View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Elsevier - Publisher Connector

Virology 471-473 (2014) 54–60

Contents lists available at ScienceDirect

Virology

journal homepage: www.elsevier.com/locate/yviro

Development of a virus detection and discovery pipeline using next generation sequencing

Thien Ho, Ioannis E. Tzanetakis *

Department of Plant Pathology, Division of Agriculture, University of Arkansas System, Fayetteville, AR, USA

ARTICLE INFO

Article history:
Received 29 July 2014
Returned to author for revisions
28 August 2014
Accepted 22 September 2014
Available online 22 October 2014

Keywords:
Virus detection
Virus discovery
Next generation sequencing
Bioinformatics

ABSTRACT

Next generation sequencing (NGS) has revolutionized virus discovery. Notwithstanding, a vertical pipeline, from sample preparation to data analysis, has not been available to the plant virology community. We developed a degenerate oligonucleotide primed RT-PCR method with multiple barcodes for NGS, and constructed VirFind, a bioinformatics tool specifically for virus detection and discovery able to: (i) map and filter out host reads, (ii) deliver files of virus reads with taxonomic information and corresponding Blastn and Blastx reports, and (iii) perform conserved domain search for reads of unknown origin. The pipeline was used to process more than 30 samples resulting in the detection of all viruses known to infect the processed samples, the extension of the genomic sequences of others, and the discovery of several novel viruses. VirFind was tested by four external users with datasets from plants or insects, demonstrating its potential as a universal virus detection and discovery tool.

© 2014 Elsevier Inc. All rights reserved.

Introduction

Next generation sequencing (NGS) has revolutionized virology with many novel viruses being discovered using popular platforms such as pyrosequencing (454 Life Sciences, Branford, CT) or Illumina dye sequencing (Illumina, San Diego, CA) (Al Rwahnih et al., 2011; Quito-Avila et al., 2013; Rwahnih et al., 2013; Thekke-Veetil et al., 2013; Vives et al., 2013) (reviewed for plant diagnostics by Massart et al. (2014)). Most commercial NGS services offer basic bioinformatics support such as de novo sequence assembly or mapping to reference genomes, but will not progress further to the specifics of virus detection and discovery. There are also various online tools designed for general sequence comparison purposes, with NCBI BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi (Altschul et al., 1997)) as the most popular application that compares a limited number of query sequences to available subject databases such as non-redundant nucleotide (nt) and amino acid (nr) collections. In the case of PLAN (http://bioinfo.noble.org/plan/ (He et al., 2007)), users can create personal projects to Blast their datasets. NGS data can also be manipulated and analyzed in Galaxy (http://galaxyproject.org (Blankenberg et al., 2010)). However, NCBI BLAST and PLAN are Blast tools and only accept limited number of sequences in flat fasta format, and Galaxy, although more flexible with NGS data, is a collection of tools designed for sequence manipulation and analysis but not for novel virus discovery purposes.

There are bioinformatics tools developed specifically for human virus detection (Bhaduri et al., 2012; Chen et al., 2013; Kostic et al., 2011; Li et al., 2013; Naeem et al., 2013; Wang et al., 2013). In general, these tools are Unix command-line standalone packages that map NGS reads to the human genome, and perform various Blast steps to remove host reads. The remaining data are analyzed to categorize into non-human, microbial, or viral integrated sequences. Metavir2 (Roux et al., 2014) and Virome (Wommack et al., 2012) are the two other web-based tools for virome analysis but focus heavily on data visualization of environmental samples and do not focus on virus discovery. As there are no bioinformatics programs that function as universal virus discovery tools, biologists often have to rely on professional bioinformaticians to process NGS data, posing a bottleneck in data analysis.

In this study, a pipeline was created, from the bench to sequence analysis for virus detection and discovery. We developed a degenerate oligonucleotide primed (DOP) RT-PCR method with multiple barcodes for NGS, and constructed VirFind, a novel and automated bioinformatics tool specifically for virus detection and discovery. The tool has been tested for the past 2 years and is available as a web-based graphical front-end interface at http://virfind.org. VirFind efficiency was evaluated for virus detection and discovery using different NGS platforms on several plant and animal samples, sequenced in-house as well as by other research groups.

Results

A DOP-RT-PCR assay for multiplexed NGS

In this study, a DOP-RT-PCR assay with two different sets of primers (Table 2 and Table S1 for complete sets) was evaluated with

* Corresponding author. Tel.: +1 479 575 3180; fax: +1 479 575 7601.
E-mail addresses: txho@uark.edu (T. Ho), itzaneta@uark.edu (I.E. Tzanetakis).

http://dx.doi.org/10.1016/j.virol.2014.09.019
0042-6822/© 2014 Elsevier Inc. All rights reserved.

29 plant dsRNA-enriched samples (sample nos. 3–31). Each primer set comprised of an RT primer (with a random hexamer at the 3′ end) and 48 barcoded PCR primers, facilitating multiplexed NGS runs without the need of further barcoding by sequencing service provider. We experimented different sample combinations, from single sample NGS (dataset nos. 1, 4–8) to multiple barcoded sample NGS (datasets #3: 2 samples; #9: 4 samples; #10: 8 samples; and #2: 11 samples), and were able to retrieve sequences from all samples based on their barcodes.

Virus detection

Sample nos. 1, 5–8, 10, 13–21 and 31 (Table 1) were employed to test the VirFind detection efficiency. The pipeline detected all known viruses, including redbud yellow ringspot virus (*Emaravirus*, unassigned family) in *Cercis canadensis* (redbud); rose rosette virus (*Emaravirus*) in *Rosa* sp. (rose); beet pseudo-yellows virus (*Crinivirus*, *Closteroviridae*) and strawberry necrotic shock virus *(Ilarvirus, Bromoviridae*) in *Fragaria* × *ananassa* (strawberry); blackberry virus E (unassigned genus, *Alphaflexiviridae*), blackberry virus X (unassigned), blackberry vein banding-associated virus (*Ampelovirus*, *Closteroviridae*), blackberry yellow vein-associated virus (*Crinivirus*) and tobacco ringspot virus (*Nepovirus*, *Secoviridae*) in *Rubus* sp. (blackberry); fig badnavirus 1 (*Badnavirus*, *Caulimoviridae*), fig mild mottle-associated virus (*Closterovirus*, *Closteroviridae*) and fig mosaic virus (*Emaravirus*) in *Ficus carica* (fig); blueberry latent virus (*Amalgavirus*, *Amalgaviridae*), blueberry necrotic ring blotch virus (unassigned) in *Vaccinium corymbosum* (blueberry) and citrus yellow vein-associated virus (unassigned) in *Citrus* × *limon*. (lemon). In *Mentha* × *gracilis* (mint), VirFind detected mint virus X (*Potexvirus*, *Alphaflexiviridae*), strawberry latent ringspot virus (unassigned genus, *Secoviridae*) and mint vein banding-associated virus (MVBaV, unassigned genus, *Closteroviridae*). VirFind extended the known MVBaV genome from 9049 nt to 13,387 nt ((Tzanetakis et al., 2005), GenBank accession KJ572575). In *Vitis vinifera* (grapevine), VirFind assembled 6416 nt of RNA 1 of peach rosette mosaic virus (PRMV, *Nepovirus*, *Secoviridae*). Currently only sequences from PRMV RNA 1 are available in GenBank. VirFind discovered two contigs with total length of 2938 nt (GenBank accessions KJ572573-4) similar to the polyprotein encoded by nepovirus RNA 2. Since no RNA 1 of other nepoviruses was found, these two contigs are presumably part of PRMV RNA 2.

VirFind was also sensitive enough to detect correctly three random 270 nt virus/viroid GenBank molecules in datasets 4–6

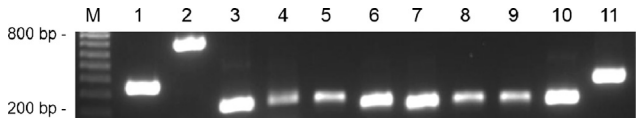

Fig. 2. Agarose gel electrophoresis of PCR confirming the presence of novel viruses identified using VirFind. Lanes 1–2: detection of novel trichovirus (DNA product = 325 bp) and novel waikavirus (DNA product = 640 bp), respectively, in blackcurrant; 3: detection of elderberry latent virus (DNA product = 217 bp) in elderberry; 4–10: detection of novel carlaviruses (DNA product = 181 bp) in elderberry; 11: detection of putative peach rosette mosaic virus RNA 2 (DNA product = 379 bp) in grape; M: Hyperladder IV molecular weight marker. Sanger sequencing confirmed virus identities.

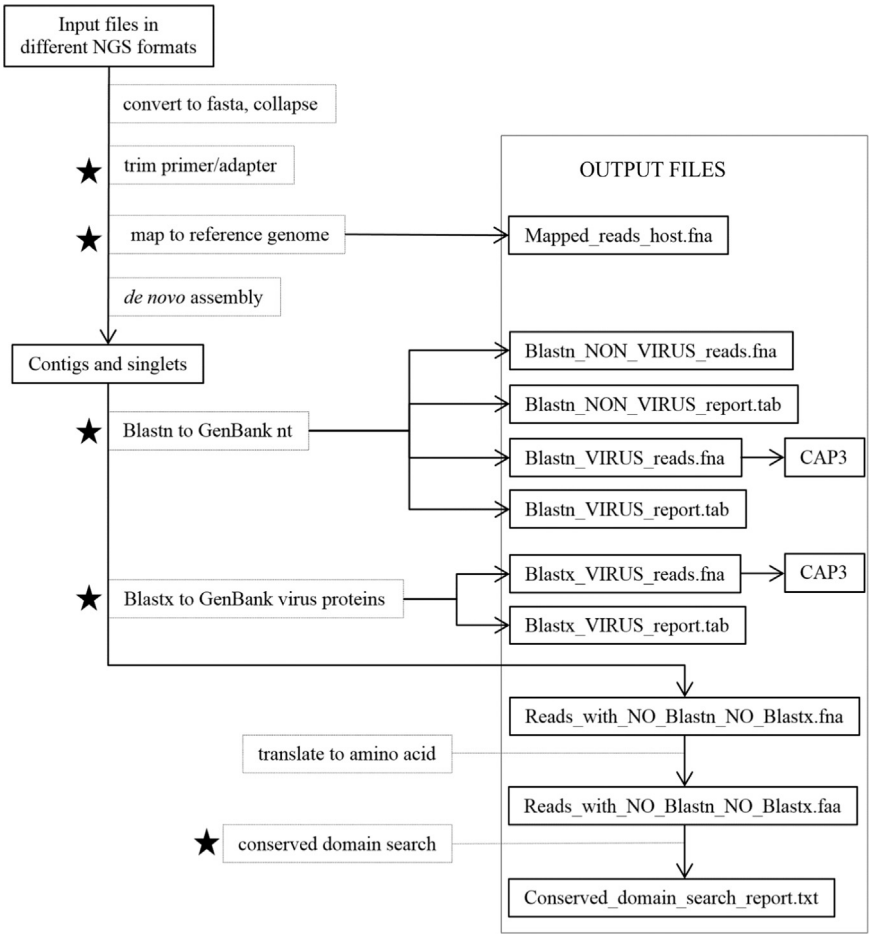

Fig. 1. VirFind flowchart for virus detection and discovery using next generation sequencing data. Each VirFind queue runs on a computer node with 64 cores and 512 Gb RAM, and uses various sequence manipulation tools, together with Bowtie 2 mapping, Velvet de novo assembler, NCBI BLAST and conserved domain search, to generate different outputs for users to find viruses in their next generation sequencing data. Stars indicate steps where users can set their own parameters.

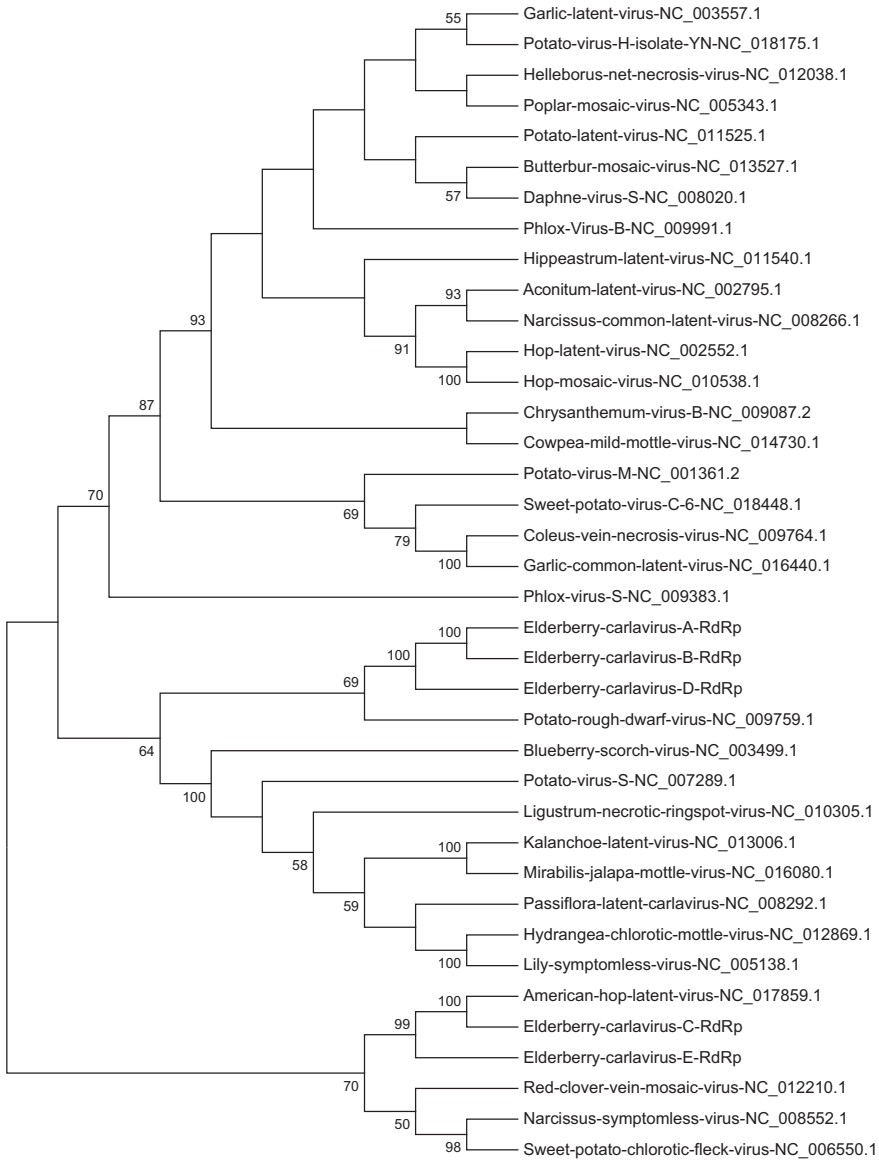

Fig. 3. Phylogenetic analysis of the RdRp amino acid sequences of five novel Elderberry carlaviruses among *Carlavirus* species with sequences deposited on GenBank. The unrooted tree was generated using the neighbor-joining algorithm of MEGA5 software (Tamura et al., 2011).

indicating that sequence subtraction by the Bowtie 2 genome mapping and Blastn steps (default cut-off e-value=0.05) did not erroneously remove any virus/viroid sequences.

Virus discovery

A novel trichovirus was discovered in *Ribes nigrum* (blackcurrant). The total contig length is 6057 nt (GenBank accessions KJ572565-6) and the partial RNA-dependent RNA polymerase (RdRp) shares about 30% identity with its trichovirus orthologs (Table S2A). A novel waikavirus, with total contig length of 6347 nt (GenBank accessions KJ572567–KJ572572), was also discovered, with the coat protein (CP) identities being less than 43% when comparing with its orthologs (Table S2B). Seven *Sambucus* (elderberry) samples were evaluated and VirFind extended the genome of elderberry latent virus (ElLV, unassigned genus, *Tombusviridae*) from 1017 nt to 3350 nt (GenBank accession KJ572576). Five novel carlaviruses with genome contig length ranging from 6044 to 8749 nt (GenBank accessions KJ572560–KJ572564) were discovered. All are new viruses given that the complete RdRp genes share less than the 80% amino acid sequence identities (ICTV species cut-off value) between them and other members of the genus. The viruses are provisionally named as elderberry carlaviruses A–E (Fig. 3 and Table S2C).

The presence of all novel viruses was verified by PCR amplification using detection primers designed from the assembled contigs (Table 2), with the exception of elderberry carlaviruses using the universal carlavirus detection primer pair kk-univ-carla (Karen Keller, personal communication), followed by Sanger sequencing (Fig. 2).

VirFind is a public tool for virus detection and discovery

Four external users evaluated VirFind using seven plants or honey bee Illumina/fasta datasets sequenced from either dsRNA, siRNA, or total RNA (NGS dataset nos. 11–17, Table 1). Without any prior knowledge about the tool, the users successfully signed up, logged in, and submitted the NGS datasets with specific parameters for the tool to execute at http://virfind.org. The users later confirmed the viruses detected by VirFind had been previously identified by their in-house bioinformatics analysis.

Discussion

VirFind, a bioinformatics pipeline for virus detection and discovery was developed. The program uses NGS data to identify known and unknown viruses and provide a robust pipeline for the end user. We evaluated different sample numbers in multiplexed NGS (2, 4, 8, and 11). In all cases, all viruses previously detected in the samples were also identified using VirFind. Unlike a chip-based method for virus detection (Chen et al., 2011), there is no need to update the hardware for DOP-RT-PCR.

Identifying a virus hit to GenBank nt or virus protein database is relatively simple in the case of long contigs with high sequence identity to known species. Still, it could be a challenging task in the case of short contigs and high e-values because of the possibility of false positives. VirFind generates Blast and conserved domain outputs with details empowering the users to make a decision on whether a known/novel virus is present in their sample. Based on the taxonomy information in the Blast reports together with e-values and sequence identity, users can infer whether a virus is a known species or a novel one.

Number of unique sequences submitted to the Blast filtering steps varied between 454 and Illumina sequencing methods (Table S3). As VirFind ignores sequences (after adapter and primer trimming) shorter than 90 nt, the majority of 454 singlets were processed further, whereas those from Illumina were filtered out. However, this will change as Illumina read length is constantly increasing.

Genomes of some of the host plants used in this study are still unavailable on the GenBank, hence the filtering steps remove a subset of host sequences, leaving a number of non-hit sequences. The majority were host sequences as identified in the conserved domain search. However, we observed cases where conserved domain search picked up a novel virus while earlier Blast steps did not (data not shown). Still, the rate of eukaryotic genome sequences available accelerates by the year and this may not be an issue altogether in a short timeframe.

Knowledge of virus evolution expands rapidly and it may be that novel virus sequences (e.g. archaea viruses) are quite different from those deposited in databases. In such case VirFind may be unable to identify them as such but as knowledge expands so will the ability of the pipeline to identify novel species.

Compared to other bioinformatics tools (Roux et al., 2014; Wommack et al., 2012), VirFind is better suited for processing of raw sequences for novel virus discovery. The tool can trim adapters/primers, and map to reference genomes before any sequence assembly and Blast steps. These steps are particularly useful with a random PCR or sample tagging protocols. Official virus taxonomy information obtained from ICTV master species list helps users identify the approximate taxa of the novel viruses.

Processing time varied between the experimental datasets, from as short as 3 min to about 70 h (Table 1) depending on the number of unique raw reads, average read length, and number of reads being processed after each filtering step. In general, 70 h are sufficient to complete the analysis of one 454 Junior sff dataset with ~200,000 reads, or an Illumina fastq dataset with ~30 million 80 nt single-end reads. Bigger datasets will need more processing power and we plan to upgrade VirFind hardware when user activity becomes significant.

A universal bioinformatics tool for virus discovery must have the ability to process datasets (i) with different NGS formats, (ii) of different read lengths, (iii) from different hosts, and (iv) infected by known and unknown virus species, closely or distantly related to those found on GenBank. The universality of VirFind was proven when the tool successfully worked with datasets in 454 sff/Illumina fastq/fasta format, processed from total nucleic acids/dsRNA, identified all categories of viruses, and generated by external users that used different preparation methods. VirFind was also used to identify viruses from siRNA datasets, which would allow the discovery of DNA viruses that leave siRNA footprints after infection. Originally VirFind was constructed to find plant viruses. However, since virus discovery by sequence comparison is the same regardless the virus species or host, and with the fact that VirFind was also tested successfully with honeybee viruses, the tool can be used for detection and discovery of viruses in any host.

Since online, VirFind was tested internally and externally using NGS datasets generated by the 454 or Illumina platforms. In all cases, VirFind produced virus detection/discovery results identical to or better than those previously identified by each individual user, demonstrating the reproducibility of the tool. Taken together, our results have shown that with VirFind, virus detection and discovery using NGS can be standardized and readily accessible to a wider audience of scientists in the absence of a designated bioinformatician.

Material and methods

Sample sources

Samples exhibiting virus-like symptoms were either plants maintained at the University of Arkansas-Fayetteville, or provided by collaborators in California, Michigan, Mississippi and Oregon. Laboratory tests (ELISA or RT-PCR) detected viruses in only a subset of samples, indicative of the presence of novel strains or species in others. Plant leaves or phloem were harvested and kept at −80 °C until nucleic acid extraction.

Sample preparation methods

>Thirty one plant samples (sample nos. 1–31, Table 1) were used for nucleic acid extraction. Sample nos. 1 and 2 were subjected to a total nucleic acid extraction protocol (Poudel et al., 2013), whereas sample nos. 3–31 were processed using a dsRNA-enrichment protocol (Yoshikawa and Converse, 1990). Samples 1, 5–8, 10, 13–21 and 31 were known to be infected by an array of viruses whereas samples 2–4, 9, 11, 12, 22–30 were never tested for viruses before. Reverse transcription was performed essentially as described before (Tzanetakis et al., 2005) using Maxima™ reverse transcriptase (Thermo Fisher Scientific, Waltham, MA) with 0.4 μM PDAP213′5 (emaravirus specific primer) (Di Bello and Tzanetakis, 2013) for samples 1 and 2, or BG4A-RT and KpnI-RT primers (0.4 μM each, Table 2) for samples 3–31. 5 μl of the cDNA were used in a 100 μl PCR reaction, with 0.8 μM PDAP213′5 primer for samples 1 and 2, or BG4A-PCR and KpnI-PCR primers (0.8 μM each, Table 2) for samples 3–31, and chemical composition as previously described (Poudel et al., 2013). The PCR program consisted of 2 min denaturation at 94 °C followed by 35 cycles of 20 s at 94 °C, 20 s at 45 °C, and 30 s at 72 °C, concluding with 10 min extension at 72 °C. The PCR products were visualized in 2% TBE-agarose gels stained with GelRed® (Biotium, Hayward, CA) and DOP-PCR products between 300 and 1000 bp were purified using GeneJET PCR Purification Kit (Thermo Fisher Scientific). DNA was quantified on a NanoDrop™ spectrophotometer (Thermo Fisher Scientific), normalized to the same amount for each sample, multiplexed as indicated in Table 1, and sequenced in 10 separate NGS reactions (NGS dataset nos. 1–10) using Illumina (Center for Genome Research and Biocomputing, Oregon State University, Corvallis, OR) or 454 Junior sequencing (Department of Biochemistry and Molecular Biology, Oklahoma State University, Stillwater, OK).

Development of VirFind

VirFind was developed as an automated tool to process NGS outputs. The pipeline is run on a Dell high performance computer node with AMD Opteron 6200 Series processors (64 cores) and 512 Gb

Table 1
List of samples and viruses detected/discovered using VirFind.

NGS dataset no.	File type	Run time (hh:mm)	Number of raw reads	Average sequence length (nt)	Sample no.	Host	Viruses detected/discovered[a]	PCR primers
1	sff	35:50	193,527	462	1	*Cercis canadensis*, *Rosa* sp.	Redbud yellow ringspot virus, rose rosette virus	PDAP213′5 (Di Bello and Tzanetakis, 2013)
2	sff	26:19	170,689	454	2	*C. canadensis*	None	PDAP213′5
					3	*Campsis* sp.	None	BG4A-I38-PCR
					4	*Campsis* sp.	None	BG4A-I39-PCR
					5	*Fragaria* × *ananassa*	Beet pseudo-yellows virus, strawberry necrotic shock virus	BG4A-I44-PCR
					6	*Fragaria* × *ananassa*	Strawberry necrotic shock virus	BG4A-I35-PCR
					7	*Fragaria* × *ananassa*	Beet pseudo-yellows virus	BG4A-I36-PCR
					8	*Fragaria* × *ananassa*	Strawberry necrotic shock virus	BG4A-I37-PCR
					9	*Lagerstroemia* sp.	None	BG4A-I43-PCR
					10	*Rubus* sp.	Blackberry vein banding-associated virus, blackberry yellow vein-associated virus, tobacco ringspot virus	BG4A-I17-PCR
					11	*Vaccinium* × *corymbosum*	None	BG4A-I31-PCR
					12	*Vaccinium* × *corymbosum*	None	BG4A-I32-PCR
3	sff	54:40	165,830	461	13	*Mentha* × *gracilis*	Mint vein banding-associated virus[b], mint virus X, strawberry latent ringspot virus	BG4A-I18-PCR
					14	*Rubus* sp.	Blackberry vein banding associated virus, blackberry virus X, blackberry yellow vein-associated virus	BG4A-I17-PCR, BG4A-I20-PCR, BG4A-I21-PCR
4[c]	fastq	02:08	29,089,718	80	15	*Ficus carica*	Fig badnavirus 1, fig mild mottle-associated virus, fig mosaic virus	KpnI-PCR
5[c]	fastq	02:41	33,294,974	80	16	*Glycine max*	Tobacco ringspot virus	KpnI-PCR
6[c]	fastq	03:03	25,632,788	80	17	*Rosa multiflora*	Blackberry chlorotic ringspot virus, rose rosette virus	KpnI-PCR
7	fastq	02:42	30,329,186	80	18	*Rubus* sp.	Blackberry virus E, blackberry yellow vein-associated virus	KpnI-PCR
8	fastq	03:35	30,820,330	80	19	*Vaccinium* × *corymbosum*	Blueberry latent virus, blueberry necrotic ring blotch virus	KpnI-PCR
9	sff	48:30	175,984	451	20	*Vaccinium* × *corymbosum*	Blueberry latent virus	BG4A-I47-PCR
					21	*Citrus* × *limon*	Citrus yellow vein-associated virus	BG4A-I7-PCR
					22	*Sambucus canadensis*	Elderberry carlavirus A[d], elderberry carlavirus B[d], elderberry carlavirus C[d], elderberry latent virus[b]	BG4A-I5-PCR
					23	*S. canadensis*	Elderberry carlavirus D[d]	BG4A-I6-PCR
10	sff	70:14	155,198	437	24	*Ribes nigrum*	Blackcurrant trichovirus A[d], blackcurrant waikavirus A[d]	BG4A-I1-PCR
					25	*S. canadensis*	Elderberry carlavirus D, elderberry latent virus	KpnI-I4-PCR
					26	*S. canadensis*	Elderberry carlavirus A, elderberry carlavirus B, elderberry carlavirus C, elderberry latent virus	BG4A-I5-PCR
					27	*S. canadensis*	Elderberry carlavirus D	BG4A-I6-PCR
					28	*S. nigra*	None	KpnI-I5-PCR
					29	*S. racemosa*	Elderberry carlavirus C, elderberry carlavirus D, elderberry carlavirus E[d]	KpnI-I3-PCR
					30	*S. racemosa* subsp. *sibirica*	Elderberry carlavirus C, elderberry carlavirus D	KpnI-I2-PCR
					31	*Vitis vinifera*	Peach rosette mosaic virus[b]	BG4A-I7-PCR
11[e]	fastq	48:29	8,483,017	50	32	Plant	*Betaflexiviridae* (5)	N/A
12[e]	fastq	46:53	9,432,449	50	33	Plant	None	N/A
13[e]	fasta	04:02	18,510,733	36	34	Plant	*Chrysovirus* (2), *Foveavirus* (1), *Maculavirus* (1), *Marafivirus* (1), *Mycovirus* (1)	N/A
14[e]	fasta	00:37	4,691,814	23	35	Plant	*Tymovirus* (1)	N/A
15[e,f]	fasta	00:03	178,519	23	36	*Apis mellifera*	Deformed wing iflavirus, varroa destructor iflavirus-1	N/A
16[e]	fasta	00:14	416,892	36	37	Plant	*Closterovirus* (1), *Idaeovirus* (1), *Potexvirus* (1), *Secoviridae* (1)	N/A
17[e]	fasta	21:23	1,501,204	100	38	Plant	*Closterovirus* (1), *Idaeovirus* (1), *Potexvirus* (1), *Secoviridae* (1)	N/A

[a] In this column, parentheses represent number of virus species.
[b] Known virus with incomplete genome sequence on GenBank, genome sequence extended using VirFind in this study.
[c] Each of these datasets was manually introduced one random virus or viroid sequence.
[d] Novel virus species.
[e] External user.
[f] Analysis of this dataset was published by the user (Wang et al., 2013).

RAM housed at Arkansas High Performance Computing Center. A detailed flowchart of the steps performed by VirFind is presented in Fig. 1. Briefly, NGS sequence files are converted to fasta format. Sequences are then trimmed at both 5′ and 3′ ends to remove any adapters and primers, and collapsed using FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit) and seq_crumbs (http://bioinf.comav.upv.es). Host sequences are removed from further processing after mapping to reference genomes using Bowtie 2 (Langmead and Salzberg, 2012). *De novo* sequence assembly is performed on unmapped reads using Velvet (Zerbino and Birney,

Table 2
List of oligos used in this study.

Primer name	Sequence
BG4A-RT[a]	CATTGCTGGGTGCCTGGTAAANNNNNN
KpnI-RT[b]	TGGTAGCTCTTGATCANNNNNN
BG4A-I1-PCR[c]	CGTGATCATTGCTGGGTGCCTGGTAAA
BG4A-I5-PCR[c]	CACTGTCATTGCTGGGTGCCTGGTAAA
BG4A-I6-PCR[c]	ATTGGCCATTGCTGGGTGCCTGGTAAA
BG4A-I7-PCR[c]	GATCTGCATTGCTGGGTGCCTGGTAAA
BG4A-I17-PCR[c]	CTCTACCATTGCTGGGTGCCTGGTAAA
BG4A-I18-PCR[c]	GCGGACCATTGCTGGGTGCCTGGTAAA
BG4A-I20-PCR[c]	GGCCACCATTGCTGGGTGCCTGGTAAA
BG4A-I21-PCR[c]	CGAAACCATTGCTGGGTGCCTGGTAAA
BG4A-I31-PCR[c]	ATCGTGCATTGCTGGGTGCCTGGTAAA
BG4A-I32-PCR[c]	TGAGTGCATTGCTGGGTGCCTGGTAAA
BG4A-I35-PCR[c]	AAAATGCATTGCTGGGTGCCTGGTAAA
BG4A-I36-PCR[c]	TGTTGGCATTGCTGGGTGCCTGGTAAA
BG4A-I37-PCR[c]	ATTCCGCATTGCTGGGTGCCTGGTAAA
BG4A-I38-PCR[c]	AGCTAGCATTGCTGGGTGCCTGGTAAA
BG4A-I39-PCR[c]	GTATAGCATTGCTGGGTGCCTGGTAAA
BG4A-I43-PCR[c]	GCTGTACATTGCTGGGTGCCTGGTAAA
BG4A-I44-PCR[c]	ATTATACATTGCTGGGTGCCTGGTAAA
BG4A-I47-PCR[c]	CTTCGACATTGCTGGGTGCCTGGTAAA
KpnI-PCR[c]	AGAGTTGGTAGCTCTTGATC
KpnI-I2-PCR[c]	ACATCGAGAGTTGGTAGCTCTTGATC
KpnI-I3-PCR[c]	GCCTAAAGAGTTGGTAGCTCTTGATC
KpnI-I4-PCR[c]	TGGTCAAGAGTTGGTAGCTCTTGATC
KpnI-I5-PCR[c]	CACTGTAGAGTTGGTAGCTCTTGATC
BCtricho-det[d]	Forward: CGGCTCTACTTCGAGTTCTTTC Reverse: CGGGCCGACAACCAATAATA
BCwaika-det[e]	Forward: CCCAAGAACTTGCTGGATAAGA Reverse: CACCACCTAGCATAGGCATTAG
ElLV-det[f]	Forward: CAGGAACTCCCGAGCTAAC Reverse: GGTCAACACCCTGACTCTT
PRMV-RNA2-det[g]	Forward: GCCAAAGAGGGCCATTATCT Reverse: GCACTCATCTCCAGAGCATTAT

[a] RT primer, used for DOP-PCR with BG4A-PCR primers.
[b] RT primer, used for DOP-PCR with KpnI-PCR primers.
[c] DOP-PCR primers. Underlines indicate barcode regions.
[d] Blackcurrant trichovirus PCR detection primer pair.
[e] Blackcurrant waikavirus PCR detection primer pair.
[f] Elderberry latent virus PCR detection primer pair.
[g] Peach rosette mosaic virus PCR detection primer pair for RNA 2.

2008) with k-mer (overlapping value)=31. For datasets with average sequence length $\leq$50 nt (primarily siRNA sequences), additional Velvet assemblies with k-mer=15 or 19 are constructed.

Short sequences may lead to false positives in Blast and for this reason only contigs and singlets of $\geq$90 nt are subjected to Blastn search against the GenBank nt database. Hits to GenBank nt are filtered out with virus and non-virus fasta reads together with their corresponding Blastn reports in tabular format. Sequences without any matches are then subjected to Blastx search against all GenBank virus protein sequences. CAP3 assemblies are also constructed on top of the Blast outputs. Official virus taxonomy information (order, family, subfamily, genus, species) derived from International Committee on Taxonomy of Viruses (ICTV) Master species list (http://talk.ictvonline.org/files/ictv_documents/m/msl/default.aspx) is presented on both reports. The remaining non-hit sequences are further processed using a Python script (http://cgpdb.ucdavis.edu/DNA_SixFrames_Translation) to translate all six frames which are consequently examined for the presence of conserved domains (Marchler-Bauer et al., 2009) against the NCBI Conserved Domain Database (CDD).

For the web interface submission, users need to complete a sequence submission form that contains the following options: (i) trimming of adapter/primer, (ii) mapping to reference genomes to remove host sequences, (iii) cut-off e-values of the Blastn and Blastx steps to define sequence relatedness to those found in GenBank, and (iv) conserved domain search of the remaining unmatched sequences. Sequence files are then uploaded to the VirFind ftp server for analysis. When all steps are completed, output files are compressed and mailed to users with information on how to download results from the server.

Virus detection and discovery

NGS dataset nos. 4–6 (Table 1) were each spiked with a random 270 nt virus/viroid GenBank sequence which was used to test the ability of VirFind to detect a single copy virus-like sequence. For sample nos. 22–27 and 29–30 where VirFind identified novel viruses, PCR primers were developed (Table 2) and used to amplify and sequence parts of the viruses' genomes, confirming their presence in individual samples.

Evaluation by external users

Four laboratories with experience analyzing NGS data as confirmed with multiple publications (Pallett et al., 2010; Quito-Avila et al., 2011; Rwahnih et al., 2013, 2012; Villamor et al., 2013; Villamor and Eastwell, 2013; Wang et al., 2013) evaluated VirFind independently using their own NGS datasets, previously analyzed and confirmed to contain an array of plant or animal viruses. The users were not provided with additional assistance other than that provided in the website.

Nucleotide sequence accession numbers

Sequences of the viruses used in this study have been deposited in GenBank under accession numbers KJ572560-76.

Funding

This work was supported by the USDA-APHIS-NCPN (Grant numbers 10-8100-1572, 11-8100-1572).

Acknowledgment

We thank Ananya Sharma and Ravi Barabote (University of Arkansas) for the advice of the pipeline automization, Patrick di Bello for the emaravirus primer, Karen Keller (USDA Corvallis, OR) for the universal carlavirus detection primers, Pawel Wolinski and the Arkansas High Performance Computing Center for providing resources utilized in this project, and users from the Tzanetakis and external labs for sending us the plant samples, testing VirFind with their NGS datasets, and critical comments about this manuscript.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.virol.2014.09.019.

References

Al Rwahnih, M., Daubert, S., Úrbez-Torres, J.R., Cordero, F., Rowhani, A., 2011. Deep sequencing evidence from single grapevine plants reveals a virome dominated by mycoviruses. Arch. Virol. 156, 397–403. http://dx.doi.org/10.1007/s00705-010-0869-8.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402. http://dx.doi.org/10.1093/nar/25.17.3389.

Bhaduri, A., Qu, K., Lee, C.S., Ungewickell, A., Khavari, P.A., 2012. Rapid identification of non-human sequences in high-throughput sequencing datasets. Bioinformatics 28, 1174–1175. http://dx.doi.org/10.1093/bioinformatics/bts100.

Blankenberg, D., Kuster, G. V., Coraor, N., Ananda, G., Lazarus, R., Mangan, M., Nekrutenko, A., Taylor, J. 2010. Galaxy: A Web-Based Genome Analysis Tool for Experimentalists. Curr. Protoc. Mol. Biol. 89:19.10:19.10.1–19.10.21.

Chen, E.C., Miller, S.A., DeRisi, J.L., Chiu, C.Y., 2011. Using a pan-viral microarray assay (Virochip) to screen clinical samples for viral pathogens. J. Vis. Exp. 50, 2536.

Chen, Y., Yao, H., Thompson, E.J., Tannir, N.M., Weinstein, J.N., Su, X., 2013. VirusSeq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue. Bioinformatics 29, 266–267. http://dx.doi.org/10.1093/bioinformatics/bts665.

Di Bello, P.L., Tzanetakis, I.E., 2013. *Rose rosette virus* is the causal agent of Rose rosette disease. Phytopathology 103 (S1), 3.

He, J., Dai, X., Zhao, X., 2007. PLAN: a web platform for automating high-throughput BLAST searches and for managing and mining results. BMC Bioinformatics 8, 53. http://dx.doi.org/10.1186/1471-2105-8-53.

Kostic, A.D., Ojesina, A.I., Pedamallu, C.S., Jung, J., Verhaak, R.G.W., Getz, G., Meyerson, M., 2011. PathSeq: software to identify or discover microbes by deep sequencing of human tissue. Nat. Biotechnol. 29, 393–396. http://dx.doi.org/10.1038/nbt.1868.

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359. http://dx.doi.org/10.1038/nmeth.1923.

Li, J.W., Wan, R., Yu, C.S., Co, N.N., Wong, N., Chan, T.F., 2013. ViralFusionSeq: accurately discover viral integration events and reconstruct fusion transcripts at single-base resolution. Bioinformatics 29, 649–651. http://dx.doi.org/10.1093/bioinformatics/btt011.

Marchler-Bauer, A., Anderson, J.B., Chitsaz, F., Derbyshire, M.K., DeWeese-Scott, C., Fong, J.H., Geer, L.Y., Geer, R.C., Gonzales, N.R., Gwadz, M., He, S., Hurwitz, D.I., Jackson, J.D., Ke, Z., Lanczycki, C.J., Liebert, C.A., Liu, C., Lu, F., Lu, S., Marchler, G.H., Mullokandov, M., Song, J.S., Tasneem, A., Thanki, N., Yamashita, R.A., Zhang, D., Zhang, N., Bryant, S.H., 2009. CDD: specific functional annotation with the Conserved Domain Database. Nucleic Acids Res. 37, D205–D210. http://dx.doi.org/10.1093/nar/gkn845.

Massart, S., Olmos, A., Jijakli, H., Candresse, T., 2014. Current impact and future directions of high throughput sequencing in plant virus diagnostics. Virus Res. 188, 90–96. http://dx.doi.org/10.1016/j.virusres.2014.03.029.

Naeem, R., Rashid, M., Pain, A., 2013. READSCAN: a fast and scalable pathogen discovery program with accurate genome relative abundance estimation. Bioinformatics 29, 391–392. http://dx.doi.org/10.1093/bioinformatics/bts684.

Pallett, D.W., 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. http://dx.doi.org/10.1016/j.jviromet.2010.06.003.

Poudel, B., Wintermantel, W.M., Cortez, A.A., Ho, T., Khadgi, A., Tzanetakis, I.E., 2013. Epidemiology of *Blackberry yellow vein associated virus*. Plant Dis. 97, 1352-–1357. http://dx.doi.org/10.1094/PDIS-01-13-0018-RE.

Quito-Avila, D.F., Brannen, P.M., Cline, W.O., Harmon, P.F., Martin, R.R., 2013. Genetic characterization of Blueberry necrotic ring blotch virus, a novel RNA virus with unique genetic features. J. Gen. Virol. 94, 1426–1434. http://dx.doi.org/10.1099/vir.0.050393-0.

Quito-Avila, D.F., Jelkmann, W., Tzanetakis, I.E., Keller, K., Martin, R.R., 2011. Complete sequence and genetic characterization of Raspberry latent virus, a novel member of the family *Reoviridae*. Virus Res. 155, 397–405. http://dx.doi.org/10.1016/j.virusres.2010.11.008.

Roux, S., Tournayre, J., Mahul, A., Debroas, D., Enault, F., 2014. Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. BMC Bioinformatics 15, 76.

Rwahnih, M.A., Dave, A., Anderson, M.M., Rowhani, A., Uyemoto, J.K., Sudarshana, M.R., 2013. Association of a DNA virus with grapevines affected by red blotch disease in California. Phytopathology 103, 1069–1076. http://dx.doi.org/10.1094/PHYTO-10-12-0253-R.

Rwahnih, M.A., Sudarshana, M.R., Uyemoto, J.K., Rowhani, A., 2012. Complete genome sequence of a novel *Vitivirus* isolated from grapevine. J. Virol. 86, 9545. http://dx.doi.org/10.1128/JVI.01444-12 (9545–9545).

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S., 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. 28, 2731–2739. http://dx.doi.org/10.1093/molbev/msr121.

Thekke-Veetil, T., Sabanadzovic, S., Keller, K.E., Martin, R.R., Tzanetakis, I.E., 2013. Molecular characterization and population structure of Blackberry vein banding associated virus, new ampelovirus associated with yellow vein disease. Virus Res. 178, 234–240.

Tzanetakis, I.E., Postman, J.D., Martin, R.R., 2005. A member of the *Closteroviridae* from mint with similarities to all three genera of the family. Plant Dis. 89, 654–658. http://dx.doi.org/10.1094/PD-89-0654.

Villamor, D.E.V., Druffel, K.L., Eastwell, K.C., 2013. Complete nucleotide sequence of a virus associated with rusty mottle disease of sweet cherry (*Prunus avium*). Arch. Virol. 158, 1805–1810. http://dx.doi.org/10.1007/s00705-013-1668-9.

Villamor, D.E.V., Eastwell, K.C., 2013. Viruses associated with rusty mottle and twisted leaf diseases of sweet cherry are distinct species. Phytopathology 103, 1287–1295. http://dx.doi.org/10.1094/PHYTO-05-13-0140-R.

Vives, M.C., Velázquez, K., Pina, J.A., Moreno, P., Guerri, J., Navarro, L., 2013. Identification of a new *Enamovirus* associated with citrus vein enation disease by deep sequencing of small RNAs. Phytopathology 103, 1077–1086. http://dx.doi.org/10.1094/PHYTO-03-13-0068-R.

Wang, H., Xie, J., Shreeve, T.G., Ma, J., Pallett, D.W., King, L.A., Possee, R.D., 2013. Sequence recombination and conservation of *Varroa destructor virus-1* and *Deformed wing virus* in field collected honey bees (*Apis mellifera*). PLoS One 8, e74508. http://dx.doi.org/10.1371/journal.pone.0074508.

Wang, Q., Jia, P., Zhao, Z., 2013. VirusFinder: software for efficient and accurate detection of viruses and their integration sites in host genomes through next generation sequencing data. PLoS One 8, e64465. http://dx.doi.org/10.1371/journal.pone.0064465.

Wommack, K.E., Bhavsar, J., Polson, S.W., Chen, J., Dumas, M., Srinivasiah, S., Furman, M., Jamindar, S., Nasko, D.J., 2012. VIROME: a standard operating procedure for analysis of viral metagenome sequences. Stand. Genomic Sci. 6, 427–439.

Yoshikawa, N., Converse, R.H., 1990. Strawberry pallidosis disease: distinctive dsRNA species associated with latent infections in indicators and in diseased strawberry cultivars. Phytopathology 80, 543–548.

Zerbino, D.R., Birney, E., 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18, 821–829. http://dx.doi.org/10.1101/gr.074492.107.