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
Available online 20 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

Table 2
List of grapevine viruses used for e-probe design.

Virus	Abbreviation	Target Accession	Near-neighbour	Near-neighbor Accession	Final number of e-probes	Total probe length
Alfalfa mosaic virus	AMV	NC_001495 NC_002024 NC_002025	Cucumber mosaic virus	NC_001440 NC_002034 NC_002035	66	4503
Arabis mosaic virus	ArMV	NC_006056 NC_006057	Tobacco ringspot virus	NC_005096 NC_005097	79	3962
Artichoke Italian latent virus[a]	AILV	X87254	Beet ringspot virus	D00322	11	562
Bean common mosaic virus	BCMV	NC_003397	Potato virus Y	NC_001616	54	2559
Beet cryptic virus 3[a]	BCV-3	S63913	Pepper cryptic virus 1	JN117276	10	1486
Blackberry virus S[a]	BlVS	FJ915122	Maize rayado fino virus	NC_002786	46	2740
Blueberry leaf mottle virus[a]	BBLMV	U20621 U20622	Tobacco ringspot virus	NC_005096 NC_005097	32	1460
Broad bean wilt virus 1	BBWV-1	NC_005289 NC_005290	Broad bean wilt virus 2	NC_003003 NC_003004	99	4407
Broad bean wilt virus 2	BBWV-2	NC_003003 NC_003004	Broad bean wilt virus 1	NC_005289 NC_005290	76	3401
Carnation mottle virus	CarMV	NC_001265	Saguaro cactus virus	NC_001780	23	2064
Cherry leafroll virus	CLRV	NC_015414 NC_015415	Tobacco ringspot virus	NC_005096 NC_005097	112	6560
Cucumber mosaic virus	CMV	NC_002034 NC_002035 NC_001440	Peanut stunt virus	NC_002038 NC_002039 NC_002040	65	3900
Grapevine Algerian latent virus	GALV	NC_011535	Tomato bushy stunt virus	NC_001554	20	1083
Grapevine Anatolian ringspot virus	GARSV	NC_018383 NC_018384	Tobacco ringspot virus	NC_005096 NC_005097	92	4845
Grapevine angular mosaic virus[a]	GAMoV	AY590305	Tobacco streak virus RNA2	NC_003842	3	212
Grapevine asteroid mosaic-associated virus[a]	GAMaV	AJ249357	Grapevine Syrah virus 1	NC_012484	15	601
Grapevine berry inner necrosis virus	GINV	NC_015220	Apple chlorotic leaf spot virus	NC_001409	65	3207
Grapevine Bulgarian latent virus	GBLV	NC_015492 NC_015493	Tobacco ringspot virus	NC_005096 NC_005097	110	6343
Grapevine chrome mosaic virus	GCMV	NC_003621 NC_003622	Tobacco ringspot virus	NC_005096 NC_005097	101	4690
Grapevine deformation virus	GDefV	NC_017938 NC_017939	Tobacco ringspot virus	NC_005096 NC_005097	59	2054
Grapevine endophyte Endornavirus	GEEV	NC_019493	Chalara endornavirus CeEV1	GQ494150	137	7620
Grapevine fanleaf virus	GFLV	KC900162 KC900163	Tobacco ringspot virus	NC_005097 NC_005096	86	3602
Grapevine fleck virus	GFkV	NC_003347	Fig fleck-associated virus	FM200426	51	1947
Grapevine leafroll-associated virus 1	GLRaV-1	NC_016509	Grapevine leafroll-associated virus 3	NC_004667	195	9449
Grapevine leafroll-associated virus 2	GLRaV-2	NC_007448	Beet yellows virus	NC_001598	150	7014
Grapevine leafroll-associated virus 3	GLRaV-3(GP18)	EU259806	Blackberry vein banding associated virus	NC_022072	192	8530
	GLRaV-3(GH24)	KM058745			199	9553
	GLRaV-3(GH30)	JQ655296			198	9214
	GLRaV-3(PL-20)	GQ352633			187	8336
	GLRaV-3(621)	GQ352631			199	8597
Grapevine leafroll-associated virus 4 (5, 6, 9)	GLRaV-4	NC_016416	Grapevine leafroll-associated virus 3	NC_004667	157	7219
Grapevine leafroll-associated virus 7	GLRaV-7	NC_016436	Little cherry virus 1	NC_001836	143	5717
Grapevine Pinot gris virus	GPGV	NC_015782	Apple chlorotic leaf spot virus	NC_001409	64	2965
Grapevine red blotch associated virus	GRBaV	NC_022002	Maize streak virus	NC_001346	6	375
Grapevine redglobe virus[a]	GRGV	AF521977	Grapevine fleck virus	NC_003347	16	755
Grapevine rupestris stempitting-associated virus	GRSPaV	NC_001948	Apple stem pitting virus	NC_003462	91	4571
Grapevine rupestris vein feathering virus[a]	GRVFV	AY706994	Maize rayado fino virus	NC_002786	58	3020
Grapevine Syrah virus 1	GSV-1	NC_012484	Maize rayado fino virus	NC_002786	57	2903
Grapevine vein clearing virus	GVCV	NC_015784	Commelina yellow mottle virus	NC_001343	72	4905
Grapevine virus A	GVA(IS151)	NC_003604	Grapevine virus B	NC_003602	61	4060
	GVA(PA3)	AF007415			66	4717
	GVA(GTR1-1)	DQ787959			64	4345
	GVA(GTR1-2)	DQ855086			72	4517
Grapevine virus B	GVB(Ref)	NC_003602	Grapevine virus A	NC_003604	63	4538
	GVB(H1)	GU733707			54	3227
	GVB(QMWH)	KF700375			71	4613
Grapevine virus D[a]	GVD	Y07764	Grapevine virus A	NC_003604	10	570
Grapevine virus E	GVE	GU903012	Grapevine virus A	NC_003604	63	4019
Grapevine virus F	GVF	NC_018458	Grapevine virus A	NC_003604	56	4210
Peach rosette mosaic virus[a]	PRMV	AF016626	Tobacco ringspot virus	NC_005097	52	3217
Petunia asteroid mosaic virus[a]	PAMV	AY500881	Tomato bushy stunt virus	NC_001554	3	123
Potato virus X	PVX	NC_011620	Potato virus Y	NC_001616	52	3185
Raphanus sativus cryptic virus 3	RsCV-3	NC_011705 NC_011706	White clover cryptic virus 1	NC_006275 NC_006276	15	1707
Raspberry ringspot virus	RpRSV	NC_005266 NC_005267	Tobacco ringspot virus	NC_005096 NC_005097	104	5941
Southern tomato virus	STV	NC_011591	Rhododendron virus A	NC_014481	17	2189
Sowbane mosaic virus	SoMV	NC_011187	Southern bean mosaic virus	NC_004060	26	3255

Table 2 (*continued*)

Virus	**Abbreviation**	**Target Accession**	**Near-neighbour**	**Near-neighbor Accession**	**Final number of e-probes**	**Total probe length**
Strawberry latent ringspot virus	SLRSV	NC_006964 NC_006965	Tobacco ringspot virus	NC_005096 NC_005097	114	6116
Tobacco mosaic virus	TMV	NC_001367	Rehmannia mosaic virus	NC_009041	18	573
Tobacco necrosis virus D	TNV-D	NC_003487	Beet black scorch virus	NC_004452	27	2563
Tobacco ringspot virus	TRSV	NC_005096 NC_005097	Grapevine fanleaf virus	KC900162 KC900163	99	5188
Tomato black ring virus	TBRV	NC_004439 NC_004440	Tobacco ringspot virus	NC_005096 NC_005097	107	5215
Tomato mosaic virus	ToMV	NC_002692	Tobacco mosaic virus	NC_001367	39	1266
Tomato ringspot virus	ToRSV	NC_003839 NC_003840	Tobacco ringspot virus	NC_005096 NC_005097	115	6174
Tomato spotted wilt virus	TSWV	NC_002050 NC_002051 NC_002052	Groundnut bud necrosis virus	NC_003614 NC_003619 NC_003620	122	4163

[a] Partial genome.

Table 3
Summary of the viruses detected with each bioinformatics pipeline.

Sample number	***De novo*** **assembly-based pipeline**			**Truffle**[a]	
	VirFind only	In-house pipeline only	Both	Viruses detected[b]	Suspected positive[c]
1	GLRaV-2, GVE, GFLV	GEEV	GLRaV-3	GLRaV-3, **GVA**, GVE, **GRSPaV**, GEEV	
2	GLRaV-2, GRSPaV, TMV, GARSV, ToMV	GEEV	GLRaV-3, GVA, GVB, GVE	GLRaV-3, GVA, GVB, GVE, **GVF**, GRSPaV, GEEV, TMV	
3	GLRaV-2, GVE		GLRaV-3	GLRaV-3, GVE, **GRSPaV, GEEV**	**GVA**
4		GVF, GEEV	GLRaV-3, GVA, GVE, GFLV	GLRaV-3, GVA, GVE, GVF, GFLV, GEEV	
5	GLRaV-2, GARSV	GEEV	GLRaV-3, GVE	GLRaV-3, GVE, **GRSPaV**, GEEV	**GVA**
6	GLRaV-2, GVB, GRSPaV, GBLV, TRSV	GVF	GLRaV-3, GVA, GVE	GLRaV-3, GVA, GVE, GVF, GRSPaV, **GEEV**	
7	GLRaV-2, GVA, GRSPaV, TMV	GVF	GLRaV-3, GVE	GLRaV-3, GVE, GVF, GRSPaV, **GEEV**	GVA
8	GLRaV-2, GRSPaV, RpRSV	GEEV	GLRaV-3, GVE	GLRaV-3, GVE, GRSPaV, GEEV	
9		GVF	GLRaV-2, GLRaV-3, GVA, GVE, GFkV	GLRaV-2, GLRaV-3, GVA, GVE, GVF, GFkV, **GEEV**	
10	GLRaV-2, GRSPaV		GLRaV-3, GVE	GLRaV-3, **GVA**, GVE, GRSPaV, GEEV	
11	TMV	GEEV	GLRaV-3, GVA, GRSPaV, GFkV	GLRaV-3, GVA, GRSPaV, GFkV, GEEV	TMV
12			GLRaV-3	GLRaV-3	
13	GVA, GFLV		GLRaV-3	GLRaV-3	
14	GVE		GLRaV-3	GLRaV-3, GVE	
15	GVA, GVE	GEEV	GLRaV-3	GLRaV-3, GVA, GVE, GEEV	
16		GEEV	GLRaV-3, GVA	GLRaV-3, GVA, GEEV	
17	GVA, GVE	GEEV	GLRaV-2, GLRaV-3, GVB,	GLRaV-2, GLRaV-3, GVB, GEEV	GVA
18		GVF	GLRaV-2, GLRaV-3, GVA	GLRaV-2, GLRaV-3, GVA, GVF	

[a] Viruses highlighted in bold were only detected with Truffle.

[b] p-value ≤ 0.05.

[c] p-value > 0.05 to 0.1.

Table 4.
Results for virus-detection analysis performed with e-probes designed for different GLRaV-3, GVA and GVB genetic variants.

Sample number	**GLRaV-3**					**GVA**				**GVB**		
	621 (group I)	GP18 (group II)	PL-20 (group III)	GH30 (group VI)	GH24 (group VII)	IS 151	PA3	GTR1-1	GTR1-2	Ref	H1	QMWH
1	+	+	+	+				+				
2	+	+	+	+	Suspected	+	+	+	+	+	+	+
3	+	+	+	+				Suspected				
4	+	+	+	+	Suspected	+	+	+	+			
5	+	+	+	+				Suspected				
6	+	+	+	+	Suspected	+	+	+	+			
7	+	+	+	+				Suspected				
8	+	+	+	+								
9	+	+	+	+		+	+	+	+			
10	+	+	+	+				+				
11	+	+	+	+	Suspected	+	+	+	+			
12					+							
13					+							
14	+	+	+	+								
15	+	+	+	+	+			+				
16	+	+	+	+	+	+	Suspected	+				
17	+	+	+	+	+		Suspected				+	
18	+	+	+	+	+	+	+	+	Suspected			

screened. Considering the targeted detection of known pathogens, the recent development of this e-probe based approach to screen for known pathogens has proven to be more effective than the *de novo* assembly-based approach (Stobbe et al., 2013). Implementation of this workflow requires running different freely available software, while users also have to create their own bioinformatic tools/scripts to parse the intermediate output data and perform statistical analysis. The software developed in this study, named Truffle, provides a user-friendly interface that can be applied, without training, for the targeted detection of known viruses based on e-probe screening against raw NGS data. The system can be executed on computers operated by OS X and Ubuntu, circumventing the need for a high-performance cluster (HPC).

Grapevine is host to a wide variety of infectious agents, which include more than 60 viruses (Martelli, 2014). Truffle was used to develop probes for 55 known grapevine viruses, which in turn were used to screen diseased grapevine plants for virus infection. The current study focussed on a recently compiled list of grapevine viruses (Martelli, 2014), however, the list of e-probes can easily be extended if probes for other viruses known to infect grapevine, are required.

Comparing the virus-detection results obtained from Truffle to those of the *de novo* assembly-based pipeline showed that the e-probe based approach was mostly comparable and in some cases (such as the detection of GRSPaV) seemed to be more sensitive. The representation of a virus genome within NGS data is dependent on the biological properties of the virus species and the amount of sequence data generated. Moreover, not all virus-derived reads (reads that can map onto a virus genome) can be assembled into contigs. Therefore, although there may be virus-derived reads in the NGS data, detectable by e-probes, these reads may not assemble into contigs making the virus undetectable by the *de novo* assembly-based pipeline. One advantage of the e-probe based pipeline over the *de novo* assembly-based pipeline is the statistical support for the presence of virus-specific regions in NGS data, while the *de novo* assembly-based pipeline relies on the discretion of the user when making a virus detection call. As expected, the focused approach of the e-probe based pipeline seems less sensitive (Table 3) to genome coverage (Supplementary Table) since it targets unique regions of a specific virus. The *de novo* assembly-based pipeline, on the other hand, may be complicated by the fact that these pipelines focus on homology-based searches that could recognise conserved regions, which are not necessarily unique to the specific virus. Virus regions covered remain to be validated for specificity.

Probe efficacy and their ability to reliably detect viruses are influenced by a number of factors during probe design. As can be seen from the results, in species where divergent variants occur (such as GLRaV-3, GVA and GVB in this study), using different genetic variants for probe design yielded different results with regard to the virus status of the plants. This result highlighted the importance of target genome choice in ensuring accurate detection results. Prior knowledge of virus species (divergent variants and possibly also their prevalence) is therefore needed to select either the appropriate variant for probe design or to design probes for multiple variants where substantial intra-species genetic variation occurs. In the current version of Truffle, multiple probe-sets, designed from different variants, cannot be used for variant calling since some probes will be universal amongst the probe-sets.

Another aspect, which may influence probe sensitivity, is the status of the genome used. To further prevent potential false negative results, it is important to use full target reference genome sequences (or as complete as possible) to represent the majority of target-specific genomic regions. The e-probes designed in this study using incomplete genomic sequence data therefore need to be redesigned once the full genomes become available. Lastly, candidate probes are filtered against NCBI's nucleotide database in order to determine virus specificity. Since only probes aligning to sequences with exactly matched virus names in the database (matches with the same spelling) are retained, it is important to include all synonyms and nomenclature conventions when designing the probes. Even with great precaution some probes may still be removed due to unforeseen mistakes in the database entries.

One of the main advantages of Truffle is that it can easily be applied for the detection of other viruses. E-probe based virus detection have previously been applied for the detection of viruses in peaches and beans (Stobbe et al., 2014). During optimisation of the software parameters, Truffle was also used to successfully develop probes for nine citrus-infecting viruses and to screen citrus plants of known and unknown infection status (unpublished). The pipeline can additionally be applied not only to dsRNA sequencing data but also to other NGS data such as RNA-Seq.

To conclude, an easy-to-use software, named Truffle, was developed for the screening of NGS data for known viruses. The analyses performed are both time and resource effective and can be run from a desktop or laptop computer by an inexperienced person. The results on grapevine virus detection presented here, support e-probe based diagnostics as an efficient approach for targeted virus detection. Truffle was made publicly available and can be applied to design tailored probes and screen users' NGS data. E-probes designed for grapevine and citrus viruses, with full genomes available, also comes preloaded in the Truffle download for users to apply.

4. Materials and methods

4.1. NGS data preparation

Using a protocol described by Burger and Maree (2015) dsRNA was extracted from the phloem tissue of 14 grapevines displaying typical grapevine leafroll disease symptoms and 4 asymptomatic rootstocks. Sequencing libraries were prepared using an adapted Illumina TruSeq Stranded Total RNA Library Prep Kit (Burger and Maree, 2015) and were sequenced on either an Illumina HiSeq, HiScanSQ or MiSeq instrument. Data were trimmed and quality filtered using Trimmomatic (Bolger et al., 2014). A head crop of 9 nts was performed and reads were trimmed at the 3′ end when the quality score was lower that 20 (slidingwindow-4, Q20).

4.2. De novo genome assembly-based virus detection

Trimmed reads were assembled into contigs using CLC Genomics Workbench 8. The minimum contig length was set to 250 nts while automatic bubble-size and word-size detection was applied.

To determine the viral status of the samples all contigs were first aligned using blastn from Blast+ (Camacho et al., 2009) against GenBank's nt database, using default parameters. Contigs, which could not be annotated with blastn, were further analysed using tblastx against the same nt database also using default parameters. Filtered reads were additionally submitted to VirFind (Ho and Tzanetakis, 2014), applying default parameters, to determine the viral status.

4.3. Truffle development

Truffle is an interface developed in Python to detect virus sequences in NGS data through designing and implementing virus-specific e-probes. The bioinformatic pipeline, based on the TOFI-derived (Satya et al., 2008) pipeline called EDNA (Stobbe et al.,

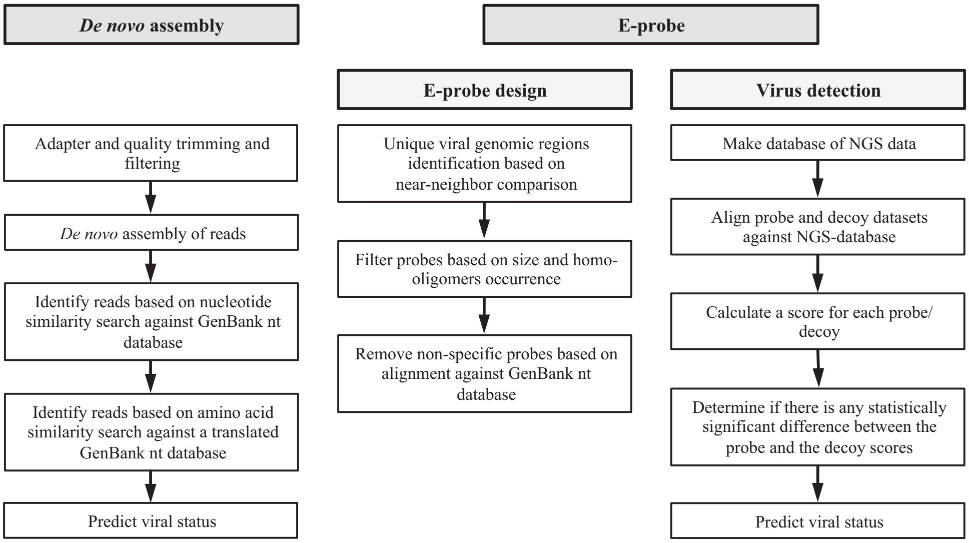
De novo assembly
Adapter and quality trimming and filtering
De novo assembly of reads
Identify reads based on nucleotide similarity search against GenBank nt database
Identify reads based on amino acid similarity search against a translated GenBank nt database
Predict viral status
E-probe
E-probe design
Unique viral genomic regions identification based on near-neighbor comparison
Filter probes based on size and homo-oligomers occurrence
Remove non-specific probes based on alignment against GenBank nt database
Virus detection
Make database of NGS data
Align probe and decoy datasets against NGS-database
Calculate a score for each probe/decoy
Determine if there is any statistically significant difference between the probe and the decoy scores
Predict viral status

Fig. 1. Diagram illustrating the workflow of the two NGS-based virus detection approaches used in the study. In the *de novo* assembly-based approach reads are first assemble into contigs followed by virus identification through similarity searches. In the e-probe based approach, virus e-probes are first designed to be specific to a target virus. E-probes are then screened against NGS data along with decoy sequences to determine the virus status of the sample. A positive call is based the statistical difference between scores calculated for probes and decoys with a positive NGS-database hit.

2013), is outlined in Fig. 1. Firstly, probes can be designed which are customised for a user's specific virus species of interest. For probe design the genome of the target virus is first compared to that of a closely related virus (Table 2) using NUCmer (-c 10, -l 10, -g 0, –noextend, –maxmatch, –nosimplify), which forms part of the MUMmer package (Delcher et al., 1999, 2002; Kurtz et al., 2004), to identity homologous genomic regions. Unique target-specific regions, 20 nts and longer, are extracted and serve as candidate probes. After removing sequences containing homo-oligomers of more than 4 nts in length the candidate probes are aligned to NCBI's online GenBank nt database (word size 7, gap cost to open 5 and to extend 2, reward 1, penalty -3), removing all probes that hit any sequence other than the virus of interest. An alignment with an e-value of 1×10^{-3} or less is considered a hit. The remaining probes form the virus-specific e-probes. A decoy set of sequences is also created, which comprises of the reverse sequences of the e-probes.

For the second application of Truffle, Blast+ (Camacho et al., 2009) is used determine the viral status of a sample. The probes and decoys are aligned, with blastn (-task blastn-short), against a database composed of the raw NGS data. A score is generated for each probe and decoy based on the number of hits, e-value and percentage of query coverage (Stobbe et al., 2013). Depending on the nature of the score-data one of the following statistical tests is performed to compare the sets of probe and decoy scores, the parametric student *t*-test (for normally distributed data with equal variance), the Welch's *t*-test (for normally distributed data with unequal variance) or the Wilcoxon Ranksum test (for data which are not normally distributed). Samples with a p-value smaller than or equal to 0.05 are considered to be positive for a specific virus, while samples with a p-value greater than of equal to 0.1 are considered to be negative (Stobbe et al., 2013). Samples rendering a p-value between these two margins are only suspected to be positive and indicated as such.

4.4. Grapevine virus probe design and implementation

Truffle was used to design probes for viruses, which are known to infect grapevine (Table 2). The viruses consist of a list of grapevine-infecting viruses generated by Martelli (2014). Generally the reference genome for a particular virus species available in GenBank was used as target genome while the type member of the genus served as the near-neighbour genome. For *Grapevine leafroll-associated virus 3*, *Grapevine fanleaf virus* and *Grapevine virus E* the full genome sequences of local isolates, available on NCBI, were used as target genomes. In the absence of a full genome the largest available sequence was used. In instances where the target species was the type member another closely related virus was chosen as near-neighbour. The final probes were screened against the raw NGS datasets of the 18 grapevine samples to determine their virus profiles.

4.5. Target genome assessment

Different e-probe sets were designed for divergent GLRaV-3, GVA and GVB variants. The results generated for the distinctive probe-sets for a species were then compared to determine the effect of intra-species genetic variation on the sensitivity of virus detection.

4.6. Read-mapping analysis

Using CLC Genomics Workbench 8, filtered reads were mapped onto all detected viruses (length fraction=0.5; similarity fraction=0.9; Non-specific reads mapped randomly) and the percentage genome coverage determined.

Acknowledgments

The authors would like to acknowledge Bernard Visser for bioinformatics support, as well as the Wine Industry Network for Expertise and Technology (Winetech, Grant GenUS11/3), Citrus Research International (CRI, Grant 1100) and the National Research Foundation (NRF) for their financial assistance towards this research. Opinions expressed and conclusions arrived at, are those of the authors and are not necessarily to be attributed to the NRF.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.virol.2016.05.008.

References

Bi, Y., Tugume, A.K., Valkonen, J.P.T., 2012. Small-RNA deep sequencing reveals *Arctium tomentosum* as a natural host of *Alstroemeria virus X* and a new putative Emaravirus. PLoS One 7, e42758. http://dx.doi.org/10.1371/journal.pone.0042758.

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120. http://dx.doi.org/10.1093/bioinformatics/btu170.

Burger, J.T., Maree, H.J., 2015. Metagenomic next-generation sequencing of viruses infecting grapevines. Methods Mol. Biol. 1302, 315–330. http://dx.doi.org/10.1007/978-1-4939-2620-6_23.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST+: architecture and applications. BMC Bioinform. 10, 421. http://dx.doi.org/10.1186/1471-2105-10-421.

Coetzee, B., Freeborough, M.-J., Maree, H.J., Celton, J.-M., Rees, D.J.G., Burger, J.T., 2010. Deep sequencing analysis of viruses infecting grapevines: virome of a vineyard. Virology 400, 157–163. http://dx.doi.org/10.1016/j.virol.2010.01.023.

Delcher, A.L., Kasif, S., Fleischmann, R.D., Peterson, J., White, O., Salzberg, S.L., 1999. Alignment of whole genomes. Nucleic Acids Res. 27, 2369–2376. http://dx.doi.org/10.1093/nar/27.11.2369.

Delcher, A.L., Phillippy, A., Carlton, J., Salzberg, S.L., 2002. Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res. 30, 2478–2483. http://dx.doi.org/10.1093/nar/30.11.2478.

Gu, Y.-H., Tao, X., Lai, X.-J., Wang, H.-Y., Zhang, Y.-Z., 2014. Exploring the polyadenylated RNA virome of sweet potato through high-throughput sequencing. PLoS One 9, e98884. http://dx.doi.org/10.1371/journal.pone.0098884.

Ho, T., Tzanetakis, I.E., 2014. Development of a virus detection and discovery pipeline using next generation sequencing. Virology 471–473, 54–60. http://dx.doi.org/10.1016/j.virol.2014.09.019.

Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., Salzberg, S.L., 2004. Versatile and open software for comparing large genomes. Genome Biol. 5, R12. http://dx.doi.org/10.1186/gb-2004-5-2-r12.

Martelli, G.P., 2014. Directory of virus and virus-like disease of the grapevine and their agents. J. Plant Pathol. 96, S1–S136. http://dx.doi.org/10.4454/JPP.V96I1SUP.

Melcher, U., Verma, R., Schneider, W.L., 2014. Metagenomic search strategies for interactions among plants and multiple microbes. Front. Plant Sci. 5, 268. http://dx.doi.org/10.3389/fpls.2014.00268.

Roux, S., Tournayre, J., Mahul, A., Debroas, D., Enault, F., 2014. Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. BMC Bioinform. 15, 76. http://dx.doi.org/10.1186/1471-2105-15-76.

Satya, R.V., Zavaljevski, N., Kumar, K., Reifman, J., 2008. A high-throughput pipeline for designing microarray-based pathogen diagnostic assays. BMC Bioinform. 9, 185. http://dx.doi.org/10.1186/1471-2105-9-185.

Stobbe, A.H., Daniels, J., Espindola, A.S., Verma, R., Melcher, U., Ochoa-Corona, F., Garzon, C., Fletcher, J., Schneider, W., 2013. E-probe Diagnostic Nucleic acid Analysis (EDNA): a theoretical approach for handling of next generation sequencing data for diagnostics. J. Microbiol. Methods 94, 356–366. http://dx.doi.org/10.1016/j.mimet.2013.07.002.

Stobbe, A.H., Schneider, W.L., Hoyt, P.R., Melcher, U., 2014. Screening metagenomic data for viruses using the e-probe diagnostic nucleic acid assay. Phytopathology 104, 1125–1129. http://dx.doi.org/10.1094/PHYTO-11-13-0310-R.

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.

Wylie, S.J., Li, H., Saqib, M., Jones, M.G.K., 2014. The global trade in fresh produce and the vagility of plant viruses: a case study in garlic. PLoS One 9, e105044. http://dx.doi.org/10.1371/journal.pone.0105044.

Zhao, G., Krishnamurthy, S., Cai, Z., Popov, V.L., Travassos da Rosa, A.P., Guzman, H., Cao, S., Virgin, H.W., Tesh, R.B., Wang, D., 2013. Identification of novel viruses using VirusHunter - an automated data analysis pipeline. PLoS One 8, e78470. http://dx.doi.org/10.1371/journal.pone.0078470.