Characterization of a novel citrus tristeza virus genotype within three cross-protecting source GFMS12 sub-isolates in South Africa by means of Illumina sequencing

Olivier Zablocki • Gerhard Pietersen

Abstract Tristeza disease (caused by citrus tristeza virus, CTV) is currently controlled in South Africa by means of cross-protection. In this study, we characterized the CTV populations of three grapefruit mild strain 12 (GFMS12) single-aphid-transmission-derived sub-isolates at the whole-genome level using Illumina sequencing technology. A novel South African isolate (CT-ZA3, of the T68 genotype) was shown to be the dominant genotype in all GFMS12 sub-isolates tested, along with reads unique to various other genotypes occurring as minor components. Uncertainty remains as to the significance of these minor components.

Keywords Tristeza virus · Next-generation sequencing · GFMS12

Nucleotide sequence data reported are available in the DDBJ/EMBL/GenBank databases under the accession numbers KC 333868 and KC 333869.

O. Zablocki · G. Pietersen
Department of Microbiology and Plant Pathology and FABI, University of Pretoria, New Agricultural Building, Pretoria 0002, South Africa
e-mail: Olivier.zablocki@fabi.up.ac.za

G. Pietersen (correspondence author)
Agricultural Research Council, Plant Protection Research Institute, Private Bag X134, Queenswood 0121, South Africa
e-mail: gerhard.pietersen@up.ac.za

Introduction

Citrus tristeza virus (CTV) (family *Closteroviridae*, genus Closterovirus) has been recognized as the causal agent of one of the most devastating diseases of citrus [1]. CTV has a host range that encompasses several members of the family Rutaceae, of which the genus *Citrus* is of particular economic importance [2]. Its genome is monopartite, composed of a positive-sense single-stranded RNA molecule approximately 19.3 kilobases (kb) in length [3]. Twelve open reading frames (ORFs) potentially encode 19 protein products, flanked by terminal 5′ and 3′ short untranslated regions (UTRs) with no 5′ methylated cap or poly (A) tail. Disease syndromes such as quick decline and stem pitting have been associated with this virus. The current method used for control of the disease in South Africa is cross-protection, but unfortunately, severe symptoms in some citrus cultivars may still occur despite pre-immunization [4]. Folimonova et al. [5] demonstrated that cross-protection, due to superinfection exclusion, is genotype specific. Ideally pre-immunizing sources should contain mild isolates of all genotypes circulating within agrospecific geographic ranges [6]. It is therefore crucial to obtain or isolate pure sources of all relevant genotypes and to assess whether they produce only mild symptoms. As a first step to achieving that goal, an accurate determination of the genotypes occurring in South African citrus orchards is essential, as well as the homogeneity status of aphid-transmitted GFMS12 sub-isolates (of which the mother tree has been used in the past for cross-protection against stem pitting of grapefruit) and candidate sources used for pre-immunization. Incongruences may be observed when different CTV gene phylogenies are compared. For example, Scott et al. [7], while characterizing the same GFMS12 sub-isolates [9] used in the current study, observed CTV

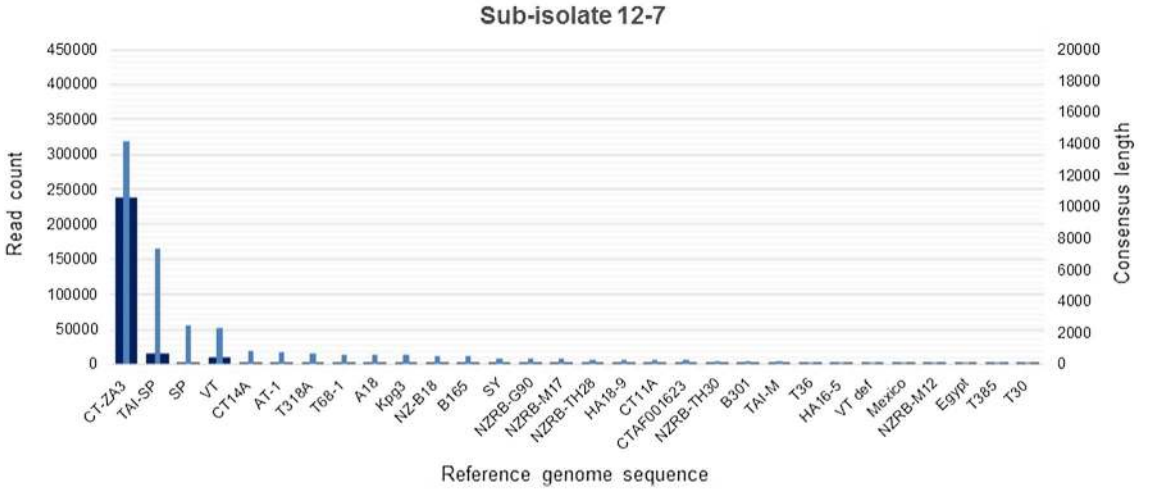
populations dominated by B165-like sequences when their analysis was based on the ORF1a region, but when p23 was used, the results suggested dominance by a VT-like genotype. These authors proposed the existence of a B165/VT recombinant amongst other genotypes in these sub-isolates. To resolve this, we randomly amplified [8] whole CTV genomes using dsRNA of the virus population via Illumina sequencing from three CTV GFMS12 sub-isolates (12-7, 12-8 and 12-9) grafted on Mexican lime indicator plants. This methodology was used to ensure that no genotype-specific bias was introduced.

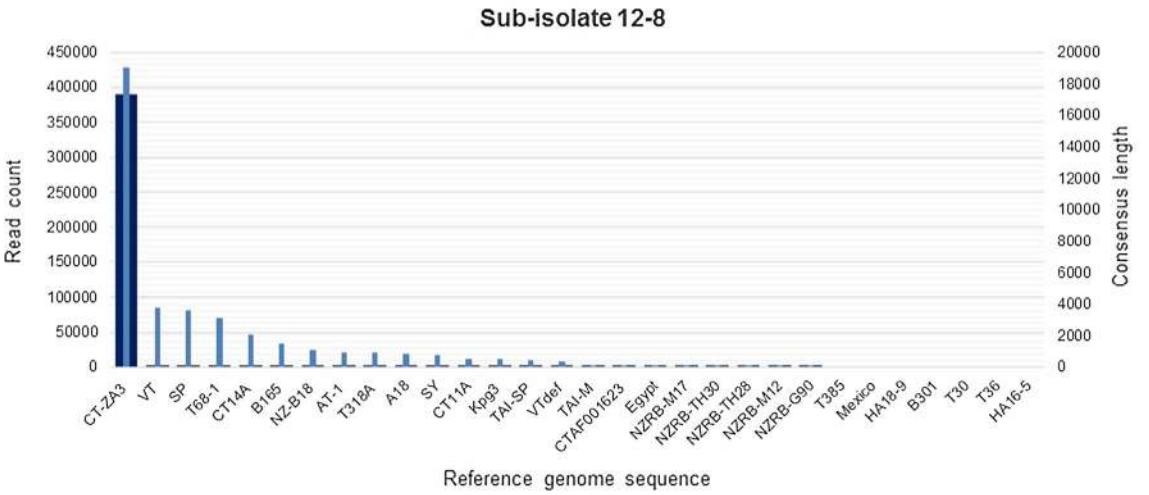
The three GFMS12 sub-isolates used in this study (12-7, 12-8 and 12-9) were maintained under glasshouse conditions on Mexican lime (*Citrus aurantifolia* (Christ.) Swing.). The sub-isolates had been derived by single-aphid transfers (*Toxoptera citricida* Kirkaldy) from a GFMS12-infected mother tree to Mexican lime seedlings [9]. Purification of double-stranded RNA was performed according to Morris and Dodds [10] with minor modifications, in which columns and bentonite were not used. dsRNA extracts were subjected to random-primed reverse transcription PCR performed according to Roosinck et al. [8] with minor modifications. Amplicons were sequenced individually on 1/15 of a lane in an Illumina HiScanSQ (Agricultural Research Council - Biotechnology Platform in Pretoria, South Africa). Paired reads were analyzed using CLC Genomics version 5.1. and aligned (reference assembly) to a set of 29 CTV full genomes obtained from the National Center for Biotechnology Information (NCBI) database (http://www.ncbi.nlm.nih.gov/). *De novo*-generated contigs were subjected to a multiBLASTn search against the nucleotide sequence collection hosted by the NCBI portal via the CLC interface. Multiple sequence alignments were made with the online version of MAFFT (version 6.952). Alignments and similarity percentages were processed in BioEdit v7.1.3.0 (Tom Hall, Isis Pharmaceuticals, Inc., 1997-2004). Phylogenetic analysis was performed using MEGA version 4 (http://www.megasoftware.net/). Unrooted dendrograms were inferred using the neighbor-joining method with a bootstrap test of a thousand pseudo-replicates. Evolutionary distances were determined by applying the Jukes-Cantor base substitution model.

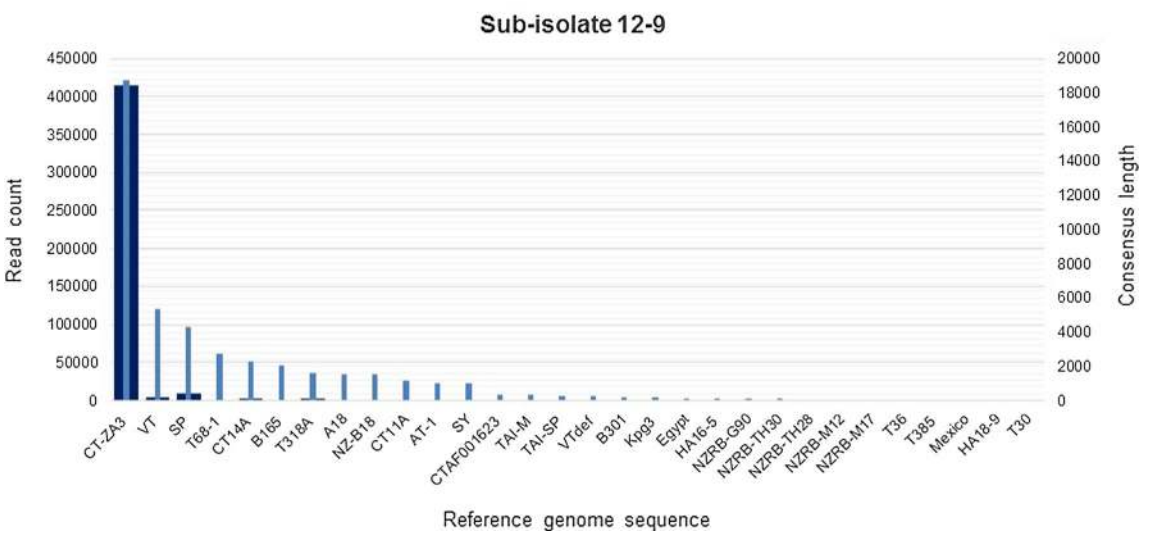
A total of 15,512,696 unprocessed 100-bp paired-end reads were obtained in total for the three samples. Three key mapping metrics were recorded: read count (RC), combined consensus length (CL) and average coverage (AC). In the *de novo* assemblies, several long contigs were obtained, with the longest ones being 17,274, 12,648 and 6668 bp. For the sake of clarity, each sample is presented separately. Genotype unique read mapping results for all sub-isolates are shown in Fig. 1, while the read genomic distribution over the most significantly mapped genotypes

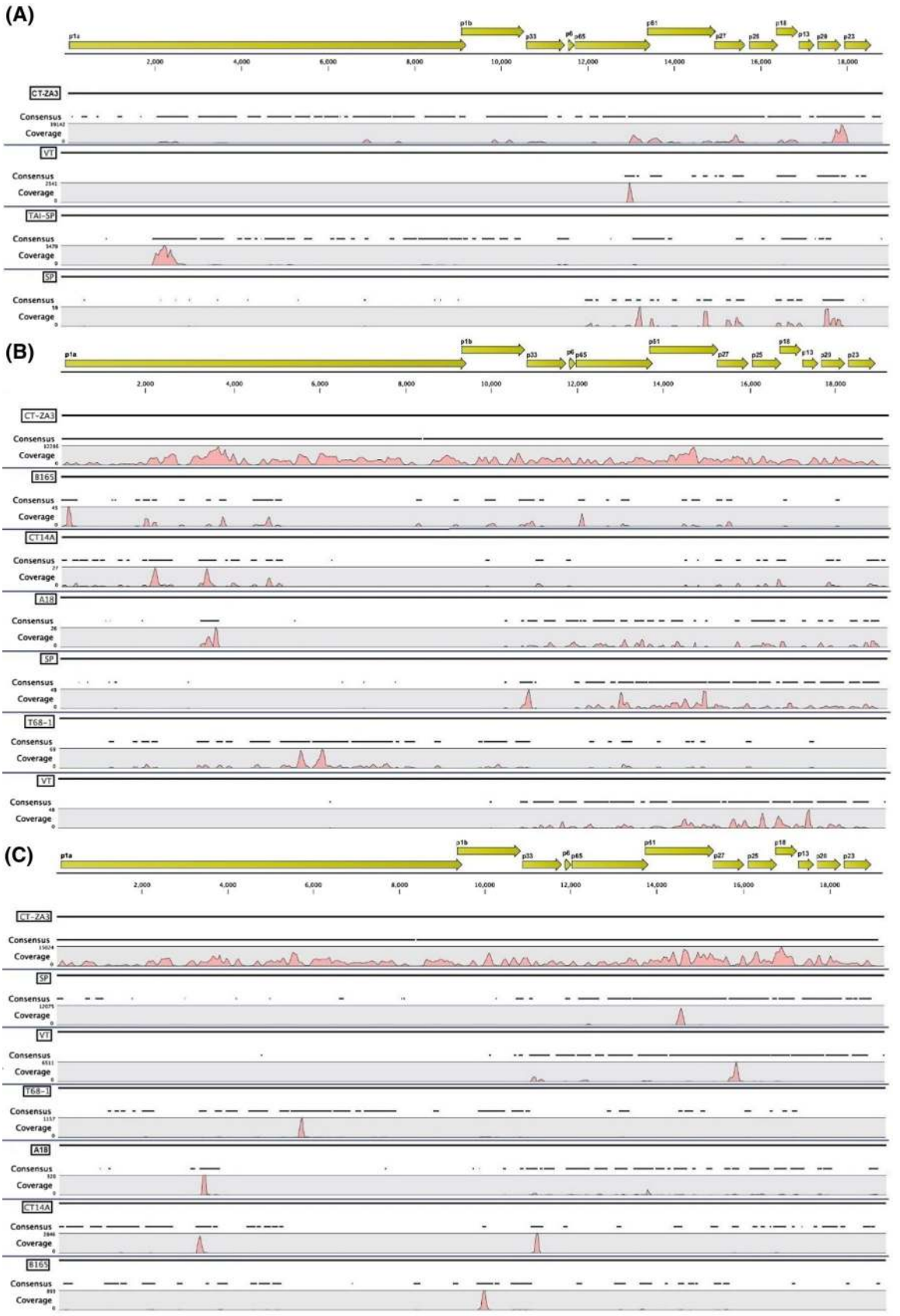
Fig. 1 Read mapping results vs. 30 CTV reference genomes (including CT-ZA3 from this study) for GFMS12 sub-isolates 12-7 (top), 12-8 (middle) and 12-9 (bottom). The mapped reference genotypes were ranked from largest to smallest using the specific combined consensus length obtained by reads. The left vertical axis scale represents the number of reads mapped. The right vertical axis represents the consensus length in nucleotides. Read count, thick dark blue bars; combined consensus length, thin light blue bars (color figure online) ▶

are included in Fig. 2. For GFMS12 sub isolate 12-8, *de novo* assembly produced three CTV-specific contigs of 12,648, 6643 and 108 bp in length. Identification via BLASTn from the NCBI database revealed that these contigs had closest homology to the T68-1 genotype (from Florida) [11]. A consensus sequence could be created from two contigs (12,648 nt and 6643 nt), which yielded a putative full genome sequence, 19,244 bp in length, which we named CT-ZA3 (GenBank accession number KC 333869). A neighbor-joining phylogeny was created (Fig. 3) which demonstrated CT-ZA3 to be most closely related to the T68-1 strain within the newly proposed T68 genotype clade [11]. A whole-genome identity matrix of CT-ZA3 against all other CTV references was conducted with the values shown next to each reference genotype in the dendrogram in Fig. 3. ORF prediction revealed that CT-ZA3 was organized into 12 ORFs, with a synteny typical of other CTV whole genomes. Reference mapping against a set of 30 full-genome CTV sequences, including the new sequence CT-ZA3 was conducted (Fig. 1, middle graph). The highest mapped read count (390,525 reads) were obtained for CT-ZA3, the longest combined consensus sequence (19,046 nt) and the highest average coverage (1776.5-fold). The next most frequently represented reference genotypes, present as low read counts but with significant total combined consensus lengths, were VT (RC = 155; CL = 3769; AC = 0.67), SP (RC = 179; CL = 3610; AC = 0.81), T68-1 (RC = 130; CL = 3094; AC = 0.53) and CT14A (RC = 71; CL2066; AC = 0.30). Distribution of the mapped reads (Fig. 2b) showed CT-ZA3 being fully mapped, with only one minor gap remaining: the VT and SP genotypes mapped mostly in their 3′ half, while T68-1 and CT14A mapped mostly in their 5′ half. Figure 1 (bottom graph) shows the reads from sub-isolate 12-9 mapped against the CTV reference set. As with sub-isolate 12-8, the majority of the reads (415,294 reads) mapped to CT-ZA3, resulting in an almost complete consensus length (18,763 nt) and an average coverage value of 1851-fold. The remaining unique reads, mapped to various other genotypes, mostly VT (RC = 4707; CL = 5340; AC = 20.82), SP (RC = 9397; CL = 4282; AC = 43.8), T68-1 (RC = 136; CL = 2742; AC = 0.54), CT14A (RC = 1859; CL = 2300; AC = 7.85), B165 (RC = 78; CL = 2038; AC = 0.31), and A18 (RC = 200;


Sub-isolate 12-7
Read count
Consensus length
Reference genome sequence
CT-ZA3
TAI-SP
SP
VT
CT14A
AT-1
T318A
T68-1
A18
Kpg3
NZ-B18
B165
SY
NZRB-G90
NZRB-M17
NZRB-TH28
HA18-9
CT11A
CTAF001623
NZRB-TH30
B301
TAI-M
T36
HA16-5
VT def
Mexico
NZRB-M12
Egypt
T385
T30


Sub-isolate 12-8
Read count
Consensus length
Reference genome sequence
CT-ZA3
VT
SP
T68-1
CT14A
B165
NZ-B18
AT-1
T318A
A18
SY
CT11A
Kpg3
TAI-SP
VTdef
TAI-M
CTAF001623
Egypt
NZRB-M17
NZRB-TH30
NZRB-TH28
NZRB-M12
NZRB-G90
T385
Mexico
HA18-9
B301
T30
T36
HA16-5


Sub-isolate 12-9
Read count
Consensus length
Reference genome sequence
CT-ZA3
VT
SP
T68-1
CT14A
B165
T318A
A18
NZ-B18
CT11A
AT-1
SY
CTAF001623
TAI-M
TAI-SP
VTdef
B301
Kpg3
Egypt
HA16-5
NZRB-G90
NZRB-TH30
NZRB-TH28
NZRB-M12
NZRB-M17
T36
T385
Mexico
HA18-9
T30


(A)
p1a
p1b
p33
p6
p65
p61
p27
p25
p18
p13
p29
p23
2,000
4,000
6,000
8,000
10,000
12,000
14,000
16,000
18,000
CT-ZA3
Consensus
Coverage
VT
TAI-SP
SP
(B)
B165
CT14A
A18
T68-1
(C)

◀**Fig. 2** Read mapping distribution for each GFMS12 sub-isolate. **A**) 12-7, **B**) 12-8, and **C**) 12-9. CTV open reading frames are shown at the top as yellow arrows with their respective gene product names. The scale below represents the full-genome length in nucleotides. The continuous, solid black lines represent the full-length reference genomes. Below each genotype name, the consensus sequences are represented by dashed black lines, where the gaps between them indicate a genome region that was not mapped by the input dataset. Under the regions where consensus sequences were obtained, the coverage achieved in any area is represented by a graph with numbers of reads represented by the scale on the left. Note that the scale of coverage values is not constant. Read maps of individual CTV reference genomes are delineated by lines

CL = 1559; AC = 0.83). Read mapping distribution against the whole genome of CTV references (Fig. 2c) confirmed that CT-ZA3 was almost completely covered, with only two minor gaps in both the 5′ and 3′ halves. Genotypes VT, SP, and A18 were mapped almost exclusively in their 3′ half, while T68-1 was mapped mostly in its 5′ half. Lastly, mapped B165 reads were scattered across the genome. Four contigs with lengths of 17,274, 345, 247 and 205 bp were shown to belong to CTV. Again, contigs were aligned with an arbitrarily selected CTV reference genome, which acted as guide for aligning contigs. The longest contig corresponded to an almost fully re-created genome (17,274 bp, 89.7 % of the length of the entire CTV genome using VT as a reference). The short contigs were able to fill most gaps, but a few still remained. Overall, contigs covered genome regions from nucleotide positions 226 to 17,803 (1a, 1b, p33, p6, p65, p61, p27, p25, p18 and p13). Only the last two genes, p20 and p23, as well as the 5′ and 3′ UTR, were not covered. We therefore determined a "partial" genome sequence, 17,407 nt long, named CT-ZA2 (GenBank accession number KC 333868). The dendrogram in Fig. 3 shows that CT-ZA2 clusters within the same clade as CT-ZA3 despite missing its terminal portions (5′ UTR, 3′ UTR, p20 and p23).

Thirty CTV contigs were assembled from GFMS 12 sub-isolate 12-7 reads, which reached a maximum length of 4715 bp. Contigs contained very few overlapping regions, making the a full-length genome construction impossible. When subjecting the contig set to mapping against the 30 reference CTV genomes which included CT-ZA3, CT-ZA3 was almost fully mapped by 20 contigs (Supplementary Figure 1a), none of which had overlapping regions, confirming the contig alignment results. The ten remaining contigs mapped to the Taiwan-PUM/SP/T1 genotype, mostly in the 1a region (Supplementary Figure 1b). Results from read assemblies against the reference genome set correlated with the contig mappings (Fig. 2, top graph). The highest metrics were obtained with the CT-ZA3 genome (CL = 14176; RC = 239212; AC = 1040.1), with read mapping across the whole genome observed. The second-best hit was Taiwan-PUM/SP/T1 (CL = 7317; RC = 14619; AC = 62.6), the genome of which was mapped mostly within its 5′ half, with a large number of reads clustered within ORF1a along with some mapping in the 3′ half as well (Fig. 2a). After discarding reads from these two references, 11,909 reads remained, distributed mostly amongst VT (9905 reads) and SP (820 reads). Although in relatively high abundance, reads that mapped specifically to VT covered a very limited area of its genome, mostly scattered in the 3′ half and with low coverage (39.12-fold). A visual representation of the distribution of mapped reads against the most frequently represented genotypes is shown in Fig. 2a.

Read datasets in this study yielded an average of 22.76 % CTV-specific reads for sub-isolates 12-8 and 12-9 and 2.65 % for sub-isolate 12-7. These percentages of CTV-specific reads were lower than expected when using a dsRNA extraction and enrichment protocol. This may be due to low titers of the virus (especially for 12-7), low active replication of virus in the tissue used, excessive host components retained after dsRNA extraction, or the use of non-optimal protocols. Sequences from the host and various microorganisms were observed during the selection of contigs following *de novo* analysis (not shown). Essentially, whole CTV genome sequences from two samples could be assembled *de novo* and were submitted to the GenBank database (accession numbers KC 333868 and KC 333869). This is the first report of a fully sequenced CTV isolate in South Africa, for which we propose the name CT-ZA3. Based on read counts, consensus lengths and average coverage, it is a member of the T68 strain group and is dominant in all of the GFMS12 sub-isolate sources tested.

The observed population structure for the sub-isolates tested partially supports and expands upon the results of a previous study [7] in which, based on sequencing multiple clones of amplicons of a segment of ORF1a (A-fragment) and the p23 gene, sub-isolate 12-7 was described as containing an B165/VT-recombinant and an RB/VT-like recombinant, 12-8 a B165/VT-like recombinant, and 12-9 a B165/VT-like recombinant and VT. It is most likely that the B165/VT-like recombinant proposed [7] within all three sub-isolates is actually CT-ZA3, which, like its most closely related isolate T68-1, has been shown by recombination analysis in this study (data not shown) and elsewhere [11] to share most of its 3′ half with VT, while its ORF1a A-fragment region would have been interpreted in the previous study [7] as a B165-like genotype member in the absence of the subsequently proposed T68 genotype [11]. The reported presence of a RB/VT-like recombinant (ORF1a/p23 gene) in sub-isolate 12-7 in addition to B165/VT [7] is not supported in the current study. In this study, a number of reads from 12-7 mapped to large portions of the 5′ half of Taiwan-PUM/SP/T1 (also an RB-like genotype)

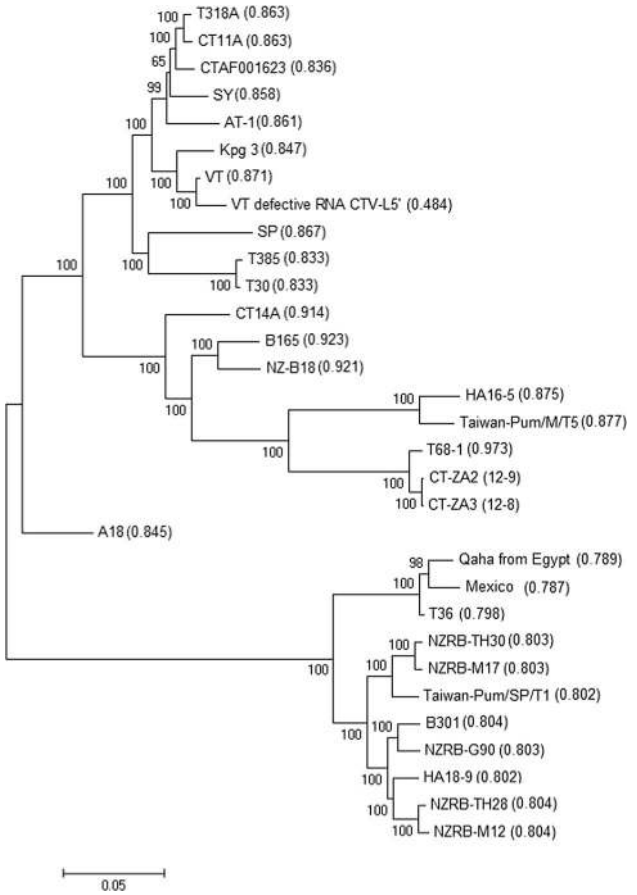


Fig. 3 Neighbor-joining dendrogram showing the phylogenetic relationship between full-length genomes of CT-ZA2/CT-ZA3 and reference CTV genomes. Confidence levels are shown as bootstrap values at each node. The numbers in brackets next to genotypes other than CT-ZA2 and CT-ZA3 are the percentage similarity value compared to CT-ZA3. Numbers in brackets next to CT-ZA2 and CT-ZA3 indicate the sub-isolates from which they were derived

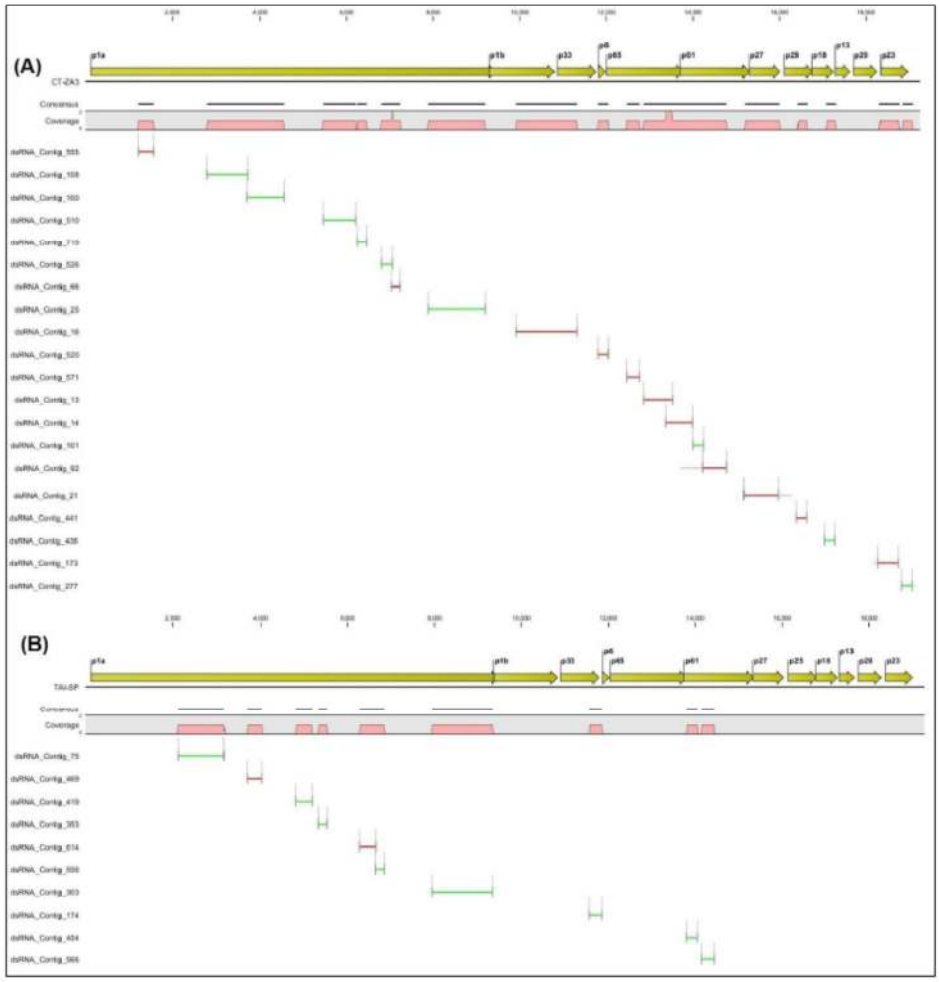
with at least one genome region in this half having high read coverage (c3450), along with some reads in its 3′ half. However, in a previous publication [8; Fig. 3a], based on direct sequencing of a portion of ORF1a, the isolate clustered with the NZRB-TH28 isolate, which occurs in a different RB-genotype clade to Taiwan-PUM/SP/T1; the initial test [7] could not be repeated subsequently. The genome region to which a large number of homologous reads mapped in Taiwan-PUM/SP/T1 corresponds to the ORF1a region commonly amplified in our laboratory [7] and may represent an amplicon contaminant in the initial dsRNA template. This possibility is supported by the inability to amplify Taiwan-PUM/SP/T1 amplicons from this plant using RB genotype-specific primers, as well as the lack of such reads in sub-isolate 12-7 when total RNA or immuno-captured viral particles were used as templates for Illumina sequencing (unpublished results). If the reads represent an ORF1a contaminant, however, it remains unclear why a number of other contigs homologous to Taiwan-PUM/SP/T1 were generated and aligned with other regions of the genome but not to other genotypes. Unique VT-like reads (9903 reads) were also found in the 3′ half. While these may suggest the presence of defective VT-like RNA in the population rather than the 5′ region of the putative "RB/VT-like" recombinant, long-read-length sequences (e.g., from Sanger or Pacific Bio platforms) are needed to create certainty. CTV stem pitting (SP)-specific reads were also present, distributed in the 3′ half of the genome but at low read counts, low coverage, and relatively low combined consensus length, and its relevance is unknown. Analysis of the genotype composition of the 12-7 population would have been aided by having obtained larger numbers of CTV-specific reads. Sub-isolate 12-8 was regarded as a homogeneous source of the "B165/VT-recombinant" in a previous study where multiple clones were sequenced [7]. In this study, in addition to CT-ZA3, very low numbers of reads (between 71 and 155) unique to VT, SP, T68-1, and CT14A isolates, occur in this sub-isolate. These reads mapped with low coverage even of specific genome areas but had combined consensus regions of between 2066 and 3769 to VT, SP, T68-1 and CT14A isolates. Whether these small numbers of reads represent very low levels of additional genotypes, multiple recombinants, defective RNA components, or quasispecies variations of specific genome regions could not be determined. This result also illustrates the inability of the multiple cloning strategy utilized previously by our group [7] (where only 60-100 clones could be analyzed in practice) to detect minor components of a population, relative to the analysis of Illumina reads, where average coverages over the whole genome of even minor components are often between 50 and 100 times, and where read coverages of specific genome areas can run into the thousands. In 12-9, in addition to the "B165/VT-like recombinant" component, an additional VT-like strain, based on both ORF1a and p23 gene sequences, was reported [7]. Based on read numbers and consensus lengths obtained, CT-ZA3 (the "B165/VT-like" genotype) is clearly dominant within this population; however, only 3′ half VT-specific reads were observed in this study, not supporting the presence of a full-length VT-like genotype found previously [7]. However, based on the NGS reads, it appears as though additional genotypes or recombinants are also present in this sub-isolate, with unique reads distributed in the 3′ half of the genome, which are SP-like and A18-like. Also, unique reads that are T68-1-like, CT14A-like or B165-like were observed distributed at various places along the entire genome length. Once again, while uncertainty exists as to what these additional reads represent, it is apparent that the

genotype composition of the CTV population in this sub-isolate is more complex than previously thought.

Acknowledgments We thank Citrus Research International (CRI) (South Africa) for the financial support of this study. We also wish to thank Prof. Fourie Joubert, Bioinformatics Unit, University of Pretoria, and Dr. Jasper Reece, ARC-Biotechnology Platform, for their bioinformatics advice. We want to thank Kirsti Snyders for doing a number of genotype-specific PCRs and Sanger sequencing to support Illumina results. We also wish to thank Dr. Fanie van Vuuren, CRI, for initially providing the GFMS12 sub-isolates studied.

References

1. Bar-Joseph M, Marcus R, Lee RF (1989) The continuous challenge of *Citrus tristeza* virus control. Annu Rev Phyopathol 27:291–316
2. Bar-Joseph M, Garsney SM, Gonsalves D, Moscovitz M, Purcifull DE, Clark MF, Loebenstein G (1979) The use of enzyme-linked immunosorbent assay for the detection of *Citrus tristeza* virus. Phytopathology 69:190–194
3. Karasev AV, Boyko VP, Gowda S, Nikolaeva OV, Hilf ME, Koonin EV, Niblett CL, Cline K, Gumpf DJ, Lee RF, Garsney SM, Lewandowski DJ, Dawson WO (1995) Complete sequence of the *Citrus tristeza* virus RNA genome. Virology 208:511–520
4. Van Vuuren SP, Collins RP, Da Graca JV (1993) Growth and production of lime trees pre-immunized with different mild *Citrus tristeza* virus isolates in the presence of natural disease conditions. Phytophylactica 25:49–52
5. Folimonova SY, Roberston CJ, Shilts T, Folimonov AS, Hilf ME, Garsney SM, Dawson WO (2010) Infection with strains of *Citrus tristeza* virus does not exclude superinfection by other strains of the virus. J Virol 84(3):1314–1325
6. Folimonova SY (2013) Developing an understanding of cross-protection by *Citrus tristeza* virus. Front Microbiol 4:1–9
7. Scott KA, Hlela Q, Zablocki O, Read D, van Vuuren S, Pietersen G (2012) Genotype composition of populations of grapefruit-cross-protecting *Citrus tristeza* virus strain GFMS12 in different host plants and aphid-transmitted sub-isolates. Arch Virol 158:27–37
8. Roossinck MJ, Saha P, Wiley GB, Quan J, White JD, Lai H, Chavarria F, Shen G, Roe BA (2010) Ecogenomics: using massively parallel pyrosequencing to understand virus ecology. Mol Ecol 19:81–88
9. Van Vuuren, SP, Van der Vyver JB, Luttig M (2000) Diversity among sub-isolates of cross-protecting *Citrus tristeza* virus isolates in South Africa. In: Proceedings 14th Conf. Intl. Org. Citrus Virol. Moreno P, da Graça JV, Timmer LW (eds) University of California, Riverside, pp 103–110
10. Morris TJ, Dodds JA, Hillman B, Jordan RL, Lommel SA, Tamaki SJ (1983) Viral specific dsRNA: diagnostic value for plant virus disease identification. Plant Mol Biol Rep 1:27–30
11. Harper SJ (2013) *Citrus tristeza* virus: evolution of complex and varied genotypic groups. Front Microbiol 4:1–18



Supplementary Figure 1 Contig mappings against A) CT-ZA3 and B) Taiwan-PUM/SP/T1 genomes in GFMS12 sub-isolate 12-7. Above each map, CTV ORFs are displayed as yellow arrows above which a nucleotide scale represents the full-length of a typical CTV genome. The mapping distribution is displayed as dashed black lines. The contig names are displayed on the left and are all represented as green (forward strand) or red bars (reverse strand) that have mapped to a reference genomic region (either CT-ZA3 or Taiwan-PUM/SP/T1). The associated coverage scale and graph for each contig are displayed below the reference genomes (continuous black line).