Virology 400 (2010) 157–163

Contents lists available at ScienceDirect

Virology

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

Rapid Communication

Deep sequencing analysis of viruses infecting grapevines: Virome of a vineyard

Beatrix Coetzee [a], Michael-John Freeborough [a], Hans J. Maree [a], Jean-Marc Celton [b], D. Jasper G. Rees [b], Johan T. Burger [a,*]

[a] Department of Genetics, Stellenbosch University, Private Bag X1 Matieland, 7602, South Africa
[b] Department of Biotechnology, University of the Western Cape, Private Bag X17, Bellville, 7535, South Africa

ARTICLE INFO

Article history:
Received 11 November 2009
Returned to author for revision
2 December 2009
Accepted 17 January 2010
Available online 20 February 2010

Keywords:
Metagenomic sequencing
Deep sequencing analysis
Grapevine viruses

ABSTRACT

Double stranded RNA, isolated from 44 pooled randomly selected vines from a diseased South African vineyard, has been used in a deep sequencing analysis to build a census of the viral population. The dsRNA was sequenced in an unbiased manner using the sequencing-by-synthesis technology offered by the Illumina Genome Analyzer II and yielded 837 megabases of metagenomic sequence data. Four known viral pathogens were identified. It was found that *Grapevine leafroll-associated virus 3* (GLRaV-3) is the most prevalent species, constituting 59% of the total reads, followed by *Grapevine rupestris stem pitting-associated virus* and *Grapevine virus A*. *Grapevine virus E*, a virus not previously reported in South African vineyards, was identified in the census. Viruses not previously identified in grapevine were also detected. The second most prevalent virus detected was a member of the *Chrysoviridae* family similar to *Penicillium chrysogenum virus*. Sequences aligning to two other mycoviruses were also detected.

Introduction

Grapevine is an important global commodity crop which is widely planted throughout temperate regions. Viruses, however, are a significant factor in reducing the quality and quantity of the yield and are known to reduce the productive life of vineyards (Martelli and Boudon-Padieu, 2006). Grapevine is subject to infection by more than 60 different viruses, the most known for any crop plant (Martelli and Boudon-Padieu, 2006). Most important grapevine virus diseases are caused by complexes of viruses, with up to nine different viruses having been identified in a single vine (Prosser et al., 2007). In South Africa, as in most grape-growing regions of the world, grapevine leafroll is regarded to be the most significant virus disease affecting grapevine, with Shiraz disease and Shiraz decline becoming more prominent as emerging diseases in the industry. Leafroll disease is well characterized and is associated with up to 10 different viruses of the *Ampelovirus* genus, with *Grapevine leafroll-associated virus 3* (GLRaV-3) being the most prominent and widely distributed virus associated with the disease. In South African vineyards it appears that *Grapevine rupestris stem pitting-associated virus* (GRSPaV) is most frequently associated with Shiraz decline (Goszczynski, pers. com.), whilst *Grapevine virus A* (GVA) seems to be most frequently associated with Shiraz disease (Goszczynski, 2007). The etiologies of these and most other grapevine virus diseases are not resolved. While these viruses have certainly been shown to be closely associated with the respective diseases, it is generally accepted that additional viruses are associated with these diseases. Moreover, anecdotal evidence exists for differing etiologies for the same disease, depending on geographical region, the stage of the growing season and the grapevine genotype.

Present disease diagnostics rely on ELISA or RT-PCR and target the viruses that have historically been associated with these diseases. While these tests are highly specific, they may not result in an accurate reflection of the etiological status of the tested plant, or of the particular disease, since none of the current diagnostic techniques address the potential contribution of other known or unknown viruses that may be involved in the etiology of a particular disease. Moreover, the error prone replication of RNA viruses leads to quasispecies, which can further complicate PCR-based detection assays as not all variants of the virus may be detected.

New and powerful technologies which are able to sequence viruses from environmental samples without the need for laborious and costly purification, cloning and screening techniques can result in the generation of sequence information for the complete virome in an unbiased fashion (Al Rwahnih et al., 2009; Cann et al., 2005; Kreuze et al., 2009; Williamson et al., 2008). These data could help to identify new viruses in disease complexes, identify the dominant variants of the viral species and give an indication of the frequency of viruses found in the infected material. Consequently, the data can be used to develop more accurate diagnostic assays. Next generation sequence data has been used successfully to evaluate the viruses present in a single grapevine plant displaying typical Syrah decline symptoms (Al Rwahnih et al., 2009), to identify an unknown virus in *Liatris spicata* (Adams et al., 2009) and to provide deep sequence analysis of virus

* Corresponding author. Fax: +27218085833.
E-mail address: jtb@sun.ac.za (J.T. Burger).

0042-6822/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.virol.2010.01.023

infected sweetpotato plants (Kreuze et al., 2009). In these three studies, potential new viruses were discovered after analysis of the generated sequence data. Al Rwahnih et al. (2009) generated long read sequences (LRSs) of ~200 bp and used BLAST analysis to confirm the expected presence of GRSPaV, to show the presence of rupestris vein-feathering virus and to identify a new marafivirus, grapevine syrah virus-1. Similarly, Adams et al. (2009) determined the sequence of *Pepino mosaic virus* from tomato using LRSs. These researchers also identified a new member of the *Cucumovirus* genus, *Gayfeather mild mottle virus*, from a diseased *Liatris* plant of unknown etiology. In a study by Kreuze et al. (2009) sweetpotato plants were inoculated with two known viruses singly or in combination. Sequence data obtained from short read sequences (SRSs) of ~24 bp yielded the expected full-length sequences of the inoculated viruses (*Sweetpotato feathery mottle virus* and *Sweetpotato chlorotic stunt virus*), as well as that of unknown viruses belonging to the *Badnavirus* and *Mastrevirus* genera. These three studies indicate the usefulness of next generation sequencing technologies and a metagenomics approach to identify known viruses and to discover possible new viruses from diseased plant material.

Adams et al. (2009) and Kreuze et al. (2009) concluded that next generation sequencing technologies can be used as a diagnostic tool to identify a plant virus when no prior knowledge of the virus is available. Therefore, it can be used as an investigative technique to generate sequences of multiple viruses, if present, in an unbiased fashion due to the use of non sequence specific primers. This could lead to the identification of the viruses and a better understanding of the disease especially when its etiology is currently unknown.

In this paper we describe the use of sequencing-by-synthesis technology (http://www.illumina.com/technology/sequencing_technology.ilmn) on the massively parallel Illumina Genome Analyzer II, to sequence an environmental sample composed of 44 randomly selected vines, to determine the viral profile of a severely diseased vineyard.

Results

Sequencing

Double stranded RNA was isolated from lignified cane material pooled from 44 vines that were randomly selected in a *Vitis vinifera* cv. Merlot vineyard. Integrity of the dsRNA was confirmed by diagnostic RT-PCR amplification of GLRaV-3, GRSPaV and GVA (data not shown).

The dsRNA was prepared for cDNA synthesis followed by sequencing. cDNA fragments of ~200 bp were selected for PCR enrichment. Quantitative PCR was used to determine the optimal amplification of the ~200 bp cDNA fragments (Supplementary data 1) and 33 PCR cycles were used for subsequent cDNA amplification. Sequencing-by-synthesis was performed using the Illumina Genome Analyzer II with 51 cycles. Approximately 13 million clusters were obtained on a single sequencing lane of the Illumina flow cell. Approximately 74% of these clusters could be analyzed and yielded quality sequence data. The paired-end sequence data yielded 19,247,026 reads, which translates to more than 837 megabases of sequence data from 1/8th of an Illumina flow cell.

De novo sequence assembly and analysis

Reads were assembled into scaffolds using the Velvet 0.7.31 *de novo* assembly algorithm (Zerbino and Birney, 2008) (http://www.ebi.ac.uk/~zerbino/velvet). A variety of parameters were tested, yielding scaffolds that varied between 100 and 8,624 nt in length. Parameters for optimal assembly were selected based on number and length of the scaffolds obtained. The following parameters were used for all further analyses: hash length of 23, coverage cut-off of 50, expected coverage of 1,000 and a minimum scaffold length of 100. In this assembly, 7,895,103 reads (41%) assembled into 449 scaffolds (Supplementary data 2). Forty-eight of the scaffolds were larger than 1,000 nt and the largest scaffold was 8624 nt in length. These scaffolds were subjected to BLAST (Altschul et al., 1997) searches against the NCBI non-redundant (nr) DNA and protein databases and classified according to the sequences they aligned to with the highest bit score. The scaffolds aligned to GLRaV-3 (124), GRSPaV (1) and GVA (7), which represent 59%, 4% and 1% respectively of the analyzed read data. Grapevine virus E (GVE) was also identified, with 2 scaffolds aligning, accounting for 1% of the read data (Fig. 1). A single scaffold aligned to GVB. However, due to the low homology (47% amino acid identity) and read content of this scaffold, it was omitted from Fig. 1.

Furthermore, BLAST searches against the non-redundant databases detected the presence of three mycoviral families. Twenty-six scaffolds, accounting for 5% of the assembled read data aligned to the fungi-infecting dsRNA viruses from the family *Chrysoviridae*: *Penicillium chrysogenum virus* (PcV), *Helminthosporium victoriae 145S virus* and *Cryphonectria nitschkei chrysovirus 1*. Two scaffolds (1.1% of read data) aligned to *Aspergillus mycovirus 1816*, a member of the family *Totiviridae* and three to the unassigned *Fusarium graminearum dsRNA mycovirus 4* (0.2% of read data). Twenty-two of the 31 scaffolds aligning to these viruses aligned at a low homology and were only identified when searching against the NCBI (nr) protein database. Bacterial, fungal and host-derived scaffolds were also identified but will not be discussed in further detail. A further 188 scaffolds representing 13% of analyzed read data did not align to any known sequences in the NCBI databases. Sequence classifications, the number of scaffolds and read counts incorporated in scaffolds of the dominant viruses, are listed in Table 1 (Velvet *de novo* assembly). To give an indication of relative abundance, the percentage read counts in each sequence classification is given in Fig. 1.

Re-assembly against reference sequences

Twenty three representative variants for the five dominant viral species identified in BLAST searches, were extracted from the NCBI database, and used as reference sequences in further re-assembly studies using the Mapping and Assembly with Quality (MAQ) assembler (Li et al., 2008) (http://MAQ.sourceforge.net/). Complete genome sequences of GLRaV-3, GRSPaV, GVA and PcV were used, and a partial sequence of GVE, since the genome for the specific strain with the highest sequence identity has not been fully sequenced yet. The extracted sequences for the 23 representative variants were used simultaneously as reference sequences in a MAQ re-assembly analysis to determine the dominant variant for each virus species based on the read count. A total of 3,860,942 reads accounting for approximately 20% of the total read data aligned to the reference sequences (Table 1). To verify the presence of these variants, reference sequences were used individually for re-assembly analysis to calculate the read count, average depth of sequence and the average coverage of the respective genomes (Table 1). The most prominent variants for each virus species could be selected based on these criteria (Fig. 2). However, the reference strains used in the assemblies are only representative of the viral groups and not identical to the variants detected in this study, therefore the random distribution of reads on the genomes with higher read counts are observed in areas with high sequence homology between the different viral variants. Data from individual re-assemblies were used for further analyses.

The GLRaV-3 variant GP18 had the highest average depth of sequence (10,009 nt coverage) followed by GRSPaV SG1 (653 nt), GVE TVP15 (178 nt) and GVA variants GTR1-1 (165 nt) and P163-1 (149 nt) (Table1, Fig. 2). The genome coverage for the four full-length genomes of the dominant variants were 100% for GLRaV-3 (GP18), 90% for GRSPaV (SG1), 93% for GVA variant GTR1-1 and 94% for P163-1 (Table 1, Fig. 2). GVE was excluded due to the lack of a full length reference genome. The average depth of sequence and genome

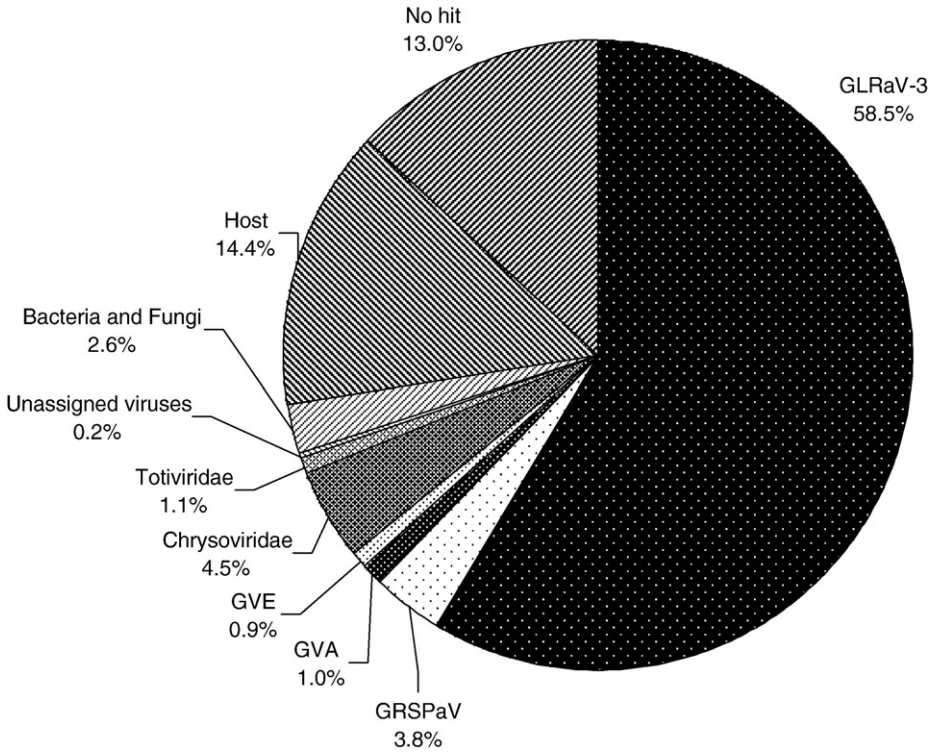

Fig. 1. Comparative percentages for read counts utilized in scaffolds for each sequence classification according to best hit with BLASTn or BLASTx searches. **GLRaV-3** *Grapevine leafroll-associated virus 3*, **GRSPaV** *Grapevine rupestris stem pitting-associated virus*, **GVA** *Grapevine virus A* and **GVE** *Grapevine virus E*.

coverage of PcV was low due to the poor sequence similarity to the reference sequence.

Despite the use of a dsRNA-specific extraction procedure, host nucleic acids were also sequenced. Ten percent of the total reads aligned to the 19 *Vitis vinifera* chromosomes (GenBank accession no: NC_012007–NC_012025), chloroplast (GenBank accession no: NC_007957) and mitochondrial sequences (GenBank accession no: NC_012119) with an average depth of sequence of 0.4 nt and genome coverage of 0.4%.

Discussion

Both *de novo* and re-assembly analyses showed the GP18 variant of GLRaV-3 (Maree et al., 2008) to be the most abundant isolate of all the viruses identified in this sample. Sixty-six of the 124 *de novo* assembled scaffolds (56% of the GLRaV-3 read data) aligned preferentially to the GP18 variant. The 58 remaining scaffolds aligned to 17 different GLRaV-3 sequences available in the NCBI database (Table 1). The lack of a full genome length scaffold for GLRaV-3 could be due to the diversity observed within the scaffolds and the possibility of two or more variants of GLRaV-3 being present in the sample. Re-assembly analysis confirmed the dominance of a GP18-like variant, as the highest number of reads could be re-assembled against this reference sequence, with 100% genome coverage and an average depth of sequence of 10,009 nt (Table 1, Fig. 2A). After re-assembly analysis, data suggested that two additional GLRaV-3 variants (Cl-766 and NY1) were present based on the read count, average depth of sequence and genome coverage (Table 1). These two variants share 97.6% identity, which suggested that one GLRaV-3 variant similar to the NY1/Cl-766 variants might be present in the vineyard. Moreover, the high variability noted in BLASTn pairwise analysis suggests that the identification of new grapevine leafroll-associated ampeloviruses from this sequence data cannot be excluded (Supplementary data 3).

A near full-length GRSPaV scaffold (8,624 nt in length, accounting for 296,352 reads) was generated in the *de novo* assembly, BLASTn alignments indicated that this viral variant shares 92.3% sequence identity with the SG1 variant (Fig. 3, Supplementary data 3) (Meng et al., 2005). Re-assembly of the reads against GRSPaV full length genomes confirmed that the GRSPaV is similar to the SG1 variant due to the read count (135,421 reads), average depth of sequence (653 nt) and 90% genome coverage (Table 1, Fig. 2B). The GRSPaV scaffold, Node 192, had complete coverage, showing that a GRSPaV variant similar but not identical to the SG1 strain was present.

Analysis of the GVA scaffolds indicated that a variant in the molecular group III (Goszczynski et al., 2008) is present in this vineyard. Seven *de novo* assembled scaffolds, accounting for 79,116 reads, aligned to GVA variants. Three scaffolds, including a scaffold of 5826 nt, aligned to the P163-1 variant in molecular group III, and four scaffolds aligned to the Is151 strain in molecular group I. Re-assemblies against the molecular group III variants, P163-1 and GTR1-1 (Goszczynski et al., 2008) were comparable with respect to read count, average depth of sequence, genome coverage (Table 1, Figs. 2C and D) and BLASTn pairwise alignment of scaffolds (Supplementary data 3). This suggests that a molecular group III variant of GVA is present. The four *de novo* assembled scaffolds which aligned preferentially to a GVA variant in molecular group I were not confirmed by re-assembly analysis as the genome coverage, depth of sequence and number of reads aligning to variants from this molecular group were poor (Table 1). These results lead us to propose that only a GVA variant of molecular group III is present in the diseased vineyard.

The sequence data identified a fourth virus, GVE, which has not previously been detected in South Africa. Subsequent PCR amplification with diagnostic primers, confirmed the presence of GVE in the samples from this vineyard. Two scaffolds were obtained for GVE, the largest being 5,172 nt in length, which aligned to and extends the partial sequence of the TvP15 strain available in the NCBI database (Nakaune et al., 2008). Although one of the scaffolds aligned with a higher bit score to the TvAQ7 strain, probably because the TvAQ7 sequence is essentially complete and allows for a larger part of the scaffold to align, the scaffold had higher homology to the TvP15 partial sequence (Supplementary data 3). This suggests that read data can be

Table 1
Comparison of *de novo* and re-assembly data for the five dominant virus species identified in this study. *De novo* assembled scaffolds are classified according to best alignment (highest bit score) in the NCBI database found with BLASTn and BLASTx searches. MAQ re-assembly data are shown for the 23 representative variants identified after *de novo* assembly analysis.

Virus	GenBank accession number	Strain[a]	Velvet *de novo* assembly			MAQ re-assembly					
			No of scaffolds	Read count[b]	Percentage read count	Simultaneous re-assembly		Individual re-assemblies			
						Read count	Percentage read count	Read count	Percentage read count	Average depth of sequence	Genome coverage
GLRaV-3	EU259806[c]	GP18	66	2,608,947	56%	3,616,585	98%	4,242,321	54%	10,009	100%
	EU344893[c]	Cl-766	25	1,048,903	23%	4,673	0%	1,678,187	21%	3,951	90%
	EF508151[d]		11	487,297	11%	–	–	–	–	–	–
	AF037268[c]	NY1	6	120,333	3%	66,943	2%	1,922,950	25%	4,558	92%
	AJ606358[d]		2	132,454	3%	–	–	–	–	–	–
	DQ314610[d]		2	32,232	1%	–	–	–	–	–	–
	AJ748516[d]		1	47,944	1%	–	–	–	–	–	–
	EF103903[d]		1	26,255	1%	–	–	–	–	–	–
	AY704412[d]		1	25,796	1%	–	–	–	–	–	–
	AY495340[d]		1	20,226	0%	–	–	–	–	–	–
	AJ748536[d]		1	17,824	0%	–	–	–	–	–	–
	EU344896[d]		1	17,542	0%	–	–	–	–	–	–
	DQ911148[d]		1	11,343	0%	–	–	–	–	–	–
	EF445656[d]		1	1,728	0%	–	–	–	–	–	–
	AJ748520[d]		1	1,189	0%	–	–	–	–	–	–
	FJ786016[d]		1	801	0%	–	–	–	–	–	–
	DQ780887[d]		1	740	0%	–	–	–	–	–	–
	ABY87025[e]		1	20,397	0%	–	–	–	–	–	–
			124	**4,621,195**	**100%**	**3,688,201**		**7,843,458**	**100%**		
GRSPaV	AY881626[c]	SG1	1	296,352	100%	120,654	94%	135,421	61%	653	90%
	AY881627[c]	BS	ND	ND	ND	4,251	3%	16,696	8%	76	35%
	AF057136[c]	GRSPaV-1	ND	ND	ND	1,605	1%	32,075	14%	152	56%
	AF026278[c]	GRSPaV-1	ND	ND	ND	1,462	1%	31,929	14%	150	55%
	AY368590[c]	SY	ND	ND	ND	163	0%	3,071	1%	14	10%
	AY368172[c]	PN	ND	ND	ND	151	0%	2,582	1%	12	10%
			1	**296,352**	**100%**	**128,286**		**221,774**			
GVA	X75433[c]	Is151	4	33,672	43%	1,230	4%	2,533	4%	14	22%
	DQ855088[c]	P163-1	3	45,444	57%	12,773	41%	26,190	38%	149	94%
	DQ787959[c]	GTR1-1	ND	ND	ND	16,217	52%	28,892	42%	166	93%
	DQ855084[c]	GTG11-1	ND	ND	ND	685	2%	2,606	4%	14	24%
	DQ855086[c]	GTR1-2	ND	ND	ND	196	1%	1,169	2%	6	10%
	DQ855082[c]	P163-M5	ND	ND	ND	83	0%	1,122	2%	6	10%
	DQ855081[c]	GTR1 SD-1	ND	ND	ND	12	0%	1,007	1%	5	12%
	DQ855083[c]	KWVMo4-1	ND	ND	ND	12	0%	1,152	2%	6	9%
	DQ855087[c]	BMO32-1	ND	ND	ND	7	0%	1,313	2%	7	11%
	AF007415[c]	PA3	ND	ND	ND	0	0%	1,084	2%	6	10%
	AY244516[c]		ND	ND	ND	0	0%	1,084	2%	6	10%
			7	**79,116**	**100%**	**31,215**		**68,152**			
GVE	AB432910[c]	TvAQ7	1	69,780	97%	18	0%	160	1%	1	3%
	AB432911[d]	TvP15	1	1,898	3%	13,204	100%	13,207	99%	178	99%
			2	**71,678**	**100%**	**13,222**		**13,367**			
PcV	AF296439[d]		6	84,403	23%	–	–	–	–	–	–
	AF296442[d]		1	751	0%	–	–	–	–	–	–
	AAM95601[e]		6	9,025	3%	–	–	–	–	–	–
	AAM95604[e]		5	12,2873	34%	–	–	–	–	–	–
	AAM95602[e]		3	70,909	20%	–	–	–	–	–	–
	AAM95603[e]		2	54,690	15%	–	–	–	–	–	–
	AAM68955[e]		1	14,017	4%	–	–	–	–	–	–
	AAM68953[e]		1	1,973	1%	–	–	–	–	–	–
	ACT79258[e]		1	537	0%	–	–	–	–	–	–
	NC_007539[c]	Seg 1 ATCC 9480	ND	ND	ND	0	0%	0	0%	0	0%
	NC_007540[c]	Seg 2 ATCC 9480	ND	ND	ND	0	0%	0	0%	0	0%
	NC_007541[c]	Seg 3 ATCC 9480	ND	ND	ND	0	0%	0	0%	0	0%
	NC_007542[c]	Seg 4 ATCC 9480	ND	ND	ND	0	0%	0	0%	0	0%
			26	**359,178**	**100%**	**0**		**0**			
Total read count				**3,860,924**		**8,146,751**					

GLRaV-3 *Grapevine leafroll-associated virus 3*, GRSPaV *Grapevine rupestris stem pitting-associated virus*, GVA *Grapevine virus A*, GVE *Grapevine virus E* and PcV *Penicillium chrysogenum virus*. ND—not detected.

[a] Strain names only shown for sequences used in MAQ re-assembly.
[b] Number of reads incorporated in scaffolds.
[c] Complete genome sequence.
[d] Partial genome sequence.
[e] Protein sequence.

assembled into scaffolds that can align to known virus sequences but can also extend the sequences of partially sequenced viral genomes, expanding existing sequence data. The presence of the TvP15 isolate was confirmed with re-assembly analysis as this GVE isolate had significantly higher read count, average depth of sequence and genome coverage than the GVE TvAQ7 strain (Table 1). This also

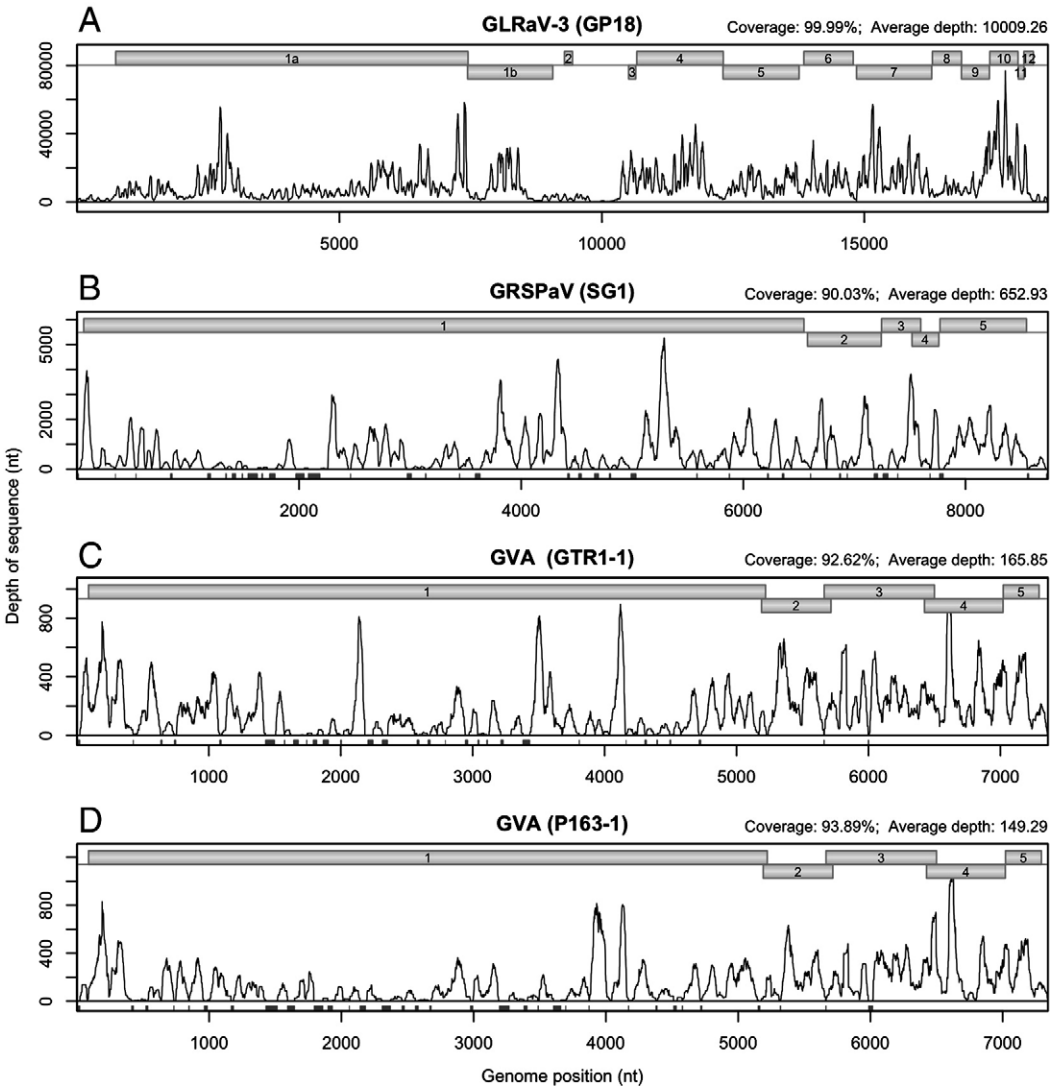

Fig. 2. MAQ-reassembly of reads on four full-length genomes representing the dominant variants for (A) GLRaV-3 (GP18), (B) GRSPaV (SG1), (C) GVA (GTR1-1) and (D) GVA (P163-1). GVE was excluded due to the lack of a full-length genome. Schematic representation of virus genomes with numbered open reading frames are shown above graph. Grey bars below graph highlight areas with no coverage. **GLRaV-3** *Grapevine leafroll-associated virus 3*, **GRSPaV** *Grapevine rupestris stem pitting-associated virus*, **GVA** *Grapevine virus A.*

shows that care must be taken when assigning the data to specific strains depending on already sequenced complete genomes.

A total of 5% of the analyzed read data aligned to members of the *Chrysoviridae* family, making it the second most abundant virus group isolated from this vineyard after GLRaV-3. Homology to known members of the *Chrysoviridae* was too low to align at the nucleotide level apart from seven scaffolds, 19 scaffolds were only revealed when searching the NCBI (nr) protein database. The yet unidentified chrysoviruses that we detected could not be classified beyond family level with either *de novo* assembly or reassembly analysis. These viruses are divergent from all other members of the family that have been sequenced to date. The PcV reference sequences used in the analysis did not correspond at the nucleotide level to the member(s) of the *Chrysoviridae* detected in this sample (Supplementary data 3)

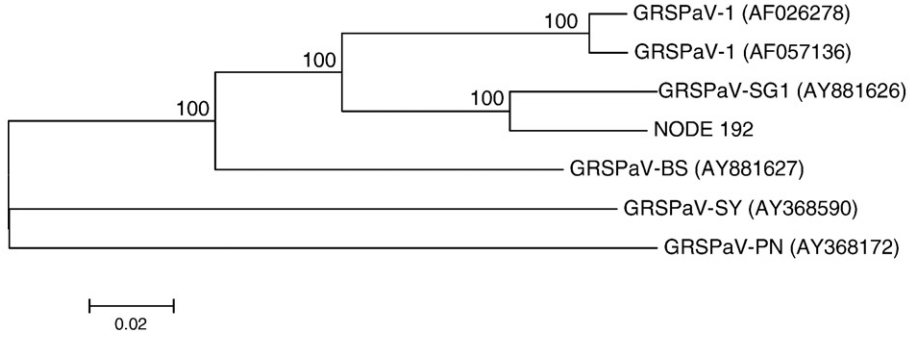

Fig. 3. Phylogenetic tree (bootstrap consensus tree) showing the relationship between the six complete genome sequences and the *de novo* generated scaffold (Node 192) for *Grapevine rupestris stem pitting-associated virus* (GRSPaV). Node 192 group with the SG1 (AY881626) strain. GenBank accession numbers are indicated in brackets. Bootstrap values (500 replicates) are indicated above the branches. The scale indicates number of substitutions per base position.

and thus no reads aligned to PcV in MAQ re-assembly analysis (Table 1). The role of these viruses in plant disease is unknown at present. It is also not clear whether members of this family use plants or fungi as the primary host. We propose that these viruses can be isolated from grapevine phloem tissue, implying that plants may act as host to the isolated virus. However, the possibility that an as yet unidentified fungal host may exist within the grapevine cannot be excluded, we therefore propose that the possible role that these dsRNA viruses play in plant diseases be investigated further. Similar ambiguous results were obtained with Amasya cherry disease (Covelli et al., 2004).

To estimate the relative abundance of the four dominant virus species in the diseased vineyard, the average depth of sequence and the number of re-assembled reads were normalized against the genome length of the respective reference sequences. When considering only these four virus species, results indicated that GLRaV-3 accounts for 90%, GRSPaV for 7% and GVA and GVE for 1.5% each of the known virus infection of this severely diseased vineyard. These percentages are similar to the relative abundance of the seven viruses when using the read counts in the *de novo* assemblies for the virus-associated scaffolds only. From this analysis GLRaV-3 accounts for 84%, unidentified *Chrysoviridae* for 7%, GRSPaV for 5%, GVA for 1%, GVE for 1%, *Totiviridae* for 2% and *Fusarium graminearum dsRNA mycovirus 4* for less than 1% of the total virus population within the vineyard sample.

In total 188 scaffolds did not align to any known GenBank sequences in either of the non-redundant DNA or protein databases. Fifty-three of these scaffolds have more than 10% unassigned nucleotides (N's) in the sequence, which may contribute to their unknown status. The remaining 135 scaffolds could be of viral, bacterial or fungal origin and need to be investigated further. In total, 41% of read data could be analyzed. This is comparable to the 50% of data with no known homology to other submitted sequences found by another metagenomic study (Kerkhof and Goodman, 2009).

Conclusion

We were able to show that sequencing-by-synthesis, using a massively parallel sequencing platform, generated read data of high quantity and quality from low amounts of viral dsRNA, extracted from infected grapevine material, to provide the first virome of a diseased vineyard. Analysis of read data identified four different viruses that infected these grapevines, as well as their most prominent variants. The data further showed the presence of at least three unknown viruses that could be assigned to different virus families.

Both *de novo* assembly of reads and re-assembly on reference sequences indicated that GLRaV-3 was the dominant virus present in the sample, accounting for greater than 84% of the viral population of the diseased vineyard based on read count in *de novo* assemblies. Assembly analysis was able to identify the dominant variant of GLRaV-3 and the presence of at least one more GLRaV-3 minor variant in this vineyard. Similarly, a dominant GVA variant within molecular group III was identified. With *de novo* assembly a single scaffold spanning 99% of the GRSPaV genome was generated with high sequence identity to the SG1 strain. We also report the presence of GVE in the vineyard and have extended the incomplete sequence of the TvP15 strain of this virus (Node 3404, Supplementary data 2).

This is the first report of sequence data for members of the dsRNA *Chrysoviridae* family isolated from grapevine. The high read counts that aligned with the *Chrysoviridae* isolate(s) suggest that these viruses were isolated directly from the phloem material of grapevine. The current analysis of sequence data also detected the presence of viruses similar to a member of the *Totiviridae* and the unclassified dsRNA virus, *Fusarium graminearum dsRNA mycovirus 4*. However, classifications and identification of these mycoviruses might become more specific as new sequence data became available on the NCBI database.

The use of LRSs generated by 454 technology has been the *de facto* choice for metagenomics and *de novo* sequencing of viral genomes. This is the first report that indicates that *de novo* and re-assembly analysis of short sequence reads generated from dsRNA can successfully distinguish between four different known viral species and gives an indication of the dominant variants of these species. Furthermore, read data analysis helped to detect at least three unknown viruses from viral families that have not previously been reported to infect grapevine, resulting in the first virome of a severely diseased vineyard being sequenced. Our results prove the feasibility of next generation high-throughput sequencing technology using Illumina technology and dsRNA as starting material in a metagenomics approach to determine the virome of a severely diseased vineyard and suggest that this approach can be used to elucidate the etiologies of the world's notorious grapevine virus diseases.

Materials and methods

Plant material

Plant material was sourced from a severely diseased vineyard (cv. Merlot) in the Stellenbosch region of South Africa. Plants were randomly selected during winter dormancy when no apparent symptoms were visible. Phloem scrapings were collected from 44 vines during August 2008. During the following growth season, symptoms were observed of these selected vines (Supplementary data 4). Vines in this severely diseased vineyard displayed typical and atypical leafroll and Shiraz disease symptoms. These symptoms include reddening of leaves while veins remain green, downwards rolling of the leaves, canes with a lack of lignification, swelling at the graft union and reduced vigor. A number of asymptomatic vines were also observed. Double stranded RNA was isolated from the samples using a cellulose extraction protocol adapted from Valverde et al. (1990). Reverse transcriptase PCR screening for the most prevalent viruses (GLRaV-3, GRSPaV and GVA) was performed to confirm the quality of the dsRNA.

Sequencing

Sequencing was performed on the Illumina Genome Analyzer II. Fifty nanograms of dsRNA was subjected to a heat denaturation step at 95 °C for 10 min and flash cooled in ice water. Fragmentation, conversion to cDNA and preparation for sequencing was performed using the Illumina mRNA Sequencing v2 kit. The 15 PCR cycles recommended by Illumina were insufficient for cDNA amplification. Quantitative-PCR was used to determine an optimal number of amplification cycles. After comparison to positive control samples, 33 cycles of PCR amplification was found to be the midpoint of the log phase of amplification for the cDNA (Supplementary data 1). PCR enrichment of the adapter-ligated cDNA fragments was therefore conducted with 33 cycles of amplification for further sequence determination. The enriched DNA template was quantified, diluted to 2.5 pM and hybridized to 1/8th of a flow cell. Clusters were generated on the flow cell using the Illumina Paired-end Cluster Generation v2 kit. The flow cell with the DNA clusters was subjected to sequencing on the Illumina Genome Analyzer II using the SBS Sequencing v3 kit.

Sequence analysis

Paired-end sequence data was assembled using the short read assembler Velvet 0.7.31 (Zerbino and Birney, 2008). Assembled scaffolds were used to search the NCBI database for sequence similarities using BLASTn and BLASTx searches (default parameters and expect value of 10^{-5} were used) (Altschul et al., 1997). Scaffolds were classified according to the sequence with the highest bit score

found with BLAST. Reference sequences were identified based on the BLAST results and used in subsequent analyses. The number of reads aligning to the reference sequences and scaffolds, coverage and average depth were determined with MAQ 0.7.1 (Li et al., 2008) using the easyrun command. Default parameters were used and only filtered data (reads aligning to the references with a high confidence) is reported on.

Acknowledgments

The financial assistance of Winetech and the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the authors and are not necessarily to be attributed to the NRF. The authors also thank Paul Krige (Kanonkop Wine Estate, Stellenbosch, South Africa) for sample material. Anthony la Grange's assistance with graphical presentation of data is appreciated.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.virol.2010.01.023.

References

Adams, I., Glover, R., Monger, W., Mumford, R., Jackeviciene, E., Navalinskiene, M., et al., 2009. Next-generation sequencing and metagenomic analysis: a universal diagnostic tool in plant virology. Mol. Plant Pathol. 10, 537–545.

Al Rwahnih, M., Daubert, S., Golino, D., Rowhani, A., 2009. Deep sequencing analysis of RNAs from a grapevine showing Syrah decline symptoms reveals a multiple virus infection that includes a novel virus. Virology 387, 395–401.

Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., et al., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389.

Cann, J., Fandrich, E., Heaphy, S., 2005. Analysis of the virus population present in equine faeces indicates the presence of hundreds of uncharacterized virus genomes. Virus Genes 30, 151–156.

Covelli, L., Coutts, R., Serio, F., Citir, A., Acikgoz, S., Hernandez, C., et al., 2004. Cherry chlorotic rusty spot and Amasya cherry diseases are associated with a complex pattern of mycoviral-like double-stranded RNAs. I. Characterization of a new species in the genus *Chrysovirus*. J. Gen. Virol. 85, 3389–3397.

Goszczynski, D., 2007. Single-strand conformation polymorphism (SSCP), cloning and sequencing reveal a close association between related molecular variants of *Grapevine virus A* (GVA) and Shiraz disease in South Africa. Plant Pathol. 56, 755.

Goszczynski, D., du Preez, J., Burger, J., 2008. Molecular divergence of *Grapevine virus A* (GVA) variants associated with Shiraz disease in South Africa. Virus Res. 138, 105–110.

Kerkhof, L., Goodman, R., 2009. Ocean microbial metagenomics. Deep-Sea Res. II (56), 1824–1829.

Kreuze, J., Perez, A., Untiveros, M., Quispe, D., Fuentes, S., Barker, I., et al., 2009. Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: a generic method for diagnosis, discovery and sequencing of viruses. Virology 388, 1–7.

Li, H., Ruan, J., Durbin, R., 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 18, 1851.

Maree, H., Freeborough, M., Burger, J., 2008. Complete nucleotide sequence of a South African isolate of *Grapevine leafroll-associated virus 3* reveals a 5′ UTR of 737 nucleotides. Arch. Virol. 153, 755–757.

Martelli, G., Boudon-Padieu, E., 2006. Directory of infectious diseases of grapevines. International centre for advanced mediterranean agronomic studies. Options Méditerr., Ser. B, Stud. Res. 55, 59–75.

Meng, B., Li, C., Wang, W., Goszczynski, D., Gonsalves, D., 2005. Complete genome sequences of two new variants of *Grapevine rupestris stem pitting-associated virus* and comparative analyses. J. Gen. Virol. 86, 1555–1560.

Nakaune, R., Toda, S., Mochizuki, M., Nakano, M., 2008. Identification and characterization of a new vitivirus from grapevine. Arch. Virol. 153, 1827–1832.

Prosser, S., Goszczynski, D., Meng, B., 2007. Molecular analysis of double-stranded RNAs reveals complex infection of grapevines with multiple viruses. Virus Res. 124, 151–159.

Valverde, R., Nameth, S., Jordan, R., 1990. Analysis of double-stranded RNA for plant virus diagnosis. Plant Dis. 74, 255–258.

Williamson, S., Rusch, D., Yooseph, S., Halpern, A., Heidelberg, K., Glass, J., et al., 2008. The sorcerer II global ocean sampling expedition: metagenomic characterization of viruses within aquatic microbial samples. PLos one 3.

Zerbino, D., Birney, E., 2008. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Res. 18, 821.

Author recommended Internet resources

http://www.illumina.com/technology/sequencing_technology.ilmn.
http://MAQ.sourceforge.net/.
http://www.ebi.ac.uk/~zerbino/velvet.