Deep sequencing of plant small RNAs: a generic method for diagnosis, discovery and sequencing of viruses

Jan F. Kreuze[1], Ana Perez[2], Milton Untiveros[1], Dora Quispe[1], Segundo Fuentes[2], Giovanna Muller[2], Ratna Kumria[3], Reinhard Simon[1], Ian Barker[2], Claude Fauquet[3] and Wilmer J. Cuellar[2]

[1]Germplasm Enhancement and Crop Improvement Division, and [2]Integrated Crop Management Division, International Potato Center, Apartado 1558, Lima 12, Peru; [3] International Laboratory for Tropical Agricultural Biotechnology, Donald Danforth Plant Science Center, 975 North Warson Road, St. Louis, MO 63132.
Corresponding author: j.kreuze@cgiar.org

Abstract

Vegetative propagated crops are prone to the build up of virus infections and new viral diseases continue to appear. Early detection of the appearance of new viruses followed by rapid and accurate identification of these agents is essential if correct control measures are to be deployed. This is particularly true for entirely new diseases where novel control strategies may have to be developed alongside characterization of novel agents. Plants defend themselves against viruses by RNA silencing which involves the generation and use of small interfering RNA (siRNA): short RNA sequences of 20-25 nt derived from the viral genomic or sub-genomic RNA. We report the first identification of novel viruses from sweetpotato, and sequence of entire viral genomes, by a single step of high-throughput parallel sequencing of short RNAs from diseased, as well as symptomless plants. Contigs were assembled from sequenced total siRNA from plants using small sequence assembly software and could positively identify RNA, ssDNA and dsDNA reverse transcribing viruses, sometimes spanning the entire genome. The results present a novel approach which cannot only identify known viral pathogens, occurring at extremely low titers, but also novel viruses, without the necessity of any prior knowledge

Keywords: plant virus, deep-sequencing, diagnosis, siRNA, virus discovery.

Introduction

Crop losses due to emerging plant diseases, including those of viral origin, are of great current concern particularly in developing countries (Anderson et al., 2004). Strategies to combat plant disease outbreaks often involve early intervention either to stop diseases spreading or to prevent their introduction in the first place. Typically, the identification of a virus requires the application of a number of approaches including physical, biological, serological and molecular methods. Recently technologies such as diagnostic microarrays and mass spectrometry have been proposed as generic tools for identifying viruses (Mumford et al., 2006) although all require some prior knowledge of the agents to be identified. With the advent of next generation high-throughput parallel sequencing ("deep sequencing") platforms, the possibility of random metagenomic sequencing of diseased samples to identify putative pathogens has emerged (Quan et al., 2008). However, elimination of host nucleic acid in these systems is critical to boosting pathogen signals toward the detection threshold. We hypothesized that it would be possible to identify viruses based on the sequences of viral defense related molecules in plants. RNA silencing (RNAi) is a cytoplasmic cell surveillance system to recognize double stranded RNA and specifically destroys single and double stranded RNA molecules homologous to the inducer, using small interfering RNAs (siRNA) as a guide (Fire et al., 1998). Viruses are both inducers and targets of RNAi that constitutes a fundamental antiviral defence mechanism in eukaryotic organisms (Haasnoot et al., 2007). It is particularly important in plants (Pantaleo et al., 2007) that use RNAi to recover from virus disease (Covey et al., 1997). We describe the use of deep sequencing of siRNAs from plants to successfully identify the viruses infecting them, including previously unknown viruses, even in extremely low titre symptomless infections.

Materials and methods

Plant material and virus strains

The following plant materials were used in this study: Sweetpotato (*Ipomoea batatas*) landrace 'Huachano', infected with *Sweet potato feathery mottle* isolate Piu (SPFMV-Piu; Genus: *Potyvirus*, Family: *Potyviridae*), *Sweet*

potato chlorotic stunt isolate M2-47 (SPCSV-M2–47; Genus: *Crinivirus*; Family: *Closteroviridae*) and with both viruses simultaneously; potato (*Solanum tuberosum*) cv. Serrana infected with *Potato virus T* from Peru (PVT-Pe; Family: *Flexiviridae*); cassava (*Manihot esculenta*) infected with *Cassava brown streak virus* isolate Ug (CBSV-Ug; Genus: *Ipomovirus*, Family: *Potyviridae*); *Nicotiana benthamiana* infected with SB29 (a novel suspected viral pathogen of potato); healthy *Physalis floridana.* All these materials were maintained in an insect-proof greenhouse at CIP, Lima, Peru, except CBSV-Ug infected cassava, which was maintained in a growth chamber at the Donald Danforth Plant Science Center, Missouri, USA.

Nucleic acid extraction and sequencing

Total RNA was isolated from 3 g of fresh leaf material using Trizol (Invitrogen, CA, USA) following the manufacturer's instructions. Lyophilized RNA was sent to Fasteris Life Sciences SA (Plan-les-Ouates, Switzerland) for processing and sequencing on the Illumina Genome Analyzer. DNA was extracted using the CTAB method. PCR amplification of virus specific fragments was performed using Taq DNA polymerase (Promega) according to the manufacturer's recommendations together with virus specific primers. Sequencing of PCR amplified fragments using the Sanger method was performed by Macrogen (Seoul, Korea).

Sequence analysis

For siRNA sequence assembly three different short read assemblers were tested: SSAKE v3.2 (Warren et al., 2007), VCAKE v2.0 (Jeck et al., 2007) and Velvet v0.6.04 (Zerbino and Birney, 2008). Different, overlapping, contigs were produced depending on the program used and the parameters set, and they could be further assembled into greater contigs using the program ContigExpress included in the Vector NTI package (Invitrogen, Carlsbad, CA). Assembled contigs were used to search the GenBank/EMBL/DDBJ database using BLASTn (nucleotide blast) or BLASTx (translated nucleotide blast). Primers (data not shown) were designed for amplification and Sanger sequencing based on the identified viral contigs using Vector NTI (Invitrogen). Guide strand mediated assembly was performed using the program MAQ (http://maq.sourceforge.net). Coverage and distribution of virus specific contigs by siRNAs were also determined using MAQ under default parameters, and results were exported to Microsoft Excel for further analysis.

Results and discusion

Sweetpotato viruses

In a first experimental setup we isolated and sequenced sRNAs from single (SPCSV or SPFMV) and double (SPFMV and SPCSV) infected sweetpotato plants using the Illumina deep sequencing platform (Kreuze et al., 2009). Between 1 and 1.2 million sRNA (1–28 nt) reads were obtained from each different sample. The far majority of sequences (>95%) were between 21 and 24 nt in size.

The obtained sRNA sequences could be assembled into contigs of up to more than 1000 nts using any of the three tested short sequence assembly programs. Among the programs tested Velvet was the fastest and generally more accurate than SSAKE or VCAKE. Velvet worked best with the hash length set between 13 and 17, whereas the ideal coverage cut-off parameter varied considerably depending on the hash length used. Searches of nucleotide and protein databases using Blast with the assembled contigs and their corresponding translated peptides successfully identified the expected viruses in each plant, and, surprisingly, also identified several contigs with similarity to badnaviruses (family *Caulimoviridae*; dsDNA reverse transcribing viruses) and mastreviruses (family *Geminiviridae*; ssDNA viruses)(Table 1).

Table 1. Number of contigs assembled by Velvet using 21–24 nt sRNA, with virus specific hits as identified using Translated Nucleotide Blast (Blastx) and % coverage and average depth of viral genome sequenced

Plant infected with	siRNAs sequenced (21-24nt)	Contigs identified	Contigs with Blastx hits All sequences k=15, cov=3	Coverage of complete genome and average sequencing depth
Sweetpotato SPFMV	1'072'019	Total contigs	1633	
		SPFMV	71	93%/93x
		SPCSV	0	-
		Badnavirus	62	A:92% /79x B: 99%/107x
		Mastrevirus	6	95%/73x
Sweetpotato SPCSV	1'169'787	Total contigs	1675	
		SPFMV	0	-
		SPCSV	64	92%/16x
		Badnavirus	63	A: 92%/90x B: 99%/135x
		Mastrevirus	10	95%/113x
Sweetpotato SPFMV + SPCSV	984'490	Total contigs	1363	
		SPFMV	43	100% / 470x
		SPCSV	41	86% /13x
		Badnavirus	63	A[1]: 92%/130x B: 99%/165x
		Mastrevirus	8	91% /123x
Cassava CBSV	879'337	Total contigs	760	
		CBSV	60	97% / 153x
		Begomovirus	55	SLCMV:93%/20x ClCuRaV: 91%/57x
		Beta satellite	4	89%/18x
		Alpha satellite	1	ND[2]
Potato PVT	1'591'500	Total contigs	2276	
		PVT	56	98.59% / 32x
		Cavemovirus	16	ND
Nicotiana benthamiana SB-29	784'718	Total contigs	276	
		Torradovirus	2	ND
Physalis floridiana Healthy	757'310	Total contigs	635	
		Cavemovirus	4	ND

[1] Two different viruses identified; [2] ND: not done

Moreover, contigs of SPFMV generated in dually infected plants were found to span the entire genome and could be further assembled to generate the complete genomic sequence at a sequencing depth of 470x (Table 1). The accuracy was confirmed by Sanger sequencing and found to concur to 99.8%. Although the contigs produced for SPCSV were too few and short to be able to assemble it's entire genome de-novo, guide strand aided assembly using the published SPCSV-Ug sequence and the MAQ software was able to assemble up to 92% of the SPCSV genome at an average depth of 16x (Table 1). Further investigation of the badna- and mastrevirus specific contigs revealed that they corresponded to at least two distinct badnaviruses and one mastrevirus and covered more than 50% of their respective genomes. Guide strand aided assembly using MAQ was not able to further assemble the genomes of these viruses because they were too different from any known

virus sequences. Therefore primers were designed based on the available contigs, which successfully amplified fragments of the expected sizes filling the gaps between the contigs found by siRNA assembly. Thus, the complete genomes of both Badnaviruses and the mastrevirus were determined. Subsequent analysis using MAQ showed that >90 % of the genomes of each of the new viruses was covered by siRNAs (Table 1). The significance of these new, apparently symptomless, viruses is currently under investigation. It is noteworthy however that the amount of siRNAs corresponding to these viruses increase in SPCSV infected plants and even more so in SPVD affected plants (Table 1), suggesting that they may have a role in the aetiology of both diseases.

Viruses of other plants

To validate our method in other plant species, deep sequencing of sRNA samples from PVT infected potato, *N. benthamiana* infected with the previously un-characterized potato virus SB29*,* CBSV infected cassava and healthy *Physalis floridiana* (as a negative control) was performed. As expected PVT and CBSV could be positively identified in infected potato and cassava plants respectively. As with SPCSV and SPFMV in sweetpotato, almost the entire genomes of both PVT and CBSV could be assembled using a combination of MAQ and Velvet (Table 1). The assembled PVT sequence showed 98% nucleotide identity with the one reported in database (NC011062). The CBSV isolate sequenced in this study however showed only about 85% nt identity with the Tanzanian highland strain MBL3 (NC012698), which was confirmed by Sanger sequencing.

Only two virus specific contigs could be identified from SB29 infected *N.benthamiana* samples and they matched to two regions separated by ~1.7 Kbp in the RNA2 of *Tomato torrado virus* (ToTV; EU563947). PCR primers designed from these two contigs amplified a product of the expected size and Sanger sequencing of this PCR product revealed a region corresponding to Vp35 (one out of 3 CP in ToTV) which had a 33-35% amino acid identity with the corresponding region in other torradoviruses reported in database. This result suggests SB29 represents a new virus that is most closely related to, but still significantly different from torradoviruses. Low sequence similarity to any known viruses is probably the reason only two torradovirus specific contigs were identified in SB29 infected *N. benthamiana* and illustrates both the potential and the limitations of this technology.

Interestingly, as in the case of sweetpotato, in all samples additional contigs with similarity to plant viruses were also identified. CBSV infected cassava samples yielded siRNA contigs with similarity to Begomoviruses and beta satellite virus (Table 1). On closer inspection these appeared to be homologous to *Sri-Lankan cassava mosaic virus* (SLCMV DNA-A and –B; Genus: *Begomovirus*, Family *Geminiviridae*), *Cotton leaf curl Rajastan virus* (ClCuRaV DNA-A; Genus: *Begomovirus*, Family *Geminiviridae*) and Cotton leaf curl beta satellite, and could be assembled to ~90% of the genome using MAQ (Table 1). siRNAs corresponding to these viruses were however very few as compared to CBSV and their presence has yet to be confirmed by independent methods.

Contigs with similarity to cavemovirus (Family: *Caulimoviridae*) sequences were recovered from both PVT infected potatoes and "healthy" *P. floridiana* samples. PCR primers designed from the contigs in *P. floridiana* to amplify the whole genome gave PCR products of expected size. Additional PCR products were also obtained, and Sanger sequencing showed that all amplified products corresponded to cavemovirus-like sequences. No follow up was made with the cavemovirus like sequences identified in potato yet. Because cavemoviruses are known to integrate into their host genomes, it will be an interesting question to determine whether these sequences correspond to ancient, non-active integrated viruses, or to infectious symptomless viruses in these plants, and how these two alternatives could be distinguished using this technology.

Distribution of siRNAs over viral genomes

The quantity of virus specific RNAs among the different size classes of sequenced RNA was determined and are shown in Fig 1. The major proportion of virus specific siRNAs were found in the 21 or 22nt class, regardless of the host plant or type of virus (DNA or RNA). A difference could however be observed between plant species, in that sweetpotato had the majority of virus specific siRNAs of the 22nt size class, and potato in the 21nt size class, whereas the situation was ambiguous for cassava. More data would be required to determine if this is due to differences in the host plants or the viruses that are targeted. Nevertheless because virus specific siRNAs represent a higher proportion of the 21 and 22nt siRNAs, sequencing only this size class would reduce the number of sequences required to identify a new virus and thus increase sensitivity of the method. Indeed analysis using the sweetpotato siRNA sequences indicated that similar results could be obtained if using only the 22nt siRNAs (Kreuze et al., 2009). Furthermore simulations using random subsets of sequences indicated that as

few as few as 30,000 total siRNA sequences were enough to assemble at least one contig recognizable as SPCSV, the virus for which the fewest siRNA sequences were found in sweetpotato.

Analysis of the distribution of siRNAs over the viral genomes revealed that they are not homogenously distributed, but concentrated in certain regions (Fig 2). This effect was highly reproducible between samples and is most likely due to a bias produced in the sample preparation technique (Linsen et al., 2009). Therefore the use of improved sample preparation methods that are currently available, and are less biased may improve the coverage and thus the ability to produce longer contigs with fewer sequences.

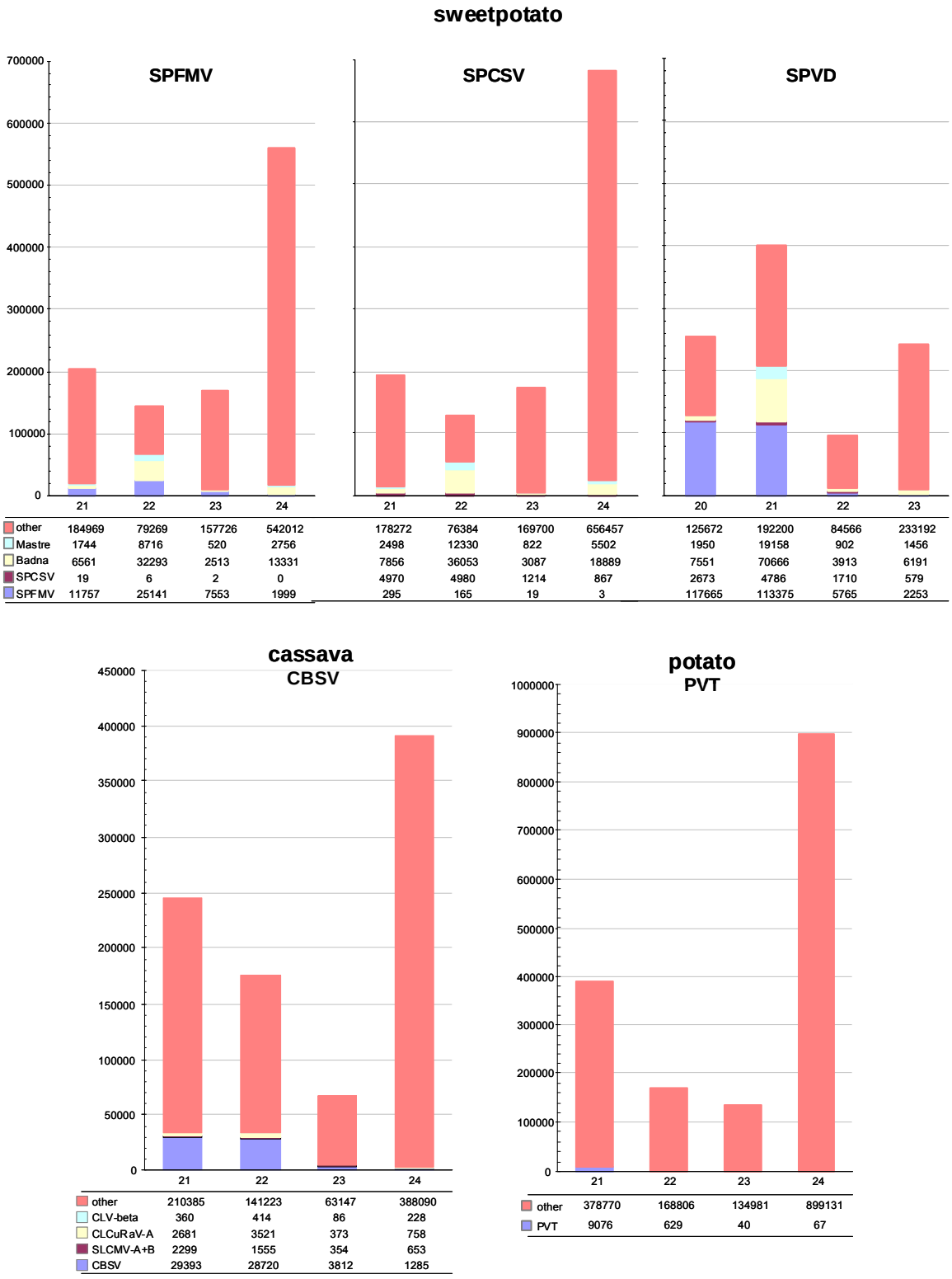

Figure 1 Bare graph showing the frequency of 21-24nt siRNAs corresponding to different viruses as compared to the total number of siRNAs sequenced in virus infected sweetpotato, potato and cassava.

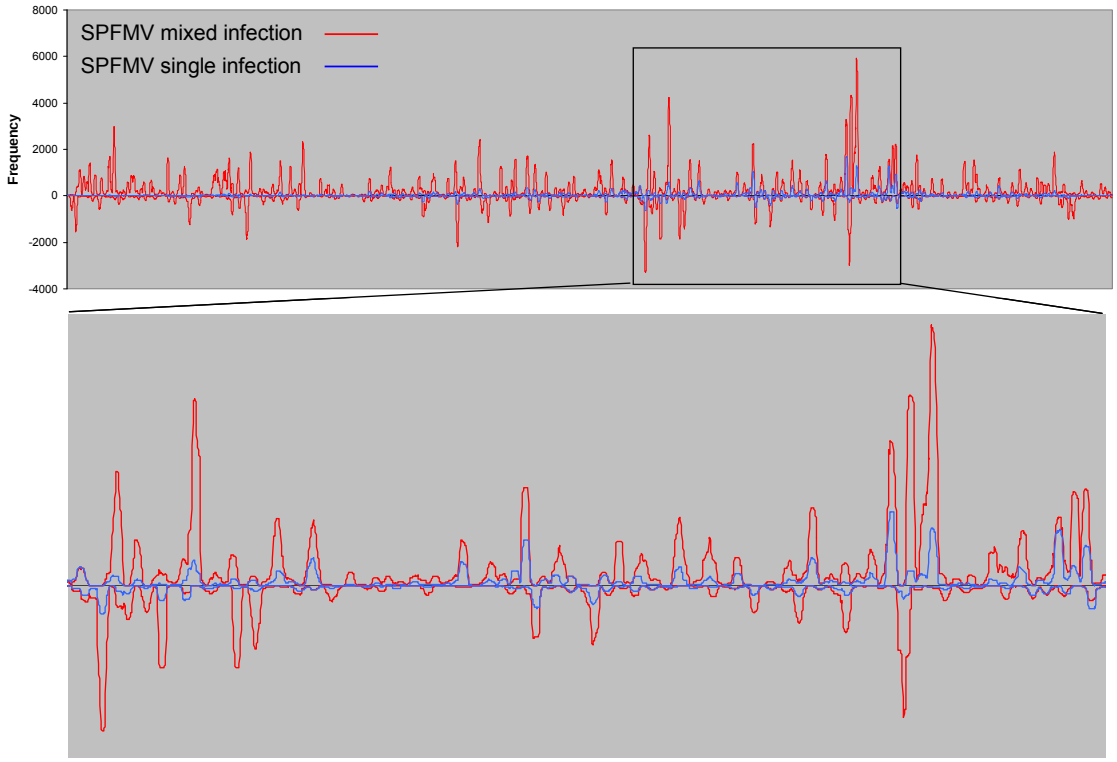

Figure 2. Graphic showing the sequence coverage of the SPFMV genome by siRNAs in single and mixed infected plants. Strong and consistent preference for certain regions over others can be observed.

Conclusions

This methodology, as applied in plants, was thus able to detect both RNA and DNA viruses, from widely different families and with different tissue trophisms and intracellular replication sites, even in extremely low titre and seemingly symptomless infections. It offers an entirely generic, specific and apparently sensitive approach to identify plant viruses, as compared to other techniques, which are all in some way limited to a subset of viruses that can be identified or require additional confirmatory steps for virus identification. The fact that similar results could be obtained with viruses from different families and in diverse plants suggests it may be universally applicable. The apparent sensitivity combined with increased throughput obtained by massive parallel sequencers may eventually lead to the technique becoming widely applicable. On the other hand the frequent identification of unexpected viral sequences even in seemingly healthy plants suggests that apparently symptomless viral infections are more common than previously thought, and poses the problem of what significance should be assigned to them.

References

Anderson, P.K.; Cunningham, A.A.; Patel, N.G.; Morales, F.J.; Epstein, P.R.; Daszak, P. 2004. Emerging infectious diseases of plants: pathogen pollution, climate change and agrotechnology drivers. Trends Ecol. Evol. 10, 535–544.

Covey, S.N.; Al-Kaff, N.S.; Langara, A.; Turner, D.S. 1997. Plants combat infection by gene silencing. Nature 285, 781-782.

Fire, A.; Xu, S.Q.; Montgomery, M.K.; Kostas, S.A.; Driver, S.E.; Mello, C.C. 1998. Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. Nature 391, 806-811.

Haasnoot, J.; Westerhout, E.M.; Berkhout, B. 2007. RNA interference against viruses: strike and counterstrike. Nature Biotech. 12, 1435-1443.

Jeck, W.R.; Reinhardt, J.A.; Baltrus, D.A.; Hickenbotham, M.T.; Magrini, V.; Mardis, E.R.; Dangl, J.L.; Jones, C.D. 2007. Extending assembly of short DNA sequences to handle error. Bioinformatics 23, 2942–2944.

Kreuze, J.F.; Perez, A.; Untiveros, M.; Quispe, D.; Fuentes, S.; Barker, I.; Simon, R. 2009 Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: a generic method for diagnosis, discovery and sequencing of viruses. Virology 388, 1-7.

Linsen, S.E.V.; de Wit, E.; Janssens, G.; Heater, S.; Chapman, L.; Parkin, R.K.; Fritz, B.; Wyman, S.K.; de Bruijn, E.; Voest, E.E.; Kuersten, S.; Tewari, M.; Cuppen, E. 2009. Limitations and possibilities of small RNA digital gene expression profiling. Nature Meth. 6, 475-476.

Mumford, R.; Boonham, N.; Tomlinson, J.; Barker, I. 2006. Advances in molecular phytodiagnostics - new solutions for old problems. European J. Plant Pathol. 116, 1–19.

Quan, P.; Briese, T.; Palacios, G.; Lipkin, W.I. 2008. Rapid sequence-based diagnosis of viral infection. Antiviral Res. 79, 1–5.

Pantaleo, V.; Szittya, G.; Burgyan, J. 2007. Molecular bases of viral RNA targeting by viral small interfering RNA-programmed RISC. J. Virol. 81, 3797-3806.

Warren, R.L.; Sutton, G.G.; Jones, S.J.M.; Holt, R.A. 2007. Assembling millions of short DNA sequences using SSAKE. Bioinformatics 4, 500–501.

Zerbino, D.R.; Birney, E. 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18, 821-829