# Comparison of multiple viral population characterization methods on a candidate cross-protection *Citrus tristeza virus* (CTV) source

Jackie Kleynhans[1,2,*] and Gerhard Pietersen[1,2,3]

[(1)] Department of Microbiology and Plant Pathology, University of Pretoria, Private Bag X20, Hatfield 0028, South Africa

[(2)] Forestry and Agricultural Biotechnology Institute (FABI) University of Pretoria, Pretoria 0002, South Africa

[(3)] Agricultural Research Council (ARC), Plant Protection Research Institute, Private Bag X134, Queenswood 0121, South Africa

*Corresponding author: Jackie Kleynhans, e-mail: jacolenelubbe@gmail.com

## Highlights

- The genotype composition of a CTV source was determined.
- Four techniques used and results compared.
- RT-PCR and Illumina sequencing provided relatively comparable results.
- The presence of RB, VT and B165 could be confirmed within the source.
- Challenges encountered discussed and solutions proposed.

## Abstract

*Citrus tristeza virus* (CTV) is the most economically important virus found on citrus and influences production worldwide. The 3' half of the RNA genome is generally conserved amongst sources, whereas the 5' portion is more divergent, allowing for the classification of the virus into a number of genotypes based on sequence diversity. The acknowledged genotypes of CTV are continually being expanded, and thus far include T36, T30, T3, VT, B165, HA16-5, T68 and RB. The genotype composition of the CTV populations of a potential cross protection source in Mexican lime was studied whilst comparing different techniques of viral population characterization. Cloning and sequencing of an ORF 1a fragment, genotype specific RT-PCRs and Illumina sequencing of the p33 gene as well as RNA template enrichment through immuno-capture was done. Primers used in the cloning and sequencing proved to be biased towards detection of the VT genotype. RT-PCR and Illumina sequencing using the two different templates provided relatively comparable results, even though the immuno-captured enriched template provided less than expected CTV specific data, while the RT-PCRs and p33 sequencing cannot be used to make inferences about the rest of the genome; which may vary due to recombination. The source was found to contain multiple genotypes, including RB and VT. When choosing a characterization method, the features of the virus under study should be considered. It was found that Illumina sequencing offers an opportunity to gain a large amount of information regarding the entire viral genome, but challenges encountered are discussed.

**Keywords:** *Citrus tristeza virus*; mild strain cross-protection; cloning; Sanger sequencing; Next generation sequencing; genotype specific RT-PCR

## 1. Introduction

*Citrus tristeza virus* (CTV) is the most economically important virus found on citrus and influences production worldwide (Kong et al., 2000). The 3’ half of the 19.3kb RNA genome (Albiach-Marti et al., 2010; Bar-Joseph, Marcus, and Lee, 1989) is generally conserved amongst sources, whereas the 5’ half is much more divergent (Mawassi et al., 1996). The virus is classified into a number of genotypes based on sequence diversity. The acknowledged genotypes of CTV are continually being expanded, and thus far include T36, T30, T3, VT, B165, HA16-5, T68 and RB (Albiach-Marti et al., 2000; Harper, Dawson, and Pearson, 2010; Harper, 2013; Karasev et al., 1995; Mawassi et al., 1996; Melzer et al., 2010; Roy and Brlansky, 2010). The CTV population of a single host may contain many different genotypes as the longevity of citrus plants allows for repeated infections of various CTV sources through aphid transmission (Rubio et al., 2001). CTV can be controlled through mild strain cross-protection. Cross-protection appears to function only amongst variants of the same genotype and not between variants of different genotypes (Folimonova et al., 2010). Since more than one genotype of CTV can occur within a plant (Albiach-Marti et al., 2000) and the severity of symptoms as well as selection of genotypes within the host is influenced by the host species and environmental conditions (Albiach-Marti et al., 1996; Fulton, 1986; Karasev et al., 1998), the success of cross-protection can be variable. Ideally the assessment of the ability of a source to serve as a cross protecting strain should be conducted on characterised, genotypically homogeneous sources on various hosts under different environmental conditions.

There are many different templates which can be used for the characterization of viral populations, including single gene amplicons, total RNA extractions, and virus particle enrichment by immuno-capture prior to RNA extraction. There are also many different techniques to characterize a viral population, including single gene amplification and sequencing, using genotype specific primers in a RT-PCR as well as doing next generation sequencing, either on the entire genome or just certain genes, where different templates can be used, for example a total RNA extract, single stranded RNA, double stranded RNA or small RNA. Virus particle enrichment through immuno-capture can also be done to enrich for encapsidated RNA. When using gene specific primers to amplify specific regions of the genome, care in the design of the primers needs to be taken to avoid the possibility of introducing amplification bias (Cook et al., 2015). This approach also only provides information on the specific region amplified. Since recombination is common within the CTV genome (Roy and Brlansky, 2010; Rubio et al., 2001; Vives et al., 2005), targeting only a specific region may provide distorted identification of the genotype as recombination events that occurred elsewhere in the genome would not be detected. This approach is however warranted when targeting a gene important for a specific biological function (Cook et al., 2015) for example the p33 gene which has been shown to be involved in the genotype specificity of the super-infection exclusion mechanism of CTV (Folimonova, 2012). While useful for the specific goal of identifying the genotypic variation within that specific gene, conclusions can only be inferred regarding that gene and not the actual genotype variation, which would require the sequence determination of the entire genome. Sequencing the entire genome has become plausible with next generation sequencing on various platforms including Illumina sequencing (Zablocki and Pietersen, 2014). One of the methods of template preparation prior to Illumina sequencing that has been used is immuno-capture of virus particles (Zablocki and Pietersen, 2014). While immuno-capture did not produce the expected results in that study (Zablocki and Pietersen, 2014), the undoubted potential of the technique suggests that it probably required some optimization (Zablocki, 2013). Immuno-capture would allow the enrichment of CTV particles with virus specific antibodies, eliminating most of the non-CTV RNA before RT-PCR.

The aim of this study was to determine the genotype composition of the CTV populations in the New Venture 41/2 source in Mexican lime, whilst comparing different techniques of viral population characterization.

## 2. Materials and methods

### 2.1 Source of CTV

The New Venture 41/2 source of CTV has been identified as a potentially useful cross-protection source by Citrus Research International, South Africa (CRI). The source was isolated when budwood was collected from 108 field grown grapefruit trees infected with CTV, but lacking any symptoms. The sources were established in a glasshouse and inoculated onto virus free Mexican lime, to assess the severity of the sources. After biological indexing in glasshouse trials the most promising sources were subjected to field trials. New Venture 41/2 was amongst the most promising of these sources (Breytenbach, Cook, and Van Vuuren, 2014a) and is currently being assessed in field trials on Marsh and Star Ruby grapefruit trees (Breytenbach, Cook, and Van Vuuren, 2014b).

Bark scrapings, leaf petioles and midribs were collected from one of the Mexican Lime trees inoculated with the New Venture 41/2 source by CRI. The plant material was macerated in liquid nitrogen with a mortar and pestle and used for RNA extraction.

### 2.2 Cloning and sequencing of the ORF 1a fragment

The ORF1a fragment of CTV, which lies within the 3' half of the CTV genome which is conserved among isolates (Albiach-Marti et al., 2010), was amplified. The Rubio *et al.* (2001) primers was used which was designed based on the T30 genotype, but has been shown to also amplify strains from the VT, T3, B165, HA16-5 and RB genotypes (Read and Pietersen, 2015). Amplicons were purified using the Promega Wizard® SV Gel and PCR Cleanup System (Promega, Madison, WI) according to the manufacturer's instructions and ligated into the p-GEM®-T Easy vector. Competent *Escherichia coli* JM109 cells (Promega, Madison, WI, USA) were transformed. For sequencing reactions, vector specific T7 (5' TAATACGACTCACTATAGGG 3') and SP6 (5' ATTTAGGTGACACTATAG 3') primers (Promega, pGEM®-T and pGEM®-T Easy Vector Systems, technical manual) were used to amplify the products followed by purification using Exonuclease and FastAP® (Fermentas, Maryland, USA). Sequencing was done using the T7 primer on the ABI Prism® 3130XL Genetic Analyser (Applied Biosystems, Foster City, California, USA). A total of 102 clones were sequenced, and together with 45 reference sequences from GenBank (Benson et al., 2005), were aligned using MAFFT online (Katoh et al., 2002) and analysed to create a phylogenetic dendrogram (Figure 1). Since many clones had identical sequences, only representatives of these were used when constructing the Neighbour Joining dendrogram using the Jukes-Cantor model with a 1000 bootstrap value in MEGA6 (Tamura et al., 2007). Sequences of clones used in the dendrogram were submitted to GenBank with accession numbers JX484742 to JX484757.

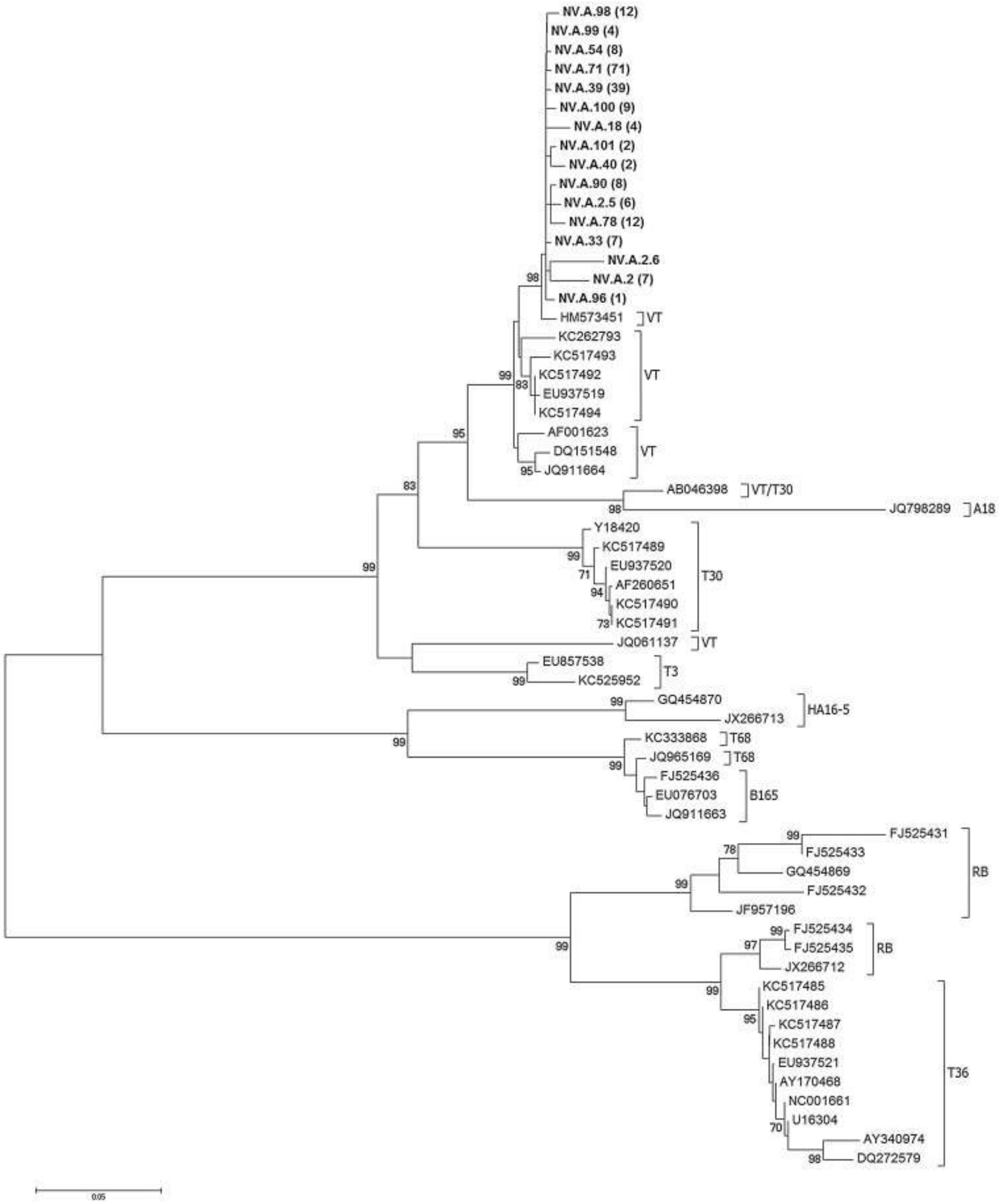

**Fig. 1** The phylogenetic relationship of cloned sequences of the New Venture 41/2 (NV.A, in bold) and 45 representative genomes ORF1a fragment using the Neighbour Joining method based on the Jukes-Cantor model with a 1000 bootstrap. Only branch support values higher than 70 are indicated on nodes. Only representative clones are shown, value following clone indicate how many clones are represented.

### 2.3 Genotype detection RT-PCR in New Venture 41/2

Specific detection of CTV genotypes was based on genotype specific primers used in a RT-PCR (Cook et al., 2015; Roy et al., 2010). The reaction used for VT, RB, B165 and T36 detection is as follows: Total RNA extract (2µl) obtained according to manufacturer's specifications with the GeneJET Plant RNA Purification Mini Kit (Thermo Scientific, Waltham, Massachusetts) was combined with 1µl of 10µM reverse primer and 7µl of nuclease free water and incubated for 3 min at 95°C and then 1 min on ice. After primer annealing, the RNA was added to 4µl RT buffer, 0.21 U AMV reverse transcriptase (Roche Diagnostics, Germany), 0.5 U RibolockRNAse Inhibitor (Thermo Scientific, Waltham, Massachusetts), 1mM of each dNTP, 10mM DTT (Thermo Scientific, Waltham, Masseuses, USA), and nuclease free water to a final volume of 20µl. Reverse transcription occurred at 42°C for 60 min followed by a 5 min incubation at 85°C to stop the reaction. Amplification was done by using 2 µl of synthesized cDNA, 10µl of the GoTaq Hot Start Green Master Mix (Promega, Madison, Wisconsin), 0.375 µM of each of the forward and reverse primers and 6.5 µl of nuclease free water for a final volume of 20 µl. Reaction conditions were as follows: 2 min at 94°C, 35 cycles of 20 sec at 94°C, 30 sec at 60°C and 20 sec at 72°C, finishing with an extension for 1 min at 72°C. Amplicons were visualized on a 1% agarose gel stained with 5 µl/L ethidium bromide (EtBr). The reactions for the HA16-5, T30 and T3 detection are as described above, except for the use of the DreamTaq Green PCR Master Mix 2X (Thermo Scientific, Waltham, Massachusetts, USA), instead of the GoTaq Green Master Mix. Annealing temperatures for the genotype specific primers (Table 1) differed; annealing temperatures for primer pairs T36, T30, T3, VT, NZ RB1 and NZRB2 was 60°C, B165 was 59°C and HA16-5 was 56°C.

**Table 1** Primers used for the characterization of the New Venture 41/2 source

| Primer | Nucleotide Sequence (5' to 3') | Product Size | Reference |
|---|---|---|---|
| A-forward | ACGTGTTCGTGAAACGCGG | 528 | (Rubio et al., 2001) |
| A-reverse | GTCGATAACTCGACAAACGAGC | | |
| T36 F | TTCCCTAGGTCGGATCCCGAGTATA | 836 | (Roy et al., 2010) |
| T36 R | CAAACCGGGAAGTGACACACTTGTTA | | |
| B165 F | GTTAAGAAGGATCACCATCTTGACGTTGA | 510 | |
| B165 R | AAAATGCACTGTAACAAGACCCGACTC | | |
| T3 F | GTTATCACGCCTAAAGTTTGGTACCACT | 409 | |
| T3 R | CATGACATCGAAGATAGCCGAAGC | | |
| VT F | TTTGAAAATGGTGATGATTTCGCCGTCA | 302 | |
| VT R | GACACCGGAACTGCYTGAACAGAAT | | |
| T30 F | TGTTGCGAAACTAGTTGACCCTACTG | 206 | |
| T30 R | TAGTGGGCAGAGTGCCAAAAGAGAT | | |
| NZRB1 F | AGTGGTGGAGATTACGTTG | 646 | (Cook et al., 2015) |
| NZRB1 R | TACACGCGACAAATCGAG | | |

| | | | |
|---|---|---|---|
| NZRB2 F | CGGAAGGGACTACGTGGT | 662 | |
| NZRB2 R | CGTTTGCACGGGTTCAATG | | |
| HA16-5 F | TAGGAAGGGTCACTGCCCTGACA | 658 | |
| HA 16-5 R | GTAAGTATCTAAAACCAGGAG | | |
| p33 Univ F | GATGTTTGCCTTCGCGAGC | 1000 | (Read, 2016) |
| 1kb Univ-p33-R | CCCGTTTAAACAGAGTCAAACGG | | |

## 2.4 Immuno-capture template preparation for Illumina sequencing

Template preparation for Illumina sequencing was done by immuno-capture of virus particles, release of RNA, and random cDNA synthesis followed by amplification. For the immuno-capture, 200 µl of a 1:5000 dilution of the CREC 29 antibody (Citrus Research and Education Center, Lake Alfred, Florida, USA) in coating buffer (15 mM $Na_2CO_3$ and 35 mM $NaHCO_3$ with a pH of 9.6) was added to the wells of the ELISA plate and incubated overnight at 4ºC. Three wash steps of 3 minutes with PBST (137 mM NaCl, 1.5 mM $KH_2PO_4$, 8 mM anhydrous $Na_2HPO_4$ and 2.7 mM KCL (pH 7.4) with 1ml/L Tween-20) was carried out before 200 µl of the sample was added to the wells. The samples (including CTV free Mexican lime as a healthy control) were prepared by macerating leaf petioles, midribs and bark shavings in coating buffer (2 ml buffer per 0.5 g sample) using the HOMEX 6 (Bioreba, Reinach, Basel-Landschaft, Switzerland). As a negative control, 200 µl of coating buffer was added to 20 wells. After 4 hours of incubation at 37ºC, a wash step of 10 minutes x 2 and 3 minutes x 5 was performed with PBST. To release virus particles from antibodies bound to the plate walls, 200 µl of virus release buffer (TE Buffer with 0.5% Tween-20) (Papayiannis et al., 2010) was placed into the wells. The ELISA plate was covered with microplate adhesive film and wet paper towel to prevent evaporation and incubated at 65ºC for 10 minutes with 3 seconds of vortexing once a minute. After virus release random cDNA synthesis was done using a reverse primer mix containing a mix of CTV degenerate primers and random hexamers (Table 2). Two microliters of this reverse primer mix was combined with 2 µl RNA, 2 µl 10x bovine serum albumin (BSA), 10µl 100% dimethyl sulfoxide (DMSO) and 7 µl water. This mixture was incubated at 95°C for 3 min and then left on ice for 1 minute for primer annealing. This 10 µl mix was combined with 1x RT Buffer, 25 U AMV Reverse Transcriptase, 10 U Protector RNAse Inhibitor (Roche Diagnostics, Mannheim, Germany), 1mM of a dNTP mix (KAPA Biosystems, Cape Town, South Africa), 10 mM DTT (Thermo Scientific, Waltham, Masseuses, USA), and PCR grade water to a total volume of 20 µl. Reverse transcription was performed at 42°C for 60 min, followed by a 85°C incubation for 5 min to inactivate the enzymes.

**Table 2** Primer mixes used for random cDNA synthesis and random PCR. For each primer mix, 10 µl of the 100 µM stock of each included primer was combined to get an 8.3 mM final concentration of each primer in the mix. IUPAC nucleotide code used, where Y=C/T, R=A/G, K=G/T, W=A/T, V=A/C/G, B=C/G/T, N=A/G/C/T.

| | **Names of primers used** | **Short Name** | **Location on Aligned Sequence**[a] | **Sequence (5' to 3')** |
|---|---|---|---|---|
| Forward primer mix | CTV Primer Pair 1 Fwd | CTV 1 F | (17397,17414) | ATCTCGTCRCTTTGTTTA |
| | CTV Primer Pair 2 Fwd | CTV 2 F | (15709,15727) | AYAARACGAAARCGGARGA |
| | CTV Primer Pair 3 Fwd | CTV 3 F | (13922,13939) | GGGGAAGYGATTTGGAAA |
| | CTV Primer Pair 4 Fwd | CTV 4 F | (11871,11888) | GTTTGYGTTTTAGTRGTK |
| | CTV Primer Pair 5 Fwd | CTV 5 F | (9782,9799) | BGYGGARGARCARATYWC |
| | CTV Primer Pair 6.1 Fwd | CTV 6.1 F | (8672,8687) | TYCACCGRAAYGAYYT |
| | CTV Primer Pair 7.1 Fwd | CTV 7.1 F | (7013,7028) | AYTTYGARCARATSGR |
| | CTV Primer Pair 8.1 Fwd | CTV 8.1 F | (5410,5428) | CTATGGARRTRGGRWCRAA |
| | CTV Primer Pair 9.1 Fwd | CTV 9.1 F | (3290,3305) | GAGTGRGARYCARCAR |
| | CTV Primer Pair 10 Fwd | CTV 10 F | (1488,1503) | GGGRCYTACACHTTTG |
| | CTV Primer Pair 11.1 Fwd | CTV 11.1 F | (54,69) | RGGAHYYGGWRTARRT |
| | Random hexamers | $(N)_6$ | n/a | NNNNNN |
| Reverse primer mix | CTV Primer Pair 1 Rev | CTV 1 R | (19354,19372) | CTTCTTTGGTTCACRCATA |
| | CTV Primer Pair 2 Rev | CTV 2 R | (17790,17807) | AAGGYAARAGCGAWGGRA |
| | CTV Primer Pair 3 Rev | CTV 3 R | (16085,16102) | GACGCTCGAAGRATRATR |
| | CTV Primer Pair 4 Rev | CTV 4 R | (14334,14353) | RGTTTTGTAAGYWTCTATYT |
| | CTV Primer Pair 5 Rev | CTV 5 R | (12219,12236) | AACTCWGARGRYGYAGCC |
| | CTV Primer Pair 6.1 Rev | CTV 6.1 R | (10742,10757) | ACRTACCAACCYCTRA |
| | CTV Primer Pair 7 Rev | CTV 7 R | (9164,9179) | GWMGCRGTMTYRTCGT |
| | CTV Primer Pair 8.1 Rev | CTV 8.1 R | (7710,7725) | WCATMARYGRRGCYYT |
| | CTV Primer Pair 9.1 Rev | CTV 9.1 R | (5600,5615) | CDGARAAYARRGAHGA |
| | CTV Primer Pair 10.1 Rev | CTV 10.1 R | (3767,3782) | RTCCARCARYTCRCCR |
| | CTV Primer Pair 11 Rev | CTV 11 R | (1910, 1925) | RCACAWVTCRTCRAAR |
| | Random hexamers | $(N)_6$ | n/a | NNNNNN |

[a] Alignment of 45 CTV reference genomes were used for primer design.

For amplification of the cDNA, 6 µl of the random cDNA mixture was used in a 60 µl PCR reaction containing 6 µl of both the forward and reverse primer mix (Table 2, final concentration of each primer in mix 8.3 mM), 30 µl GoTaq Hot Start Green Master Mix (Promega, Madison, Wisconsin) and 12 µl PCR grade water. Reaction conditions were as follows: 94°C for 2 min, followed by 60 cycles of 94°C for 20 sec, 55°C for 30 sec and 72°C for 20 sec. The final extension period was 1 min at 72°C. To ensure synthesis of cDNA across the CTV genome, PCR (i.e. no reverse transcriptase step) was successfully conducted using several different primers to amplify different regions of the genome (data not shown).

### 2.5 Amplification of p33 gene for Illumina sequencing

Total RNA (2 µl) was extracted according to the manufacturers specifications with the GeneJET Plant RNA Purification Mini Kit (Thermo Scientific, Waltham, Massachusetts) from Mexican lime leaves infected with New Venture 41/2 and was combined with 1 µl of 10 µM 1kb Univ-p33-R (Table 1) and 7µl of nuclease free water and incubated at 95°C for 3 min, 1 min on ice.  This was added to 4 µl RT buffer, 0.21 U AMV reverse transcriptase (Roche Diagnostics, Germany), 0.5 U Ribolock RNAse Inhibitor (Thermo Scientific, Waltham, Massachusetts), 1 mM of each dNTP, 10 mM DTT and nuclease free water to a final volume of 20 ul and incubated 42°C for 60 min followed by a 5 min incubation at 85°C. Amplification was done by using 4 µl of synthesized cDNA, 20 µl of the GoTaq Hot Start Green Master Mix (Promega, Madison, Wisconsin), 0.375 µM of each of the p33 forward and reverse primers (Table 1) and 13 µl of nuclease free water for a final volume of 40 µl. Reaction conditions were as follows: 2 min at 94°C, 35 cycles of 20 sec at 94°C, 45 sec at 65°C and 1 min at 72°C, finishing with an extension for 10 min at 72°C. Amplicons were visualized on a 1 % agarose gel stained with 5 µl/L EtBr and the gel was purified using the NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel, Düren, Germany) according to the manufacturer's instructions.

### 2.6 Illumina sequencing and data analysis

Immuno-captured and randomly amplified cDNA as well as the p33 amplicons were sequenced on the Illumina MiSeq platform (Illumina Inc, San Diego, California, USA) at the Agricultural Research Council Bioinformatics platform at Onderstepoort, Pretoria.

Data sets were analyzed on CLC Genomics Workbench 6 (CLC Bio, Aarhus, Denmark) with default parameters except where stated otherwise. Sequence reads were imported as paired end data and trimmed based on quality and sequences of the TruSeq adapters which were used during sequencing. Trimmed reads were mapped to a reference data set containing the 45 complete CTV genomes (for immuno-capture dataset) and cognate p33 gene region derived from whole CTV genomes (for amplicons dataset) available on GENBANK at that stage. Accession numbers for these genomes are: AB046398.1, AF001623.1, AF260651.1, AY170468.1, AY340974.1, DQ151548.1, DQ272579.1, EU076703.3, EU857538.1, EU937519.1, EU937520.1, EU937521.1, FJ525431.1, FJ525432.1, FJ525433.1, FJ525434.1, FJ525435.1, FJ525436.1, GQ454869.1, GQ454870.1, HM573451.1, JF957196.1, JQ061137.1, JQ798289.1, JQ911663.1, JQ911664.1, JQ965169.1, JX266712.1, JX266713.1, KC262793.1, KC333868.1, KC517485.1, KC517486.1, KC517487.1, KC517488.1, KC517489.1, KC517490.1, KC517491.1, KC517492.1, KC517493.1, KC517494.1, KC525952.1, NC001661.1, U16304.1 and Y18420.1. Mapping was done with length fraction set to 0.5, similarity fraction set to 0.75 for the immuno-capture template and set to 0.9 for both parameters for the amplicons template. Non-specific match handling was set to 'random'. Furthermore, *de novo* assembly was performed on the immuno-capture dataset and all contigs larger than 500 bp were blasted on the NCBI website using the blastn platform (Altschul et al., 1990).

## 3. Results

### 3.1 Cloning and sequencing of the ORF 1a fragment

The phylogenetic dendrogram (Figure 1) constructed from the sequences of the ORF 1a fragment clones indicated a low diversity of the sequences with the VT genotype being predominant with all clones (n = 102) grouping together with the Kpg3 (HM573451.1) reference sequence.

Although the sequences of all of the clones group with VT, slight variations (0.5% divergence) between the sequences might reflect errors occurring during amplification or are due to the quasispecies nature of a typical RNA viral population (Domingo et al., 1998).

### 3.2 Genotype detection RT-PCR in New Venture 41/2

Genotype specific RT-PCRs on the New Venture 41/2 source revealed that it consists of the T36, VT, RB, B165 and HA16-5 genotypes. However published T36 primers also amplify RB isolates and hence the positive result obtained with these primers might be due to the presence of either T36 primers or RB, and was reassessed by Illumina sequencing.

### 3.3 Immuno-capture template preparation for Illumina sequencing

Just over 4 million reads were obtained with Illumina sequencing of the New Venture 41/2 source following virus particle enrichment by immuno-capture. When doing reference mapping, only a very small percentage (1.23%) of the total reads mapped back to the CTV reference genomes. While the CTV specific reads constituted only a small portion of the total reads obtained, they still allow valuable analysis. From the mapping data it is evident that this source contains multiple genotypes (Figure 2).

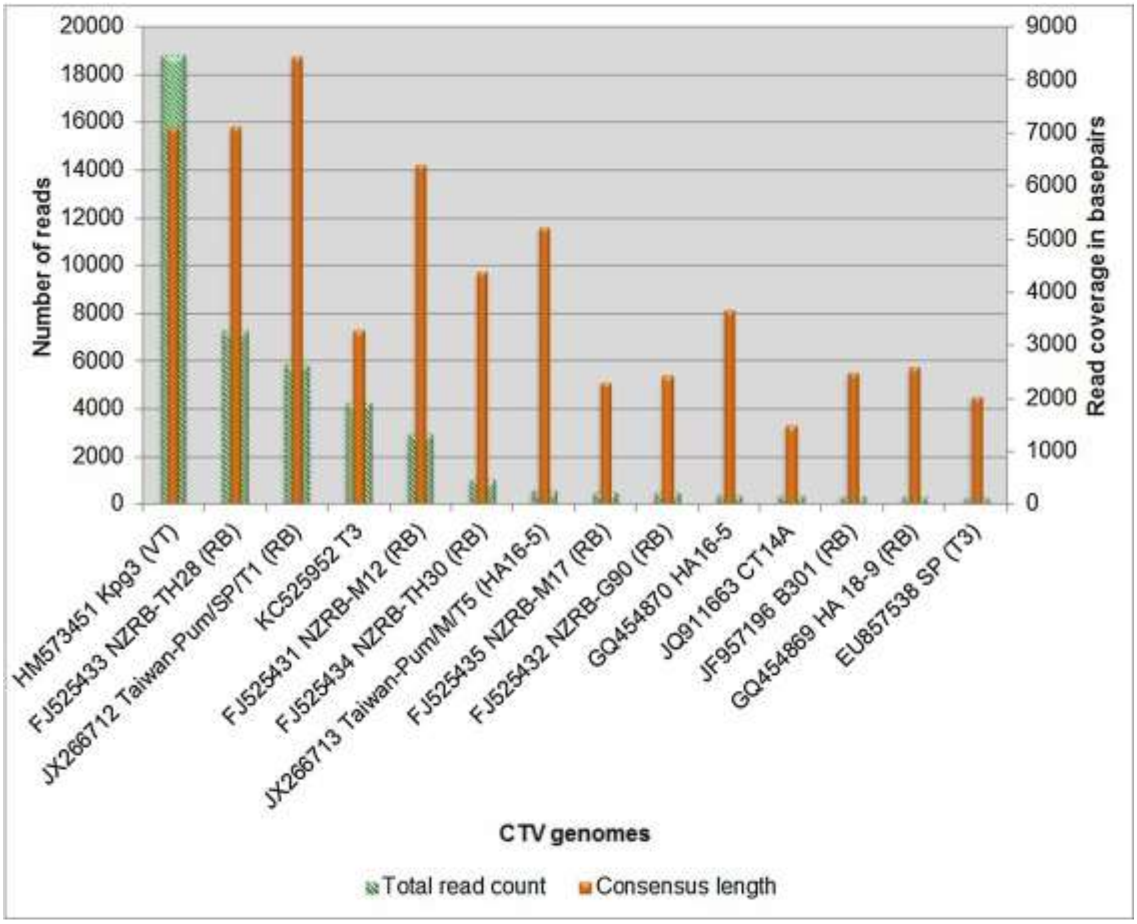

**Fig. 2** Reference mapping of a CTV New Venture 41/2 source, enriched for virus particles by immuno-capture template sequence reads against CTV genomes, showing total number of mapped reads and consensus length of the mapped reads in base pairs. Due to lack of space, figure shows only first 14 genomes of 45 genomes mapped to, arranged in descending order by total read count. Complete figure can be found in supplementary data (Sup Figure 1)

The majority of reads (18 817, 37.4% of total CTV reads) mapped to the Kpg3 strain (accession HM573451) which groups within the VT genotype. Also, a considerable number of reads (19 139, 38% of total CTV reads) mapped back to different strains from the Resistance Breaking (RB) genotype (accessions FJ525433, JX266712, FJ525431, FJ525434, FJ525435, FJ525432, JF957196 and GQ454869), while 4 292 reads (8.5% of total CTV reads) also mapped to the T3 strain (accession KC525952). Illumina results obtained in a previous study demonstrated that for relative homogenous sources the consensus length was high (~size of the fragment mapped to) for the single strain to which the vast majority of reads (80 – 90%) mapped (Lubbe, 2015). In contrast in this case we observed that the consensus length of reads obtained was notable for multiple of the reference sequences, confirming the heterogeneity of this source. It is not just small fragments of heterologous genomes that were sequenced, which may be suggestive of recombination events within a single genotype, but rather presence of a second complete genome of which a large  region was sequenced. Visualization of the distribution of reads mapping to different genomes can be seen in Figure 3.

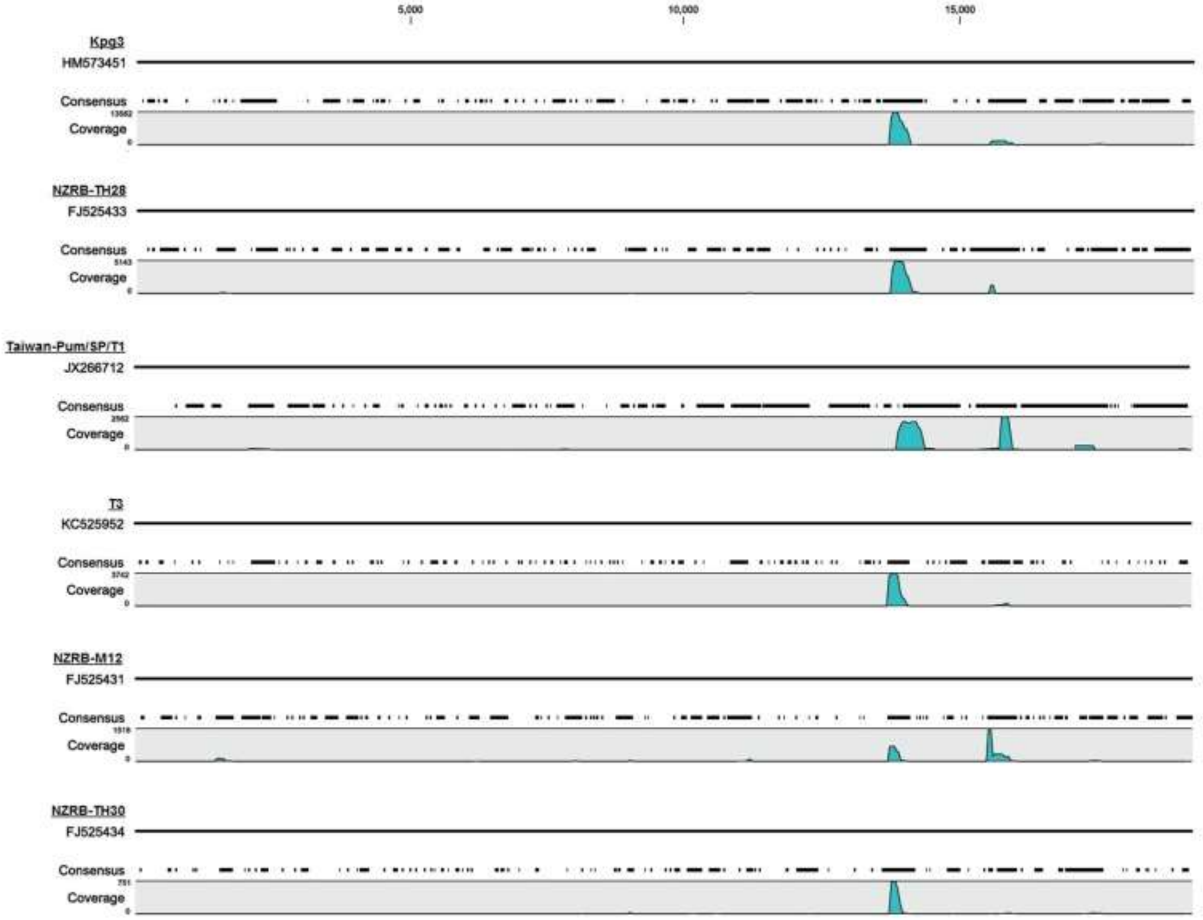

**Fig. 3** Distribution of New Venture 41/2 immuno-capture template sequence reads to those reference CTV genomes (line next to GenBank accession number) with the highest identity, showing areas where a consensus sequence can be created (interrupted line under reference genome) and the coverage of these areas (graph under consensus sequence).

After *de novo* assembly, contigs greater than 300bp were subjected to BLAST against the GenBank sequence database. Sixteen of the 1217 contigs returned as CTV related, ranging in size from 300 bp to 598 bp. Seven of the CTV related contigs, ranging from 312 bp to 534 bp were most closely related to the Kpg3 strain (HM573451.1) while one other contig (598 bp) was most similar to the FS701-VT (KC517494.1), both being part of the VT genotype clade. There were five contigs that were similar to strains within the RB genotype, three being similar to NZRB-TH28 (FJ525433.1) and Taiwan-Pum/SP/T1 (JX266712.1) and NZRB-M12 (FJ525431.1), respectfully. All these contigs range from 300 bp to 335 bp. Two other contigs of about the same size were most similar to the Taiwan-Pum/M/T5 (JX266713.1) strain which forms part of the HA16-5 genotype.

The last contig of 508 bp was most similar to the CT14A (JQ911663.1) isolate, which is closely related to the B165 genotype.

The remaining contigs were similar to host DNA, but also among others Bacteriophage S13, *Helianthus annuus* (sun flower) mitochondrion, *Rhipicephalus appendiculatus* (tick*)*, and *Mannheimia haemolytica* (bacterium causing epizootic pneumonia in cattle).

**3.4 Amplification of p33 gene for Illumina sequencing**

Reference mapping of the p33 gene reads obtained yielded 185 776 (25.8%) CTV specific reads of the total of 719 490 reads. This is a higher percentage CTV specific reads than obtained when the immuno-captured viral particles was used as template, but still low when considering that these were amplicons being sequenced which were gel purified to eliminate all unspecific amplification.

The majority of CTV-specific reads (80 300, 43.2% of total CTV reads) obtained from sequencing the p33 gene mapped to a strain within the RB genotype clade (NZRB-M12, accession FJ525431), followed by reads mapping to the Kpg3 strain (accession HM573451) which is part of the VT genotype (40 778 reads, 22% of total CTV reads) and then to a second  isolate within the RB genotype clade, Taiwan-Pum/SP/T1 (accession JX266712) (35 982 reads, 19.4%  of total CTV reads). Other strains to which small numbers of reads mapped covering the entire 977 bp of the reference length, belong to the RB, T3, HA 16-5 and A 18 genotypes. Reads mapping to less than the entire expected 977bp of references were not considered indicative of the presence of that genotype and were not analysed further (Figure 4).

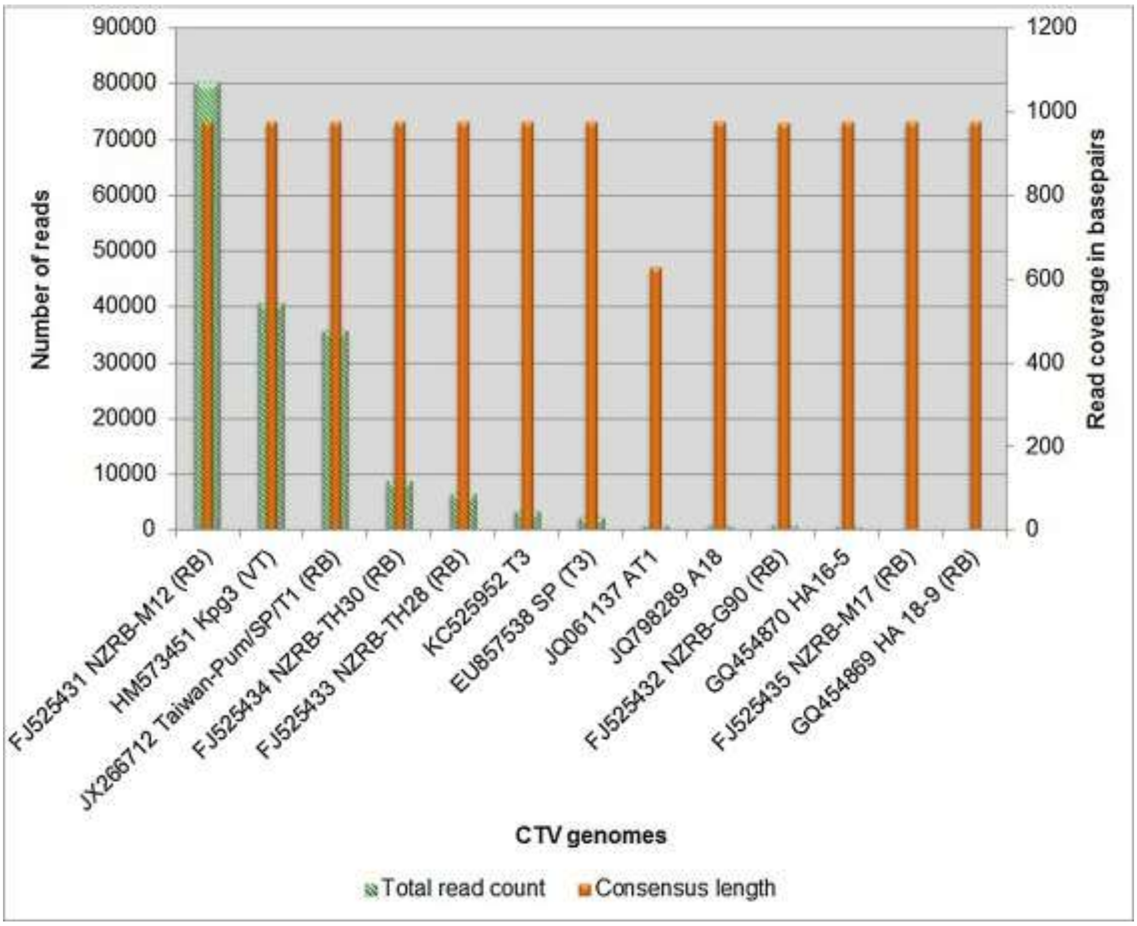

**Fig. 4** Reference mapping of New Venture 41/2 when amplicons of p33 gene was used as  template for Illumina sequencing and mapped to the cognate CTV p33 amplicon region from published CTV whole genome references area, showing total number of mapped reads (primary axis) and consensus length of the mapped reads (secondary axis) in base pairs. Due to lack of space, figure shows only first 14 genomes of 45 genomes mapped to, arranged by descending order of total read count. Complete figure can be found in supplementary data (Sup Figure 2)

## 4. Discussion

Although the different methods used to determine the genotype composition of the New Venture 41/2 source did not provide the same results, it is very clear that this is a source containing multiple genotypes. Genotype specific primers focussing on the 5' end of the genome suggest that the source contains VT, RB, B165, HA16-5 and possibly T36 genotypes as part of its population. The presence of VT, RB, B165 and HA16-5 could be confirmed with Illumina sequencing when looking at both data sets. Reference mapping indicated that VT and RB were dominant within the source with only a few reads mapping to HA16-5 (2% for IC template and 0.5% for p33 amplicon) although the full consensus sequence was covered in the case of the p33 amplicon. A *de novo* contig obtained when immune-capture of particles preceded Illumina sequencing blasted to an isolate that is most closely related to the B165 genotype isolates. Although some reads mapped to T36, the read count and consensus length were very low for this genotype. If not for the genotype specific RT-PCR results it would not have been considered to be present within the source. The positive reaction of the T36 primers is most probably due to the ability of the T36 primers to also amplify the RB strains. The establishment of a threshold regarding the presence/absence of a genotype remains vexing as reads from the Illumina data mapped back to every one of the reference genomes, but PCR results only confirm the presence of 4 different genotypes. This may be explained by the fact that some regions of the genome are conserved amongst strains (Mawassi et al., 1996) and that mapping of small fragments can occur with all of these regions. The BLAST results from the contigs obtained from the immuno-capture enriched template with *de novo* assembly also confirmed the presence of RB and VT, in addition to the HA16-5 genotype, and a contig similar to the CT14A strain which is most closely related to the B165 genotype. The Illumina data obtained when using amplicons as template showed a complete consensus length for reads mapping to the T3 and A18 genotypes. The presence of T3 could not be confirmed with the immuno-capture enriched genomic RNA data or the CTV genotype specific RT-PCRs. The A18 genotype could be seen in both Illumina datasets, but it was not possible to confirm its presence with RT-PCR as there is no primers specific to this genotype available yet. Primers used in the RT-PCRs do not allow detection of this isolate, as assessed with CLC Genomics Workbench 6 (CLC Bio, Aarhus, Denmark) with the "find primer binding site" function.

The 1a fragment clones proved the least informative, confirming that the 1a fragment primers are biased towards certain genotypes and that, in general, impractically large numbers of clones need to be sequenced for the approach to be accurate when dealing with multiple infections (Cook et al., 2015). While it was shown that using these primers, four CTV genotypes (VT, T30, B165 and RB) could be detected after the analysis of 117 clones from one source (Cook et al., 2015), the current study could only detect VT in the mixed source tested here.

Use of immuno-capture enriched genomic RNA as template did not yield the high quality results expected; with the number of CTV specific reads received still only being a small component of total reads obtained. A much higher value was expected since the immuno-capture was used for the purpose of enriching for CTV particles. It remains possible however that many other RNAs could have been present during the random cDNA synthesis and amplification since ELISA is a crude technique. Some of the CTV RNA might also have been damaged or lost during the process of virus release by heating. Although we could detect CTV cDNA within the sample, it might still have been present at very small amounts. A real time RT-PCR was used to test which virus release buffers yielded the best results (data not shown), but is a very sensitive method and it is possible that the amount of CTV RNA after release was insufficient for downstream processes and sequencing, while still being detectable with real time RT-PCR.

This method of enrichment is therefore not recommended for future use. Although the immuno-capture itself is a very efficient way of enriching for the virus, the inefficient conversion of the RNA to DNA and the amplification of the DNA without specific primers may be the problem. It might be more valuable in systems where the virus being sequenced has a poly(A) tail that can be used for cDNA synthesis, or is not as diverse that would allow the use of specific primers without introducing bias.

Using amplicons as template for Illumina sequencing produced more reads mapping to CTV than with the immuno-capture particle enrichment template. Amplicon templates however only allow conclusions limited to the specific gene fragment amplified, and no information on whether recombination took place in any other region of the virus genome and hence if the gene sequenced is representative of the entire genome. While it is not known whether the p33 gene is the only gene playing a role in cross-protection, it does appear as though it is responsible for the genotype specificity of super-infection exclusion (Folimonova, 2012). This makes it particularly important to characterize this gene region of circulating CTV strains to identify potential cross-protection sources, as was done in this study.

The low number of reads mapping back to CTV genomes after using Illumina sequencing was unexpected. When less stringent mapping conditions were used, reads mapped to areas of the genome that was not sequenced (in the case of p33 amplicon sequencing), indicating unspecific mapping. Some of the remaining sequences do blast back to host DNA, indicating that host genetic material may be contaminating the sample. A method by Adams *et al. (2009)* suggests subtractive hybridization to remove any host genetic material after cDNA synthesis, and should be considered in future studies.

The presence of reads unrelated to citrus may indicate index leaching took place. During Illumina sequencing, multiplexing is done to increase the amount of samples that can be done in one run. This is established through ligating sequences to the template that functions as barcodes or indexes. These indexes are used after sequencing to sort samples that were multiplexed. It is possible that cross-contamination can take place during multiplexing. Advances like double-indexing have been made to prevent this and can be incorporated in future studies (Kircher, Sawyer, and Meyer, 2011).

Genotype specific RT-PCRs allow for a relatively quick and economical identification of genotypes within a viral population. The largest concern however is the lack of exploration of the rest of the genome as with amplicon sequencing. During this study it was also found that this method is prone to contamination necessitating the need to duplicate reactions to confirm results.

When setting out to characterize a viral population there are multiple methods that can be used. It is crucial to consider the virus properties when choosing a characterization method. If dealing with a virus where the genome is conserved and recombination between different genotypes does not occur commonly, a PCR and conventional sequencing method can be used. Illumina sequencing is an appealing option to gain large amounts of information regarding the population, but there are several challenges that have to be overcome. If a specific gene is targeted for sequencing, results obtained may be biased and the area of the genome to which inferences can be made is limited. When targeting the entire viral genome, one of the challenges are enriching for the virus in an unbiased manner, which may prove difficult when dealing with RNA viruses where poly(A) tails does not occur.

## 5. Acknowledgments

We thank Citrus Research International (CRI) (South Africa), the University of Pretoria for the financial support of this study as well as the National Research Foundation-THRIP program for partial funding. We also wish to thank Kobus Breytenbach, CRI, for providing the New Venture 41/2 source studied as well as Glynnis Cook, CRI, and Kirsti Snyders, UP, for their advice on genotype specific RT-PCRs.

## 6. References

Adams, I.P., Glover, R.H., Monger, W.A., Mumford, R., Jackeviciene, E., Navalinskiene, M., Samuitiene, M. and Boonham, N., 2009. Next-generation sequencing and metagenomic analysis: a universal diagnostic tool in plant virology. Mol Plant Pathol 10, 537-545.

Albiach-Marti, M.R., da Graça, J.V., van Vuuren, S.P., Guerri, J., Cambra, M., Laigret, F. and Moreno, P., 1996. The effects of different hosts and natural disease pressure on molecular profiles of mild isolates of *Citrus tristeza virus* (CTV). Proc 13th Conference of the IOCV, 147-153.

Albiach-Marti, M.R., Mawassi, M., Gowda, S., Satyanarayana, T., Hilf, M.E., Shanker, S., Almira, E.C., Vives, M.C., Lopez, C., Guerri, J., Flores, R., Moreno, P., Garnsey, S.M. and Dawson, W.O., 2000. Sequences of *Citrus tristeza virus* Separated in Time and Space Are Essentially Identical. The Journal of Virology 74, 6856-6865.

Albiach-Marti, M.R., Robertson, C., Gowda, S., Tatineni, S., Belliure, B., Garnsey, S.M., Folimonova, S.Y., Moreno, P. and Dawson, W.O., 2010. The pathogenicity determinant of *Citrus tristeza virus* causing the seedling yellows syndrome maps at the 3′-terminal region of the viral genome. Mol Plant Pathol 11, 55-67.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J., 1990. Basic local alignment search tool. J Mol Biol 215, 403-410.

Bar-Joseph, M., Marcus, R. and Lee, R.F., 1989. The Continuous Challenge of *Citrus tristeza virus* Control. Annu Rev Phytopathol 27, 291-316.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Wheeler, D.L., 2005. GenBank. Nucleic Acids Res 33, 34-38.

Breytenbach, J.H.J., Cook, G. and Van Vuuren, S.P. 2014a. Progress report: Cross-protection of  Star Ruby using Beltsville sub-isolates of Nartia Mild Strain for the Orange River Valley, Citrus Research International, Nelspruit, South Africa, pp. 24.

Breytenbach, J.H.J., Cook, G. and Van Vuuren, S.P. 2014b. Progress report: Cross-protection of Marsh and Star Ruby by using the best field isolates collected in the different grapefruit production areas of southern Africa., Citrus Research International, Nelspruit, South Africa, pp. 105.

Cook, G., van Vuuren, S.P., Breytenbach, J.H.J., Burger, J.T. and Maree, H.J., 2015. Expanded Strain-Specific RT-PCR Assay for Differential Detection of Currently Known *Citrus tristeza virus* Strains: a Useful Screening Tool. J Phytopathol, 12454.

Domingo, E., Baranowski, E., Ruiz-Jarabo, C.M., Escarmis, C., Martin-Hernandez, A.M. and Saiz, J.C., 1998. Quasispecies structure and persistence of RNA viruses. Emerg Infect Dis 4, 521.

Folimonova, S.Y., 2012. Superinfection Exclusion Is an Active Virus-Controlled Function that Requires a Specific Viral Protein. J Virol 86, 5554-5561.

Folimonova, S.Y., Robertson, C.J., Shilts, T., Folimonov, A.S., Hilf, M.E., Garnsey, S.M. and Dawson, W.O., 2010. Infection with Strains of *Citrus tristeza virus* Does Not Exclude Superinfection by Other Strains of the Virus. J Virol 84, 1314-1325.

Fulton, R.W., 1986. Practices and precautions in the use of cross protection for plant virus disease control. Annu Rev Phytopathol 24, 67-81.

Harper, S., Dawson, T. and Pearson, M., 2010. Isolates of *Citrus tristeza virus* that overcome *Poncirus trifoliata* resistance comprise a novel strain. Arch Virol 155, 471-480.

Harper, S.J., 2013. *Citrus tristeza virus*: Evolution of complex and varied genotypic groups. Front Micro 4, 93.

Karasev, A.V., Boyko, V.P., Gowda, S., Nikolaeva, O.V., Hilf, M.E., Koonin, E.V., Niblett, C.L., Cline, K., Gumpf, D.J., Lee, R.F., Garnsey, S.M., Lewandowski, D.J. and Dawson, W.O., 1995. Complete Sequence of the *Citrus tristeza virus* RNA Genome. Virology 208, 511-520.

Karasev, A.V., Dawson, W.O., Hilf, M.E. and Garnsey, S.M., 1998. Molecular Biology of *Citrus tristeza virus*: Implications for disease diagnosis and control. ACTA Horticulturae 472, 333-350.

Katoh, K., Misawa, K., Kuma, K.i. and Miyata, T., 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30, 3059-3066.

Kircher, M., Sawyer, S. and Meyer, M., 2011. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Res 40, e3.

Kong, P., Rubio, L., Polek, M. and Falk, B.W., 2000. Population Structure and Genetic Diversity within California *Citrus tristeza virus* (CTV) Isolates. Virus Genes 21, 139-145.

Lubbe, J. 2015. Molecular and biological characterization of three *Citrus tristeza virus* candidate cross-protection sources, Faculty of Natural and Agricultural Sciences Department of Microbiology and Plant Pathology, University of Pretoria, Pretoria.

Mawassi, M., Mietkiewska, E., Gofman, R., Yang, G. and Bar-Joseph, M., 1996. Unusual Sequence Relationships Between Two Isolates of *Citrus tristeza virus*. J Gen Virol 77, 2359-2364.

Melzer, M., Borth, W., Sether, D., Ferreira, S., Gonsalves, D. and Hu, J., 2010. Genetic diversity and evidence for recent modular recombination in Hawaiian *Citrus tristeza virus*. Virus Genes 40, 111-118.

Papayiannis, L.C., Hunter, S.C., Iacovides, T. and Brown, J.K., 2010. Detection of *Cucurbit yellow stunting disorder virus* in Cucurbit Leaves Using Sap Extracts and Real-time TaqMan® Reverse Transcription (RT) Polymerase Chain Reaction (PCR). J Phytopathol 158, 487-495.

Read, D.A. 2016. Overcoming bias in Citrus tristeza virus (CTV) genotype detection and a population study of CTV within Southern African Star Ruby grapefruit orchards, Department of Microbiology, University of Pretoria.

Read, D.A. and Pietersen, G., 2015. Genotypic diversity of Citrus tristeza virus within red grapefruit, in a field trial site in South Africa. Eur J Plant Pathol 142, 531-545.

Roy, A., Ananthakrishnan, G., Hartung, J.S. and Brlansky, R.H., 2010. Development and Application of a Multiplex Reverse-Transcription Polymerase Chain Reaction Assay for Screening a Global Collection of *Citrus tristeza virus* Isolates. Phytopathology 100, 1077-1088.

Roy, A. and Brlansky, R.H., 2010. Genome analysis of an orange stem pitting *Citrus tristeza virus* isolate reveals a novel recombinant genotype. Virus Res 151, 118-130.

Rubio, L., Ayllon, M.A., Kong, P., Fernandez, A., Polek, M., Guerri, J., Moreno, P. and Falk, B.W., 2001. Genetic Variation of *Citrus tristeza virus* Isolates from California and Spain: Evidence for Mixed Infections and Recombination. J Virol 75, 8054-8062.

Tamura, K., Dudley, J., Nei, M. and Kumar, S., 2007. Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol 24, 1596-1599.

Vives, M.C., Rubio, L., Sambade, A., Mirkov, T.E., Moreno, P. and Guerri, J., 2005. Evidence of multiple recombination events between two RNA sequence variants within a *Citrus tristeza virus* isolate. Virology 331, 232-237.

Zablocki, O. 2013. Unbiased, next-generation sequencing for the characterization of *Citrus tristeza virus* populations, Faculty of Natural and Agricultural Sciences Department of Microbiology and Plant Pathology, University of Pretoria.

Zablocki, O. and Pietersen, G., 2014. Characterization of a novel *Citrus tristeza virus* genotype within three cross-protecting source GFMS12 sub-isolates in South Africa by means of Illumina sequencing. Arch Virol 159, 2133-9.