

siRNA Deep Sequencing and Assembly: Piecing Together Viral Infections

Jan Kreuze

Abstract RNA silencing constitutes a fundamental antiviral defense mechanism in plants in which host enzymes cut viral RNA into pieces of 20–24 nt. When isolated, sequenced en-mass and properly assembled or aligned these virus-derived small RNA (sRNA) sequences can reconstitute genomic sequence information of the viruses being targeted in the plant. This approach is independent of the ability to culture or purify the virus and does not require any specific amplification or enrichment of viral nucleic acids as it automatically enriches for small RNAs of viral origin by tapping into a natural antiviral defense mechanism. Using this technique known and novel DNA and RNA viruses as well as viroids have been identified at sensitivity levels comparable to PCR. This chapter will examine the strength and caveats of small RNA sequencing and assembly (sRSA), utilizing examples from literature as well as our own unpublished experiences and analysis of publically available plant sRNA sequence datasets.

Keywords Diagnostics • Small-RNA • Sequencing • Assembly • Virus • Viroid • Detection • Plants

Introduction

Next generation sequencing (NGS) technologies have led to a revolution in sequencing based applications in biological sciences. Viral diagnostics has not lagged behind with a number of NGS based approaches for pathogen identification emerging over the last 5 years. In principal NGS could enable universal diagnostics, allowing the identification of any pathogen in a sample based on its sequence. However even though throughput is ever increasing, sensitivity can be an issue

J. Kreuze (✉)

Laboratory of Virology, Peru International Potato Center (CIP),

Avenida la Molina, 1895 La Molina, Peru

e-mail: j.kreuze@cgiar.org

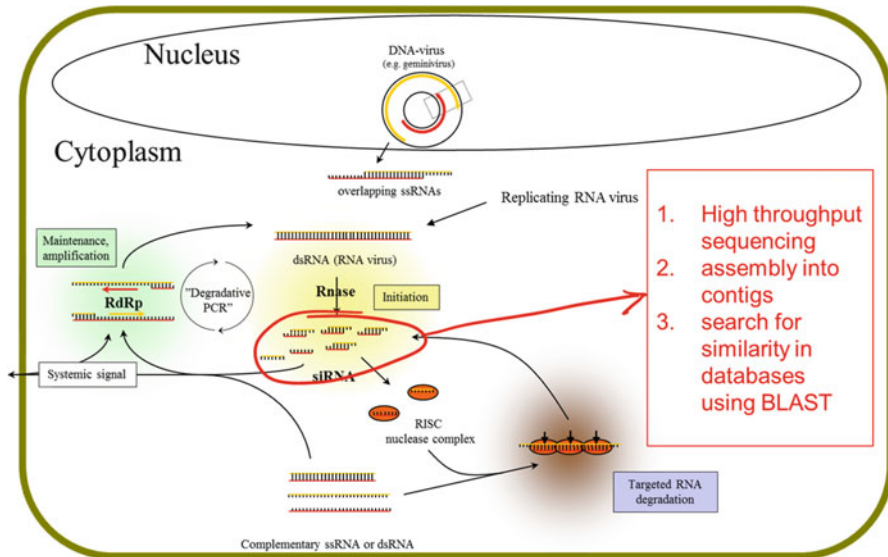


Fig. 1 Antiviral RNA silencing leads to the accumulation of virus derived siRNAs, which can be isolated, sequenced and assembled to posteriorly identify viruses by homology to known virus sequences

when low titer pathogens are involved, especially when taking price into consideration. While different sequencing platforms have been used, such as 454 and Illumina, which may influence results due to differences in sequence length and volume, key differences lie in type of nucleic acids that have been used for sequencing as well as methods to enrich for viral nucleic acids. The most straight forward approach is probably simply to sequence total mRNA and this approach has been successfully applied to identify a number of viruses (Wylie and Jones 2011; Wylie et al. 2012a, b, 2013; Al Rwahnih et al. 2009). Adams et al. (2009) used RNA from a healthy plant for subtractive hybridization, thereby enriching the RNA for pathogen sequences. Other studies sequenced purified double stranded RNA (dsRNA) (Al Rwahnih et al. 2011; Coetzee et al. 2010; Roossinck et al. 2010; Candresse et al. 2013) and a combination of both techniques has also been applied (Adams et al. 2013). NGS has also been combined with rolling circle amplification to efficiently identify circular DNA plant viruses (Hagen et al. 2012; Ng et al. 2011).

The most extensively utilized approach to date however relies on enrichment of viral sequences by the plant itself (Kreuze et al. 2009): small RNA sequencing and assembly (sRSA; Fig. 1). The approach exploits a natural anti-viral defence system called RNA silencing or RNA interference (RNAi), which is present in all eukaryotic organisms and generates 21–24 nt small RNA (sRNA) molecules homologous to viruses that may be infecting them (Mlotshwa et al. 2008). An additional benefit of this approach is that it makes ideal use of sequencing platforms that generate

enormous quantities, but short sequences. It has been successfully employed to identify many new plant viruses of all possible genome types (Kashif et al. 2012; Loconsole et al. 2012a, b; Bi et al. 2012; Wu et al. 2012; Zhang et al. 2011; Li et al. 2012; Hwang et al. 2013; Cuellar et al. 2011; De Souza et al. 2013; Fuentes et al. 2012; Kreuze et al. 2009, 2013; Untiveros et al. 2010; Sela et al. 2012; Roy et al. 2013), including viroids (Li et al. 2012; Wu et al. 2012) but has also been applied to identify viruses from worms, insects and even humans (Huang et al. 2013; Ma et al. 2011; Isakov et al. 2011; Wu et al. 2010).

sRSA: An Historic Overview

Viruses are both inducers and targets of RNA silencing, a fundamental antiviral defense mechanism in eukaryotic organisms (Mlotshwa et al. 2008; Haasnoot et al. 2007) (Fig. 1). RNA silencing is a cytoplasmic cell surveillance system that recognizes double-stranded RNA (dsRNA) and specifically eliminates RNAs homologous to the inducer RNA by cleavage using RNase III endonucleases called dicers. Plants encode several Dicer-like enzymes that recognize and cleave long dsRNA molecules to 21-, 22-, and 24-bp fragments that act as small interfering RNAs (siRNAs). siRNAs bind to ribonuclease H-like proteins in the RNA induced silencing complex (RISC) and are used to detect homologous single-stranded RNA (ssRNA) molecules for cleavage, producing more siRNAs. In plants, RNAi becomes amplified when the cleaved RNA recruits an RNA-directed RNA polymerase to generate more dsRNA, which is again cleaved by a dicer protein to produce secondary siRNAs, that are once again able to detect and cleave homologous RNA in a type of 'degradative PCR' cycle (Fig. 1). This leads to the accumulation of large amounts of siRNAs with homology to the invading virus. Also DNA viruses have been found to induce the production of copious amounts of siRNAs in plants, although the exact mechanism by which the system is initially triggered is not yet fully elucidated. Because accumulation of virus-derived siRNAs appeared to be a common feature of defense against viral infection in diverse eukaryotic hosts, and seemed particularly abundant in plants, using these defense molecules, which could be isolated based on their size, to identify viruses infecting a host had become a plausible concept (Kreuze et al. 2009). Next generation sequencing platforms generating massive amounts, but rather short reads such as Illumina (Solexa) and SOLiD were rapidly developing, and with them a set of bioinformatics tools for *de novo* assembly of such short reads. However, even if these tools had been developed for short reads, they were generally longer than the 21–24 nucleotides of siRNAs, and it was not clear how well they would function using reads of that size. It was also unclear to what extent viral siRNAs overlap and thus whether they were suitable for the assembly into larger contigs that could be used for virus detection. Our study using experimentally infected sweetpotatoes (Kreuze et al. 2009) provided a positive answer to both questions since all three assembly tools tested (Velvet, VCAKE and SSAKE) were able to assemble viral

contigs from five different viruses efficiently, even if a full genome could only be assembled *de novo* from one of the two RNA viruses (a potyvirus) and only after manual re-assembly. Although only relatively short contigs could be assembled for the other RNA virus (crinivirus), guide strand mediated assembly by alignment using MAQ and using the siRNA reads to the genome of a related virus isolate was able to assemble 98 % of the genome of that virus as well. A notable aspect of that study was that besides the two expected RNA viruses a number of additional contigs corresponding to DNA viruses (two strains of a new badnavirus and a mastrevirus) were identified and further confirmed by PCR and Sanger sequencing. Although we have subsequently been able to confirm that the badnaviruses are not genome-integrated and can be graft transmitted (Kreuzer, unpublished), identification of sequences related to para-retroviruses may not always signify infection by a virus as they may represent silencing of endogenous integrated virus sequences (see section “[Reverse transcribing viruses](#)” of this chapter below).

Soon afterwards the principal was demonstrated in invertebrate animals by Wu et al. (2010), who were able to assemble a known virus and five previously described viruses from nematode, fruit fly and mosquitoes. Later Isakov et al. (2011) identified Human immunodeficiency virus (HIV) and a mycoplasma in human cells through the application of sRSA. Each of these studies identified unexpected novel pathogens underlining not only the ability to detect infectious agents without a priori knowledge, but also the fact that organisms considered to be healthy, may not always be virus free, including well studied cell lines.

Several studies have followed characterizing plant viruses with various types of genomes, including single- (Loconsole et al. 2012b; Kashif et al. 2012), and double- (Kashif et al. 2012; Zhang et al. 2011) stranded DNA genomes, positive- (Bi et al. 2012; Cuellar et al. 2011; De Souza et al. 2013; Fuentes et al. 2012; Kashif et al. 2012; Kreuze et al. 2013; Li et al. 2012; Loconsole et al. 2012a; Untiveros et al. 2010), negative- (Bi et al. 2012) and double- (Sela et al. 2012) stranded RNA genomes, and viroids (Li et al. 2012; Wu et al. 2012).

Whereas these studies indicate the broad applicability of sRSA among virus species, they are still relatively limited in the range of plant species to which sRSA has been applied (principally dicotyledonous plants) and a more extended analysis across different taxonomic groups of the plant kingdom would be beneficial.

General Lessons Regarding sRSA: Analysis of siRNA Sequence Datasets from Across the Plant Kingdom

To explore the efficiency of sRSA to identify viruses on a broad range of plant species we examined published plant small RNA sequence datasets available on the internet. The website of the project “comparative sequencing of plant small RNAs” at <http://smallrna.udel.edu/> contains siRNA sequence data from plants throughout the plant kingdom (Table 1). We included into this analysis also three samples of

Table 1 Virus like sequences identified using siRNA datasets from 36 different species

	Source plant (common name)	Tissue	Contigs found (# and minimum e-value) ^a	Assembly ^b
<i>Eudicots</i>	Cucurbita maxima (pumpkin)	Leaves, phloem sap	–	–
	Populus trichocarpa (poplar)	Leaves	Caulimovirid (6; $\geq 9\text{e-}11$)	54–144
	Gossypium arbore- tum (tree cotton)	Leaves, flowers, boll fiber	Totivirus (12, 0, 3; $\geq 5\text{e-}17$)	51–204
	Citrus sinensis (sweet orange)	Leaves	Caulimovirid (11; $\geq 3\text{e-}07$)	54–144
	Cairica papaya (papaya) [PRSV] ^c	Leaves	PRSV	99.6 %
		Leaves	Partitivirus (8, 9, 0; $\geq 7\text{e-}19$)	57–210
		Flowers	Totivirus (2, 0, 3; $\geq 2\text{e-}5$)	54–99
	Silene latifolium (white campion)	Smutted flowers	Chrysovirus (0, 0, 6; $\geq 1\text{e-}10$) ^d	67–198
			Caulimovirid (2, 4, 0; $\geq 4\text{e-}12$)	57–102
	Physalis floridiana (Physalis) ^c	Leaves	TVCV-like (4; ≥ 0.012)	32–50
	Petunia hybrida (petunia)	Leaves	–	–
	Nicotiana tabaccum (tobacco)	Leaves	TVCV-like (23; $\geq 2\text{e-}10$)	42–198
	N. benthamiana [SB-29] ^{d,e}	Leaves	Torrado-like virus (5; $\geq 6\text{e-}5$)	59–119
	Capsicum annum (sweet pepper)	Leaves, flowers, fruits	BPeV (29, 6, 9; $\geq 1\text{e-}35$)	48–267
			TVCV-like (29, 11, 20; $\geq 5\text{e-}14$)	42–240
	Solanum lycopersicum (tomato)	Leaves	TVCV-like (10; $\geq 4\text{e-}8$)	45–111
	S. tuberosum (potato) [PVT] ^{d,e}	Leaves	PVT	98.6 %
			TVCV-like (9; $\geq 4\text{e-}8$)	42–105
	Lactuca sativa (lettuce)	Leaves, B. lactucea inoculated leaves, flowers	LBVaV RNA1-2 (56, 0, 3)	97.2 % (96.4–98)
			MiLV RNA1-4 (33, 0, 2)	93.3 % (85.6– 99.6)
			BPYV (0, 0, 42)	96.9 % (96.1– 97.8)
			TICV (0, 0, 14)	86.4 % (87.8– 84.9)
			Badnavirus (1, 2, 17; $\geq 4\text{e-}13$)	

(continued)

Table 1 (continued)

	Source plant (common name)	Tissue	Contigs found (# and minimum e-value) ^a	Assembly ^b
<i>Magnolid</i>	Mimulus guttatus (Spotted mon- key flower)	Leaves	–	–
	Persea americana (avocado)	Leaves	–	–
	Zostera marina (sea grass)	Leaves	–	–
<i>Monocots</i>	Musa Acuminata (banana)	Leaves	Badnavirus (251; ≥8e-31)	42–207
	Oryza sativa (rice)	Leaves	RRSV RNA1-10	99.1 % (98.5– 99.7)
	Triticum aestivum (wheat)	Leaves	–	–
	Hordeum vulgare (barley)	Leaves	–	–
	Panicum virgatum (switch grass)	Leaves	–	–
	Setaria italica (fox- tail millet)	Leaves	–	–
	Zea mays (maiz)	Leaves	–	–
	Sorghum bicolor (Sorghum)	Leaves	–	–
	Miscanthus giganteus (Miscanthus)	Leaves	–	–
	Nuphar advena (water lilly)	Leaves, roots, flowers	CMV RNA1-3 Cytorhabdovirus (27,9,9; ≥9e-21)	99.5 % ^f 42–228
			Caulimovirid (23,16,8; ≥6e-33)	42–315
	Amborella trichopoda (amborella)	Leaves	Caulimovirid (15; ≥2e-23)	51–228
	Picea abies (Norway spruce)	Leaves	–	–
<i>Gymnosperms</i>	Ginko biloba (maidenhair tree)	Leaves	–	–
	Cycas rumphii (Cycas)	Leaves	Caulimovirid (1; ≥4e-4)	69
	Marsilea quadrifolia (European water clover)	Leaves, roots, draught treated leaves	TuYV	99.6 % ^g

(continued)

Table 1 (continued)

	Source plant (common name)	Tissue	Contigs found (# and minimum e-value) ^a	Assembly ^b
<i>Non-seed plants</i>	Physcomitrella patens		Caulimovirid (4; $\geq 1\text{e-}3$)	60–102
	Chara coralline (common stonewort)	Thallus	Caulimovirid (12; $\geq 4\text{e-}6$)	39–120
<i>Algae</i>	Chlamydomonas reinhardtii (chlamydo monas)	Cells	–	–
	Volvox carteri (volvox)	Cells	–	–

^aNumber of contigs found, when more than one number is provided they corresponds to various tissues in the same order as named in the tissue column. Number of contigs or e-value not given when near complete genomes were obtained. –: no virus detected
^bSize range of contigs obtained or % coverage of complete genome (in case of multiple genome components, range of coverage is given between *brackets*)
^cPlant deliberately infected with virus between *square brackets*
^dChrysovirus contigs were only found in smutted leaves, while other viruses were found in all tissues
^eData from our own study
^fExcluding 16–150 nt and 81–90 nt missing from 5' and 3' ends respectively
^gExcluding 673 nt deletion and ~100 nt missing from 5' end

our own: healthy *Physalis foridiana*, *Potato virus T* (PVT, *Flexiviridae: Tepovirus*) infected potato (*Solanum tuberosum*), and *Nicotiana benthamiana* infected with an uncharacterized virus SB-29. siRNA sequences assembled by Velvet for each of these datasets were submitted to BLASTX against the GenBank database, limiting the search to sequences of viral origin. For samples where known viruses were identified, full genomes of the most related sequence were downloaded and used for guide strand mediated assembly using MAQ (Fig. 2). The sequences and e-values of hits with significant similarity to viruses were extracted and are summarized in Table 1. Contig sequences are available from <https://research.cip.cgiar.org/confluence/display/cpx/RsA+Plant+Kingdom>. Analysis of this data provided a number of general lessons regarding BLASTX search results against the virus sequence database.

Reverse Transcribing Viruses

In most cases contigs with low level similarity (E values > 0.1) to animal retro- and pararetroviruses were found. These are unlikely to be viruses, but rather may represent unspecific hits, or sequences corresponding to degenerated or

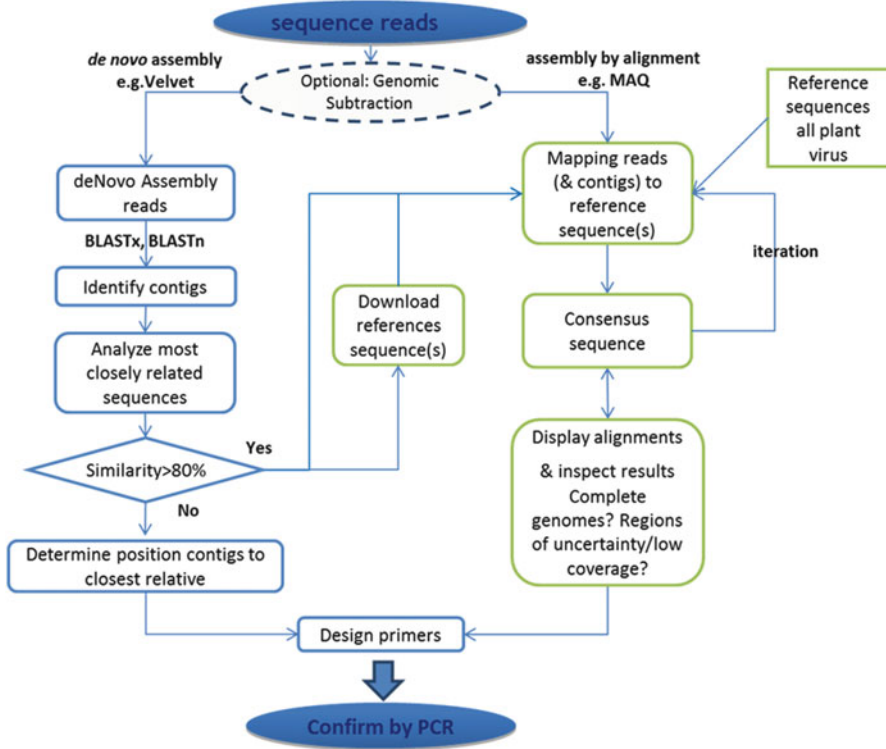


Fig. 2 General workflow for identifying viruses using sRSA performed in our lab

uncharacterized retro-elements in the plant genome. Furthermore, in many cases also contigs with significant similarity to viruses of the family *Caulimoviridae* were identified (Table 1). Although these may represent replicating viruses, they could also be the result of endogenous pararetrovirus like sequences or retrotransposons. Indeed, the contigs found in the moss *Physcomitrella patens* and tree *Populus trichocarpa* (Poplar) corresponded 100 % to genomic sequences when blasted against their respective genomes. Whereas in the case of *P. patens* the corresponding genomic fragment was annotated as an LTR retrotransposon (Genbank: GQ294565.1), the *P. trichocarpa* caulimovirus-like sequences corresponded to several predicted proteins, e.g. on scaffold_2615 (Genbank: NW_001490516.1), which contained two predicted ORFs corresponding to the coat protein and reverse transcriptase in the same gene order as found in viruses of the genera *Caulimovirus* and *Sobemovirus*. Numerous *Banana streak virus* (BSV, *Caulimoviridae*: *Badnavirus*) like sequences were identified in Banana (*Musa acuminata*, Table 1), a plant which is well known to contain endogenous BSV forms (Geering et al. 2005). In addition, contigs with high level of similarity to *Tobacco vein clearing virus* (TVCV; *Caulimoviridae*: *Cavemovirus*), were identified in all plants from the family *Solanaceae*, except petunia (*Petunia hybrida*).

Again, these are unlikely to represent replicating virus, but rather are genome integrated elements which have previously been shown to occur in potato (Hansen et al. 2005). This was confirmed by BLAST searches of the TVCV like contigs against the potato genome (www.potatogenome.net), which resulted in many hundreds of hits. Because TVCV like sequences have been reported also from other solanaceous plants such as tomato (Staginnus et al. 2007) and various tobaccos (Lockhart et al. 2000; Gregor et al. 2004; Jakowitsch et al. 1999) these sequences may represent an ancient integration event maintained during the evolution within this plant family. Using primers designed to the contigs found in *P. floridana*, we were indeed able to amplify a 1,651 bp fragment encompassing partial ORF3 and ORF4 of a TVCV like sequence, which however contained numerous stop codons, suggesting it was indeed an inactive integrated sequence. On the other hand, no siRNA contigs with similarity to *Petunia vein clearing virus* (PVCV; *Caulimoviridae: Petuvirus*), or endogenous rice tungro bacilliform virus (ERTBV) were found in petunia or rice (*Oryza sativa*) respectively, although they have been shown to be integrated into genomes of these species (Kunii et al. 2004; Richert-Poggeler et al. 2003). This suggests that such sequences do not always lead to production of siRNAs, irrespective if they can be activated (PVCV) or not (ERTBV). Together these results show that pararetrovirus like sequences are common throughout the plant kingdom including algae, but that the identification of pararetrovirus like siRNA contigs does not unambiguously signify viral infection. Conversely, lack of pararetrovirus like contigs may not necessarily signify absence of integrated, or even activatable, pararetroviral sequences. Whereas the presence of stop codons in the translated reading frame of assembled contigs may suggest that inactive integrated viruses are present, it doesn't exclude extra-chromosomal viral replicons, and further analyses will be required to verify each particular case, particularly the first time a plant species is analyzed using this method. Availability of host genome sequences may help alleviate these problems through genomic subtraction (see section "[Bioinformatics](#)").

Known Viruses

Beyond contigs corresponding to reverse transcribing viruses, five new putative viruses as well as eight known viruses were identified among the 34 species downloaded (Table 1).

After identifying contigs with high similarity (>90 %) to known viruses in Papaya (*Papaya ring spot virus*, PRSV; *Potyviridae: Potyvirus*), Lettuce (*Mirafiori lettuce virus*, MiLV, *Ophioviridae: Ophiovirus*; *Lettuce big-vein associated virus*, LBVaV, *Varicosavirus*; *Beet pseudo yellows virus*, BPYV & *Lettuce infectious yellows virus*, LIYV, *Closteroviridae: Crinivirus*), Rice (*Rice ragged stunt virus*, RRSV; *Reoviridae: Oryzavirus*), Water lily (*Cucumber mosaic virus*, CMV, *Bromoviridae: Cucumovirus*), and European waterclover (*Turnip yellows virus*,

TuYV, *Luteoviridae: Polerovirus*), their complete sequences were downloaded and used for guide strand assisted assembly with MAQ and/or Novoalign. Nearly complete genome sequences could thus be assembled by combining *de novo* and guide strand mediated assembly (Table 1). It was notable that in most cases, assembly of 3' and 5' un-translated regions (UTRs) was poor and thus both UTRs of the CMV, MiLV and the 5' UTR of TuYV were partly or entirely missing from the assemblies. Our experience is that viral UTRs are often, but not always, poor sources of siRNA as compared to coding regions and may reflect the evolution of these sequences to avoid degradation/targeting by RNA silencing. Also, in the multipartite viruses MiLV and RRSV, clear differences in coverage could be observed between different genome components, possibly reflecting their relative abundances in the plant. Whereas CMV, MiLV, LBVaV and RRSV were >98 % identical to published genomes, TuYV represented a new strain (Mq) of the virus as it had only ~90 % nt identity to the published sequence and also contained a 673 nt deletion eliminating the N-terminal 161 amino acids from the CP read-through cistron. The PRSV isolate showed the highest level of similarity (96 %) to isolates of Hawaii, where it was isolated (Emanuele De Paoli; personal communication). Comparison of the assemblies of the same virus from different tissues in the case of LBVaV, MiLV, CMV and TuYV, yielded practically (>99 %) identical consensus sequences where bases had been called, even though coverage varied significantly between tissues, e.g. the average number of unique siRNAs corresponding to TuYV were 18.9, 57.4 and 34 per 1,000 reads in leaf, root or draught stressed leaf samples, respectively.

New Viruses

Several new RNA viruses were also putatively predicted in some samples: a new totivirus from tree cotton, a new cytorhabdovirus from water lily, and three new viruses from white campion: another totivirus, a partitivirus and a chrysovirus. Chrysoviruses are typically fungal viruses and accordingly contigs with homology to chrysoviruses were only detected from smut infected flowers, and not healthy flowers or leaves, suggesting that the siRNAs originated from the smut pathogen (*Microbotryum violaceum*, Fuckel). Assuming this is the case, it implies that the method also works for fungi, as was previously predicted (Kreuze et al. 2009). The availability of data from different plant organs also enabled us to verify if the same viruses could be detected in different tissues. This was not always the case for the novel viruses when using BlastX searches of *de novo* assembled contigs (Table 1), even though MAQ was able to identify siRNAs homologous to viral contigs assembled from other tissues of the same plant, indicating virus specific siRNA were present in these tissues. This discrepancy is most likely due to that siRNAs were present at a too low frequency for efficient *de novo* assembly of contigs that were long enough to provide significant hits to the distantly related virus sequences available. Similar problems were indeed not experienced with the

known viruses, which were readily detectable from any tissues with either BlastX or BlastN. The only exceptions were BPYV and LIYV, which could only be detected in the flower samples but they were the result of siRNA libraries taken from different plants (which were probably infected with additional viruses) than the leaf samples (Emanuele de Pauli, personal communication).

Although we are unable to confirm the results of this analysis because the data was generated by other scientist, and the original samples were not available to us, the analysis has been useful in highlighting both the scope of its applicability among plant species and tissue types, but also the challenges that are still faced, i.e., dealing with short contigs resulting from low coverage and how to distinguish integrated sequences from replicating viruses. The following section explores some of these limitations and the approaches that have been developed to improve them.

Factors Affecting the Efficacy of sRSA

Several factors can influence the efficiency of sRSA, and range from the sample preparation method to issues involving the bioinformatics analysis. Effects of sample preparation include method of RNA isolation, sample bulking and the siRNA cloning method. Although total RNA extraction kits can be used (Roy et al. 2013), most standard column based RNA extraction methods do not efficiently isolate small RNAs. Specialized kits for isolating miRNA can be used (Hwang et al. 2013; Sela et al. 2012), but alternatives such as Trizol (Kashif et al. 2012; Bi et al. 2012; Kreuze et al. 2009; Untiveros et al. 2010; De Souza et al. 2013; Fuentes et al. 2012) and CTAB (our lab, unpublished) protocols followed by gel purification of sRNAs have both proven to be adequate depending on the plant species and tissue type to be used. siRNA preparation steps prior to sequencing, which include sequential steps of adapter ligations and PCR enrichment can be fully outsourced to sequencing providers or performed in-house to save costs (e.g., Chen et al. 2012).

Effect of siRNA Cloning Method on Sequence Distribution and Assembly

Analysis of data from our first sRSA experiments (Kreuze et al. 2009) indicated that siRNA coverage of viral genomes was very unevenly distributed with short regions of extremely high coverage separated by regions with relatively low coverage (Fig. 3B). Linsen et al. (2009) showed that the method of siRNA library preparation could introduce a strong and consistent bias towards certain sequences depending on the method used and we suspected this was the case for our data as well. Uneven distribution of reads could potentially inhibit efficient *de novo* assembly of siRNAs.

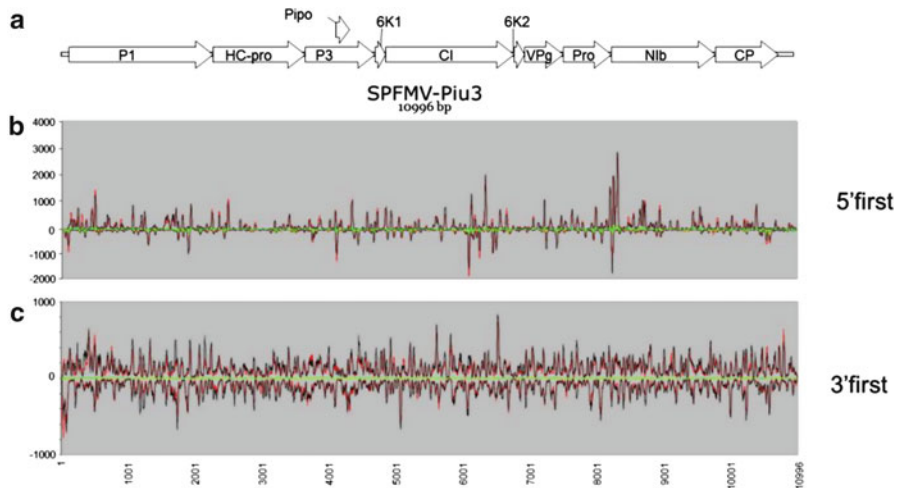


Fig. 3 Coverage of the sweet potato feathery mottle virus (SPFMV) genome by siRNAs following Illumina sequencing of the same samples using two different sample preparation techniques. (a) SPFMV genome structure. (b) Graph showing coverage for a library prepared by ligating the 5' adaptor first. (c) Graph showing the coverage for a library prepared by ligating the 3' adaptor first. The *lines* above and below the graph indicate the coverage of siRNAs corresponding to the positive or negative strand of the virus. The *different colors* of the *lines* indicate the different siRNA sizes: 21 nt black; 22 nt red; 23 nt green; 24 nt yellow

Therefore we decided to explore the effect of reversing the order of RNA adapter ligation in the sample preparation process of the exact same RNA extract from virus infected sweetpotato used previously (Kreuze et al. 2009). Whereas in the previous study the 5' adaptor was ligated first followed by the 3' adaptor, this time the 3' adaptor was ligated first followed by the 5' adaptor. Results showed that this simple change of procedure resulted in a much more homogenous distribution of siRNA reads over all five viral genomes analyzed and in all three samples (Fig. 3). Indeed Velvet was able to generate much larger viral contigs from each of the three samples and was able to assemble the whole genome of SPFMV without the need for manual re-assembly.

siRNA library preparation protocols have since been improved and currently recommended protocols are probably close to optimum. Nevertheless improvements may still be possible through adaptor and primer design, and should be kept in mind when choosing or developing in-house protocols.

Mixed Infections, Defective RNA/DNAs and Contamination

To make cost effective use of the current high throughput of NGS sequencers, samples are bulked and may be individually tagged by including a short DNA

“barcode” or index during sample preparation. After a sequencing run each sample can then be disaggregated from the bulk based on their DNA index using computational algorithms. Care must be taken when preparing multiple samples as experience has shown the risk of contamination is significant and even minute amounts of cross sample contaminations can lead to assembly of significant portions of a virus genome, albeit mostly at low coverage. If siRNA libraries are prepared in-house it is recommended to always leave two empty lanes between each sample when running gel purifications. Still a recent study suggests ‘virtual contamination’ between indexed samples may occur at a rate of about 0.3 % during the computational processing of indexed libraries, and suggest double indexing as an efficient solution (Kircher et al. 2012).

Besides the issue of physical or virtual contaminations our experiences with sweetpotato field samples are that mixed infection with related viral strains are common. This is easily recognized by the assembly of contigs of the same region but differing in nucleotide sequence. Whereas this doesn’t prevent virus identification, it makes assembly of reliable full genomes of each individual strain extremely difficult. Particularly when stretches of more than 20 nt with perfect identity exist between the two sequences, it may be difficult to exclude *in silico* recombination. Another confounding factor may be the presence of defective interfering (DI) DNA or RNA molecules, which usually replicate to much higher levels than the virus itself and produce large amounts of siRNA (Wu et al. 2010), which can confuse assembly software.

Bioinformatics

From the beginning, principally two different bioinformatics approaches have been followed to assemble viral genomes from sRNA sequences: *de novo* assembly of siRNAs (e.g. using Velvet), and alignment/mapping of sequences (e.g. with MAQ, BWA, NovoAlign). Often, the two are used in combination in an iterative process of *de novo* assembly and alignment to come to a final consensus (Fig. 2). Nevertheless, full or near complete virus genomes are often only achieved by assembly of siRNAs and or partial, *de novo* assembled, contigs onto a closely related reference genome. Vodovar et al. (2011), using *Flock house virus* as a model, developed a script, named Paparazzi, that reliably reconstitutes viral genomes through an iterative alignment/consensus call procedure using a related reference sequence as a scaffold. Still when nucleotide similarity drops below ~80 % efficiency and reliability of alignments of siRNAs are significantly reduced. Thus *de novo* assembly of siRNAs to achieve a complete genome for viruses where reference genomes with sufficient similarity are not available can be difficult.

One of the reasons limiting the ability of short sequence assemblers to generate full virus genomes *de novo*, may be interference in the assembly process by endogenous small RNA sequences, which are abundant and extremely polymorphic, often outnumbering viral siRNAs by orders of magnitude. Because

mammalian antiviral siRNA are much less abundant than other eukaryotic organisms Isakov et al. (2011) used sRSA together with the concept of short RNA subtraction (or genomic subtraction) to address the problem of interference by host sRNAs. The method boosts pathogen-related siRNA signal through subtraction of host-derived sRNAs and led to successful identification of HIV in infected human cell cultures. Li et al. (2012) applied genomic subtraction to identify *Pepino mosaic virus* (Potyviridae: Potyvirus) and *Potato spindle tuber viroid* (Pospiviroidae: Pospiviroid) in tomato which resulted in a several-fold increase in the length of viral contigs and large reduction in the number of non-viral assembled contigs, as compared to assembly without genomic subtraction. Our own unpublished experiences with various viruses in potato also confirmed the increased assembly efficiency resulting from genomic subtraction, and it therefor seems recommendable to apply genomic subtraction whenever a genome sequence is available. However, care needs to be taken when perform genomic subtraction as some plant genome assemblies contain sequences of viral contaminations, besides endogenous integrated virus sequences.

One of the limitations for detecting new viruses is the need for sequences with a minimum level of homology in the database to achieve recognition. BlastX is efficient at identifying more distantly related sequences, but the success of identification of distantly related sequences will rapidly decrease with reduced contig size and for non-coding regions such as 5' and 3' UTRs, highlighting the benefit of assembling large contigs *de novo*. Still, considering the genetic diversity of known viruses, it is likely that there are still undiscovered viruses that exhibit no similarity to any of the currently known genera and families and these would escape detection by the current homology-dependent approaches. Genomic subtraction addresses this limitation to some extent, because any sequences remaining after subtraction of the host genome should automatically be considered as suspicious. One must however keep in mind that genomic subtraction should not be 100 % efficient, as the genomes of plant genotypes being analyzed may vary significantly from published genomes, particularly in heterozygous and polyploidy crops. For example, in potato the efficiency of genomic subtraction using the DM potato genome is only about 40–50 % when applied to other genotypes (nevertheless an improvement in *de novo* assembly of viral contigs is often observed). In addition, due to the high complexity of plant genome especially in centromere regions, currently no plant genomes have been sequenced and assembled to 100 % coverage. Wu et al. (2012) took another route to address the problem of homology requirement by developing a homology independent computational algorithm to identify replicating circular RNAs (referred to as progressive filtering of overlapping small RNAs: PFOR) and demonstrated its efficiency in identifying known viroids, but also a new viroid with no significant similarity to any previously known molecules. This approach seems promising, not only to identify viroids, but potentially also unknown circular DNA viruses. However, this approach requires the complete (100 %) coverage of the virus genome by siRNAs to ensure the identification of the circular structure of the genome and in our hands we were unable to recover the sequences of *Grapevine yellow speckle viroid 1* (Pospiviroidae: Apiscaviroid) and *Hop stunt viroid*

(*Pospiviroidae: Hostuviroid*) from an infected grapevine, or *Sweet potato leaf curl virus* (*Geminiviridae: Begomovirus*) from infected sweetpotato plants using PFOR, whereas they were readily identified by assembly and BLAST as well as alignment with MAQ.

Concluding Remarks

Evidenced by the number of reports during the past several years sRSA has proven its power to identify and sequence diverse sets of viral and sub-viral agents including many new ones. The mining of published plant siRNA datasets presented in this chapter further expand on that, confirming its broad applicability but also revealing the remaining challenges. For sRSA to emerge from the research realm and become a mainstream generic viral detection method however important improvements are still required, particularly on the side of the sequence analysis. Whereas different bioinformatics solutions have been presented to address specific situations, they don't yet present a unified solution addressing each of the different issues. Development of automated and universally applicable bioinformatics pipelines for analysis of sRNA data, specifically for virus identification, should be a priority for future development and are required before the method can become part of regular viral diagnostics.

Acknowledgments I acknowledge the contributions from the various present and previous members of my lab and colleagues at CIP, and Emanuele De Paoli and Manfred Ruddat for sharing information about the samples from their small RNA libraries. Sincere thanks also to Zhangjun Fei for critical reading and useful comments of this chapter. The work on sRSA in our lab has been funded by grants from the Howard Buffet Foundation (subgrant 20127-C), Bill and Melinda Gates Foundation through the SASHA project (Grant OPP53344), the National Science Foundation through its BREAD program (Award No. 1110080), ICGEB-TWAS-UNESCO/IBSP Joint Programme on Capacity Building in Basic Molecular Biology (Contract No.: CRP/101005) and EU FP7 through the QBOL project (Grant No. 226482).

References

- Adams IP, Glover RH, Monger WA, Mumford R, Jackeviciene E, Navalinskiene M, Samuitiene M, Boonham N (2009) Next-generation sequencing and metagenomic analysis: a universal diagnostic tool in plant virology. *Mol Plant Pathol* 10(4):537–545. doi:[10.1111/j.1364-3703.2009.00545.x](https://doi.org/10.1111/j.1364-3703.2009.00545.x)
- Adams IP, Miano DW, Kinyua ZM, Wangai A, Kimani E, Phiri N, Reeder R, Harju V, Glover R, Hany U, Souza-Richards R, Nath PD, Nixon T, Fox A, Barnes A, Smith J, Skelton A, Thwaites R, Mumford R, Boonham N (2013) Use of next-generation sequencing for the identification and characterization of maize chlorotic mottle virus and sugarcane mosaic virus causing maize lethal necrosis in Kenya. *Plant Pathol* 62(4):741–749. doi:[10.1111/j.1365-3059.2012.02690.x](https://doi.org/10.1111/j.1365-3059.2012.02690.x)

- 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(2):395–401. doi:[10.1016/j.virol.2009.02.028](https://doi.org/10.1016/j.virol.2009.02.028)
- Al Rwahnih M, Daubert S, Urbez-Torres JR, Cordero F, Rowhani A (2011) Deep sequencing evidence from single grapevine plants reveals a virome dominated by mycoviruses. *Arch Virol* 156(3):397–403. doi:[10.1007/s00705-010-0869-8](https://doi.org/10.1007/s00705-010-0869-8)
- 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. doi:[10.1371/journal.pone.0042758](https://doi.org/10.1371/journal.pone.0042758)
- Candresse T, Marais A, Faure C, Gentil P (2013) Association of little cherry virus 1 (LChV1) with the Shirofugen stunt disease and characterization of the genome of a divergent LChV1 isolate. *Phytopathology* 103(3):293–298. doi:[10.1094/phyto-10-12-0275-r](https://doi.org/10.1094/phyto-10-12-0275-r)
- Chen Y-R, Zheng Y, Liu B, Zhong S, Giovannoni J, Fei Z (2012) A cost-effective method for Illumina small RNA-Seq library preparation using T4 RNA ligase 1 adenylated adapters. *Plant Methods* 8:4
- Coetzee B, Freeborough MJ, Maree HJ, Celton JM, Rees DJ, Burger JT (2010) Deep sequencing analysis of viruses infecting grapevines: virome of a vineyard. *Virology* 400(2):157–163. doi:[10.1016/j.virol.2010.01.023](https://doi.org/10.1016/j.virol.2010.01.023)
- Cuellar WJ, Cruzado RK, Fuentes S, Untiveros M, Soto M, Kreuze JF (2011) Sequence characterization of a Peruvian isolate of sweet potato chlorotic stunt virus: further variability and a model for p22 acquisition. *Virus Res* 157(1):111–115. doi:[10.1016/j.virusres.2011.01.010](https://doi.org/10.1016/j.virusres.2011.01.010)
- De Souza J, Fuentes S, Savenkov S, Cuellar W, Kreuze J (2013) The complete nucleotide sequence of sweet potato C6 virus: a carlavirus lacking a cysteine-rich protein. *Arch Virol* 158(6):1393–1396. doi:[10.1007/s00705-013-1614-x](https://doi.org/10.1007/s00705-013-1614-x)
- Fuentes S, Heider B, Tasso RC, Romero E, Zum Felde T, Kreuze JF (2012) Complete genome sequence of a potyvirus infecting yam beans (*Pachyrhizus* spp.) in Peru. *Arch Virol* 157(4):773–776. doi:[10.1007/s00705-011-1214-6](https://doi.org/10.1007/s00705-011-1214-6)
- 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(Pt 2):511–520. doi:[10.1099/vir.0.80261-0](https://doi.org/10.1099/vir.0.80261-0)
- Gregor W, Mette MF, Staginnus C, Matzke MA, Matzke AJM (2004) A distinct endogenous pararetrovirus family in *Nicotiana tomentosiformis*, a diploid progenitor of polyploid tobacco. *Plant Physiol* 134(3):1191–1199. doi:[10.1104/pp.103.031112](https://doi.org/10.1104/pp.103.031112)
- Haasnoot J, Westerhout EM, Berkhout B (2007) RNA interference against viruses: strike and counterstrike. *Nat Biotechnol* 12:1435–1443
- 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. doi:[10.1007/s00705-012-1253-7](https://doi.org/10.1007/s00705-012-1253-7)
- Hansen CN, Harper G, Heslop-Harrison JS (2005) Characterisation of pararetrovirus-like sequences in the genome of potato (*Solanum tuberosum*). *Cytogenet Genome Res* 110(1–4):559–565. doi:[10.1159/000084989](https://doi.org/10.1159/000084989)
- Huang Y, Mi ZQ, Zhuang L, Ma MJ, An XP, Liu W, Cao WC, Tong YG (2013) Presence of entomobirnaviruses in Chinese mosquitoes in the absence of Dengue virus co-infection. *J Gen Virol* 94:663–667. doi:[10.1099/vir.0.048231-0](https://doi.org/10.1099/vir.0.048231-0)
- Hwang YT, Kalischuk M, Fusaro AF, Waterhouse PM, Kawchuk L (2013) Small RNA sequencing of potato leafroll virus-infected plants reveals an additional subgenomic RNA encoding a sequence-specific RNA-binding protein. *Virology* 438(2):61–69. doi:[10.1016/j.virol.2012.12.012](https://doi.org/10.1016/j.virol.2012.12.012)
- Isakov O, Modai S, Shomron N (2011) Pathogen detection using short-RNA deep sequencing subtraction and assembly. *Bioinformatics* 27(15):2027–2030. doi:[10.1093/bioinformatics/btr349](https://doi.org/10.1093/bioinformatics/btr349)

- Jakowitsch J, Mette MF, van der Winden J, Matzke MA, Matzke AJM (1999) Integrated pararetroviral sequences define a unique class of dispersed repetitive DNA in plants. *Proc Natl Acad Sci U S A* 96(23):13241–13246. doi:[10.1073/pnas.96.23.13241](https://doi.org/10.1073/pnas.96.23.13241)
- Kashif M, Pietilä S, Artola K, Jones RAC, Tugume AK, Mäkinen V, Valkonen JPT (2012) Detection of viruses in sweetpotato from Honduras and Guatemala augmented by deep-sequencing of small-RNAs. *Plant Dis* 96(10):1430–1437. doi:[10.1094/pdis-03-12-0268-re](https://doi.org/10.1094/pdis-03-12-0268-re)
- Kircher M, Sawyer S, Meyer M (2012) Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res* 40(1):e3. doi:[10.1093/nar/gkr771](https://doi.org/10.1093/nar/gkr771)
- Kreuze JF, Pérez A, Untiveros M, Quispe D, Fuentes S, Barker I, Simon R (2009) Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: a generic method for diagnosis, discovery and sequencing of viruses. *Virology* 388(1):1–7. doi:[10.1016/j.virol.2009.03.024](https://doi.org/10.1016/j.virol.2009.03.024)
- Kreuze J, Koenig R, De Souza J, Vetten HJ, Muller G, Flores B, Ziebell H, Cuellar W (2013) The complete genome sequences of a Peruvian and a Colombian isolate of Andean potato latent virus and partial sequences of further isolates suggest the existence of two distinct potato-infecting tymovirus species. *Virus Res* 173(2):431–435
- Kunii M, Kanda M, Nagano H, Uyeda I, Kishima Y, Sano Y (2004) Reconstruction of putative DNA virus from endogenous rice tungro bacilliform virus-like sequences in the rice genome: implications for integration and evolution. *BMC Genomics* 5(10):80. doi:[10.1186/1471-2164-5-80](https://doi.org/10.1186/1471-2164-5-80)
- 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(5):e37127. doi:[10.1371/journal.pone.0037127](https://doi.org/10.1371/journal.pone.0037127)
- Linsen SE, de Wit E, Janssens G, Heater S, Chapman L, Parkin RK, Fritz B, Wyman SK, de Bruijn E, Voest EE, Kuersten S, Tewari M, Cuppen E (2009) Limitations and possibilities of small RNA digital gene expression profiling. *Nat Methods* 6(7):474–476. doi:[10.1038/nmeth0709-474](https://doi.org/10.1038/nmeth0709-474)
- Lockhart BE, Menke J, Dahal G, Olszewski NE (2000) Characterization and genomic analysis of tobacco vein clearing virus, a plant pararetrovirus that is transmitted vertically and related to sequences integrated in the host genome. *J Gen Virol* 81:1579–1585
- Loconsole G, Onelge N, Potere O, Giampetruzzi A, Bozan O, Satar S, De Stradis A, Savino V, Yokomi RK, Saponari M (2012a) Identification and characterization of citrus yellow vein clearing virus, a putative new member of the genus Mandarivirus. *Phytopathology* 102(12):1168–1175. doi:[10.1094/PHYTO-06-12-0140-R](https://doi.org/10.1094/PHYTO-06-12-0140-R)
- Loconsole G, Saldarelli P, Doddapaneni H, Savino V, Martelli GP, Saponari M (2012b) Identification of a single-stranded DNA virus associated with citrus chlorotic dwarf disease, a new member in the family Geminiviridae. *Virology* 432(1):162–172. doi:[10.1016/j.virol.2012.06.005](https://doi.org/10.1016/j.virol.2012.06.005)
- Ma M, Huang Y, Gong Z, Zhuang L, Li C, Yang H, Tong Y, Liu W, Cao W (2011) Discovery of DNA viruses in wild-caught mosquitoes using small RNA high throughput sequencing. *PLoS One* 6(9):e24758. doi:[10.1371/journal.pone.0024758.t001](https://doi.org/10.1371/journal.pone.0024758.t001)
- Mlotshwa S, Pruss GJ, Vance V (2008) Small RNAs in viral infection and host defense. *Trends Plant Sci* 13(7):375–382
- Ng TFF, Duffy S, Polston JE, Bixby E, Vallad GE, Breitbart M (2011) Exploring the diversity of plant DNA viruses and their satellites using vector-enabled metagenomics on whiteflies. *PLoS One* 6(4):e19050. doi:[10.1371/journal.pone.0019050](https://doi.org/10.1371/journal.pone.0019050)
- Richert-Poggeler KR, Noreen F, Schwarbacher T, Harper G, Hohn T (2003) Induction of infectious petunia vein clearing (pararetro) virus from endogenous provirus in petunia. *EMBO J* 22(18):4836–4845. doi:[10.1093/emboj/cdg443](https://doi.org/10.1093/emboj/cdg443)
- Roossinck MJ, Saha P, Wiley GB, Quan J, White JD, Lai H, Chavarria F, Shen G, Roe BA (2010) Ecogenomics: using massively parallel pyrosequencing to understand virus ecology. *Mol Ecol* 19(1):81–88

- Roy A, Choudhary N, Guillermo LM, Shao J, Govindarajulu A, Achor D, Wei G, Picton DD, Levy L, Nakhla MK, Hartung JS, Brlansky RH (2013) A novel virus of the genus Cilevirus causing symptoms similar to *Citrus leprosis*. *Phytopathology* 103(5):488–500. doi:[10.1094/phyto-07-12-0177-r](https://doi.org/10.1094/phyto-07-12-0177-r)
- Sela N, Luria N, Dombrovsky A (2012) Genome assembly of bell pepper endornavirus from small RNA. *J Virol* 86(14):7721–7721. doi:[10.1128/jvi.00983-12](https://doi.org/10.1128/jvi.00983-12)
- Staginnus C, Gregor W, Mette MF, Teo CH, Borroto-Fernandez EG, Machado ML, Matzke M, Schwarzacher T (2007) Endogenous pararetroviral sequences in tomato (*Solanum lycopersicum*) and related species. *BMC Plant Biol* 7:24. doi:[10.1186/1471-2229-7-24](https://doi.org/10.1186/1471-2229-7-24)
- Untiveros M, Quispe D, Kreuze J (2010) Analysis of complete genomic sequences of isolates of the *Sweet potato feathery mottle virus* strains C and EA: molecular evidence for two distinct potyvirus species and two P1 protein domains. *Arch Virol* 155(12):2059–2063. doi:[10.1007/s00705-010-0805-y](https://doi.org/10.1007/s00705-010-0805-y)
- 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(21):11016–11021. doi:[10.1128/JVI.05647-11](https://doi.org/10.1128/JVI.05647-11)
- Wu Q, Luo Y, Lu R, Lau N, Lai EC, Li W-X, Ding S-W (2010) Virus discovery by deep sequencing and assembly of virus-derived small silencing RNAs. *Proc Natl Acad Sci U S A* 107(4):1606–1611. doi:[10.1073/pnas.0911353107](https://doi.org/10.1073/pnas.0911353107)
- Wu Q, Wang Y, Cao M, Pantaleo V, Burgyan J, Li W-X, Ding S-W (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(10):3938–3943. doi:[10.1073/pnas.1117815109](https://doi.org/10.1073/pnas.1117815109)
- Wylie SJ, Jones MGK (2011) The complete genome sequence of a Passion fruit woodiness virus isolate from Australia determined using deep sequencing, and its relationship to other potyviruses. *Arch Virol* 156(3):479–482. doi:[10.1007/s00705-010-0845-3](https://doi.org/10.1007/s00705-010-0845-3)
- Wylie SJ, Luo H, Li H, Jones MGK (2012a) Multiple polyadenylated RNA viruses detected in pooled cultivated and wild plant samples. *Arch Virol* 157(2):271–284. doi:[10.1007/s00705-011-1166-x](https://doi.org/10.1007/s00705-011-1166-x)
- Wylie SJ, Tan AJY, Li H, Dixon KW, Jones MGK (2012b) Caladenia virus A, an unusual new member of the family Potyviridae from terrestrial orchids in Western Australia. *Arch Virol* 157(12):2447–2452. doi:[10.1007/s00705-012-1452-2](https://doi.org/10.1007/s00705-012-1452-2)
- Wylie SJ, Li H, Dixon KW, Richards H, Jones MGK (2013) Exotic and indigenous viruses infect wild populations and captive collections of temperate terrestrial orchids (*Diuris* species) in Australia. *Virus Res* 171(1):22–32. doi:[10.1016/j.virusres.2012.10.003](https://doi.org/10.1016/j.virusres.2012.10.003)
- 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(9):1081–1090. doi:[10.1094/PHYTO-02-11-0034](https://doi.org/10.1094/PHYTO-02-11-0034)

Detection and Diagnostics of Plant Pathogens

Gullino, M.L.; Bonants, P.J.M. (Eds.)

2014, XI, 200 p. 22 illus., 19 illus. in color., Hardcover

ISBN: 978-94-017-9019-2