Journal of
Data Mining in Genomics & Proteomics

Roy et al., J Data Mining Genomics Proteomics 2013, 4:2
DOI: 10.4172/2153-0602.1000129

Review Article Open Access

A Case Study on Discovery of Novel *Citrus Leprosis* Virus Cytoplasmic Type 2 Utilizing Small RNA Libraries by Next Generation Sequencing and Bioinformatic Analyses

Avijit Roy[1,3*,#], Jonathan Shao[2#], John S Hartung[2], William Schneider[3] and RH Brlansky[1]

[1]*University of Florida, IFAS, Plant Pathology Department, Citrus Research and Education Center, 700 Experiment Station Road, Lake Alfred, FL, USA*
[2]*USDA-ARS, MPPL, Beltsville, MD, USA*
[3]*USDA-ARS, FDWSRU, Ft. Detrick, MD, USA*
[#]*Avijit Roy and Jonathan Shao should be considered as co-first authors.*

Abstract

The advent of innovative sequencing technology referred to as "Next-Generation" Sequencing (NGS), provides a new approach to identify the 'unknown known' and 'unknown unknown' viral pathogens without *a* priori knowledge. The genomes of plant viruses can be rapidly determined even when occurring at extremely low titers in the infected host. The method is based on massively parallel sequencing of the population of small RNA molecules 18-35 nucleotides in length produced by RNA silencing host defense. Improvements in chemistries, bioinformatic tools and advances in engineering has reduced the costs of NGS, increased its accessibility, and enabled its application in the field of plant virology. In this review, we discuss the utilization of the Illumina GA IIX platform combined with the application of molecular biology and bioinformatic tools for the discovery of a novel cytoplasmic *Citrus leprosis* virus (CiLV). This new virus produced symptoms typical of CiLV but was not detected with either serological or PCR-based assays for the previously described virus. The new viral genome was also present in low titer in sweet orange (*Citrus sinensis*), an important horticultural crop with incomplete genomic resources. This is a common situation in horticultural research and provides an example of the broader utility of this approach. In addition to the discovery of novel viruses, the sequence data may be useful for studies of viral evolution and ecology and the interactions between viral and host transcriptomes.

Introduction

The identification of pathogens remains a major focus of the fields of plant pathology and medicine. Causal agent identification and detection remains particularly important for the field of plant virology, where the lack of host immune systems and post infection therapy options mean that early and correct pathogen detection is still the best tool for the prevention of disease epidemics. Prior to the development of molecular tools, the identification of unknown viruses in infected plants relied on conventional methods such as identification of a suitable experimental plant host, purification of the virus by precipitation and differential or gradient centrifugation followed by electron microscopy [1]. These methods were remarkably effective, as evidenced by the large number of characterized plant viruses identified prior to the molecular age. However, these methods were fairly labor intensive and required longer periods of time between problem identification and causal pathogen characterization. In addition, morphology and biological characteristics can be misleading when multiple viruses share similar traits.

The development of immuno-based detection techniques such as ELISA [2] revolutionized virus detection for both plant and animal viruses. This was followed by detection of pathogen nucleic acids using a number of techniques that fall into two basic categories: hybridization and amplification [i.e., PCR and reverse transcription (RT)-PCR] based techniques [3]. Both immunological and nucleic acid based detection systems require some *a* priori characterization of the pathogen. Immunological methods require the production of an antibody and nucleic acid detection systems require knowledge of the pathogen sequence. This is seldom a problem for characterized diseases, but does present a limitation for the identification and detection of unknown pathogens.

Unknown pathogens can be divided into two classes: "unknown knowns", which are pathogens that have well characterized relatives (e.g., a novel unknown potyvirus) [4], and "unknown unknowns", which are pathogens with no characterized relatives. There are a number of approaches for dealing with unknown known novel plant viruses, including general antibodies [5,6], nucleic acid signatures that can identify many members of plant virus families [7-10] and other commercially available virus group detection assays. However, these reagents are much less likely to be available for poorly characterized plant virus families. The detection of "unknown unknown" plant viruses is particularly problematic, as there are no reference sequences or reagents (e.g., antibodies) available for comparison or identification.

In recent years technologies have been developed that do not require significant prior knowledge of the virus in order to identify it. A researcher may use SDS-PAGE to separate proteins from diseased and healthy plants. Bands unique to the infected plants are excised and identified by peptide mass fingerprints. This technique relies on MALDI-TOF mass spectrometry [11]. Another interesting approach is to interrogate a microarray of viral genomic sequences with an unknown sample to identify an unidentified virus by sequence homology [12]. An added alternative pathway for virus discovery has recently been developed that exploits novel 'next generation'

***Corresponding author:** Avijit Roy, University of Florida, Citrus Research and Education Center, 700, Experiment Station Road, Lake Alfred, Florida 33850, USA, Tel: +1-863-956-8703; Fax: +1-863-956-4631; E-mail: avijitroy@ufl.edu

Received May 09, 2013; **Accepted** May 28, 2013; **Published** June 05, 2013

Citation: Roy A, Shao J, Hartung JS, Schneider W, Brlansky RH (2013) A Case Study on Discovery of Novel *Citrus Leprosis* Virus Cytoplasmic Type 2 Utilizing Small RNA Libraries by Next Generation Sequencing and Bioinformatic Analyses. J Data Mining Genomics Proteomics 4: 129. doi:10.4172/2153-0602.1000129

sequencing (NGS) technology. The NGS methods are more sensitive even than the microarray based assay and have potential to detect the complete genome of either unidentified or unexpected viruses [13-15].

In the new era of NGS technology, hundreds of gigabases of sequences can be obtained directly utilizing total nucleic acid extracts from infected and healthy hosts [16]. The revolutionary NGS technologies have opened new perspectives for research in the areas of sequencing of viral genomes, evolution, ecology, diagnostics and interactions between viral and host transcriptomes. In the last seven years improved chemistries, software and advances in engineering were combined, which has led to the NGS technologies. The reduced costs and increased accessibility of the technologies have resulted in increased application of this novel technology in the various fields of plant virology. NGS methods can be classified into three categories. (1) The 'second generation' system comprised of (i) the first commercially available NGS system that was developed by 454 Life Sciences in 2005 and later purchased by Roche. This sequencing works on the principal of 'pyrosequencing' (ii) the Illumina Genome Analyzer based on the concept of 'sequencing by synthesis' developed by Solexa GA in 2006 and (iii) SOLiD is a unique sequencing process catalyzed by DNA ligase and was commercially released in October 2007, by Life Technologies. (2) The 'third generation' sequencing consists of the 'PacBio Rs' platform developed by Pacific Biosciences. (3) Additional or 'intermediate' platforms for 'second and third generation' sequencing include (i) 'Ion Torrent' by Life Technologies and (ii) 'Helicos' sequencing by Helicos Biosciences. All of these methods are different in terms of their sequencing chemistries and protocols (type of amplification and separation); approximate sequence read lengths, the estimated maximum amount of data generated per run, the main sources of errors, error rates and applications. The 'second generation' system amplified template DNA molecules with a typical 'wash and scan technique' and is able to sequence the populations [17]. Third generation sequencing provides sequence data for single long DNA molecules without gaps. The intermediate category of NGS technologies uses a hybrid approach, with features of both second and third generation sequencing. All NGS methods follow a three step process that includes preparation of a library of short molecules, labeling short molecules with primers to capture and convert RNA into DNA followed by the sequencing phase itself. In this review, we discuss the small RNA (sRNA) NGS technology utilizing Illumina GA IIX platform for discovery of novel *Citrus leprosis* virus cytoplasmic type 2 (CiLV-C2).

A prominent host defense mechanism against plant viruses is based on RNA interference (RNAi) or silencing, the specific cleavage and degradation of the virus genome. When the plant host is infected with a virus, the plant RNA Induced Silencing Complex (RISC) is directed to destroy the invading viral genome [18,19]. Viral genomes produce double stranded (ds) intermediates during replication and these serve as substrates for a Dicer-like ribonuclease, which cleaves the dsRNA into small viral-(siRNAs) which binds with the Argonaut protein (Ago) to form the RISC complex [18,20,21]. The RISC complex is also used by plants to regulate expression of endogenous genes. The RNA cleavage products of both the viral and endogenous genes accumulate as a pool of sRNAs in the size range of 15-35 nucleotides. All of these sRNAs are prime targets for sequencing using the massively parallel approach and Illumina® sequencing technologies. Bioinformatic tools are then used to assemble these sRNA sequences into contigs. Genbank is searched for matches to these sequences and a tentative identification of viral genes and a genome can be made. Here we present a bioinformatics and laboratory workflow to identify nucleic acid sequences of known and unknown viruses within the sRNA fragment pool of their plant host. This protocol identifies unknown virus genomes with low concentrations, typically 1-2% of sRNA reads in the sample. It also has been shown to identify multiple viruses present in the same plant, viruses which do not produce symptoms, and both RNA and DNA viruses [22].

Accurate and rapid identification of the nucleic acid sequences of unknown viruses in nucleic acid extracts of plants is a key to the application of this technology. With the rapid and cost-effective development of NGS technology, elucidation of pathogens using sRNA sequencing has become a practical and thoroughly documented method [23]. Some recent examples in plant virology include identifying *Pepino mosaic virus* in tomato [24], *Citrus yellow vein clearing virus* [25] *Citrus chlorotic dwarf virus* [26], CiLV-C2 [27] and *Citrus leprosis virus* nuclear type (Roy et al. Unpublished) infecting citrus. A similar approach was applied to identify and detect viral pathogens in sweet potato [22,28], grapevine [29-31] and *Dactylis glomerata*, the wild cocksfoot grass [32].

General Considerations

A flowchart that summarizes the work flow is provided (Figure 1). The bioinformatic and computational tools needed for this analysis are as follows. A short read aligner such as Bowtie [33] for alignment of short RNA sequence reads to known genome sequences of interest. Sequence assembly tools such as Velvet and the Oases suites [34,35] are developed to assemble the short RNA sequences into longer contigs and transcripts. The BLAST (Basic Local Alignment Search Tool)

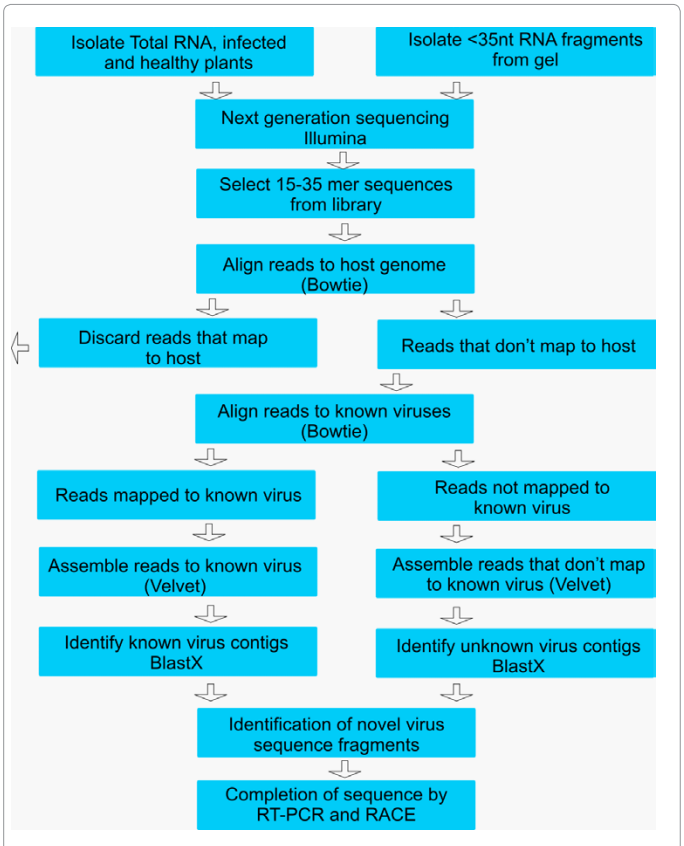

Figure 1: Work plan for the discovery of unknown viruses through next generation sequencing.

[36] is utilized to identify the contigs by sequence similarity searches against Genbank. Results will vary depending on which alignment tool or assembler is used, since different algorithms are used when performing each task by the different software packages. Bowtie and Velvet are relatively fast command line programs which are run in the Linux operating system, allowing effortless execution of options that each program offers. Bowtie is a robust program that is specifically designed to map short reads on to large genomes. Velvet is based on the de-bruijn graph algorithm that excels in aligning very short reads [35] that allows the graph to be traversed by the computationally efficient Eulerian cycle. The alternative assembly strategy is a Hamiltonian cycle used in traditional Sanger Sequence assembly that is computationally much harder to solve for massively parallel datasets and does not scale well [37]. Furthermore an adequate computer system is also essential. In this work, our programs were installed on a Redhat Linux Enterprise 5 system with 12 processors and about 100 GB of RAM available for analysis. The time required to run the analysis is based on a number factors such as the number of replicates or size of your libraries, and the sizes of the databases you want to blast against. It is also affected by the relative purity of your sequence library, for example, whether or not it contains sequences from other viruses, bacteria, or fungi. In most cases it takes between 2-4 hours to complete this analysis. The starting point for examples in this bioinformatic analysis were sequences 15-35 nucleotides in length produced with the Illumina GA IIX platform. The sequencing can be performed by a number of commercial vendors starting with purified RNA. Costs will vary and can be reduced by multiplexing of samples. The resulting short sequence libraries typically have 8-20 million reads. The number of starting reads will vary depending on how many samples are pooled in the sequencing step, which is at the discretion of the researcher.

RNA Isolation and Quality Test for sRNA Sequencing

Before any sequencing of the sRNA pool can be attempted, the RNA from the infected plants must be purified. We have successfully used the Trizol® reagent (Invitrogen, Carlsbad, CA) and RNeasy® Plant Mini Kit (Qiagen) for this purpose, starting with infected leaves ground to a powder in liquid nitrogen. The quantity and quality of the resulting RNA preparation must be sufficient to support the sRNA sequencing process. RNA is quantified after purification using the Qubit fluorometer (Invitrogen, Carlsbad, CA). This fluorometric assay is preferred over spectrophotometric quantitation because DNA, nucleotides, proteins and other chemical contaminants do not interfere with the assay. The quality of the RNA is determined using the 2100 Bioanalyser (Agilent, Santa Clara, CA). This instrument performs a microcapillary electrophoresis of the RNA and provides a digitized result with a qualitative assessment in the form of an RNA Integrity Number (RIN) [38]. A RNA extract with higher RIN will provide a better representation in the resulting sequence library.

sRNA sequencing protocol

Samples of citrus leaves with typical symptoms of *Citrus leprosis virus* (CiLV) were received from Colombia for testing at the United States Department of Agriculture-Animal and Plant Health Inspection Service (USDA-APHIS)-PPQ-CPHST, Beltsville, MD under APHIS permits. Interestingly, indistinguishable leprosis symptoms were known to be caused by two very different leprosis causing viruses, one of which had virions in the cytoplasm [(CiLV cytoplasmic type (CiLV-C)] and the other in the nucleus [CiLV nuclear type (CiLV-N)] of infected cells [39]. However, both CiLV-C RT-PCR and serologically-based assays that were expected to identify these viruses failed to detect them in these samples. Therefore we applied sRNA NGS technology to solve this problem, with a hypothesis that the causal virus in this case was a previously unknown variant of CiLV. Total RNA was isolated from a number of leprosis affected leaf samples that were negative for CiLV-C by RT-PCR. 50 µl of total RNA (L147V1@148ng/µl and GD1234@172 ng/µl) was sent to a NGS service using the Illumina GA IIX platform (Fasteris SA, Switzerland) for further processing. The protocol included converting the sRNA into a cDNA library. To do this, the RNA was separated by acrylamide gel electrophoresis, the sRNA (<35 nt) fragments were eluted, and single-stranded ligations with 3′- and 5′-adapters were followed by acrylamide gel purification of the ligation products to prepare purified sRNA. RT-PCR amplification was used to generate a template library. Quality controlled libraries were then quantified and diluted to a concentration of 10 nM. The libraries were then multiplexed in a single-read channel of a 1×50 bp run on the high-throughput Illumina GA IIX platform.

Criterion for bioinformatic analyses

In some cases the full genome sequence of the host plant may be available to facilitate the bioinformatic analysis via removal of host sequences from the assembly process, but this is not essential. The host plants in the following examples have been members of the genus *Citrus*. Only incomplete genomic sequence data for three citrus species are available, which is a very common situation. The genome of *Citrus sinensis* (sweet orange) (scaffold v1.0) was used in this study and had 12,574 scaffolds, over-lapping contigs with gaps of known length derived through inferences from pair-end sequencing. These scaffolds had a total length of 252,507,433 bases and a GC content of 34.62%. Genome sequence data for *C. clementina* ('Mandarin' orange) is also available with 1,389 scaffolds with a total length of 295,168,965 bases and a GC content of 34.96% (www.citrusgenomedb.org). If relatively complete genome sequence data for the plant host is not available, then the most closely related plant genome should be used. We have used this approach in other experiments when the genome of our citrus host, *C. macrophylla* was not available.

In a number of cases, neither the host plant genome nor the genome of any closely related host genome will be readily available. Though not ideal, there are also solutions to handle this case as well. Experiments should always include a comparable sequence library from healthy plants in order to validate results. Sequences shared between the healthy and diseased plant libraries can be removed from further analysis, since they will not include the viral sequences of interest. The population of sequences that remain will be enriched for viral sequences of interest. As shown in the example below, when starting with field samples, a virus free control sample may not be easily obtained. In this situation, an additional pathway to the discovery of a virus genome in a plant sample is to align the sequence reads from the diseased plant to sequences of known plant viruses. This will identify known virus sequences present in your plant, but these virus sequences may be unrelated to the disease of interest. These sequences can be removed from further consideration. For example, we often find sequences of *Citrus tristeza* virus (CTV) and citrus viroids present in our plant samples obtained from field sources. The next step is to assemble the remaining reads and use blast to identify contigs that match known plant genomes. The assembly process using Velvet in this case takes only a few minutes, while blasting could take many hours depending on how many candidate sequences are discovered. These sequences can be removed from further consideration. The remaining sequences should be enriched for sequences of unknown viruses of interest. These virus sequences are then identified by further blast searches against

Genbank looking for matches to conserved virus genes, such as a coat protein, replicase or movement protein. At this point the e-values may be slightly positive (1.5) though lower e-values indicate a better match (e^{-1} to e^{-20}) whereas in other cases there won't be any hit at all to any known virus genome structure. These contigs or transcripts should also be sequestered for further analysis.

Discovery of Novel *Citrus Leprosis Virus* Using sRNA NGS

As noted above, NGS produces a very large number of reads. In the current study, the two samples from diseased plants (GUB1 and GUB2) produced 22,226,197 and 17,387,445 reads, respectively. In addition, each library was made with insert sizes ranging from 15-35 nucleotides (Figure 2). The reads in this case were single, rather than paired-end or mate-paired. Viral sequences present in the degradome as a result of gene silencing tend to be 21 nt in length [22,23,40]. When the sequence reads were sorted by size class, a peak at 21 nucleotides in length was observed (Figure 2). Although virus RNA fragments derived from the gene silencing pathway are most frequently found in the 21 nucleotide class, when looking for signature sequences from an unknown viral pathogen, we take all sequences that can be reasonably be assembled. Thus, all of the single end reads ranging from 15-35 nt in size were used in this analysis: 20,935,577 (GUB1) and 16,122,595 (GUB2) (Figure 3). Bowtie indexes were built using the Bowtie-build program from the Bowtie suite for both the citrus genome and for known viruses. A Bowtie index is a Burrow-Wheeler index that is used to conserve memory. Using Bowtie version 0.12.7, the reads in each library were mapped to the citrus genome. The mapped reads for each library, 12,512,869 (GUB1) and 10,170,825 (GUB2), were then removed from further consideration, since we were looking for sequences of an unknown virus. The remaining unmapped reads (4,366,936 (GUB1) and 2,833,216 (GUB2) were then mapped to genomic sequences of known viruses using Bowtie (Figure 3).

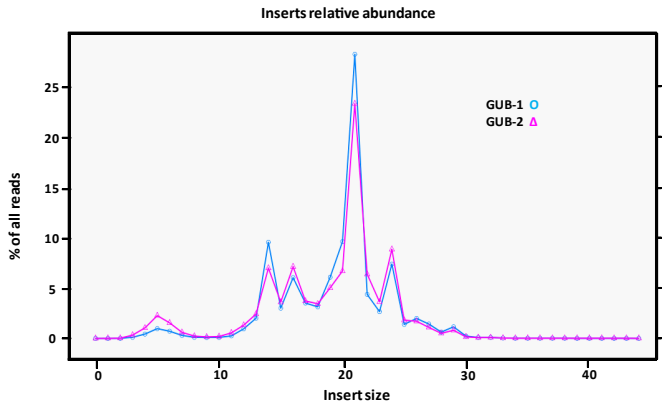

Figure 2: A graphical representation of the relative abundance of the small RNA inserts from two libraries (GUB-1 and GUB-2) of citrus leprosis affected samples.

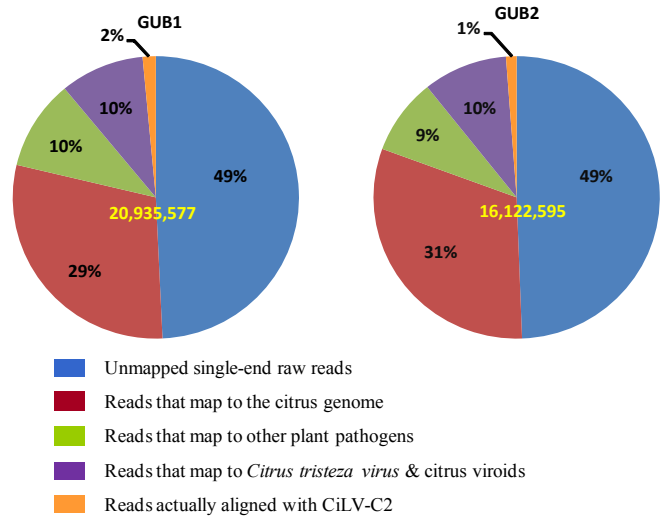

Figure 3: Distribution of reads obtained from small RNA libraries in the process of next generation sequencing. The method of subtraction followed by assembly of reads acquired from the GUB1 and GUB2 libraries was used to deduce the genome sequence of the novel *Citrus leprosis virus* cytoplasmic type 2 (CiLV-C2). Total single-end raw reads were included in the each pie chart.

The reads that mapped to known viruses were assembled using Velvet version 1.2.03 and blast was used to search the non-redundant protein database at NCBI in order to identify the virus genes present in the library. In this case blastx was used, since it is the fastest database search tool. Blastx, takes a nucleotide sequence and translates it into amino acids before it searches a protein database. An additional critical advantage of blastx is that viral structures are more conserved at the amino acid level than at the nucleotide level. Each alignment with a negative or slightly positive e-value (cut off e-value no greater than 1.5), to the non-redundant protein database must be examined closely to determine if a pathogen has been found. In the present case study, most amino acid sequence matches were to various strains of CTV, a widely distributed citrus virus that we were certain was not responsible for the disease symptoms observed. Since CTV was not the pathogen of interest, and viral sequences mutate rapidly, contigs assembled from the unmapped reads were examined. Using the Velvet and Oases 0.2.06 tools, reads (4,055,772 for GUB1 and 3,118,554 for GUB2) that did not map to any of the citrus genomes or to known virus sequences were assembled into contigs. The blastx tool was then used to compare these contigs against the non-redundant protein database. Some contigs matched (negative to slightly positive e-values) known viral structures such as a coat protein, movement protein, methyltransferase, helicase and RNA dependent RNA polymerase. These sRNA contigs sequences were retained for further assembly. Note that many authors will estimate the validity of their sequence reads by counting the numbers of identical copies of each read. Those reads with higher copy numbers are deemed to be more reliable. In this case we felt this was not necessary because the reads tiled to form contigs which were clearly viral in nature. Subsequent laboratory work was performed to complete the assembly and determine the sequence of the novel virus genome. Furthermore, if the unknown virus genome is present in low titer or if the RNA is degraded prior to sequencing, there may be valid reads those are not redundant, and so it is wise not to exclude reads at this stage.

However, other contigs did not match sequences in the non-redundant protein database. As noted above, in most cases a blastx search of the protein sequence database is sufficient to identify similar proteins, since most viral structures match at the amino acid level even between evolutionary divergent viruses. On the other hand, nucleotide databases are more comprehensive than the protein database. Therefore blast was used to search the nucleotide databases, including 'nt' (nucleotide database), 'wgs' (whole genome shotgun), and 'EST'

(expressed sequence tags) to identify matches between the unmapped contigs and the nucleotide sequences in Genbank. For this purpose, an additional tool, tblastx, which compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database, was useful. This analysis is more time-consuming but meets the objective, and it can be used to search many different databases. Sequences of fungal, bacterial and plant origin may also be identified at this stage of the analysis. If genome sequences from an alleged viral pathogen are not discovered, then sequences that have been identified as of fungal and bacterial origin may be removed from consideration in another round of assembly and database searching. This can be helpful because assembly programs such as Velvet tend to give a better assembly with less complex input.

In the present case study after subtracting the contigs of CTV and citrus viroids, the total number of virus contigs was reduced from 231 to 46 for GUB1 and from 185 to 27 for GUB2. Both of the libraries when combined had 73 partially matched contigs with CiLV-C. Of these 73, only 33 contigs (23 from RNA1 and 10 from RNA2) in the size range of 101-807 nt which had the most significant expect (*e)* values (e^{-24} to $e^{1.5}$) were saved for further analysis. Approximately 2.30-2.88% of both RNA-1 and RNA-2 reads were recovered from the libraries and were mapped to the CiLV-C2 genomes. Overlapping sequences were removed and the sequence of the CiLV-C2 genome was identified. As both the 5′ and 3′-UTRs sequences were missing from the sRNA sequencing analysis, RACE techniques were used to sequence the UTR regions of RNA-1 and RNA-2 and successfully complete the viral genome sequence. The sRNA NGS data provided coverage of ~33% (RNA1) and ~55% (RNA2) of the CiLV-C2 genome. To verify the NGS data, primers were designed and the expected amplicons were obtained using RT-PCR. Cloning and Sanger sequencing were performed to obtain overlaps and close gaps in the sequences.

The complete nucleotide sequence of a new bipartite RNA virus and its structure was determined. RNA1 (GenBank Accession Number JX000024) and RNA2 (JX000025) of CiLV-C2 had only 58% and 50% nucleotide identities, respectively, with previously described CiLV-C sequences [41,42]. Conserved domains of CiLV-C2 RNA1 (Helicase and RNA dependent RNA polymerase) and RNA2 (Movement protein) also consistently placed into the CiLV-C cluster and showed that the CiLV-C2 was more closely related to CiLV-C than other viruses [27]. Based on all experimental data, we recommended the unequivocal classification of CiLV-C2 as a definitive member of the genus *Cilevirus* [27].

Conclusions and Perspectives

The sRNA NGS method can provide initial sequence information about both 'unknown known' and 'unknown unknown' viruses. In this study, the assembled contigs generated from two sRNA libraries provided a tentative identification of the virus associated with the leprosis symptoms observed, which is an example of an 'unknown known' virus. In general, the sequence will not be fully assembled by tiling of the short sequences and the gaps will need to be filled. If a reference map of a related virus is available, the sequence contigs can be tiled onto the reference genome and PCR used to amplify the sequence between the contigs for Sanger sequencing [27]. In addition the terminal sequences of linear RNA viruses may not be obtained directly. These can be determined by using 5′ and 3′ RACE technologies [27]. Alternatively, if the viral genome is expected to be circular, then the NGS techniques can be supplemented with rolling circle amplification to enhance turnaround time and sensitivity [43], and verification can be done by inverse PCR methods to amplify the entire genome [25]. At this point the relationship of the new virus to other related viruses can be established by comparing homologous gene sequences. Based on the new sequence information, PCR or serologically based assays can also be used to establish a tight linkage between the novel virus and the symptoms observed. The best way to establish a causal link between a virus and its host is to complete Koch's postulates with a cloned viral genome [44]. The sRNA sequencing approach has potential for sequencing any plant virus [27], as even DNA viruses are subject to degradation of viral mRNA [22,23,45]. Genomes of viruses that are particularly adept at avoiding host imposed silencing may not be easily detected with these methods, but for the most part sRNA sequencing seems to be a particularly effective universal approach to generating sequence for unknown viruses.

References

1. Hamilton RI, Edwardson JR, Francki RIB, Hsu HT, Hull R, et al. (1981) Guidelines for the Identification and Characterization of Plant Viruses. J Gen Virol 54: 223-241.
2. Clark MF, Adams AN (1977) Characteristics of the microplate method of enzyme-linked immunosorbent assay for the detection of plant viruses. J Gen Virol 34: 475-483.
3. Schaad NW, Frederick RD, Shaw J, Schneider WL, Hickson R, et al. (2003) Advances in molecular-based diagnostics in meeting crop biosecurity and phytosanitary issues. Ann Rev Phytopath 41: 305-324.
4. Damsteegt VD, Stone AL, Smith OP, McDaniel L, Sherman DJ, et al. (2013) A previously undescribed potyvirus isolated and characterized from arborescent Brugmansia. Arch. Virol.
5. Bar-Joseph M, Garnsey S, Gonsalves D, Moscovitz M, Purcifull DE, et al. (1979) The Use of Enzyme-Linked Immunosorbent Assay for Detection of *Citrus Tristeza Virus*. Phytopathology 69: 190-194.
6. Jordan R, Hammond J (1991) Comparison and differentiation of potyvirus isolates and identification of strain-, virus-, subgroup-specific and potyvirus group-common epitopes using monoclonal antibodies. J Gen Virol 72: 25-36.
7. Badge J, Brunt A, Carson R, Dagless E, Karamaglioli M, et al. (1996) A carlavirus-specific PCR primer and partial nucleotide sequence provides further evidence for the recognition of cowpea mild mottle virus as a whitefly-transmitted carlavirus. Euro J Pl Path 102: 305-310.
8. Dovas CI, Efthimiou K, Katis NI (2004) Generic detection and differentiation of tobamoviruses by a spot nested RT-PCR-RFLP using dI-containing primers along with homologous dG-containing primers. J Virol Methods 117: 137-144.
9. Untiveros M, Perez-Egusquiza Z, Clover G (2010) PCR assays for the detection of members of the genus Ilarvirus and family Bromoviridae. J Virol Methods 165: 97-104.
10. Zheng L, Rodoni BC, Gibbs MJ, Gibbs AJ (2010) A novel pair of universal primers for the detection of potyviruses. Plant Pathology 59: 211-220.
11. Luo H, Wylie SJ, Jones MG (2010) Identification of plant viruses using one-dimensional gel electrophoresis and peptide mass fingerprints. J Virol Methods 165: 297-301.
12. Boonham N, Tomlinson J, Mumford R (2007) Microarrays for rapid identification of plant viruses.. Ann Rev Phytopathology 45: 307-328.
13. Adams IP, Glover RH, Monger WA, Mumford R, Jackeviciene E, et al. (2009) Next-generation sequencing and metagenomic analysis: a universal diagnostic tool in plant virology. Mol Pl Pathol 10: 537-545.
14. Cox-Foster DL, Conlan S, Holmes EC, Palacios G, Evans JD, et al. (2007) A metagenomic survey of microbes in honey bee colony collapse disorder. Science 318: 283-287.
15. Palacios G, Druce J, Du L, Tran T, Birch C, et al. (2008) A new arenavirus in a cluster of fatal transplant-associated diseases. N Engl J Med 358: 991-998.
16. Tucker T, Marra M, Friedman JM (2009) Massively Parallel Sequencing: The next big thing in genetic medicine. Am J Hum Gent 85: 142-154.
17. Schadt EE, Turner S, Kasarskis A (2010) A window into third-generation sequencing. Hum Mol Genet 19: R227–R240.

18. Baulcombe D (2004) RNA silencing in plants. Nature 431: 356-363.

19. Vance V, Vaucheret H (2001) RNA silencing in plants--defense and counterdefense.. Science 292: 2277-2280.

20. van Mierlo JT, van Cleef KWR. van Rij RP (2010) Small Silencing RNAs: Piecing Together a Viral Genome. Cell Host Microbe 7: 87-89.

21. Wang XB, Jovel J, Udomporn P, Wang Y, Wu Q, et al. (2011) The 21-nucleotide, but not 22-nucleotide, viral secondary small interfering RNAs direct potent antiviral defense by two cooperative argonautes in Arabidopsis thaliana. Plant Cell 23: 1625-1638.

22. Kreuze JF, Perez A, Untiveros M, Quispe D, Fuentes S, et al. (2009) Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: A generic method for diagnosis, discovery and sequencing of viruses. Virology 388: 1-7.

23. Donaire L, Wang Y, Gonzalez-Ibeas D, Mayer KF, Aranda MA, et al. (2009) Deep-sequencing of plant viral small RNAs reveals effective and widespread targeting of viral genomes. Virology 392: 203-214.

24. Li R, Gao S, Hernandez AG, Wechter WP, Fei Z, et al. (2012) Deep sequencing of small RNAs in Tomato for virus and viroid identification and strain differentiation. PLoS One 7: e37127.

25. Loconsole G, Onelge N, Potere O, Giampetruzzi A, Bozan O, et al. (2012) Identification and characterization of *Citrus yellow vein clearing virus*, a putative new member of the genus Mandarivirus. Phytopathology 102: 1168-1175.

26. Loconsole G, Saldarelli P, Doddapaneni H, Savino V, Martelli GP, et al. (2012) Identification of a single-stranded DNA virus associated with citrus chlorotic dwarf disease, a new member in the family Geminiviridae. Virology 432: 162-172.

27. Roy A, Choudhary N, Guillermo LM, Shao J, Govindarajulu A, et al. (2013) A novel virus of the genus *Cilevirus* causing symptoms similar to citrus leprosis. Phytopathology 103: 488-500.

28. Kashif M, Pietilä S, Artola K, Jones R, Tugume AK, Mäkinen V, Valkonen J (2012) Detection of Viruses in Sweetpotato from Honduras and Guatemala Augmented by Deep-Sequencing of Small-RNAs. Plant Dis. 96:1430-1437.

29. Coetzee B, Freeborough MJ, Maree HJ, Celton JM, Rees DJ, et al. (2010) Deep sequencing analysis of viruses infecting grapevines: virome of a vineyard. Virology 400: 157-163.

30. Pantaleo V, Saldarelli P, Miozzi L, Giampetruzzi A, Gisel A, et al. (2010) Deep sequencing analysis of viral short RNAs from an infected Pinot Noir grapevine. Virology 408: 49-56.

31. Zhang Y, Singh K, Kaur R, Qiu W (2011) Association of a novel DNA virus with the grapevine vein-clearing and vine decline syndrome. Phytopathology 101: 1081-1090.

32. Pallett DW, Ho T, Cooper I, Wang H (2010) Detection of *Cereal yellow dwarf virus* using small interfering RNAs and enhanced infection rate with *Cocksfoot streak virus* in wild cocksfoot grass (*Dactylis glomerata*). J Virol Methods 168: 223-227.

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

34. Schulz MH, Zerbino DR, Vingron M, Birney E (2012) Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 15: 1086-1092.

35. Zerbino DR, Birney E (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18: 821-829.

36. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25: 3389-3402.

37. Compeau PE, Pevzner PA, Tesler G (2011) How to apply de Bruijn graphs to genome assembly. Nat Biotechnol. 29: 987-991.

38. Schroeder A, Mueller O, Stocker S, Salowsky R, Leiber M, Gassmann M, et al. (2006) The RIN: an RNA integrity number for assigning integrity values to RNA measurements. BMC Mol Biol 7: 3.

39. Bastianel M, Noveli VM, Kubo KS, Kitajima EW, Bassanezi R, et al. (2010) Citrus leprosis: Centennial of an unusual mite–virus pathosystem. Plant Dis 94: 284-292.

40. Wu Q, Luo Y, Lua R, Laub N, Lai EC, et al. (2010) Virus discovery by deep sequencing and assembly of virus-derived small silencing RNAs. PNAS USA 107:1606-1611.

41. Locali-Fabris EC, Freitas-Astúa J, Souza AA, Takita MA, Astúa-Monge G, et al. (2006) Complete nucleotide sequence, genomic organization and phylogenetic analysis of *Citrus leprosis virus* cytoplasmic type. J Gen Virol 87: 2721-2729.

42. Pascon RC, Kitajima JP, Breton MC, Assumpcao L, Greggio C, Zanca AS, et al. (2006) The complete nucleotide sequence and genomic organization of Citrus Leprosis associated Virus, cytoplasmatic type (CiLV-C). Virus Genes 32: 289-298.

43. Dean FB, Nelson JR, Giesler TL, Lasken RS (2001) Rapid amplification of plasmid and phage DNA using Phi29 DNA polymerase and multiply-primed rolling circle amplification. Genome Research 11: 1095-1099.

44. Huang Q, Hartung JS (2001) Cloning and Sequence analysis of an infectious clone of *Citrus yellow mosaic virus* that can infect sweet orange via *Agrobacterium*-mediated inoculation. J Gen Virol 82: 2549-2558.

45. Aregger M, Borah BK, Seguin J, Rajeswaran R, Gubaeva EG, et al. (2012) Primary and Secondary siRNAs in Geminivirus-induced Gene Silencing. PLoS Pathog 8: e1002941.

This article was originally published in a special issue, **Bioinformatics for Highthroughput Sequencing** handled by Editor: Dr. Heinz Ulli Weier, Lawrence Berkeley National Laboratory, USA