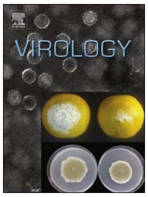
Virology 495 (2016) 122–128

Contents lists available at ScienceDirect

Virology

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



Targeted virus detection in next-generation sequencing data using an automated e-probe based approach

Marike Visser [a,b], Johan T. Burger [b], Hans J. Maree [a,b,*]

[a] Agricultural Research Council, Infruitec-Nietvoorbij: Institute for Deciduous Fruit, Vines and Wine, Stellenbosch, South Africa
[b] Department of Genetics, Stellenbosch University, Stellenbosch, South Africa

ARTICLE INFO

Article history:
Received 19 February 2016
Returned to author for revisions
5 May 2016
Accepted 12 May 2016

Keywords:
Bioinformatics
Grapevine
Grapevine leafroll disease
High-throughput sequencing
Virus diagnostics

ABSTRACT

The use of next-generation sequencing for plant virus detection is rapidly expanding, necessitating the development of bioinformatic pipelines to support analysis of these large datasets. Pipelines need to be easy implementable to mitigate potential insufficient computational infrastructure and/or skills. In this study user-friendly software was developed for the targeted detection of plant viruses based on e-probes. It can be used for both custom e-probe design, as well as screening preloaded probes against raw NGS data for virus detection. The pipeline was compared to *de novo* assembly-based virus detection in grapevine and produced comparable results, requiring less time and computational resources. The software, named Truffle, is available for the design and screening of e-probes tailored for user-specific virus species and data, along with preloaded probe-sets for grapevine virus detection.

1. Introduction

Efficient virus detection plays an important role in securing agricultural crop health. Metagenomic analyses of samples through next-generation sequencing (NGS) have been applied successfully to study virus populations in various plant species (Bi et al., 2012; Coetzee et al., 2010; Gu et al., 2014; Wylie et al., 2014). However, the introduction of NGS brought about large datasets, which pose various challenges to many biologists. Analyses may be limited by the lack of bioinformatic skills or due to inadequate computational resources. Several groups have developed pipelines addressing the limitations of NGS data analysis, which include publically available tools for virus detection (Ho and Tzanetakis, 2014; Roux et al., 2014; Wang et al., 2013; Zhao et al., 2013). The majority of these use a workflow, which include either the mapping of sequence reads against virus reference genomes, or the *de novo* assembly of reads and the subsequent identification of assembled contigs aligning to virus sequences present in databases. The latter has the advantage of discovering novel viruses. Both methods, however, are relatively time-consuming and require extensive computational resources and pre-processing of the data.

A novel approach for pathogen detection was recently developed which screens for viruses in NGS data with short unique pathogen-specific reference sequences, known as electronic-probes (e-probes) (Stobbe et al., 2013). E-probe design was based on an approach used for developing microarray probes, where unique pathogen regions are identified through sequence comparison to a closely related organism (Satya et al., 2008). Pathogen-specific regions were verified through subsequent sequence similarity-based screening of databases. Screening of highly specific e-probes against NGS data presented a faster and computationally less resource-intensive solution for focused virus detection (Stobbe et al., 2013). Implementation of this workflow still requires substantial bioinformatic skills.

In this study all the steps for e-probe based virus detection in NGS data were compiled into a single pipeline and packaged in a user-friendly interface, named Truffle (http://truffle.sourceforge.net). The software can design e-probes based on user-defined virus targets, or be used with preloaded probes. Probes were developed for 55 grapevine-infecting viruses with reference sequence data available on GenBank. Compared to virus detection based on *de novo* assembly, the simplified design and screening of these e-probes proved not only to be more time and computationally efficient, but also provided statistical strength for the presence of virus-specific sequences in NGS data.

2. Results

2.1. NGS datasets

Eighteen NGS datasets were generated from dsRNA extracted from grapevines displaying typical grapevine leafroll disease (GLD)

[*] Corresponding author at: Agricultural Research Council, Infruitec-Nietvoorbij: Institute for Deciduous Fruit, Vines and Wine, Stellenbosch, South Africa.
E-mail addresses: visserm@arc.agric.za (M. Visser), jtb@sun.ac.za (J.T. Burger), hjmaree@sun.ac.za (H.J. Maree).

http://dx.doi.org/10.1016/j.virol.2016.05.008
0042-6822/© 2016 Elsevier Inc. All rights reserved.

Table 1.
Summary of the raw data for each sample as well as processed data used for the in-house *de novo* assembly-based pipeline.

Sample number	Raw reads	In-house *de novo* assembly-based pipeline		
		Filtered reads	Contigs	Contigs (tblastx)
1	14,857,338	11,932,469	10,346	7080
2	35,618,188	32,911,522	624	303
3	12,472,948	10,294,369	2927	1734
4	18,365,984	17,397,432	264	40
5	16,442,566	13,322,984	7043	4651
6	22,011,476	18,868,794	13,807	8871
7	43,406,332	40,548,428	556	224
8	8,790,738	7,124,429	8182	5109
9	22,413,050	21,102,596	443	114
10	25,135,320	20,289,927	3744	2408
11	26,324,518	25,096,338	705	219
12	10,989,196	10,764,178	2199	188
13	6,972,098	5,836,204	473	195
14	1,442,480	1,310,503	3643	82
15	11,968,451	8,746,839	120	73
16	10,106,920	6,827,678	103	20
17	11,455,574	10,471,147	125	16
18	2,645,420	2,522,476	111	2

symptoms, as well as from asymptomatic rootstocks. The raw datasets range from ~1.4 million to ~43.4 million reads per sample and between ~1.3 million and ~40.5 million reads per sample after adapter removal, and quality trimming and filtering (Table 1).

2.2. De novo assembly, e-probe design and virus detection

Filtered reads were assembled into contigs, which were subsequently aligned against the GenBank nt database for virus identification. The number of contigs (250 nts or longer) ranged from 111 to 13,807 per sample (Table 1), with the largest contig being 18,571 nts in length. More than half (56.5%) of all contigs could not be annotated based on nucleotide identity (blastn) and were further analysed based on amino acid similarity (tblastx).

Truffle was used to design e-probes for 55 virus species (44 with complete genomes available) known to infect grapevine (Table 2). The number of probes varied from three to 199 with a cumulative probe length ranging from 123 to 9553 nts per virus. Due to the lack of reference sequence data or a suitable near-neighbour, probes could not be designed for grapevine Ajinashika virus (GAV), grapevine labile rod-shaped virus (GLRSV), grapevine line pattern virus (GLPV), grapevine stunt virus (GSV), grapevine Tunisian ringspot virus (GTRV) or raspberry bushy dwarf virus (RBDV).

Probe-based grapevine virus detection was compared to the *de novo* assembly-based detection pipeline. Together, the detection results revealed the presence of potentially 16 viruses in the samples (Table 3). All samples tested positive for grapevine leafroll-associated virus 3 (GLRaV-3) using both approaches. Grapevine virus A (GVA) and grapevine endophyte endornavirus (GEEV) were also prevalent in the samples. There were some discrepancies between the results of the two approaches. VirFind detected grapevine leafroll-associated virus 2 (GLRaV-2) homologous sequences in more samples than both the in-house *de novo* assembly-based pipeline and the e-probe based pipeline along with other viruses such as grapevine anatolian ringspot virus (GARSV), tomato mosaic virus (ToMV) and tobacco ringspot virus (TRSV). In samples with conflicting results the genome coverage for these four viruses was particularly low (Supplementary Table). VirFind, however, failed to detect grapevine virus F (GVF) and grapevine endophyte endornavirus (GEEV), despite up to 90% and 100% genome coverage obtained in some samples for these viruses, respectively. The e-probe based approach, on the other hand, identified more samples infected with grapevine rupestris stem-pitting-associated virus (GRSPaV) than *de novo* assembly and sequence similarity searches, despite relatively low genome coverage (~10%). Samples suspected to be positive for GVA or tobacco mosaic virus (TMV), in most cases, had lower genome coverage than positive samples.

2.3. Intra-species genetic variation and virus detection

To determine the effect of host genome selection on the sensitivity of genetic variant detection, the samples were screened with e-probes designed for divergent variants of GLRaV-3, GVA and grapevine virus B (GVB). For each of these species the results were variable (Table 4). Some samples had the same predicted result for a virus species, irrespective of the variant probe-set used. However, for other samples the result depended on the probe-set used. For GLRaV-3, it was clear that group VII variants, in particular, are too divergent for a single probe-set to detect all variant groups.

2.4. Truffle: a user-friendly pipeline and interface for targeted virus detection

Truffle provides a bioinformatic pipeline and graphical user interface (GUI) to a previously described workflow (Stobbe et al., 2013). It is functional on computers operated by OS X or Ubuntu, with at least 4 GB RAM. To initiate the screening of a sample takes less than a minute hands-on time. Using an OS X operated laptop with 16 GB RAM and a 2.5 GHz Intel Core i7 processor, sample 7 (with 43,406,332 raw reads) could be screened with the 69 probe-sets (listed in Table 2) in 2 h and 27 min, while sample 14 (with 1,442,480 raw reads) could be screened in only 6 min. The software, along with the grapevine virus e-probes, and previously designed citrus virus probes (unpublished), have been made available online for download (http://truffle.sourceforge.net). Truffle can be used to design custom, virus-specific e-probes, and to search for viruses in NGS data with these or pre-loaded probes.

3. Discussion

Currently the identification of viruses through NGS comprises either large-scale alignment of reads against nucleotide databases or *de novo* assembly thereof, followed by alignment analysis of numerous contigs against a large database. The latter approach decreases the number of query sequences, thus reducing the scale of alignment analysis, as well as the number of potential false-positives, which could result from short query lengths. While these traditional approaches enable the discovery of unexpected or novel viruses in existing NGS data, they have a few shortcomings. Extensive computational power is required for both assemblies and sequence similarity searches. Aligning NGS reads or contigs against large databases may take days to complete, while submitting data online to available servers can be as time-consuming. Self-implementation of these pipelines often require computational skills such as running command-line based programs or, even more challenging, parsing data to extract relevant information.

Other approaches to enhance the analysis of NGS datasets have been developed and are discussed in a review by Melcher et al. (2014). These include optimising computational speed through parallelizing analyses, the screening of data against focused databases, as well as the implementation of the NGS data as a searchable database against which target-specific e-probes are