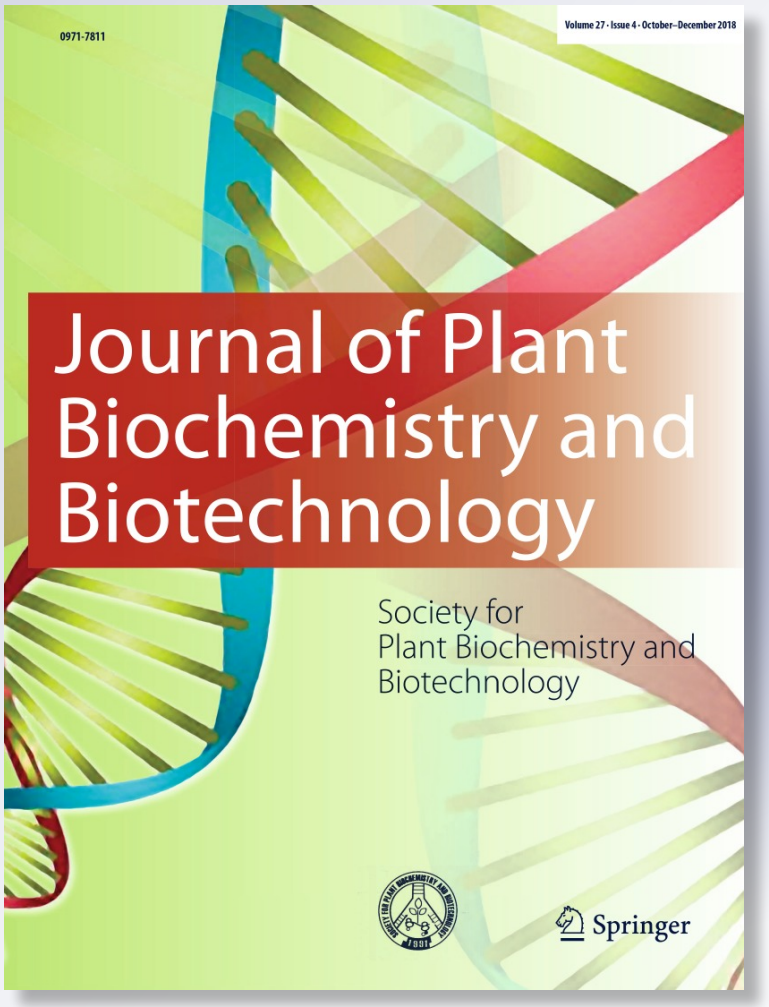
Comparative characterization of small RNAs derived from an emaravirus and a geminivirus infecting pigeonpea

Basavaprabhu L. Patil & Deepika Arora

Journal of Plant Biochemistry and Biotechnology

ISSN 0971-7811
Volume 27
Number 4

J. Plant Biochem. Biotechnol. (2018)
27:382-392
DOI 10.1007/s13562-018-0447-9



Springer

Journal of Plant Biochemistry and Biotechnology (October–December 2018) 27(4):382–392
https://doi.org/10.1007/s13562-018-0447-9

ORIGINAL ARTICLE

Comparative characterization of small RNAs derived from an emaravirus and a geminivirus infecting pigeonpea

Basavaprabhu L. Patil[1] · Deepika Arora[1]

Received: 7 July 2017 / Accepted: 6 February 2018 / Published online: 3 March 2018

Abstract

High throughput sequencing technologies, supported by bioinformatics tools are employed to retrieve small RNA sequence information derived from the nucleic acids of plant infecting viruses. In addition to characterization of the small RNAs to understand the biology of the virus, the small RNA sequence can be assembled to reconstitute viral genome sequence. For the first time the semiconductor based Ion Proton sequencing technology is used to sequence the small RNAs from pigeonpea (*Cajanus cajan*) plants infected by two distinct viruses with RNA and DNA as their genomes. The reconstitution of the viral genome sequence revealed that the pigeonpea plant from Kalaburagi (erstwhile Gulbarga, Karnataka state) was infected by an emaravirus species *Pigeonpea sterility mosaic emaravirus* 1 (PPSMV-1) and another plant from New Delhi was infected by a begomovirus species *Mungbean yellow mosaic India virus* (MYMIV). Characterization and comparison of small RNA sequences derived from both the viruses showed vast differences in their pattern of accumulation and their size classes. In the case of PPSMV-1, the 21 nt sized siRNAs accumulated at far greater levels followed by 22 and 24 nt siRNAs. Whereas in MYMIV, the proportion of accumulation of each size class of siRNAs was similar. Further the distribution of small RNAs across the genomes of PPSMV-1 and MYMIV was mapped and the density of small RNA accumulation showed a positive correlation with the GC content of viral sequence.

Keywords Emaravirus · Geminivirus · Pigeonpea · Small RNA · Next generation sequencing · Ion Proton

Introduction

In a virus infected plant a significant proportion of small interfering RNA (siRNA) population is derived from the viral genome or its transcripts through a mechanism called RNA-interference (RNAi). RNAi provides antiviral immunity in majority of eukaryotic organisms including plants, by production of virus-derived siRNAs (vsiRNAs) through the action of Dicer-Like (DCL) proteins which are subsequently loaded on to an Argonaute protein complex for antiviral silencing (Wang et al. 2012). RNAi-mediated viral immunity in plants requires host RNA-directed RNA polymerase (RDR), such as RDR1, RDR2 and RDR6 to produce viral secondary siRNAs following viral RNA replication-triggered primary siRNA production (Wang et al. 2012).

Plants code for four types of DCLs, namely DCL1, DCL2, DCL3 and DCL4 which process dsRNA to produce different sizes of siRNAs required for antiviral defense (Wang et al. 2012). The 21 nt siRNAs, also known as primary siRNAs are processed by DCL4, whereas DCL2 produces 22 nt or 23 nt siRNAs and DCL3 forms 24 nt siRNAs (Axtell 2013; Deleris et al. 2006). The 21, 22 and 24 nt siRNAs are known to be involved in antiviral defence in plants, but the 23 nt siRNA is not known to have an antiviral activity (Axtell 2013; Deleris et al. 2006). DCL1 also can produce both 21 and 24 nt siRNAs, but none of them are involved in antiviral defence (Deleris et al. 2006), however the most important function of DCL1 is to produce microRNAs of size 21 nt, which regulates gene expression in the plants (Matzke and Birchler 2005). The vsiRNAs play a major role in defense against plant viruses and modification of the host genome and hence are key to the understanding of pathogenicity of plant viruses. Studies

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s13562-018-0447-9) contains supplementary material, which is available to authorized users.

✉ Basavaprabhu L. Patil
blpatil2046@gmail.com

[1] ICAR-National Research Centre on Plant Biotechnology, LBS Centre, IARI Campus, Pusa, New Delhi 110012, India

using RNA and DNA viruses have revealed that the vsiRNAs could be derived from several regions of viral genome (Zhang et al. 2015).

The revolution in next generation sequencing (NGS) technologies has helped in the discovery of different sized vsiRNAs which unravels an accurate scenario about the abundance, complexity and diversity of vsiRNAs in a virus infected organism (Niu et al. 2015). Hence one of the important strategies for identification of hitherto unknown viruses is the cloning and sequencing of small RNAs from a plant infected by an unknown virus by NGS technologies, supported by bioinformatics tools (Barba et al. 2014; Boonham et al. 2014; Hagen et al. 2012; Idris et al. 2014). Of the several NGS technologies, Ion Proton system is the first benchtop sequencing system capable of sequencing in a few hours of time and is relatively cheaper than illumina-based NGS. Ion Proton system combines the semiconductor sequencing technology which directly translates the chemical information into digital data (Rothberg et al. 2011), generating high level of scalability and flexibility to cater to the diverse range of high throughput sequencing applications (Liu et al. 2012; Yuan et al. 2016). The Ion semiconductor sequencing technique is based on the detection of hydrogen ions (H^+) produced during the polymerization of DNA and hence it is termed as "sequencing by synthesis", during which the complementary strand is synthesized. Recently Ion Proton based NGS has been successfully used for sequencing human immunodeficiency virus (HIV) from different clinical samples (Ameur et al. 2014). This study is the first report on the use of Ion Proton system for sequencing small RNAs derived from plant viruses.

In this study, Ion Proton system has been used to sequence the small RNA libraries constructed from pigeonpea (*Cajanus cajan*) plants infected with two distinct plant viral species, namely *Pigeonpea sterility mosaic emaravirus* 1 (PPSMV-1) belonging to the genus *Emaravirus* of family *Fimoviridae* and *Mungbean yellow mosaic India virus* (MYMIV) belonging to the genus *Begomovirus* of family *Geminiviridae* (Hema et al. 2014; Patil and Kumar 2015, 2017). The genus *Emaravirus* consists of viruses with segmented negative sense RNA viral genome, a characteristic feature of the order Bunyavirales (Kormelink et al. 2011; Mielke-Ehret and Muehlbach 2012). PPSMV causes sterility mosaic disease (SMD) of pigeonpea, which is a major viral disease of pigeonpea occurring in the Indian subcontinent (Patil and Kumar 2015, 2017). SMD is caused by two distinct emaravirus species, namely *Pigeonpea sterility mosaic emaravirus* 1 (PPSMV-1; Elbeaino et al. 2014) and *Pigeonpea sterility mosaic emaravirus* 2 (PPSMV-2; Elbeaino et al. 2015). PPSMV-1 was the first to be identified (Elbeaino et al. 2014) and subsequently another distinct emaravirus PPSMV-2 was also reported to be involved in SMD (Elbeaino et al. 2015). Both PPSMV-1 and PPSMV-2 consist of six genomic RNA segments, namely, RNA1 (7022 nt in length) encoding for RNA-dependent RNA polymerase (RdRp, 2295 amino acids); RNA2 (2223 nt) coding for glycoprotein (GP, 649 amino acids); RNA3 (1442 nt) coding for nucleocapsid protein (NP, 309 amino acids); RNA4 (1563 nt) coding for a putative movement protein p4 (MP, 362 amino acids); RNA5 (1689 nt) coding for a protein with unknown function (474 amino acids), and RNA6 (1194 nt) also coding for a protein with unknown function (238 amino acids) (Elbeaino et al. 2014, 2015; Patil et al. 2017). Although recent publications have reported presence of only five segments in PPSMV-1, in contrast to six segments in PPSMV-2 (Elbeaino et al. 2014, 2015), our studies have shown association of RNA6 with PPSMV-1 (Patil et al. 2017). The *Fig mosaic virus* (FMV), *European mountain ash ringspot-associated virus* (EMARaV), *Raspberry leaf blotch virus* (RLBV), *Redbud yellow ringspot associated virus* (RYRSaV), *Rose rosette virus* (RRV), and *Wheat mosaic virus* (WMoV) are some of the emaravirus species which encompass 4–8 genomic RNA segments and are known to cause important plant viral diseases (Mielke-Ehret and Mühlbach 2012; Patil and Kumar 2015; Tatineni et al. 2014). Whereas the begomovirus MYMIV, consisting of two single stranded circular DNA components namely, DNA-A and DNA-B, each of about $\sim$ 2.7 kb, is known to cause a major viral disease called as yellow mosaic disease (YMD) in several legumes such as mungbean, soybean and cowpea, however at present, YMD is a minor disease of pigeonpea (Hema et al. 2014; Kumar et al. 2017). The DNA-A component codes for six proteins, namely Replication associated protein (AC1 or Rep), Transcription activator protein (AC2), Replication enhancer protein (AC3) and AC4 known to be a silencing suppressor on the complementary strand and coat protein (AV1) and pre-coat protein (AV2) on the virion strand required for replication and encapsidation. Whereas the DNA-B component encodes for movement protein (BC1) and nuclear shuttle protein (BV1), required for inter- and intra-cellular movement respectively (Hanley-Bowdoin et al. 2000; Jeske 2009).

In this study, the vsiRNAs derived from both the viruses infecting pigeonpea are characterized and compared. Further, by assembling the vsiRNA sequences we have reconstituted the entire viral genome sequence for both of these DNA and RNA viruses infecting pigeonpea (Seguin et al. 2014).

Materials and methods

Field collection of pigeonpea leaf samples and isolation of total RNA

The leaf sample of a pigeonpea plant showing characteristic symptoms of SMD was collected from Kalaburagi (formerly Gulbarga) in Karnataka state of India. Similarly, leaf sample of another pigeonpea plant with symptoms of YMD was collected from IARI campus, New Delhi (India), along with a healthy control plant, which was confirmed to be negative for presence of above viruses by PCR and RT-PCR. These leaf samples were snap frozen in liquid nitrogen and stored in − 80 °C deep freezer for further use. The leaf samples from each plant were pooled and the total RNA was isolated using Spectrum^{TM} Total RNA Kit from Sigma (Sigma-Aldrich, USA), following manufacturer's instructions. The amount of binding solution was increased for each extraction to maintain the yield of small RNAs. The concentration and quality of RNA was checked using Qubit RNA assay kit and Bioanalyzer Total RNA QC (RNA Nano Kit).

Small RNA library preparation and sequencing using Ion Proton sequencer

The small RNA population was enriched from the total RNA and its QC (Quality Control) was checked using RNA Qubit Assay Kit and further its RNA integrity number (RIN10) values were determined using the Agilent 2100 Bioanalyzer System expert software. Subsequently the good quality small RNA was subjected to library preparation using the Total RNA Seq Kit v2 (Thermo Fischer Scientific), as per the instructions in the user manual and the library QC was checked using the Qubit HS assay kit. Template preparation was done using an Ion PI^{TM} OT2 200 kit V2 and the pooled (equimolar pool of the 3 libraries) concentration was determined to 20 pico moles (pM).

Data analysis

Raw reads were downloaded from the Ion Proton system in fastq format and filtered using fastx_toolkit (v0.0.14); later the reads with length less than 16 nt and more than 35 nt were deleted (http://hannonlab.cshl.edu/fastx_toolkit/). The filtered reads were mapped onto the reference genome of pigeonpea (Singh et al. 2011; Varshney et al. 2012) using Shrimp (v2.2.3) mapping software (Rumble et al. 2009) and reads remaining unmapped were converted into fastq file. Mapped data was utilized to perform statistical study whereas the unmapped reads were used for further viral sequence analysis.

The unmapped reads were subjected to de novo assembly to derive contigs using MIRA (v3.9.18) (Chevreux et al. 1999) which provides different contig counts for all the samples. These contigs were also aligned on published PPSMV-1 and MYMIV sequences to study the similarity using MUMmer (Delcher et al. 2002) and consensus sequences were generated using ABACAS (v1.3.1) (Assefa et al. 2009). Further, the unmapped reads were used to align on the published PPSMV-1 (Elbeaino et al. 2014) and MYMIV sequences using Shrimp (v2.2.3) and the mapped data was used to perform downstream analysis. Different statistics were generated along with graphs and charts for representations. After mapping of reads on reference sequences, its distribution was calculated across all the references along with segregation of mapped reads based on their lengths. The software Samtools (0.1.19) (Li et al. 2009) was used to perform various tasks for calculating and generating statistical summary.

The small RNA reads mapping on the reference viral genome were segregated based on their read length of 21, 22, and 24 nt. After segregation, the small RNA reads were counted and reads per million (RPM) were calculated for all the categories of small RNA read lengths. The RPM graphs were generated based on mapped small RNA reads covering the genomes of PPSMV-1 and MYMIV. The programs such as Samtools and bamutils (NGSutils—v0.5.7) (Breese and Liu 2013) were used to extract and generate counts and statistics for different small RNA read lengths.

RT-PCR, cloning and sequencing of PPSMV-1 and MYMIV

From the total RNA isolated using Spectrum^{TM} Plant Total RNA Kit (Sigma-Aldrich, USA), 2 µg was reverse transcribed for cDNA synthesis using MultiScribe Reverse Transcriptase, RT Random Primers and other components from the High-Capacity cDNA Reverse Transcription kit (Applied Biosystems, USA) as per manufacturers guidelines. The diluted cDNA was used for PCR amplification of various segments of PPSMV-1 using specific primer pairs (Suppl. Table 2). The eluted PCR amplicons were ligated in pGEM-T Easy vector (Promega, USA) according to manufacturer's guidelines and transformed in *E. coli*. The recombinant plasmids containing target sequences were subjected to Sanger sequencing. The edited raw sequences were manually assembled by identifying overlapping sequences, further the nucleotide homology searches were done with BLASTN Sequence Analysis of NCBI (http://blast.ncbi.nlm.nih.gov/Blast.cgi).

To clone the genomic components of MYMIV, total DNA was isolated from pigeonpea leaf sample with YMD symptoms using CTAB method and then it was subjected

to ϕ29 DNA polymerase-mediated rolling circle amplification (RCA) using illustra TempliPhi DNA Amplification Kits (TempliPhi^{TM}, GE Healthcare) (Haible et al. 2006; Johne et al. 2009). The amplified DNA was digested with the restriction enzyme *Eco*RI to linearise the ~ 2.7 kb sized DNA-A and DNA-B components and were subsequently cloned in pGreen0029 vector and sequenced by Sanger sequencing.

Results

The assembled and validated sequences of PPSMV-1 and MYMIV were submitted to the NCBI sequence database, with GenBank accession numbers: KX363886–KX363891 for six RNA (RNA1-RNA6) segments of PPSMV-1, and KX363947 and KX363948 for DNA-A and DNA-B of MYMIV. A total number of raw reads of 9,503,998 for healthy control pigeonpea plant, 25,168,817 for PPSMV-1 infected pigeonpea from Kalaburagi (Gulbarga), and 22,518,970 for MYMIV infected pigeonpea from Delhi were obtained by Ion Proton based NGS for small RNAs. The number of small RNA reads that remained after trimming of adaptor and filtering were 6,869,125 for control plant, 17,679,838 for PPSMV-1 infected pigeonpea and 16,144,586 for MYMIV infected pigeonpea. Of these 6,371,918, 16,623,248 and 15,194,545 respectively mapped to the pigeonpea genome and the rest remained unmapped. For these two virus infected pigeonpea samples, 1,134,902 small RNA reads in the size range of 16–25 nts mapped to the PPSMV-1 genome and 792,665 reads mapped to the genome of MYMIV, accounting for 6.47 and 4.99% of total small RNA population, respectively (Table 1, Fig. 1).

Except for RNA1 of PPSMV-1 with a small RNA coverage of 94%, the other five RNA segments had ~ 100% coverage by the vsiRNAs. For MYMIV DNA-A component the vsiRNA coverage was 99.74%, while for DNA-B component it was 98.62% (Table 2). The relative content (%) of vsiRNAs derived from the six segments of PPSMV-1 was 3.6% for RNA1, 11.7% for RNA2, 36.5% for RNA3, 19.8% for RNA4, 10.7% for RNA5, and 17.6% for RNA6 (Table 2). Whereas for MYMIV, the relative content (%) of small RNA population for DNA-A and DNA-B components was 48 and 52% respectively (Table 2).

Of the six RNA segments of PPSMV-1, the RNA3 segment produced the highest amount of vsiRNAs, followed by RNA4, RNA6, RNA2, RNA5 and RNA1 respectively (Table 2). For all the six RNA segments of PPSMV-1, the 21 nt vsiRNAs accumulated at the highest level, followed by 22 nt and 24 nt. However the relative content (%) or percentage share of the three different size classes of vsiRNAs varied among the six PPSMV-1 segments (Table 2). In particular the content of 21 nt vsiRNAs was highest for RNA1 segment.

In contrast to PPSMV-1, the relative proportion of accumulation of the three different size classes of vsiRNAs namely 21, 22 and 24 nt was fairly equivalent for both the MYMIV genomic components (Table 2). Although the MYMIV DNA-B component accumulated relatively higher levels of vsiRNAs compared to the DNA-A component, the levels of 21 nt vsiRNA was highest for DNA-A, whereas for DNA-B the 22 nt vsiRNA were at highest level, followed by 24 nt and 21 nt vsiRNAs. In contrast to a positive correlation between the GC content and the levels of vsiRNAs produced for PPSMV-1, no such correlation was observed for MYMIV. Despite the higher (43.7%) GC content of MYMIV DNA-A component, than its

Table 1 Summary of small RNA deep sequencing results from three different pigeonpea plants from two locations (Gulbarga/Kalaburagi and New Delhi), infected by Pigeonpea sterility mosaic emaravirus 1 (PPSMV-1) and Mungbean yellow mosaic India virus (MYMIV)

Experimental samples (library)	Control (healthy)	Gulbarga (PPSMV-1)	New Delhi (MYMIV)
Total raw reads of small RNAs	9,503,998	25,168,817	22,518,970
Pre-processing reads after trimming and filtering	6,869,125	17,679,838	16,144,586
Small RNA reads mapping to Pigeonpea genome	6,371,918	16,623,248	15,194,545
Small RNA reads remaining unmapped	497,207	1,056,590	950,041
Pigeonpea small RNA annotation			
Reads mapping within siRNA region	72,694	131,687	267,101
Reads mapping outside siRNA region	6,299,224	16,491,561	14,927,444
Total small RNA reads	–	17,548,145	15,877,483
Small RNA reads mapping to viral genome (16–25 nts)	–	1,134,902	792,665
Percent small RNA reads mapping (%)	0	6.47	4.99

The virus free healthy pigeonpea plant from New Delhi is used as control

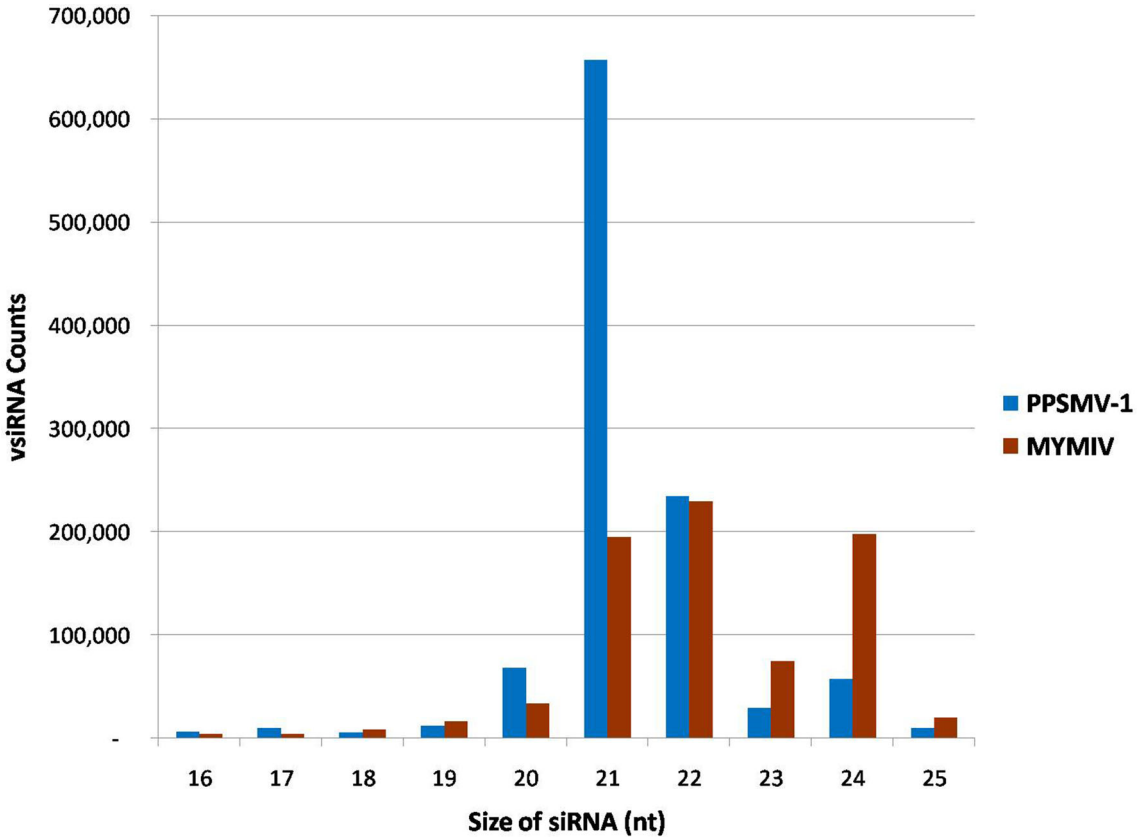


Fig. 1 Size distribution of total small RNAs (16–25 nts) in libraries prepared from pigeonpea plants infected by an emaravirus, Pigeonpea sterility mosaic emaravirus 1 (PPSMV-1) and a begomovirus, Mungbean yellow mosaic India virus (MYMIV)

corresponding DNA-B component (40.4%), DNA-B produced higher levels of vsiRNAs (Suppl. Table 1). However the levels of the primary vsiRNAs of 21 nt accumulated at much higher level in MYMIV DNA-A than the DNA-B component (Table 2). Similar to the PPSMV-1 segments, within each MYMIV genomic components, the GC rich regions showed peaks for vsiRNA accumulation and in contrast the GC poor regions accumulated low levels of vsiRNAs in both of the DNA components (Suppl. Table 1).

For both PPSMV-1 and MYMIV, the coding regions (ORFs) produced significantly higher levels of vsiRNA than the non coding regions and coincidentally these regions had low GC content (Fig. 3 and 4). This was evident for PPSMV-1 segments RNA-2, 3, 4, 5 and 6 (Fig. 3). Although the density of both sense (+ve strand) and antisense (−ve strand) vsiRNAs showed similar pattern of concentration on certain regions of both PPSMV-1 and MYMIV genomes, however there were exceptions to this (Figs. 3, 4). For all the PPSMV-1 segments, the 21 nt and 22 nt vsiRNAs derived from the sense strand were more than the vsiRNAs derived from the antisense strand for both vsiRNAs, but not for 24 nt vsiRNAs (Table 3 and Fig. 2). In contrast to PPSMV-1, higher levels of vsiRNAs were produced from the antisense (−ve) strand as compared to the sense (+ve) strand for the three size classes of vsiRNAs in MYMIV (Table 3).

Discussion

Next generation sequencing (NGS) technology has revolutionized the discovery of unknown viruses and the sequence information of virus derived small RNAs is key to reconstitution of virus sequence. Since last several years, various NGS technologies have been developed for high throughput sequencing and these diverse NGS platforms employ different sequencing biochemistries and differ in sequencing protocol. These different NGS platforms are under constant improvement to make them faster, efficient and cost-effective in order to enhance the accessibility of high throughput sequencing and also to accelerate their applications. Thus selection of appropriate NGS platform is critical to cater to specific sequencing requirements and also for the economic feasibility of sequencing costs. Ion Proton, a semiconductor based bench-top high throughput sequencing technology is one of the most economical NGS platform and in addition the sequencing reaction needs a short span of time, than the other platforms.

Table 2 Number of 21, 22 and 24 nt sized vsiRNA reads that map to the RNA segments (RNA1–RNA6) of Pigeonpea sterility mosaic emaravirus 1 (PPSMV-1) and the two DNA components (DNA-A and DNA-B) of Mungbean yellow mosaic India virus (MYMIV)

Size of vsiRNAs	21 nt	22 nt	24 nt	% share of vsiRNAs across segments	% viral genome covered by vsiRNAs
Pigeonpea from Kalaburagi (Gulbarga)					
Total reads mapped	656,562	234,681	57,506	–	–
PPSMV-1 RNA1	31,937 (82.24%)	5919 (15.24%)	978 (2.52%)	3.6	94.13
PPSMV-1 RNA2	86,929 (69.27%)	33,482 (26.68%)	5085 (4.05%)	11.7	100
PPSMV-1 RNA3	273,045 (69.58%)	94,073 (23.97%)	25,287 (6.45%)	36.5	100
PPSMV-1 RNA4	142,172 (66.71%)	59,424 (27.89%)	11,511 (5.4%)	19.8	99.74
PPSMV-1 RNA5	79,079 (68.8%)	28,373 (24.69%)	7483 (6.51%)	10.7	100
PPSMV-1 RNA6	133,324 (70.49%)	39,315 (20.79%)	16,496 (8.72%)	17.6	100
Pigeonpea from New Delhi					
MYMIV DNA-A	106,611 (35.68%)	104,674 (35.03%)	87,500 (29.29%)	48%	99.89
MYMIV DNA-B	88,251 (27.29%)	125,154 (38.70%)	110,010 (34.01%)	52%	98.62

The relative content (%) of vsiRNAs derived from each viral genomic segment/component with reference to the total number of vsiRNAs is summarized. The percentage of viral genome covered by the vsiRNAs is also given

Table 3 Number of vsiRNAs (21, 22 and 24 nt) that map to sense and antisense strands of Pigeonpea sterility mosaic emaravirus 1 (PPSMV-1) and Mungbean yellow mosaic India virus (MYMIV) from two different locations (Kalaburagi and New Delhi)

Location	Virus and their strand	Total vsiRNAs counts		
		21 nt	22 nt	24 nt
Kalaburagi (Gulbarga)	PPSMV-1	656,562	234,681	57,506
	+ve strand	376,375	130,061	24,256
	−ve strand	280,187	104,620	33,250
New Delhi	MYMIV	194,862	229,828	197,510
	+ve strand	85,538	92,839	71,927
	−ve strand	109,324	136,989	125,583

In addition to the discovery of novel viral sequences through vsiRNA sequence information obtained through NGS, the detailed characterization of small RNA sequence information also helps in understanding the role of vsiRNAs in antiviral defense and host genome modifications. Through high throughput sequencing of the small RNAs one can learn about the diversity and complexity of virus derived small RNAs, which helps in understanding of viral pathogenicity and interaction with their host plants. Hitherto deep sequencing has mostly been employed to characterize the virus derived small RNAs in the experimental host plants or in natural hosts maintained in experimental conditions. Some of the examples are *Cucumis melo* plants infected with Melon necrotic spot virus, Watermelon mosaic virus (Donaire et al. 2009) and tomato plants infected with Tomato yellow leaf curl virus (Yang et al. 2011), grapevine plants infected with different viruses (Pantaleo et al. 2010), rice plants infected with Rice stripe virus (RSV) (Yan et al. 2010; Xu et al. 2012) and tomato spotted wilt virus (TSWV) infected tomato (Mitter et al. 2013). Most of the past studies are based on positive sense RNA viruses and there are few reports on characterization of small RNAs derived from negative sense RNA viruses, except for RSV and TSWV, which have −ve sense RNA genomes (Mitter et al. 2013; Yan et al. 2010; Xu et al. 2012). *Emaravirus* is a newly identified genus consisting of negative sense segmented RNA genome (Mielke-Ehret and Muehlbach 2012). Most recently some of the emaravirus

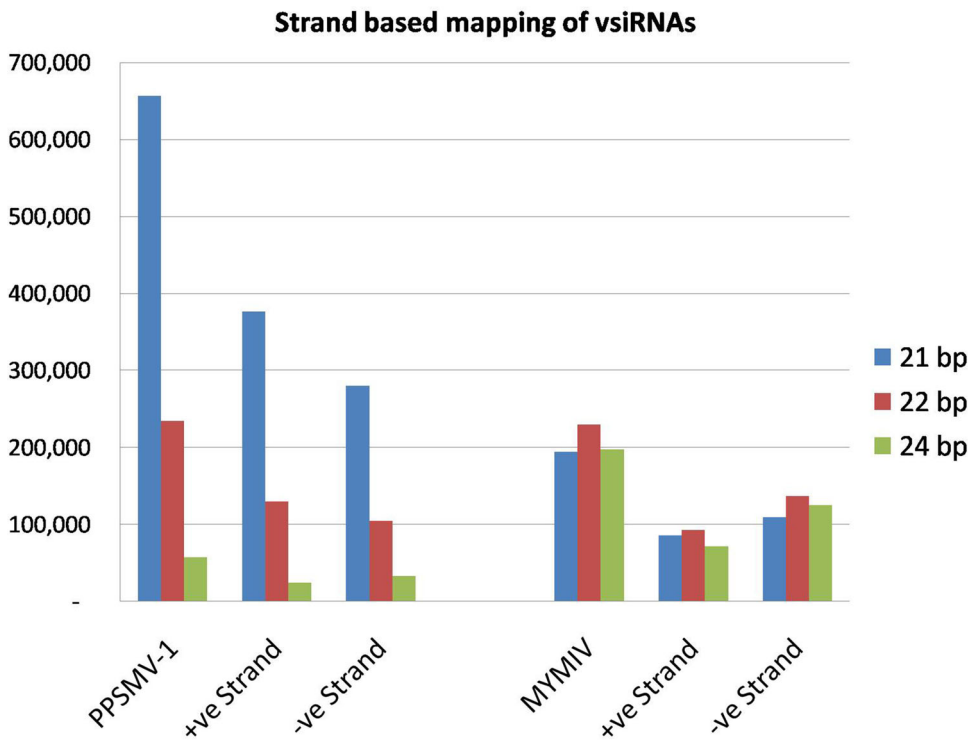


Fig. 2 Mapping of sense and antisense strand derived 21, 22 and 24 nt siRNAs specific Pigeonpea sterility mosaic emaravirus 1 (PPSMV-1) and Mungbean yellow mosaic India virus (MYMIV)

genome sequences, such as Actinidia chlorotic ringspot-associated virus (AcCRaV; Zheng et al. 2016) and Woolly burdock yellow vein virus (WBYVV; Bi et al. 2012) are discovered by high throughput sequencing of their small RNAs.

This study has made the first molecular characterization of the small RNAs derived from PPSMV-1 an emaravirus and MYMIV a geminivirus (or begomovirus), infecting two pigeonpea plants in two different locations. These studies show that the RNA1 segment of PPSMV-1 accumulated lowest levels of vsiRNAs despite being the largest RNA segment in contrast to RNA3 and RNA4 segments, which generated highest levels of vsiRNAs. Northern hybridization studies with WMoV indicated that the RNA1 accumulated at lower levels compared to other RNA segments (Tatineni et al. 2014) and also a recent report on electrophoresis of PPSMV dsRNA has revealed that the levels of RNA-3 and RNA-4 are much higher than the levels of RNA1 segment (Kumar et al. 2017). In the past similar reports are made for the tospovirus TSWV, in which case the M-RNA produced highest levels of vsiRNA followed by S-RNA and L-RNA, in both tomato and the model host *Nicotiana benthamiana* (Mitter et al. 2013). Similarly for RSV, maximum vsiRNAs were obtained from RNA4 and minimum from the RNA1 segment (Yan et al. 2010; Xu et al. 2012). The effect of particular virus species/strain/isolate, the host plant species and the implications of host–virus–environment interaction on the pattern of vsiRNA accumulation cannot be ruled out (Fletcher et al. 2016; Kuria et al. 2017).

The GC content of the PPSMV-1 segments ranged from 30.8 to 33.9% and the majority of RNA segments such as RNA4, RNA3, RNA6 and RNA2 with higher GC content accumulated higher levels of vsiRNAs, while RNA1 and RNA5 accumulated relatively lower levels of vsiRNAs (Table 2). Across each RNA segment, the GC rich regions showed higher levels of vsiRNA accumulation and the GC poor regions accumulated low levels of vsiRNAs (Suppl. Table 1). Such positive correlations were also observed between the GC% and levels of vsiRNA in TSWV (Mitter et al. 2013). Several past reports also show a positive correlation between the GC content and the levels of vsiRNAs they accumulate (Ho et al. 2007; Rudnick et al. 2008; Patil and Fauquet 2015). Such a variation in the levels of accumulation of vsiRNAs with no positive correlation to the size of the genomic RNA segment could also be because of less number of RNA1 copies when compared to other RNA segments, such as RNA3, RNA4 and RNA6. A closer look at the peaks of vsiRNA accumulation within each of the PPSMV-1 RNA segments and the MYMIV DNA components showed that these regions were enriched with GC bases compared to their adjacent regions with lower levels of vsiRNA accumulation (Figs. 3, 4 and Suppl. Table 1). Such hotspots of siRNAs are clusters of multiple reads and previous studies have reported higher GC content in such hotspots (Donaire et al. 2009; Mitter et al. 2013).

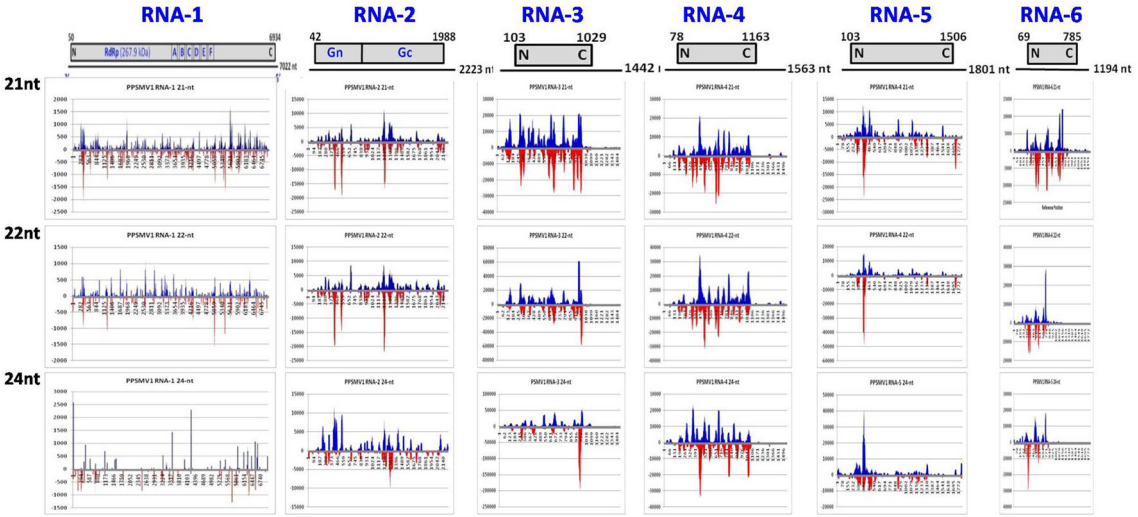


Fig. 3 Distribution maps three size classes of vsiRNAs, viz. 21 nt, 22 nt and 24 nt across the six RNA segments (RNA1–RNA6) of Pigeonpea sterility mosaic emaravirus 1 (PPSMV-1). The genome organization of the PPSMV-1 RNA segments is schematically represented on the top. The vsiRNAs originating from the sense strand of the viral genome is shown in blue colour and those originating from the antisense strand are in red colour. The number of reads per million (RPM) of the vsiRNAs is given on the *X* axis and the viral genomic coordinates are marked on the *Y* axis

Such a variation could also be due to variation in the secondary or tertiary structures of the viral RNA segments or their transcripts and their differential accessibility to the RNAi machinery (Donaire et al. 2008). This difference could also be because of an important role of RNA1 segment encoding RdRp, critical for replication of RNA segments and hence may have evolved to have maximum protection from RNAi machinery. This information should help in identifying preferred emaravirus sequences to be targeted by RNAi and for efficient management of PPSMV, either by RNAi or dsRNA based strategy (Patil et al. 2016; Voloudakis et al. 2015).

Characterization of vsiRNAs from the MYMIV infected pigeonpea revealed that the DNA-B component produced higher levels of vsiRNAs than the corresponding DNA-A, despite a higher GC content of DNA-A (Tables 2, 3). Such a difference could also be because of higher copy number of DNA-B component (Briddon et al. 2010). However within each DNA component of MYMIV, the GC rich regions accumulated higher levels of siRNAs (Fig. 4 and Suppl. Table 1). In case of DNA-A component the ORFs *AV*1 and *AC*3 accumulated higher levels of siRNAs (Fig. 4). However, the un-transcribed intergenic region always accumulated lower levels of siRNAs (Yang et al. 2011). Strong fold-back structures within the viral transcripts, or overlapping transcripts placed in opposite orientations, and also large numbers of aberrant transcripts are the major precursors for the host RNA Dependent RNA polymerases (RDRs), which eventually trigger PTGS (Chellappan et al. 2004; Vanitharani et al. 2005).

In case of both PPSMV-1 and MYMIV, the ORFs accumulated higher levels of vsiRNA, when compared to the non-coding untranslated regions (UTRs). This pattern was conspicuous in the PPSMV-1 segments RNA3, RNA4, RNA5 and RNA6. Similar such observations are also made for the recently characterised emaravirus AcCRaV (Zheng et al. 2016). The GC content of these PPSMV-1 segments was higher in the coding region when compared to the non-coding 5′ UTR (Fig. 3 and Suppl. Table 1) and such correlations are also made for AcCRaV (Zheng et al. 2016). Similar to AcCRaV, there was a fairly uniform distribution of vsiRNAs across RNA1 and RNA2 of PPSMV-1 (Zheng et al. 2016).

It is interesting to note that the 21 nt sized vsiRNAs accumulated at far greater levels for PPSMV-1, followed by 22 nt and 24 nt. However, the accumulation of 20 nt sized vsiRNAs was relatively more than 24 nt. Although there is no role assigned to 20 nt sized vsiRNAs, certain microRNAs of 20 nt size are reported from the plants (Axtell 2013). The 23 nt sized siRNAs, which are not known to have antiviral activity were less abundant for PPSMV-1 compared to MYMIV, similar observations were made for the emaraviruses, AlsVX and WBYVV (Bi et al. 2012). Although the function of 23 nt sized siRNAs in plants is not known, for yeast and drosophila they are implicated in heterochromatin formation and regulation of RNA Polymerase II (Castel and Martienssen 2013).

The vsiRNA size distribution pattern for MYMIV contrasts with the pattern observed for PPSMV-1, although both viruses target the same host i.e. pigeonpea. In the case of MYMIV there was not much contrasting variation in the

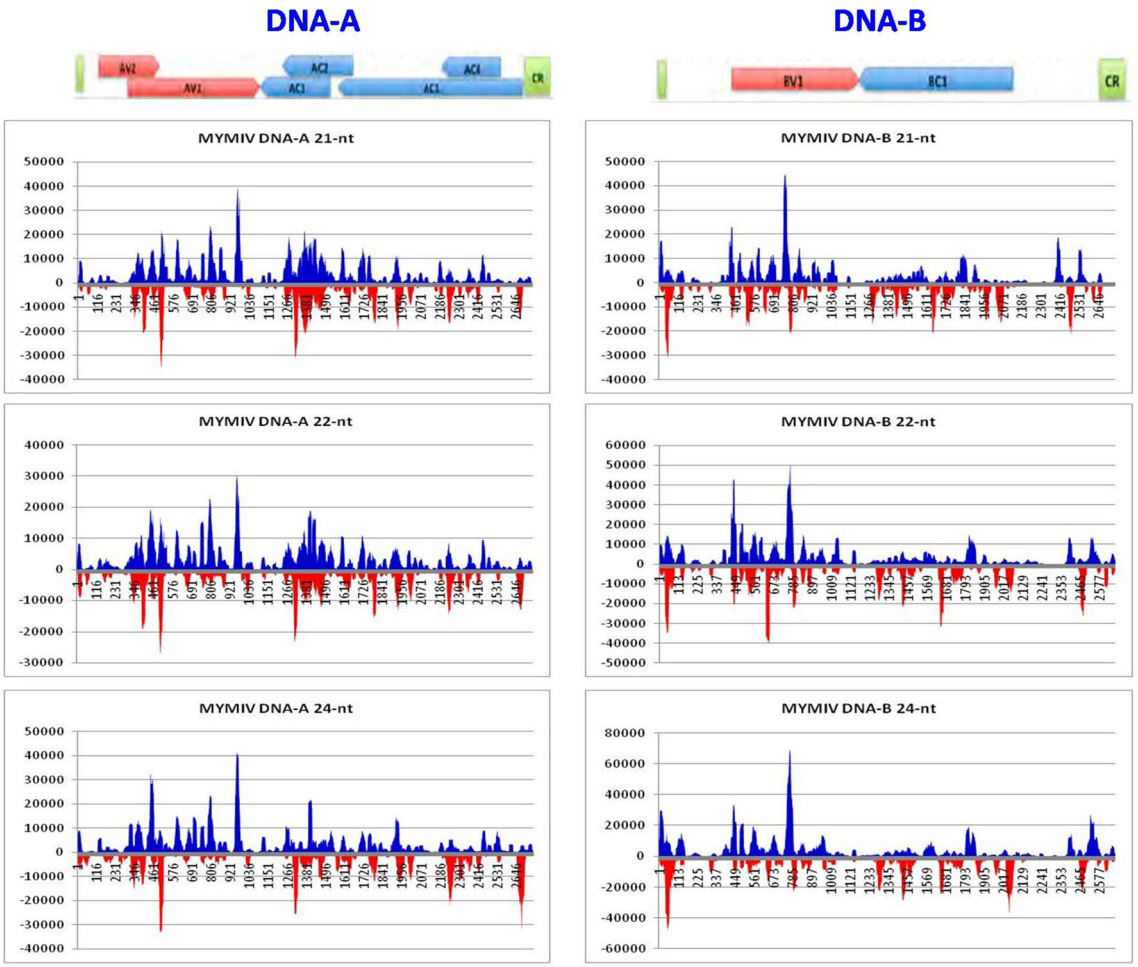


Fig. 4 Distribution maps three size classes of vsiRNAs, viz. 21 nt, 22 nt and 24 nt across the two genomic components (DNA-A and DNA-B) of MYMIV. The genome organization of the two MYMIV genomic components is schematically represented on the top. The vsiRNAs originating from the sense strand of the viral genome is shown in blue colour and those originating from the antisense strand in red colour. The number of reads per million (RPM) of the vsiRNAs is given on the *X* axis and the viral genomic coordinates are marked on the *Y* axis

levels of accumulation of 21, 22 and 24 nt sized siRNAs, when compared to PPSMV-1. In the case of MYMIV DNA-A component 21 nt sized siRNAs accumulated at relatively higher levels compared to 22 and 24 nt, whereas for DNA-B the 22 nt sized siRNAs occurred at relatively higher levels compared to 21 and 24 nt siRNAs. This may imply that there are certain structural and functional differences in the way the transcripts are processed from the two different DNA components of bipartite begomoviruses, to generate different size classes of siRNAs. Thus this study provides the first report on characterization of vsiRNAs derived from two distinct viruses infecting pigeonpea, namely PPSMV-1 an emaravirus with negative sense segmented RNA genome and MYMIV a geminivirus with circular DNA. These studies also provide further evidence that Ion Proton based high throughput sequencing technology could be used for characterization of virus derived small RNAs and also for virus diagnosis and reconstitution of hitherto unknown viral genome sequences.

Acknowledgements BLP acknowledges the research funding and DA acknowledges the young professional fellowship from ICAR-NRCPB. We thank Mr. Pravin Nilawe from Thermo Fisher Scientific for his help in bioinformatics analysis.

Author contributions BLP designed and carried out all the experiments and wrote the manuscript. DA helped in validation of PPSMV-1 sequence by RT-PCR.

Compliance with ethical standards

Conflict of interest Both authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Ameur A, Meiring TL, Bunikis I, Häggqvist S, Lindau C, Lindberg JH, Gustavsson I, Mbulawa ZZ, Williamson AL, Gyllensten U (2014) Comprehensive profiling of the vaginal microbiome in HIV positive women using massive parallel semiconductor sequencing. Sci Rep 4:4398

Assefa S, Keane TM, Otto TD, Newbold C, Berriman M (2009) ABACAS: algorithm-based automatic contiguation of assembled sequences. Bioinformatics 25(15):1968–1969

Axtell MJ (2013) Classification and comparison of small RNAs from plants. Annu Rev Plant Biol 64:137–159

Barba M, Czosnek H, Hadidi A (2014) Historical perspective, development and applications of next-generation sequencing in plant virology. Viruses 6(1):106–136

Bi Y, Tugume AK, Valkonen JP (2012) Small-RNA deep sequencing reveals *Arctium tomentosum* as a natural host of Alstroemeria virus X and a new putative Emaravirus. PLoS ONE 7(8):e42758

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

Breese MR, Liu Y (2013) NGSUtils: a software suite for analyzing and manipulating next-generation sequencing datasets. Bioinformatics 29(4):494–496

Briddon RW, Patil BL, Bagewadi B, Nawaz-ul-Rehman MS, Fauquet CM (2010) Distinct evolutionary histories of the DNA-A and DNA-B components of bipartite begomoviruses. BMC Evol Biol 10:97

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 78:7465–7477

Chevreux B, Wetter T, Suhai S (1999) Genome sequence assembly using trace signals and additional sequence information. Comput Sci Biol: Proc Ger Conf Bioinform (GCB) 99:45–56

Delcher AL, Phillippy A, Carlton J, Salzberg SL (2002) Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res 30(11):2478–2483

Deleris A, Gallego-Bartolome J, Bao JS, Kasschau KD, Carrington JC, Voinnet O (2006) Hierarchical action and inhibition of plant Dicer-like proteins in antiviral defense. Science 313:68–71

Donaire L, Barajas D, Martínez-García B, Martínez-Priego L, Pagán I, Llave C (2008) Structural and genetic requirements for the biogenesis of tobacco rattle virus-derived small interfering RNAs. J Virol 82:5167–5177

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

Elbeaino T, Digiaro M, Uppala M, Sudini H (2014) Deep sequencing of pigeonpea sterility mosaic virus discloses five RNA segments related to emaraviruses. Virus Res 188:27–31

Elbeaino T, Digiaro M, Uppala M, Sudini H (2015) Deep sequencing of dsRNAs recovered from mosaic-diseased pigeonpea reveals the presence of a novel emaravirus: pigeonpea sterility mosaic virus 2. Arch Virol 160(8):2019–2029

Fastx Toolkit. http://hannonlab.cshl.edu/fastx_toolkit/

Fletcher SJ, Shrestha A, Peters JR, Carroll BJ, Srinivasan R, Pappu H, Mitter N (2016) The tomato spotted wilt virus genome is processed differentially in its plant host *Arachis hypogaea* and its thrips vector *Frankliniella fusca*. Front Plant Sci 7:1349

Hagen C, Frizzi A, Gabriels S, Huang M, Salati R, Gabor B, Huang S (2012) Accurate and sensitive diagnosis of geminiviruses through enrichment, high-throughput sequencing and automated sequence identification. Arch Virol 157(5):907–915

Haible D, Kober S, Jeske H (2006) Rolling circle amplification revolutionizes diagnosis and genomics of geminiviruses. J Virol Methods 135(1):9–16

Hanley-Bowdoin L, Settlage SB, Orozco BM, Nagar S, Robertson D (2000) Geminiviruses: models for plant DNA replication, transcription, and cell cycle regulation. Crit Rev Biochem Mol Biol 35(2):105–140

Hema M, Sreenivasulu P, Patil BL, Lava Kumar P, Reddy DVR (2014) Tropical food legumes: virus diseases of economic importance and their control. In: Loebenstein G, Katis N (eds) Control of plant virus diseases: seed-propagated crops. Adv Virus Res 90(9), 431–505

Ho T, Wang H, Pallett D, Dalmay T (2007) Evidence for targeting common siRNA hotspots and GC preference by plant Dicer-like proteins. FEBS Lett 581:3267–3272

Idris A, Al-Saleh M, Piatek MJ, Al-Shahwan I, Ali S, Brown JK (2014) Viral metagenomics: analysis of begomoviruses by illumina high-throughput sequencing. Viruses 6(3):1219–1236

Jeske H (2009) Geminiviruses. Curr Top Microbiol Immunol 331:185–226

Johne R, Müller H, Rector A, van Ranst M, Stevens H (2009) Rolling-circle amplification of viral DNA genomes using phi29 polymerase. Trends Microbiol 17(5):205–211

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

Kumar S, Subbarao BL, Hallan V (2017) Molecular characterization of emaraviruses associated with Pigeonpea sterility mosaic disease. Sci Rep. 7(1):11831

Kuria P, Ilyas M, Ateka E, Miano D, Onguso J, Carrington JC, Taylor NJ (2017) Differential response of cassava genotypes to infection by cassava mosaic geminiviruses. Virus Res 227:69–81

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) 1000 Genome project data processing subgroup. The sequence alignment/map format and SAMtools. Bioinformatics 25(16):2078–2079

Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M (2012) Comparison of next generation sequencing systems. J Biomed Biotechnol 2012:251364

Matzke MA, Birchler JA (2005) RNAi-mediated pathways in the nucleus. Nat Rev Genet 6:24–35

Mielke-Ehret N, Mühlbach HP (2012) Emaravirus: a novel genus of multipartite, negative strand RNA plant viruses. Viruses 4:1515–1536

Mitter N, Koundal V, Williams S, Pappu H (2013) Differential expression of Tomato spotted wilt virus-derived viral small RNAs in infected commercial and experimental host plants. PLoS ONE 8:e76276

Niu D, Wang Z, Wang S, Qiao L, Zhao H (2015) Profiling of small RNAs involved in plant–pathogen interactions. Methods Mol Biol 1287:61–79

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 a Pinot Noir infected grapevine. Virology 408:49–56

Patil BL, Fauquet C (2015) Differential behaviour of the genomic components of cassava mosaic geminiviruses and the diversity of their small RNA profiles. Virus Genes 50:474–486

Patil BL, Kumar PL (2015) Pigeonpea sterility mosaic virus: a legume-infecting Emaravirus from South Asia. Mol Plant Pathol 16:775–786

Patil BL, Kumar PL (2017) Pigeonpea sterility mosaic emaravirus: a journey from a mysterious disease to a classic emaravirus. In: Mandal B, Rao GP, Baranwal VK, Jain RK (eds) A century of

plant virology in India. Springer, Berlin, pp 255–270 **(Chapter 10)**

Patil BL, Dangwal M, Mishra M (2017) Variability of emaravirus species associated with sterility mosaic disease of pigeonpea in India provides evidence of segment reassortment. Viruses 9(7):E183. https://doi.org/10.3390/v9070183

Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M, Hoon J, Simons JF, Marran D, Myers JW, Davidson JF, Branting A, Nobile JR, Puc BP, Light D, Clark TA, Huber M, Branciforte JT, Stoner IB, Cawley SE, Lyons M, Fu Y, Homer N, Sedova M, Miao X, Reed B, Sabina J, Feierstein E, Schorn M, Alanjary M, Dimalanta E, Dressman D, Kasinskas R, Sokolsky T, Fidanza JA, Namsaraev E, McKernan KJ, Williams A, Roth GT, Bustillo J (2011) An integrated semiconductor device enabling non-optical genome sequencing. Nature 475:348–352

Rudnick SI, Swaminathan J, Sumaroka M, Liebhaber S, Gewirtz AM (2008) Effects of local mRNA structure on posttranscriptional gene silencing. Proc Natl Acad Sci USA 105:13787–13792

Rumble SM, Lacroute P, Dalca AV, Fiume M, Sidow A, Brudno M (2009) SHRiMP: accurate mapping of short color-space reads. PLoS Comput Biol 5(5):e1000386. https://doi.org/10.1371/journal.pcbi.1000386

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(2):e88513

Singh NK, Gupta DK, Jayaswal PK, Mahato AK, Dutta S, Singh S, Bhutani S, Dogra V, Singh BP, Kumawat G, Pal JK, Pandit A, Singh A, Rawal H, Kumar A, Rama Prashat G, Khare A, Yadav R, Raje RS, Singh MN, Datta S, Fakrudin B, Wanjari KB, Kansal R, Dash PK, Jain PK, Bhattacharya R, Gaikwad K, Mohapatra T, Srinivasan R, Sharma TR (2011) The first draft of the pigeonpea genome sequence. J Plant Biochem Biotechnol 21:98–112

Tatineni S, McMechan AJ, Wosula EN, Wegulo SN, Graybosch RA, French R, Hein GL (2014) An eriophyid mite-transmitted plant virus contains eight genomic RNA segments with unusual heterogeneity in the nucleocapsid protein. J Virol 88:11834–11845

Vanitharani R, Chellappan P, Fauquet CM (2005) Geminiviruses and RNA silencing. Trends Plant Sci 10:144–151

Varshney RK, Chen W, Li Y, Bharti AK, Saxena RK, Schlueter JA, Donoghue MT, Azam S, Fan G, Whaley AM, Farmer AD, Sheridan J, Iwata A, Tuteja R, Penmetsa RV, Wu W, Upadhyaya HD, Yang SP, Shah T, Saxena KB, Michael T, McCombie WR, Yang B, Zhang G, Yang H, Wang J, Spillane C, Cook DR, May GD, Xu X, Jackson S (2012) Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource poor farmers. Nat Biotechnol 30:83–89

Voloudakis AE, Holeva MC, Sarin LP, Bamford DH, Vargas M, Poranen MM, Tenllado F (2015) Efficient double-stranded RNA production methods for utilization in plant virus control. Methods Mol Biol 1236:255–274

Wang MB, Masuta C, Smith NA, Shimura H (2012) RNA silencing and plant viral diseases. Mol Plant Microbe Interact 25(10):1275–1285

Xu Y, Huang L, Fu S, Wu J, 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(9):e46238

Yan F, Zhang HM, Adams MJ, Yang J, Peng JJ, Antoniw JF, Zhou Y, Chen J (2010) Characterization of siRNAs derived from rice stripe virus in infected rice plants by deep sequencing. Arch Virol 155:935–940

Yang X, Wang Y, Guo W, Xie Y, Xie Q, Fan L, Zhou X (2011) Characterization of siRNAs derived from the geminivirus/betasatellite complex using deep sequencing. PLoS ONE 6:e16928

Yuan Y, Xu H, Leung RK (2016) An optimised protocol and analysis of Ion Proton sequencing read for RNA-Seq. BMC Genom 17:403

Zhang C, Wu Z, Li Y, Wu J (2015) Biogenesis, function, and applications of virus-derived small RNAs in plants. Front Microbiol 6:1237

Zheng Y, Navarro B, Wang G, Wang Y, Yang Z, Xu W, Zhu C, Wang L, Serio FD, Hong N (2016) Actinidia chlorotic ringspot-associated virus: a novel emaravirus infecting kiwifruit plants. Mol Plant Pathol. https://doi.org/10.1111/mpp.12421