRNA contigs to the host genome. Such host sRNAs or sRNAs contigs were found to interfere with the *de novo* assembly of the complete viral genome. The mapping tools are currently the best to perform the filter step because their output Bam/Sam files are easy to parse and keep the sequencing quality information, which facilitates a subsequent creation of the input fastq files for *de novo* assembly steps. The position of the filter step in a *de novo* reconstruction pipeline is problematic because it can be performed either early or late in the pipeline. An early filter eliminates the host sequences and consequently reduces the computing time for the *de novo* assembly step, with the risk to remove common contigs. With a late filter, the probability that a contig belongs to both the host and the viral genomes is close to zero. Our best strategies place the filter step after the Velvet/Oases or Velvet/Metavelvet assembly steps. Nevertheless, the filter step can be performed directly after the small RNA assembly using Velvet or equivalent assemblers which allows to use contigs smaller than contigs created by Oases or Metavelvet. But our obtained results with the strategies 5 and 6 showed that the best results were obtained with contigs generated with the Velvet/Oases or Velvet/Metavelvet associations. Metavelvet is an interesting complementary assembler for Velvet: It yields the best result when multiple viruses and viroids are present in a sample, like for example in our infected grapevine samples.

Using the siRomics approach, we reconstructed the complete genomes for all the viruses from experimentally infected *Arabidopsis thaliana* as a single contig, except for a DNA-A component of CaLCuV. The difficulty for CaLCuV is due to the common region sequence involved in replication of both DNA-A and DNA-B components. This near identical sequence of ca. 200-nt prevented a correct assembly of the DNA-A and the DNA-B as single contigs, generating three contigs representing 99% of the complete genome of CaLCuV. For the grapevine samples, the strategies used generated diverse contigs corresponding to concatenated viroids, parts of viruses and unknown sequences. Even though we identified and reconstructed the DNA virus and the two viroids, the unknown sequences may represent other micro-organisms or unknown host

sequences. Consequently, the universality of our siRomics approach is limited to the completely sequenced plant genomes such as Arabidopsis thaliana and the specific cultivars of crop plants such as grapevine for which the complete reference genome is available. With the current development of NGS technologies, the genomes of all economically important crop plants and their cultivars will be sequenced within the next 10 years, which will be important for identification of the pathogenic and non-pathogenic viruses and viroids persisting in the respective crop plants. Nevertheless, even with an incomplete genome, our approach allows for universal diagnostics of unknown and recently emerging viruses such as the grapevine geminivirus identified and reconstructed in this study. Furthermore, the healthy non-infected samples can be used for the removal of the contigs corresponding to the host genome. However, as we found for CaLCuV, the sRNAs can be shared between the host and the viral genomes, and so the presence of the virus can influence the host sRNA transcriptome (Dunoyer and Voinnet, 2005), which may result in great variations of the host siROme population in infected vs healthy plants. Moreover, the profile of host miRNAs is affected during infections, leading to the alterations in plant development and resulting in the appearance of disease symptoms (Bazzini et al., 2007). These variations of miRNA population can result from the activity of viral silencing suppressor proteins (Jay et al., 2011). Consequently, the subtraction of sRNA contigs derived from the siRomes of non-infected plants is less efficient than the use of complete host genome sequence as a filtering step, because it keeps the virus-induced host contigs, which can prevent the complete *de novo* assembly of viral genomes.

4.1.2 Reconstruction of consensus master genome

In this thesis work, we demonstrate that the complete viral genome sequence reconstructed using the siRomic approach corresponds to the consensus master genome of the wild type virus under the given conditions. However, being a quasispecies, any virus continues to evolve in the changing environment or in a new host plant. Mutations can produce a substantial population of the virus genomes that deviate from the consensus master genome at some positions and eventually create a quasispecies with a different consensus genome sequence which will be processed into viral sRNAs. The concept of quasi-species is used to group the population of all the viral genome sequences which have closely-related genomes and are subjected to a continuous process of genetic variation, competition and selection (Matthews and Hull, 2002).

4.2 Antiviral mechanisms based on siRNA-directed gene silencing

The bioinformatics analysis of small RNA populations from virus-infected wild type and silencing-deficient mutant plants allows to deduce the host genes involved in the vsiRNA biogenesis (Blevins et al. 2006). Based on the findings in *Arabidopsis thaliana* and rice, the analysis of sRNA size classes allows to hypothesize which DCLs are involved in processing of viral siRNA precursors according to the length of sRNA, while the analysis of the 5' nucleotide identity can predict which AGO proteins are associated with viral siRNAs. In this PhD study, we performed such an analysis for DNA and RNA viruses as well as viroids that infect model (Arabidopsis) and crop (grapevine, banana, rice) plants, and our main findings were summarized in four research articles (Aregger et al. 2012; Seguin et al. 2014a; Rajeswaran et al. 2014a; Rejeswaran et al. 2014b).

4.2.1 RNA viruses

We found that ORMV-derived siRNAs belong predominantly to the 21-nt class, followed by 20-nt and 22-nt 5'A -classes. Only a slight bias to 5'U was observed in these major size classes. This indicates that the PTGS pathway, mediated by DCL4 and DCL2 together with multiple AGOs, is targeting this RNA virus in *A. thaliana*. In the triple mutant dcl2/3/4, the production of vsiRNAs was drastically reduced but not eliminated, suggesting that in the absence of siRNA-degenerating DCLs, the miRNA pathway mediated by DCL1 can also generate vsiRNAs with similar proportions between the different size-classes. However, a strong bias of vsiRNAs to the plus strand of the viral genome in *dcl2/3/4* (but not in wild type) Arabidopsis indicates that DCL1 is processing mostly secondary structures of the highly-abundant viral RNAs of the plus polarity, whereas DCL2 and DCL4 process vsiRNAs from double stranded RNA intermediates generated during replication of viral genomic RNAs (Malpica et al. in preparation).

4.2.2 Pararetroviruses

In CaMV-infected *A. thaliana* wild type plants, the majority of vsiRNAs belong to 24-nt class, followed by 21-nt and 22-nt classes. A high proportion of these sRNAs starts with a 5'A or 5'U. This indicates that the TGS pathway mediated by DCL3 and AGO4 clades proteins is preferably used to

protect the host plant against this DNA virus. Nevertheless, the presence of 21 and 22-nt sRNAs indicates that the PTGS pathway is also used. The involvement of both TGS or PTGS pathways is consistent with the CaMV replication cycle, in which the viral minichromosomes are accumulated in the nucleus for Pol II-mediated transcription, whereas the pregenomic 35S RNA is transported from the nucleus to the cytoplasm for translation and reverse transcription (reviewed in Pooggin 2013). The majority of all size-classes sRNAs maps to the leader region which folds in a complex hairpin structure (Hemmings-Mieszczak and Hohn, 1999; Hemmings-Mieszczak et al., 1997). However, this structure is not the main target of DCLs to produce the corresponding vsiRNAs, because this leader region produces a decoy dsRNA processed by all the four DCLs into siRNAs of both sense and antisense polarities as was previously demonstrated (Blevins et al., 2011).

For most banana streak pararetroviruses (BSV), the sRNA size class profile is different: 21 and 22-nt sRNAs constitute the majority of vsiRNAs followed by the 24-nt sRNAs. Thus, the PTGS pathway appears to be preferably used by the host banana plants. Healthy *M. balbisiana* banana plants have 24-nt sRNA derived from the integrated BSV sequences which are predominant, suggesting the maintenance of their silencing by DNA methylation. In contrast, episomal BSV species are largely non-methylated, even in the case of BSOLV which spawns relatively high levels of 24-nt siRNAs (Rajeswaran et al. 2014a). Thus, contrary to the integrated virus, the episomal BSV evades siRNA-directed DNA methylation and TGS. The substantial proportion of 21 and 22-nt vsiRNAs starting with a 5' U allows us to hypothesize that AGO1-like protein(s) in banana are associated with vsiRNAs in antiviral RISCs.

For RTBV, vsiRNAs share some similarities with those of CaMV and BSV. The sRNA size-class profile is dominated with 21 and 22-nts classes followed by 24-nt class. 21 and 22-nt sRNAs exhibit a bias to 5'U, while 24-nt viral sRNAs to 5'A. Thus, both PTGS and TGS are likely involved in silencing RTBV. The absence of mapped RTBV sRNAs in the control healthy plant suggests that there is no integrated RTBV sequences in the rice genome. However, it is possible that the MSU6 rice genome sequence is not complete. Similar to CaMV, the coverage of the RTBV genome is heterogeneous with the majority of vsiRNAs derived from both strands of the pregenomic RNA leader region which generates dsRNA precursor of vsiRNAs (Rajeswaran et al. 2014b). Thus, a decoy strategy of silencing evasion appears to be conserved in CaMV and RTBV.

4.2.3 Geminiviruses

In contrast to pararetroviruses which have different sRNA profiles in different host plants, geminiviruses show more similarities in vsiRNA populations. In CaLCuV-infected *A thaliana*, 21-nt viral siRNAs are predominant, followed by slightly less abundant 24-nt siRNAs and finally 22-nt

vsiRNAs. No strong bias in 5' nucleotide identity was observed in these major size-classes. Likewise, in GVGV-infected grapevine, 21-nt viral sRNAs are the most abundant, followed by 24-nt and 22-nt vsiRNAs. The majority of 24-nt vsiRNAs start with 5' A, while the majority 21-nt vsiRNAs start with 5'U. For these two viruses, there is no major hotspot of vsiRNA production like in CaMV and RTBV, although the ORF regions produce more abundant siRNAs for each size class.

For ICMV/SLCMV, the siRNA size class profile was found to be different in that 24-nt siRNAs is the third most abundant after 21-nt and 22-nt siRNAs. 21 and 22-nt sRNA exhibit bias to 5' U. In comparison with GVGV and CaLCuV, the hotspots of vsiRNAs are concentrated to several short regions of ICMV/SLCMV genomes.

In summary, PTGS is a major pathway targeting geminiviruses, while these viruses appear to evade siRNA-directed methylation of viral DNA and transcriptional silencing of viral genes owing to their specialized replication mechanisms in the nucleus (reviewed in Pooggin 2013).

4.3 The vaccine strategy

The main objective of the COST Action, which provided the grant for this thesis work, was to develop an RNAi-based vaccine to immunize the crop plants against viral infections. To this end, the regions of low viral sRNA production along the viral genome had to be identified to allow the design of antiviral siRNA-generating constructs targeting these cold spot regions. Then, methods of delivering these antiviral siRNAs to infected plants would be developed. This delivery would boost the endogenous silencing pathways to immunize the plant against the virus.

Previously, RNAi transgenic crop plants resistant to viruses have been generated. In collaboration with Dr. Fuentes, samples of RNAi transgenic tomatoes transformed with a hairpin dsRNA construct targeting the C1 gene of the geminivirus TYLCV were analyzed (Fuentes et al., 2006). Bioinformatics analysis of transgene- and virus-derived siRNAs showed that the hairpin structure allows to produce a highly abundant siRNAs with a strong bias to 5' U, compared to the corresponding region of the TYLCV which spawns low levels of vsiRNAs with no strong bias to 5'U or 5'A. This explains the fact that the transgenic tomato plants exhibit immunity to TYLCV disease even under the extreme open-field conditions with large amounts of viruliferous whiteflies that transmit the virus in non-transgenic plants causing severe disease. The success of this RNAi approach targeting the vsiRNA coldspot regions proves that the RNAi based vaccine strategy is promising (Fuentes et al. in preparation). However, it is not sure that this strategy can be applied for all crop plants.

In fact, some crop plants, such as banana, possess the integrated viral sequences in their

genomes. In the case of banana, these integrations may potentially develop an antiviral immunity against episomal BSV (Chabannes et al., 2013; Iskra-Caruana et al., 2014). However, episomal BSV quasispecies can evolve very fast to brake the sequence-specific resistance provided by the integrated BSV sequences. Moreover, these integrants themselves can give rise to episomal BSV infections upon various stresses including breeding (Iskra-Caruana et al., 2014). If the integrated BSVs were really involved in the banana immunity against other BSVs species, evolution would select the most efficient BSV sequences to immunize banana plants rather than keep the infectious integrants in the banana genome.

Likewise, the RNAi transgenic plants will unlikely be efficient in the long term, because the targeted virus can evolve to overcome the sequence specificity of the RNA transgene. The best strategy would be to use a transient siRNA vaccine based on the consensus master genome reconstructed using the siRomics approach. The sequence of such vaccine can be modified depending on the on-going evolution of the targeted virus quasispecies.

There are several parameters to be considered for designing an effective antiviral vaccine. Firstly, it should contain viral sequences that produce little siRNAs in virus-infected plants. Secondly, the viral gene essential for replication (such as Rep/C1 in the case of TYLCV) should be targeted. Thirdly, the host genomic sequence should not be targeted by the vaccine, because certain viral disease symptoms can be due to the repression of these host genes by vsiRNAs. Consequently, the viral sequence selected for the vaccine can potentially produce the disease symptoms even if the virus is absent. Moreover, the disease symptoms have an impact on the vector behaviour. For example, aphids can be attracted by discoloured and yellow or by special volatile organic compounds from infected plant (Ingwell et al., 2012; Mauck et al., 2012). So, the infected plants will not be discriminated from non-infected plants, and the vectors containing perhaps other viruses will be attracted by the immunized plants, leading to disease by a different virus.

To summarize, although further research on target viral sequences is required for designing an effective siRNA-based antiviral vaccine, the target sequence should correspond to a siRNA cold spot region in the evolutionary-conserved (essential) sequence of the viral genome which shares no homology with the host plant genome.

5. Conclusion and outlook

The antiviral mechanism based on siRNA-directed gene silencing is evolutionarily conserved in all land plants. Their multiple DCL, AGO and RDR genes seem to keep their activity and specificity among the different plant species. This mechanism is efficient to silence the viruses at post-transcriptional (PTGS) and, in the case of DNA viruses, also at transcriptional (TGS) levels. The development of siRNA-based vaccines is therefore a promising approach to control viral diseases in economically-important crop plants.

The RNA silencing mechanisms in crop plants are not sufficiently known to deduce general rules for siRNA-based vaccines to be applicable for all plant and viral families. With the improvement of NGS technologies, we hope to have more information on the genome sequences and endogenous and vsiRNA profile for all major crop plants in the next 10 years.

Bioinformatics tools allow us to identify and reconstruct genomes of new viruses using deeply sequenced small RNA populations. Currently, the NGS technologies do not allow us to sequence sRNA samples directly in the fields, but the miniaturization of this technology and new deep-sequencing methods may allow this at a later stage.

One of the most interesting findings in this thesis work concerns the interaction between the viral and the host genomes. During viral infection, different phenomena are observed. The easiest to understand is the antiviral mechanisms based on small RNAs. Now, it is possible to identify which regions of the viral genome generate vsiRNAs, to determine the size-classes and the 5' nucleotide identity of viral sRNAs, and to deduce (with the existing knowledge about DCL, RDR and AGO genes in *A. thaliana* and rice) the genetic requirements for vsiRNA biogenesis and function. The second phenomenon is the integration of viruses in the host plant genome which may be involved in the antiviral immunity. Finally, the targeting of the host plant genes by vsiRNAs can be a potential cause of some viral disease symptoms. All these phenomena illustrate the complexity and co-evolution of the virus-plant interactions. To counteract the siRNA-based antiviral defense, viruses evolved different mechanisms of silencing suppression and evasion. The silencing suppression mechanisms can potentially interfere with endogenous silencing pathways that regulate plant gene expression and chromatin states and thereby contribute to the development of viral disease symptoms, which might be required to attract the (insect) vectors to transmit the virus from plant to plant.

In the case of integrated viruses, the most interesting question is how and why a pararetrovirus (which differs from a retroviruses by having no obligatory genome-integration step in the replication cycle) can be integrated and kept in the host genome during the plant evolution. We suppose that this integration provides an advantage for both the virus and the host. Normally, viruses are considered as cellular parasite, a pure “organism” corresponding to the concatenation

of “selfish genes”. The genome integration may be a strategy that resembles the prisoner's dilemma as it was exposed by Richard Dawkins in his book (Dawkins, 1990). For this moment, we can just hypothesize that the integration of a viral sequence could be used by the plant to produce antiviral siRNAs to protect itself from an incoming virus. But what is the advantage for the virus to be integrated and silenced in the host genome? To be able to propagate in the plant by vertical transmission and eventually to re-emerge following intra- and interspecific hybridization, like in the case of BSV integrants that are released from the banana B genome in the hybrid banana plants?

Adams, M.J., Antoniw, J.F., and Kreuze, J. (2009). Virgaviridae: a new family of rod-shaped plant viruses. Arch. Virol. *154*, 1967–1972.

Aguilar, I., Sánchez, F., Martin Martin, A., Martinez-Herrera, D., and Ponz, F. (1996). Nucleotide sequence of Chinese rape mosaic virus (oilseed rape mosaic virus), a crucifer tobamovirus infectious on Arabidopsis thaliana. Plant Mol. Biol. *30*, 191–197.

Akbergenov, R., Si-Ammour, A., Blevins, T., Amin, I., Kutter, C., Vanderschuren, H., Zhang, P., Gruissem, W., Meins, F., Jr, Hohn, T., et al. (2006). Molecular characterization of geminivirus-derived small RNAs in different plant species. Nucleic Acids Res. *34*, 462–471.

Allen, E., and Howell, M.D. (2010). miRNAs in the biogenesis of trans-acting siRNAs in higher plants. Semin. Cell Dev. Biol. *21*, 798–804.

Al Rwahnih, M., Dave, A., Anderson, M.M., Rowhani, A., Uyemoto, J.K., and Sudarshana, M.R. (2013). Association of a DNA virus with grapevines affected by red blotch disease in California. Phytopathology *103*, 1069–1076.

Aregger, M., Borah, B.K., Seguin, J., Rajeswaran, R., Gubaeva, E.G., Zvereva, A.S., Windels, D., Vazquez, F., Blevins, T., Farinelli, L., et al. (2012). Primary and Secondary siRNAs in Geminivirus-induced Gene Silencing. PLoS Pathog. *8*, e1002941.

Banerjee, A., Roy, S., and Tarafdar, J. (2011). Phylogenetic analysis of Rice tungro bacilliform virus ORFs revealed strong correlation between evolution and geographical distribution. Virus Genes *43*, 398–408.

Bao, Y., Federhen, S., Leipe, D., Pham, V., Resenchuk, S., Rozanov, M., Tatusov, R., and Tatusova, T. (2004). National center for biotechnology information viral genomes project. J. Virol. *78*, 7291–7298.

Bartel, D.P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. Cell *116*, 281–297.

Bazzini, A.A., Hopp, H.E., Beachy, R.N., and Asurmendi, S. (2007). Infection and coaccumulation of tobacco mosaic virus proteins alter microRNA levels, correlating with symptom and plant development. Proc. Natl. Acad. Sci. *104*, 12157–12162.

Blevins, T., Rajeswaran, R., Shivaprasad, P.V., Beknazariants, D., Si-Ammour, A., Park, H.-S., Vazquez, F., Robertson, D., Meins, F., Jr, Hohn, T., et al. (2006). Four plant Dicers mediate viral small RNA biogenesis and DNA virus induced silencing. Nucleic Acids Res. *34*, 6233–6246.

Blevins, T., Rajeswaran, R., Aregger, M., Borah, B.K., Schepetilnikov, M., Baerlocher, L., Farinelli, L., Meins, F., Jr, Hohn, T., and Pooggin, M.M. (2011). Massive production of small RNAs from a non-coding region of Cauliflower mosaic virus in plant defense and viral counter-defense. Nucleic Acids Res.

Bol, J.F. (1999). Alfalfa mosaic virus and ilarviruses: involvement of coat protein in multiple steps of the replication cycle. J. Gen. Virol. *80*, 1089–1102.

Boonham, N., Kreuze, J., Winter, S., van der Vlugt, R., Bergervoet, J., Tomlinson, J., and Mumford, R. (2014). Methods in virus diagnostics: From ELISA to next generation sequencing. Virus Res. *186*, 20–31.

Bottcher, B., Unseld, S., Ceulemans, H., Russell, R.B., and Jeske, H. (2004). Geminate Structures of African Cassava Mosaic Virus. J. Virol. *78*, 6758–6765.

Burrows, M., and Wheeler, D.J. (1994). A Block-sorting Lossless Data Compression Algorithm (Digital, Systems Research Center).

Chabannes, M., Baurens, F.-C., Duroy, P.-O., Bocs, S., Vernerey, M.-S., Rodier-Goud, M., Barbe, V., Gayral, P., and Iskra-Caruana, M.-L. (2013). Three infectious viral species lying in wait in the banana genome. J. Virol. *87*, 8624–8637.

Clark, M.F., and Adams, A.N. (1977). Characteristics of the Microplate Method of Enzyme-Linked Immunosorbent Assay for the Detection of Plant Viruses. J. Gen. Virol. *34*, 475–483.

Cock, P.J.A., Fields, C.J., Goto, N., Heuer, M.L., and Rice, P.M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. *38*, 1767–1771.

Compeau, P.E.C., Pevzner, P.A., and Tesler, G. (2011). How to apply de Bruijn graphs to genome assembly. Nat. Biotechnol. *29*, 987–991.

Creager, A.N., Scholthof, K.B., Citovsky, V., and Scholthof, H.B. (1999). Tobacco mosaic virus. Pioneering research for a century. Plant Cell *11*, 301–308.

Dawkins, R. (1990). The Selfish Gene (Oxford ; New York: Oxford University Press).

Deleris, A., Gallego-Bartolome, J., Bao, J., Kasschau, K.D., Carrington, J.C., and Voinnet, O. (2006). Hierarchical Action and Inhibition of Plant Dicer-Like Proteins in Antiviral Defense. Science *313*, 68–71.

D'Hont, A., Denoeud, F., Aury, J.-M., Baurens, F.-C., Carreel, F., Garsmeur, O., Noel, B., Bocs, S., Droc, G., Rouard, M., et al. (2012). The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. Nature.

Dolja, V.V., Kreuze, J.F., and Valkonen, J.P.T. (2006). Comparative and functional genomics of closteroviruses. Virus Res. *117*, 38–51.

Drucker, M., Froissart, R., Hébrard, E., Uzest, M., Ravallec, M., Espérandieu, P., Mani, J.-C., Pugnière, M., Roquet, F., Fereres, A., et al. (2002). Intracellular distribution of viral gene products regulates a complex mechanism of cauliflower mosaic virus acquisition by its aphid vector. Proc. Natl. Acad. Sci. U. S. A. *99*, 2422–2427.

Dunoyer, P., and Voinnet, O. (2005). The complex interplay between plant viruses and host RNA-silencing pathways. Curr. Opin. Plant Biol. *8*, 415–423.

Franck, A., Guilley, H., Jonard, G., Richards, K., and Hirth, L. (1980). Nucleotide sequence of cauliflower mosaic virus DNA. Cell *21*, 285–294.

Fuentes, A., Ramos, P.L., Fiallo, E., Callard, D., Sánchez, Y., Peral, R., Rodríguez, R., and Pujol, M.

(2006). Intron-hairpin RNA derived from replication associated protein C1 gene confers immunity to tomato yellow leaf curl virus infection in transgenic tomato plants. Transgenic Res. *15*, 291–304.

Geering, A.D.W., Pooggin, M.M., Olszewski, N.E., Lockhart, B.E.L., and Thomas, J.E. (2005). Characterisation of Banana streak Mysore virus and evidence that its DNA is integrated in the B genome of cultivated Musa. Arch. Virol. *150*, 787–796.

Ghildiyal, M., and Zamore, P.D. (2009). Small silencing RNAs: an expanding universe. Nat. Rev. Genet. *10*, 94–108.

Gray, S., and Gildow, F.E. (2003). Luteovirus-aphid interactions. Annu. Rev. Phytopathol. *41*, 539–566.

Gronenborn, B. (2004). Nanoviruses: genome organisation and protein function. Vet. Microbiol. *98*, 103–109.

Haas, M., Bureau, M., Geldreich, A., Yot, P., and Keller, M. (2002). Cauliflower mosaic virus: still in the news. Mol. Plant Pathol. *3*, 419–429.

Hadfield, J., Linderme, D., Shepherd, D.N., Bezuidenhout, M., Lefeuvre, P., Martin, D.P., and Varsani, A. (2011). Complete genome sequence of a dahlia common mosaic virus isolate from New Zealand. Arch. Virol. *156*, 2297–2301.

Haible, D., Kober, S., and Jeske, H. (2006). Rolling circle amplification revolutionizes diagnosis and genomics of geminiviruses. J. Virol. Methods *135*, 9–16.

Hamilton, A.J., and Baulcombe, D.C. (1999). A Species of Small Antisense RNA in Posttranscriptional Gene Silencing in Plants. Science *286*, 950–952.

Hanley-Bowdoin, L., Settlage, S.B., Orozco, B.M., Nagar, S., and Robertson, D. (1999). Geminiviruses: Models for Plant DNA Replication, Transcription, and Cell Cycle Regulation. Crit. Rev. Plant Sci. *18*, 71–106.

Hanley-Bowdoin, L., Bejarano, E.R., Robertson, D., and Mansoor, S. (2013). Geminiviruses: masters at redirecting and reprogramming plant processes. Nat. Rev. Microbiol. *11*, 777–788.

Harper, G., and Hull, R. (1998). Cloning and Sequence Analysis of Banana Streak Virus DNA. Virus Genes *17*, 271–278.

Harper, G., Osuji, J.O., Heslop-Harrison, J.S., and Hull, R. (1999). Integration of banana streak badnavirus into the Musa genome: molecular and cytogenetic evidence. Virology *255*, 207–213.

Harper, G., Hull, R., Lockhart, B., and Olszewski, N. (2002). Viral Sequences Integrated into Plant Genomes. Annu. Rev. Phytopathol. *40*, 119–136.

Hay, J.M., Jones, M.C., Blakebrough, M.L., Dasgupta, I., Davies, J.W., and Hull, R. (1991). An analysis of the sequence of an infectious clone of rice tungro bacilliform virus, a plant pararetrovirus. Nucleic Acids Res. *19*, 2615–2621.

Hemmings-Mieszczak, M., and Hohn, T. (1999). A stable hairpin preceded by a short open reading frame promotes nonlinear ribosome migration on a synthetic mRNA leader. RNA *5*, 1149–1157.

Hemmings-Mieszczak, M., Steger, G., and Hohn, T. (1997). Alternative structures of the cauliflower mosaic virus 35 S RNA leader: implications for viral expression and replication. J. Mol. Biol. *267*, 1075–1088.

Herzog, E., Guerra-Peraza, O., and Hohn, T. (2000). The rice tungro bacilliform virus gene II product interacts with the coat protein domain of the viral gene III polyprotein. J. Virol. *74*, 2073–2083.

Hoh, F., Uzest, M., Drucker, M., Plisson-Chastang, C., Bron, P., Blanc, S., and Dumas, C. (2010). Structural Insights into the Molecular Mechanisms of Cauliflower Mosaic Virus Transmission by Its Insect Vector. J Virol *84*, 4706–4713.

Hohn, T., and Rothnie, H. (2013). Plant pararetroviruses: replication and expression. Curr. Opin. Virol. *3*, 621–628.

Hohn, T., Corsten, S., Dominguez, D., Fütterer, J., Kirk, D., Hemmings-Mieszczak, M., Pooggin, M., Schärer-Hernandez, N., and Ryabova, L. (2001). Shunting is a translation strategy used by plant pararetroviruses (Caulimoviridae). Micron Oxf. Engl. 1993 *32*, 51–57.

Hull, R. (1996). Molecular Biology of Rice Tungro Viruses. Annu. Rev. Phytopathol. *34*, 275–297.

Hull, R. (2001). Caulimoviridae (Plant Pararetroviruses). In eLS, (John Wiley & Sons, Ltd),.

Hull, R. (2013). Plant Virology (Academic Press).

Ihaka, R., and Gentleman, R. (1996). R: A Language for Data Analysis and Graphics. J. Comput. Graph. Stat. *5*, 299–314.

Ingwell, L.L., Eigenbrode, S.D., and Bosque-Pérez, N.A. (2012). Plant viruses alter insect behavior to enhance their spread. Sci. Rep. *2*.

Ishibashi, K., Miyashita, S., Katoh, E., and Ishikawa, M. (2012). Host membrane proteins involved in the replication of tobamovirus RNA. Curr. Opin. Virol. *2*, 699–704.

Iskra-Caruana, M., Chabannes, M., Duroy, P.-O., and Muller, E. (2014). A possible scenario for the evolution of Banana streak virus in banana. Virus Res. *186*, 155–162.

Iskra-Caruana, M.-L., Baurens, F.-C., Gayral, P., and Chabannes, M. (2010). A Four-Partner Plant–Virus Interaction: Enemies Can Also Come from Within. Mol. Plant. Microbe Interact. *23*, 1394–1402.

Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., Jubin, C., et al. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature *449*, 463–467.

James, A. p., Geijskes, R. j., Dale, J. l., and Harding, R. m. (2011). Molecular characterisation of six badnavirus species associated with leaf streak disease of banana in East Africa. Ann. Appl. Biol. *158*, 346–353.

Jay, F., Wang, Y., Yu, A., Taconnat, L., Pelletier, S., Colot, V., Renou, J.-P., and Voinnet, O. (2011). Misregulation of AUXIN RESPONSE FACTOR 8 Underlies the Developmental Abnormalities Caused by Three Distinct Viral Silencing Suppressors in Arabidopsis. PLoS Pathog *7*, e1002035.

Ji, L., Liu, X., Yan, J., Wang, W., Yumul, R.E., Kim, Y.J., Dinh, T.T., Liu, J., Cui, X., Zheng, B., et al. (2011). ARGONAUTE10 and ARGONAUTE1 Regulate the Termination of Floral Stem Cells through Two MicroRNAs in Arabidopsis. PLoS Genet. *7*.

Jiwan, S.D., and White, K.A. (2011). Subgenomic mRNA transcription in Tombusviridae. RNA Biol. *8*, 287–294.

Jones, M.C., Gough, K., Dasgupta, I., Rao, B.L.S., Cliffe, J., Qu, R., Shen, P., Kaniewska, M., Blakebrough, M., Davies, J.W., et al. (1991). Rice tungro disease is caused by an RNA and a DNA virus. J. Gen. Virol. *72*, 757–761.

Kapoor, M., Arora, R., Lama, T., Nijhawan, A., Khurana, J.P., Tyagi, A.K., and Kapoor, S. (2008). Genome-wide identification, organization and phylogenetic analysis of Dicer-like, Argonaute and RNA-dependent RNA Polymerase gene families and their expression analysis during reproductive development and stress in rice. BMC Genomics *9*, 451.

Khelifa, M., Massé, D., Blanc, S., and Drucker, M. (2010). Evaluation of the minimal replication time of Cauliflower mosaic virus in different hosts. Virology *396*, 238–245.

King, A.M., Dr, E.L., Adams, M.J., and Carstens, E.B. (2011). Virus Taxonomy: Ninth Report of the International Committee on Taxonomy of Viruses (London ; Waltham, MA: Elsevier).

Kormelink, R., Garcia, M.L., Goodin, M., Sasaya, T., and Haenni, A.-L. (2011). Negative-strand RNA viruses: the plant-infecting counterparts. Virus Res. *162*, 184–202.

Krenz, B., Thompson, J.R., Fuchs, M., and Perry, K.L. (2012). Complete genome sequence of a new circular DNA virus from grapevine. J. Virol. *86*, 7715.

Kreuze, J.F., Perez, A., Untiveros, M., Quispe, D., Fuentes, S., Barker, I., and 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.

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

Lheureux, F., Laboureau, N., Muller, E., Lockhart, B.E.L., and Iskra-Caruana, M.-L. (2007). Molecular characterization of banana streak acuminata Vietnam virus isolated from Musa acuminata siamea (banana cultivar). Arch. Virol. *152*, 1409–1416.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinforma. Oxf. Engl. *25*, 1754–1760.

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics *26*, 589–595.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009a). The Sequence Alignment/Map format and SAMtools. Bioinforma. Oxf. Engl. *25*, 2078–2079.

Li, R., Yu, C., Li, Y., Lam, T.-W., Yiu, S.-M., Kristiansen, K., and Wang, J. (2009b). SOAP2: an improved ultrafast tool for short read alignment. Bioinforma. Oxf. Engl. *25*, 1966–1967.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., and Law, M. (2012). Comparison of Next-Generation Sequencing Systems. BioMed Res. Int. *2012*.

Love, A.J., Geri, C., Laird, J., Carr, C., Yun, B.-W., Loake, G.J., Tada, Y., Sadanandom, A., and Milner, J.J. (2012). Cauliflower mosaic virus protein P6 inhibits signaling responses to salicylic acid and regulates innate immunity. PloS One *7*, e47535.

Lu, Y., Gan, Q., Chi, X., and Qin, S. (2008). Roles of microRNA in plant defense and virus offense interaction. Plant Cell Rep. *27*, 1571–1579.

Lutz, L., Raikhy, G., and Leisner, S.M. (2012). Cauliflower mosaic virus major inclusion body protein interacts with the aphid transmission factor, the virion-associated protein, and gene VII product. Virus Res. *170*, 150–153.

Mansilla, C., Sánchez, F., Padgett, H.S., Pogue, G.P., and Ponz, F. (2009). Chimeras between oilseed rape mosaic virus and tobacco mosaic virus highlight the relevant role of the tobamoviral RdRp as pathogenicity determinant in several hosts. Mol. Plant Pathol. *10*, 59–68.

Marmey, P., Bothner, B., Jacquot, E., de Kochko, A., Ong, C.A., Yot, P., Siuzdak, G., Beachy, R.N., and Fauquet, C.M. (1999). Rice tungro bacilliform virus open reading frame 3 encodes a single 37-kDa coat protein. Virology *253*, 319–326.

Martelli, G.P., Adams, M.J., Kreuze, J.F., and Dolja, V.V. (2007). Family Flexiviridae: A Case Study in Virion and Genome Plasticity. Annu. Rev. Phytopathol. *45*, 73–100.

Matthews, R.E.F., and Hull, R. (2002). Matthews' Plant Virology (Academic Press).

Mauck, K., Bosque-Pérez, N.A., Eigenbrode, S.D., De Moraes, C.M., and Mescher, M.C. (2012). Transmission mechanisms shape pathogen effects on host–vector interactions: evidence from plant viruses. Funct. Ecol. *26*, 1162–1175.

Mayer, P., Farinelli, L., and Kawashima, E.H. (2013). Method of nucleic acid amplification.

Metzker, M.L. (2010). Sequencing technologies - the next generation. Nat. Rev. Genet. *11*, 31–46.

Meyer, R.S., DuVal, A.E., and Jensen, H.R. (2012). Patterns and processes in crop domestication: an historical review and quantitative analysis of 203 global food crops. New Phytol. *196*, 29–48.

Miller, J.R., Koren, S., and Sutton, G. (2010). Assembly Algorithms for Next-Generation Sequencing Data. Genomics *95*, 315–327.

Montgomery, T.A., Howell, M.D., Cuperus, J.T., Li, D., Hansen, J.E., Alexander, A.L., Chapman, E.J., Fahlgren, N., Allen, E., and Carrington, J.C. (2008). Specificity of ARGONAUTE7-miR390 interaction and dual functionality in TAS3 trans-acting siRNA formation. Cell *133*, 128–141.

Namiki, T., Hachiya, T., Tanaka, H., and Sakakibara, Y. (2012). MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. Nucleic Acids Res. *40*, e155.

Navarro, B., Gisel, A., Rodio, M.-E., Delgado, S., Flores, R., and Di Serio, F. (2012). Viroids: how to infect a host and cause disease without encoding proteins. Biochimie *94*, 1474–1480.

Nawaz-ul-Rehman, M.S., and Fauquet, C.M. (2009). Evolution of geminiviruses and their satellites. FEBS Lett. *583*, 1825–1832.

Ouyang, S., Zhu, W., Hamilton, J., Lin, H., Campbell, M., Childs, K., Thibaud-Nissen, F., Malek, R.L., Lee, Y., Zheng, L., et al. (2007). The TIGR Rice Genome Annotation Resource: improvements and new features. Nucleic Acids Res. *35*, D883–D887.

Patil, B.L., and Fauquet, C.M. (2009). Cassava mosaic geminiviruses: actual knowledge and perspectives. Mol. Plant Pathol. *10*, 685–701.

Plisson, C., Uzest, M., Drucker, M., Froissart, R., Dumas, C., Conway, J., Thomas, D., Blanc, S., and Bron, P. (2005). Structure of the Mature P3-virus Particle Complex of Cauliflower Mosaic Virus Revealed by Cryo-electron Microscopy. J. Mol. Biol. *346*, 267–277.

Pooggin, M.M. (2013). How can plant DNA viruses evade siRNA-directed DNA methylation and silencing? Int. J. Mol. Sci. *14*, 15233–15259.

Pooggin, M.M., Rajeswaran, R., Schepetilnikov, M.V., and Ryabova, L.A. (2012). Short ORF-dependent ribosome shunting operates in an RNA picorna-like virus and a DNA pararetrovirus that cause rice tungro disease. PLoS Pathog. *8*, e1002568.

Poojari, S., Alabi, O.J., Fofanov, V.Y., and Naidu, R.A. (2013). A leafhopper-transmissible DNA virus with novel evolutionary lineage in the family geminiviridae implicated in grapevine redleaf disease by next-generation sequencing. PloS One *8*, e64194.

Project, I.R.G.S. (2005). The map-based sequence of the rice genome. Nature *436*, 793–800.

Rajeswaran, R., Seguin, J., Chabannes, M., Duroy, P.-O., Laboureau, N., Farinelli, L., Iskra-Caruana, M.-L., and Pooggin, M.M. (2014a). Evasion of siRNA-directed antiviral silencing in Musa acuminata persistently infected with six distinct banana streak pararetroviruses. J. Virol.

Rajeswaran, R., Golyaev, V., Seguin, J., Zvereva, A., Farinelli, L., and Pooggin, M. (2014b). Interactions of rice tungro bacilliform pararetrovirus and its protein P4 with plant RNA silencing machinery. Mol. Plant-Microbe Interact. MPMI.

Rogers, K., and Chen, X. (2012). microRNA Biogenesis and Turnover in Plants. Cold Spring Harb. Symp. Quant. Biol. *77*, 183–194.

Rohozková, J., and Navrátil, M. (2011). P1 peptidase--a mysterious protein of family Potyviridae. J. Biosci. *36*, 189–200.

Roossinck, M.J. (2011). The big unknown: plant virus biodiversity. Curr. Opin. Virol. *1*, 63–67.

Roossinck, M.J., Sabanadzovic, S., Okada, R., and Valverde, R.A. (2011). The remarkable evolutionary history of endornaviruses. J. Gen. Virol. *92*, 2674–2678.

Sanfaçon, H., Wellink, J., Le Gall, O., Karasev, A., van der Vlugt, R., and Wetzel, T. (2009). Secoviridae: a proposed family of plant viruses within the order Picornavirales that combines the families Sequiviridae and Comoviridae, the unassigned genera Cheravirus and Sadwavirus, and the proposed genus Torradovirus. Arch. Virol. *154*, 899–907.

Saunders, K., Salim, N., Mali, V.R., Malathi, V.G., Briddon, R., Markham, P.G., and Stanley, J.

(2002). Characterisation of Sri Lankan Cassava Mosaic Virus and Indian Cassava Mosaic Virus: Evidence for Acquisition of a DNA B Component by a Monopartite Begomovirus. Virology *293*, 63–74.

Scholthof, K.-B.G., Adkins, S., Czosnek, H., Palukaitis, P., Jacquot, E., Hohn, T., Hohn, B., Saunders, K., Candresse, T., Ahlquist, P., et al. (2011). Top 10 plant viruses in molecular plant pathology. Mol. Plant Pathol. *12*, 938–954.

Schulz, M.H., Zerbino, D.R., Vingron, M., and Birney, E. (2012). Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics.

Seguin, J., Rajeswaran, R., Malpica-López, N., Martin, R.R., Kasschau, K., Dolja, V.V., Otten, P., Farinelli, L., and Pooggin, M.M. (2014a). De Novo Reconstruction of Consensus Master Genomes of Plant RNA and DNA Viruses from siRNAs. PloS One *9*, e88513.

Seguin, J., Otten, P., Baerlocher, L., Farinelli, L., and Pooggin, M.M. (2014b). 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.

Shimura, H., Pantaleo, V., Ishihara, T., Myojo, N., Inaba, J., Sueda, K., Burgyán, J., and Masuta, C. (2011). A viral satellite RNA induces yellow symptoms on tobacco by targeting a gene involved in chlorophyll biosynthesis using the RNA silencing machinery. PLoS Pathog. *7*, e1002021.

Song, X., Li, P., Zhai, J., Zhou, M., Ma, L., Liu, B., Jeong, D.-H., Nakano, M., Cao, S., Liu, C., et al. (2012). Roles of DCL4 and DCL3b in rice phased small RNA biogenesis. Plant J. *69*, 462–474.

Strange, R.N., and Scott, P.R. (2005). Plant Disease: A Threat to Global Food Security. Annu. Rev. Phytopathol. *43*, 83–116.

The Arabidopsis Initiative (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature *408*, 796–815.

Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. (2012). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinform.

Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinform. *14*, 178–192.

Trejo-Saavedra, D.L., Vielle-Calzada, J.P., and Rivera-Bustamante, R.F. (2009). The infective cycle of Cabbage leaf curl virus (CaLCuV) is affected by CRUMPLED LEAF (CRL) gene in Arabidopsis thaliana. Virol. J. *6*, 169.

Van Regenmortel, M.H. (1999). The antigenicity of tobacco mosaic virus. Philos. Trans. R. Soc. B Biol. Sci. *354*, 559–568.

Voinnet, O. (2008). Use, tolerance and avoidance of amplified RNA silencing by plants. Trends Plant Sci. *13*, 317–328.

Vunsh, R., Rosner, A., and Stein, A. (1990). The use of the polymerase chain reaction (PCR) for the detection of bean yellow mosaic virus in gladiolus. Ann. Appl. Biol. *117*, 561–569.

Wu, Y.-W., and Ye, Y. (2011). A Novel Abundance-Based Algorithm for Binning Metagenomic

Sequences Using l-tuples. J. Comput. Biol. *18*, 523–534.

Wu, L., Zhang, Q., Zhou, H., Ni, F., Wu, X., and Qi, Y. (2009). Rice MicroRNA effector complexes and targets. Plant Cell *21*, 3421–3435.

Wu, L., Zhou, H., Zhang, Q., Zhang, J., Ni, F., Liu, C., and Qi, Y. (2010). DNA methylation mediated by a microRNA pathway. Mol. Cell *38*, 465–475.

Zerbino, D.R., and Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. *18*, 821–829.

Zhang, W., Chen, J., Yang, Y., Tang, Y., Shang, J., and Shen, B. (2011). A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies. PloS One *6*, e17915.

Zhou, X., Ren, L., Meng, Q., Li, Y., Yu, Y., and Yu, J. (2010). The next-generation sequencing technology and application. Protein Cell *1*, 520–536.

Annex: (Aregger et al., 2012)

Primary and Secondary siRNAs in Geminivirus-induced Gene Silencing

Michael Aregger, Basanta K. Borah, Jonathan Seguin, Rajendran Rajeswaran, Ekaterina G. Gubaeva, Anna S. Zvereva, David Windels, Franck Vazquez, Todd Blevins, Laurent Farinelli, Mikhail M. Pooggin

PLOS Pathogens (2012), Vol.8, Issue 9

OPEN ACCESS Freely available online

PLOS | PATHOGENS

Primary and Secondary siRNAs in Geminivirus-induced Gene Silencing

Michael Aregger[1☯¤], Basanta K. Borah[1☯], Jonathan Seguin[1,2☯], Rajendran Rajeswaran[1], Ekaterina G. Gubaeva[1], Anna S. Zvereva[1], David Windels[1], Franck Vazquez[1], Todd Blevins[3], Laurent Farinelli[2], Mikhail M. Pooggin[1]*

1 Institute of Botany, University of Basel, Basel, Switzerland, **2** Fasteris SA, Plan-les-Ouates, Switzerland, **3** Biology Department, Indiana University, Bloomington, Indiana, United States of America

Abstract

In plants, RNA silencing-based antiviral defense is mediated by Dicer-like (DCL) proteins producing short interfering (si)RNAs. In *Arabidopsis* infected with the bipartite circular DNA geminivirus *Cabbage leaf curl virus* (CaLCuV), four distinct DCLs produce 21, 22 and 24 nt viral siRNAs. Using deep sequencing and blot hybridization, we found that viral siRNAs of each size-class densely cover the entire viral genome sequences in both polarities, but highly abundant siRNAs correspond primarily to the leftward and rightward transcription units. Double-stranded RNA precursors of viral siRNAs can potentially be generated by host RDR-dependent RNA polymerase (RDR). However, genetic evidence revealed that CaLCuV siRNA biogenesis does not require RDR1, RDR2, or RDR6. By contrast, CaLCuV derivatives engineered to target 30 nt sequences of a *GFP* transgene by primary viral siRNAs trigger RDR6-dependent production of secondary siRNAs. Viral siRNAs targeting upstream of the *GFP* stop codon induce secondary siRNAs almost exclusively from sequences downstream of the target site. Conversely, viral siRNAs targeting the *GFP* 3′-untranslated region (UTR) induce secondary siRNAs mostly upstream of the target site. RDR6-dependent siRNA production is not necessary for robust *GFP* silencing, except when viral siRNAs targeted *GFP* 5′-UTR. Furthermore, viral siRNAs targeting the transgene enhancer region cause *GFP* silencing without secondary siRNA production. We conclude that the majority of viral siRNAs accumulating during geminiviral infection are RDR1/2/6-independent primary siRNAs. Double-stranded RNA precursors of these siRNAs are likely generated by bidirectional readthrough transcription of circular viral DNA by RNA polymerase II. Unlike transgenic mRNA, geminiviral mRNAs appear to be poor templates for RDR-dependent production of secondary siRNAs.

Citation: Aregger M, Borah BK, Seguin J, Rajeswaran R, Gubaeva EG, et al. (2012) Primary and Secondary siRNAs in Geminivirus-induced Gene Silencing. PLoS Pathog 8(9): e1002941. doi:10.1371/journal.ppat.1002941

Editor: Shou-Wei Ding, University of California Riverside, United States of America

Received April 12, 2012; **Accepted** August 18, 2012; **Published** September 27, 2012

Funding: The work was financed through Swiss National Science Foundation grants (31003A_127514 to MMP and 31003A_122469 to Thomas Hohn and MMP), European Cooperation in Science and Technology (COST) grant SER No. C09.0176 to LF and MMP, and European Commission grant (a Marie Curie fellowship PIIF-237493-SUPRA to RR). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: Mikhail.Pooggin@unibas.ch

¤ Current address: College of Life Sciences, University of Dundee, Dundee, United Kingdom

☯ These authors contributed equally to this work.

Introduction

RNA silencing directed by miRNAs, short interfering (si)RNAs and PIWI-interacting RNAs is involved in regulation of gene expression and chromatin states and in defense against invasive nucleic acids such as transposons, transgenes and viruses [1–3]. Virus-infected plants accumulate high levels of viral siRNAs (vsRNAs) of three major size-classes: 21-nt, 22-nt and 24-nt [4,5]. In *Arabidopsis thaliana* infected with DNA viruses, all four Dicer-like (DCL) enzymes are involved in processing of vsRNA duplexes from longer double-stranded RNA (dsRNA) precursors: DCL4 and DCL1 generate 21-nt class, DCL2 generates 22-nt class and DCL3 generates 24-nt class; 21-nt and 24-nt vsRNAs accumulate at higher levels than 22-nt vsRNAs [6–8]. By contrast, in RNA virus-infected *Arabidopsis*, DCL4-dependent 21-nt vsRNAs and/or DCL2-dependent 22-nt vsRNAs are the most abundant species, whereas DCL3-dependent 24-nt vsRNAs accumulate at much lower levels [7,9,10]. This reflects the difference in viral life cycles: DNA viruses transcribe their genomes in the nucleus, whereas RNA viruses are generally restricted to the cytoplasm. Likewise, plant endogenous genes and transgenes that undergo transcriptional silencing spawn predominantly DCL3-dependent 24-nt siRNAs, whereas those that undergo post-transcriptional silencing spawn predominantly DCL4-dependent 21-nt siRNAs and, in certain cases, DCL2-dependent siRNAs [1,11,12].

In endogenous and transgene-induced silencing pathways, dsRNA precursors of siRNAs can be generated by RNA-dependent RNA-polymerase (RDR). The *Arabidopsis thaliana* genome encodes six RDRs, three of which have been implicated in siRNA biogenesis [13]. RDR2 is required for biogenesis of 24-nt heterochromatic siRNAs (hcsiRNAs) mainly originating from repetitive DNA loci including transposons. RDR6 is required for biogenesis of *trans*-acting siRNAs (tasiRNAs), natural antisense transcript siRNAs and siRNAs derived from posttranscriptionally-silenced transgenes [1]. RDR6 is also involved in production of secondary siRNAs from some protein-coding genes targeted by

Author Summary

RNA silencing directed by small RNAs (sRNAs) regulates gene expression and mediates defense against invasive nucleic acids such as transposons, transgenes and viruses. In plants and some animals, RNA-dependent RNA polymerase (RDR) generates precursors of secondary sRNAs that reinforce silencing. Most plant mRNAs silenced by miRNAs or primary siRNAs do not spawn secondary siRNAs, suggesting that they may have evolved to be poor templates for RDR. By contrast, silenced transgenes often produce RDR-dependent secondary siRNAs. Here we demonstrate that massive production of 21, 22 and 24 nt viral siRNAs in DNA geminivirus-infected *Arabidopsis* does not require the functional RDRs RDR1, RDR2, or RDR6. Deep sequencing analysis indicates that dsRNA precursors of these primary viral siRNAs are likely generated by RNA polymerase II-mediated bidirectional readthrough transcription on the circular viral DNA. Primary viral siRNAs engineered to target a *GFP* transgene trigger robust, RDR6-dependent production of secondary siRNAs, indicating that geminivirus infection does not suppress RDR6 activity. We conclude that geminiviral mRNAs, which can potentially be cleaved by primary viral siRNAs, are resistant to RDR-dependent amplification of secondary siRNAs. We speculate that, like most plant mRNAs, geminiviral mRNAs may have evolved to evade RDR activity.

miRNAs [14,15]. RDR1 has so far been implicated in viral siRNA biogenesis (see below) and its function in endogenous or transgene-induced silencing is not known. Presumptive single-stranded RNA templates for RDR2 are produced by plant-specific RNA polymerases Pol IV and/or Pol V, but little is known about Pol IV and Pol V transcripts and RDR2-dependent dsRNAs [16]. dsRNA precursors of tasiRNAs originate from Pol II transcripts of *TAS* genes, which are cleaved by a miRNA::Argonaute (AGO) protein complex [17–20]. Either the 3′ cleavage product or the 5′ cleavage product is converted by RDR6 to dsRNA: RDR6 recruitment to only one of the two cleavage products is determined by 22-nt size of the initiator miRNA produced from a bulged hairpin precursor [21–23] or a second binding site of the miRNA::AGO complex [17,19], respectively.

The possible role of RDRs in vsRNA biogenesis has been extensively studied using *A. thaliana* single, double and triple null mutants for RDR1, RDR2 and RDR6 [8,24–28]. These studies produced rather conflicting results, but in many cases, wild type viruses were shown to predominantly spawn RDR-independent vsRNAs [29]. However, mutant RNA viruses with deletion or point mutation in the viral silencing suppressor gene spawn RDR6- and/or RDR1-dependent vsRNAs [26–28]. As a consequence the suppressor-deficient RNA viruses could establish systemic infection only on *A. thaliana* mutant plants lacking RDR6 and/or RDR1 activity. Nevertheless, suppressor-deficient RNA viruses spawn substantial amounts of RDR-independent vsRNAs. Thus, one of the major precursors of RNA virus-derived vsRNAs is likely a double-stranded replicative intermediate, transiently produced by viral RNA-dependent RNA-polymerase (vRdRP). Primary vsRNAs generated from such precursors may trigger RDR-dependent production of secondary siRNAs.

Plant DNA viruses do not encode a vRdRP. However, the biogenesis of DNA virus-derived vsRNAs does not appear to involve host RDRs. Thus, *Cauliflower mosaic virus* (CaMV)-derived vsRNAs of all major classes accumulate at comparable high levels in *A. thaliana* wild-type and *rdr1 rdr2 rdr6* triple mutant plants and their long dsRNA precursors are likely generated by Pol II [8]. The lack of RDR-dependent vsRNAs can be explained by the ability of a CaMV silencing suppressor protein to interfere with DCL4-mediated processing of dsRNAs produced by RDR6 [30,31]. Silencing suppressor proteins of DNA geminiviruses have not been reported to interfere with RDR activity or DCL-mediated processing of RDR-dependent dsRNAs. In *A. thaliana* null mutants for Pol IV, RDR2, or RDR6 activity, the biogenesis of vsRNAs from *Cabbage leaf curl virus* (CaLCuV; a member of genus *Begomovirus* of the family *Geminiviridae*) was not affected, suggesting that RDR2 and RDR6 are not involved in production of dsRNA precursors of vsRNAs [7]. However, involvement of RDR1 in this process or possible redundancy in activities of distinct RDRs were not investigated so far.

Geminiviruses encapsidate circular single-stranded (ss)DNA of ca. 2.5-to-2.7 kb in geminate virions and accumulate in the nucleus as multiple circular dsDNA minichromosomes. The minichromosomes are both the intermediates of rolling circle replication and the templates for Pol II-mediated bidirectional transcription [32]. Like many members of the genus *Begomovirus*, CaLCuV has a bipartite genome comprising 2.6 kb DNA-A and 2.5 kb DNA-B [33]. The DNA-A encodes proteins involved in replication (AC1 and AC3), transcription (AC2) and encapsidation (AV1), while the DNA-B encodes BC1 and BV1 proteins with movement functions. A large intergenic region on DNA-A and DNA-B contains a 192 bp common region of nearly identical sequence with the origin of replication and bidirectional promoter elements. By analogy with other begomoviruses [34], the bidirectional promoter is expected to drive Pol II transcription of the leftward (*AC1/AC4/AC2/AC3* and *BC1*) and rightward (*AV1* and *BV1*) genes. In addition, a monodirectional promoter is expected to drive Pol II transcription of a short *AC2/AC3* transcript, which is co-terminal with the long *AC1/AC4/AC2/AC3* transcript. On both DNAs, the leftward and rightward transcription is terminated by poly(A) signals located in a close vicinity on the virion (sense) and complementary (antisense) strands, respectively. In CaLCuV DNA-A, this juxtaposition of the poly(A) signals creates a ca. 25-nt overlap of the sense and antisense transcripts. Such overlap was proposed to form a dsRNA precursor of primary vsRNAs [35], which may initiate RDR-dependent production of vsRNAs from other regions of the viral transcripts.

Such phenomenon of transitivity has been described for posttranscriptional and transcriptional silencing of a transgene targeted by vsRNAs (virus-induced gene silencing; VIGS) or by primary siRNAs derived from an inverted-repeat transgene. In these cases, RDR6- or RDR2-dependent production of secondary siRNAs outside of the target region was detected, respectively [36,37]. Notably, posttranscriptional silencing of endogenous plant genes by virus- or transgene-derived primary siRNAs was not associated with secondary siRNA production [36,38,39], suggesting that endogenous mRNAs are not good templates for RDRs.

In this study, we used Illumina deep sequencing of short RNAs, combined with blot hybridization and genetic analysis, to investigate the biogenesis of primary and secondary siRNAs. To this end, *Arabidopsis* wild-type, *RDR*-mutant and transgenic plants were infected with CaLCuV or its derivatives carrying fragments of an endogenous gene or a transgene. We found that, like most endogenous plant mRNAs, viral mRNAs are not prone to transitivity: the majority of vsRNAs are RDR1-, RDR2- and RDR6-independent primary siRNAs. By contrast, a transgene mRNA targeted by primary vsRNAs is subject to RDR6-dependent production of secondary siRNAs. We also found that silencing of the transgene driven by a CaMV 35S promoter can be triggered by primary vsRNAs targeting an enhancer (but not core

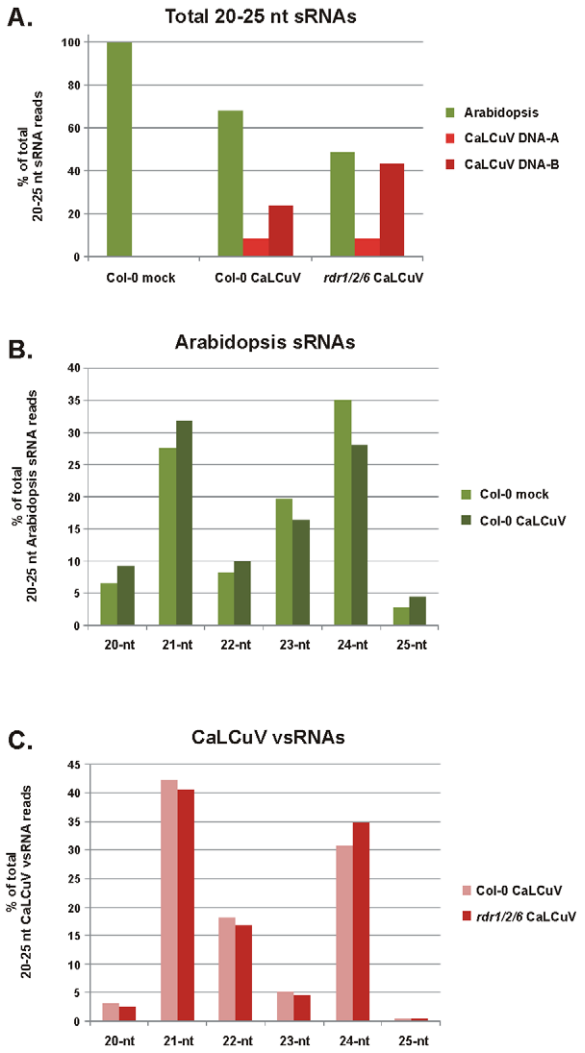


Figure 1. Illumina deep-sequencing of sRNAs from mock-inoculated and CaLCuV-infected *Arabidopsis* wild-type (Col-0) and *rdr1/2/6* triple mutant plants. The graphs show the percentages of *Arabidopsis* and vsRNAs in the pool of 20–25 nt reads mapped to the *Arabidopsis* and CaLCuV DNA-A and DNA-B genomes with zero mismatches (**A**), of each size-class of 20–25 nt host sRNA reads mapped to the *Arabidopsis* genome with zero mismatches (**B**), and of each size-class of 20–25 nt vsRNA reads mapped to the CaLCuV DNA-A and DNA-B with zero mismatches (**C**).
doi:10.1371/journal.ppat.1002941.g001

promoter) region and this, presumably transcriptional, silencing was not associated with accumulation of secondary siRNAs.

Results/Discussion

21, 22 and 24 nt vsRNAs accumulate at high levels in CaLCuV-infected *Arabidopsis*

To analyze begomovirus interactions with the host small RNA (sRNA)-generating silencing pathways, we deep-sequenced sRNA populations from mock-inoculated and CaLCuV-infected *A. thaliana* wild-type (Col-0) plants and CaLCuV-infected *rdr1 rdr2 rdr6* triple null mutant plants (*rdr1/2/6* in Col-0 background; [8]). The protocol was designed to sequence short RNAs with 5′-phosphate and 3′-hydroxyl groups, which include DCL products. Samples of total RNA extracted from pools of three plants were processed in parallel and the resulting cDNA libraries sequenced in one channel of an Illumina Genome Analyzer, thus allowing quantitative comparison of changes in the profile of host sRNAs upon virus infection and the profile of vsRNAs in wild-type versus mutant plants.

A total number of reads in the high-coverage libraries was ranging from 9.3 to 10.4 million, of which 7.3 million ('Col-0 mock'), 5.3 million ('Col-0 CaLCuV') and 5.0 million ('*rdr1/2/6* CaLCuV') of 20–25 nt reads mapped to the *Arabidopsis thaliana* Col-0 or CaLCuV genomes with zero mismatches (Table S1A). Two additional low-coverage libraries with 0.45 million ('Col-0 mock*') and 0.43 million ('Col-0 CaLCuV*') of 20–25 nt reads with zero mismatches (Table S1A) were obtained in an independent experiment.

In mock-inoculated plants, most of the 20–25 nt sRNAs mapped to the *A. thaliana* genome (Figure 1A; Table S1A). The 24-nt and 21-nt classes were predominant (35% and 28%, respectively), whereas other size-classes were less abundant (23-nt – 19%; 22-nt – 8%; 20-nt – 7%; 25-nt – 3%) (Figure 1B). This is consistent with the previous studies showing that 24-nt hcsiRNAs and 21-nt miRNAs are the most abundant sRNA classes in *A. thaliana* [40,41]. Upon CaLCuV infection, the host sRNA profile was slightly altered in that the 21-nt class became the largest (32%) and the 24-nt class the second largest (28%) (Figure 1B; Table S1A). A similar shift in the host sRNA profile was also detected in the low coverage experiment (Table S1A). By contrast, *A. thaliana* infection with the pararetrovirus CaMV results in overaccumulation of 24-nt host sRNAs [8]. The biological significance of the opposite effects of geminivirus and pararetrovirus infections on host sRNAs remains to be investigated.

In CaLCuV-infected Col-0 plants, a large fraction of 20–25 nt reads mapped to the virus genome with zero mismatches (ca. 32% and 62% in the high- and low-coverage libraries, respectively; Figure 1A and Table S1A). Notably, the viral DNA-B was the major source of vsRNAs (70% and 85% of 20–25 nt viral reads, respectively; Table S1A). On both DNA-A and DNA-B, vsRNA reads were almost equally distributed between the virion and complementary strands (Table S1A; Figures 2 and S1). Similar to the host sRNAs in infected plants, 21-nt and 24-nt vsRNAs represent the first (42%) and the second (31%) largest fractions of 20–25 nt viral reads, respectively. But unlike the host sRNAs, 22-nt viral reads represent the third largest fraction (18%), while 20-nt, 23-nt and 25-nt classes are significantly underrepresented (Figure 1C). This size-class profile of CaLCuV vsRNAs agrees with our blot hybridization analysis using short probes and confirms the involvement of distinct DCLs in vsRNA biogenesis (Figure S2; [7]).

Interestingly, the host sRNAs of 21-nt and 24-nt classes exhibit a strong bias to 5′-terminal uridine (5′U; 69%) and 5′-terminal adenosine (5′A; 52%), respectively (Table S1A), owing to the preferential association of miRNAs with AGO1 and hcsiRNAs with AGO4 [17,42–44]. By contrast, vsRNAs of 21-nt and 24-nt classes are less strongly enriched in 5′U (46%) and 5′A (32%), respectively, and the second most dominant nucleotide is 5′A for 21-nt class (25%) and 5′U for 24-nt class (32%) (Table S1A). Both the diversity in nucleotide composition and size of CaLCuV vsRNAs and the lack of any strong 5′-nucleotide bias imply the involvement of multiple AGOs in sorting vsRNAs.

vsRNA species densely tile along the entire circular viral DNAs and accumulate at high levels in several large hotspot regions

Inspection of single-nucleotide resolution maps of 20–25 nt vsRNAs revealed that unique vsRNA species of each major class (21-nt, 22-nt and 24-nt) cover the entire genome of CaLCuV in

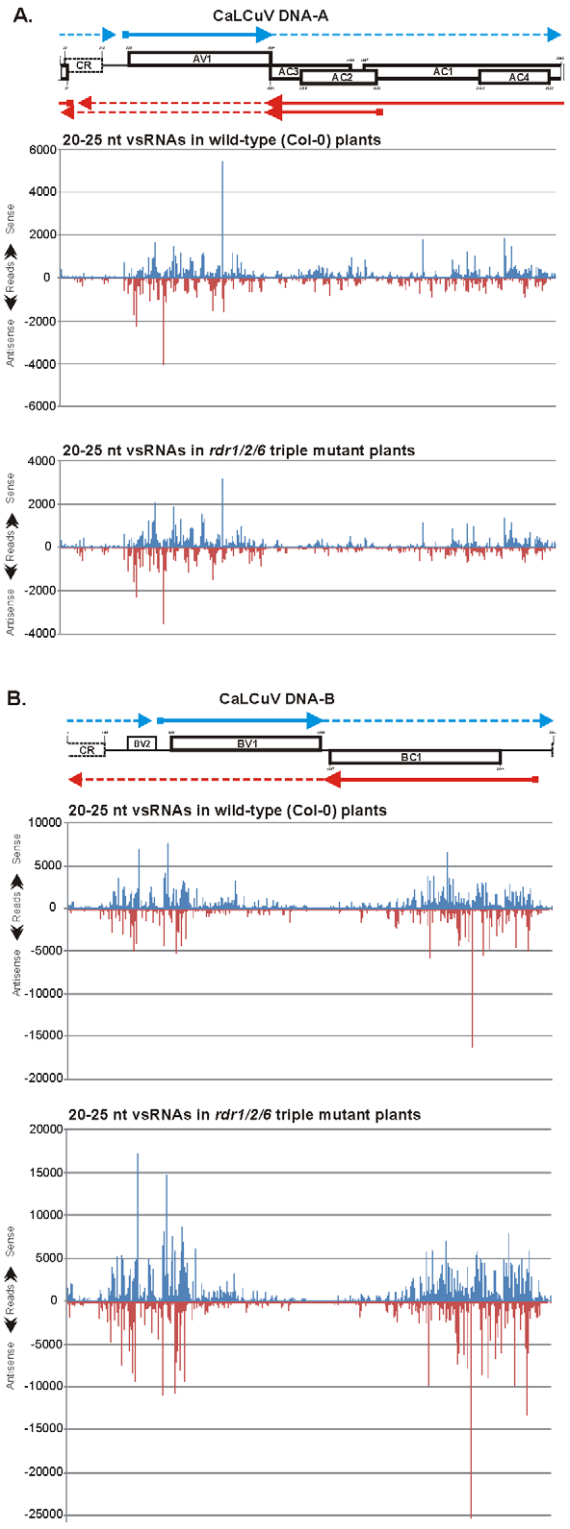


Figure 2. Maps of vsRNAs from CaLCuV-infected wild type (Col-0) and *rdr1/2/6* triple mutant plants at single-nucleotide resolution. The graphs plot the number of 20–25 nt vsRNA reads at each nucleotide position of the 2583 bp DNA-A (**A**) and the 2513 bp DNA-B (**B**); Bars above the axis represent sense reads starting at each respective position; those below represent antisense reads ending at the respective position (Tables S2 and S3). The genome organizations of DNA-A and DNA-B are shown schematically above the graphs, with leftward (AC1, AC4, AC2, AC3 and BC1) and rightward (AV1 and BV1) ORFs and common region (CR) indicated. The predicted rightward and rightward mRNAs are shown as respectively blue and red solid lines with arrowheads. Potential readthrough transcripts are shown as dotted thin lines.
doi:10.1371/journal.ppat.1002941.g002

both sense and antisense polarity as dense tiling arrays without gaps on the circular sequences of 2583 bp DNA-A and 2513 bp DNA-B (Tables S2 and S3). Hence, dsRNA precursors of vsRNAs of each class should cover the entire circular viral DNAs. However, the relative abundance of vsRNAs varies drastically: several large regions of DNA-A and DNA-B are densely covered in both polarities with vsRNA hotspots (defined here arbitrarily as short sequence segments spawning several vsRNA species with more than 300 reads each) (Figure 2 and Figure S1). This implies the existence of several overlapping dsRNA precursors that accumulate at high and low levels. Interestingly, vsRNA hotspots on both virion and complementary strands are interrupted with short sequences that spawn vsRNAs of lower abundance (Figure 2 and Figure S1; Table S2 and Table S3). This implies differential stability of vsRNA duplexes processed consequently from ends of long dsRNA precursors or, alternatively, preferential internal excisions of vsRNA duplexes from certain regions of a long dsRNA. We also found that most vsRNA hotspots contain all the three major size-classes of vsRNAs (Figure S1; Table S2 and Table S3), indicating that same dsRNA precursors are processed by different DCLs. This conclusion is consistent with our genetic analysis coupled with blot-hybridization of DNA virus-derived sRNAs [6,7] (Figure S2) and sRNA deep-sequencing studies of other viruses [8,45–48].

In DNA-A, the most abundant vsRNAs of both sense and antisense polarities, which include those with more than 1000 reads, originate from the *AV1* ORF (Figure 2A and Figure S1A). The left border of this vsRNA hotspot region is at position 331 (Table S2), where the transcription start site can be predicted, i.e. at an optimal distance downstream of the TATA box (TATATAA at positions 228–305) and 9 nts upstream of the *AV1* start codon (339–341). The right border of this vsRNA hotspot is at around position 1060 (Table S2), i.e. just upstream of the *AV1* stop codon (1092–1094). After a short gap of 55 bp (1061–1116) lacking highly abundant vsRNAs, a large region spanning all the leftward ORFs is also covered with vsRNA hotspots, albeit at lower density than in the *AV1* region. In this region, the most abundant vsRNAs originate from the large portion of the *AC1* ORF including the nested AC4 ORF and less abundant vsRNAs from the *AC2* ORF (Figure 2A; Table S2). Notably, the 25 nt region (1089–1113), in which the rightward (*AV1*) and the leftward (*AC1/AC4/AC2/AC3* and *AC2/AC3*) viral mRNAs are expected to overlap and potentially form a dsRNA substrate for DCL, is not a vsRNA hotspot. Likewise, the 240 bp intergenic region between the predicted leftward and rightward transcription start sites (at positions 93 and 331, respectively), which contains the bidirectional promoter elements and overlaps the common region (22–213), is also devoid of vsRNA hotspots: it has only two islands covered with vsRNAs of 100–250 reads. Furthermore, the promoter region in front of the predicted transcription start site of *AC2/AC3* mRNA (position 1651, downstream of TATATAA at 1683–1677) does not contain any prominent vsRNA hotspots (Figure 2A and Figure S1A; Table S2). Taken together, the promoter and terminator regions of CaLCuV DNA-A are devoid of highly abundant vsRNAs. Thus, the virus may have evolved a mechanism to evade transcriptional silencing which could potentially be directed by vsRNAs.

In DNA-B, two large regions are covered with extreme hotspots containing multiple vsRNA species with more than 1000 reads on

both sense and antisense strands. The first is located downstream of the common region and it spans a large portion of the *BV1* ORF. The second is located upstream of the common region and it spans a large portion of the *BC1* ORF (Figure 2B and Figure S1B; Table S3). Like in DNA-A, the terminator region of rightward (*BV1*) and leftward (*BC1*) genes is devoid of vsRNA hotspots. Note that the DNA-B poly(A) signals AATAAA are located at positions 1305–1310 and 1356–1361 of the virion and complementary stands, respectively, and therefore the *BV1* and *BC1* mRNAs are not expected to overlap. A predicted BC1 promoter region with the TATA-box at positions 2471–2463 (TATATAA) is devoid of vsRNA hotspots and the border of the vsRNA hotspot region corresponds to the predicted transcription start site at 2439. Thus, *BC1* mRNA can form one of the strands of a vsRNA precursor. In contrast, a predicted *BV1* promoter region with the TATA-box at position 442–447 (TATATAA) is covered with vsRNA hotspots on both strands. This suggests that the region upstream of the *BV1* ORF might be actively transcribed. Interestingly, it contains an ORF at positions 319 to 471 (Figure 2B). Such active transcription could in turn lead to production of abundant vsRNAs that can potentially direct transcriptional silencing of the *BV1* promoter. This may represent either a host antiviral defense or a viral strategy of gene regulation.

Based on close inspection of cold versus hot spots of viral siRNAs, AU-rich sequences can generally be considered as a poor source of siRNAs, possibly owing to relatively low stability of AU-rich siRNA duplexes processed by DCLs from long dsRNA precursors. Other features of RNA primary or secondary structure which might potentially influence siRNA biogenesis or stability remain to be further investigated.

vsRNA biogenesis is not affected drastically in plants lacking RDR1, RDR2 and RDR6

The *Arabidopsis* sRNA profile is drastically altered in *rdr1/2/6* triple mutant compared to wild-type plants: 24-nt and 23-nt classes are selectively and strongly reduced, mainly owing to the loss of RDR2-dependent hcsiRNAs [40]. Thus, 21-nt class becomes the most predominant, followed by 20-nt and 22-nt classes (Table S1A): these three classes are mainly populated with RDR-independent miRNAs, whereas RDR6-dependent tasiRNAs and secondary siRNAs are much less abundant [41]. By contrast, the CaLCuV vsRNA profile was only slightly altered in *rdr1/2/6* compared to wild-type (Figure 1C).

The overall accumulation level of 20–25 nt vsRNAs was higher in *rdr1/2/6* than wild-type plants. If normalized by the levels of 21-nt host sRNAs (1.22 million in 'Col-0 CaLCuV' versus 1.21 million in '*rdr1/2/6* CaLCuV'), this ca. 1.5-fold increase is mainly owing to higher accumulation of DNA-B vsRNAs of all the major classes (Table S1A; Figure 1A).

The single-nucleotide resolution maps of vsRNAs from Col-0 and *rdr1/2/6* are remarkably similar. The vsRNA hotspots occur in the same regions and the relative abundance of vsRNA species is very similar within most hotspots (Figure 2 and Figure S1; Table S2 and Table S3). For DNA-A, the levels of 20–25 nt vsRNAs derived from the *AC2* hotspot region are relatively lower in *rdr1/2/6* than in Col-0, whereas those derived from the *AV1* region are generally similar in *rdr1/2/6* and Col-0 (Figure 2A), with an exception of 24-nt vsRNAs that accumulate at relatively higher levels in *rdr1/2/6* (Figure S1A; Table S1A). For DNA-B, the levels of 20–25 nt vsRNAs in most hotspots are 1.5- to 2.5-fold higher in *rdr1/2/6* than in Col-0, with an exception of the middle part and the 3′ part of *BV1* ORF, in which vsRNA levels are generally similar in *rdr1/2/6* and Col-0 or, at some locations in the 3′ part, lower in *rdr1/2/6* (Figure 2B). No drastic difference in the relative abundance of vsRNA size-classes along the DNA-B sequence was observed (Figure S2B; Table S3).

Analysis of 5′-terminal nucleotides of vsRNAs revealed no substantial difference between Col-0 and *rdr1/2/6* (Table S1A), further supporting that vsRNA biogenesis is not drastically affected by null mutations in *RDR1*, *RDR2* and *RDR6*.

The above-described deep sequencing findings for vsRNA size-classes, relative abundance and distribution along the viral genome and RDR1/2/6-independence of vsRNA biogenesis were confirmed by blot hybridization analysis of sRNAs from CaLCuV-infected wild-type and *rdr1/2/6* mutant plants using several short probes specific to DNA-A or DNA-B (Figure S2 and Figure 3B). In addition, analysis of CaLCuV-infected *dcl1 dcl2 dcl3 dcl4* quadruple mutant plants (*dcl1/2/3/4*) confirmed our previous findings that the majority of vsRNAs are generated by four DCLs [7]. We further established that a mutant DCL1 protein produced from the *dcl1-9/caf1* allele in *dcl1/2/3/4* plants [8] appears to be capable of generating 21-nt vsRNA from dsRNA precursors derived from vsRNA hotspot regions of DNA-B (Figure S2). Likewise, a major fraction of 21-nt vsRNAs derived from the leader region of CaMV, which is an extreme hotspot of 21-24 nt vsRNA production, requires DCL1 for their biogenesis and residual accumulation of 21-nt vsRNAs was observed in *dcl1/2/3/4* [8].

Taken together, our findings indicate that CaLCuV vsRNA biogenesis does not require RDR1, RDR2, or RDR6. However, there appears to be a quantitative difference in relative abundance of dsRNA precursors derived from the vsRNA hotspot regions of DNA-A and DNA-B in wild-type versus *rdr1/2/6* plants.

Accumulation of viral long nucleic acids in wild type versus *rdr1/2/6* plants

To test if the observed differences in relative abundance of vsRNAs correlate with relative levels of viral transcripts and/or viral DNA, we measured the accumulation of viral long nucleic acids in wild-type and *rdr1/2/6* plants by RNA and DNA blot hybridization as well as real time PCR (Figure 3). The results of total RNA (Figure 3A) and polyadenylated mRNA (Figure 3D) analyses revealed that the relative accumulation of viral transcripts positively correlates the relative abundance of vsRNAs in the major hot spot regions. Indeed, AV1 mRNA, the most readily detectable viral transcript, accumulated at slightly higher levels in *rdr1/2/6* than wild type plants, whereas accumulation of the less abundant AC2/AC3 mRNA was slightly reduced in *rdr1/2/6*. This resembles the profile of DNA-A derived vsRNAs and its alteration in *rdr1/2/6*. Furthermore, accumulation of BC1 and BV1 polyadenylated mRNAs was increased ca. 1.2- and 1.4-fold, respectively, in *rdr1/2/6* compared to wild type plants, which correlates with slightly increased accumulation of DNA-B derived vsRNAs in *rdr1/2/6*. Notably, in addition to viral mRNAs, shorter viral transcripts also accumulate at high levels and appear as a smear on the total RNA blot (Figure 3A). These aberrant RNAs may represent degradation products of viral mRNAs or prematurely terminated viral transcripts. In the case of DNA-B, the aberrant RNAs appear to be much more abundant than BV1 and BC1 mRNAs, since the latter are barely detectable (Figure 3A). This correlates with much higher accumulation of vsRNAs from DNA-B than DNA-A (Figure 1A). The higher abundance of aberrant RNAs transcribed from DNA-B can be explained by higher accumulation of total DNA-B compared to total DNA-A as estimated by Southern (Figure 3C).

Real time PCR analysis (Figure 3D) revealed that total viral DNA accumulates at higher levels in *rdr1/2/6* compared to wild type plants (ca. 1.4- and 2-fold increase for DNA-A and DNA-B,

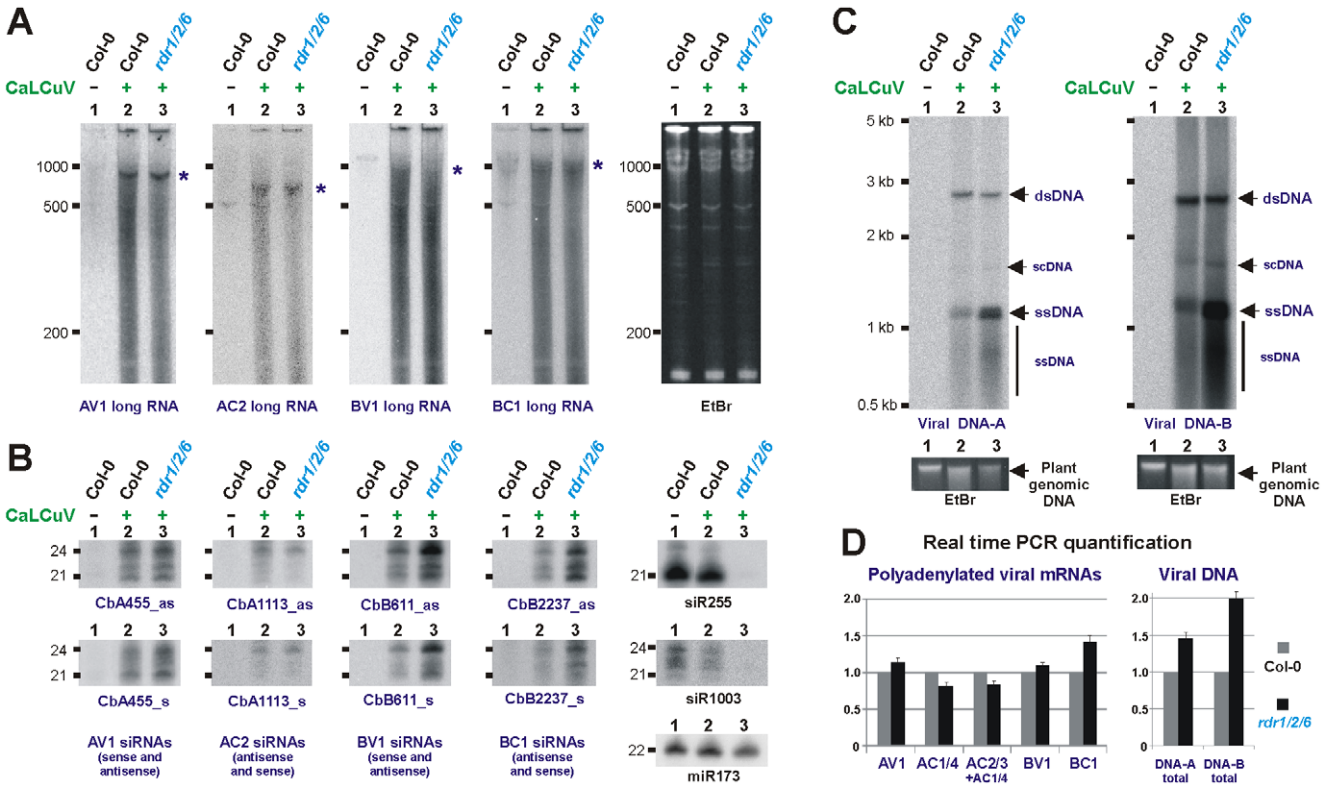


Figure 3. Accumulation of long viral nucleic acids and vsRNAs in wild type versus *rdr1/2/6* triple mutant plants. Total RNA and total DNA from CaLCuV-infected *Arabidopsis* wt (Col-0) and *rdr1/2/6* plants was analyzed by RNA blot hybridization using 5% (**A**) and 15% (**B**) PAGE and by Southern blot hybridization (**C**). The RNA blot membranes were successively hybridized with mixtures of DNA oligonucleotide probes complementary to respective viral mRNAs (for sequences, see Protocol S1) and, in the case of sRNA analysis, single DNA oligonucleotide probes specific for vsRNA of sense or antisense polarity and the endogenous *Arabidopsis* miRNA (22 nt miR173), tasiRNA (21 nt siR255) and hcsiRNA (24 nt siR1003). The Southern blot membranes were hybridized with long dsDNA probes specific for DNA-A or DNA-B. Positions of co-migrating forms of viral DNA including open-circular double-stranded (dsDNA), supercoiled (scDNA) and single-stranded (ssDNA) are indicated by arrows; the smear of shorter (than monomeric) ssDNA is also indicated. EtBr staining of total RNA (**A**) or plant genomic DNA (**C**) is shown as loading control. The size markers are indicated on each scan. Positions of viral mRNAs are indicated by asterisks. (**D**) Real time qPCR measurement of relative accumulation of viral polyadenylated mRNAs (left) and total viral DNAs A and B (right) in wild type versus *rdr/1/2/6* mutant plants. For each mRNA and each DNA, the accumulation level in the wild type sample is set to 1.
doi:10.1371/journal.ppat.1002941.g003

respectively). However, Southern blot hybridization analysis (Figure 3C) showed that this increase is mainly owing to increased accumulation of viral single-stranded DNA (ssDNA). By contrast, the levels of viral dsDNA, which serves as a template for both transcription and replication, are similar in wild type and *rdr1/2/6* plants. Thus, rolling circle and/or recombination-dependent replication mechanisms [32] produce increased levels of viral ssDNA (but not dsDNA) in the absence of RDR1, RDR2 and RDR6. This finding implicates an RDR activity in the regulation of geminiviral DNA replication. Interestingly, homologous recombination-dependent, double-stranded DNA brake (DSB) repair in *Arabidopsis* involves DSB-induced small RNAs (diRNAs) [49]. RDR2 and RDR6 play redundant roles in the biogenesis of diRNAs, implicating RDR activity in DSB repair.

Silencing of a host gene directed by CaLCuV-derived primary siRNAs is not associated with production of secondary siRNAs

Our above-described results suggested that CaLCuV vsRNAs are primary siRNAs (i.e. RDR-independent) and that secondary siRNAs (i.e. RDR-dependent) may comprise only a small fraction of vsRNAs (if any). To investigate if primary vsRNAs are capable of triggering production of secondary siRNAs in CaLCuV-infected plants, we used a virus-induced gene silencing (VIGS) vector based on the CaLCuV DNA-A derivative lacking most of the *AV1* ORF sequence (positions 350–1032) [50].

When a 354 bp fragment of the *A. thaliana Chlorata I* (*ChlI/CH42*; At4g18480) gene ORF is inserted in place of the AV1 ORF, the resulting recombinant virus CaLCuV::Chl knocks down *ChlI* mRNA levels in all tissues of CaLCuV::Chl-infected *A. thaliana* plants [7] and causes whitening of newly growing tissues due to the loss of chlorophyll ("*chlorata*" phenotype; [50]). The recombinant virus spawns abundant 21, 22, and 24 nt siRNAs from the *ChlI* insert, whose biogenesis does not require RDR6 or RDR2. However, an extensive chlorata phenotype is nearly abolished in *rdr6* and *dcl4* null mutant plants [7], suggesting that RDR6-/DCL4-dependent secondary siRNAs might be involved in total silencing the *ChlI* gene. To test this hypothesis we deep-sequenced sRNAs from CaLCuV::Chl-infected Col-0 plants exhibiting an extensive chlorata phenotype.

Of 2.28 million total 20–25 nt reads, 1.58 million mapped to the *A. thaliana* genome and 0.61 million to CaLCuV::Chl genome (A+B) with zero mismatches. Of the latter reads, 0.45 million originate from the circular CaLCuV::Chl DNA and 0.16 million from DNA-B (Table S1B). This is in contrast to our above observation for wild-type CaLCuV which spawns more abundant vsRNAs from DNA-B.

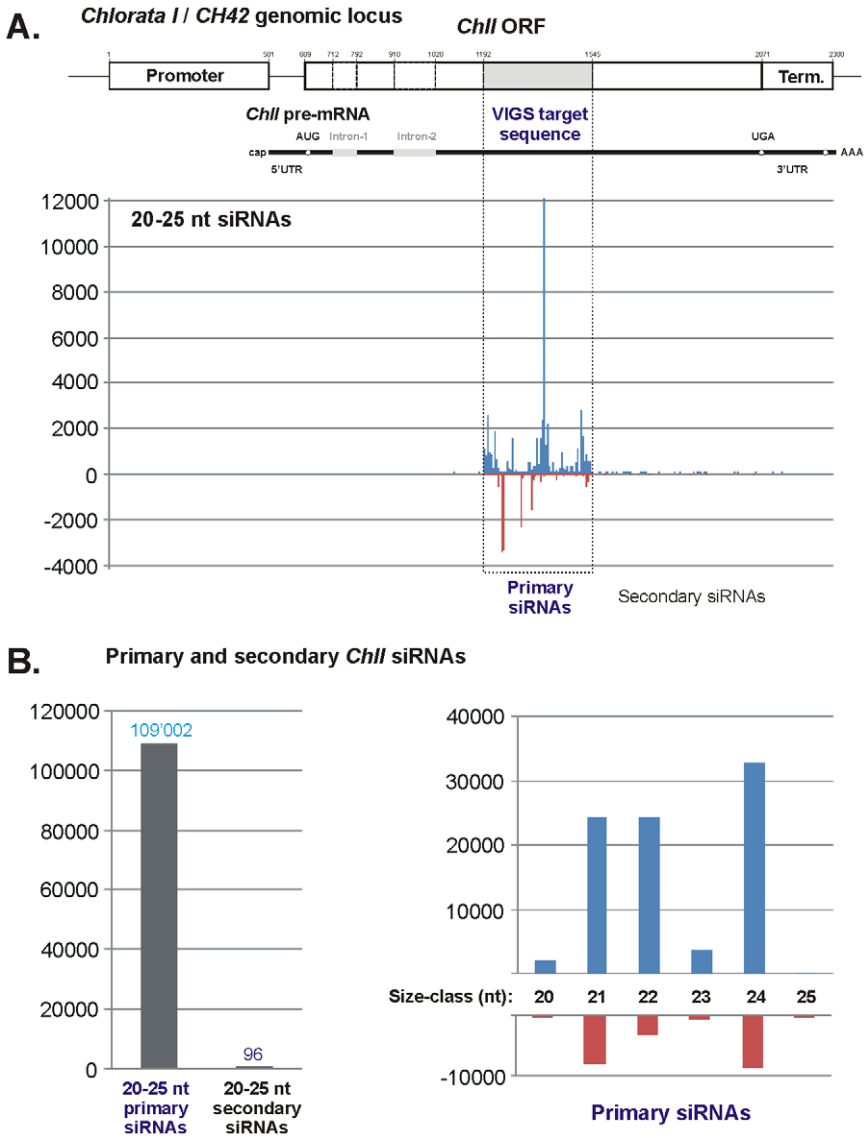


Figure 4. Primary and secondary siRNAs in CaLCuV::Chl virus-infected wild type (Col-0) plants. (**A**) The 2300 bp region of the *Arabidopsis* genome, which contains *Chlorata I/CH42* gene (*ChlI*), is shown schematically with positions of *ChlI* promoter, pre-mRNA with two introns, and terminator sequences indicated; numbering starts 500 nucleotides upstream of the transcription start site. The VIGS target sequence (inserted in CaLCuV::Chl virus) is highlighted with grey. The graph plots the number of 20–25 nt siRNA reads at each nucleotide position of the *ChlI* gene; Bars above the axis represent sense reads starting at each respective position; those below represent antisense reads ending at the respective position (Table S4). (**B**) The left bar graph shows the total numbers of 20–25 nt primary (CaLCuV::Chl-derived) and secondary siRNAs derived from *ChlI* sequences outside of the VIGS target region, while the right bar graph shows the number of primary siRNAs for each size class and polarity.
doi:10.1371/journal.ppat.1002941.g004

Inspection of the single-nucleotide resolution map of 20–25 nt sRNAs perfectly matching to a 3298 bp region of the *A. thaliana* genome, which contains the *ChlI* gene, revealed that of the 109′098 redundant reads, 109′002 originate from the 354 bp segment (positions 1192–1545) that corresponds exactly to the *ChlI* segment inserted in CaLCuV::Chl. The remaining sRNAs (91 reads) originate mostly from the *ChlI* sequence downstream of this segment (Figure 4; Table S1B and Table S4). We conclude that accumulation of secondary siRNAs outside of the vsRNA target region is negligible compared to primary siRNAs. This is consistent with the previous studies that detected no transitivity when endogenous plant genes were knocked down by RNA virus- or transgene-induced silencing [36,38,39].

Within the *ChlI* target region the sRNA profile resembles the global profile of CaLCuV vsRNAs in that the three size-classes are predominant (21-nt – 30%; 22-nt – 25%; 24-nt – 38%). However, the distribution of sRNAs is unequal between the strands: 80% of 20–25 nt reads map to the coding strand, and 21-nt and 22-nt classes derived from the coding strand are equally abundant (28% each). This strong bias is due to a bigger number of sRNA hotspots and higher accumulation levels of sRNA species within the hotspots on the coding strand (Figure 4; Table S4). The significance of this bias for *ChlI* silencing remains to be investigated.

In *A. thaliana*, the *ChlI* gene has a close homolog *ChlI-2* (At5g45930), silencing of which is likely required for the chlorata phenotype. To address if potential silencing of *ChlI-2* is associated

with secondary siRNA production we created a map of *ChlI-2* sRNAs (Figure S3A). Of 3′093 reads of 20–25 nt sRNAs matching the *ChlI-2* genomic locus with zero mismatches in CaLCuV::Chl-infected plants, 2′987 reads map within the 354 bp VIGS-target sequence and only 104 (ca. 3%) map downstream of the target. Moreover, within the target sequence almost all the reads (2′977) match two sequence stretches of >20 nts in length which are identical in *ChlI* and *ChlI-2* (Figure S3A; Table S4). Thus, similar to *ChlI*, only small amounts of secondary siRNAs are generated on *ChlI-2* target gene. Presently, we cannot exclude that these small amounts of secondary siRNAs are required for total chlorata silencing. As we hypothesized earlier [7], total *Chl* silencing is likely established in newly emerging leaves by mobile RDR6- and DCL4-dependent *Chl* siRNAs. Recent studies indicate that 21–24 nt siRNAs act as mobile silencing signals and can direct mRNA cleavage and DNA methylation in recipient cells, even though they accumulate in recipient tissues at much lower levels than in source tissues [51,52].

Notably, vsRNAs targeting *ChlI-2* mRNA at two potentially cleavable sites separated by ca. 100 nts do not trigger any robust secondary siRNA production from the intervening region. This indicates that a two-hit model for the RDR6-dependent biogenesis of tasiRNAs and other secondary siRNAs [14,19,53] does not apply for *ChlI-2* and *ChlI*.

Like in the wild-type DNA-A, vsRNAs cover the entire circular CaLCuV::Chl DNA in both orientations without gaps (Table S4). However, vsRNA hotspots are more evenly distributed along the CaLCuV::Chl sequence compared to the wild-type DNA-A: in fact, new hotspots appear in the intergenic region between the transcription start sites as well as in the terminator region (Figure S3; Table S4). This finding was confirmed by blot hybridization (Figure S2, compare CaLCuV wt and CaLCuV::Chl). Furthermore, genetic analysis revealed that production of vsRNAs from any region of CaLCuV::Chl including the *ChlI* insert does not require RDR6 or RDR2, since vsRNAs of all classes accumulated at similar levels in wild type and *rdr2 rdr6* double mutant plants (*rdr2/6*; Figure S2). The latter finding indicates that RDR6-dependent secondary siRNA production does not occur within the VIGS target region and that potential cleavage of endogenous (*ChlI* or *ChlI-2*) and CaLCuV mRNAs at two sites is not sufficient to attract RDR6 activity.

Taken together, our findings for both wild-type and CaLCuV::Chl viruses suggest that dsRNA precursors of vsRNAs originate from the entire circular viral DNAs including "non-transcribed" intergenic regions. Therefore, these precursors might be produced by Pol II-mediated readthrough transcription far beyond the poly(A) signals, thus encircling the viral DNA in sense and antisense orientation. It can be further suggested that such readthrough transcription is more efficient on CaLCuV::Chl DNA-A than wild-type DNA-A, owing to the smaller size and the chimeric configuration of the rightward transcription unit carrying the *ChlI* segment. This would explain prominent hotspots in the promoter and terminator regions and also much higher production of vsRNAs from CaLCuV::Chl DNA-A than DNA-B, which is not the case for wild-type CaLCuV. Notably, CaLCuV::Chl is an attenuated virus which produces much less severe symptoms than wild type CaLCuV [49]. Whether vsRNA-directed silencing contributes to the attenuated symptom development of this recombinant virus remains to be investigated.

Targeting a transgene transcribed region by CaLCuV-derived primary siRNAs triggers robust production of secondary siRNAs

The apparent paucity of secondary siRNAs derived from CaLCuV mRNAs or *ChlI* and *ChlI-2* mRNAs could be explained by two scenarios. In the first scenario, the products of potential vsRNA-directed cleavage of host and viral mRNAs are not optimal templates for RDR activity. In the second one, CaLCuV infection blocks RDR activity and thereby prevents RDR-dependent amplification of siRNAs. To distinguish between these scenarios, we used the CaLCuV VIGS vector for targeting a transgene in the *A. thaliana* line L2 expressing green fluorescence protein (GFP) under the control of the CaMV 35S promoter and terminator (35S::GFP; [54]; Figure 5). Like other transgenes, 35S promoter-driven *GFP* transgenes in *A. thaliana* and *N. benthamiana* are prone to transitivity in which secondary siRNAs are generated outside of the region targeted by primary sRNAs [36,38,55]. An aberrant nature of transgenic transcripts appears to attract RDR activity.

We inserted in the CaLCuV vector a full-length (FL), 771 bp *GFP* coding sequence (designated '*CodFL*') or 30-bp sequences of the *GFP* transgene transcribed region. The latter is defined here as the *GFP* mRNA region from the transcription start site to the mRNA processing/poly(A) addition site. As depicted in Figure 5, the short inserts included the sequences from within the 5′-untranslated region (5′UTR) (designated '*Lead*'), the beginning, middle and end of the coding sequence ('*CodB*', '*CodM*' and '*CodE*'), and the 3′UTR ('*Trail*' and '*PolyA*') and the sequences surrounding the ATG start codon ('*Start*') or the TAA stop codon ('*Stop*'). Inoculation of L2 plants with the resulting recombinant viruses by biolistic delivery of viral DNA led to development of local *GFP* silencing on inoculated leaves followed by systemic *GFP* silencing on newly-emerging infected tissues (both leaves and inflorescence; Figure S4B). *GFP* silencing in infected tissues, which was manifested under UV light as red fluorescence areas on otherwise green fluorescent tissues (Figure 5B and Figure S4B), well correlated with knockdown of *GFP* mRNA levels as measured by real time PCR (Figure S4D).

All the recombinant viruses carrying an insert from the *GFP* transcribed region induced systemic *GFP* silencing, although to various degrees (Figure 5B). Furthermore, in all these cases, *GFP* silencing correlated with accumulation of *GFP* siRNAs derived from both the short insert/target sequences and the *GFP* mRNA sequences outside of the target sequence (Figure 5C and Figure S4C). Notably, the 30 bp *GFP* insert/target sequences generally gave rise to abundant siRNAs of 21-nt, 22-nt and 24-nt classes, resembling those derived from the virus genome and therefore likely originating from the replicating virus carrying the insert rather than from the transgene. By contrast, secondary siRNAs derived from non-target sequences of the *GFP* transgene were generally represented by a dominant 21-nt class, although 22-nt and 24-nt classes were also detected (Figure 5C; also see below). Furthermore, targeting the *GFP* sequences upstream of the translation stop codon (*Lead, Start, CodB, CodM* and *CodE*) induced the production of abundant secondary siRNAs exclusively from sequences downstream of the target site, whereas targeting the 3′UTR sequences (*Stop, Trail* and *PolyA*) resulted in secondary siRNAs from the sequences upstream and downstream of the target site (Figure 5C). Such directionality in secondary siRNA biogenesis resembles that in RDR6-/DCL4-dependent biogenesis of tasiRNAs [17,18]. Our findings further suggest that, following vsRNA-directed cleavage of *GFP* mRNA, the 5′-cleavage product might be protected by translating ribosomes from being converted to dsRNA precursor of secondary siRNAs. However, if it contains the translation stop codon, the ribosomes can terminate translation and be released. Thus, following vsRNA-directed cleavage downstream of the stop codon, both 5′ and 3′ cleavage products of *GFP* mRNA enter the secondary siRNA-generating pathway.

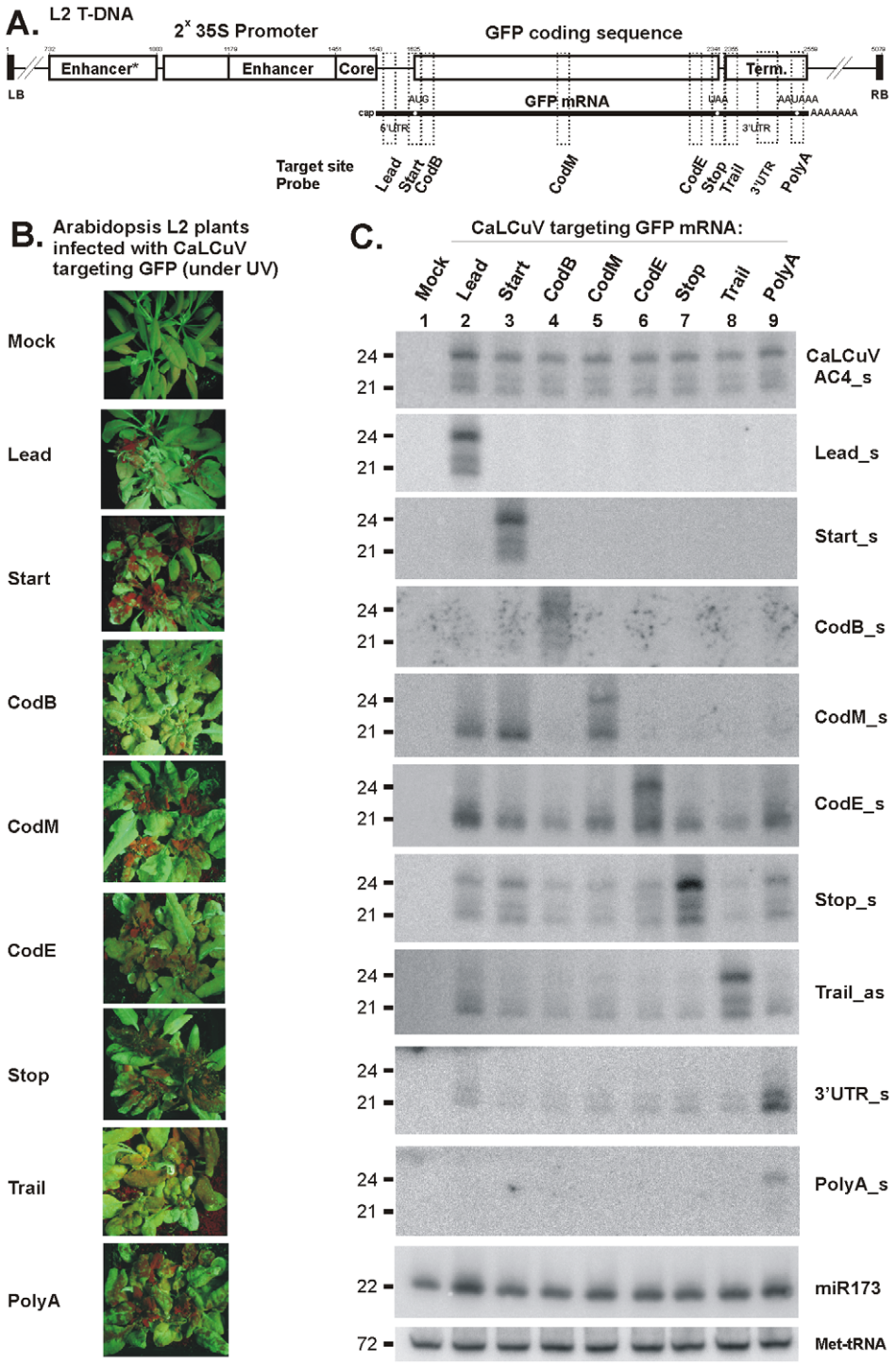


Figure 5. VIGS phenotypes and accumulation of primary and secondary siRNAs in L2 transgenic plants infected with CaLCuV::GFP viruses targeting the *GFP* transcribed region. (**A**) The L2 T-DNA region containing the 35S-GFP transgene is shown schematically. Positions of the duplicated CaMV 35S enhancer and core promoter elements, *GFP* mRNA elements including 5′UTR, translation start (AUG) and stop (UAA) codons and 3′UTR with poly(A) signal (AAUAAA), and 35S terminator sequences indicated. Numbering is from the T-DNA left border (LB). The VIGS target sequences, inserted in the CaLCuV::GFP viruses *Lead*, *CodB, CodM, CodE, Trail* and *polyA*, are indicated with dotted boxes; (**B**) Pictures under UV light of L2 transgenic plants infected with the above viruses; (**C**) Blot hybridization analysis of total RNA isolated from plants shown in Panel B. The blot was successively hybridized with short DNA probes specific for CaLCuV *AC4* gene (AC4_s) and 35S::GFP transgene sequences inserted in the CaLCuV::GFP viruses (*Lead, CodB, CodM, CodE, Trail* and *polyA*), the *GFP* mRNA 3′UTR non-target sequence (3′UTR) and *Arabidopsis* miR173 and Met-tRNA (the latter two serve as loading control).
doi:10.1371/journal.ppat.1002941.g005

The above findings based on blot hybridization analysis (Figure 5C) were fully validated by Illumina sequencing of sRNAs from L2 plants infected with *Lead*, *CodM*, *Trail* and *polyA* viruses (Figure 6 and Figure S5; Table S5 and Table S6). In addition, analysis of the deep sequencing data showed that vsRNAs targeting the 3′UTR induce production of much more abundant secondary siRNAs from the region upstream of the target site than from downstream sequences (Figure 6). Interestingly, secondary siRNA hotspots are non-randomly distributed along the *GFP* transcribed region: in all the four cases the siRNA hotspots occur in the region comprising the 3′ portion of the *GFP* ORF and the beginning of the 3′UTR. The size-class profile and relative abundance of siRNA species in this siRNA hotspot region are very similar. In the case of *Lead* and *polyA* viruses, additional siRNA hotspots occur in the middle of *GFP* ORF and the 3′UTR, respectively (Figure 6 and Figure S5). Interestingly, vsRNAs targeting the 5′UTR does not induce abundant secondary siRNA production from the region immediately downstream of the target site, which contains the 5′ portion of *GFP* ORF. This region also appears to be a poor source/target of primary vsRNAs (see *CodB* in Figure 5). Furthermore, robust production of secondary siRNAs does not appear to depend on the accumulation levels of any major size-class of primary vsRNAs of antisense polarity that have the potential to cleave *GFP* mRNA and initiate secondary siRNA biogenesis (Figure S5; Table S1, Table S5 and Table S6). We assume that, once initiated by primary vsRNAs, secondary siRNA biogenesis might be reinforced by feedback loops in which certain secondary siRNAs of antisense polarity target the *GFP* mRNA. Such feedback loops regulate tasiRNA production from *TAS1c* gene, in which certain tasiRNAs cleave its own precursor transcript to initiate RDR6-dependent production of additional dsRNAs [20], and potentially occur in transgene-induced silencing systems [56,57].

Targeting a transgene enhancer region by CaLCuV-derived primary siRNAs causes silencing without secondary siRNA production

Contrary to what we observed for the transcribed region, targeting of the *GFP* non-transcribed regions with short sequences inserted into the CaLCuV VIGS vector did not lead to *GFP* silencing or secondary siRNA production in systemically-infected L2 plants (Figure 7 and Figure S4). The 30-bp sequences which surround the 35S core promoter elements including the CAAT and TATA boxes ('*CAAT*' and '*TATA*') and the transcription start site ('*Plus1*'), or sequences that occur in a distal region of the 35S enhancer ('*EnhSh*') and just downstream of the mRNA processing/poly(A) addition site ('*Post*') gave rise to abundant siRNAs of the three major classes but no secondary siRNAs were detected outside of the target sequence. Furthermore, insertion of the 90-bp 35S core promoter region ('*Core*') did not result in *GFP* silencing or secondary siRNA production, despite abundant primary siRNAs targeting this region. However, insertions of the entire 35S enhancer region of 272 bp ('*Enh*') or the full-length promoter of 382 bp ('*ProFL*') resulted in systemic *GFP* silencing. But also in these two cases no secondary siRNAs were detected outside of the target region (Figure 7). These findings were confirmed by Illumina sequencing of sRNAs from L2 plants systemically infected with *Core, Enh* and *ProFL* viruses (Figures 8 and Figure S6; Table S5 and Table S6). In addition, the deep sequencing revealed that, besides extremely low levels of secondary siRNA accumulation outside of the target region, there appear to be almost no secondary siRNA amplification within the target region. Thus, the duplicated 273-bp Enhancer* region shares 94% nucleotide identity with the target Enhancer region, since these sequences originate from two different strains of CaMV, and we found only negligible numbers of reads in the three stretches of the Enhancer* sequence that have mismatches to corresponding stretches of the Enhancer sequence (Figure 8; Table S5, see positions 760–781, 803–837 and 869–905).

Taken together, we conclude that production of abundant secondary siRNAs can be triggered by primary virus-derived siRNAs that target *GFP* mRNA. Hence, CaLCuV infection does not block amplification of secondary siRNAs likely mediated by RDR activities (see below). This is also supported by our blot hybridization analysis showing that accumulation of RDR6-dependent tasiRNAs is not significantly affected by CaLCuV infection (Figure S2; siR255). Both primary (virus-derived) and secondary siRNAs correlate with efficient *GFP* silencing. However, targeting of the non-transcribed, 35S enhancer region by primary siRNAs induces efficient *GFP* silencing without any substantial production of secondary siRNAs. Hence, secondary siRNAs do not appear to be necessary for silencing *GFP* transgene, at least at the transcriptional level. Previously, transcriptional VIGS through targeting the 35S promoter region of 35S::*GFP* transgene was observed but its dependence on primary or secondary siRNAs was not tested in that case [58].

GFP secondary siRNAs are RDR6-dependent

To investigate genetic requirements for the biogenesis of *GFP* secondary siRNAs, the L2 transgenic line was crossed with the Col-0 mutant lines carrying point mutations in *RDR6* (*rdr6-14*; [59]) and *DCL4* (*dcl4-2*; [60]). The resulting homozygous mutant lines L2 x *rdr6* and L2 x *dcl4* expressed high levels of GFP, similar to those of the parental L2 plants (not shown).

Systemic infection of L2 x *rdr6* and L2 x *dcl4* plants with the recombinant viruses *Lead*, *CodM* and *Trail* resulted in *GFP* silencing in all cases, except L2 x *rdr6* plants infected with the *Lead* virus. Consistent with our findings for wild-type CaLCuV (Figure S2) and CaLCuV::Chl ([7]; Figure S2), blot hybridization analysis revealed that the biogenesis of 21, 22 and 24 nt vsRNAs derived from the *AC4* ORF region of the three recombinant viruses was not affected in L2 x *rdr6* plants lacking RDR6 (Figure 9). By contrast, probes specific for the target transgene revealed a major contribution of RDR6 in secondary siRNA production. In fact, production of secondary siRNAs of all size-classes outside of the target region was nearly abolished in L2 x *rdr6* plants infected with *Lead*, *CodM* and *Trail* viruses (Figure 9). For the latter two viruses, accumulation of siRNAs from the insert/target sequence was also reduced: interestingly, the reduced accumulation was observed for siRNAs of sense but not antisense polarity in *CodM* virus, while siRNAs of both polarities were strongly reduced in *Trail* virus. By contrast, accumulation of siRNAs from the *Lead* insert/target sequence was not altered in L2 x *rdr6* plants infected with *Lead* virus (Figure 9). We conclude that RDR6-independent primary vsRNAs represent the majority of siRNAs derived from the *Lead* sequence, whereas the *CodM* and *Trail* sequences also spawn RDR6-dependent secondary siRNAs in addition to primary vsRNAs. These secondary siRNAs could potentially be produced from the transgene and/or the viral insert. We therefore used the probes specific to the viral sequence located just downstream of the insert (CbA1063_s and CbA1063_as), i.e. present in the chimeric rightward viral transcript. The results revealed that, in the case of *Lead* and *CodM* viruses, RDR6 is not involved in production of vsRNAs from this region (Figure 9). Thus, the contribution of RDR6 to siRNA production from the *CodM* insert/target sequence of antisense polarity can be explained by RDR6-dependent siRNA production from the target gene rather than the chimeric virus. However, accumulation of vsRNAs derived from

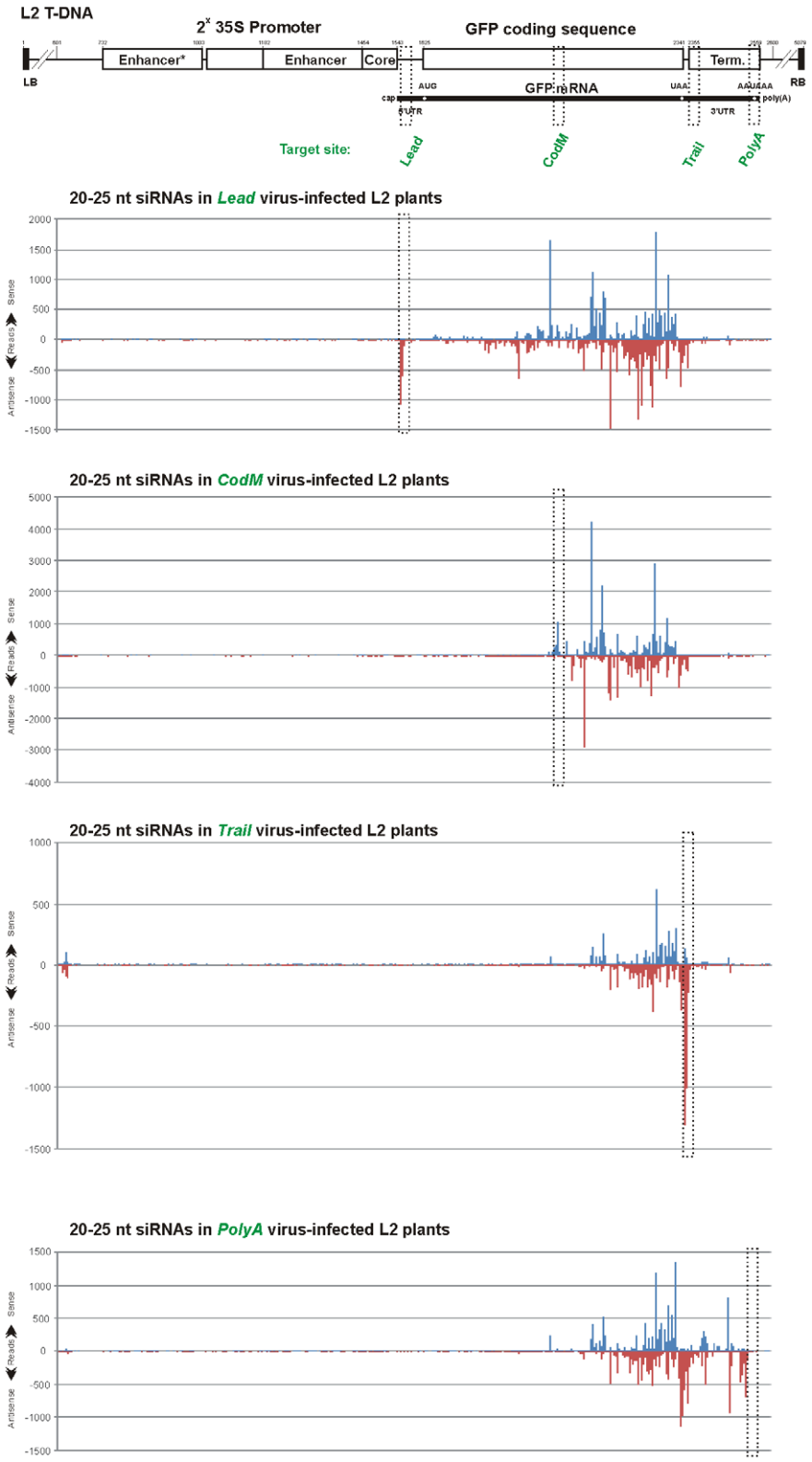


Figure 6. Maps of primary and secondary siRNAs accumulating in L2 transgenic plants infected with CaLCuV::GFP viruses that target the *GFP* transcribed region. The graphs plot the number of 20–25 nt vsRNA reads at each nucleotide position of the L2 T-DNA-based 35S::GFP transgene; Bars above the axis represent sense reads starting at each respective position; those below represent antisense reads ending at the respective position (Table S5). The 35S-GFP transgene is shown schematically above the graphs. Positions of the duplicated 35S enhancer and core promoter, *GFP* mRNA elements and 35S terminator are indicated. Numbering is from the T-DNA left border (LB). The VIGS target sequences inserted in the CaLCuV::GFP viruses Lead, CodM, Trail and polyA are indicated with dotted boxes.
doi:10.1371/journal.ppat.1002941.g006

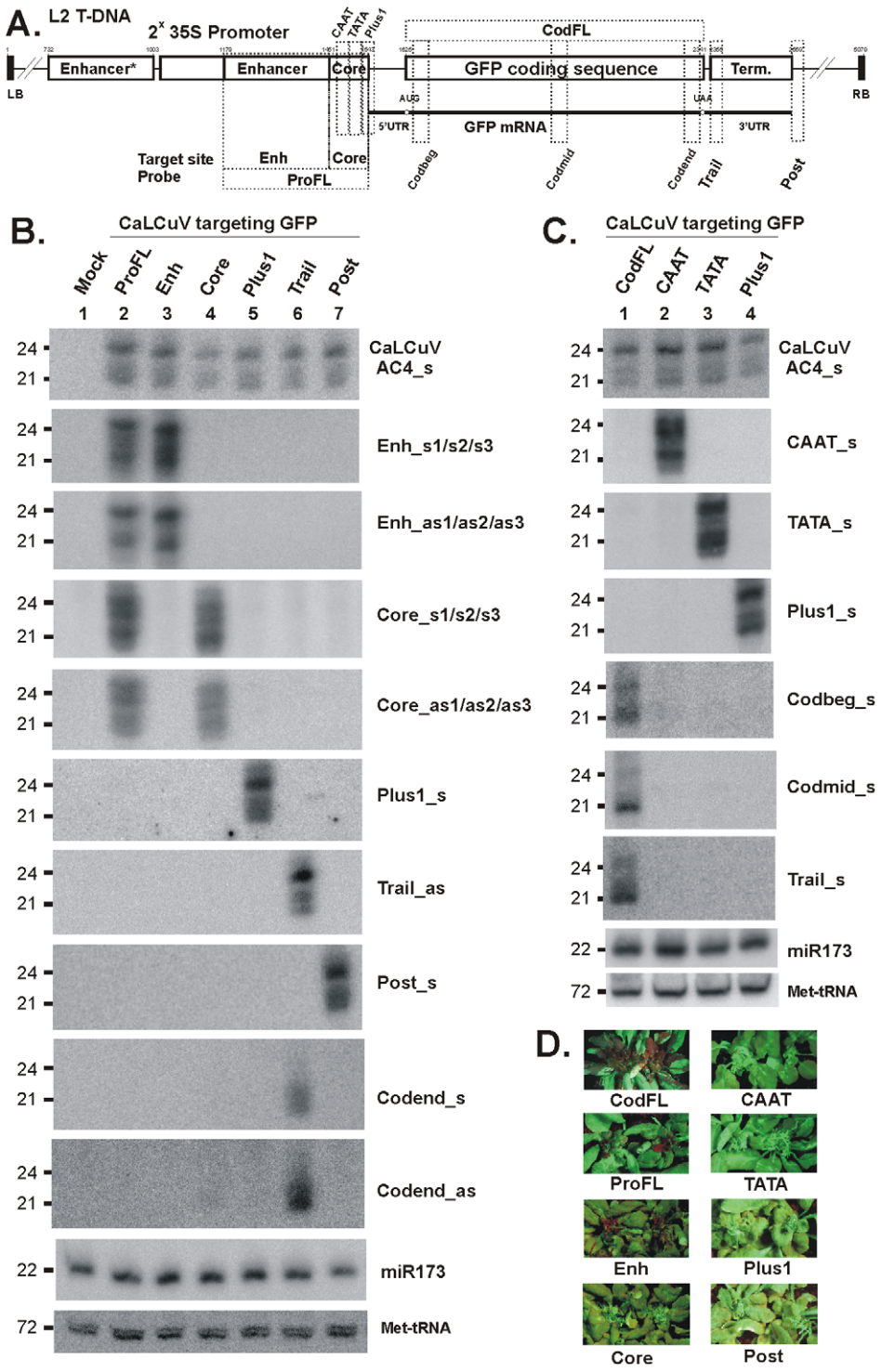


Figure 7. VIGS phenotypes and primary siRNA accumulation in L2 transgenic plants infected with CaLCuV::GFP viruses that target the *GFP* promoter and terminator elements. (**A**) The L2 T-DNA region containing the 35S-GFP transgene is shown schematically. Positions of the duplicated CaMV 35S enhancer (*Enh*) and core promoter (*Core*) elements (CAAT and TATA boxes and transcription start *Plus1*), the *GFP* mRNA elements (5′UTR, AUG and UAA codons and 3′UTR, and 35S terminator are indicated. Numbering is from the T-DNA left border (LB). The VIGS target sequences, inserted in the CaLCuV::GFP viruses *ProFL, Enh, CAAT, TATA, Plus1, CodFL, Trail* and *Post* are indicated with dotted boxes; (**B**) and (**C**) Blot hybridization analysis of total RNA isolated from L2 plants infected with the above viruses. The two blots were successively hybridized with short DNA probes specific for CaLCuV *AC4* gene (AC4_s) and the 35S::GFP transgene sequences inserted in CaLCuV::GFP viruses and *Arabidopsis* miR173 and Met-tRNA (the latter two serve as loading control). (**D**) Pictures under UV light of L2 transgenic plans infected with the CaLCuV::GFP viruses (names indicated).

doi:10.1371/journal.ppat.1002941.g007

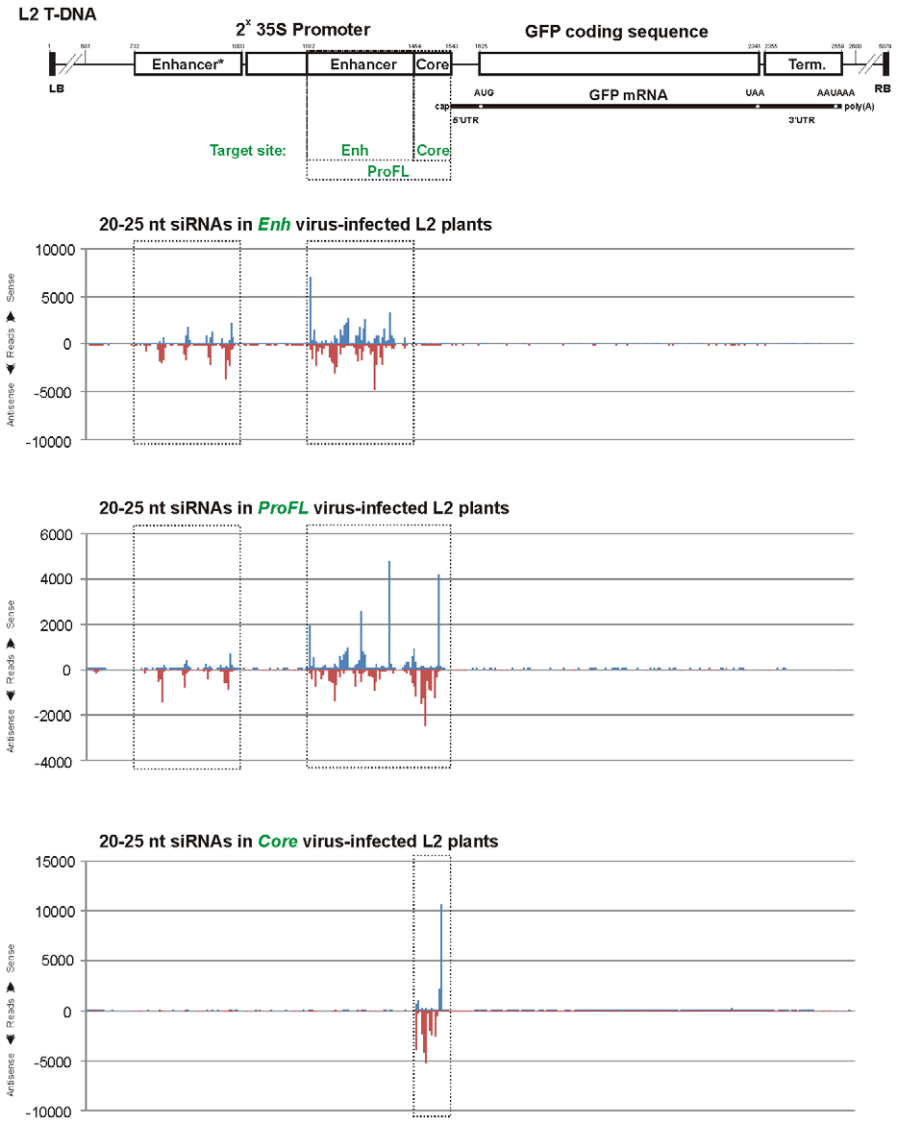


Figure 8. Maps of primary siRNAs accumulating in L2 transgenic plants infected with CaLCuV::GFP viruses that target the *GFP* promoter elements. The graphs plot the number of 20–25 nt vsRNA reads at each nucleotide position of the L2 T-DNA-based 35S::GFP transgene; Bars above the axis represent sense reads starting at each respective position; those below represent antisense reads ending at the respective position (Table S5). The 35S-*GFP* transgene is shown schematically above the graphs. Positions of the duplicated 35S enhancer and core promoter, *GFP* mRNA elements and 35S terminator are indicated. Numbering is from the T-DNA left border (LB). The VIGS target sequences inserted in the CaLCuV::GFP viruses *CodFL, Enh* and *Core* are indicated with dotted boxes. Note that the duplicated 35S promoter sequences Enhancer* and Enhancer (each 273 nt long) share 94% nucleotide identity, since they originate from two different strains of CaMV. Therefore, primary siRNA reads are unequally distributed between the two VIGS target regions.
doi:10.1371/journal.ppat.1002941.g008

the chimeric transcript region of *Trail* virus was substantially reduced (24-nt) or nearly abolished (21-nt and 22-nt) in L2 x *rdr6* plants. This indicates that, in addition to the transgenic mRNA, the chimeric viral transcript can also be used for RDR6-dependent production of secondary siRNAs. But the insert sequence itself appears to regulate relative contribution of RDR6. Notably, the *ChlI* insert sequence does not make the chimeric viral transcript prone to RDR6-dependent vsRNA production (Figure S2). It remains to be further investigated why the *Trail* (but not *Lead*, *CodM or ChlI*) sequence makes the viral chimeric transcript prone to RDR6-dependent amplification of secondary siRNAs. Interestingly, this sequence originates from the CaMV terminator/leader region and contains two stretches of AG-repeats (Protocol S1).

It is puzzling that, in the absence of RDR6-dependent secondary siRNAs in L2 x *rdr6* plants, the *GFP* silencing is efficiently triggered by *CodM* and *Trail* viruses but not by *Lead* virus. We speculate that *GFP* mRNA cleaved by primary siRNAs within its 5′UTR can still be translated, unless it enters the RDR6 pathway converting the coding and 3′UTR sequences to secondary siRNAs. By contrast, primary siRNA-directed cleavage within the coding sequence or 3′UTR would block productive translation and could therefore be sufficient for *GFP* silencing.

In L2 x *dcl4* plants, we detected reduced accumulation of 21-nt primary siRNAs from the viral *AC4* region and 21-nt primary and secondary siRNAs from the *GFP* sequences. Unexpectedly, accumulation of 22-nt and 24-nt primary and secondary siRNAs was increased: this increase was more prominent for secondary

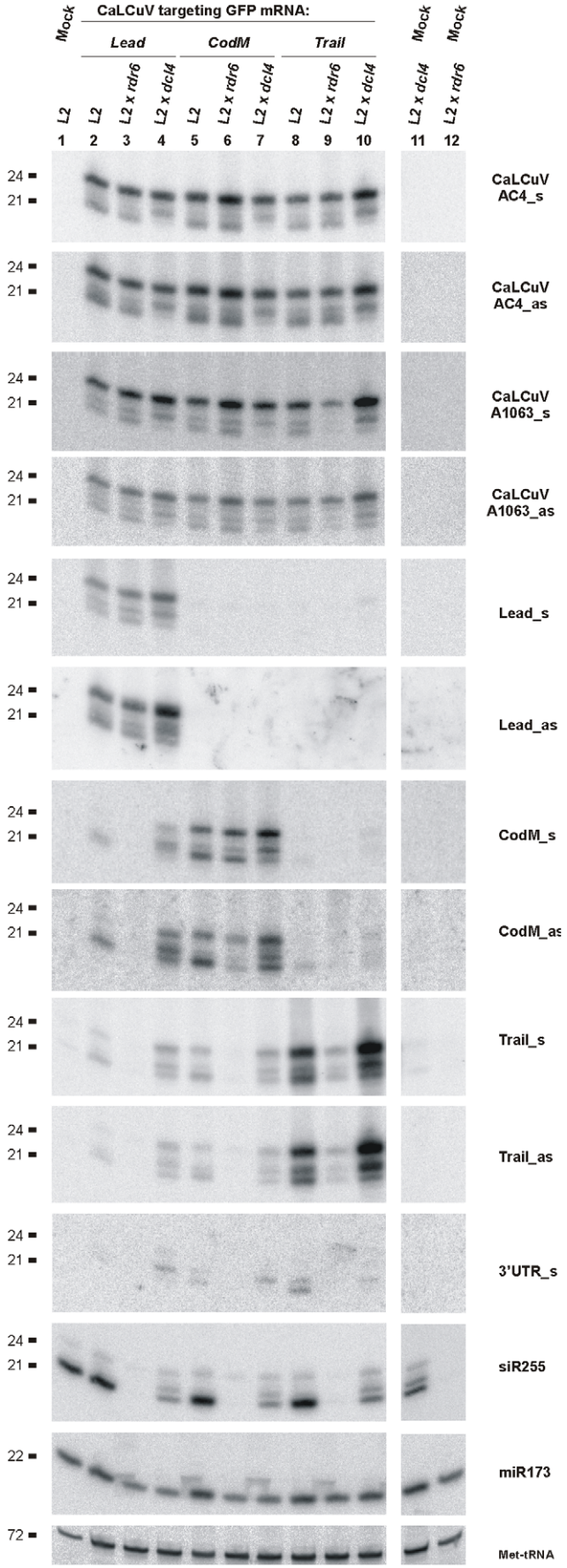


Figure 9. Genetic requirements for primary and secondary siRNA accumulation in L2 transgenic plants. Blot hybridization analysis of total RNA isolated from L2, L2 x *rdr6* and L2 x *dcl4* plants infected with CaLCuV::GFP viruses *Lead, CodM* and *Trail*. The blot was successively hybridized with short DNA probes specific for CaLCuV genes AC4 (AC4_s and AC4_as) and AV1 (A1063_s and A1063_as), 35S::GFP transgene sequences inserted in the CaLCuV::GFP viruses (*Lead, CodM, Trail*), *GFP* mRNA 3′UTR non-target sequence (3′UTR_s) and *Arabidopsis* miR173 and Met-tRNA (the latter two serve as loading control).
doi:10.1371/journal.ppat.1002941.g009

GFP siRNAs (Figure 9). This resembles the shift in the profile of RDR6-dependent 21-nt tasiRNAs in this particular mutant background ([60]; Figure 9, see tasiRNA siR255). Thus, a mutated DCL4 protein expressed from the *dcl4-2* allele appears to promote processing of RDR6-dependent dsRNAs by alternate DCLs that generate longer siRNAs (i.e. DCL2 and DCL3).

Taken together, our findings confirm a major role of DCL4 in processing 21-nt secondary siRNAs from RDR6-dependent dsRNA precursors derived from the transgene and 21-nt primary vsRNAs from RDR6-independent viral dsRNA precursors. In addition, our results reveal that RDR6-dependent dsRNA can be efficiently processed by alternate DCL activities if the DCL4 protein is mutated by an amino acid substitution in the helicase domain. These alternate DCLs produce primary and secondary siRNAs which are equally potent in *GFP* silencing, since we did not observe any substantial difference in systemic silencing phenotypes between wild-type and *dcl4-2* plants infected with any of the recombinant viruses. This is in line with our previous findings for CaLCuV::Chl-derived primary vsRNAs of distinct classes produced in single, double and triple *dcl* mutant plants, which could efficiently knockdown *ChlI* mRNA [7]. Previously, a major role of DCL2 was established for production of secondary siRNAs in a transgene targeted by primary siRNAs from another transgene [11]. Here, in addition to DCL2, we find the apparent involvement of DCL3 which normally generates 24-nt nuclear siRNAs in secondary siRNA production. Thus, a fraction of dsRNA precursors of the *GFP* transgene-derived secondary siRNAs might be localized in the nucleus. Alternatively, a fraction of DCL3 protein might also be cytoplasmic.

Concluding remarks

Secondary siRNAs are involved in various silencing pathways in plants, fungi and some animals. In *C. elegans*, RDR-dependent amplification of secondary siRNAs appears to reinforce silencing triggered by primary siRNAs which are processed by dicer from endogenous or exogenous dsRNA [61,62]. In plants, some of the endogenous mRNAs targeted by miRNAs spawn RDR6-dependent secondary RNAs, a contribution of which to miRNA-directed gene silencing is not fully clarified [14,15]. In most cases, plant miRNA-directed cleavage or translational repression is sufficient for robust gene silencing without production of secondary siRNAs [14]. Likewise, most plant mRNAs silenced by transgene- or virus-derived primary siRNAs do not spawn secondary siRNAs. This suggests that plant mRNAs could have evolved to be poor templates for RDR activity. Our study supports this notion by demonstrating that *Arabidopsis ChlI* and *ChlI-2* mRNAs that undergo robust VIGS spawn only small amounts of secondary siRNAs. Furthermore, we demonstrate that geminiviral mRNAs, which can potentially be targeted by highly abundant vsRNAs of antisense polarity (Figure 2), are not templates for RDR1-, RDR2-, or RDR6-dependent siRNA amplification. By contrast, the transgenic *GFP* mRNA targeted by primary viral siRNAs spawns massive amounts of secondary siRNAs whose production requires RDR6. Our findings suggest that some aberrant feature(s) of the

transgenic *GFP* mRNA possessing non-self UTR sequences may attract RDR6 activity. Notably, the involvement of RDR6 and RDR1 in production of viral siRNAs in RNA virus-infected plants was revealed only by using the mutant RNA viruses carrying deletions or point mutations in viral silencing suppressor genes: unlike wild-type RNA, the mutated viral RNA spawned RDR-dependent vsRNAs. What makes mutant/chimeric viral mRNAs and transgenic mRNAs good templates for RDR activity remains unclear. One possibility is that viral and plant mRNAs could have evolved primary sequence or secondary structure elements that block RDR activity. Such elements may accidentally be disrupted by mutations in the suppressor-deficient RNA viruses. Likewise, transgene transcripts might lack some of the naturally evolved sequence or structure elements.

Our findings suggest that the precursors of geminiviral siRNAs are most likely produced by Pol II-mediated bidirectional readthrough transcription in both sense and antisense orientations on the circular viral DNA. Such transcripts (or their degradation products) can potentially pair viral mRNAs and thus form perfect dsRNAs to be processed by multiple DCLs into vsRNAs. Readthrough transcription far beyond a poly(A) signal is a known property of Pol II. In pararetroviruses, it represents an obligatory mechanism by which a pregenomic RNA covering the entire circular genome is generated. The poly(A) signal of plant pararetroviruses is located at a relatively short distance (e.g. 180 bp in CaMV) downstream of the pregenomic RNA promoter: this allows efficient readthrough transcription at the first encounter by the Pol II complex and termination of transcription at the second encounter [63,64]. Thus, substantial readthrough transcription can also be expected in geminiviruses which possess relatively short transcription units. Evidence for the existence of readthrough transcripts was obtained earlier for a related geminivirus [34] and is also provided here by deep sequencing showing that vsRNAs of both sense and antisense polarities densely tile along the entire CaLCuV genome including ''non-transcribed'' intergenic region of both DNA-A and DNA-B. Pol II readthrough transcription downstream of a canonical poly(A) signal of the endogenous *A. thaliana* gene *FCA* was recently shown to be repressed by a DCL4-dependent mechanism [12]. In a *dcl4* mutant, the increased transcriptional readthrough far beyond the *FCA* poly(A) signal triggered silencing of a transgene containing the same 3′ region. Notably, the transgene silencing was caused by RDR6-dependent production of very abundant 22-nt siRNAs by DCL2 and less abundant 24-nt siRNAs by DCL3. This siRNA pattern resembles the pattern of *GFP* transgene-derived secondary siRNAs that we observed in L2 x *dcl4* plants (Figure 9). Also in line with our observations, robust siRNA-directed silencing of the transgene and *FCA* did not spread to a converging gene that overlaps with the *FCA* readthrough transcript [12], further supporting the notion that most endogenous genes are not prone to RDR6-dependent transitivity.

Materials and Methods

Plant mutants and virus infection

Arabidopsis thaliana wild-type (Col-0) and *rdr2/6*, *rdr1/2/6* and *dcl1/2/3/4* mutant lines used in this study, their growth conditions and infection with wild-type CaLCuV (the DNA-A clone 'CLCV-A dimer' [33] and the DNA-B clone pCPCbLCVB.002 [50]) and CaLCuV::Chl (pMTCbLCVA::CH42 and pCPCbLCVB.002 [50]) using biolistic delivery of viral DNA have been described earlier [7,8]. Using the same protocols, L2 transgenic plants (Line 2; [54]) were grown and inoculated with CaLCuV::GFP viruses.

L2 plants [54] were crossed with the *dcl4-2* and *rdr6-14* mutants [59,60]. L2 homozygosity was determined by PCR in the F2 populations using 5′-TTGCTGCAACTCTCTCAGGGCC-3′ and 5′-GATAAATGTGGAGGAGAAGACTGCC-3′ for detecting the presence of the T-DNA and 5′-ACACTCTCTCTCCTT-CATTTTCA-3′ and 5′-TCTGCAACACTCTGTCATTGG-3′ for detecting the absence of intact genomic region. RDR6-14 homozygosity was determined by visual observation of the typical epinastic leaf phenotype of the *rdr6* mutants and was further confirmed using a dCAPS marker consisting of *NcoI* digestion of the PCR product obtained using 5′-AAGATTTGATCCCTGAGc-CAT-3′ and 5′-GTTCGCCTTGTTTCTTGCTT-3′. DCL4-2 homozygosity was determined by the typical epinastic leaf phenotype of the *dcl4* mutants. Homozygosity for L2 and the respective mutations were confirmed in F3 plants following the same procedures.

Construction of recombinant viruses

The CaLCuV::GFP viruses *EnhSh*, *CAAT*, *TATA*, *Plus1*, *Lead*, *Start*, *CodB*, *CodM*, *CodE*, *Stop*, *Trail*, *PolyA* and *Post* were generated by cloning preannealed sense and antisense oligonucleotides (listed in Protocol S1) into *XbaI* and *XhoI* sites of the CaLCuV VIGS vector pCPCbLCVA.007 [50]. The CaLCuV::GFP viruses *Enh*, *Core* and *ProFL* were generated by subcloning into *XbaI* and *XhoI* sites of pCPCbLCVA.007 the corresponding regions of the L2 T-DNA 35S promoter using PCR with primers listed in Protocol S1 on total DNA isolated from L2 transgenic plants. In all the above derivatives of the CaLCuV VIGS vector the insert sequences are in antisense orientation with respect of the *AV1* gene promoter.

sRNA analysis

For both blot hybridization and Illumina deep-sequencing, aerial tissues of three virus-infected (or mock-inoculated) plants were harvested one month post-inoculation and pooled for total RNA preparation using a Trizol method [7]. sRNA blot hybridization analysis was performed as in Blevins et al. [7] using short DNA oligonucleotide probes listed in Protocol S1. cDNA libraries of the 19–30 nt RNA fraction of total RNA samples were prepared as we described previously [8]. The high-coverage libraries of wild-type CaLCuV were sequenced on an Illumina Genome Analyzer (GA) *Hi-Seq 2000* using a *TruSeq v5* kit, while the low coverage libraries on a GA-II using *Chrysalis v2*. The libraries of CaLCuV::Chl and CaLCuV::GFP viruses were sequenced on a GA-IIx using *Chrysalis v4* and *TruSeq v5*, respectively. After trimming the adaptor sequences, the datasets of reads were mapped to the reference genomes of *Arabidopsis thaliana* Col-0 (TAIR9), CaLCuV (U65529.2 for DNA-A and U65530.2 for DNA-B) and other references using a Burrows-Wheeler Alignment Tool (BWA version 0.5.9) [65] with zero mismatches to the reference sequence. The reference sequences of CaLCuV DNA-A and its derivatives, CaLCuV DNA-B, L2 T-DNA and *ChlI*/CH42 and *ChlI-2* genomic loci are given in Protocol S1. Reads mapping to several positions on the references were attributed at random to one of them. To account for the circular virus genome the first 50 bases of the viral sequence were added to its 3′-end. For each reference genome/sequence and each sRNA size-class (20 to 25 nt), we counted total number of reads, reads in forward and reverse orientation, and reads starting with A, C, G and T (Table S1). In the single-base resolution maps of 20, 21, 22, 23, 24 and 25 nt vsRNA (Tables S2, S3, S4, S5, S6 and S7), for each position on the sequence (starting from the 5′ end of the reference sequence), the number of matches starting at this position in forward (first base of the read) and reverse (last base

of the read) orientation for each read length is given. Note that the reads mapped to the last 50 bases of the extended viral sequence were added to the reads mapped to the first 50 bases.

Analysis of long viral RNA and DNA by blot hybridization

The detailed protocol for high-resolution analysis of long RNA using total RNA and 5% PAGE followed by blot hybridization was described previously [30]. To detect the viral mRNAs AV1, AC2/AC3, BV1 and BC1 (Figure 3A), the membrane was successively hybridized with mixtures of DNA oligonucleotides complementary to each given mRNA (for sequences, see Protocol S1).

Southern blot analysis was performed as in [66]. In short, total DNA from the plants were extracted by CTAB-based protocol. Five µg of total DNA was electrophoresed in 1% agarose gel prepared in 1× Tris-sodium acetate-EDTA buffer. Full-length linear DNA of CaLCuV was loaded as a positive control for Southern hybridization. After EtBr staining, the DNA in the gel was alkali-denatured and transferred to the Hybond N+ nylon membrane (GE healthcare lifesciences). PCR fragments of DNA-A (900 bp obtained with the primers Cb_AV1_qPCR_s and Cb_AC3_qPCR_as) and DNA-B (862 bp Cb_BV1_qPCR_s and Cb_BC1_qPCR_as), which do not contain the common region of the virus, were labeled with [α-32P]dCTP using Rediprime II DNA labeling system (GE healthcare lifesciences) and used as probes. Hybridization with the labeled probe was performed at 65°C for 16–20 hours using PerfectHyb Plus Hybridization Buffer (Sigma-Aldrich) and the membrane was washed thrice at 65°C with 2× SSC/0.5% SDS. The signal was detected after 5 days exposure to a phosphor screen using a Molecular Imager (Typhoon FLA 7000, GE healthcare lifesciences).

Real time PCR

Relative accumulation of polyadenylated viral mRNAs and total viral DNA in wild type versus *rdr1/2/6* (Figure 3D) was measured using real time PCR as in [8]. For polyadenylated RNA, cDNA was synthesized from 5 µg of total RNA using 100 pmoles of oligo d(T)16 primer. The RNA-primer mixture was heated to 70°C for 10 min and chilled on ice for 5 min. 4 µl of 5× first-strand synthesis buffer (250 mM Tris-HCl [pH 8.3], 375 mM KCl, 15 mM MgCl2, 0.1 M DTT), 2 µl 0.1 M DTT, 1 µl 10 mM deoxynucleoside triphosphate mix and 1 µl (200 U) of Superscript III reverse transcriptase (Invitrogen) were added and incubated at 50°C for 60 min. The reaction was stopped by heating the mixture to 95°C for 5 min. 2 µl of the 10 times diluted reverse transcription reaction mix or 2 µl of total DNA (2 ng) were taken for PCR in LightCycler 480 Real-Time PCR System (Roche applied sciences) using FastStart Universal SYBR Green Master (Rox) mix (Roche) and primers designed using Beacon designer 2 software (PREMIER Biosoft International). PCR primers specific for viral DNAs A and B and each viral mRNA as well as internal controls (18S rDNA and *ACT2* mRNA) are given in Protocol S1. Cycling parameters were 95°C for 10 min, followed by 45 cycles: 95°C for 10 s, 56°C for 10 s, 72°C for 20 s. Amplification efficiency of primers was determined by means of a calibration curve (Ct value vs. log of input cDNA/DNA) prepared in triplicate. The Ct values obtained for viral genes were normalized with internal control values and the delta Ct values were obtained. The normalized values for CaLuCV-infected wild type Col-0 were set to 1. To quantify the L2 GFP mRNA levels, poly-dT cDNAs were made as described above. Real-time PCR was performed in 96-well titer plates on an ABI PRISM 7000 SDS apparatus with SYBR GREEN PCR Master Mix (ABI) following manufacturers' recommendations (95°C for 5 min., followed by 40 cycles: 95°C for 30 s, 60°C for 45 s). Primers are given in Protocol S1. Uncertainties were propagated from standard errors for triplicate measurements of cDNA pools (derived from column-purified RNA of 3–4 plants).

Supporting Information

Figure S1 Maps of 21, 22 and 24 nt vsRNAs from CaLCuV-infected wild type (Col-0) and *rdr1/2/6* triple mutant plants at single-nucleotide resolution. The graphs plot the number of 21-nt, 22-nt, or 24-nt vsRNA reads at each nucleotide position of the 2583 bp DNA-A (**A**) and the 2513 bp DNA-B (**B**); Bars above the axis represent sense reads starting at each respective position; those below represent antisense reads ending at the respective position (Tables S2 and S3). The genome organizations of DNA-A and DNA-B are shown schematically above the graphs, with leftward (AC1, AC4, AC2, AC3 and BC1) and rightward (AV1 and BV1) ORFs and common region (CR) indicated.
(PDF)

Figure S2 Validation of vsRNA deep-sequencing data and genetic requirements for vsRNA biogenesis. Total RNA isolated from CaLCuV wild type (wt) virus- or CaLCuV::Chl-infected *Arabidopsis* wt (Col-0) plants and various mutants (*rdr2*, *rdr6*, *rdr2/6*, *rdr1/2/6* and *dcl1/2/3/4-caf*; described in Blevins et al, 2006) was analyzed by RNA blot hybridization using 15% PAGE. Membranes were successively hybridized with CaLCuV DNA-A (**A**) and CaLCuV DNA-B (**B**) derived DNA oligonucleotide probes (for sequences, see Protocol S1) or probes specific the endogenous *Arabidopsis* small RNAs (**C**) 22 nt miR173, 21 nt siR255 and 24 nt siR1003. The probes Chl_s and Chl_as in panel A are specific for the *ChlI* gene segment inserted in CaLCuV::Chl DNA-A. EtBr staining of total RNA is shown as loading control. The sizes are indicated on each scan.
(PDF)

Figure S3 Viral and target gene siRNAs in CaLCuV::Chl virus-infected wild type (Col-0) plants. (**A**) The 1961 bp *ChlI-2* genomic locus is shown schematically; numbering starts from the transcription start site. The VIGS target region is highlighted in grey, with the two stretches of >20 nts in length which are identical in *ChlI* and *ChlI-2* shown in red. The graph plots the number of 20–25 nt siRNA reads at each nucleotide position of the *ChlI-2* gene; Bars above the axis represent sense reads starting at each respective position; those below represent antisense reads ending at the respective position (Table S4). (**B**) Alignment of the *ChlI and ChlI-2* sequences containing the VIGS target region is shown below the graph; (**C**) Virus-derived siRNAs. The graphs plot the number of 20–25 nt, 21-nt, 22-nt, or 24-nt vsRNA reads at each nucleotide position of the 2298 bp CaLCuV::Chl DNA-A. Bars above the axis represent sense reads starting at each respective position; those below represent antisense reads ending at the respective position (Tables S4). The genome organization of CaLCuV::Chl DNA is shown schematically above the graphs, with leftward (AC1, AC4, AC2, AC3 and BC1) ORFs and the rightward AV1::Chl chimeric gene and the common region (CR) indicated. The 353 bp *ChlI* gene segment inserted in the multiple cloning site (MCS) of the CaLCuV VIGS vector is highlighted in grey.
(PDF)

Figure S4 VIGS phenotypes and accumulation of primary and secondary siRNAs in L2 *GFP* transgenic plants infected with CaLCuV::GFP viruses. (**A**) The L2 T-DNA region containing the 35S-GFP transgene is shown schematically.

Positions of the duplicated CaMV 35S enhancer and core promoter elements, *GFP* mRNA elements including 5′UTR, translation start (AUG) and stop (UAA) codons and 3′UTR with poly(A) signal (AAUAAA), and 35S terminator sequences are indicated. Numbering is from the T-DNA left border (LB). The VIGS target sequences, inserted in the CaLCuV::GFP viruses *EnhSh*, *CodM*, *CodE* and *CodFL* are indicated with dotted boxes. (**B**) Pictures under UV light of the L2 transgenic plant infected with the *CodFL* virus at 7, 12, 19, 26 and 33 days post-inoculation (dpi) and of the same plant at 40 dpi under UV and day light. Below are pictures under UV light of L2 plants infected with the CaLCuV empty vector and its derivatives *EnhSh*, *CodM* and *CodE*. Sampling of infected tissues of lower leaves (LL) and upper leaves (UL) for RNA preparation was performed as indicated on the left image. (**C**) Blot hybridization analysis of total RNA isolated from plants shown in Panel B. The blot was successively hybridized with short DNA probes specific for 35S::GFP transgene sequences inserted in the CaLCuV::GFP viruses *EnhSh*, *CodM* and *CodE* and for the *GFP* mRNA 3′UTR non-target sequence (3′UTR). EtBr staining serves as loading control. (**D**) Real time quantitative RT-PCR (qPCR) analysis of *GFP* mRNA accumulation in upper leaves of L2 plants infected with infected with the CaLCuV empty vector and its derivatives *EnhSh*, *CodM*, *CodE* (shown in Panel B). Total RNA from non-transgenic wild type *Arabidopsis* (Col-0) was used as a negative control.
(PDF)

Figure S5 Maps of primary and secondary siRNAs accumulating in L2 transgenic plants infected with CaLCuV::GFP viruses that target the *GFP* transcribed region. The graphs plot the number of 21-nt, 22-nt and 24-nt vsRNA reads at each nucleotide position of the L2 T-DNA-based 35S::GFP transgene in L2 transgenic plants infected with the CaLCuV::GFP viruses *Lead* (**A**), *CodM* (**B**), *Trail* (**C**), or *PolyA* (**D**). Bars above the axis represent sense reads starting at each respective position; those below represent antisense reads ending at the respective position (Table S5). The 35S-GFP transgene is shown schematically above the graphs. Positions of the duplicated 35S enhancer and core promoter, *GFP* mRNA elements and 35S terminator are indicated. Numbering is from the T-DNA left border (LB). The VIGS target sequences inserted in the CaLCuV::GFP viruses *Lead*, *CodM*, *Trail* or *polyA* are indicated with dotted boxes.
(PDF)

Figure S6 Maps of primary siRNAs accumulating in L2 transgenic plants infected with CaLCuV::GFP viruses that target the *GFP* promoter elements. The graphs plot the number of 21-nt, 22-nt and 24-nt vsRNA reads at each nucleotide position of the L2 T-DNA-based 35S::GFP transgene in L2 transgenic plants infected with the CaLCuV::GFP viruses *Enh* (**A**), *ProFL* (**B**) or *Core* (**C**). Bars above the axis represent sense reads starting at each respective position; those below represent antisense reads ending at the respective position (Table S5). The 35S-GFP transgene is shown schematically above the graphs. Positions of the duplicated 35S enhancer and core promoter, *GFP* mRNA elements and 35S terminator are indicated. Numbering is from the T-DNA left border (LB). The VIGS target sequences inserted in the CaLCuV::GFP viruses *Enh*, *CodFL* and *Core* are indicated with dotted boxes. Note that the duplicated 35S promoter sequences Enhancer* and Enhancer (each 273 nt long) share 94% nucleotide identity, since they originate from two different strains of CaMV. Therefore, primary siRNA reads are unequally distributed between the two VIGS target regions.
(PDF)

Protocol S1 The file contains the list of DNA oligonucleotides probes for RNA and DNA blot hybridization, primers for subcloning of the 35S::GFP tarnsgene-derived sequences into CaLCuV VIGS vector and for real time PCR as well as Reference sequences used for bioinformatic analysis.
(PDF)

Table S1 Counts of viral and endogenous small RNAs in the Illumina small RNA deep-sequencing libraries for mock-inoculated and wild type CaLCuV-infected Col-0 and *rdr1/2/6* plants (S1A), mock inoculated and CaLCuV::Chl-infected Col-0 plants (S1B), mock inoculated and CaLCuV::*GFP-Pro-FL*-infected Col-0 plants (S1C), mock inoculated and CaLCuV::*GFP-Enh*-infected Col-0 plants (S1D), mock inoculated and CaLCuV::*GFP-Core*-infected Col-0 plants (S1E), mock inoculated and CaLCuV::*GFP-Lead*-infected Col-0 plants (S1F), mock inoculated and CaLCuV::*GFP-CodM*-infected Col-0 plants (S1G), mock inoculated and CaLCuV::*GFP-Trail*-infected Col-0 plants (S1H), and mock inoculated and CaLCuV::*GFP-PolyA*-infected Col-0 plants (S1I).
(XLSX)

Table S2 Single-base resolution maps of 20–25 nt DNA-A derived viral siRNAs in CaLCuV-infected wild type (Col-0) and *rdr1/2/6* triple mutant *Arabidopsis* plants.
(XLS)

Table S3 Single-base resolution maps of 20–25 nt DNA-B derived viral siRNAs in CaLCuV-infected wild type (Col-0) and *rdr1/2/6* triple mutant *Arabidopsis* plants.
(XLS)

Table S4 Single-base resolution maps of 20–25 nt *ChlI/CH-42* and *ChlI-2* derived siRNAs as well as CaLCuV::Chl virus-derived siRNAs in mock inoculated and CaLCuV::Chl-infected *Arabidopsis* plants.
(XLS)

Table S5 Single-base resolution maps of 20–25 nt *L2 GFP T-DNA* derived siRNAs in mock inoculated and CaLCuV::*GFP* virus (*ProFL, Enh, Core, Lead, CodM, Trail*, or *PolyA*)-infected *Arabidopsis* plants.
(XLS)

Table S6 Single-base resolution maps of 20–25 nt viral siRNAs in CaLCuV::*GFP* virus (*ProFL, Enh, Core, Lead, CodM, Trail*, or *PolyA*)-infected *Arabidopsis* plants.
(XLSX)

Acknowledgments

We thank Nachelli Malpica for technical assistance, Thomas Hohn for critical reading of the manuscript and stimulating discussions, Fred Meins for supporting initial experiments of TB at the Friedrich Miescher Institute and Thomas Boller for supporting our research and hosting the groups of MMP and FV at the Botanical Institute.

Author Contributions

Conceived and designed the experiments: MMP TB LF. Performed the experiments: MA BKB JS EGG ASZ RR TB. Analyzed the data: JS LF MA MMP. Contributed reagents/materials/analysis tools: DW FV. Wrote the paper: MMP.

References

1. Vaucheret H (2006) Post-transcriptional small RNA pathways in plants: mechanisms and regulations. Genes Dev 20: 759–71.
2. Ding SW, Voinnet O (2007) Antiviral immunity directed by small RNAs. Cell 130: 413–26.
3. Ghildiyal M, Zamore PD (2009) Small silencing RNAs: an expanding universe. Nat Rev Genet 10: 94–108.
4. Llave C (2010) Virus-derived small interfering RNAs at the core of plant-virus interactions. Trends Plant Sci 15: 701–7.
5. Pantaleo V (2011) Plant RNA silencing in viral defence. Adv Exp Med Biol 722: 39–58.
6. Akbergenov R, Si-Ammour A, Blevins T, Amin I, Kutter C, et al. (2006) Molecular characterization of geminivirus-derived small RNAs in different plant species. Nucleic Acids Res. 34: 462–471.
7. Blevins T, Rajeswaran R, Shivaprasad PV, Beknazariants D, Si-Ammour A, et al. (2006) Four plant Dicers mediate viral small RNA biogenesis and DNA virus induced silencing. Nucleic Acids Res. 34: 6233–6246.
8. Blevins T, Rajeswaran R, Aregger M, Borah BK, Schepetilnikov M, et al. (2011) Massive production of small RNAs from a non-coding region of Cauliflower mosaic virus in plant defense and viral counter-defense. Nucleic Acids Res 39: 5003–14.
9. Deleris A, Gallego-Bartolome J, Bao J, Kasschau KD, Carrington JC, et al. (2006) Hierarchical action and inhibition of plant Dicer-like proteins in antiviral defense. Science 313: 68–71.
10. Bouché N, Lauressergues D, Gasciolli V, Vaucheret H (2006) An antagonistic function for Arabidopsis DCL2 in development and a new function for DCL4 in generating viral siRNAs. EMBO J 25: 3347–56.
11. Mlotshwa S, Pruss GJ, Peragine A, Endres MW, Li J, et al. (2008) DICER-LIKE2 plays a primary role in transitive silencing of transgenes in Arabidopsis. PLoS One 3: e1755.
12. Liu F, Bakht S, Dean C (2012) Cotranscriptional role for Arabidopsis DICER-LIKE 4 in transcription termination. Science 335: 1621–3.
13. Wassenegger M, Krczal G (2006) Nomenclature and functions of RNA-directed RNA polymerases. Trends Plant Sci 11: 142–51.
14. Howell MD, Fahlgren N, Chapman EJ, Cumbie JS, Sullivan CM, et al. (2007) Genome-wide analysis of the RNA-DEPENDENT RNA POLYMERASE6/DICER-LIKE4 pathway in Arabidopsis reveals dependency on miRNA- and tasiRNA-directed targeting. Plant Cell 19: 926–42.
15. Si-Ammour A, Windels D, Arn-Bouldoires E, Kutter C, Ailhas J, et al. (2011) miR393 and secondary siRNAs regulate expression of the TIR1/AFB2 auxin receptor clade and auxin-related development of Arabidopsis leaves. Plant Physiol 157: 683–91.
16. Haag JR, Pikaard CS (2011) Multisubunit RNA polymerases IV and V: purveyors of non-coding RNA for plant gene silencing. Nat Rev Mol Cell Biol 12: 483–92.
17. Montgomery TA, Howell MD, Cuperus JT, Li D, Hansen JE, et al. (2008) Specificity of ARGONAUTE7-miR390 interaction and dual functionality in TAS3 trans-acting siRNA formation. Cell 133: 128–41.
18. Montgomery TA, Yoo SJ, Fahlgren N, Gilbert SD, Howell MD, et al. (2008) AGO1-miR173 complex initiates phased siRNA formation in plants. Proc Natl Acad Sci U S A 105:20055–62.
19. Rajeswaran R, Pooggin MM (2012) RDR6-mediated synthesis of complementary RNA is terminated by miRNA stably bound to template RNA. Nucleic Acids Res 40: 594–9.
20. Rajeswaran R, Aregger M, Zvereva AS, Borah BK, Gubaeva EG, et al. (2012) Sequencing of RDR6-dependent double-stranded RNAs reveals novel features of plant siRNA biogenesis. Nucleic Acids Res 40: 6241–54.
21. Cuperus JT, Carbonell A, Fahlgren N, Garcia-Ruiz H, Burke RT, et al. (2010) Unique functionality of 22-nt miRNAs in triggering RDR6-dependent siRNA biogenesis from target transcripts in Arabidopsis. Nat Struct Mol Biol 17: 997–1003.
22. Chen HM, Chen LT, Patel K, Li YH, Baulcombe DC, et al. (2010) 22-Nucleotide RNAs trigger secondary siRNA biogenesis in plants. Proc Natl Acad Sci U S A 107: 15269–74.
23. Manavella PA, Koenig D, Weigel D (2012) Plant secondary siRNA production determined by microRNA-duplex structure. Proc Natl Acad Sci U S A 109: 2461–6.
24. Donaire L, Barajas D, Martínez-García B, Martínez-Priego L, Pagán I, et al. (2008) Structural and genetic requirements for the biogenesis of tobacco rattle virus-derived small interfering RNAs. J Virol 82: 5167–77.
25. Qi X, Bao FS, Xie Z (2009) Small RNA deep sequencing reveals role for Arabidopsis thaliana RNA-dependent RNA polymerases in viral siRNA biogenesis. PLoS One 4: e4971.
26. Wang XB, Wu Q, Ito T, Cillo F, Li WX, et al. (2010) RNAi-mediated viral immunity requires amplification of virus-derived siRNAs in Arabidopsis thaliana. Proc Natl Acad Sci U S A 107: 484–9.
27. Wang XB, Jovel J, Udomporn P, Wang Y, Wu Q, et al. (2011) The 21-nucleotide, but not 22-nucleotide, viral secondary small interfering RNAs direct potent antiviral defense by two cooperative argonautes in Arabidopsis thaliana. Plant Cell 23:1625–38.
28. Garcia-Ruiz H, Takeda A, Chapman EJ, Sullivan CM, Fahlgren N, et al. (2010) Arabidopsis RNA-dependent RNA polymerases and dicer-like proteins in antiviral defense and small interfering RNA biogenesis during Turnip Mosaic Virus infection. Plant Cell 22:481–96.
29. Rajeswaran R, Pooggin MM (2012) Role of virus-derived small RNAs in plant antiviral defense: insights from DNA viruses. In MicroRNAs in Plant development and Stress Response. R. Sunkar, ed. Heidelberg: Springer. pp. 261–289.
30. Shivaprasad PV, Rajeswaran R, Blevins T, Schoelz J, Meins F Jr, et al. (2008) The CaMV transactivator/viroplasmin interferes with RDR6-dependent trans-acting and secondary siRNA pathways in Arabidopsis. Nucleic Acids Res 36: 5896–909.
31. Haas G, Azevedo J, Moissiard G, Geldreich A, Himber C, et al. (2008) Nuclear import of CaMV P6 is required for infection and suppression of the RNA silencing factor DRB4. EMBO J 27: 2102–12.
32. Jeske H (2009) Geminiviruses. Curr Top Microbiol Immunol 331: 185–226.
33. Hill JE, Strandberg JO, Hiebert E, Lazarowitz SG (1998) Asymmetric infectivity of pseudorecombinants of cabbage leaf curl virus and squash leaf curl virus: implications for bipartite geminivirus evolution and movement. Virology 250:283–92.
34. Shivaprasad PV, Akbergenov R, Trinks D, Rajeswaran R, Veluthambi K, et al. (2005) Promoters, transcripts, and regulatory proteins of Mungbean yellow mosaic geminivirus. J Virol 79: 8149–63.
35. Chellappan P, Vanitharani R, Pita J, Fauquet CM (2004) Short interfering RNA accumulation correlates with host recovery in DNA virus-infected hosts, and gene silencing targets specific viral sequences. J Virol 78:7465–77.
36. Vaistij FE, Jones L, Baulcombe DC (2002) Spreading of RNA targeting and DNA methylation in RNA silencing requires transcription of the target gene and a putative RNA-dependent RNA polymerase. Plant Cell 14: 857–67.
37. Daxinger L, Kanno T, Bucher E, van der Winden J, Naumann U, et al. (2009) A stepwise pathway for biogenesis of 24-nt secondary siRNAs and spreading of DNA methylation. EMBO J 28:48–57.
38. Himber C, Dunoyer P, Moissiard G, Ritzenthaler C, Voinnet O (2003) Transitivity-dependent and -independent cell-to-cell movement of RNA silencing. EMBO J 22: 4523–33.
39. Petersen BO, Albrechtsen M (2005) Evidence implying only unprimed RdRP activity during transitive gene silencing in plants. Plant Mol Biol 58:575–83.
40. Lu C, Kulkarni K, Souret FF, MuthuValliappan R, Tej SS, et al. (2006) MicroRNAs and other small RNAs enriched in the Arabidopsis RNA-dependent RNA polymerase-2 mutant. Genome Res 16: 1276–88.
41. Kasschau KD, Fahlgren N, Chapman EJ, Sullivan CM, Cumbie JS, et al. (2007) Genome-wide profiling and analysis of Arabidopsis siRNAs. PLoS Biol 5: e57.
42. Mi S, Cai T, Hu Y, Chen Y, Hodges E, et al. (2008) Sorting of small RNAs into Arabidopsis argonaute complexes is directed by the 5′ terminal nucleotide. Cell 133: 116–27.
43. Takeda A, Iwasaki S, Watanabe T, Utsumi M, Watanabe Y (2008) The mechanism selecting the guide strand from small RNA duplexes is different among argonaute proteins. Plant Cell Physiol 49:493–500.
44. Havecker ER, Wallbridge LM, Hardcastle TJ, Bush MS, Kelly KA, et al. (2010) The Arabidopsis RNA-directed DNA methylation argonautes functionally diverge based on their expression and interaction with target loci. Plant Cell 22: 321–34.
45. Donaire L, Wang Y, Gonzalez-Ibeas D, Mayer KF, Aranda MA, et al. (2009) Deep-sequencing of plant viral small RNAs reveals effective and widespread targeting of viral genomes. Virology 392: 203–14.
46. Kreuze JF, Perez A, Untiveros M, Quispe D, Fuentes S, 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.
47. Yang X, Wang Y, Guo W, Xie Y, Xie Q, et al. (2011) Characterization of small interfering RNAs derived from the geminivirus/betasatellite complex using deep sequencing. PLoS One 6: e16928.
48. Hu Q, Hollunder J, Niehl A, Kørner CJ, Gereige D, et al. (2011) Specific impact of tobamovirus infection on the Arabidopsis small RNA profile. PLoS One 6: e19549.
49. Wei W, Ba Z, Gao M, Wu Y, Ma Y, et al. (2012) A role for small RNAs in DNA double-strand break repair. Cell 149: 101–12.
50. Turnage MA, Muangsan N, Peele CG, Robertson D (2002) Geminivirus-based vectors for gene silencing in Arabidopsis. Plant J 30: 107–114.
51. Molnar A, Melnyk CW, Bassett A, Hardcastle TJ, Dunn R, et al. (2010) Small silencing RNAs in plants are mobile and direct epigenetic modification in recipient cells. Science 328:872–5.
52. Molnar A, Melnyk C, Baulcombe DC (2011) Silencing signals in plants: a long journey for small RNAs. Genome Biol 12:215.
53. Axtell MJ, Jan C, Rajagopalan R, Bartel DP (2006) A two-hit trigger for siRNA biogenesis in plants. Cell 127: 565–77.
54. Glazov E, Phillips K, Budziszewski GJ, Schöb H, Meins F Jr, et al. (2003) A gene encoding an RNase D exonuclease-like protein is required for post-transcriptional silencing in Arabidopsis. Plant J 35:342–9. Erratum in: Plant J. 36: 741.
55. Moissiard G, Parizotto EA, Himber C, Voinnet O (2007) Transitivity in Arabidopsis can be primed, requires the redundant action of the antiviral Dicer-

like 4 and Dicer-like 2, and is compromised by viral-encoded suppressor proteins. RNA 13:1268–78.
56. García-Pérez RD, Houdt HV, Depicker A (2004) Spreading of post-transcriptional gene silencing along the target gene promotes systemic silencing. Plant J 38: 594–602.
57. Vermeersch L, De Winne N, Depicker A (2010) Introns reduce transitivity proportionally to their length, suggesting that silencing spreads along the pre-mRNA. Plant J 64: 392–401.
58. Jones L, Ratcliff F, Baulcombe DC (2001) RNA-directed transcriptional gene silencing in plants can be inherited independently of the RNA trigger and requires Met1 for maintenance. Curr Biol 11: 747–57.
59. Peragine A, Yoshikawa M, Wu G, Albrecht HL, Poethig RS (2004) SGS3 and SGS2/SDE1/RDR6 are required for juvenile development and the production of trans-acting siRNAs in Arabidopsis. Genes Dev 18: 2368–79.
60. Yoshikawa M, Peragine A, Park MY, Poethig RS (2005) A pathway for the biogenesis of trans-acting siRNAs in Arabidopsis. Genes Dev 19: 2164–75.
61. Pak J, Fire A (2007) Distinct populations of primary and secondary effectors during RNAi in C. elegans. Science 315: 241–4.
62. Gu SG, Pak J, Guang S, Maniar JM, Kennedy S, et al. (2012) Amplification of siRNA in Caenorhabditis elegans generates a transgenerational sequence-targeted histone H3 lysine 9 methylation footprint. Nat Genet 44: 157–64.
63. Sanfaçon H, Hohn T (1990) Proximity to the promoter inhibits recognition of cauliflower mosaic virus polyadenylation signal. Nature 346: 81–4.
64. Rothnie HM, Chapdelaine Y, Hohn T (1994) Pararetroviruses and retroviruses: a comparative review of viral structure and gene expression strategies. Adv Virus Res 44: 1–67.
65. Li H. and Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics 25: 1754–60.
66. Mahajan N, Parameswari C, Veluthambi K (2011) Severe stunting in blackgram caused by the Mungbean yellow mosaic virus (MYMV) KA27 DNA B component is ameliorated by co-infection or post-infection with the KA22 DNA B: MYMV nuclear shuttle protein is the symptom determinant. Virus Res 157: 25–34.

Annex: (Rajeswaran et al., 2014a)

Evasion of Short Interfering RNA-Directed Antiviral Silencing in Musa acuminata Persistently Infected with Six Distinct Banana Streak Pararetroviruses

Rajendran Rajeswaran, Jonathan Seguin, Matthieu Chabannes, Pierre-Olivier Duroy, Nathalie Laboureau, Laurent Farinelli, Marie-Line Iskra-Caruana and Mikhail M. Pooggin

Journal of Virology (2014), Vol.88, Issue 19

JVI
Journals.ASM.org

Evasion of Short Interfering RNA-Directed Antiviral Silencing in *Musa acuminata* Persistently Infected with Six Distinct Banana Streak Pararetroviruses

Rajendran Rajeswaran,[a]* Jonathan Seguin,[a,b] Matthieu Chabannes,[c] Pierre-Olivier Duroy,[c]* Nathalie Laboureau,[c] Laurent Farinelli,[b] Marie-Line Iskra-Caruana,[c] Mikhail M. Pooggin[a]

University of Basel, Department of Environmental Sciences, Botany, Basel, Switzerland[a]; FASTERIS SA, Plan-les-Ouates, Switzerland[b]; CIRAD, UMR BGPI, Montpellier, France[c]

ABSTRACT

Vegetatively propagated crop plants often suffer from infections with persistent RNA and DNA viruses. Such viruses appear to evade the plant defenses that normally restrict viral replication and spread. The major antiviral defense mechanism is based on RNA silencing generating viral short interfering RNAs (siRNAs) that can potentially repress viral genes posttranscriptionally through RNA cleavage and transcriptionally through DNA cytosine methylation. Here we examined the RNA silencing machinery of banana plants persistently infected with six pararetroviruses after many years of vegetative propagation. Using deep sequencing, we reconstructed consensus master genomes of the viruses and characterized virus-derived and endogenous small RNAs. Consistent with the presence of endogenous siRNAs that can potentially establish and maintain DNA methylation, the banana genomic DNA was extensively methylated in both healthy and virus-infected plants. A novel class of abundant 20-nucleotide (nt) endogenous small RNAs with 5′-terminal guanosine was identified. In all virus-infected plants, 21- to 24-nt viral siRNAs accumulated at relatively high levels (up to 22% of the total small RNA population) and covered the entire circular viral DNA genomes in both orientations. The hotspots of 21-nt and 22-nt siRNAs occurred within open reading frame (ORF) I and II and the 5′ portion of ORF III, while 24-nt siRNAs were more evenly distributed along the viral genome. Despite the presence of abundant viral siRNAs of different size classes, the viral DNA was largely free of cytosine methylation. Thus, the virus is able to evade siRNA-directed DNA methylation and thereby avoid transcriptional silencing. This evasion of silencing likely contributes to the persistence of pararetroviruses in banana plants.

IMPORTANCE

We report that DNA pararetroviruses in *Musa acuminata* banana plants are able to evade DNA cytosine methylation and transcriptional gene silencing, despite being targeted by the host silencing machinery generating abundant 21- to 24-nucleotide short interfering RNAs. At the same time, the banana genomic DNA is extensively methylated in both healthy and virus-infected plants. Our findings shed light on the siRNA-generating gene silencing machinery of banana and provide a possible explanation why episomal pararetroviruses can persist in plants whereas true retroviruses with an obligatory genome-integration step in their replication cycle do not exist in plants.

Viruses are often described as causing acute and persistent infections: acute virus infections cause severe disease symptoms and, in many cases, kill the host plant, whereas persistent virus infections normally cause mild symptoms and allow the host to recover from the disease; persistent virus infections can sometimes be beneficial for plant physiology (1). Depending on the environmental conditions and host plant species, the viral disease symptoms can oscillate from severe to nonvisible (recovery) and back. The molecular mechanisms underlying such oscillations in persistent virus infections are poorly understood, although an antiviral defense system based on RNA silencing has been implicated in plant recovery from DNA and RNA virus infections (2, 3).

RNA silencing, also known as RNA interference (RNAi), is an evolutionarily conserved sequence-specific mechanism which regulates gene expression and chromatin states and also defends against invasive nucleic acids such as transposons, transgenes, and viruses (4–7). It is induced by double-stranded RNA (dsRNA), which can be produced by DNA-dependent RNA polymerase II (Pol II) transcribing inverted repeats or certain genomic regions in both sense orientation and antisense orientation and, in some organisms, including plants, also by RNA-dependent RNA polymerase (RDR). In most eukaryotes, Dicer or Dicer-like (DCL) enzymes catalyze processing of dsRNA into small RNA (sRNA) duplexes, which are then sorted by Argonaute (AGO) family proteins. AGO forms the RNA-induced silencing complex (RISC) with one of the duplex strands. This single-stranded sRNA guides RISC to a complementary sequence in a target RNA. Following the complementary interaction, AGO catalyzes cleavage and/or translational repression of the target RNA, which results in posttran-

Received 24 May 2014 Accepted 18 July 2014
Published ahead of print 23 July 2014
Editor: A. Simon
Address correspondence to Mikhail M. Pooggin, Mikhail.Pooggin@unibas.ch.
* Present address: Rajendran Rajeswaran, Swiss Federal Institute of Technology Zurich (ETH-Zurich), Department of Biology, Zurich, Switzerland; Pierre-Olivier Duroy, Bayer, Lyon, France.
Supplemental material for this article may be found at http://dx.doi.org/10.1128/JVI.01496-14.

doi:10.1128/JVI.01496-14

scriptional gene silencing (PTGS). In plants, fungi, and some animals, the RNA silencing machinery can also repress genes transcriptionally through DNA cytosine methylation and/or chromatin histone modification. In plants, transcriptional gene silencing (TGS) through sRNA-directed DNA methylation (RdDM) plays a key role in inactivation of transposons (8, 9). RdDM and TGS have also been implicated in defense against DNA geminiviruses (10) and pararetroviruses (11), although an involvement of RdDM is a matter of ongoing debate (12, 13).

Plants have evolved diverse sRNA-generating silencing pathways mediated by four distinct DCL enzymes. In *Arabidopsis*, DCL1 catalyzes processing of 21-to-22-nucleotide (nt) microRNAs (miRNAs) from hairpin structures of miRNA gene transcripts, while DCL4, DCL2, and DCL3 produce short interfering RNAs (siRNAs) of three major size classes (21 nt, 22 nt, and 24 nt, respectively) from perfect dsRNA precursors. DCL4 and DCL2 play a primary role in defense against RNA viruses by producing 21-nt and 22-nt viral siRNAs, respectively (14), whereas all the four DCLs generate DNA virus-derived 21-, 22-, and 24-nt siRNAs (15–18). Nuclear DCL3 generates highly abundant 24-nt viral siRNAs in *Arabidopsis* infected with the DNA geminivirus *Cabbage leaf curl virus* (CaLCuV) (17) and the pararetrovirus *Cauliflower mosaic virus* (CaMV) (16), which can potentially direct viral DNA methylation and TGS. However, CaMV-derived 24-nt siRNAs were not found to be associated with AGO4, the major effector protein in RdDM (16). This finding and other lines of evidence suggest that DNA viruses may evade siRNA-directed DNA methylation (13). The genetic requirements for viral siRNA biogenesis and antiviral defense have been investigated mostly in model plants such as *Arabidopsis thaliana* and *Nicotiana benthamiana*. Nonetheless, the antiviral silencing pathways appear to be conserved across the plant kingdom, since model and crop plants infected with various RNA and DNA viruses accumulate abundant viral siRNAs of the three major size classes (15–21).

However, very little is known about RNA silencing pathways and antiviral defenses in crop plants, including banana. Although the genome of *Musa acuminata* has been sequenced and conserved miRNAs have been identified *in silico* (22), banana RNA silencing genes and sRNA profiles have not been reported so far. Recently, RNAi transgenic banana plants which express inverted-repeat transgene-derived siRNAs cognate to *Banana bunchy top virus* (BBTV; a single-stranded DNA virus from the family *Nanoviridae*) were generated and shown to be resistant to BBTV infection (23). This suggests that the banana silencing machinery has the potential for antiviral defense.

Most cultivars of dessert bananas, plantains, and cooking bananas (*Musa* spp.) are derived from intra- or interspecific hybrids of the seedy banana species *Musa acuminata* (denoted the A genome) and *Musa balbisiana* (denoted the B genome). Banana streak disease was first described more than 50 years ago in Ivory Coast as the cause of characteristic leaf stripes on *M. acuminata* (24). Subsequently, the disease was recorded in most *Musa*-producing regions and in many *Musa* spp., and the causal agent was named banana streak virus (BSV) (25). BSV disease is transmitted by mealybugs (26, 27) but can also be spread by vegetative propagation of infected material. Significant variation in the severity of banana streak disease has been observed in different regions of the world, but the relative contributions of plant, virus, and environment to this variation are largely unknown (28). *Banana streak Obino l'Ewai virus* (BSOLV) from the triploid AAB plantain cv. Obino l'Ewai was the first sequenced BSV species associated with the disease (29). The circular 7.4-kb genome of BSOLV contains three consecutive open reading frames (ORFs), the first two encoding small (20.8-kDa and 14.5-kDa) proteins of unknown function and the third coding for a large (208-kDa) polyprotein consisting of a putative movement protein, an RNA binding coat protein (CP [analogous to retroviral Gag]), aspartyl proteinase, and a viral replicase (analogous to retroviral Pol) consisting of reverse transcriptase (RT) and RNase H domains. Based on the genome organization, protein functions, and other features, BSV is classified into genus *Badnavirus* (bacilliform DNA virus) of the family *Caulimoviridae*. This family comprises nine genera of plant pararetroviruses that replicate via reverse transcription of pregenomic RNA (pgRNA) and encapsidate open circular double-stranded (ds) genomic DNA into bacilliform or icosahedral virions (30, 31). According to an interpretation based mostly on extensive studies of CaMV from genus *Caulimovirus* (31, 32), the viral life cycle begins with a release of viral double-stranded DNA (dsDNA) from virions into the nucleus, where the gaps on both strands remaining after reverse transcription are repaired and the resulting covalently closed dsDNA associates with histones to form a minichromosome (episome). Pol II-mediated transcription of the circular minichromosome generates a capped and polyadenylated pgRNA with terminal repeats which is transported to the cytoplasm for translation of viral proteins and subsequently for reverse transcription. As demonstrated for CaMV and *Rice tungro bacilliform virus* (RTBV; genus *Tungrovirus*), translation of pgRNA is initiated by a shunt mechanism in which ribosomes bypass a long leader sequence containing multiple short ORFs (sORFs) and folding into a stable stem-loop structure (33–35), both features conserved in plant pararetroviruses (36). Several consecutive viral ORFs on polycistronic pgRNA are then translated by reinitiation (CaMV) (37) or leaky-scanning (RTBV) (38) mechanisms. The reverse transcription takes place in cytoplasmic inclusion bodies and presumably is initiated by CP-mediated packaging of pgRNA (39). The resulting open circular dsDNA can reenter the nucleus for the next round of transcription or be encapsidated into mature virions for movement within the host plant and for transmission to a new host plant by insect vectors. At late stages of pararetrovirus infection, plant cells accumulate high copy numbers of both circular covalently closed viral dsDNA in the nucleus and open circular dsDNA in the virions (31).

In contrast to retroviruses, pararetroviruses do not encode integrase and their life cycle does not require an integration step into the host genome. However, some plant pararetroviruses, including BSV, were able to get integrated into the host genomes, likely through a process of illegitimate recombination, and most of these integrants appear to be noninfectious relics of ancient infection events (40, 41). This is the case in the banana A genome, which contains multiple integrants of badnavirus-like sequences, but those contain only incomplete, highly rearranged, and fragmented genomes that exhibit distant similarity to the infectious BSV species described to date, limiting the possibility of giving rise to episomal virus infections (22, 40, 42, 43). In contrast, more recent BSV integrants in the B genome, including BSOLV (44), *Banana streak Goldfinger virus* (BSGFV) (45), *Banana streak Imove virus* (BSIMV) (41, 46), and, likely, *Banana streak Mysore virus* (BSMYV) (27), still retain infectivity. Taking the findings together, natural BSV infections of *M. acuminata* spp. and their intraspecific hybrids are possible only through insect transmission

of a virus that had recombined out of the banana B genome in interspecific hybrids (e.g., BSOLV, BSGFV, BSIMV, or BSMYV) and/or of a nonintegrated virus persisting in *Musa* spp. or other hosts. Such persistent BSV species that seem to have no infectious counterparts in the genomes of *Musa* spp. include *Banana streak Vietnam virus* (BSVNV) (47) and *Banana streak Cavendish virus* (BSCAV) (48). The defense mechanisms induced in BSV-infected *Musa* spp. and viral counterdefense strategies have not been investigated so far.

In this study, we characterized the sRNA-generating silencing machinery in persistently infected banana plants and examined whether or not episomal pararetroviruses are able to evade siRNA-directed DNA methylation. We used rolling circle amplification (RCA) and sequencing of viral DNA combined with sRNA deep sequencing to reconstruct the complete genomes and siRNA profiles for six episomal BSV species (BSOLV, BSGFV, BSIMV, BSMYV, BSVNV, and BSCAV) that persisted in *M. acuminata* triploid (AAA) banana plants over 10 years of vegetative propagation following transmission by mealybugs. We found that BSV infection induces multiple silencing pathways generating 21-, 22-, and 24-nt viral siRNAs which can potentially be associated with AGOs to target the viral genome. Despite the accumulation of 24-nt siRNAs covering the entire virus genome in sense and antisense orientations, the bulk of viral circular covalently closed dsDNA serving as the template for pgRNA transcription in the nucleus is largely not methylated. This implies that BSV evades RdDM and thereby accumulates multiple transcriptionally active minichromosomes in infected banana plants. Such evasion of viral DNA methylation and transcriptional silencing may contribute to persistence of episomal BSV in host plants.

MATERIALS AND METHODS

Banana plants and viruses. *M. acuminata* dwarf Cavendish plants (triploid AAA genome) were inoculated in 2000 by mealybugs (*Planococcus citri*) fed on fresh leaves of banana plants infected with a single BSV species each (Ben Lockhart, personal communication). Ben Lockhart's collection of the BSV-infected banana plants established in 1992 to 1993 was used as a source of inoculums for BSOLV, BSGFV, BSIMV, BSCAV, and BSMYV, while BSVNV came from plant ITC 1431 (*M. acuminata siamea*) (47). The resulting BSV-infected dwarf Cavendish collection was maintained at CIRAD (Montpellier, France) in a tropical greenhouse by vegetative propagation (i.e., growing of infected suckers) under the following conditions: 12 h of daylight with luminosity not exceeding 400 W/m^2, 75% relative humidity, and temperatures of 26°C during the day and 24°C at night. The plants were regularly checked for BSV infection by immune capture-PCR as described earlier (49). The disease severity and virus load were oscillating depending on the season, with strong decreases in winter and intense multiplication in spring and summer. The leaves displaying the characteristic streak disease symptoms (Fig. 1A) were collected in November 2010. One of the two plants infected with BSGFV (the plant sample designated BPO-61 in Datasets S1 to S3 in the supplemental material) was found to be coinfected with BSCAV, which had accumulated much lower levels of viral DNA and viral siRNAs than BSGFV. Other plants proved to be infected with single BSV species.

Total RNA and total DNA preparations. The same banana leaf tissue was ground in liquid nitrogen and taken for preparation of total RNA and total DNA. Total RNA for both Illumina deep-sequencing (Fig. 2 and 3) and blot hybridization (Fig. 4) analyses of sRNAs was extracted from the banana leaves as described in reference 50. Briefly, 2 g of banana leaf tissue ground in liquid nitrogen was added to 10 ml of extraction buffer (100 mM Tris, 500 mM NaCl, 25 mM EDTA, 1.5% SDS, 2% polyvinylpyrrolidone [PVP], 0.7% 2-mercaptoethanol). The mixture was subjected to a vortex procedure, incubated at room temperature for 10 min, and centrifuged at 3,700 rpm for 15 min to pellet the debris. A one-third volume of precooled 5 M sodium acetate (pH 6) was added to the supernatant, mixed and incubated on ice for 30 min, and centrifuged at 10,000 rpm for 15 min at 4°C. The supernatant was treated with an equal volume of phenol-chloroform-isoamyl alcohol and centrifuged at 10,000 rpm for 10 min at 4°C. The supernatant was extracted with an equal volume of chloroform-isoamyl alcohol and centrifuged as described above. RNA in the aqueous phase was precipitated by addition of 2 to 3 volumes of cold ethanol, incubated at −70°C for 30 min, and pelleted by spinning at 10,000 rpm for 15 min at 4°C. RNA was dissolved in 500 μl of diethyl pyrocarbonate (DEPC)-treated water, extracted once again with chloroform-isoamyl alcohol, and precipitated with ethanol as described above. The RNA pellet was washed with 70% ethanol, air dried, dissolved in DEPC-treated water, and stored at −80°C until use.

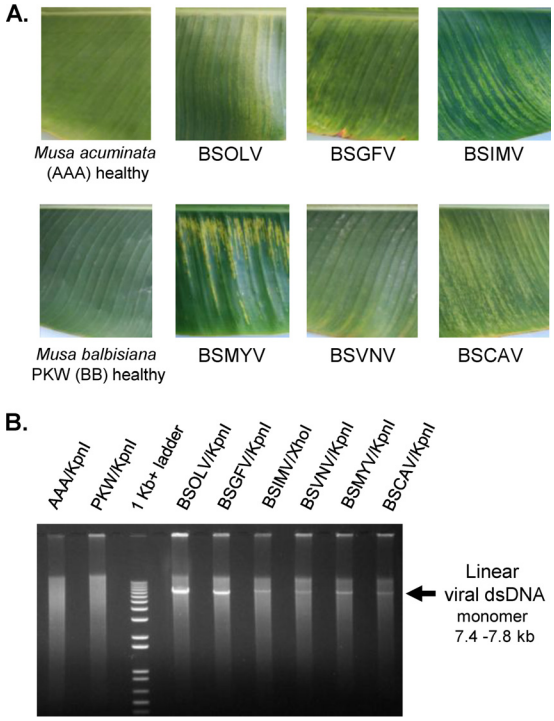


FIG 1 Leaf samples of BSV-infected and healthy control banana plants (A) and RCA analysis of viral DNA (B). Cropped pictures of the leaves from the healthy control *M. acuminata* cv. Cavendish (AAA) and *M. balbisiana* Pisang Klutuk Wulung (PKW) (BB) plants and the *M. acuminata* plants individually infected with BSOLV, BSGFV, BSIMV, BSMYV, BSVNV, and BSCAV are shown. Total DNA was taken for rolling circle amplification (RCA) reactions supplemented with a set of degenerate primers which are listed in Table S1 in the supplemental material. The resulting RCA products were digested with KpnI (except for BSIMV, where XhoI was used), separated on a 1% agarose gel, and stained with ethidium bromide (EtBr). As the DNA size marker, a 1-Kb+ ladder was used. The linear monomeric viral dsDNAs of expected sizes (from 7.4 to 7.8 kb) are indicated by arrows.

Total DNA for both rolling circle amplification (Fig. 1B) and Southern blot hybridization (Fig. 5) analyses was isolated as described by Gawel and Jarret (51). In short, 100 mg of banana leaf tissues ground in liquid nitrogen was incubated with 500 μl of extraction buffer (100 mM Tris, 1.4 M NaCl, 20 mM EDTA, 2% mixed alkyl trimethyl ammonium bromide (MATAB; Sigma), 1% polyethylene glycol [PEG] 6000, 0.5% sodium sulfite) preheated up to 74°C. The slurry was mixed and incubated at 74°C for 20 min, and 500 μl of chloroform-isoamyl alcohol was added. The contents were mixed and centrifuged at 14,000 rpm for 15 min. DNA in the aqueous phase was precipitated with an equal volume of cold isopropanol and pelleted by centrifugation at 14,000 rpm for 30 min at 4°C. After a

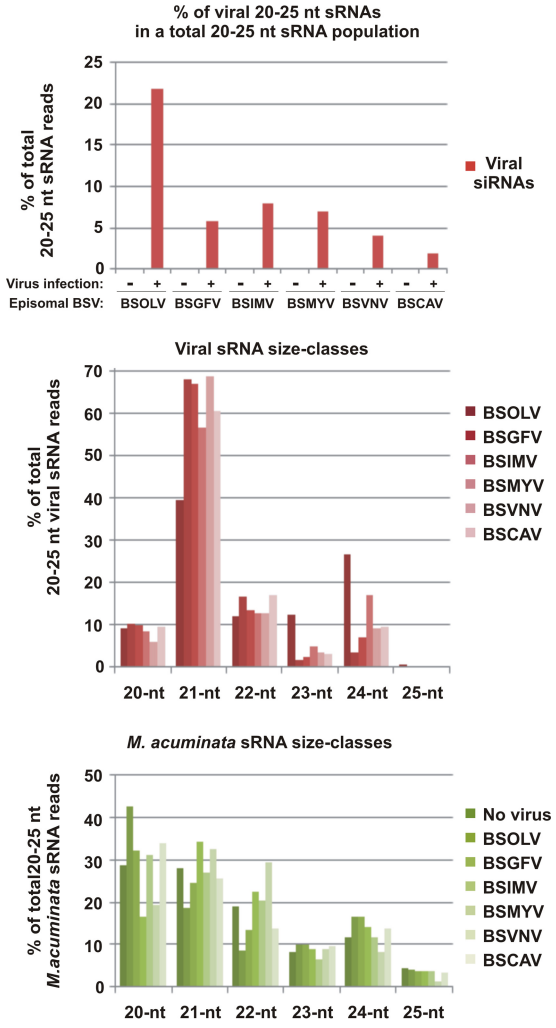


FIG 2 Characterization of viral and endogenous banana sRNAs in *M. acuminata* plants individually infected with six distinct BSV species and healthy control plants. The graphs show the percentages of 20-to-25-nt viral sRNAs in the pool of total (host and viral) 20-to-25-nt reads mapped to the *M. acuminata* and the genomes of the respective viruses (BSOLV, BSGFV, BSIMV, BSMYV, BSVNV, and BSCAV) with zero mismatches, the percentages of each size class of 20-to-25-nt viral sRNA reads mapped to each of the BSV species genomes with zero mismatches, and the percentages of each size class of 20-to-25-nt banana sRNA reads mapped to the genome of virus-free and BSV-infected plants with zero mismatches.

wash with 75% ethanol, the DNA pellet was dried by the use of a speed vacuum and dissolved in 100 μl of water.

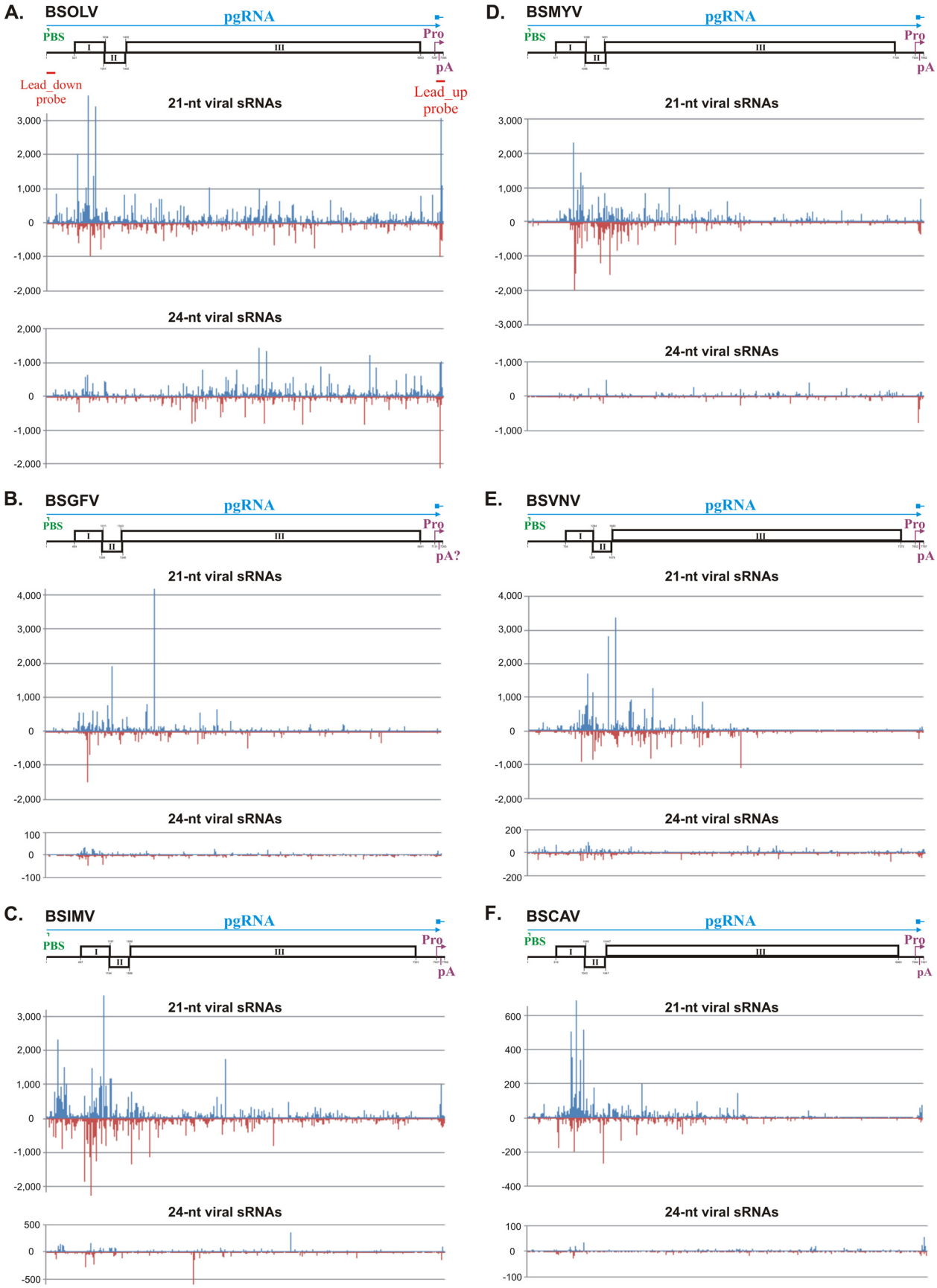
RCA and Southern blot hybridization. Total DNA was taken for rolling circle amplification (RCA) using an illustra TempliPhi amplification kit (GE Healthcare Life Sciences) and a set of degenerate primers (see Table S1 in the supplemental material) according to the protocol described in reference 48. Briefly, ~25 ng of total DNA and 1 μl of degenerate primer mix (4.16 pmol/μl of each primer; see Table S1) were added to 5 μl of sample buffer from the kit and the mixture was heated to 95°C for 3 min. The mixture was cooled on ice, 5 μl of reaction buffer from the kit (premixed with 0.2 μl of bacteriophage Phi29 DNA polymerase) was added, and the reaction was allowed to proceed for 18 h at 30°C. The reaction was stopped by incubation at 65°C for 10 min, and an aliquot of the RCA products was digested with KpnI (except for BSIMV, where XhoI was used) and separated on 1% agarose gel (Fig. 1B).

For McrBC (New England BioLabs) treatment and subsequent Southern blot hybridization, 10 μg total DNA was taken and digested with 30 U of McrBC enzyme overnight at 37°C as recommended by the manufacturer. As a positive control for McrBC analysis, 0.5 μg of methylated plasmid (with one McrBC site; supplied by the manufacturer) was spiked in 10 μg of total DNA isolated from the *M. balbisiana* plant (see Fig. 5; arrows indicate the two products of McrBC digestion of the methylated plasmid). The nontreated total DNA samples represented in Fig. 5 were incubated in parallel under the same conditions as the McrBC-treated total DNA samples but without the McrBC enzyme.

Following the treatment with or without McrBC, the DNA of each total reaction mixture was separated in one 1% agarose gel in 1× TNE buffer (40 mM Tris-acetate, 20 mM sodium acetate, 2 mM EDTA, pH 7.5), stained with EtBr (Fig. 5, lower part), and then transferred onto a Hybond N+ membrane (Amersham). The membrane was hybridized overnight at 45°C in UltraHyb-oligo buffer (Ambion) with a mixture of probes specific for each BSV. Two 40-nt (or 41-nt) oligonucleotides of same GC content and melting point (see Table S1 in the supplemental material) were designed for each of the six BSV species, and all the 12 oligonucleotides were pooled and end labeled with P32 by the use of polynucleotide kinase for hybridization. After 16 h of hybridization, the blot was washed two times with 2× SSC (1× SSC is 0.15 M NaCl plus 0.015 M sodium citrate)–0.5% SDS for 30 min at 45°C and the signal was detected after 20 h to 5 days of exposure to a phosphor screen using a Molecular Imager (Typhoon FLA 7000; GE Healthcare Life Sciences). For repeated hybridizations, the membrane was stripped with 0.5× SSC–0.5% SDS for 30 min at 80°C and then with 0.1× SSC–0.5% SDS for 30 min at 80°C. To enhance the signal of the viral supercoiled DNA form, the membrane was stripped and rehybridized at 35°C. Note that, due to low titers of viral DNA, the level of supercoiled DNA for two of the six viruses analyzed (namely, BSIMV and BSVNV) was below the detection threshold.

Illumina deep-sequencing and bioinformatic analysis of viral nucleic acids. The RCA products were fragmented by the use of a Bioruptor with settings for 350-bp fragments. The DNA fragments were separated using a 2% agarose gel, extracted, and sequenced following the Illumina protocols using an Illumina Hi-Seq 2000 Genome Analyzer (GA) (a 1-by-50-bp run in a v3 flow cell) and a TruSeq SBS kit (v3). After the removal of adaptors, the data sets of 50 reads were used for *de novo* reconstruction of BSV genomes and identification of single nucleotide polymorphisms (SNPs) and indels (insertions/deletions). To reconstruct viral genomes, the reads were assembled into contigs using Velvet 1.2.07 (https://www.ebi.ac.uk/~zerbino/velvet/) (52) followed by Oases 0.2.08 (http://www.ebi.ac.uk/~zerbino/oases/) (53). Oases contigs were merged using the Seqman module of the Lasergene DNASTAR 8.1.2 Core Suite (DNAStar, Madison, WI). SNP/indel calling and correction of errors in the viral genomes were done using Integrative Genomics Viewer (IGV; www.broadinstitute.org/igv) (54) with redundant and nonredundant reads. The reconstructed virus genomes and SNPs were further verified using the data sets of sRNA reads obtained by deep sequencing (see below).

For sRNA deep sequencing, cDNA libraries of 19-to-30-nt RNA fractions of the total RNA samples were prepared as described previously (16). The libraries were sequenced on a Hi-Seq 2000 GA using a TruSeq kit (v5). After the adaptor sequences were trimmed, the data sets of reads were mapped to the reference genome sequences of the DH-Pahang subspecies *malaccensis Musa acuminata* plant (22) and the six reconstructed BSV species using a Burrows-Wheeler Alignment Tool (BWA version 0.5.9) (55) with zero mismatches to each reference sequence. The results of the bioinformatics analysis of the mapped reads are summarized in Fig. 2 and 3 and Datasets S1 to S3 in the supplemental material. Reads mapping to several positions on the reference genome were attributed at random to one of them. To account for the circular BSV genome, the first 50 bases of the viral sequence were added to its 3′ end. For each reference genome or sequence and each sRNA size class (20 to 25 nt), we counted the total number of reads, the numbers of reads in the forward and reverse orientations, and the numbers of reads starting with A, C, G, and T (see Dataset



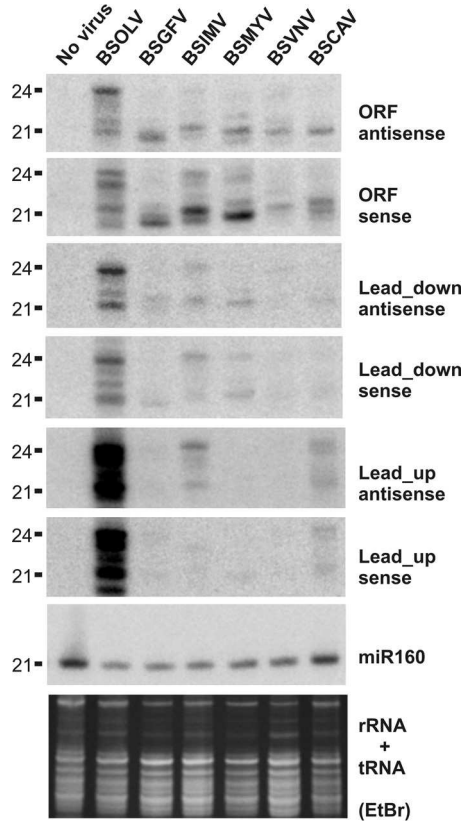


FIG 4 RNA blot hybridization analysis of viral and endogenous banana sRNAs in *M. acuminata* plants infected with six distinct BSV species. Total RNA samples from the healthy control *M. acuminata* plant and *M. acuminata* plants individually infected with BSOLV, BSGFV, BSIMV, BSMYV, BSVNV, and BSCAV were analyzed by RNA blot hybridization using 15% polyacrylamide gel electrophoresis (PAGE). The RNA blot membranes were successively hybridized with mixtures of 12 DNA oligonucleotide probes complementary to the respective viruses (2 probes per target region for each BSV species; for sequences and genome positions, see Table S1 in the supplemental material) and then to the evolutionary conserved miRNA (miR160). Positions of the BSV-specific probes Lead_up and Lead_down with respect of the reverse transcription start site are indicated in Fig. 3 for BSOLV. Ethidium bromide (EtBr) staining of ribosomal and transfer RNAs (rRNA + tRNA) is shown as a loading control. The 21-nt and 24-nt size markers are indicated. Note that sRNA mobility in 15% PAGE depends on the molecular weight: purine-rich sRNAs migrate more slowly by a ca. 0.5-to-1-nucleotide distance than pyrimidine-rich sRNAs of the same size.

S2). The single-base resolution maps of 20-, 21-, 22-, 23-, 24-, and 25-nt viral sRNAs (Fig. 3; see also Dataset S3) were generated by a map tool, MISIS (http://www.fasteris.com/apps/) (56). In these maps, for each position on the sequence (starting from the 5′ end of the reference sequence), the numbers of matches starting at this position in the forward (first base of the read) and reverse (last base of the read) orientations for each read length are given. Note that the reads mapped to the last 50 bases of the extended viral sequence were added to the reads mapped to the first 50 bases. By default, MISIS generates two maps, one with zero mismatches to the reference genome (see Dataset S3) and another with up to two mismatches. The comparison of the two maps was informative for initial identification of SNPs in the reference sequence, which could then be confirmed using the IGV tool described above.

The sequence analyses of the reconstructed BSV genomes (see Sequence Analysis S1 in the supplemental material) were performed using the following tools: NCBI nucleotide BLAST (blastn), the NCBI ORF finder with a protein BLAST option (http://www.ncbi.nlm.nih.gov/gorf/gorf.html), the EMBOSS Needle for pairwise sequence alignment (http://www.ebi.ac.uk/Tools/psa/emboss_needle/nucleotide.html), ClustalW 2.0.12 for multiple alignment (http://mobyle.pasteur.fr/cgi-bin/portal.py?#forms::clustalw-multialign), and MFold for prediction of RNA secondary structure (http://mfold.rna.albany.edu/?q = mfold/RNA-Folding-Form).

sRNA blot hybridization analysis. sRNA blot hybridization analysis was performed as described by Blevins et al. (18) using the short DNA oligonucleotide probes listed in Table S1 in the supplemental material. For each of the several successive hybridizations shown in Fig. 4, a mixture of 12 p32-labeled DNA oligonucleotides (six pairs specific for the corresponding region of each virus) was used as a probe (see Table S1).

Nucleotide sequence accession numbers. The NCBI accession numbers for the Montpellier (MP) isolates are as follows: for BSOLV-MP, KJ013506; for BSGFV-MP, KJ013507; for BSIMV-MP, KJ013508; for BSMYV-MP, KJ013509; for BSVNV-MP, KJ013510; and for BSCAV-MP, KJ013511.

RESULTS AND DISCUSSION

***De novo* reconstruction of consensus master genomes of six BSV species causing individual persistent infections in banana plants.** To begin the investigation of persistent viral infections in *M. acuminata*, we first reconstructed consensus master genomes of BSOLV, BSGFV, BSIMV, BSMYV, BSVNV, and BSCAV from the leaf samples of individual banana plants, which had been inoculated with single BSV species by insect transmission and then propagated vegetatively for 10 years. Each sample corresponded to the third leaf before the last emerged leaf that showed characteristic banana streak disease symptoms. Total DNA and RNA were extracted from the leaf tissues of six distinct virus-infected plants and of virus-free plants of *M. acuminata* triploid (AAA) cultivars and seedy *M. balbisiana* (BB) as controls (Fig. 1A). To amplify each viral circular genomic DNA, total DNA was subjected to rolling circle amplification (RCA) supplemented with degenerate primers specific for all the six investigated BSV species (designed based on their full-length reference genomes from NCBI GenBank; see Table S1 in the supplemental material). The resulting RCA products were analyzed by the use of restriction endonucleases. As expected, the amplified viral genomic DNA from infected plants was monomerized by single cutters and, after gel separation, appeared as a single band of linear dsDNA of ge-

FIG 3 Single-nucleotide-resolution maps of 21-nt and 24-nt viral sRNAs from *M. acuminata* plants infected with six distinct episomal BSV species. The graphs plot the number of 21-nt and 24-nt viral sRNA reads at each nucleotide position of the reconstructed genome sequences of BSOLV (A), BSGFV (B), BSIMV (C), BSMYV (D), BSVNV (E), and BSCAV (F). Bars above the axis represent sense reads starting at each of the respective positions; those below the axis represent antisense reads ending at the respective positions (for the map details, see Dataset S3 in the supplemental material). The genome organization of each BSV species is shown schematically above the graph, starting with the Met-tRNA primer binding site (PBS) indicated in green and ending with the transcription start site indicated with a pink bent arrow (Pro), followed by the poly(A) signal (pA); the three ORFs (I, II, and III) are shown with boxes. The predicted pgRNA transcript is shown as a blue line (interrupted at the PBS). The positions of the probes used for sRNA blot hybridization (Lead-up and Lead-down; see Fig. 4) are indicated with red lines below the genome of BSOLV. Note that the putative BSOLV decoy region is located between the starts of transcription and reverse transcription, as indicated with the position of the "Lead-up probe."

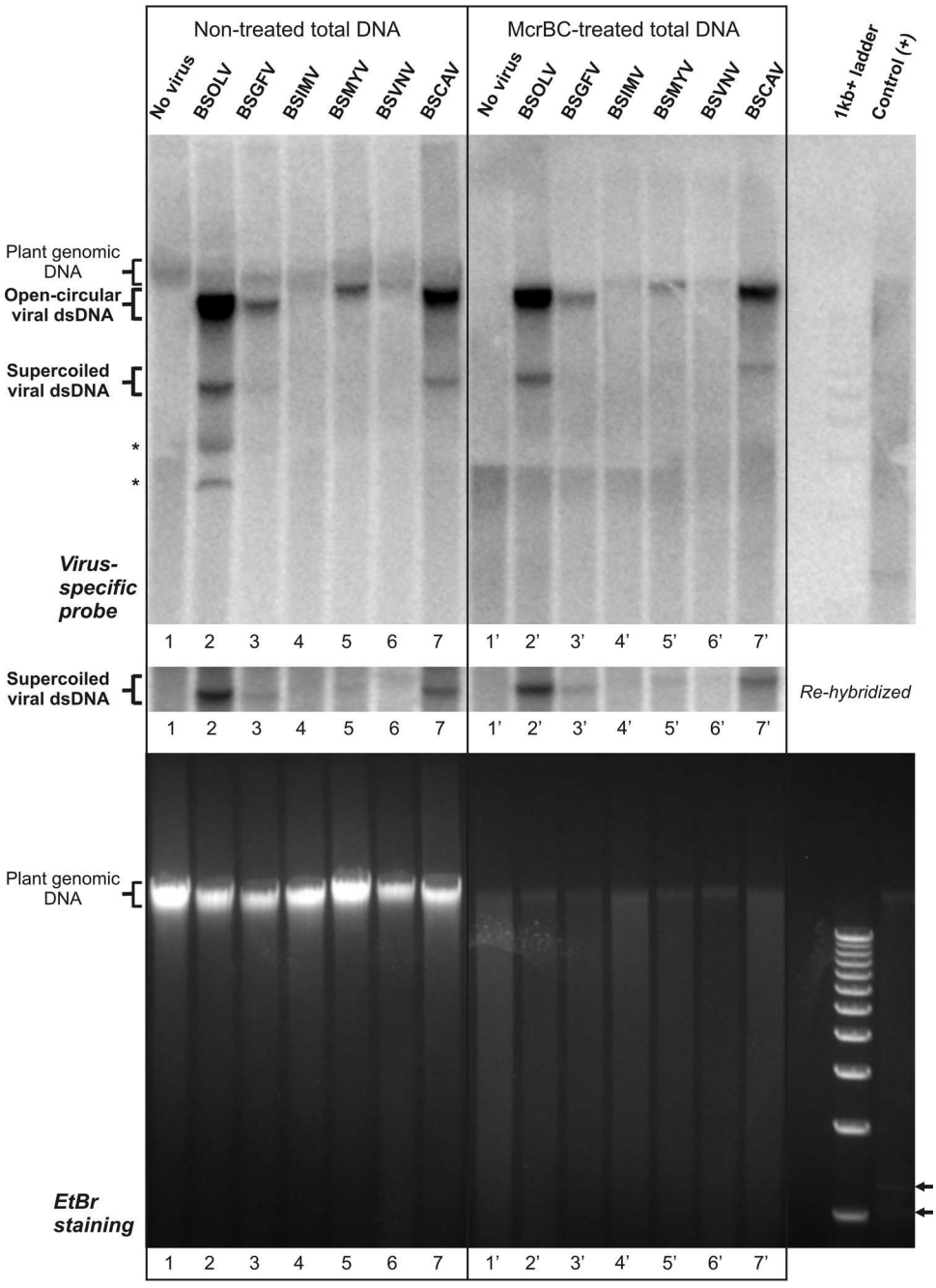


FIG 5 Analysis of relative accumulations of viral DNA and methylation statuses of the supercoiled and open circular forms of viral dsDNA using Southern blot hybridization. Total DNA samples from the healthy control *M. acuminata* plant and *M. acuminata* plants individually infected with BSOLV, BSGFV, BSIMV, BSMYV, BSVNV, and BSCAV were treated with the methylation-dependent enzyme McrBC and then analyzed by Southern blotting hybridization using 1% agarose gel electrophoresis as described in Materials and Methods. As a "nontreated" control, aliquots of the total DNA samples were incubated in parallel in the same buffer but without enzyme (left side of the blot). As a positive control [Control (+)], a methylated plasmid DNA was spiked in total DNA from the *M. balbisiana* plant (the digestion products are indicated by arrowheads in lower panel). As a DNA size marker, a 1-Kb+ ladder was used. Following separation, the gel was stained with EtBr (lower panel) and then the DNA was transferred to the membrane for hybridization with a mixture of the BSV-specific probes (see Materials and Methods). The hybridized membrane image is shown in the upper panel, with the positions of plant genomic DNA, viral dsDNA forms (open circular and supercoiled), and two McrBC-sensitive bands (asterisks) indicated. The membrane was rehybridized at 35°C to enhance the hybridization signal of the supercoiled form of viral DNA (cropped image in the middle panel).

nome size; no band of corresponding size was detected in the control plants (Fig. 1B). To reconstruct consensus master genomes and identify their potential variants in the viral quasispecies, the undigested RCA products were then deep sequenced using Illumina technology and the complete circular viral genomes were reconstructed *de novo* from the sequencing reads using bioinformatic tools (see Materials and Methods). The reconstructed genomes were validated by SNP (single nucleotide poly-

morphism) and indel (insertion/deletion) calling with redundant reads. In each case, several SNPs were identified (see Dataset S1 in the supplemental material), highlighting the quasispecies nature of persistent BSV species which, like all viruses, exist in "clouds" of microvariants deviating from a consensus master genome (57). Interestingly, most of the SNPs in all the BSV species were found to be "silent": SNPs in the intergenic region did not affect conserved *cis*-acting elements (see below), while most of SNPs in the coding sequences did not change the encoded amino acids (most of them occurred in the wobble position of codons) and only a few SNPs resulted in amino acid change. This suggests that the majority of SNPs belong to presumably viable variants of the viral genomes. To account for potential errors of the RCA method, the master genomes and SNPs were further verified by deep sequencing and bioinformatic analysis of viral siRNAs (see Dataset S1 and below) using a siRomics approach, proven to be applicable for *de novo* reconstruction of consensus master genomes of RNA and DNA viruses (58), and were deposited to NCBI GenBank as isolates of the previously identified BSVs (see Materials and Methods).

Sequence analysis revealed that the reconstructed viral genomes of our BSV isolates share 93.8% to 99.7% nucleotide identities with the GenBank reference genomes (see Sequence Analysis S1A in the supplemental material). For BSOLV, BSGFV, BSIMV, BSMYV, and BSVNV, which share 99.7%, 99.6%, 99.2%, 99.7%, and 98.9% identity, respectively, the differences are rather minor (except for BSVNV; see below) and include several single nucleotide substitutions, some of which result in amino acid substitutions of the viral ORF products, and a few short indels in the intergenic region (see Sequence Analysis S1A). Some or all these differences may represent the mutations that each original virus genome had accumulated following transmission to a new host and adaptation to a changing environment during the long persistence in vegetatively propagated host plants. Consistent with this hypothesis, *de novo* reconstruction of a BSGFV master genome from a second persistently infected and independently propagated banana plant revealed 10 single nucleotide substitutions (see Dataset S1 in the supplemental material). Interestingly, all these substitutions were present as variants deviating from the consensus master genome in one or both BSGFV-infected plants (see Dataset S1), indicating the ongoing evolution of the viral quasispecies. Among other alterations, our isolate of BSVNV has two 1-nt deletions in ORF I, which shift the frame and thereby elongate the ORF I product by 31 amino acids (see Sequence Analysis S1A). However, the elongated ORF I overlaps with ORF II, similarly to other BSVs (see Sequence Analysis S1B), and encodes a protein similar to those encoded by its homologs from other BSVs. This suggests that the previously described BSVNV genome (47) may contain cloning or sequencing errors that resulted in the truncation of ORF I. This hypothesis remains to be validated through generation of an infectious clone of BSVNV. In the case of BSCAV, the two isolates share 93.8% nucleotide identity and have much more substantial differences in the intergenic regions and the amino acid contents of three ORFs (see Sequence Analysis S1A), which classifies our isolate as a new strain of the virus.

Further sequence analysis revealed that the reconstructed genomes preserve the *cis* elements predicted to regulate transcription and translation of BSV pgRNA, which include the Pol II promoter and terminator elements and the elements driving ribosome shunting and leaky scanning. For the promoter and terminator elements, the only exception is BSGFV, for which both of our isolates and the GenBank reference sequence lack a CCAAT-box upstream of the TATA box and a recognizable poly(A) signal between the transcription start site and the Met-tRNA primer binding site (see Sequence Analysis S1 in the supplemental material). In all the BSV species, the ribosome shunt configuration is preserved in the pgRNA leader region, despite the drastic differences in the lengths and nucleotide compositions of the leader sequence (see Sequence Analysis S1B). Similarly to other plant pararetroviruses (36), this configuration comprises (i) a short ORF (sORF 1) terminating at a short distance (6 or 7 nt) in front of a large stem-loop structure, (ii) a stable bottom helix of the structure, and (iii) an unstructured landing site downstream of the structure followed by the ORF I start codon (AUG or CUG) (see Sequence Analysis S1B). The shapes and stabilities of the stem-loop structure, which is bypassed by shunting ribosomes, differ between the BSV species and also differ from those in CaMV and RTBV (shown schematically in Sequence Analysis S1B), for which the ribosome shunting mechanism has been dissected (reference 35 and references therein). Remarkably, and consistent with the shunt model, the only primary sequences which were found to be nearly identical in all the six BSV leaders include (i) the Met-tRNA primer binding site, (ii) sORF 1 (of 18 nt), and (iii) a 22-nt conserved motif just downstream of sORF 1 (see Sequence Analysis S1A). sORF 1 and the downstream motif likely constitute a takeoff site for shunting ribosomes. The primary leader sequence downstream of the takeoff site elements is not conserved and contains variable numbers of sORFs (see Sequence Analysis S1A), possibly because the function of this sequence is to form a secondary structure that brings the takeoff site in close proximity to the unstructured landing site just upstream of the ORF I start codon. Furthermore, a top section of the stem-loop structure in all the BSV leaders contains a purine-rich sequence which may serve as a packaging signal, in accordance with a model for CP-mediated packaging of CaMV pgRNA (39).

Consistent with the leaky-scanning model for initiation of translation of ORF II and ORF III, the ORF I and ORF II sequences are devoid of AUG triplets, except for suboptimal start codons of the ORFs I and II themselves. ORF I contains a few sORFs only in BSCAV and BSOLV (see Sequence Analysis S1B in the supplemental material). These sORFs, however, are not expected to block downstream translation. Thus, an artificial sORF introduced in ORF II of RTBV decreased leaky-scanning-mediated translation initiation of ORF III only slightly, showing that ribosomes can reinitiate translation following a short translational event (38).

In summary, the reconstructed genomes of six episomal BSV species encode a complete set of (potentially functional) viral proteins and, with the exception of BSGFV, possess all the conserved *cis* elements known to be required for viral replication and gene expression. This indicates that they represent infectious master genomes. The lack of recognizable CCAAT-box and poly(A) signal in the two BSGFV isolates and the GenBank reference sequence implies that alternative mechanisms of pgRNA transcription regulation evolved in this virus.

Persistent BSV infection induces multiple siRNA-generating silencing pathways targeting the entire virus genome. To examine the activity of sRNA-generating silencing machinery in persistently infected banana plants, a fraction of 19-to-30-nt sRNAs from the BSV-infected and noninfected banana plants was deep

sequenced using Illumina technology and analyzed by mapping of sRNA reads to the banana genome and the reconstructed BSV genomes (see Materials and Methods). The sRNA mapping and counting results are detailed in Datasets S2 and S3 in the supplemental material and summarized in Fig. 2 and 3. We restricted our analysis to sRNAs ranging in size from 20 to 25 nt, the size range known to correspond to functional plant miRNAs and siRNAs. The size profile of sRNAs matching the banana genome with zero mismatches was found to be interesting, as the relative accumulation of 20-nt sRNAs was comparable to that of 21-nt sRNAs, whereas 22-nt and 24-nt sRNAs represented the third- and fourth-most-abundant classes, respectively (Fig. 2; see also Dataset S2). In contrast, the endogenous sRNA profiles in other monocot and dicot plants, e.g., rice (59), maize (60), poplar (61), and *Arabidopsis* (16, 17, 62), differ in that the 24-nt class (mostly populated with heterochromatic siRNAs) is predominant, followed in most cases by the 21-nt class (populated with miRNAs and secondary-phase siRNAs). The analysis of 5′ nucleotide identities of the banana sRNAs revealed strong biases in each of the four major classes, with 5′ G dominating in 20-nt sRNAs (82%), 5′ U in 21-nt sRNAs (71%) and 22-nt sRNAs (74%), and 5′ A in 24-nt sRNAs (54%). These biases suggest that sRNAs of different size classes are sorted by different AGO family proteins based in most cases on the 5′ nucleotide as established in *A. thaliana* (63–65). The predominance of 20-nt 5′-G sRNAs in banana suggests that they represent a novel class of plant sRNAs, the biogenesis and function of which remain to be investigated. Our preliminary results indicated that most of the 20-nt 5′-G sRNAs are derived from so-called "unanchored" sequences of the recently released *M. acuminata* DH Pahang genome (22) which together constitute one-third of the banana genome and cannot be anchored to any of the 11 chromosomes.

Analysis of sRNAs matching the consensus master genomes of BSV species with zero mismatches revealed that the viral sRNAs constitute a significant fraction of the total 20-to-25-nt sRNAs in BSV-infected banana plants (~2% for BSCAV, ~4% for BSVNV, ~6% for BSGFV, ~7% for BSMYV, 8% for BSIMV, and 22% for BSOLV) but not in healthy banana plants (Fig. 2; see also Dataset S2 in the supplemental material). Given the much smaller size of the virus genome (7.3 to 7.8 Kb) compared to the banana genome (~473 Mb), banana plants produce large amounts of virus genome-derived siRNAs, especially in the case of BSOLV. The differences in the relative accumulations of viral siRNAs may reflect different proportions of the infected cells in the analyzed leaf tissues or different stages of virus replication (with different copy numbers of the viral DNA per cell) in the oscillating persistent infections. Thus, deep sequencing of sRNAs from the second plant infected with BSGFV revealed lower levels of viral siRNAs (~2% of total 20 to 25 sRNAs; see Dataset S2). The size profile of viral sRNAs differs substantially from that of the endogenous banana sRNAs: 21-nt viral sRNAs are predominant, followed by the 22-nt class (BSGFV, BSIMV, BSVNV, and BSCAV) or the 24-nt class (BSOLV and BSMYV), whereas the 20-nt class is only the third or the fourth most abundant (Fig. 2). The relative accumulation of 25-nt viral sRNAs is negligible. Interestingly, viral 20-nt sRNAs do not exhibit any strong bias to 5′ G as observed for endogenous 20-nt sRNAs (see Dataset S2), suggesting that the banana 20-nt 5′-G sRNA pathway is not involved in targeting BSV. For most BSV species, viral siRNAs of the 21-nt and 22-nt classes exhibit bias to 5′ U, albeit the bias is less pronounced than that seen with endogenous sRNAs of these classes, while viral 24-nt sRNAs are dominated by abundant 5′ A and 5′ U species (see Dataset S2). Taking the data together, viral sRNAs can potentially be associated with multiple AGOs because of the high diversity of sRNA species in each size class (see below). The predominance of 21-nt viral sRNAs together with the underrepresentation of 24-nt viral sRNAs in BSGFV, BSIMV, BSVNV, and BSCAV infections suggested that these viruses induce mostly PTGS pathways in banana plants. These presumptive PTGS pathways of *Musa acuminata* would be analogous to the *Arabidopsis thaliana* DCL4- and DCL2-dependent pathways that generate 21-nt and 22-nt siRNAs, respectively, and play a primary role in defense against RNA viruses (14, 18). In BSOLV and BSMYV infections, however, the presence of relatively high levels of 24-nt viral sRNAs implies the involvement of a nuclear TGS pathway in defense against episomal BSVs in banana plants. In *Arabidopsis*, the DCL3-dependent nuclear pathway generates 24-nt siRNAs and thereby contributes together with the PTGS pathways to defense against DNA viruses, as established for the geminivirus CaLCuV and the pararetrovirus CaMV (15–18).

Analysis of single-nucleotide-resolution maps of virus-derived sRNAs revealed that nonredundant 20-to-25-nt sRNA species cover the entire circular virus genome without gaps (see the "total" column in Dataset S3 in the supplemental material). A few strand-specific gaps of more than 24 nt in the 5′ and 3′ nucleotide coverage were observed only in BSCAV, which leaves less than 100 nt not covered with the viral reads on each strand, but all these small gaps were covered on the opposite stand (see the "total_forward" and "total_reverse" columns in Dataset S3). Strand-specific gaps in sRNA coverage were not observed for other BSV species. Since BSCAV has the lowest number of total reads (47,473), whereas other viruses have much higher numbers (ranging from 134,023 reads in BSIMV to 658,896 reads in BSOLV; see Dataset S2), it could be expected that deeper sequencing of BSCAV sRNAs would close the gaps. In conclusion, the entire circular genome of BSV is covered with sense and antisense precursors of viral sRNAs. Similar findings have been reported for CaLCuV (17), grapevine geminivirus (58), and CaMV (58). For CaLCuV and other geminiviruses that transcribe circular viral DNA bidirectionally, these findings have implicated Pol II in production of sense and antisense transcripts forming dsRNA precursors of viral siRNAs (13). Indeed, the genetic evidence using CaLCuV-infected *Arabidopsis* mutants for silencing-related RNA polymerases ruled out involvement of Pol IV, Pol V, RDR1, RDR2, or RDR6 in the biogenesis of viral siRNAs (17, 18). In CaMV and other pararetroviruses with monodirectional Pol II transcription, the mechanism of dsRNA production remains to be investigated. For CaMV, genetic evidence indicates that Pol IV, Pol V, RDR1, RDR2, and RDR6 are not required for viral siRNA biogenesis (16, 18), suggesting that Pol II might be involved (13).

Analysis of redundant sRNA reads revealed that viral sRNA hotspots are unevenly distributed along the genome (see Fig. 3 for 21-nt and 24-nt sRNAs and Dataset S3 in the supplemental material for all the size classes). In most BSVs (except BSOLV), the hotspots of 21-nt and 22-nt viral sRNAs of both sense and antisense polarity tend to concentrate within ORF I, ORF II, and the 5′ portion of ORF III, while the 3′ half of ORF III and the promoter portion of the intergenic region upstream of the transcription start site are relatively poorly covered with 21-nt or 22-nt viral sRNAs. In contrast, 24-nt viral sRNA hotspots are more evenly

distributed along the viral genomes (Fig. 3). This suggests that, in the hot spot regions, 21-nt and 22-nt viral sRNAs are produced from the more abundant dsRNA precursors which may preferentially be amplified by an RDR activity. In BSOLV, 24-nt siRNAs are of exceptionally high abundance and their hot spots are slightly depleted from the ORF I-to-ORF II region that generates the most abundant 21-nt and 22-nt viral sRNAs (Fig. 3). This implies that 24-nt and 21-to-22-nt viral sRNAs may be processed from different dsRNA precursors.

In CaMV, the pgRNA leader region between the start site for Pol II transcription and the primer binding site for reverse transcription generates the most abundant viral siRNAs of the three major size classes and both polarities (16). This region was proposed to produce a highly abundant dsRNA decoy that diverts components of the plant silencing machinery from the upstream promoter region and the downstream coding regions (13, 16). In all the BSV species, the primer binding site is located much closer to the transcription start site than in CaMV, and therefore much of the pgRNA leader sequence which contains the ribosome shunt elements and forms the large stem-loop structure is located downstream of it (see Sequence Analysis S1 in the supplemental material). The structured leader region is populated with the hotspots of 21/22-nt and 24-nt viral sRNAs only in BSIMV, while the corresponding region in other BSV species is a relatively poor source of sRNAs (Fig. 3). The latter finding indicates that a stem-loop secondary structure *per se* may not be a major determinant for massive production of viral sRNAs. Moreover, the high abundance of viral sRNA of both polarities implies that viral sRNA duplexes are produced from perfect dsRNA precursors rather than from the structured regions of pgRNA. In BSOLV, a "decoy" region between the transcription and reverse transcription starts spawns highly abundant 21-nt and 24-nt viral sRNAs (Fig. 3A; Lead_up). This and most other findings from the deep sequencing analysis were validated by sRNA blot hybridization analysis (Fig. 4; see Lead_up probes). Thus, BSOLV may have evolved a CaMV-like decoy strategy of silencing evasion (16), which may explain why the accumulation of BSOLV DNA is higher than that of other species (Fig. 5). However, such an evasion mechanism appears to be less efficient in persistently infected banana plants in diverting the siRNA-generating machinery from other regions of the viral genome (Fig. 3A and 4) than the CaMV decoy mechanism in *Arabidopsis* plants that exhibit severe CaMV disease with no recovery (16).

We noted that BSV infection alters the relative accumulations of endogenous banana sRNAs, but no clear tendency for either increased or decreased accumulation of any particular sRNA size class was observed (Fig. 2). Interestingly, BSOLV, whose sRNA size class profile resembles that of CaMV (the two species produce comparable amounts of 21-nt and 24-nt viral siRNAs), has a CaMV-like impact on the endogenous sRNA profile, boosting endogenous 24-nt sRNA production at the expense of 21-nt sRNAs (16). The biological significance of this alteration remains to be investigated.

In summary, the RNA silencing machinery of banana plants infected with BSV generates abundant 21-to-22-nt and 24-nt viral sRNAs, possibly in the cytoplasm and the nucleus, respectively. Based on the characteristic sizes, distributions along the entire genome in both orientations, and 5′ nucleotide biases, these viral sRNAs can be classified as typical siRNAs that have the potential to bind distinct AGO proteins and target viral nucleic acids for post-transcriptional and transcriptional silencing, respectively. The prevalence of 21-nt viral siRNAs with their hotspots in a 5′ portion of pgRNA in most BSV infections is reminiscent of the siRNA profile of an *Evade* retrotransposon activated in certain epigenetic mutants of *Arabidopsis* (8). Highly abundant 21-nt siRNAs derived from this retrotransposon are generated by the RDR6- and DCL4-dependent mechanism and loaded onto AGO1 and AGO2. However, these siRNAs are not effective in silencing the retrotransposon transcript, because PTGS is evaded via Gag protein-mediated packaging of the transcript into virus-like particles (8). By analogy, pararetroviral CP/Gag that encapsidates pgRNA for reverse transcription might also protect the pgRNA from viral siRNA-directed cleavage and/or translational repression. Whether or not BSV and other badnaviruses have evolved specialized suppressors of silencing remains to be investigated.

Viral circular dsDNA evades cytosine methylation in persistently infected banana plants. To address whether or not the viral siRNAs accumulating in BSV-infected banana plants direct methylation of viral dsDNA, we exploited the cleavage activity of the McrBC methylation-dependent enzyme. This enzyme recognizes 5-methylcytosines (a hallmark of plant DNA methylation) in an R^mC context (where R = A or G) and cleaves between two recognition sites separated by ~30 to ~3,000 bp at distance of 25 to 30 bp from one of the two sites. A circular plasmid DNA containing a single R^mC site on one strand could also be linearized by McrBC (66). Thus, any circular viral dsDNA that contains at least one 5′ methylcytosine in an R^mC context should be digested by McrBC. We found that each of the six BSV genomes contains numerous McrBC recognition sites. These sites can potentially be targets for methylation, presumably by a banana RdDM pathway.

Total DNA extracted from the banana leaf samples was treated overnight with McrBC and then separated on a 1% agarose gel along with total DNA aliquots of the same samples treated in parallel under the same conditions but without McrBC. As a control, a methylated plasmid was mixed with total banana DNA and the mixture was treated with McrBC. Ethidium bromide (EtBr) staining revealed that the methylated plasmid was fully digested, yielding two fragments as expected. The banana genomic DNA in the control reaction and all the other reactions was almost fully digested by McrBC (Fig. 5; compare the McrBC-treated and non-treated sample results). This indicates that McrBC enzymatic activity is not inhibited by total banana genomic DNA and that the bulk of banana DNA is extensively methylated. The smear of McrBC digestion products (Fig. 5) suggests that the cytosine methylation sites are scattered along the banana genome.

The methylation status of viral DNA was then evaluated by Southern blotting hybridization using a mixture of six pairs of DNA oligonucleotide probes with equal annealing temperatures, each pair being designed to hybridize specifically to one of the six BSV species. This experimental design allows simultaneous detection of major forms of viral DNA of all the six BSV species and measurement of their relative levels of accumulation. The results revealed two major forms of circular viral dsDNA, the more abundant open circular dsDNA and the less abundant covalently closed (supercoiled) dsDNA, before and after McrBC treatment. Compared to the banana genomic DNA, both major forms of viral DNA appeared to be resistant to McrBC (Fig. 5, upper panel). This indicates that the major fraction of viral genomic DNA is not methylated. More specifically, the resistance to McrBC digestion of the supercoiled form of viral dsDNA, which is clearly detectable

in BSOLV, BSCAV, BSGFV, and BSMYV (Fig. 5, cropped image in the middle panel), suggests that viral minichromosomes accumulating in the nucleus for Pol II-mediated transcription of pgRNA are largely not methylated. We cannot exclude the possibility that a fraction of the episomal DNA is methylated. In fact, some reduction in open circular dsDNA levels after the enzymatic treatment was observed (Fig. 5). However, this form of viral DNA is packaged in the virions following reverse transcription in the cytoplasm and cannot be methylated by the nuclear RdDM machinery. A potential nonspecific activity of McrBC during the overnight incubation may account for minor loss of viral DNA. Since no substantial decrease in the ratio of the supercoiled form to the open circular form was observed (Fig. 5), this argues for little (if any) methylation of the supercoiled DNA. It is difficult to evaluate the methylation status of episomal DNA in BSIMV and BSVNV, for which the supercoiled form was barely detectable even under low-stringency hybridization and washing conditions (cropped portion in the middle panel of Fig. 5). We expect, however, that if these two low-titer viruses were to replicate more actively and accumulate supercoiled DNA at detectable levels, this DNA could also be largely unmethylated, like the supercoiled DNA for the four higher-titer viruses.

The Southern blot hybridization analysis combined with the viral sRNA profiles shows that the relative accumulation of viral DNA does not correlate with the relative abundance of viral siRNAs. Indeed, BSOLV, spawning siRNAs that were 10 times more abundant than those spawned by BSCAV (Fig. 2), accumulated viral DNA at an abundance only three times greater than that seen with BSCAV. The other four BSV species produced larger amounts of viral siRNAs than BSCAV and accumulated smaller amounts of viral DNA than BSCAV (Fig. 5), but these amounts of DNA differed dramatically, unlike the amounts of viral siRNAs, which were comparable. Finally, BSVNV, with barely detectable levels of viral DNA, accumulated viral siRNAs that were only five times less abundant than those seen with BSOLV, which accumulated the largest amounts of DNA. This implies that viral siRNAs do not necessarily restrict accumulation of viral DNA, although the total abundance of viral siRNAs may not reflect the proportion of viral siRNAs incorporated into functional RISCs. Moreover, the fact that similar ratios of the supercoiled form to the open circular form were observed in BSOLV and BSCAV, which produced the largest and the smallest amounts of siRNAs, respectively, indicates that viral siRNAs do not influence the distribution of viral DNA between the cytoplasm (populated with the open circular form) and the nucleus (populated mostly with the supercoiled form). Therefore, the efficiency of viral siRNA production (or the relative abundance of 24-nt viral siRNAs) does not appear to affect the viral replication cycle, in which the nucleus supplies viral pgRNA to the cytoplasm for reverse transcription whereas the cytoplasm supplies the open circular viral DNA to the nucleus for repair (i.e., production of covalently closed, supercoiled DNA) and subsequent Pol II-mediated transcription of pgRNA on the covalently closed DNA (reviewed in reference 13).

In the case of BSOLV (but not other BSVs), two additional bands of viral DNA were detected which migrated faster than the supercoiled form, and both bands disappeared after the McrBC treatment (Fig. 5; asterisks). Note that these bands cannot be an artifact of DNA overload, since equal amounts of total DNA were taken for treatment in the presence and absence of McrBC followed by Southern blotting hybridization. The nature of these bands and why these are the only apparently methylated forms of viral DNA remain to be investigated. In geminiviruses, the only form of viral DNA which contains a substantial proportion of methylated cytosines is heterogeneous linear dsDNA, a byproduct of recombination-dependent replication (12, 13). Pararetroviruses are not known to exploit this replication mechanism.

In conclusion, most of the circular covalently closed viral dsDNA in BSV-infected *M. acuminata* plants is nonmethylated. Thus, multiple BSV minichromosomes appear to evade siRNA-directed DNA methylation in the nucleus and thereby retain the potential for active Pol II transcription. One can argue that 24-nt viral siRNA levels are not sufficient to exert DNA methylation. However, the relatively highly abundant 24-nt viral siRNAs accumulating in BSOLV-infected plants, which cover the entire virus genome in both orientations, also fail to direct methylation of viral DNA. The mechanism of RdDM evasion by episomal BSV remains to be investigated. In *Arabidopsis*, the establishment and maintenance of cytosine methylation at the RdDM loci require Pol V and Pol IV, respectively. Pol V generates a scaffold transcript and recruits *de novo* methyltransferase through interaction with 24-nt siRNA-AGO4 complexes, while Pol IV and RDR2 together generate dsRNA precursors of 24-nt siRNAs (reviewed in reference 13). Genetic evidence indicates that Pol V, Pol IV, and RDR2 are not required for the biogenesis of 24-nt viral siRNAs in CaMV-infected *Arabidopsis* (16, 18), which effectively uncouples the viral siRNAs from the RdDM machinery (13). This may explain the lack of detectable cytosine methylation of episomal viral DNA in CaMV-infected turnip plants (11) or in kohlrabi plants that recovered from CaMV disease symptoms (2). Note, however, that those earlier studies of CaMV made use of methylation-sensitive enzymes, which would report cytosine methylation only at the respective enzyme recognition sites. Investigation of genetic requirements for viral siRNA biogenesis in banana plants is currently difficult, because no information on the banana RNA silencing genes and no respective gene mutant lines are available. Our analysis of the viral and endogenous sRNA profiles and the finding that the banana genome is extensively methylated suggest the existence of conserved nuclear and cytoplasmic components of RNA silencing machinery that can mediate both PTGS and RdDM/TGS in banana, similarly to other plant species. Our finding that the banana sRNA population contains a dominant 20-nt 5′-G RNA class which is absent in the viral siRNA population raises the intriguing possibility that those sRNAs might be indirectly involved in the establishment or maintenance of DNA cytosine methylation in banana plants.

The findings reported here may also be relevant for further understanding and better control of the banana streak disease during industrial banana cultivation, but it will be important to investigate the banana silencing machinery and BSV infections under open field conditions and in other banana genotypes.

ACKNOWLEDGMENTS

The work was supported by the Swiss National Science Foundation (31003A_143882/1 to M.M.P.), the European Commission Marie Curie fellowship (PIIF-237493-SUPRA to R.R.), the European Cooperation in Science and Technology (COST) action FA0806 (SER no. C09.0176 to L.F. and M.M.P.), and the CIRAD (Ph.D. grant to P.-O.D).

We thank Thomas Boller for supporting research of the M.M.P. group at the University of Basel and Ben Lockhart, who performed the mealybug-mediated BSV infection of all the banana plants used in this study.

R.R., M.M.P., M.-L.I.-C., M.C., and L.F. designed the research, R.R., J.S., M.C., P.-O.D., and N.L. performed the research, J.S., M.M.P., R.R., M.C., and M.-L.I.-C. analyzed the data, and M.M.P. wrote the paper.

REFERENCES

1. **Roossinck MJ.** 2013. Plant virus ecology. PLoS Pathog. **9:**e1003304. http://dx.doi.org/10.1371/journal.ppat.1003304.
2. **Covey SN, Al-Kaff NS, Lángara A, Turner DS.** 1997. Plants combat infection by gene silencing. Nature **385:**781–782. http://dx.doi.org/10.1038/385781a0.
3. **Ratcliff FG, MacFarlane SA, Baulcombe DC.** 1999. Gene silencing without DNA. RNA-mediated cross-protection between viruses. Plant Cell **11:**1207–1216.
4. **Ghildiyal M, Zamore PD.** 2009. Small silencing RNAs: an expanding universe. Nat. Rev. Genet. **10:**94–108. http://dx.doi.org/10.1038/nrg2504.
5. **Carthew RW, Sontheimer EJ.** 2009. Origins and mechanisms of miRNAs and siRNAs. Cell **136:**642–655. http://dx.doi.org/10.1016/j.cell.2009.01.035.
6. **Law JA, Jacobsen SE.** 2010. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat. Rev. Genet. **11:**204–220. http://dx.doi.org/10.1038/nrg2719.
7. **Ding SW, Voinnet O.** 2007. Antiviral immunity directed by small RNAs. Cell **130:**413–426. http://dx.doi.org/10.1016/j.cell.2007.07.039.
8. **Marí-Ordóñez A, Marchais A, Etcheverry M, Martin A, Colot V, Voinnet O.** 2013. Reconstructing de novo silencing of an active plant retrotransposon. Nat. Genet. **45:**1029–1039. http://dx.doi.org/10.1038/ng.2703.
9. **Panda K, Slotkin RK.** 2013. Proposed mechanism for the initiation of transposable element silencing by the RDR6-directed DNA methylation pathway. Plant Signal. Behav. **8:**e25206. http://dx.doi.org/10.4161/psb.25206.
10. **Raja P, Wolf JN, Bisaro DM.** 2010. RNA silencing directed against geminiviruses: post-transcriptional and epigenetic components. Biochim. Biophys. Acta **1799:**337–351. http://dx.doi.org/10.1016/j.bbagrm.2010.01.004.
11. **Al-Kaff NS, Covey SN, Kreike MM, Page AM, Pinder R, Dale PJ.** 1998. Transcriptional and posttranscriptional plant gene silencing in response to a pathogen. Science **279:**2113–2115. http://dx.doi.org/10.1126/science.279.5359.2113.
12. **Paprotka T, Deuschle K, Metzler V, Jeske H.** 2011. Conformation-selective methylation of geminivirus DNA. J. Virol. **85:**12001–12012. http://dx.doi.org/10.1128/JVI.05567-11.
13. **Pooggin MM.** 2013. How can plant DNA viruses evade siRNA-directed DNA methylation and silencing? Int. J. Mol. Sci. **14:**15233–15259. http://dx.doi.org/10.3390/ijms140815233.
14. **Deleris A, Gallego-Bartolome J, Bao J, Kasschau KD, Carrington JC, Voinnet O.** 2006. Hierarchical action and inhibition of plant Dicer-like proteins in antiviral defense. Science **313:**68–71. http://dx.doi.org/10.1126/science.1128214.
15. **Akbergenov R, Si-Ammour A, Blevins T, Amin I, Kutter C, Vanderschuren H, Zhang P, Gruissem W, Meins F, Jr, Hohn T, Pooggin MM.** 2006. Molecular characterization of geminivirus-derived small RNAs in different plant species. Nucleic Acids Res. **34:**462–471. http://dx.doi.org/10.1093/nar/gkj447.
16. **Blevins T, Rajeswaran R, Aregger M, Borah BK, Schepetilnikov M, Baerlocher L, Farinelli L, Meins F, Jr, Hohn T, Pooggin MM.** 2011. Massive production of small RNAs from a non-coding region of Cauliflower mosaic virus in plant defense and viral counter-defense. Nucleic Acids Res. **39:**5003–5014. http://dx.doi.org/10.1093/nar/gkr119.
17. **Aregger M, Borah BK, Seguin J, Rajeswaran R, Gubaeva EG, Zvereva AS, Windels D, Vazquez F, Blevins T, Farinelli L, Pooggin MM.** 2012. Primary and secondary siRNAs in geminivirus-induced gene silencing. PLoS Pathog. **8:**e1002941. http://dx.doi.org/10.1371/journal.ppat.1002941.
18. **Blevins T, Rajeswaran R, Shivaprasad PV, Beknazariants D, Si-Ammour A, Park HS, Vazquez F, Robertson D, Meins F, Jr, Hohn T, Pooggin MM.** 2006. Four plant Dicers mediate viral small RNA biogenesis and DNA virus induced silencing. Nucleic Acids Res. **34:**6233–6246. http://dx.doi.org/10.1093/nar/gkl886.
19. **Donaire L, Wang Y, Gonzalez-Ibeas D, Mayer KF, Aranda MA, Llave C.** 2009. Deep-sequencing of plant viral small RNAs reveals effective and widespread targeting of viral genomes. Virology **392:**203–214. http://dx.doi.org/10.1016/j.virol.2009.07.005.
20. **Pantaleo V, Saldarelli P, Miozzi L, Giampetruzzi A, Gisel A, Moxon S, Dalmay T, Bisztray G, Burgyan J.** 2010. Deep sequencing analysis of viral short RNAs from an infected Pinot Noir grapevine. Virology **408:**49–56. http://dx.doi.org/10.1016/j.virol.2010.09.001.
21. **Yang X, Wang Y, Guo W, Xie Y, Xie Q, Fan L, Zhou X.** 2011. Characterization of small interfering RNAs derived from the geminivirus/betasatellite complex using deep sequencing. PLoS One **6:**e16928. http://dx.doi.org/10.1371/journal.pone.0016928.
22. **D'Hont A, Denoeud F, Aury JM, Baurens FC, Carreel F, Garsmeur O, Noel B, Bocs S, Droc G, Rouard M, Da Silva C, Jabbari K, Cardi C, Poulain J, Souquet M, Labadie K, Jourda C, Lengellé J, Rodier-Goud M, Alberti A, Bernard M, Correa M, Ayyampalayam S, Mckain MR, Leebens-Mack J, Burgess D, Freeling M, Mbéguié-A-Mbéguié D, Chabannes M, Wicker T, Panaud O, Barbosa J, Hribova E, Heslop-Harrison P, Habas R, Rivallan R, Francois P, Poiron C, Kilian A, Burthia D, Jenny C, Bakry F, Brown S, Guignon V, Kema G, Dita M, Waalwijk C, Joseph S, Dievart A, et al.** 2012. The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. Nature **488:**213–217. http://dx.doi.org/10.1038/nature11241.
23. **Shekhawat UK, Ganapathi TR, Hadapad AB.** 2012. Transgenic banana plants expressing small interfering RNAs targeted against viral replication initiation gene display high-level resistance to banana bunchy top virus infection. J. Gen. Virol. **93:**1804–1813. http://dx.doi.org/10.1099/vir.0.041871-0.
24. **Yot-Dauthy D, Bové JM.** 1966. Mosaïque du bananier. Identification et purification de diverses souches du virus. Fruit **21:**449–465.
25. **Lockhart BEL.** 1986. Purification and serology of a bacilliform virus associated with banana streak disease. Phytopathology **76:**995–999. http://dx.doi.org/10.1094/Phyto-76-995.
26. **Kubiriba J, Legg JP, Tushemereirwe W, Adipala E.** 2001. Vector transmission of Banana streak virus in the screenhouse in Uganda. Ann. Appl. Biol. **139:**37–43. http://dx.doi.org/10.1111/j.1744-7348.2001.tb00128.x.
27. **Geering AD, Pooggin MM, Olszewski NE, Lockhart BE, Thomas JE.** 2005. Characterisation of Banana streak Mysore virus and evidence that its DNA is integrated in the B genome of cultivated Musa. Arch. Virol. **150:**787–796. http://dx.doi.org/10.1007/s00705-004-0471-z.
28. **Geering ADW, Thomas JE.** 2002. Banana streak virus. Web descriptions of plant viruses, description no. 390. Assoc. Appl. Biol. http://www.dpvweb.net/dpv/showdpv.php?dpvno=390.
29. **Harper G, Hull R.** 1998. Cloning and sequence analysis of banana streak virus DNA. Virus Genes **17:**271–278. http://dx.doi.org/10.1023/A:1008021921849.
30. **Hull R.** 2007. Caulimoviridae (plant pararetroviruses). eLS, John Wiley & Sons Ltd., Chichester, United Kingdom. http://www.els.net. http://dx.doi.org/10.1002/9780470015902.a0000746.pub2.
31. **Hohn T, Rothnie H.** 2013. Plant pararetroviruses: replication and expression. Curr. Opin. Virol. **3:**621–628. http://dx.doi.org/10.1016/j.coviro.2013.08.013.
32. **Hohn T.** 2013. Plant pararetroviruses: interactions of cauliflower mosaic virus with plants and insects. Curr. Opin. Virol. **3:**629–638. http://dx.doi.org/10.1016/j.coviro.2013.08.014.
33. **Fütterer J, Kiss-László Z, Hohn T.** 1993. Nonlinear ribosome migration on cauliflower mosaic virus 35S RNA. Cell **73:**789–802. http://dx.doi.org/10.1016/0092-8674(93)90257-Q.
34. **Pooggin MM, Ryabova LA, He X, Fütterer J, Hohn T.** 2006. Mechanism of ribosome shunting in Rice tungro bacilliform pararetrovirus. RNA **12:**841–850. http://dx.doi.org/10.1261/rna.2285806.
35. **Pooggin MM, Fütterer J, Hohn T.** 2008. Cross-species functionality of pararetroviral elements driving ribosome shunting. PLoS One **3:**e1650. http://dx.doi.org/10.1371/journal.pone.0001650.
36. **Pooggin MM, Fütterer J, Skryabin KG, Hohn T.** 1999. A short open reading frame terminating in front of a stable hairpin is the conserved feature in pregenomic RNA leaders of plant pararetroviruses. J. Gen. Virol. **80:**2217–2228.
37. **Ryabova LA, Pooggin MM, Hohn T.** 2002. Viral strategies of translation initiation: ribosomal shunt and reinitiation. Prog. Nucleic Acid Res. Mol. Biol. **72:**1–39. http://dx.doi.org/10.1016/S0079-6603(02)72066-7.
38. **Fütterer J, Rothnie HM, Hohn T, Potrykus I.** 1997. Rice tungro bacilliform virus open reading frames II and III are translated from polycistronic pregenomic RNA by leaky scanning. J. Virol. **71:**7984–7989.
39. **Guerra-Peraza O, de Tapia M, Hohn T, Hemmings-Mieszczak M.** 2000. Interaction of the cauliflower mosaic virus coat protein with the pre-

genomic RNA leader. J. Virol. **74:**2067–2072. http://dx.doi.org/10.1128/JVI.74.5.2067-2072.2000.
40. **Geering AD, Olszewski NE, Harper G, Lockhart BE, Hull R, Thomas JE.** 2005. Banana contains a diverse array of endogenous badnaviruses. J. Gen. Virol. **86:**511–520. http://dx.doi.org/10.1099/vir.0.80261-0.
41. **Chabannes M, Iskra-Caruana ML.** 2013. Endogenous pararetroviruses–a reservoir of virus infection in plants. Curr. Opin. Virol. **3:**615–620. http://dx.doi.org/10.1016/j.coviro.2013.08.012.
42. **Geering AD, Olszewski NE, Dahal G, Thomas JE, Lockhart BE.** 2001. Analysis of the distribution and structure of integrated Banana streak virus DNA in a range of Musa cultivars. Mol. Plant Pathol. **2:**207–213. http://dx.doi.org/10.1046/j.1464-6722.2001.00071.x.
43. **Gayral P, Iskra-Caruana ML.** 2009. Phylogeny of Banana Streak Virus reveals recent and repetitive endogenization in the genome of its banana host (Musa sp.). J. Mol. Evol. **69:**65–80. http://dx.doi.org/10.1007/s00239-009-9253-2.
44. **Chabannes M, Baurens FC, Duroy PO, Bocs S, Vernerey MS, Rodier-Goud M, Barbe V, Gayral P, Iskra-Caruana ML.** 2013. Three infectious viral species lying in wait in the banana genome. J. Virol. **87:**8624–8637. http://dx.doi.org/10.1128/JVI.00899-13.
45. **Gayral P, Noa-Carrazana JC, Lescot M, Lheureux F, Lockhart BE, Matsumoto T, Piffanelli P, Iskra-Caruana ML.** 2008. A single Banana streak virus integration event in the banana genome as the origin of infectious endogenous pararetrovirus. J. Virol. **82:**6697–6710. http://dx.doi.org/10.1128/JVI.00212-08.
46. **Geering AD, Parry JN, Thomas JE.** 2011. Complete genome sequence of a novel badnavirus, banana streak IM virus. Arch. Virol. **156:**733–737. http://dx.doi.org/10.1007/s00705-011-0946-7.
47. **Lheureux F, Laboureau N, Muller E, Lockhart BE, Iskra-Caruana ML.** 2007. Molecular characterization of banana streak acuminata Vietnam virus isolated from Musa acuminata siamea (banana cultivar). Arch. Virol. **152:**1409–1416. http://dx.doi.org/10.1007/s00705-007-0946-9.
48. **James AP, Geijskes RJ, Dale JL, Harding RM.** 2011. Molecular characterisation of six badnavirus species associated with leaf streak disease of banana in East Africa. Ann. Appl. Biol. **158:**346–353. http://dx.doi.org/10.1111/j.1744-7348.2011.00466.x.
49. **Le Provost G, Iskra-Caruana ML, Acina I, Teycheney PY.** 2006. Improved detection of episomal Banana streak viruses by multiplex immunocapture PCR. J. Virol. Methods **137:**7–13. http://dx.doi.org/10.1016/j.jviromet.2006.05.021.
50. **Liu JJ, Goh CJ, Loh CS, Liu P, Pua EC.** 1998. A method for isolation of total RNA from fruit tissues of banana. Plant Mol. Biol. Rep. **16:**1–6. http://dx.doi.org/10.1023/A:1017158311412.
51. **Gawel NJ, Jarret RL.** 1991. Chloroplast DNA restriction fragment length polymorphisms (RFLPs) in Musa species. Theor. Appl. Genet. **81:**783–786.
52. **Zerbino DR, Birney E.** 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. **18:**821–829. http://dx.doi.org/10.1101/gr.074492.107.
53. **Schulz MH, Zerbino DR, Vingron M, Birney E.** 2012. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics **28:**1086–1092. http://dx.doi.org/10.1093/bioinformatics/bts094.
54. **Thorvaldsdóttir H, Robinson JT, Mesirov JP.** 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinform. **14:**178–192. http://dx.doi.org/10.1093/bib/bbs017.
55. **Li H, Durbin R.** 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics **26:**589–595. http://dx.doi.org/10.1093/bioinformatics/btp698.
56. **Seguin J, Otten P, Baerlocher L, Farinelli L, Pooggin MM.** 2014. 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. http://dx.doi.org/10.1016/j.jviromet.2013.10.013.
57. **Domingo E, Sheldon J, Perales C.** 2012. Viral quasispecies evolution. Microbiol. Mol. Biol. Rev. **76:**159–216. http://dx.doi.org/10.1128/MMBR.05023-11.
58. **Seguin J, Rajeswaran R, Malpica-López N, Martin RR, Kasschau K, Dolja VV, Otten P, Farinelli L, Pooggin MM.** 2014. De novo reconstruction of consensus master genomes of plant RNA and DNA viruses from siRNAs. PLoS One **9:**e88513. http://dx.doi.org/10.1371/journal.pone.0088513.
59. **Jeong DH, Park S, Zhai J, Gurazada SG, De Paoli E, Meyers BC, Green PJ.** 2011. Massive analysis of rice small RNAs: mechanistic implications of regulated microRNAs and variants for differential target RNA cleavage. Plant Cell **23:**4185–4207. http://dx.doi.org/10.1105/tpc.111.089045.
60. **Jiao Y, Song W, Zhang M, Lai J.** 2011. Identification of novel maize miRNAs by measuring the precision of precursor processing. BMC Plant Biol. **11:**141. http://dx.doi.org/10.1186/1471-2229-11-141.
61. **Klevebring D, Street NR, Fahlgren N, Kasschau KD, Carrington JC, Lundeberg J, Jansson S.** 2009. Genome-wide profiling of populus small RNAs. BMC Genomics **10:**620. http://dx.doi.org/10.1186/1471-2164-10-620.
62. **Kasschau KD, Fahlgren N, Chapman EJ, Sullivan CM, Cumbie JS, Givan SA, Carrington JC.** 2007. Genome-wide profiling and analysis of Arabidopsis siRNAs. PLoS Biol. **5:**e57. http://dx.doi.org/10.1371/journal.pbio.0050057.
63. **Mi S, Cai T, Hu Y, Chen Y, Hodges E, Ni F, Wu L, Li S, Zhou H, Long C, Chen S, Hannon GJ, Qi Y.** 2008. Sorting of small RNAs into Arabidopsis argonaute complexes is directed by the 5′ terminal nucleotide. Cell **133:**116–127. http://dx.doi.org/10.1016/j.cell.2008.02.034.
64. **Montgomery TA, Howell MD, Cuperus JT, Li D, Hansen JE, Alexander AL, Chapman EJ, Fahlgren N, Allen E, Carrington JC.** 2008. Specificity of ARGONAUTE7-miR390 interaction and dual functionality in TAS3 trans-acting siRNA formation. Cell **133:**128–141. http://dx.doi.org/10.1016/j.cell.2008.02.033.
65. **Havecker ER, Wallbridge LM, Hardcastle TJ, Bush MS, Kelly KA, Dunn RM, Schwach F, Doonan JH, Baulcombe DC.** 2010. The Arabidopsis RNA-directed DNA methylation argonautes functionally diverge based on their expression and interaction with target loci. Plant Cell **22:**321–334. http://dx.doi.org/10.1105/tpc.109.072199.
66. **Panne D, Raleigh EA, Bickle TA.** 1999. The McrBC endonuclease translocates DNA in a reaction dependent on GTP hydrolysis. J. Mol. Biol. **290:**49–60. http://dx.doi.org/10.1006/jmbi.1999.2894.

Annex: (Rajeswaran et al., 2014b)

Interactions of rice tungro bacilliform pararetrovirus and its protein P4 with plant RNA silencing machinery.

Rajendran Rajeswaran, Victor Golyaev, Jonathan Seguin, Anna Zvereva, Laurent Farinelli, and Mikhail Pooggin

Molecular Plant-Microbe Interactions (2014), proof

MPMI Vol. 27, No. 12, 2014, pp. 1370–1378. http://dx.doi.org/10.1094/MPMI-07-14-0201-R

e-Xtra*

Interactions of Rice Tungro Bacilliform Pararetrovirus and Its Protein P4 with Plant RNA-Silencing Machinery

Rajendran Rajeswaran,[1] Victor Golyaev,[1] Jonathan Seguin,[1,2] Anna S. Zvereva,[1] Laurent Farinelli,[2] and Mikhail M. Pooggin[1]

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

Submitted 7 July 2014. Accepted 4 August 2014.

Small interfering RNA (siRNA)-directed gene silencing plays a major role in antiviral defense. Virus-derived siRNAs inhibit viral replication in infected cells and potentially move to neighboring cells, immunizing them from incoming virus. Viruses have evolved various ways to evade and suppress siRNA production or action. Here, we show that 21-, 22-, and 24-nucleotide (nt) viral siRNAs together constitute up to 19% of total small RNA population of *Oryza sativa* plants infected with *Rice tungro bacilliform virus* (RTBV) and cover both strands of the RTBV DNA genome. However, viral siRNA hotspots are restricted to a short noncoding region between transcription and reverse-transcription start sites. This region generates double-stranded RNA (dsRNA) precursors of siRNAs and, in pregenomic RNA, forms a stable secondary structure likely inaccessible to siRNA-directed cleavage. In transient assays, RTBV protein P4 suppressed cell-to-cell spread of silencing but enhanced cell-autonomous silencing, which correlated with reduced 21-nt siRNA levels and increased 22-nt siRNA levels. Our findings imply that RTBV generates decoy dsRNA that restricts siRNA production to the structured noncoding region and thereby protects other regions of the viral genome from repressive action of siRNAs, while the viral protein P4 interferes with cell-to-cell spread of antiviral silencing.

Rice tungro disease is one of the major constraints for rice cultivation in Asia. The disease is caused by a complex of DNA pararetrovirus *Rice tungro bacilliform virus* (RTBV) and RNA picorna-like virus *Rice tungro spherical virus* (RTSV). RTBV is the main determinant of disease symptoms, while RTSV accentuates the symptoms and is essential for disease transmission from plant to plant by leafhoppers (Borah et al. 2013; Hull 1996). Thus far, no information on interactions of RTBV or RTSV with the plant antiviral defense system based on RNA silencing has been reported.

R. Rajeswaran and V. Golyaev contributed equally to this work.

Present address of R. Rajeswaran: Swiss Federal Institute of Technology Zurich (ETH-Zurich), Department of Biology, Zurich, Switzerland.

Corresponding author: M. M. Pooggin; Telephone: +41-61-2672314; E-mail: Mikhail.Pooggin@unibas.ch

*The *e*-Xtra logo stands for "electronic extra" and indicates that two supplementary figures, two supplementary datasets, and supplementary sequence information are published online.

RNA silencing, also known as RNA interference (RNAi), is an evolutionarily conserved sequence-specific mechanism that regulates gene expression and chromatin states and defends against invasive nucleic acids such as transposons, transgenes, and viruses (Ghildiyal and Zamore 2009; Joshua-Tor and Hannon 2011; Rajeswaran and Pooggin 2012a). RNA silencing is mediated by Dicer or Dicer-like (DCL) enzymes that catalyze processing of perfect or nearly perfect double-stranded RNA (dsRNA) into small RNA (sRNA) duplexes. One of the duplex strands gets associated with an Argonaute (AGO) family protein and guides the resulting RNA-induced silencing complex (RISC) to complementary target RNA. Following the complementary interaction, AGO cleaves target RNA or represses its translation, thereby causing post-transcriptional gene silencing (PTGS). In plants, fungi, and invertebrate animals, RISC can also target chromatin for DNA cytosine methylation or histone modification, thereby causing transcriptional gene silencing (TGS). Both PTGS and TGS can be reinforced and maintained by RNA-dependent RNA polymerase (RDR) activity that amplifies sRNA precursors. In plants, the multigene families encode DCL, AGO, and RDR with specialized functions in diverse PTGS and TGS pathways. Depending on the mechanisms of biogenesis and function, plant sRNAs are classified into miRNAs and short interfering RNAs (siRNAs). The *Oryza sativa* genome codes for at least five DCL. OsDCL1 and OsDCL3a generate canonical 21- or 22-nucleotide (nt) miRNAs and noncanonical 24-nt miRNAs, respectively, from hairpin dsRNA structures of miRNA genes' transcripts (Wu et al. 2009, 2010). Diverse populations of siRNAs are produced from multiple loci of the rice genome. OsDCL4 generates 21-nt phased (secondary) siRNAs, including trans-acting siRNAs from presumable OsRDR6-dependent dsRNA precursors (Liu et al. 2007; Song et al. 2012), whereas OsDCL3b and OsDCL3a generate 24-nt phased siRNAs and 24-nt repeat-associated siRNAs, respectively, from presumable OsRDR2-dependent dsRNA precursors. Both 24-nt miRNAs and 24-nt siRNAs associate with OsAGO4 clade proteins and direct cytosine methylation of complementary target DNA, while 21-nt miRNAs associate with OsAGO1 clade proteins and direct cleavage of target mRNAs (Song et al. 2012; Wei et al. 2014; Wu et al. 2010).

Once initiated in a single cell, plant RNA silencing can spread locally from cell to cell and systemically via phloem tissues. In dicots such as *Arabidopsis* and *Nicotiana* spp., sRNAs are part of the mobile silencing signals, with 21-nt siRNAs involved in the local spread and 24-nt siRNAs in the systemic spread of silencing (Dunoyer et al. 2010; Hamilton et al. 2002; Himber et al. 2003; Molnar et al. 2010).

The RNA silencing machinery plays a major role in defense against RNA and DNA viruses. Multiple plant DCL generate

viral siRNAs which can potentially direct PTGS and TGS (Akbergenov et al. 2006; Blevins et al. 2006; Deleris et al. 2006). Viruses have evolved various ways to evade silencing as well as actively suppress the biogenesis and action of viral siRNAs (Pooggin 2013; Rajeswaran and Pooggin 2012a). The mechanisms of silencing-based antiviral defense and viral counter-defense have been extensively studied in *Arabidopsis thaliana* and other model plants but insufficient information is available for rice. *O. sativa* plants infected with *Rice stripe virus* (RSV), an RNA tenuivirus, accumulate predominantly 21-nt viral siRNAs and less abundant 22- and 20-nt siRNAs; the viral siRNAs constitute from 0.6% (Yan et al. 2010) to 15% (Xu et al. 2012) of total sRNAs and are unevenly distributed along the segmented RSV genome, with the siRNA hotspots derived from more abundant viral mRNAs. Genetic requirements for RSV siRNA biogenesis are unknown, except that their accumulation is reduced in *OsRDR6* knock-down plants (Jiang et al. 2012). In *Arabidopsis*, RDR6 is involved in defense against RNA viruses, presumably by generating dsRNA precursors of secondary 21-nt viral siRNAs (Garcia-Ruiz et al. 2010; Wang et al. 2011). Consistent with possible involvement of OsRDR6 in antiviral defense, P6 protein of RNA rhabdovirus *Rice yellow stunt virus* interacts with OsRDR6 and interferes with production of RDR6-dependent secondary siRNAs in *Nicotiana benthamiana*-based transient assays (Guo et al. 2013). *OsDCL* that mediate antiviral defense remain unknown. Interestingly, knock down of *OsDCL2* negatively affected maintenance of an endogenous dsRNA virus by increasing the accumulation of viral siRNAs (Urayama et al. 2010). This suggests that OsDCL generating viral siRNAs are negatively regulated by an OsDCL2-dependent mechanism. In *Arabidopsis* infected with DNA viruses, four DCL produce viral siRNAs of three major size classes: 21 nt (DCL4 and DCL1), 22 nt (DCL2), and 24 nt (DCL3) (Akbergenov et al. 2006; Aregger et al. 2012; Blevins et al. 2006, 2011). In contrast, RNA viruses with cytoplasmic replication cycles are targeted mainly by DCL4 and DCL2, which generate 21- and 22-nt viral siRNAs, respectively (Blevins et al. 2006; Deleris et al. 2006). Notably, RDR do not appear to be required for the biogenesis of DNA virus-derived siRNAs, as demonstrated for *Cabbage leaf curl virus* and *Cauliflower mosaic virus* (CaMV) in *Arabidopsis* (Aregger et al. 2012; Blevins et al. 2011), and dsRNA precursors of viral siRNAs are presumably produced by Pol II-mediated sense and antisense transcription of circular viral DNA (Pooggin 2013).

RTBV and CaMV belong to distinct genera of the family *Caulimoviridae*, also known as plant pararetroviruses (Hohn and Rothnie 2013). They share the mechanisms of host Pol II-mediated transcription of viral DNA into pregenomic RNA (pgRNA, also known as 35S RNA) and viral replicase-mediated reverse transcription of pgRNA. In both cases, translation of pgRNA is initiated by a ribosome shunt mechanism by which the ribosome bypasses a long and highly structured leader sequence preceding the first large viral open reading frame (ORF) (Pooggin et al. 2006, 2008). However, translation of the further downstream ORF on polycistronic pgRNA proceeds by different mechanisms: leaky scanning in RTBV (Fütterer et al. 1997) and reinitiation in CaMV (Ryabova et al. 2006). The gene most distal from the pgRNA promoter (ORF IV and ORF VI, respectively) is also expressed by different mechanisms. In RTBV, splicing of pgRNA joins the first short ORF in the leader with ORF IV, creating a monocistronic mRNA for P4 (Fütterer et al. 1994). In CaMV, Pol II drives transcription of a subgenomic 19S RNA, the monocistronic mRNA for P6. Studies of RNA silencing-based antiviral defense in *A. thaliana* imply that CaMV has evolved protein-based and RNA-based counter-defense strategies. The CaMV P6 protein interacts with dsRNA-binding protein DRB4 and interferes with DCL4-mediated processing of RDR6-dependent dsRNAs into secondary siRNAs (Haas et al. 2008; Shivaprasad et al. 2008). The CaMV leader region generates decoy dsRNA (8S RNA) engaging all four DCL in massive production of viral siRNAs and, thereby, protecting other regions of the viral genome from repressive action of siRNAs (Blevins et al. 2011). In this work, we investigated whether RTBV also expresses decoy dsRNA and whether the RTBV protein P4 of previously unknown function has antisilencing activity.

RESULTS

Massive production of viral siRNAs is restricted between transcription and reverse-transcription start sites of the RTBV genome.

Using Illumina sequencing, we analyzed sRNA populations from leaves of mock-inoculated and RTBV-infected *O. sativa* japonica plants. We restricted our bioinformatics analysis to a size range from 20 to 25 nt, containing known rice miRNAs and siRNAs and constituting the majority of reads in our sRNA libraries (Supplementary Dataset S1). Of the total 20- to 25-nt reads in two biological replicates of RTBV-infected rice plants, 17% (1,732,734 reads) and 19% (1,637,597 reads) mapped to the 8-kb viral genome with zero mismatches (Fig. 1A). Given a much bigger size of the rice genome (approximately 500 Mb) and limitation of the virus to vascular tissues (Cruz et al. 1993), we conclude that viral sRNA production in RTBV-infected cells is massive. In this respect, RTBV resembles the distantly related pararetrovirus CaMV that spawns massive quantities of viral siRNAs, comparable with the entire population of endogenous miRNAs and siRNAs in *Arabidopsis* (Blevins et al. 2011).

In the absence of viral infection, size profile of sRNAs mapped to the rice genome with zero mismatches was dominated by 24-nt reads (55%), followed by 21-nt (13%), 23-nt (12%), 20-nt (10%), 22-nt (7%), and 25-nt (3%) reads. Upon RTBV infection, the proportions of 24- and 23-nt sRNAs were slightly reduced, whereas those of 20- to 22-nt sRNAs were slightly increased (Fig. 1B). The biological significance of this alteration in the rice sRNA profile remains to be investigated. Interestingly, 79 to 89% of 20-nt sRNAs in both mock-inoculated and virus-infected plants were found to possess 5′-guanosine (5′G). These 5′G-sRNAs (20G-RNA) may represent a previously uncharacterized class of rice sRNAs whose biogenesis and function remain to be investigated. In a parallel deep-sequencing study, we found that 20G-RNA is a major class of endogenous sRNAs in *Musa acuminata* banana plants (Rajeswaran et al. 2014), suggesting that 20G-RNAs might be a common feature of monocots. As could be expected from previous studies of rice sRNAs (Song et al. 2012; Wei et al. 2014; Wu et al. 2010), other major size classes also exhibit 5′-nucleotide biases, with 5′-uridine (5′U) being predominant in 21-nt sRNAs (66 to 72%), and 5′-adenosine (5′A) in 24-nt (49 to 51%) and 22-nt (46 to 50%) sRNAs. Similar biases were found in the presence of RTBV. The latter finding suggests that preferential association of endogenous sRNAs with specific OsAGO proteins may not be affected by virus infection. It should be mentioned, however, that 5′-nucleotide biases may also be an artifact of the sRNA library preparation for Illumina sequencing (Sorefan et al. 2012).

Bioinformatics analysis of redundant reads mapped to the RTBV genome with zero mismatches revealed that, in both replicates, 21-nt sRNAs are the most abundant (approximately 45%), followed by 22-nt (30 to 31%) and 24-nt (15 to 18%) sRNAs. Other size classes were below 5% of total 20- to 25-nt viral sRNAs (Fig. 1C). Compared with endogenous sRNAs, viral sRNAs exhibit similar but less pronounced 5′-nucleotide biases in each of the three major size classes. In addition to be-

ing a minor fraction, 20-nt viral sRNAs do not exhibit any bias to 5′G, suggesting that the 20G-RNA pathway does not target RTBV. Distribution of viral sRNAs along the RTBV genome was found to be uneven: approximately 50 to 60% of total 20- to 25-nt reads fell into the pgRNA noncoding leader region. More specifically, most of the viral sRNA hotspots in both polarities concentrated within a portion of the leader region between the start sites for transcription and reverse transcription (Fig. 1D; Supplementary Dataset S2). In other regions of the 8-kb genome, sRNA reads were distributed more or less evenly, although their density is somewhat higher in the P4 mRNA region (Fig. 1D). Sense and antisense reads tiled along both the

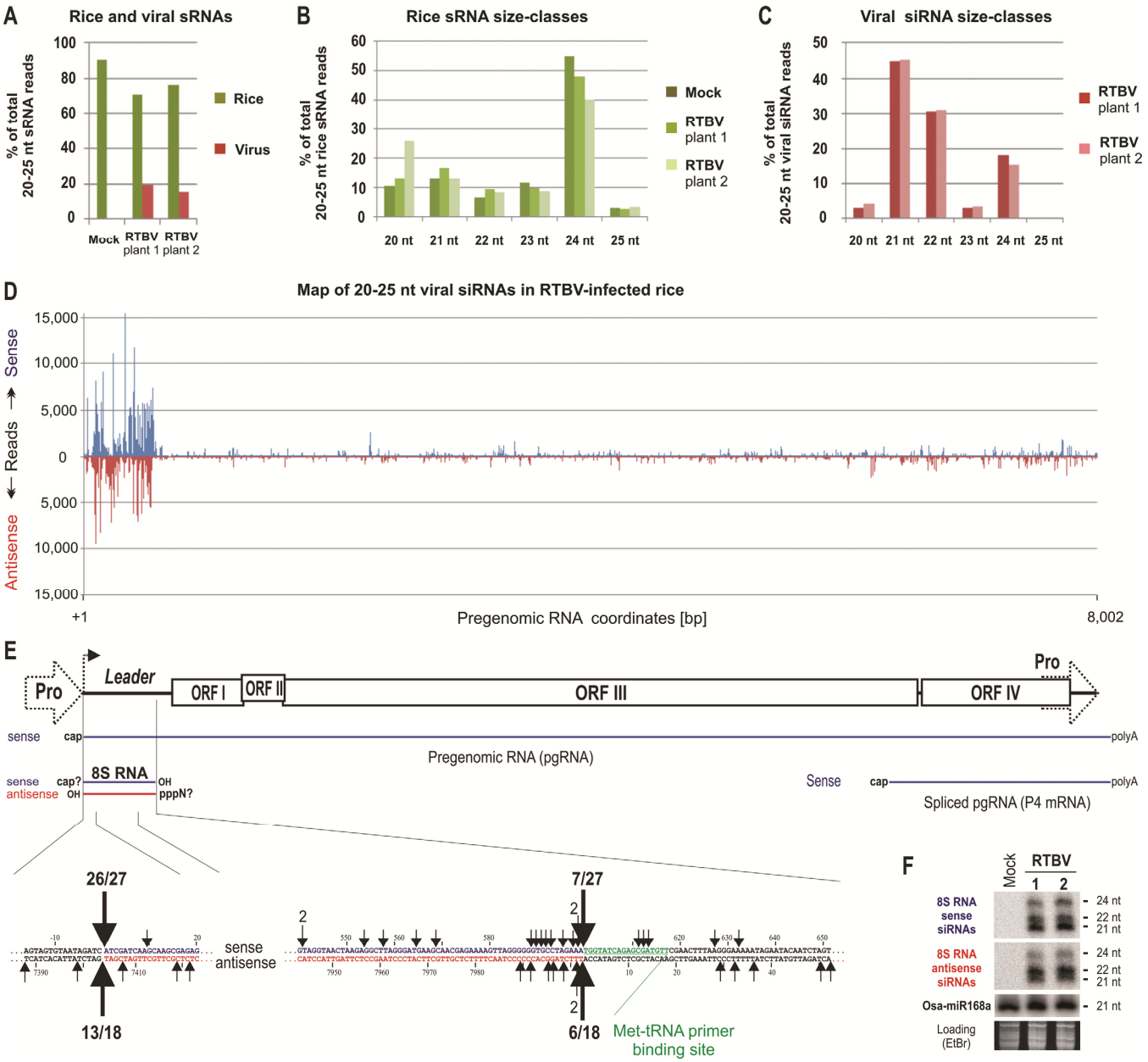


Fig. 1. Deep-sequencing and blot hybridization analyses of *Rice tungro bacilliform virus* (RTBV) siRNAs and their precursor mapping. Bar graphs show **A,** percentages of 20- to 25-nucleotide (nt) viral sRNAs in the pool of total (host and viral) 20- to 25-nt reads mapped to the *Oryza sativa* and RTBV genome with zero mismatches; **B,** percentages of each size-class of 20- to 25-nt rice sRNA reads mapped to the genome of virus-free and RTBV-infected plants (1 and 2); **C,** percentages of each size-class of 20- to 25-nt viral sRNA reads mapped to the viral genome. **D,** Genome-wide map of viral siRNAs from RTBV-infected rice (plant 1) at single-nucleotide resolution. The graph plots the number of 20- to 25-nt siRNA reads at each position of the 8,002-bp RTBV genome; the numbering starts at the 5′ terminus of the pgRNA (genome position 7404) as depicted in E. Bars above the axis represent sense reads starting at each respective position; those below represent antisense reads ending at the respective position. **E,** Circularization reverse-transcription polymerase chain reaction (cRT-PCR) mapping of RTBV 8S RNA from virus-infected rice. The circular RTBV genome organization is shown schematically with viral genes (boxes) and the promoter (dotted arrow) that drives Pol II transcription of pgRNA (depicted below the genome). The position and termini of the RTBV leader region-derived 8S RNA and its antisense counterpart are indicated, as determined for sense and antisense polarities by cRT-PCR product sequencing. Regions surrounding the pgRNA start site and the Met-tRNA primer binding site are enlarged. Termini of sequenced 8S RNA clones are indicated by arrows above the sequence for sense RNAs (27 clones) and below the sequence for antisense RNAs (18 clones). The number of clones is given when more than one clone had the same 5′ or 3′ terminus. Thick arrows indicate the major start and termination sites for sense 8S RNA and its antisense counterpart, which fall on the same genome positions. Position of the Met-tRNA primer binding site for reverse transcription is indicated with bent lines. **F,** Blot hybridization analysis of RTBV leader region-derived siRNAs. Total RNA samples from mock-inoculated and RTBV-infected rice (plant 1 and plant 2) were analyzed by 15% polyacrylamide gel electrophoresis followed by blot hybridization. The membrane was successively hybridized with sense and antisense probes for the RTBV 8S dsRNA and the probe for the rice miRNA Osa-miR168 (21 nt). Ethidium bromide staining serves as a loading control. Sizes are indicated on each data image.

hot and cold regions of sRNA production (Fig. 1D), with an overall approximately 60% bias of total 20-to 25-nt reads to the sense strand. Notably, this sRNA profile is very similar for each of the three major size classes of viral sRNAs. Taken together, these findings indicate that the rice silencing machinery responds to RTBV infection by generating typical viral siRNAs of three major size classes which resemble viral siRNAs previously characterized in *Arabidopsis* infected with DNA viruses (Aregger et al. 2012; Blevins et al. 2011). By analogy with DNA virus-derived siRNAs in *Arabidopsis* (Aregger et al. 2012; Blevins et al. 2006, 2011), RTBV-derived siRNAs are likely produced by multiple OsDCL and potentially associated with multiple OsAGO in active RISC to direct silencing of the viral genes. However, RTBV appears to evade the repressive action of viral siRNAs by restricting their production to the noncoding region (discussed below).

Reconstruction and precise mapping of dsRNA precursors of viral siRNAs.

Bioinformatics analysis of nonredundant reads revealed that viral siRNA species cover both sense and antisense strands of the circular RTBV genome without gaps, which allows for de novo reconstruction of the entire dsDNA genome from siRNAs, as we previously demonstrated for CaMV and other DNA and RNA viruses (Seguin et al. 2014b) (the results of RTBV reconstruction are described below). This finding indicates that multiple viral siRNA duplexes are processed from perfect dsRNA precursors covering the entire circular RTBV genome. The overall bias of redundant reads to the sense strand (approximately 60% of total siRNAs) can be explained by differential stability of single-stranded siRNA species following the interaction of siRNA duplexes with OsAGO and incorporation of one of the two strands into RISC. However, association of OsAGO with viral siRNAs remains to be confirmed experimentally. Furthermore, we cannot exclude that a fraction of viral siRNAs of sense polarity are produced from secondary structures of pgRNA or its spliced version. It has been reported that positive-sense RNA virus-derived siRNA populations exhibit strong bias to the sense polarity, suggesting their processing from the secondary structures of viral genomic RNA (Molnar et al. 2005; Szittya et al. 2010).

Restriction of the viral siRNA hotspots to the RTBV leader region suggested that these siRNAs could be generated from a separate dsRNA precursor, which is efficiently expressed and processed by OsDCL. To confirm the expression of this presumptive dsRNA in rice plants, we used circularization reverse-transcription polymerase chain reaction (cRT-PCR) that allows simultaneous sequencing of 5′ and 3′ termini of any given RNA (Supplementary Fig. S1A). The results of cRT-PCR analysis (Fig. 1E) revealed that sense transcripts in the leader region start predominantly at position A7404 (26 of 27 cRT-PCR clones), coinciding with the 5′ terminus of pgRNA as mapped previously (Bao and Hull 1993), and terminate at position A8002 (seven of 27 clones), coinciding with the DNA gap site just upstream of the Met-tRNA primer binding site (Bao and Hull 1992) or in the close vicinity of this position. The antisense transcripts mapped by cRT-PCR in the leader region start at position U8002 (six of 18 clones) or in the close vicinity of this position, and terminate predominantly at position U7404 (13 of 18 clones) (Fig. 1E), thus mirroring the sense transcripts. Attempts to identify longer antisense transcripts using a primer annealing downstream of the reverse-transcription start site did not yield any distinct cRT-PCR product. Taken together, a large fraction of the antisense transcripts is fully comple-

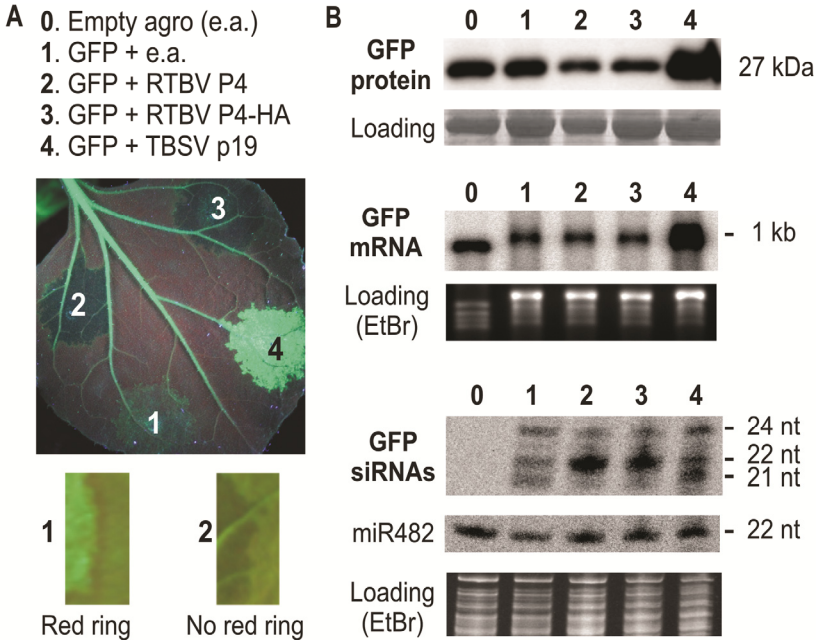


Fig. 2. *Rice tungro bacilliform virus* (RTBV) P4 suppresses cell-to-cell spread of green fluorescent protein (GFP) silencing but enhances cell-autonomous GFP silencing in *Nicotiana benthamiana* 16c plants. **A,** Analysis of *Agrobacterium*-infiltrated leaf tissues in line 16c under UV light. Four leaf tissue patches shown in the image were co-infiltrated with the agro-strain carrying the GFP expression cassette (GFP) in combination with the agro-strain carrying no vector (patch 1), the RTBV P4 (patch 2), the RTBV P4-HA (patch 3), or the TBSV p19 (patch 4) expression cassettes, and the picture was taken at 8 days post-infiltration. Below, the cropped images of the leaf patches co-infiltrated with GFP + empty vector (patch 1) and GFP + RTBV P4 (patch 2) are enlarged. A thin border of red tissue (red ring) is visible in the absence of P4 but not in its presence. **B,** Molecular analysis of the agroinfiltrated tissues from A (lane numbers correspond to the patch numbers) by Western, Northern, and sRNA blot hybridization. The Western blot membrane was probed with GFP protein-specific antibody and then stained with amidoblack for loading control (Rubisco). The Northern blot membrane was probed with the GFP mRNA specific probe; ethidium bromide (EtBr) staining of the agarose gel before blotting was used as a loading control. The sRNA blot membrane was successively hybridized with DNA oligonucleotide probes specific for the GFP 3′ untranslated region sequence-derived siRNAs and the *N. benthamiana* miRNA miR482; EtBr staining of the gel before blotting was used as a loading control.

mentary to the sense transcripts, suggesting the existence of a 599-bp blunt-ended dsRNA spanning the region between the start sites for transcription and reverse transcription. This dsRNA and its slightly shorter or longer versions appear to be the major precursors of the highly abundant viral siRNAs.

In CaMV, the pgRNA leader region also generates dsRNA, although its terminus next to the Met-tRNA primer binding site was found to be much more heterogeneous than that in RTBV (Blevins et al. 2011). Notably, the primer binding site for reverse transcription in the pgRNA leader sequence is located within the stem-loop structure in RTBV and downstream of the structure in CaMV. The 90-nt sequence located just downstream of reverse-transcription start site, which is involved in formation of the most stable section of the RTBV pgRNA leader secondary structure (Pooggin et al. 2006), does not produce highly abundant siRNAs. Thus, stable secondary structure per se is not a main determinant of massive siRNA production in RTBV. Note that the existence and functional role of this secondary structure were validated by sequence replacement analyses (Pooggin et al. 2006, 2008).

RTBV P4 can suppress cell-to-cell spread of RNA silencing.

Functional analysis of RTBV P4 has not been reported thus far. Other pararetroviruses from a closely related genus (*Badnavirus*) or other genera of the family *Caulimoviridae* do not possess any *P4*-related gene. Therefore, we hypothesized that *Tungrovirus P4* may not be essential for viral replication, assembly, or movement and this gene may have been acquired to counteract the host plant defenses based on RNA silencing or innate immunity (Zvereva and Pooggin 2012). To test whether P4 protein has antisilencing activity, we employed a classical transient assay in leaves of the *N. benthamiana* transgenic line 16c expressing green fluorescent protein (GFP). In this system, cell-autonomous and mobile silencing of the GFP transgene can be induced by *Agrobacterium*-mediated ectopic expression of the cognate GFP gene, which triggers production of 21- to 24-nt GFP siRNAs, and suppressed by co-expression of a protein that interferes with siRNA production or action (Hamilton et al. 2002; Himber et al. 2003). As a positive control for silencing suppression, we used the well-studied protein p19 from *Tomato bushy stunt virus* (TBSV) that binds 21-nt siRNA duplexes and thereby suppresses PTGS (Vargason et al. 2003). Confirming previous findings, co-expression of TBSV p19 with the GFP silencing trigger resulted in strong suppression of both cell-autonomous GFP silencing and its cell-to-cell spread (Fig. 2; Supplementary Fig. S2). Surprisingly, co-expression of the RTBV protein P4 (or its HA- and FLAG-tagged versions) enhanced cell-autonomous GFP silencing, as was manifested by significant reduction in both GFP fluorescence and GFP protein accumulation compared with the control (Fig. 2). Because GFP mRNA levels were not significantly affected by P4 (Fig. 2), the observed enhancement of cell-autonomous GFP silencing appears to occur at the mRNA translation level. Nonetheless, RTBV P4 abolished cell-to-cell spread of GFP silencing because no development of a characteristic red ring around the infiltration zone was observed (Fig. 2). The red ring development in this system depends on a mobile silencing signal moving from cell to cell for a short distance (10 to 15 cells) and triggering silencing of the GFP transgene outside of the infiltrated cells' perimeter (Himber et al. 2003). Because 21-nt siRNAs appear to be (part of) this mobile signal (Dunoyer et al. 2010; Himber et al. 2003), we examined accumulation of GFP siRNAs in the infiltrated cells by sRNA blot hybridization. Indeed, the accumulation of 21-nt siRNAs was strongly reduced by RTBV P4, accounting for the absence of red ring. The reduced accumulation of 21-nt siRNAs was accompanied by increased accumulation of 22-nt siRNAs, while 24-nt siRNAs were only slightly affected (Fig. 2). Thus, RTBV P4-mediated enhancement of cell-autonomous GFP silencing is correlated with the increased production of 22-nt siRNAs.

To suppress silencing, the TBSV p19 binds preferentially to 21-nt siRNA duplexes and, thereby, prevents their interaction with AGO and assembly of RISC. Because the affinity of TBSV p19 for 21-nt duplexes is sixfold higher than its affinity for 22-nt duplexes (Vargason et al. 2003), we reasoned that RTBV P4 might counteract TBSV p19-mediated suppression of silencing by promoting the production of 22-nt siRNAs. Indeed, co-expression of RTBV P4 substantially reduced, although not abolished, TBSV p19-mediated suppression of cell-autonomous GFP silencing. Interestingly, the co-expression of these proteins in combination still abolished cell-to-cell spread of GFP silencing and red ring development, consistent with the abilities of RTBV P4 and TBSV p19 to block the biogenesis or the function of 21-nt siRNAs, respectively. Taken together, these findings imply that 22-nt siRNAs can potentially replace 21-nt siRNAs in cell-autonomous silencing but not in cell-to-cell spread of silencing. To our knowledge, RTBV P4 exhibits a unique combination of anti- and prosilencing activities in *N. benthamiana* 16c leaves, which was not reported for other proteins studied in this classical assay.

DISCUSSION

Massive production of viral siRNAs in plant defense and virus counter-defense.

Our results demonstrate that, in response to RTBV infection, the rice silencing machinery generates massive quantities of 21-, 22-, and 24-nt viral siRNAs. This indicates that multiple OsDCL are involved in robust antiviral defense. By analogy with DNA viruses in *Arabidopsis* (Aregger et al. 2012; Blevins et al. 2006, 2011) and based on the reported activities of OsDCL (Song et al. 2012; Wei et al. 2014; Wu et al. 2010), the 24-nt viral siRNAs might be produced by OsDCL3a or OsDCL3b and the 21-nt viral siRNAs by OsDCL4 or OsDCL1. However, this hypothesis needs to be validated experimentally. Furthermore, it remains to be investigated whether OsDCL2 is involved in siRNA biogenesis and, in particular, in production of the 22-nt viral siRNAs, similar to *Arabidopsis* DCL2.

The 24-nt siRNA class is characteristic for DNA viruses and viroids, which have a nuclear phase in their replication cycle (Aregger et al. 2012; Akbergenov et al. 2006; Blevins et al. 2006, 2011; Pooggin 2013; Seguin et al. 2014a). This class is underrepresented in rice plants infected with the RNA virus RSV (Xu et al. 2012; Yan et al. 2010) and in other plant species infected with cytoplasmic RNA viruses (Rajeswaran and Pooggin 2012a). The 5′-nucleotide biases observed for each of the three major size classes of RTBV siRNAs indicate that, like endogenous siRNAs, the viral siRNAs of each class may preferentially associate with specific OsAGO proteins and guide the resulting RISC to cognate viral RNA and DNA. However, production of highly abundant viral siRNAs is restricted to the 599-bp noncoding region just downstream of the Pol II transcription start site, which protects other regions of the viral genome from the repressive action of RISC. Our findings imply that this noncoding region generates decoy dsRNA engaging multiple OsDCL in massive siRNA production, and we assume that the decoy-derived siRNAs themselves may not be effective in directing silencing of the viral genes.

The RTBV decoy-derived siRNAs of antisense polarity can potentially target the complementary sequences of pgRNA leader that precede ORF I. However, the 700-nt pgRNA leader sequence folds into a stable secondary structure bypassed by shunting ribosomes (Pooggin et al. 2006, 2008) and this structure is unlikely to be accessible for RISC. Thus, plant viroids

with highly structured RNA genomes are resistant to cleavage and degradation by viroid-derived siRNAs (Itaya et al. 2007). At the DNA level, the RTBV decoy region is located downstream of the pgRNA promoter and, therefore, potential cytosine methylation at this region directed by viral siRNAs may not interfere with the promoter activity. Moreover, siRNA-directed DNA methylation (RdDM) requires the interaction of RISC with a nascent RNA transcript at the RdDM target loci (Pooggin 2013). Given that the leader secondary structure can be formed in the nascent pgRNA transcript, viral siRNA-RISC may not be able to establish a complementary interaction with target RNA sequences and bring about the de novo methyltransferase. Consistent with this hypothesis, in CaMV-infected plants, the viral circular dsDNA that serves as a template for Pol II transcription of pgRNA was found to be nonmethylated (Covey et al. 1997). Taken together, the decoy strategy previously proposed for CaMV and here for RTBV may allow these and possibly other plant pararetroviruses to evade both PTGS and TGS.

The long and highly structured leader sequence is a common feature of plant pararetroviruses which regulates translation initiation of pgRNA and contains a signal of pgRNA packaging for reverse transcription (Guerra-Peraza et al. 2000; Pooggin et al. 1999). However, the secondary structure per se is not likely to be the main determinant for the biogenesis of decoy dsRNA precursor of viral siRNAs, as was suggested in the previous studies of CaMV (Blevins et al. 2011; Moissiard and Voinnet 2006). Here, we found that the hotspots of siRNA production do not encompass the entire secondary structure of the RTBV leader. Furthermore, the termini of the decoy dsRNA were mapped precisely to the start sites of transcription and reverse transcription. This finding strongly suggests that the sense strand of decoy dsRNA is generated by abrupt termination of Pol II-mediated transcription of pgRNA on a fraction of viral dsDNA, in which the minus strand gap (Bao and Hull 1992) had not yet been repaired after reverse transcription. Such run-off transcript (8S RNA) was identified in CaMV-infected plants and precisely mapped by cRT-PCR (Blevins et al. 2011). In further support of Pol II involvement, we found that a small fraction of the RTBV sense transcripts have a short stretch of adenines, suggesting that the run-off transcript might be inefficiently polyadenylated by the Pol II machinery. An RNA polymerase that generates the antisense strand of decoy dsRNA remains to be identified. Genetic evidence in *Arabidopsis* ruled out involvement of three functional RDR (RDR1, RDR2, and RDR6) or DNA-dependent RNA polymerase Pol IV and Pol V in production of CaMV leader-derived siRNAs (Blevins et al. 2006, 2011). Because a major fraction of the mapped dsRNA precursors of RTBV leader-derived siRNAs is blunt ended (Fig. 1E), OsRDR6 (and possibly other OsRDR) is not likely to be involved. In *Arabidopsis* and tomato, RDR6 initiates synthesis of complementary RNA preferentially at the third nucleotide from the template's 3′ end and adds one or two nontemplate nucleotides to the run-off transcript, thereby creating dsRNA with 1- or 2-nt 3′ overhangs at both ends (Rajeswaran and Pooggin 2012b; Rajeswaran et al. 2012; Schiebel et al. 1993).

Dual strategy in viral counter-defense.

In addition to the RNA-based decoy strategy of silencing evasion, RTBV appears to have evolved a protein-based strategy of silencing suppression, thus further resembling its distant relative CaMV. In both cases, protein suppressors of silencing may serve as a backup in case the decoy-mediated evasion fails. In fact, the production of decoy dsRNA would rely on incomplete repair of the discontinuous viral dsDNA following its delivery to the nucleus (Pooggin 2013). Furthermore, in addition to highly abundant siRNAs produced from the decoy dsRNA, siRNAs of low abundance are generated from other regions of the viral genome in both CaMV (Blevins et al. 2011; Seguin et al. 2014a) and RTBV (this study). Viral siRNA species cover both strands of the CaMV and RTBV genomes without gaps and can potentially direct cleavage of complementary viral transcripts. Products of inefficient cleavage events directed by those low-abundant primary siRNAs in unstructured sequences of viral pgRNA or spliced or subgenomic RNA can potentially be converted by RDR6 (or other RDR) into dsRNA precursors of secondary siRNAs. In CaMV, the viral protein P6 interferes with amplification of 21-nt secondary siRNAs by blocking DCL4-mediated processing of RDR6-dependent dsRNAs (Haas et al. 2008; Shivaprasad et al. 2008). Here, we demonstrate that the RTBV protein P4, of previously unknown function, interferes with the biogenesis of transgene-derived 21-nt siRNAs in *N. benthamiana* and blocks cell-to-cell spread of transgene silencing likely mediated by 21-nt siRNAs. This implies that RTBV P4 interferes with DCL4 activity, although *N. benthamiana* DCL remain to be functionally characterized. Furthermore, RTBV P4 seems to have unique properties, not reported for other viral silencing suppressors: while suppressing cell-to-cell spread of silencing, it enhanced cell-autonomous silencing in *N. benthamiana* by promoting accumulation of 22-nt siRNAs. Because *N. benthamiana* is a dicot plant and cannot support RTBV infection, it remains to be investigated whether or not the P4 properties uncovered in *N. benthamiana* are relevant in the context of RTBV infection in rice. Both 21- and 22-nt viral siRNAs accumulate at comparable levels in RTBV-infected plants. Thus, the biogenesis of 21-nt viral siRNAs does not appear to be blocked. We assume that the RTBV *P4* gene is expressed only during early stages of viral infection, because P4 protein is translated from the spliced pgRNA (Fütterer et al. 1994). The splicing is likely repressed at the late stages of infection to promote production of the full-length pgRNA for reverse transcription. Analysis of the P4 protein activities at the early stages of viral infection would be important to further investigate its interactions with the rice defense system.

Taken together, our findings imply that the pararetrovirus RTBV has evolved a dual counter-defense strategy, in which the viral decoy dsRNA restricts siRNA production to a highly structured noncoding region and thereby protects other regions of the viral genome from the repressive action of siRNAs, while the viral protein P4 interferes with cell-to-cell spread of antiviral silencing.

MATERIALS AND METHODS

Plant growth conditions and infection with RTBV.

Seedlings of *O. sativa* japonica 'Taipei 309' were grown in a phytochamber with approximately 80% humidity at 28 to 30°C and, at approximately 3 weeks postgermination, inoculated with an infectious clone of RTBV isolate Philippines (GenBank accession X57924) (Hay et al. 1991) using *Agrobacterium tumefaciens* GV3859 harboring pRTRB1162 (or the empty vector pBin19 for mock inoculation), as described previously (Dasgupta et al. 1991). At 50 days postinoculation, nine of 12 inoculated plants showing the characteristic disease symptoms (slight stunting of the plant and weak yellowing of the leaves) were taken for total RNA preparation.

Total RNA preparation and sRNA blot hybridization analysis.

Total RNA was extracted from leaves of the mock-inoculated and RTBV-infected plants using Trizol reagent (Life Technologies) as per the manufacturer's protocol. Total RNA

preparation and high-resolution blot hybridization were performed as described in detail previously (Akbergenov et al. 2006; Rajeswaran et al. 2012). All nine plants that showed the disease symptoms scored positive for RTBV viral RNA and virus-derived siRNAs (data not shown). Total RNA from the plants designated as 1 and 2 (Fig. 1F) were taken for deep sequencing of the sRNA population using Illumina sequencing.

Illumina sequencing and bioinformatic analysis of viral and endogenous sRNAs.

Complementary DNA libraries of the 19- to 30-nt RNA fraction of the total RNA were prepared following Illumina protocols, as described previously (Aregger et al. 2012). The libraries were sequenced on the Genome Analyzer HiSeq 2000 using a TruSeq kit v5. After trimming the adaptor sequences, the datasets of reads were mapped to the reference genome sequences of *O. sativa* subsp. *japonica* MSU6 (Kawahara et al. 2013) and RTBV using a Burrows-Wheeler Alignment Tool (BWA version 0.5.9) with zero mismatches to each reference sequence. The bioinformatics analysis of the mapped reads is summarized in Figure 1. Reads mapping to several positions on the reference genome were attributed at random to one of them. To account for the circular RTBV genome, the first 50 bases of the viral sequence were added to its 3′ end. For each reference genome or sequence and each sRNA size class (20 to 25 nt), we counted total number of reads, reads in forward and reverse orientation, and reads starting with A, C, G, and T. The single-base resolution maps of 20-, 21-, 22-, 23-, 24-, and 25-nt viral sRNAs were generated by map tool MISIS (Seguin et al. 2014b). In these maps, for each position on the sequence (starting from the 5′ end of the reference sequence), the number of matches starting at this position in forward (first base of the read) and reverse (last base of the read) orientation for each read length is given. Note that the reads mapped to the last 50 bases of the extended viral sequence were added to the reads mapped to the first 50 bases. By default, MISIS generates two maps: one map with zero mismatches and another map with up to two mismatches to the reference genome. The comparison of the two maps was informative for initial identification of mismatches or errors in the RTBV reference sequence as well as single nucleotide polymorphisms (SNPs).

We then applied a recently developed siRomics approach that allows for de novo reconstruction of consensus master genomes of RNA and DNA viruses from viral siRNAs (Seguin et al. 2014b). This reconstruction revealed 13 single-nucleotide mismatches between the reference RTBV genome and the viral genome reconstructed form both biological replicates of RTBV-infected plants (RTBV1 and RTBV2). These mismatches were verified by resequencing the complete RTBV infection clone used for inoculation and found to be errors in the reference RTBV sequence deposited to the GenBank. The correct sequence of the RTBV infectious clone is given in Supplementary Sequence Information S1 (designated RTBV1). In addition, in one of the two biological replicates (designated RTBV2), one mismatch was identified at position 7968 of the RTBV genome: G to A transition, supported by 75% redundant reads and 60% nonredundant reads. This SNP illustrates the quasi-species nature of RTBV and other viruses, which can accumulate mutations even within one systemic infection cycle in a single plant, following inoculation. With an arbitrary threshold of 10% redundant reads, several other SNPs were identified in each of the two replicates but those were supported by a maximum of 13% of sRNA reads (not shown).

cRT-PCR mapping.

cRT-PCR mapping of transcripts from the RTBV leader region was performed as described in detail previously (Blevins et al. 2011; Shivaprasad et al. 2005). Briefly, decapping was performed on 10 μg of total RNA using tobacco acid pyrophosphatase (Epicentre Technologies, Madison, WI, U.S.A.) in the presence of RNase inhibitor RNasin (Promega Corp., Madison, WI, U.S.A.). After chloroform extraction, RNA was precipitated with ethanol and circularized using T4 RNA ligase 1 (NEB) in the presence of RNasin. Following extraction with chloroform, ligation products were precipitated with ethanol. Circular RNA was reverse transcribed using SuperScriptIII reverse transcriptase (Invitrogen) with either RtbvL_as2 or RtbvL_s1 primer. The cDNA synthesized with the RtbvL_s1 primer was PCR amplified using Taq DNA polymerase (NEB) with the RtbvL_s1 primer together with RtbvL_as1 or RtbvL_as2 primers, and cDNA synthesized with the RtbvL_as2 primer was PCR amplified with the RtbvL_as2 primer together with RtbvL_s1, RtbvL_s2, or RtbvL s3 primers. PCR products were analyzed by agarose gel electrophoresis, excised from the gel, cloned in pGEM-T Easy vector (Promega Corp.), and sequenced.

Transient expression assay in leaves of *N. benthamiana* line 16c.

The RTBV P4 ORF was subcloned from the RTBV infectious clone into the pEarlyGate vectors 100 (no tag), 201 (HA tag), and 202 (FLAG tag), using PCR primers AttB1_Rtbv4_s and AttB2_Rtbv4_as. To account for the pgRNA splicing that creates P4 mRNA (Fütterer et al. 1994), the sequence upstream of the splice junction, which contains the P4 start codon, was imbedded in forward primer AttB1_Rtbv4_s. The resulting plasmids, which carry the CaMV 35S promoter-driven P4 protein expression cassettes (designated RTBV P4, RTBV P4-HA, and RTBV P4-FLAG), were mobilized to *A. tumefaciens* C58C1 for agroinfiltration assays.

Transient expression experiments using *N. benthamiana* 16c line were performed as described previously (Hamilton et al. 2002; Himber et al. 2003). The plants were grown at 23 to 24°C under natural light and, after 4 to 5 weeks postgermination, infiltrated with agrobacteria. The agrobacteria strains were inoculated into 2 ml of Luria-Bertani media supplemented with kanamycin at 50 mg/liter and rifampicin at 10 mg/liter and grown at 28°C for 16 h. Cells were pelleted and resuspended to an optical density of 0.3 at 600 nm in a solution containing 10 mM 2-(N-morpholino)ethanesulphonic acid, 10 mM $MgCI_2$, and 100 μM acetosyringone and, before infiltration, mixed in equal proportions. GFP fluorescence was monitored under UV light at 3, 5, 8, 10, and 20 days postinfiltration (dpi). Samples of the infiltrated tissues were taken at 8 dpi and used for the molecular analysis shown in Figure 2. The same tissue sample was ground in liquid nitrogen and aliquots were taken for protein and RNA analyses.

GFP protein accumulation was analyzed by Western blotting. The infiltrated leaf tissues ground in liquid nitrogen were mixed with 100 to 200 μl of 6× sodium dodecyl sulfate (SDS) sample buffer (0.35M Tris [pH 6.8], 22.4% glycerol, 10% SDS, 0.6% dithiothreitol, and bromphenol blue), boiled for 5 min, and centrifuged at 12,000 × *g* for 5 min, and the supernatant was loaded onto 12% SDS polyacrylamide gel electrophoresis. After protein separation and blotting to Hybond-P membrane (Amersham), the membrane was incubated with α-GFP antibody (1:1000; Sigma) followed by α-mouse HRP secondary antibody (1:10000; SouthernBiotech). GFP-specific bands were visualized using the ECL Prime Western Blot Detection Reagent (Amersham Bioscience). For loading control, the membrane was stained with amidoblack.

GFP mRNA accumulation analysis by classical Northern blot hybridization and GFP siRNA analysis by high-resolution sRNA blot hybridization were performed as described previ-

ously (Akbergenov et al. 2006; Blevins et al. 2006). Total RNA was extracted from the infiltrated leaf tissues ground in liquid nitrogen using TRIzol reagent (Sigma), following the manufacturer's instructions, with an additional first step using GHCL extraction buffer (6.5 M guanidinium hydrochloride, 100 mM Tris-HCI [pH 8.0], 0.1 M sodium acetate [pH 5.5], and 0.1 M β-mercaptoethanol). For analysis of GFP mRNA, 10-µg samples of total RNA were separated on formaldehyde denaturing 1.2% agarose gel. For sRNA analysis, 10-µg samples of total RNA were separated on 15% polyacrylamide gel. Ethidium bromide staining of the gels was used for loading control. In both cases, RNAs were transferred to Hybond N+ membrane (Amersham) and the membrane was hybridized with ^{32}P-ATP-labeled, GFP 3′ untranslated region-specific probe mGFP5_Nosterm_as. The sRNA membrane was then stripped and hybridized with miR482-specific probe.

ACKNOWLEDGMENTS

We thank T. Boller for supporting research of M. M. Pooggin's group at the University of Basel, T. Hohn for helpful discussions and support of R. Rajeswaran, and I.-R. Choi for providing the infectious clone of RTBV. The work was supported by the European Commission Marie Curie fellowship PIIF-237493-SUPRA to R. Rajeswaran, the Swiss Government Excellence Scholarship to V. Golyaev, the Swiss National Science Foundation grants 31003A_143882 to M. M. Pooggin and 31003A_122469 to T. Hohn and M. M. Pooggin, and the European Cooperation in Science and Technology action FA0806 grant SER No. C09.0176 to L. Farinelli and M. M. Pooggin. R. Rajeswaran, V. Golyaev, L. Farinelli and M. M. Pooggin conceived and designed research; R. Rajeswaran, V. Golyaev and A. Zvereva performed research; J. Seguin, R. Rajeswaran and M. M. Pooggin analyzed the data; and M. M. Pooggin wrote the paper.

LITERATURE CITED

Akbergenov, R., Si-Ammour, A., Blevins, T., Amin, I., Kutter, C., Vanderschuren, H., Zhang, P., Gruissem, W., Meins, F., Jr., Hohn, T., and Pooggin, M. M. 2006. Molecular characterization of geminivirus-derived small RNAs in different plant species. Nucleic Acids Res. 34:462-471.

Aregger, M., Borah, B. K., Seguin, J., Rajeswaran, R., Gubaeva, E. G., Zvereva, A. S., Windels, D., Vazquez, F., Blevins, T., Farinelli, L., and Pooggin, M. M. 2012. Primary and secondary siRNAs in geminivirus-induced gene silencing. PLoS Pathog. 8:e1002941.

Bao, Y., and Hull, R. 1992. Characterization of the discontinuities in rice tungro bacilliform virus DNA. J. Gen. Virol. 73:1297-1301.

Bao, Y., and Hull, R. 1993. Mapping the 5′-terminus of rice tungro bacilliform viral genomic RNA. Virology 197:445-448.

Blevins, T., Rajeswaran, R., Shivaprasad, P. V., Beknazariants, D., Si-Ammour, A., Park, H. S., Vazquez, F., Robertson, D., Meins, F., Jr., Hohn, T., and Pooggin, M. M. 2006. Four plant Dicers mediate viral small RNA biogenesis and DNA virus induced silencing. Nucleic Acids Res. 34:6233-6246.

Blevins, T., Rajeswaran, R., Aregger, M., Borah, B. K., Schepetilnikov, M., Baerlocher, L., Farinelli, L., Meins, F., Jr., Hohn, T., and Pooggin, M. M. 2011. Massive production of small RNAs from a non-coding region of *Cauliflower mosaic virus* in plant defense and viral counter-defense. Nucleic Acids Res. 39:5003-5014.

Borah, B. K., Sharma, S., Kant, R., Johnson, A. M., Saigopal, D. V., and Dasgupta, I. 2013.Bacilliform DNA-containing plant viruses in the tropics: Commonalities within a genetically diverse group. Mol. Plant Pathol. 14:759-771.

Covey, S. N., Al-Kaff, N. S., Lángara, A., and Turner, D. S. 1997. Plants combat infection by gene silencing. Nature 385:781-782.

Cruz, F. C. S., Koganezawa, H., and Hibino, H. 1993. Comparative cytology of rice tungro viruses in selected rice cultivars. J. Phytopathol. 138:274-282.

Dasgupta, I., Hull, R., Eastop, S., Poggi-Pollini, C., Blakebrough, M., Boulton, M. I., and Davies, J. W. 1991. Rice tungro bacilliform virus DNA independently infects rice after *Agrobacterium*-mediated transfer. J. Gen. Virol. 72:1215-1221.

Deleris, A., Gallego-Bartolome, J., Bao, J., Kasschau, K. D., Carrington, J. C., and Voinnet, O. 2006. Hierarchical action and inhibition of plant Dicer-like proteins in antiviral defense. Science 313:68-71.

Dunoyer, P., Schott, G., Himber, C., Meyer, D., Takeda, A., Carrington, J. C., and Voinnet, O. 2010. Small RNA duplexes function as mobile silencing signals between plant cells. Science 328:912-916.

Fütterer, J., Potrykus, I., Valles Brau, M. P., Dasgupta, I., Hull, R., and Hohn, T. 1994. Splicing in a plant pararetrovirus. Virology 198:663-670.

Fütterer, J., Rothnie, H. M., Hohn, T., and Potrykus, I. 1997. Rice tungro bacilliform virus open reading frames II and III are translated from polycistronic pregenomic RNA by leaky scanning. J. Virol. 71:7984-7989.

Garcia-Ruiz, H., Takeda, A., Chapman, E. J., Sullivan, C. M., Fahlgren, N., Brempelis, K. J., and Carrington, J. C. 2010. *Arabidopsis* RNA-dependent RNA polymerases and dicer-like proteins in antiviral defense and small interfering RNA biogenesis during *Turnip mosaic virus* infection. Plant Cell 22:481-496.

Ghildiyal, M., and Zamore, P. D. 2009. Small silencing RNAs: An expanding universe. Nat. Rev. Genet. 10:94-108.

Guerra-Peraza, O., de Tapia, M., Hohn, T., and Hemmings-Mieszczak, M. 2000. Interaction of the *Cauliflower mosaic virus* coat protein with the pregenomic RNA leader. J. Virol. 74:2067-2072.

Guo, H., Song, X., Xie, C., Huo, Y., Zhang, F., Chen, X., Geng, Y., and Fang, R. 2013. Rice yellow stunt rhabdovirus protein 6 suppresses systemic RNA silencing by blocking RDR6-mediated secondary siRNA synthesis. Mol. Plant-Microbe Interact. 26:927-936.

Haas, G., Azevedo, J., Moissiard, G., Geldreich, A., Himber, C., Bureau, M., Fukuhara, T., Keller, M., and Voinnet, O. 2008. Nuclear import of CaMV P6 is required for infection and suppression of the RNA silencing factor DRB4. EMBO (Eur. Mol. Biol. Organ.) J. 27:2102-2112.

Hamilton, A., Voinnet, O., Chappell, L., and Baulcombe, D. 2002. Two classes of short interfering RNA in RNA silencing. EMBO (Eur. Mol. Biol. Organ.) J. 21:4671-4679.

Hay, J. M., Jones, M. C., Blakebrough, M. L., Dasgupta, I., Davies, J. W., and Hull, R. 1991. An analysis of the sequence of an infectious clone of rice tungro bacilliform virus, a plant pararetrovirus. Nucleic Acids Res. 19:2615-2621.

Himber, C., Dunoyer, P., Moissiard, G., Ritzenthaler, C., and Voinnet, O. 2003. Transitivity-dependent and -independent cell-to-cell movement of RNA silencing. EMBO (Eur. Mol. Biol. Organ.) J. 22:4523-4533.

Hohn, T., and Rothnie, H. 2013. Plant pararetroviruses: Replication and expression. Curr. Opin. Virol. 3:621-628.

Hull, R. 1996. Molecular biology of Rice tungro viruses. Annu. Rev. Phytopathol. 34:275-297.

Itaya, A., Zhong, X., Bundschuh, R., Qi, Y., Wang, Y., Takeda, R., Harris, A.R., Molina, C., Nelson, R. S., and Ding, B. 2007. A structured viroid RNA serves as a substrate for dicer-like cleavage to produce biologically active small RNAs but is resistant to RNA-induced silencing complex-mediated degradation. J. Virol. 81:2980-2994.

Jiang, L., Qian, D., Zheng, H., Meng, L. Y., Chen, J., Le, W. J., Zhou, T., Zhou, Y. J., Wei, C. H., and Li, Y. 2012. RNA-dependent RNA polymerase 6 of rice (*Oryza sativa*) plays role in host defense against negative-strand RNA virus, *Rice stripe virus*. Virus Res. 163:512-519.

Joshua-Tor, L., and Hannon, G. J. 2011. Ancestral roles of small RNAs: An Ago-centric perspective. Cold Spring Harb. Perspect. Biol. 3:a003772.

Kawahara, Y., de la Bastide, M., Hamilton, J. P., Kanamori, H., McCombie, W. R., Ouyang, S., Schwartz, D. C., Tanaka, T., Wu, J., Zhou, S., Childs, K. L., Davidson, R. M., Lin, H., Quesada-Ocampo, L., Vaillancourt, B., Sakai, H., Lee, S. S., Kim, J., Numa, H., Itoh, T., Buell, C. R., and Matsumoto, T. 2013. Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. Rice 6:4.

Liu, B., Chen, Z., Song, X., Liu, C., Cui, X., Zhao, X., Fang, J., Xu, W., Zhang, H., Wang, X., Chu, C., Deng, X., Xue, Y., and Cao, X. 2007. *Oryza sativa* dicer-like4 reveals a key role for small interfering RNA silencing in plant development. Plant Cell 19:2705-2718.

Moissiard, G., and Voinnet, O. 2006. RNA silencing of host transcripts by *Cauliflower mosaic virus* requires coordinated action of the four *Arabidopsis* Dicer-like proteins. Proc. Natl. Acad. Sci. U.S.A. 103:19593-19598.

Molnár, A., Csorba, T., Lakatos, L., Várallyay, E., Lacomme, C., and Burgyán, J. 2005. Plant virus-derived small interfering RNAs originate predominantly from highly structured single-stranded viral RNAs. J. Virol. 79:7812-7818.

Molnar, A., Melnyk, C. W., Bassett, A., Hardcastle, T. J., Dunn, R., and Baulcombe, D. C. 2010. Small silencing RNAs in plants are mobile and direct epigenetic modification in recipient cells. Science 328:872-875.

Pooggin, M. M. 2013. How can plant DNA viruses evade siRNA-directed DNA methylation and silencing? Int. J. Mol. Sci. 14:15233-15259.

Pooggin, M. M., Fütterer, J., Skryabin, K. G., and Hohn, T. 1999. A short open reading frame terminating in front of a stable hairpin is the conserved feature in pregenomic RNA leaders of plant pararetroviruses. J. Gen. Virol. 80:2217-2228.

Pooggin, M. M., Ryabova, L. A., He, X., Fütterer, J., and Hohn, T. 2006. Mechanism of ribosome shunting in Rice tungro bacilliform pararetrovirus. RNA 12:841-850.

Pooggin, M. M., Fütterer, J., and Hohn, T. 2008. Cross-species functionality of pararetroviral elements driving ribosome shunting. PLoS One 3:e1650.

Rajeswaran, R., and Pooggin, M. M. 2012a. Role of virus-derived small RNAs in plant antiviral defense: Insights from DNA viruses. Pages 261-289 in: MicroRNAs in Plant development and Stress Response. R. Sunkar, ed. Springer, Berlin.

Rajeswaran, R., and Pooggin, M. M. 2012b. RDR6-mediated synthesis of complementary RNA is terminated by miRNA stably bound to template RNA. Nucleic Acids Res. 40:594-599.

Rajeswaran, R., Aregger, M., Zvereva, A. S., Borah, B. K., Gubaeva, E. G., and Pooggin, M. M. 2012. Sequencing of RDR6-dependent double-stranded RNAs reveals novel features of plant siRNA biogenesis. Nucleic Acids Res. 40:6241-6254.

Rajeswaran, R., Seguin, J., Chabannes, M., Duroy, P. O., Laboureau, N., Farinelli, L., Iskra-Caruana, M. L, and Pooggin, M. M. 2014. Evasion of siRNA-directed antiviral silencing in *Musa acuminata* persistently infected with six distinct banana streak pararetroviruses. J. Virol. 88:11516-11528

Ryabova, L. A., Pooggin, M. M., and Hohn, T. 2006. Translation reinitiation and leaky scanning in plant viruses. Virus Res. 119:52-62.

Schiebel, W., Haas, B., Marinković, S., Klanner, A., and Sänger, H. L. 1993. RNA-directed RNA polymerase from tomato leaves. II. Catalytic in vitro properties. J. Biol. Chem. 268:11858-11867.

Seguin, J., Otten, P., Baerlocher, L., Farinelli, L., and 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., and Pooggin, M. M. 2014b. De novo reconstruction of consensus master genomes of plant RNA and DNA viruses from siRNAs. PLoS One 9:e88513.

Shivaprasad, P. V., Akbergenov, R., Trinks, D., Rajeswaran, R., Veluthambi, K., Hohn, T., and Pooggin M. M. 2005. Promoters, transcripts, and regulatory proteins of Mungbean yellow mosaic geminivirus. J. Virol. 79:8149-8163.

Shivaprasad, P. V., Rajeswaran, R., Blevins, T., Schoelz, J., Meins, F., Jr., Hohn, T., and Pooggin, M. M. 2008. The CaMV transactivator/viroplasmin interferes with RDR6-dependent trans-acting and secondary siRNA pathways in *Arabidopsis*. Nucleic Acids Res. 36:5896-5909.

Song, X., Li, P., Zhai, J., Zhou, M., Ma, L., Liu, B., Jeong, D. H., Nakano, M., Cao, S., Liu, C., Chu, C., Wang, X. J., Green, P. J., Meyers, B. C., and Cao, X. 2012. Roles of DCL4 and DCL3b in rice phased small RNA biogenesis. Plant J. 69:462-474.

Sorefan, K., Pais, H., Hall, A. E., Kozomara, A., Griffiths-Jones, S., Moulton, V., and Dalmay, T. 2012. Reducing ligation bias of small RNAs in libraries for next generation sequencing. Silence 3:4.

Szittya, G., Moxon, S., Pantaleo, V., Toth, G., Rusholme Pilcher, R. L., Moulton, V., Burgyan, J., and Dalmay, T. 2010. Structural and functional analysis of viral siRNAs. PLoS Pathog. 6:e1000838.

Urayama, S., Moriyama, H., Aoki, N., Nakazawa, Y., Okada, R., Kiyota, E., Miki, D., Shimamoto, K., and Fukuhara, T. 2010. Knock-down of OsDCL2 in rice negatively affects maintenance of the endogenous dsRNA virus, *Oryza sativa* endornavirus. Plant Cell Physiol. 51:58-67.

Vargason, J. M., Szittya, G., Burgyán, J., and Hall, T. M. 2003. Size selective recognition of siRNA by an RNA silencing suppressor. Cell 115:799-811.

Wang, X. B., Jovel, J., Udomporn, P., Wang, Y., Wu, Q., Li, W. X., Gasciolli, V., Vaucheret, H., and Ding, S. W. 2011. The 21-nucleotide, but not 22-nucleotide, viral secondary small interfering RNAs direct potent antiviral defense by two cooperative argonautes in *Arabidopsis thaliana*. Plant Cell 23:1625-1638.

Wei, L., Gu, L., Song, X., Cui, X., Lu, Z., Zhou, M., Wang, L., Hu, F., Zhai, J., Meyers, B. C., and Cao, X. 2014. Dicer-like 3 produces transposable element-associated 24-nt siRNAs that control agricultural traits in rice. Proc. Natl. Acad. Sci. U.S.A. 111:3877-3882.

Wu, L., Zhang, Q., Zhou, H., Ni, F., Wu, X., and Qi, Y. 2009. Rice MicroRNA effector complexes and targets. Plant Cell 21:3421-3435.

Wu, L., Zhou, H., Zhang, Q., Zhang, J., Ni, F., Liu, C., and Qi, Y. 2010. DNA methylation mediated by a microRNA pathway. Mol. Cell 38:465-475.

Xu, Y., Huang, L., Fu, S., Wu, J., and Zhou, X. 2012. Population diversity of rice stripe virus-derived siRNAs in three different hosts and RNAi-based antiviral immunity in *Laodelphgax striatellus*. PLoS One 7:e46238.

Yan, F., Zhang, H., Adams, M. J., Yang, J., Peng, J., Antoniw, J. F., Zhou, Y., and Chen, J. 2010. Characterization of siRNAs derived from rice stripe virus in infected rice plants by deep sequencing. Arch Virol. 155:935-940.

Zvereva, A. S., and Pooggin, M. M. 2012. Silencing and innate immunity in plant defense against viral and non-viral pathogens. Viruses 4:2578-2597.

AUTHOR-RECOMMENDED INTERNET RESOURCE

MISIS map tool: www.fasteris.com/apps

Annex: (Seguin et al., 2014a)

De Novo Reconstruction of Consensus Master Genomes of Plant RNA and DNA Viruses from siRNAs

Jonathan Seguin, Rajendran Rajeswaran, Nachelli Malpica-Lopez, Robert R. Martin, Kristin Kasschau, Valerian V. Dolja, Patricia Otten, Laurent Farinelli, Mikhail M. Pooggin

PLOS One (2014), Vol.9, Issue 2

OPEN ACCESS Freely available online

PLOS ONE

De Novo Reconstruction of Consensus Master Genomes of Plant RNA and DNA Viruses from siRNAs

Jonathan Seguin[1,4], Rajendran Rajeswaran[1], Nachelli Malpica-López[1], Robert R. Martin[2,3], Kristin Kasschau[3], Valerian V. Dolja[3], Patricia Otten[4], Laurent Farinelli[4], Mikhail M. Pooggin[1]*

1 University of Basel, Department of Environmental Sciences, Institute of Botany, Basel, Switzerland, **2** United States Department of Agriculture–Agricultural Research Service, Horticultural Crops Research Laboratory, Corvallis, Oregon, United States of America, **3** Oregon State University, Department of Botany and Plant Pathology, Center for Genome Research and Biocomputing, Corvallis, Oregon, United States of America, **4** Fasteris SA, Plan-les-Ouates, Geneva, Switzerland

Abstract

Virus-infected plants accumulate abundant, 21–24 nucleotide viral siRNAs which are generated by the evolutionary conserved RNA interference (RNAi) machinery that regulates gene expression and defends against invasive nucleic acids. Here we show that, similar to RNA viruses, the entire genome sequences of DNA viruses are densely covered with siRNAs in both sense and antisense orientations. This implies pervasive transcription of both coding and non-coding viral DNA in the nucleus, which generates double-stranded RNA precursors of viral siRNAs. Consistent with our finding and hypothesis, we demonstrate that the complete genomes of DNA viruses from *Caulimoviridae* and *Geminiviridae* families can be reconstructed by deep sequencing and *de novo* assembly of viral siRNAs using bioinformatics tools. Furthermore, we prove that this 'siRNA omics' approach can be used for reliable identification of the consensus master genome and its microvariants in viral quasispecies. Finally, we utilized this approach to reconstruct an emerging DNA virus and two viroids associated with economically-important red blotch disease of grapevine, and to rapidly generate a biologically-active clone representing the wild type master genome of *Oilseed rape mosaic virus*. Our findings show that deep siRNA sequencing allows for *de novo* reconstruction of any DNA or RNA virus genome and its microvariants, making it suitable for universal characterization of evolving viral quasispecies as well as for studying the mechanisms of siRNA biogenesis and RNAi-based antiviral defense.

Citation: Seguin J, Rajeswaran R, Malpica-López N, Martin RR, Kasschau K, et al. (2014) *De Novo* Reconstruction of Consensus Master Genomes of Plant RNA and DNA Viruses from siRNAs. PLoS ONE 9(2): e88513. doi:10.1371/journal.pone.0088513

Editor: Hanu Pappu, Washington State University, United States of America

Received October 4, 2013; **Accepted** January 6, 2014; **Published** February 11, 2014

Funding: The work was supported by European Cooperation in Science and Technology [grant SERI No. C09.0176 to L.F. and M.M.P.]; Swiss National Science Foundation [grant 31003A_143882/1 to M.M.P.]; Vinoculate, Inc. [contract 2010-744 to V.V.D.], USDA-NIFA [subcontract 2009-04401 to V.V.D.]; USDA-NIFA-SCRI [2009-51181-06027 subaward to R.R.M.]; and Bard [award IS-4314-10C to V.V.D.]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The work in VVD lab is partially supported by a contract with Vinoculate, Inc. that also holds exclusive rights to an OSU patent "Closterovirus Vectors and Methods". No. 8,415,147 Issued April 9, 2013; OSU Ref. No. 06-57; Klarquist Ref. No. 245-79793-10. Jonathan Seguin, Patricia Otten and Laurent Farinelli are employed by Fasteris SA. There are no further patents, products in development or marketed products to declare. This does not alter the authors' adherence to all the PLOS ONE policies on sharing data and materials, as detailed online in the guide for authors.

* E-mail: Mikhail.Pooggin@unibas.ch

Introduction

Owing to error-prone replication, viruses accumulate microvariants which deviate from a consensus master genome by one or more SNPs (single-nucleotide polymorphisms) and/or indels (insertions/deletions) and comprise a viral quasispecies that can rapidly evolve in changing environment [1]. Resistance-breaking strains often emerge from such microvariants and from recombination events involving distinct viral strains or viruses. Existing methods of viral diagnostics using antibodies and PCR often fail to identify new pathogenic strains and are not applicable for emerging viruses with unknown genomes. Therefore, next generation deep sequencing approaches and *de novo* assembly of virus genomes from sequencing reads hold a great promise for universal diagnostics of viral pathogens and reliable characterization of causative agent(s) of any given disease [2,3]. In a pioneering work, Kreuze *et al.* [2] have demonstrated that a complete genome of a known plant RNA virus can be reconstructed *de novo* from multiple contigs of short interfering RNAs (siRNAs) which are generated in infected plants by the evolutionarily conserved RNA silencing/RNA interference (RNAi) machinery [4–6]. This and the follow-up studies have proven that deep siRNA sequencing and bioinformatics are applicable for identification and at least partial genome reconstruction of plant viruses and viroids [7–14] as well as insect viruses [15,16]. Here we extend these findings by demonstrating that the complete genomes of plant DNA viruses of two major families – *Caulimoviridae* and *Geminiviridae* – can be reconstructed without a reference genome as a single contig or a few overlapping contigs of viral siRNAs. Furthermore, we show that bioinformatics analysis of viral siRNA population allows for the identification of the master genome and its microvariants in viral quasispecies. We also used this technology to reconstruct a newly emerged single-stranded DNA virus and two viroids associated with the red blotch disease of grapevines in the United States. Thus, deep siRNA sequencing can be used for identification and reconstruction of the consensus master genome of any plant virus or viroid, and for studying virus diversity and evolution. Moreover, our analysis of siRNAs derived from DNA viruses and viroids contributes to further understanding the mechanisms of

siRNA biogenesis and RNAi-based antiviral defense and raises new questions for future research.

Results and Discussion

Growing evidence indicates that 21–24 nt virus-derived siRNAs are produced from double-stranded (ds) RNA precursors covering the entire viral genome sequences [4,17,18]. Accordingly, complete or near complete genomes of new RNA virus strains have been reconstructed using a reference strain sequence as a scaffold for assembly of overlapping siRNA contigs which were generated by the short sequence read assembler Velvet [2,7,8,10,11–13]. Similar approaches have succeeded in detection and partial genome reconstruction of plant DNA viruses [2,9,10,19]. Because DNA viruses do not replicate through dsRNA intermediates, viral siRNA precursors are likely produced by the host Polymerase II-mediated, bidirectional transcription of a circular viral DNA beyond the poly(A) sites, thereby including both coding (mRNA) and non-coding (promoter) sequences [4,17,18]. To validate this hypothesis and test if complete genomes of DNA viruses could also be reconstructed as single contigs of viral siRNAs we used the small RNA (sRNA) deep-sequencing libraries obtained from *Arabidopsis* plants infected with *Cauliflower mosaic virus* (CaMV) [18] and *Cabbage leaf curl virus* (CaLCuV) [17] (Datasets S1*A* and S1*B*), which represent the *Caulimoviridae* and the *Geminiviridae* families, respectively. Bioinformatic analysis of the redundant sRNA reads revealed that the hotspots of viral siRNA production cover only portions of the viral genome in both cases. The non-redundant reads, however, cover the entire circular genomes of CaMV and CaLCuV in the sense and antisense orientations without gaps (Figure 1; Dataset S2), albeit the density of non-redundant reads is somewhat higher in the hotspot regions (Dataset S2). This suggests that both coding and non-coding regions of circular viral dsDNA are transcribed in both orientations to generate dsRNA precursors of viral siRNAs. Since genetic evidence revealed that the silencing-related, DNA-dependent RNA polymerases Pol IV and Pol V, or RNA-dependent RNA polymerases RDR1, RDR2 and RDR6 are not required for the biogenesis of CaMV- and CaLCuV-derived siRNAs [17,18], these dsRNA precursors are likely generated by Pol II-mediated sense and antisense transcription. However, potential involvement of RDR3, RDR4 or RDR5 in viral siRNA biogenesis was not ruled out yet. In conclusion, similar to RNA viruses, the entire genomes of DNA viruses from the families *Caulimoviridae* and *Geminiviridae* are densely covered with non-redundant viral siRNA species and therefore can potentially be reconstructed as single contigs of the viral siRNAs.

To *de novo* assemble viral siRNAs, we tested different algorithms using Velvet followed by Oases or Metavelvet for assembling redundant or non-redundant reads and Seqman for merging the resulting contigs. In some cases, we also used mapping to the plant genome as a filtering step before Seqman to separate the viral siRNA contigs from the plant sRNA contigs (Figure 2). As a result, with both Oases and Metavelvet, the complete 8,031 nt genome of CaMV was reconstructed as a single terminally-redundant contig (Figure 1). The bipartite genome of CaLCuV was assembled as one terminally-redundant contig covering 2,512 nt DNA-B and two contigs covering 2,583 nt DNA-A (Figure 1). In the latter case, the filtering step was required. Because DNA-A and DNA-B of CaLCuV share a near identical common region of 195 nts (with 7 SNPs), during *de novo* assembly the DNA-A siRNA contig gets split in two contigs within this region. Generally, Oases generated longer contigs, while Metavelvet more precise contigs. Non-redundant reads assembled in longer contigs. SNPs and short indels that occurred in some of the contigs could be identified and corrected by SNP calling with redundant reads (see Materials and Methods for further details of the bioinformatics analysis). Thus, using Oases or Metavelvet followed by Seqman, the complete viral genomes were assembled *de novo* as single contigs of non-redundant siRNAs. This is unlike most of the above mentioned reports of virus or viroid reconstruction, in which multiple contigs of redundant siRNAs generated by Velvet were assembled using a reference genome as a scaffold. Furthermore, we found that the filtering step before Velvet, which was applied in some previous studies in efforts to remove host small RNAs interfering with assembly of viral siRNAs, often generates gaps in viral siRNA contigs. In contrast, the filtering step before Seqman used in our study enables reconstruction of complete genomes, especially in the case of DNA viruses. Moreover, we found that SNP calling with redundant siRNA reads can be applied to correct potential errors in *de novo* assembly algorithms.

To determine if such 'siRNA omics' (siRomics) approach is applicable for identification of a master genome in viral 'quasispecies cloud', we sequenced sRNAs from *Arabidopsis* infected with *Oilseed rape mosaic virus* (ORMV), the RNA virus for which an available cDNA clone was not infectious [20] because of potential cloning errors or because it represented a defective microvariant from the ORMV quasispecies. Using Oases followed by Seqman, the 6,303 nt ORMV genome was reconstructed *de novo* as a single contig from two independent sRNA libraries (Dataset S1*C* and Dataset S2). This reconstructed genome differed from the available cDNA sequence at three positions (G-to-A at position 12, A-to-T at position 231, and G-to-T at position 5612; Figure 1). SNP calling using the two sRNA libraries confirmed these mismatches in 96.5–97.7% (A12), 99.5–99.9% (T231) and 88.6–93.8% (T5612) viral redundant reads and highlighted the overall variation in the ORMV quasispecies (Dataset S3*A*). We corrected these mismatches (presumably cloning errors) in the cDNA clone and tested it for infectivity. Strikingly, the resulting clone was fully biologically active, causing the disease symptoms indistinguishable from those of the wild type ORMV sap (Figure 3). Thus, a common problem of virology, often taking years to overcome [21], was solved in one step. Our unpublished results suggest that G at position 12 in the original cDNA clone had a drastic impact on ORMV infectivity, possibly because it affects initiation of positive stand synthesis during the viral replication process; the nucleotide substitutions at the positions 231 and 5612 likely represent viable variants in the virus quasispecies (N.M.L., R.R., and M.M.P., in preparation).

Interestingly, SNP calling revealed that ORMV, CaMV and CaLCuV do not differ drastically in the frequency of SNPs or the average degree of deviation (in %) from the master genome nucleotides (Dataset S3*A-D*). This implies that distinct replication mechanisms of these viruses, involving the viral RDR (ORMV), the viral reverse transcriptase (CaMV), or the host DNA polymerase (CaLCuV), may have a comparable error rate. The error rate of the host DNA polymerase possessing proof-reading activity might be as high as the error rates of the viral RDR and reverse transcriptase lacking any proof-reading activity, because it replicates viral DNA via a rolling circle mechanism involving a viral Rep protein (recently reviewed by Pooggin [22]). Alternatively, comparable accumulation of the microvariants in all the three viruses may reflect their adaptation to the experimental host plant *Arabidopsis thaliana*. Note, that for identification of the SNPs listed in Dataset S3, we used a quite conservative cut-off, 10% non-redundant reads, to account for both an error of Illumina sequencing of sRNAs (0.1–0.5%) and sequence-specific biases that may lead to overrepresentation of certain sRNAs in redundant

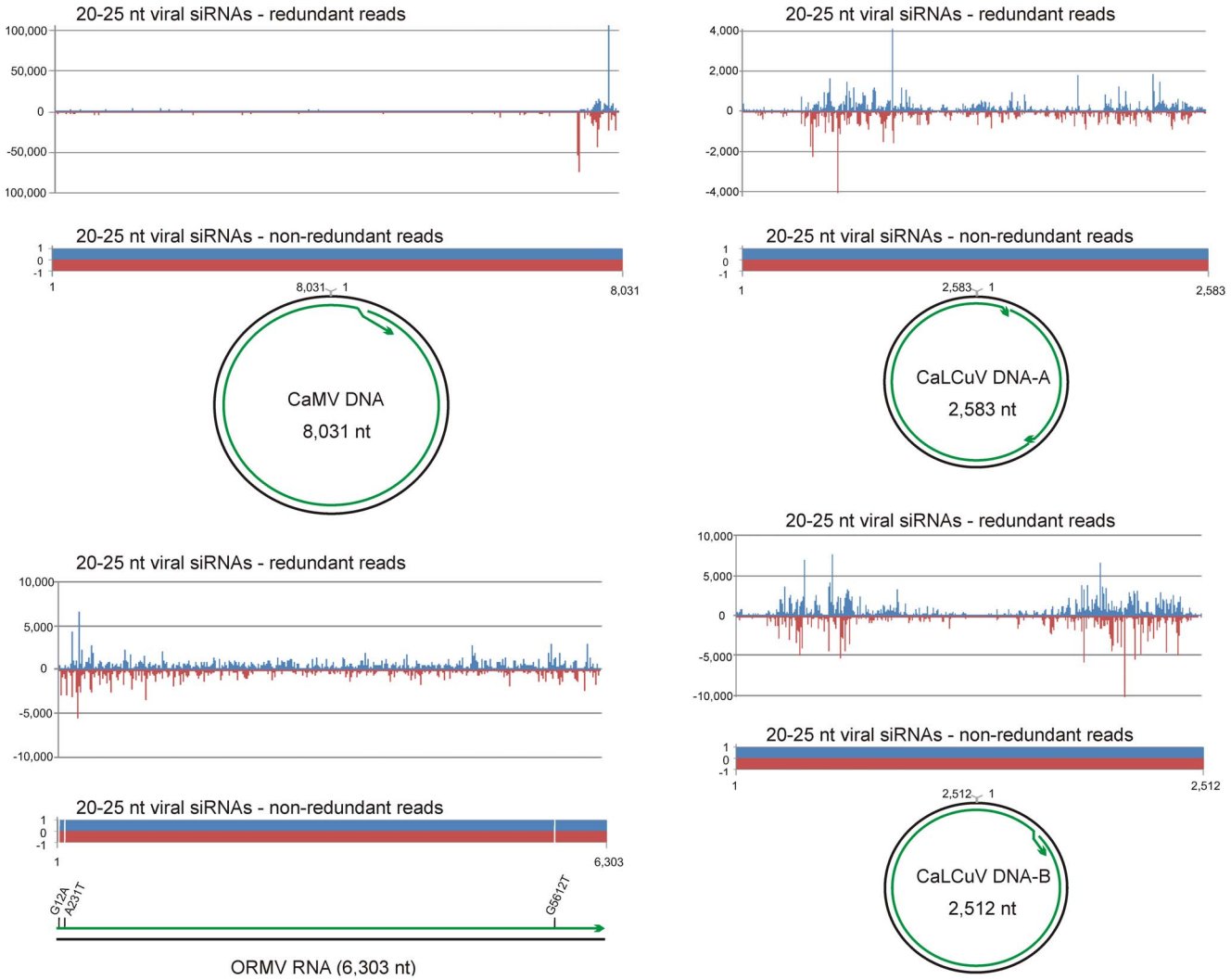


Figure 1. Maps of viral siRNAs and their contigs. The graphs plot the number of 20–25 nt viral siRNA reads (redundant and non-redundant) at each nucleotide position of the genomes of CaMV, ORMV and CaLCuV (DNA-A and DNA-B); Bars above the axis represent sense reads starting at respective positions; those below the antisense reads ending at respective positions. Circular DNA genomes of CaMV and CaLCuV and linear RNA genome of ORMV are shown below the graphs, with the siRNA contigs covering the genomes depicted as green lines with arrowheads. Mismatches between the ORMV contig and the reference genome are indicated.
doi:10.1371/journal.pone.0088513.g001

reads. Since host RDR activity may amplify CaLCuV siRNAs and thereby contribute to the observed deviations from the CaLCuV master genome, we compared the viral microvariant accumulation in CaLCuV-infected wild-type plants and *rdr1/2/6* triple mutant plants with diminished RDR activities [17]: no drastic difference was observed in the frequency of SNPs or the average degree of deviation from the master genome nucleotides (Dataset S3*C-D*). Our findings for CaLCuV are consistent with the observations that geminiviruses have high mutation frequency and evolve as fast as RNA viruses (see [23] and references therein).

To evaluate the potential of siRomics for diagnostics of an unknown disease, we deep sequenced sRNAs from grapevines (*Vitis vinifera* cv. Pinot noir) grown at vineyards in Oregon, some of which exhibited severe leaf red blotch disease symptoms of unknown etiology, and from control plants with green, healthy-looking leaves. *De novo* reconstruction revealed that both infected and healthy-looking vines harbored *Grapevine yellow speckle viroid 1* (GYSVd-1) and *Hop stunt viroid* (HSVd), whose small circular RNA genomes were assembled as single terminally-redundant contigs and verified by redundant sRNA coverage (Figure 4; Dataset S1*D* and Dataset S2). Previously, these two viroids have been identified in grapevines in Oregon and elsewhere (see Materials and Methods). In addition, the disease-affected vines harbored a virus with a 3,206 nt circular genome reconstructed *de novo* from four contigs and validated by redundant sRNA coverage (Figure 4A; Dataset S1*D* and Dataset S2). Phylogenomic analysis showed that the virus is distantly related to circular DNA genomes of plant geminiviruses. The only closely related genome at the time of our analysis was that of a recently identified DNA virus from a declining Cabernet franc vineyard in New York State [24] (named 'grapevine geminivirus' (GVGV); the NCBI Genbank accession NC_017918). Despite their occurrence across the continent in distinct grapevine cultivars, the New York and Oregonian isolates differed by only 11 SNPs (Dataset S3*E*). Notably, all these 11 nucleotides in the Oregonian isolate are supported by at least 90% redundant reads in three independent red blotch leaf samples and therefore represent quite stable nucleotide positions in the master genome, unlike some of the other positions identified by SNP

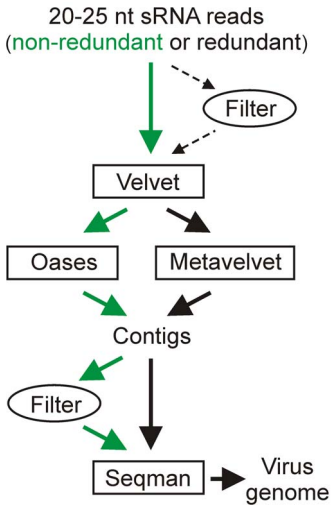


Figure 2. Bioinformatics algorithms for *de novo* reconstruction of viral/viroid genomes from siRNAs. The *de novo* assembler programs are boxed. Green arrows indicate the algorithm which in many cases generated the longest contigs.
doi:10.1371/journal.pone.0088513.g002

calling (Dataset S3*B*). We designed PCR primers specific to the virus (5′-TGCAAGTGGACATACGTTTA and 5′-GGGATCC-CATCAATTGTTCT) and confirmed its presence in DNA samples from 12 of 16 symptomatic vines from the same vineyard, but not from any of 18 symptomless vines. Intriguingly, the most recent reports describe *Grapevine red blotch-associated virus* (GRBaV) and *Grapevine red leaf-associated virus* (GRLaV) that severely affect vineyards in States of California [25] and Washington [26], respectively. Both GRBaV and GRLaV sequences were found to be very similar to the geminivirus from New York [25,26]. Thus, the virus that we identified in Oregon appears to be geographically wide-spread and associated with the emerging disease that threatens the high cash-value crop, grapevine. Whether this virus causes red blotch disease alone or in complex with the viroids identified in our study remains to be investigated. Interestingly, GRLaV was found to be associated with two RNA viruses and four viroids including HSVd and GYSVd-1 [26]. Thus, both or one of the latter two viroids may contribute to the disease.

Unlike most RNA viruses that generate predominantly 21-nt siRNAs (e.g. ORMV [4]; Dataset S1*C*), DNA viruses that are transcribed in the nucleus spawn 21-, 22- and 24-nt siRNAs (e.g. CaMV and CaLCuV; Dataset S1*A* and Dataset S1*B*), which are processed by distinct Dicer-like (DCL) enzymes from long dsRNA precursors [4,17,18]. We found that the grapevine geminivirus also induces formation of predominantly 21-, 22- and 24-nt siRNAs (Dataset S1*D*), thus resembling other geminiviruses. Furthermore, both coding and non-coding regions of this virus are covered with viral siRNA species in both orientations without gaps (Figure 4; Dataset S2). This suggests that bi-directional readthrough transcription of circular viral DNA may generate dsRNA precursors covering the entire virus genome, like it was proposed in the case of CaLCuV [17]. Analysis of 5′-nucleotides of the grapevine geminivirus-derived siRNAs and the host sRNAs (Dataset S1*D*) revealed similar biases to 5′U in 21-nt sRNAs and 5′A in 24-nt sRNAs. This indicates that, similar to plant miRNAs and siRNAs [27,28], viral siRNAs are also sorted by Argonaute (AGO) proteins based on 5′-nucleotide identity to form silencing complexes. Bioinformatics analysis of siRNAs derived from the grapevine viroids HSVd and GYSVd-1 (Dataset S1*C*) revealed

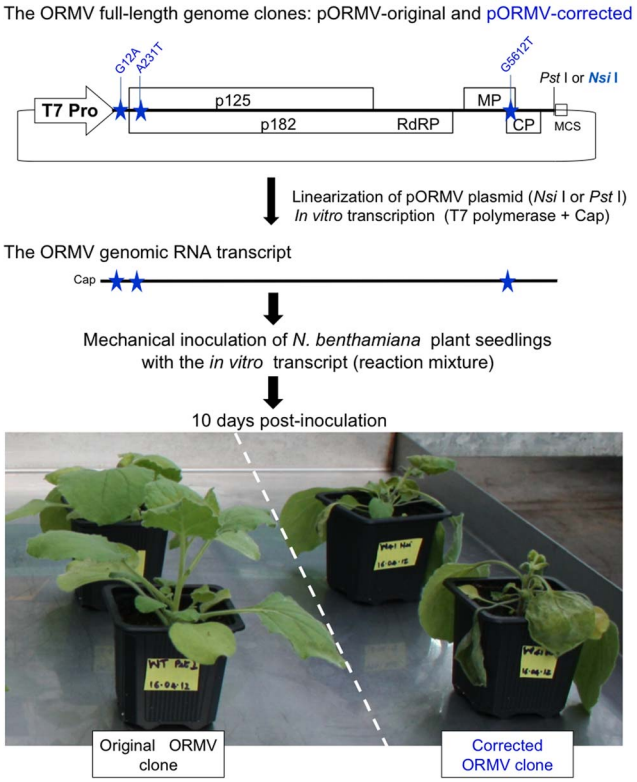


Figure 3. Test of the original and the corrected ORMV clones for infectivity. The plasmid containing the full-length ORMV genome sequence (original or corrected) behind the T7 promoter is depicted schematically: the restriction site *Pst* I or *Nsi* I, respectively, just downstream of the genome (located in multiple cloning site; MCS) was used for linearization of the plasmid, followed by run-off transcription by T7 polymerase in the presence of a cap analog. The resulting *in vitro* transcript (ORMV genomic RNA) was taken for mechanical inoculation of *N. benthamiana* plants. The picture shows the inoculated plants at 10 days post-inoculation.
doi:10.1371/journal.pone.0088513.g003

that these nucleus-localized viroids produce predominantly 21-, 22- and 24-nt siRNAs reminiscent to those produced by DNA viruses transcribed in the nucleus. In further similarity to DNA viruses, siRNA species of each size-class and polarity densely cover the entire circular RNA genomes of these viroids (Dataset S2; Figure 4). This implies that viroid siRNAs are processed from dsRNA intermediates of Pol II-mediated replication of circular single-stranded viroid RNA [29]. The hotspots of viroid/virus siRNA production that map to similar locations for each siRNA size-class and polarity (Dataset S2) may result from preferential internal processing of the dsRNA precursors by distinct DCLs and/or stabilization of siRNAs with certain nucleotide compositions by AGO proteins.

In summary, using deep siRNA sequencing and bioinformatics, dubbed siRomics, we developed a pipeline for nucleotide sequencing of the entire genomes of DNA and RNA viruses or viroids and for identification of the consensus master genome and its microvariants in viral or viroid quasispecies. Moreover, we demonstrated utility of siRomics for the universal diagnostics of known and emerging viral diseases, as well as for rapid generation of the biologically-active clones of problematic viruses. In addition to highlighting the potential of siRomics for applied research, our findings contribute to further understanding the siRNA silencing mechanisms targeting DNA viruses as well as RNA viruses and viroids.

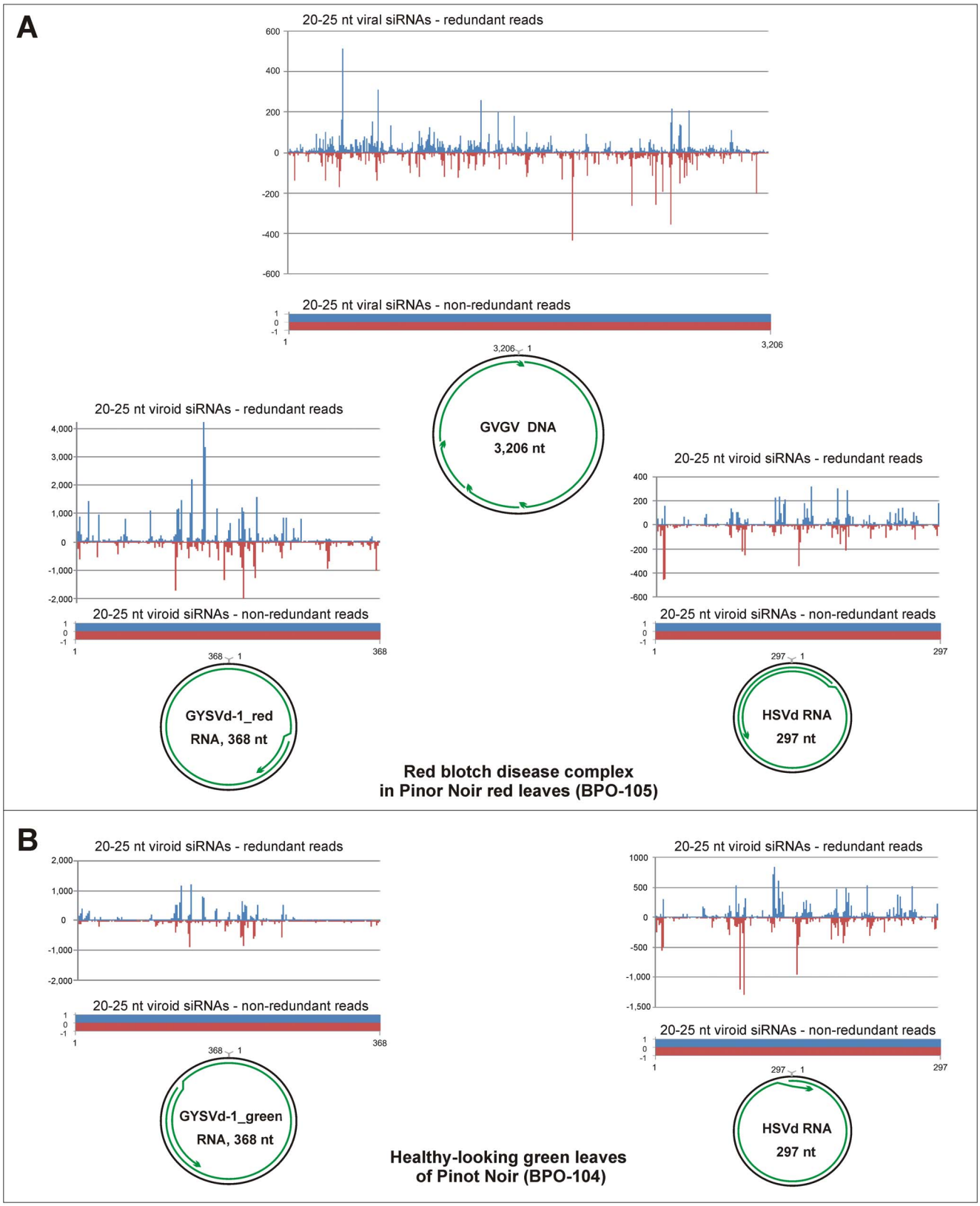


Figure 4. Maps of viral and viroid siRNAs and their contigs from red blotch disease-infected and healthy-looking leaves of grapevine. (**A**) Red blotch disease-infected leaves. (**B**) Healthy-looking green leaves. The graphs plot the number of 20–25 nt viral or viroid siRNA reads (redundant and non-redundant) at each nucleotide position of the genomes of the grapevine geminivirus (GVGV; also named GRBaV and GRLaV) and the viroids HSVd, GYSVd1_red and GYSVd1_green; Bars above the axis represent sense reads starting at respective positions; those below represent antisense reads ending at respective positions. The circular DNA genome of GVGV and the circular RNA genomes of HSVd, GYSVd1_red and

GYSVd1_green are shown below the graphs, with the siRNA contigs covering the genomes depicted as green lines with arrowheads.
doi:10.1371/journal.pone.0088513.g004

Materials and Methods

Plants and viruses

Growth conditions and virus infections of *Arabidopsis thaliana* Col-0 plants were described in detail previously [4]. Briefly, seedlings were infected either by biolistic inoculation with DNA clones of *Cauliflower mosaic virus* (CaMV; the NCBI Genbank accession V00140) and *Cabbage leaf curl virus* (CaLCuV; U65529.2 for DNA-A and U65530.2 for DNA-B) or by mechanical inoculation with sap from *Oilseed rape mosaic virus* (ORMV)-infected *Nicotiana benthamiana*. A previously constructed plasmid containing ORMV cDNA downstream of the T7 promoter (kindly provided by Dr. Fernando Ponz) was modified using synthetic DNA fragments and suitable restriction sites to correct the cloning errors and obtain the reconstructed wild type ORMV genome clone (deposited to the Genbank as KF137561). The resulting and the original clones were linearized downstream of the ORMV sequence and used as templates for *in vitro* transcription reactions (MEGAscript T7 kit, Ambion) to produce a capped viral genomic RNA. The reaction mixtures were used for mechanical inoculation. Symptom development at day 10 post-inoculation is shown in Figure 3.

Samples of the red leaves displaying leafroll-like disease symptoms (named 'red blotch' disease) and healthy-looking green leaves of grapevine cv. Pinot noir plants were collected in a privately owned vineyard near Newberg, Oregon, USA in summer, 2011. The samples were scion clone 777 grafted onto rootstock 44–53 and collected with permission of the owner.

Deep sequencing and bioinformatics analysis of viral/viroid siRNAs

Total RNA from infected and control tissue samples was extracted with Trizol and used for Illumina sequencing of 19–30 nt RNAs as described for CaLCuV by Aregger *et al.* [17]. The resulting small RNA (sRNA) libraries (detailed in Dataset S1) were taken for bioinformatics analysis and for *de novo* reconstruction of the viral genomes using the algorithms summarized in Figure 2. The results of bioinformatics analysis of the viral and host sRNA populations are summarized in Datasets S1, S2 and S3. To reconstruct viral and viroid genomes, the non-redundant or redundant sRNA reads ranging from 20 to 25 nts were assembled into contigs using Velvet 1.2.07 (www.ebi.ac.uk/~zerbino/velvet) [30] followed by Oases 0.2.08 (www.ebi.ac.uk/~zerbino/oases) [31] or Metavelvet 1.2.01 (metavelvet.dna.bio.keio.ac.jp) [32]. Number and size of the resulting contigs varied depending on the choice of Velvet *k*-mer values (13 through 23). 100% coverage of a virus genome could be achieved either with single *k*-mers or certain combinations thereof, as exemplified for CaMV, CaLCuV and ORMV in Dataset S4. SNP calling and correction of errors in viral contigs/genomes was done using Integrative Genomics Viewer (IGV; www.broadinstitute.org/igv) [33]. Oases and Metavelvet contigs obtained with all *k*-mer values or their selected combinations were merged using the Seqman module of Lasergene DNASTAR 8.1.2 Core Suite (DNAStar, Madison, WI). If required, the filtering step before Seqman (or Velvet) was done by mapping contigs (or sRNA sets) to the *Arabidopsis thaliana* genome (TAIR9) or *Vitis vinifera* genome (PRJNA33471) using Burrows-Wheeler Aligner (BWA) 0.5.9 [34]. The *de novo* reconstructed viral genomes were scanned for SNPs and indels using IGV with redundant reads. Finally, single-base resolution maps of viral sRNAs on the virus genomes were created using BWA and a sRNA map tool MISIS (www.fasteris.com/apps; [35]). Reads mapping to several positions on the reference sequence were attributed at random to one of them. To account for a circular virus/viroid genome the first 25 bases of the genome sequence were added to its 3′-end. For each reference genome and each sRNA size (20 to 25 nt), MISIS counted total number of reads, reads in forward and reverse orientation (Dataset S1) and thus generated single-base resolution maps (Dataset S2), where for each position starting from the 5′ end of the reference genome, the number of matches starting at this position in forward (first base of the read) and reverse (last base of the read) orientation for each read length is given. The reads mapped to the last 25 bases of the extended genome sequence were added to the reads mapped to the first 25 bases. MISIS generated two types of counts tables, one with zero mismatches and another with up to two mismatches. Comparison of the two tables was informative for identification and correction of potential mismatches between a reference sequence and the master genome sequence as well as for initial identification of SNPs and short indels in viral quasispecies (see Dataset S2, for all the viruses and viroids analyzed in this study). The positions of SNPs and the degree of deviation (in %) from the master genome nucleotide at each position in viral and viroid quasispecies were identified by IGV analysis of redundant and non-redundant sRNAs mapped to the reference genome with up to 2 mismatches. For identification of the SNPs listed in Dataset S3, we set an arbitrary cutoff value of 10% non-redundant reads: in other words, ten or more percent of the reads support the deviation from the master genome nucleotide for each SNP.

The analysis of siRNAs and complete genome contigs revealed that the infectious DNA-B clone of CaLCuV differs from its reference sequence U65530.2 by a single nucleotide deletion at the last position of the reference (making the genome 1 nt shorter), the infectious clone of CaMV differs from its reference sequence V00140 by two substitutions (C6175A and T6281C), while the original non-infectious ORMV clone differs from its reference sequence (NC_004422; named *Youcai mosaic virus*) at several positions. These apparent sequencing errors were confirmed by re-sequencing of the three clones. The original ORMV clone (confirmed by re-sequencing) differs at the three positions (Dataset S3*A*) from the reconstructed wild-type ORMV genome (KF137561) described in this study. *Hop stunt viroid* (HSVd; deposited to the Genbank as KF137565), which we reconstructed from each of the two green and three red leaf samples of grapevine cv. Pinot noir (Dataset S1*D*), is 100% identical to the sequences of other HSVd isolates, e.g. from grapevine cultivars Lumunage and Thompson Seedless in China (DQ371455, DQ371459) and a citrus tree in Tunisia (GU825977). In addition, the green leaves contained a variant of this viroid with the two SNPs supported by ca. 60% reads (Dataset S3*H*). *Grapevine yellow speckle viroid 1* (GYSVd-1) reconstructed from the two green leaf samples (GYSVd1_green; deposited to the Genbank as KF137564) is most closely related to the GYSVd-1 isolate from Germany (X87911; two SNPs), while GYSVd-1 reconstructed from the three red leaf samples (GYSVd1_red; deposited to the Genbank as KF137563) to the GYSVd-1 isolate from Japan (AB028466; two SNPs). The genome sequence of grapevine geminivirus (GVGV) reconstructed from the red leaves was deposited to the Genbank as KF137562.

Acknowledgments

We thank Thomas Boller for supporting the research of M.M.P. group at the Botanical Institute, and Fernando Ponz and Manfred Heinlein for providing ORMV materials.

Supporting Information

Dataset S1 Counts of viral and endogenous sRNAs in the sRNA deep-sequencing libraries from mock-inoculated and CaMV-infected *Arabidopsis* (Table S1A), CaLCuV-infected *Arabidopsis* (Table S1B), ORMV-infected *Arabidopsis* (Table S1C), and healthy-looking green and red blotch disease-infected leaves of grapevine cv. Pinot noir plants (Table S1D).
(XLSX)

Dataset S2 MISIS-generated, single-base resolution maps of 20–25 nt viral siRNAs from CaMV (BPO-20, BPO-22)-, CaLCuV (BPO-57)- and ORMV (BPO-38, BPO-44)- infected *Arabidopsis* plants and of 20–25 nt viral (GVGV) and viroid (HSVd, GYSVd1_red, GYSVd1_green) siRNAs from red blotch disease-infected red leaves (BPO-105) and healthy-looking green leaves (BPO-104) of grapevine cv. Pinot noir plants. The numbering of nucleotide positions are according to the NCBI Genbank reverence sequences of CaMV (V00140; note that two corrections C6175A and T6281C were introduced in this sequence based on the sRNA and DNA sequencing), CaLCV DNA-A (U65529.2), CaLCuV DNA-B (U65530.2; the last nucleotide of this reference sequence, position 2513, was deleted based on the sRNA and DNA sequencing), ORMV (KF137561), GVGV (KF137562), HSVd (KF137565), GYSVd1_green (KF137564) and GYSV-d1_red (KF137563). Note that the positions of 5′-terminal nucleotide of sense sRNAs and 3′-terminal nucleotide of antisense sRNAs along the reference sequence are given, and the read counts are given for each sRNA of 20-, 21-, 22-, 23-, 24- and 25-nt classes mapped to the forward strand (X20, X21, X22, X23, X24, X25) and the reverse strand (X20_rev, X21_rev, X22_rev, X23_rev, X24_rev, X25_rev) with zero mismatches, along with the total counts of 20–25 nt sRNAs mapped on the forward (total_forward) and reverse (total_reverse) strands and on both strands (total). The last column contains the total number of 20–25 nt sRNA mapped to the reference sequence with up to two mismatches.
(XLSX)

Dataset S3 SNPs in viral and viroid quasispecies. **Table S3A**: SNPs at positions of the mismatches between the reconstructed wild-type ORMV genome and the original ORMV genome clone as well as SNPs in the wild type ORMV viral quasispecies; **Table S3B**: SNPs in CaMV; **Table S3C**: SNPs in CaLCuV DNA-A; **Table S3D**: SNPs in CaLCuV DNA-B; **Table S3E**: SNPs at the positions of the mismatches between the GVGV-Oregon and the GVGV-New York genomes as well as other SNPs in the GVGV-Oregon quasispecies; **Table S3F**: SNPs in the GYSVd-1 (red) and green) quasispecies; **Table S3G**: SNPs in the GYSVd-1 (green) quasispecies; **Table S3H**: SNPs in the HSVd quasispecies.
(XLSX)

Dataset S4 Analysis of the contigs generated by Velvet and Oases using non-redundant 20–25 nt sRNA libraries from CaMV (BPO-20 and BPO-21)-, CaLCuV (BPO-57)- and ORMV (BPO38 and BPO44)-infected *Arabidopsis*. Coverage (in %) of the viral genome sequences with the siRNA contigs is calculated for single k-mer values and combination thereof.
(XLSX)

Author Contributions

Conceived and designed the experiments: MMP LF VVD. Performed the experiments: JS RR NML RRM KK PO. Analyzed the data: MMP JS PO VVD. Wrote the paper: MMP VVD.

References

1. Domingo E, Sheldon J, Perales C (2012) Viral quasispecies evolution. Microbiol Mol Biol Rev 76: 159–216.
2. Kreuze JF, Perez A, Untiveros M, Quispe D, Fuentes S, 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.
3. Hagen C, Frizzi A, Gabriels S, Huang M, Salati R, et al. (2012) Accurate and sensitive diagnosis of geminiviruses through enrichment, high-throughput sequencing and automated sequence identification. Arch Virol. 157: 907–15.
4. Blevins T, Rajeswaran R, Shivaprasad PV, Beknazariants D, Si-Ammour A, et al. (2006) Four plant Dicers mediate viral small RNA biogenesis and DNA virus induced silencing. Nucleic Acids Res 34: 6233–6246.
5. Donaire L, Wang Y, Gonzalez-Ibeas D, Mayer KF, Aranda MA, et al. (2009) Deep-sequencing of plant viral small RNAs reveals effective and widespread targeting of viral genomes. Virology 392: 203–214.
6. Llave C (2010) Virus-derived small interfering RNAs at the core of plant-virus interactions. Trends Plant Sci 15: 701–707.
7. Al Rwahnih M, Daubert S, Golino D, Rowhani A (2009) Deep sequencing analysis of RNAs from a grapevine showing Syrah decline symptoms reveals a multiple virus infection that includes a novel virus. Virology 387: 395–401.
8. Cuellar WJ, Cruzado RK, Fuentes S, Untiveros M, Soto M, et al. (2011) Sequence characterization of a Peruvian isolate of Sweet potato chlorotic stunt virus: further variability and a model for p22 acquisition. Virus Res 157: 111–115.
9. Zhang Y, Singh K, Kaur R, Qiu W (2011) Association of a novel DNA virus with the grapevine vein-clearing and vine decline syndrome. Phytopathology 101: 1081–90.
10. Hagen C, Frizzi A, Kao J, Jia L, Huang M, et al. (2011) Using small RNA sequences to diagnose, sequence, and investigate the infectivity characteristics of vegetable-infecting viruses. Arch Virol 156: 1209–1216.
11. Li R, Gao S, Hernandez AG, Wechter WP, Fei Z, Ling KS (2012) Deep sequencing of small RNAs in tomato for virus and viroid identification and strain differentiation. PLoS One 7: e37127.
12. Fuentes S, Heider B, Tasso RC, Romero E, Zum Felde T, et al. (2012) Complete genome sequence of a potyvirus infecting yam beans (Pachyrhizus spp.) in Peru. Arch Virol 157: 773–776.
13. Giampetruzzi A, Roumi V, Roberto R, Malossini U, Yoshikawa N, et al. (2012) A new grapevine virus discovered by deep sequencing of virus- and viroid-derived small RNAs in Cv Pinot gris. Virus Res 163: 262–268.
14. Wu Q, Wang Y, Cao M, Pantaleo V, Burgyan J, et al. (2012) Homology-independent discovery of replicating pathogenic circular RNAs by deep sequencing and a new computational algorithm. Proc Natl Acad Sci U S A 109: 3938–3943.
15. Wu Q, Luo Y, Lu R, Lau N, Lai EC, 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.
16. Vodovar N, Goic B, Blanc H, Saleh MC (2011) In silico reconstruction of viral genomes from small RNAs improves virus-derived small interfering RNA profiling. J Virol 85: 11016–11021.
17. Aregger M, Borah BK, Seguin J, Rajeswaran R, Gubaeva EG, et al. (2012) Primary and secondary siRNAs in geminivirus-induced gene silencing. PLoS Pathog 8: e1002941.
18. Blevins T, Rajeswaran R, Aregger M, Borah BK, Schepetilnikov M, et al. (2011) Massive production of small RNAs from a non-coding region of Cauliflower mosaic virus in plant defense and viral counter-defense. Nucleic Acids Res 39: 5003–5014.
19. Loconsole G, Saldarelli P, Doddapaneni H, Savino V, Martelli GP, et al. (2012) Identification of a single-stranded DNA virus associated with citrus chlorotic dwarf disease, a new member in the family Geminiviridae. Virology 432: 162–172.
20. Mansilla C, Sánchez F, Padgett HS, Pogue GP, Ponz F (2009) Chimeras between oilseed rape mosaic virus and tobacco mosaic virus highlight the relevant role of the tobamoviral RdRp as pathogenicity determinant in several hosts. Mol Plant Pathol 10: 59–68.
21. Kurth EG, Peremyslov VV, Prokhnevsky AI, Kasschau KD, Miller M, et al. (2012) Virus-derived gene expression and RNA interference vector for grapevine. J Virol 86: 6002–6009.

22. Pooggin MM (2013) How can plant DNA viruses evade siRNA-directed DNA methylation and silencing? Int J Mol Sci 14: 15233–15259.
23. Duffy S, Holmes EC (2009) Validation of high rates of nucleotide substitution in geminiviruses: phylogenetic evidence from East African cassava mosaic viruses. J Gen Virol 90: 1539–1547.
24. Krenz B, Thompson JR, Fuchs M, Perry KL (2012) Complete genome sequence of a new circular DNA virus from grapevine. J Virol 86: 7715.
25. Al Rwahnih M, Dave A, Anderson MM, Rowhani A, Uyemoto JK, et al. (2013) Association of a DNA virus with Grapevines affected by Red Blotch disease in California. Phytopathology 103: 1069–1076.
26. Poojari S, Alabi OJ, Fofanov VY, Naidu RA (2013) A leafhopper-transmissible DNA virus with novel evolutionary lineage in the family geminiviridae implicated in grapevine redleaf disease by next-generation sequencing. PLoS One 8: e64194.
27. Mi S, Cai T, Hu Y, Chen Y, Hodges E, et al. (2008) Sorting of small RNAs into Arabidopsis argonaute complexes is directed by the 5′ terminal nucleotide. Cell 133: 116–127.
28. Havecker ER, Wallbridge LM, Hardcastle TJ, Bush MS, Kelly KA, et al. (2010) The Arabidopsis RNA-directed DNA methylation argonautes functionally diverge based on their expression and interaction with target loci. Plant Cell 22: 321–334.
29. Gómez G, Pallás V (2013) Viroids: a light in the darkness of the lncRNA-directed regulatory networks in plants. New Phytol 198: 10–15.
30. Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18: 821–829.
31. Schulz MH, Zerbino DR, Vingron M, Birney E (2012) Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 28: 1086–1092.
32. Namiki T, Hachiya T, Tanaka H, Sakakibara Y (2012) MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. Nucleic Acids Res 40: e155.
33. Thorvaldsdóttir H, Robinson JT, Mesirov JP (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 14: 178–192.
34. Li H, Durbin R (2009) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26: 1754–1760.
35. Seguin J, Otten P, Baerlocher L, Farinelli L, Pooggin MM (2014) 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.

Annex: (Seguin et al., 2014b)

MISIS: A bioinformatics tool to view and analyze maps of small RNAs derived from viruses and genomic loci generating multiple small RNAs

Jonathan Seguin, Patricia Otten, Loïc Baerlocher, Laurent Farinelli, Mikhail M. Pooggin

Journal of Virological Methods (2014), Vol.195, 120-122

Journal of Virological Methods 195 (2014) 120–122

Contents lists available at ScienceDirect

Journal of Virological Methods

journal homepage: www.elsevier.com/locate/jviromet

Short communication

MISIS: A bioinformatics tool to view and analyze maps of small RNAs derived from viruses and genomic loci generating multiple small RNAs

Jonathan Seguin[a,b], Patricia Otten[b], Loïc Baerlocher[b], Laurent Farinelli[b], Mikhail M. Pooggin[a,*]

[a] Zürich-Basel Plant Science Center, University of Basel, Department of Environmental Sciences, Botany, Hebelstrasse 1, 4056 Basel, Switzerland
[b] Fasteris SA, Ch. du Pont-du-Centenaire 109, 1228 Plan-les-Ouates, Switzerland

Article history:
Received 18 July 2013
Received in revised form
21 September 2013
Accepted 1 October 2013
Available online 14 October 2013

Keywords:
siRNA
miRNA
piRNA
Virus
Plant virus
Bioinformatics tool

A B S T R A C T

In eukaryotes, diverse small RNA (sRNA) populations including miRNAs, siRNAs and piRNAs regulate gene expression and repress transposons, transgenes and viruses. Functional sRNAs are associated with effector proteins based on their size and nucleotide composition. The sRNA populations are currently analyzed by deep sequencing that generates millions of reads which are then mapped to a reference sequence or database. Here we developed a tool called MISIS to view and analyze sRNA maps of genomic loci and viruses which spawn multiple sRNAs. MISIS displays sRNA reads as a histogram where the *x*-axis indicates positions of the 5′- or 3′-terminal nucleotide of sense and antisense sRNAs, respectively, along a given 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. MISIS reads standard BAM/SAM files outputted by mapping tools and generates table files containing counts of sRNA reads at each position of the reference sequence forward and reverse strand and for each of the chosen size-classes of sRNAs. These table files can be used by other tools such as Excel for further quantitative analysis and visualization. MISIS is a Java standalone program. It is freely available along with the source code at the following website: http://www.fasteris.com/apps.

1. Introduction

Small RNAs (sRNA) play a key role in regulation of gene expression and defense against invasive nucleic acids. In eukaryotes, sRNAs of 20–30 nts are classified in miRNAs, siRNAs and piRNAs based on the mechanisms of biogenesis and action (Ghildiyal and Zamore, 2009). Functional sRNAs are associated with the AGO/PIWI family proteins that sort sRNAs mostly by size and nucleotide composition (Cenik and Zamore, 2011). Populations of sRNAs can be extremely complex. Thus, siRNAs and piRNAs are produced as multiple species of different size-classes and abundances from various genomic loci as well as RNA and DNA viruses. The sRNA coverage profile with distinct size-classes and polarities (sense or antisense) and hotspot distribution along a given sequence can be indicative of the particular silencing mechanism that targets the corresponding sequences. Thus, detailed analysis of sRNA profiles is important for our understanding of gene regulation and defense mechanisms.

Currently, sRNA populations are analyzed using deep sequencing technologies such as Illumina, which generate millions of sRNA reads. Different mapping tools such as BWA (Li and Durbin, 2009) and bowtie (Langmead et al., 2009) align sRNA reads along a given reference sequence and generate results in a standard 'SAM/BAM' format (Li et al., 2009). The SAM/BAM files can be loaded by IGV (Robinson et al., 2011) to view the maps and Seqmonk (http://www.bioinformatics.bbsrc.ac.uk/projects/seqmonk/) to perform a quantitative analysis. To facilitate visualization of sRNA maps, a new tool called MISIS was developed which reads SAM/BAM files and generates table files containing sRNA read counts at each position of the reference sequence forward and reverse strands and for each of the chosen sRNA size-classes. These files can be converted into interactive histograms showing sRNA distribution along the reference sequence and providing detailed information on the origin and abundance of single sRNA species of different size-classes.

* 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).

0166-0934/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jviromet.2013.10.013

A

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	position	X20	X21	X22	X23	X24	X25	X20_rev	X21_rev	X22_rev	X23_rev	X24_rev	X25_rev	total_forward	total_reverse	total
2	1	1	3	2	1	2	0	-2	-2	-1	0	0	0	9	-5	14
3	2	0	0	0	0	1	0	-5	-9	0	-1	-8	0	1	-23	24
4	3	2	0	0	1	0	0	0	-2	0	-1	-2	0	3	-5	8
5	4	3	5	2	1	2	0	-2	-1	-1	0	0	0	13	-4	17
6	5	1	2	0	1	0	0	-1	0	-3	0	-5	0	4	-9	13
7	6	3	3	1	0	0	0	0	-17	-1	-1	-3	0	7	-22	29
8	7	1	0	1	0	10	0	0	-3	-1	-1	0	0	12	-5	17
9	8	0	0	0	0	0	0	-1	-5	-1	0	-1	0	0	-8	8
10	9	0	1	0	0	0	0	0	-3	0	0	0	0	1	-3	4
11	10	0	6	4	0	1	0	-1	-2	-2	0	0	0	11	-5	16
12	11	4	93	0	2	1	0	0	-5	0	-1	0	0	100	-6	106
13	12	10	0	1	0	1	0	0	0	0	-1	0	0	12	-1	13

B

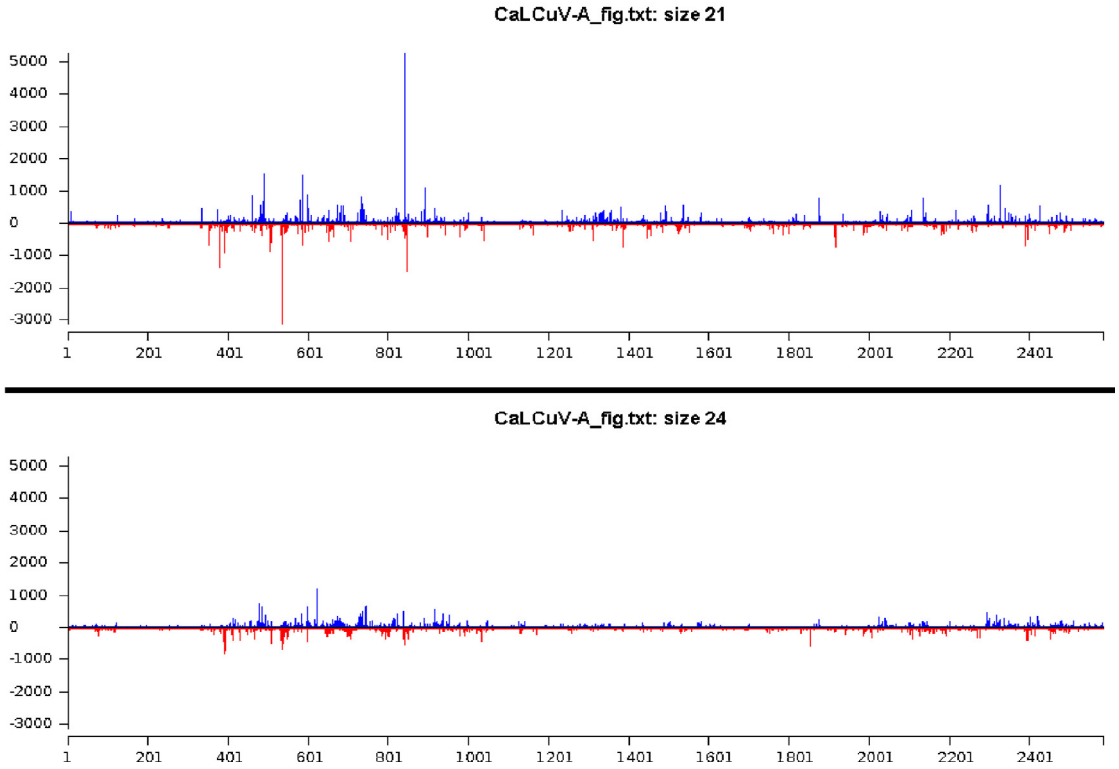


Fig. 1. (A) An example of the intermediate table generated by MISIS and visualized by Excel. The columns indicate/contain the nucleotide position along the reference sequence (position), the read counts for each sRNA of 20-, 21-, 22-, 23-, 24- and 25-nt classes mapped to the forward strand (X20, X21, X22, X23, X24, X25) and the reverse strand (X20_rev, X21_rev, X22_rev, X23_rev, X24_rev, X25_rev), and the total counts of sRNAs mapped on the forward (total_forward) and reverse (total_reverse) strands and on both strands (total). Only an upper part of the table is shown. (B) An example of the histograms created by MISIS using the table shown in panel A. The user selected two size-classes (21-nt and 24-nt), fixed the scales of the two histograms and saved the resulting image. The *x*-axis indicates the reference sequence positions (from 1 to 2500) against which multiple sRNA populations are mapped. The *y*-axis indicates the counts of each mapped sRNA (from 0 to 5000 reads). The sense sRNAs are displayed as blue bars and the antisense sRNAs as red bars. For further details of this particular analysis of viral siRNAs in plants infected with *Cabbage leaf curl virus* (CaLCuV), see Aregger et al. (2012). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

2. Results and discussion

MISIS was written in Java language to be compatible with Windows, Mac and Linux. This tool is intuitive for the user and has a limited number of simple but necessary functions. MISIS reads a BAM/SAM file containing sRNA mapping data for a given reference sequence (the maximal size of 300 kb is recommended, if the user's computer memory is limited to 8 Gb RAM), extracts the information on position, size, orientation and quality of each mapped sRNA and creates an output table. In this 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 (Fig. 1A). By reading CIGAR and tag "MD" values which indicate the number of mismatches in a mapped read, the algorithm simultaneously creates two output tables: one containing the counts of sRNAs mapped to the reference sequence without mismatch and another the combined counts for sRNAs mapped with up to two mismatches. Comparison of the two tables is informative of nucleotide difference (SNP or INDEL) between the sRNAs and the reference sequence. These small-size intermediate tables can be loaded by spreadsheet software such as Excel for further quantitative analysis and visualization.

As a second step, MISIS reads the table 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. 1B). The user can select sRNA size-classes separately or in any chosen combination, which then appear as single or multiple histograms, respectively. When two or more histograms are displayed, the user can change the scale of *y*-axis by clicking on "fix scale". This allows quantitative comparison of sRNAs. Furthermore, the user can zoom into any region of the reference sequence and move along the reference sequence. The histograms can be saved in jpeg, png or gif formats.

MISIS was tested for viewing and analysing various sRNA maps generated by the BWA tool. For example, the populations of viral siRNAs generated in DNA virus-infected plants were deep-sequenced and analyzed (Blevins et al., 2011; Aregger et al., 2012). The MISIS-aided analysis revealed that DNA viruses spawn massive quantities of siRNAs comparable to the entire population of endogenous plant miRNAs and siRNAs. The viral siRNAs belong to three major size-classes (21, 22 and 24 nt) and are distributed along the entire virus genome in both sense and antisense orientations, but the most abundant viral siRNA reads are confined to certain regions of the virus genome. The intermediate table files generated by MISIS were used for quantitative analysis of relative accumulation of viral siRNAs of different size-classes, polarities and 5′-nucleotide identities. This analysis along with genetic and biochemical evidence enabled us to uncover the molecular mechanisms of viral siRNA biogenesis and function in plant antiviral defense (Blevins et al., 2011; Aregger et al., 2012).

In summary, MISIS is a simple and user-friendly tool for visualizing and analyzing sRNA maps of the viruses and genomic regions generating multiple sRNAs. It complements other sRNA map-view and analysis methods such as 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

The sRNA libraries used for testing MISIS were prepared as described in Aregger et al. (2012), using the well-established protocols for sRNA extraction from virus-infected plants (Blevins et al., 2011; Aregger et al., 2012) and the current Illumina and Fasteris (www.fasteris.com) protocols and algorithms for cDNA preparation, sequencing and adaptor removal from raw reads. For mapping of the resulting sRNA datasets, BWA (Li and Durbin, 2009) was used with the defaults parameters (http://bio-bwa.sourceforge.net/bwa.shtml).

Acknowledgements

We thank Mihaela Zavolan for critical reading of the manuscript and Thomas Boller for supporting our research. 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 SER No. C09.0176 to L.F. and M.M.P.).

References

Aregger, M., Borah, B.K., Seguin, J., Rajeswaran, R., Gubaeva, E.G., Zvereva, A.S., Windels, D., Vazquez, F., Blevins, T., Farinelli, L., Pooggin, M.M., 2012. Primary and secondary siRNAs in geminivirus-induced gene silencing. PLoS Pathog. 8, e1002941.

Blevins, T., Rajeswaran, R., Shivaprasad, P.V., Beknazariants, D., Si-Ammour, A., Park, H.S., Vazquez, F., Robertson, D., Meins Jr., F., Hohn, T., Pooggin, M.M., 2011. Massive production of small RNAs from a non-coding region of Cauliflower mosaic virus in plant defense and viral counter-defense. Nucleic Acids Res. 39, 5003–5014.

Cenik, E.S., Zamore, P.D., 2011. Argonaute proteins. Curr. Biol. 21, R446–R449.

Ghildiyal, M., Zamore, P.D., 2009. Small silencing RNAs: an expanding universe. Nat. Rev. Genet. 10, 94–108.

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.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078–2079.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., Mesirov, J.P., 2011. Integrative genomics viewer. Nat. Biotech. 29, 24–26.

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.