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.

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/

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

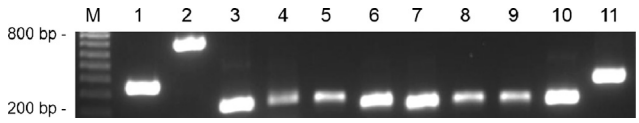

Fig. 2. Agarose gel electrophoresis of PCR confirming the presence of novel viruses identified using VirFind. Lanes 1–2: detection of novel trichovirus (DNA product = 325 bp) and novel waikavirus (DNA product = 640 bp), respectively, in blackcurrant; 3: detection of elderberry latent virus (DNA product = 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.

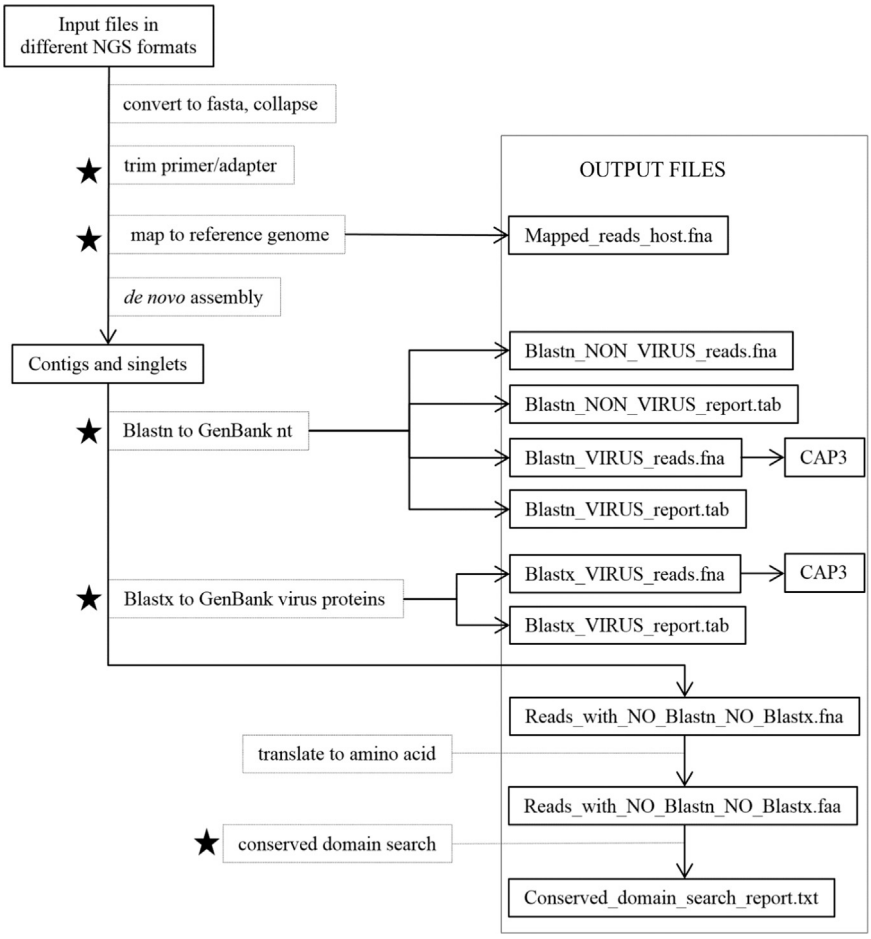

Fig. 1. VirFind flowchart for virus detection and discovery using next generation sequencing data. Each VirFind queue runs on a computer node with 64 cores and 512 Gb RAM, and uses various sequence manipulation tools, together with Bowtie 2 mapping, Velvet de novo assembler, NCBI BLAST and conserved 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.