

PCR bias associated with conserved primer binding sites, used to determine genotype diversity within *Citrus tristeza virus* populations

David Alan Read¹ and Gerhard Pietersen^{1, 2*}

¹Department of Microbiology and Plant Pathology, Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Pretoria 0002, South Africa. david.read@fabi.up.ac.za

²Agricultural Research Council—Plant Protection Research Institute, Pretoria 0002, South Africa. gerhard.pietersen@up.ac.za

* - Corresponding author.

Highlights

- The PCR amplification bias of a number of primers that target CTV genes was determined.
- The value of using a next generation sequencing platform for viral population studies is shown.
- Primers that are used for determining viral population diversities should be under constant revision to avoid preferential amplification of target sequences.

ABSTRACT

Citrus tristeza virus (CTV) is present in almost all of the major citrus production areas where it continues to reduce the profitability of citriculture. The accurate characterisation of CTV populations, which are usually made up of a number of disparate strains, requires the use of robust PCR protocols. Mismatches between primers and their corresponding binding sites may introduce primer-associated bias during amplification. The primer-associated bias of four sets of CTV specific primers, targeting the A and F regions and the p33 and p23 genes, were evaluated. This was done through the amplification of defined templates followed by their characterisation using the sequencing of multiple clones, as well as Illumina next generation sequencing. High levels of bias were found to be associated with the primer pairs

targeting the A and F regions. The p33 gene primers were found to be biased against two genotypes and suggestions for preventing this apparent bias are discussed. The primer pair targeting the conserved p23 gene was found to have very little associated bias. Primers should undergo rigorous screening before being used to characterize virus populations that are known to exhibit high levels of variation, especially within primer binding sites.

Keywords: PCR bias; viral populations; *Citrus tristeza virus*;

1. INTRODUCTION

Polymerase chain reaction (PCR) based amplification and the subsequent sequencing of genes have been valuable tools for determining the diversity and taxonomic relationships of populations of fungi, bacteria and viruses. Polyspecific primers are usually designed to amplify multiple templates present in samples, often including a variety of organisms (Farris and Olsen, 2007). It is critical to design polyspecific primers that are specific enough to target only the taxonomic groups of interest but at the same time not biasing the PCR reaction toward only certain components of a population (Anderson et al, 2003). In order to detect multiple components of viral populations, primers are designed to target conserved primer binding sites. These conserved sites can only be targeted by comparing known sequences of a specific virus or group of viruses. The efficacy of the resulting primers will be based on how well they correspond to the conserved sites of both the known and unknown sequences (Zheng et al, 2008). As the pool of known sequence information increases the conserved nature of target sites begins to degrade through process termed “consensus decay” (Zheng et al, 2008). Continued use of primers that no longer represent conserved sites may result in the preferential amplification of sequences from populations.

Citrus tristeza virus (CTV) is a *Closterovirus*, whose virions are flexuous, thread-like (Karasev et al, 1995) and contain a positive sense genome of ~20kb (Bar-Joseph et al, 1985). CTV is the most severe viral pathogen of citrus (Ghorbel et al, 2001) and has become well established within almost all of the major citrus production areas worldwide, along with its aphid vector species (Moreno et al, 2008).

There is considerable sequence variation between strains and little agreement regarding their classification. However, based on extensive phylogenetic and evolutionary analyses of extant CTV genomes, Harper, (2013) proposed that the virus exists as a complex of strains based on six distinct lineages (Harper, 2013) called genotypes, that usually exist as mixtures within a single infection (Rubio et al, 2001). In areas where CTV is endemic the economically productive lifespan of trees in the field is increased by inoculating budwood with mild-strain populations of CTV (Lee and Keremane, 2013) in a process known as cross-protection, intended to prevent secondary infections through aphid inoculations, of potentially severe strains of the virus (Gal-On and Shibolet, 2005). Folimonova et al, 2010 showed that CTV cross-protection functions via a genotype specific mechanism. Previously, cross-protecting sources were characterised on an empirical basis through grafting onto indicator hosts without knowledge of the genotypes present (Moreno et al., 2008). Since CTV cross-protection is genotype-specific, it is essential to determine which CTV genotypes are circulating within a given production area, as well as the genotype composition of candidate cross-protection sources.

A number of gene regions have been used previously to characterise CTV populations. These include the A-region and the F-region (Rubio et al, 2001), the p23 gene (Sambade et al, 2003), the p27 gene (Iglesias et al, 2008) and the coat protein gene (CP) (Sekiya et al, 1991).

In this study, the potential primer-directed bias of primer sets targeting four different genomic regions of various CTV, previously used in our laboratory, was evaluated. A set of genotype-specific primers were designed for each gene region and then used to amplify plasmid CTV inserts of known identity. Mixtures containing equimolar concentrations of each genotype-true template were produced and then re-amplified, using the polyspecific primers targeting the consensus sites. The resulting amplicons were then subjected to both Sanger sequencing of multiple clones and Illumina MiSeq sequencing.

2. MATERIALS AND METHODS

2.1. Gene selection and primer design

Four CTV gene regions were selected for the PCR bias experiment. The A-region and F-region primer binding sites (Rubio et al, 2001) are located toward the 5' end of

the genome in ORF1a. The primer-binding sites associated with the p33 and p23 (Sambade et al, 2003) genes are located near the midpoint and near the 3' end of the genome respectively. It should be noted that the absence of any variability within the p23 gene forward site, does not reflect reality and was used in this study to determine possible experimental variability.

The published primer sequences, used to amplify these gene regions, were aligned against reference sequences, using BioEdit Sequence alignment editor 7.1.3 (Hall, 1999). Forty six whole genome reference sequences were obtained from GenBank and are listed as follows: NC_001661 (T36); AY 340974 (Qaha); U16304 (T36); DQ272579 (Mexico); AY170468 (T36); EU937521 (T36); KC517485 (FS674-T36); KC517486 (FS701-T36); KC517487 (FS703-T36); KC517488 (FS577); JX266713 (Taiwan-Pum/M/T5); AF001623 (SY568); AF260651 (T30); Y18420 (T385); KC517489 (FS701-T30); KC517490 (FL278-T30); KC517491 (FS703-T30); JF957196 (B301); FJ525432 (NZRB-G90); GQ454869 (HA 18-9); FJ525435 (NZRB-M17); JX266712 (Taiwan-Pum/SP/T1); FJ525431 (NZRB-M12); FJ525433 (NZRB-TH28); FJ525434 (NZRB-TH30); JQ798289 (A18); KC525952 (T3); HM573451 (Kpg3); EU857538 (SP); GQ454870 (HA 16-5); DQ151548 (T318A); AB0463981 (NUagA); JQ911664 (CT11A); KC517493 (FL202-VT); U56902 (VT); KC517492 (FS703-VT); EU937519 (VT); KC517494 (FS701-VT); KC262793 (L192GR); JQ911663 (CT14A); FJ525436 (NZ-B18); JQ965169 (T68); EU076703 (B165); JQ061137 (AT-1); KC333868 (12-8); KC333869 (12-9).

A total of 7 A-region genotype-true forward primers and 9 reverse primers were used to incorporate genotype-true primer binding sites into amplicons that were generated through the amplification of plasmid clones representing the following genotypes VT, T3, T30, HA 16-5, RB1, RB2, B165 and T36. Six forward and six reverse F-region genotype-true primers were used to target the inserts of plasmid clones representing the T30, VT, B165, HA 16-5, RB1 and T36 genotypes. The p33 gene genotype-true forward primer binding sites of the RB genotype was incorporated into the amplicons representing Taiwan-Pum/SP/T1 and RB, while the CTZA true forward primer binding site was incorporated into an amplicons representing CTZA. The original p33 reverse primer was used in all amplification steps. The sequences of the primers are listed in *Table 1*.

Table 1: Genotype specific primer sequences used to amplify the insert of each specific plasmid to generate genotype true amplicons. Bases that differ from the published generic CTV primers are shown in bold text, with the last column indicating the total number of differences between genotype specific primers and the generic primers. ^aBased on the primer binding positions on the sequence of T385 (Y18420). ^bBased on the primer binding positions on the sequence of T36 (U16304)

Gene (location on genome)	Primer orientation	Primer name	Strain represented	Primer sequence (5' to 3')	Number of differences
A-region (2021 – 2548bp) ^a	Forward	A-F standard (Rubio et al, 2001)	Consensus /VT	AGCTGTTCTGAAACGCGG	-
		A-F T3	T3	ACGTGTT T GTGAAACGCGG	1
		A-F T30	T30	ACGTGTTCTGAAACG T GG	1
		A-R HA16-5	HA16-5	ACGTGTT TA TGAAACGCGG	2
		A-F RB	RB1 and RB2	ACGTGTT T GTGAA G CG T GG	3
		A-F B165	B165	ACGTGTT TA C GAAACG T GG	4
	A-F T36	T36	ACGTGTT T GT A AA G CG T GG	4	
	Reverse	A-R standard (Rubio et al, 2001)	Consensus/T385	GTCGATAACTCGACAAACGAGC	-
		A-R VT	VT	G C CGATAACTCGACAAACGAGC	1
		A-R T30	T30	GTCGATAACTCGACA G ACGAGC	1
		A-R RB2	RB-TH30	GTC TG TAACTCGACAAACGAGC	2
		A-R T3	T3	G C CGATAACTCGA T AAACGAGC	2
		A-R RB1	RB-TH28	G C C TG TAACTCGACAAACGAGC	3
		A-R T36	T36	GTC TG TAAC C CGACAAACGAGC	3
		A-R B165	B165	AG C TG TAACTCGACAGACGAGC	4
A-R HA16-5		HA16-5	CC C TG TA G CTCGACA G ACG G GC	7	
F-region (3561 – 3998bp) ^a	Forward	F-F standard (Rubio et al, 2001)	Consensus	GTGTTATCATGCGTCTGAAGCG	-
		F-F T30	T30	GTGTT T TCATGCGTCTGAAGCG	1
		F-F VT	VT	GTGTT T TCATGCGTCTGA T GCG	2
		F-F B165	B165	A TGTT T TCATGCGTCTGA T GCG	3
		F-F HA16-5	HA16-5	GTGTT T TCATGCGTC C GA T GCG	3
		F-F RB-TH28	RB-TH28	GTGTTATCATG A GTC G GA G GCG	3
		F-F T36	T36	GTGTTATCATGC A TC G GA G GCG	3
	Reverse	F-R standard (Rubio et al, 2001)	Consensus	GGAATCTTAATCCTAATCAAG	-
		F-R T30	T30	GGAATCTTAATCCTAATCAAG	0
		F-R RB-TH28	RB-TH28	GGAATCTT G AT T CCTAATCAAG	2
		F-R B165	B165	GGAATCTT G ATCCT G ATCAAG	2
		F-R VT	VT	GGAATCTT G ATCCT G ATCAAG	2
		F-R T36	T36	GGAAT T TT G AT T CCTAATCAAG	3
		F-R HA16-5	HA16-5	GG G ATCTTAAT T CCTAAT T AA A	4
		p33 gene (10905 – 11882bp) ^b	Forward	Univ p33 F	Kpg3 HA 16-5 T36 CT-3719 AT-1
p33-F-RB	Taiwan-Pum/SP/T1 RB			GATGTTTGC T TTCGCGAGC	1
P33-F-CTZA	CTZA			GATGTTTGCCTTCGCGAG T	1
Reverse	Univ p33 R		Kpg3 HA 16-5 T36 CT-3719 AT-1 Taiwan-Pum/SP/T1 RB CTZA	CCCCTTTAAACAGAGTCAAACGG	-
p23 gene (18347 – 18365bp) ^b	Forward/ Reverse		PM50/PM51 (Sambade et al, 2003)	T36 T30 SP Kpg3 RB1	ACTAACTTTAATTCGAACA / AACTTATTCGCTCCACTTC

2.2. Plasmid insert amplification

The inserts of appropriate plasmids, with sequences that could be differentiated phylogenetically, were amplified with the sets of the genotype-true primers (*Table 1*), by adding 1 µl of each recombinant plasmid to a reaction mix containing: 5 µl of 1x PCR reaction buffer, 2.5 mM MgCl₂, 2.5 U DNA polymerase (Bioline, London, United Kingdom), 0.14 mM dNTP mix, 2 µM forward primer, 2 µM reverse primer and PCR grade water to a final reaction volume of 50 µl, using the following PCR conditions: 1 cycle of 92°C for 2 min. 35 cycles of 92°C for 30 s, 55°C for 45 s and 72°C for 1 min. 1 cycle of 72°C for 10 min. The concentrations of the PCR products were determined using the Quant-IT™ dsDNA BR Assay Kit in conjunction with the Qubit fluorimeter (Life Technologies, Grand Island, NY, United States). Equimolar concentrations of each genotype true template were pooled to produce a single mixture of genotype templates for each gene region. These templates were amplified with standard A, F, p33 and p23 gene primers, by adding 5 µl of template to each reaction mix and the same PCR reaction conditions used for the genotype specific primer amplifications. The products were purified using the Wizard® SV Gel and PCR Clean-Up System (Promega, Madison, WI, United States) and quantified using a NanoDrop 2000 spectrophotometer (Thermo Scientific, Wilmington, DE, United States).

2.3. Cloning of amplified products

Quantified amplicons were cloned using the pGEM-T Easy vector (Promega, Madison, WI, United States) according to the manufacturer's specifications and used to transform competent *E. coli* JM109 cells. Putative recombinants were selected using blue/white selection and plasmid extractions were performed using alkaline lysis.

Plasmid inserts were amplified using the T7 (5'- TAA TAC GAC TCA CTA TAG GG-3') and SP6 (5'- ATT TAG GTG ACA CTA TAG AA -3') vector specific primers and PCR conditions described earlier. Amplicons displaying the correct amplicon size were purified by adding 2 µl of FastAP and 0.5 µl ExoI enzymes (Thermo Scientific, Vilnius, Lithuania) to 19 µl of amplified PCR product and sequenced in one direction with T7 primer. Sequencing reactions contained: 1 µl BigDye® Terminator mix v3.1 (Applied Biosystems, Foster City, CA, United States), 2.25 µl 5x BigDye® v3.1 sequencing buffer (Applied Biosystems, Foster City, CA, United States), 0.75 µl

T7 primer (2 μ M), and molecular grade water to a final reaction volume of 10 μ l. Sequencing reaction conditions were: 94°C for 1 minute, 25 cycles of 94°C for 10 seconds, 50°C for 5 seconds, 60°C for 4 minutes.

2.4. Illumina MiSeq sequencing of amplified products

Next generation sequencing (NGS), using the Illumina MiSeq platform (Illumina, San Diego, CA, United States), was carried out in parallel for each amplicon that was produced from the equimolar templates. Paired-end DNA libraries were prepared using the Nextera V2 sample kit (Epicentre, Madison, WI, United States) and run on 1/8th of a lane on an Illumina MiSeq flow cell. The samples were sequenced at the Agricultural Research Council (ARC), Biotechnology Platform, Pretoria, South Africa.

2.5. Data analysis

Sanger sequencing chromatograms were viewed and corrected using Chromas Lite 2.1 (Technelysium, Brisbane, Australia). Alignments were performed with CLUSTAL W alignment software (EBI, Cambridgeshire, England) within the BioEdit Sequence alignment editor 7.1.3 program (Hall, 1999). Phylogenetic dendrograms were produced using the MEGA 4.1 software package (Tamura et al, 2007), as neighbour-joining trees with a bootstrap value of 1000.

The NGS data was analyzed using the CLC Genomics Workbench 5.5.1 (CLC bio, Aarhus, Denmark). Data was imported as paired-end reads with a distance range of 180-250. Adapter and quality trimming was done using the Fast QC function settings with Nextera V2 transposase adapter sequences (Transposase 1: GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG; Transposase 2: TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG). Assemblies were performed on the trimmed datasets, using the sequences of their respective clones as references. The reference assemblies were carried out using settings. Length fraction: 0.9; similarity fraction: 0.9; global alignment: off; non-specific match handling: map randomly.

3. RESULTS

3.1. *Multiple clone sequence data*

The results of sequencing multiple clones or performing Illumina sequencing from each amplification reaction are listed in *Table 2*. The clone sequence data for the A-region and p23 gene were previously published by Read and Pietersen (2015). Six genotype-specific forward and reverse primers were used for the F-region primer bias experiment, in which a total of 156 clone sequences were recovered.

A total of three genotype-true p33 gene forward primers and only one reverse primer were used to amplify the cloned inserts of plasmids representing the genotypes listed in *Table 1* and *2*. The Univ-p33-F forward primer amplified cloned inserts of Kpg3, HA 16-5, T36, CT-3719 and AT-1, while the p33-F-RB primer amplified the cloned inserts of Taiwan-Pum/SP/T1 and RB and the cloned insert of CTZA was amplified using the p33-F-CTZA forward primer. A cloned sequence, putatively termed CT-3719, was obtained from a sample from a South African Star Ruby tree (unpublished data) and forms a unique phylogenetic branch (supported by bootstrapping) that did not cluster with any of the currently available reference sequences. The numbers of clones obtained for the p33 gene suggest that a significant primer-directed bias existed for the primer pair described here. This bias was directed against the Taiwan-Pum/SP/T1, RB and CTZA sequence types, which were all represented at levels less than expected. This coincides with the single nucleotide difference near the mid-point of the forward primer binding site of the Taiwan-Pum/SP/T1 and RB and the single nucleotide difference at the 3' position of the CTZA primer.

Table 2: Cloned sequence and Illumina MiSeq data showing the primer-directed bias, associated with the A and F region, p33 and p23 gene PCR amplification protocols. The Illumina MiSeq data for the A-region amplicon is an amalgamation of two separate sequencing runs based on the same template. The first column shows the particular primer pairs that were analysed. Column 4 and 7 show the number of sequenced clone inserts and Illumina MiSeq sequencing reads respectively that correspond back to their particular sequence type. Data is ranked by diminishing number of either cloned sequences that clustered with their particular reference sequence or the number of Illumina MiSeq reads that mapped back to particular references (columns 3 and 6).

Region/gene	Clone sequence data			MiSeq sequence data		
	Number of differences within F and R primers	Strain	Nr. of seqs grouping with ref clone (% of total number of recombinants)	Number of differences within F and R primers	Strain	Nr. Of reads mapping to reference (% of total reads)
A-region	3	T3	44 (34)	3	T3	41043 (32)
	1	VT	44 (34)	1	VT	31369 (24)
	8	B165	19 (15)	8	B165	22919 (18)
	2	T30	8 (6)	2	T30	18831 (15)
	5	HA 16-5	7 (5)	6	RB-TH28	6800 (5)
	7	RB-TH30	6 (4)	7	RB-TH30	3967 (3)
	6	RB-TH28	3 (2)	9	T36	2656 (2)
	9	T36	0 (0)	5	HA16-5	604 (1)
F-region	1	T30	134 (86)	1	T30	402861 (54)
	4	VT	11 (7)	4	VT	239504 (32)
	7	HA16-5	6 (4)	7	HA16-5	49839 (7)
	5	B165	4 (2)	5	B165	30305 (4)
	6	RB-TH28	1 (1)	5	T36	13218 (2)
	5	T36	0 (0)	6	RB-TH28	9360 (1)
p33 gene	-	Kpg3	18 (26)	-	AT-1	195802 (24)
	-	HA 16-5	14 (21)	-	CT-3719	172031 (21)
	-	AT-1	14 (21)	-	T36	166143 (20)
	-	CT-3719	12 (18)	-	Kpg3	113290 (14)
	-	T36	6 (9)	-	HA 16-5	100392(12)
	1	Taiwan-Pum/SP/T1	2 (3)	1	CTZA	40167 (5)
	1	RB	1 (2)	1	RB	14129 (2)
	1	CTZA	0	1	Taiwan-Pum/SP/T1	12140 (2)
p23 gene	-	T36	41 (35)	-	T30	32210 (23)
	-	SP	21 (18)	-	T36	30100 (22)
	-	Kpg3	19 (16)	-	RB-TH28	26854 (19)
	-	RB-TH28	19 (16)	-	SP	25486 (18)
	-	T30	18 (15)	-	Kpg3	25138 (18)

3.2. *Illumina MiSeq data*

The number of reads mapping back to each specific reference sequence can be seen in *Table 2*. A total of 128 459 reads mapped back to the A-region reference sequences. The rank order among genotypes by MiSeq read numbers was the same as that of the multiple clone sequence data for the first 5 genotypes with differences in the rank order of reads mapped and clones obtained were observed for the last three genotypes of the A-region.

A total of 754 723 reads mapped back to the F-region reference sequences. The ranking of genotypes based on reads mapping back to their respective reference sequences, as well as by numbers of genotype specific clones obtained yielded minor differences in the apparent rank order of genotypes.

The forward and the reverse primer binding site sequences used in the p23 gene analysis have identical nucleotides amongst the five genotypes tested. MiSeq read mapping for the p23 gene amplicon sequencing suggests that this region lacks primer-directed PCR bias.

4. DISCUSSION

In this study, the usefulness and potential bias of four CTV-specific primer sets used in our laboratory for population studies was determined by amplifying target sequences from an artificially assembled heterogeneous genotype template through cloning of resultant amplicons and Sanger sequencing of multiple clones and comparing the data obtained with the abundance of genotype-specific reads obtained through Illumina MiSeq sequencing of the same amplicons. Primers targeting the; 1) A-, and 2) F fragments of ORF1a (Rubio et al, 2001), 3) p23 gene (Sambade et al, 2003), and 4) p33 gene (this study). The results suggest that all of the primer pairs, with the exception the one targeting the p23 gene, show significant bias toward the amplification of a number of genotype-specific sequences within a defined template.

PCR primers can be designed to target at the domain, genus or species level of various organisms (Del Portillo et al, 1991; van Kuppeveld et al, 1992; Watanabe et al, 2001). The recent advances in next generation sequencing has allowed for a massively increased output of data, allowing analysis of microbial communities at an unprecedented depth (Waud et al, 2014). This has increased the demand on primers

to be able to amplify molecular markers in an unbiased fashion, in order to accurately represent the microbial population diversity (Logan et al, 2014). Primers generally used for microbial community analyses are often not evaluated in terms of their specificity or ability to accurately amplify target sequences (Waud et al, 2014).

Viral populations (quasi-species or mixtures of strains or genotypes) can be characterised through PCR, using primers specific to highly conserved primer binding sites that flank variable regions that are able to differentiate sequence variants and genotypes. Amplicons are then usually cloned followed by the sequencing of multiple inserts. The number of sequenced inserts sequenced per population varies but twenty has typically been deemed sufficient (Mullan et al, 2004). However, the ability of the technique to detect minor components of the viral population is directly related to the number of clones sequenced and because of the labour intensive nature of the task, it is seldom done with sufficient clones to achieve high enough resolution of the population structure (Waud et al, 2014). In this study between 67 and 156 cloned sequences were analysed for each of the primer pairs, which is considerably more than that usually used in viral quasi-species analysis (Mullan et al, 2004). With the advent of next generation sequencing (NGS) technologies, parallel sequencing of millions of sequences in parallel without cloning is possible (van Dijk et al, 2014) and should provide a more accurate depiction of viral population structures.

While multiple factors, such as the choice of polymerases (Mullan et al, 2001), could potentially contribute to bias within population studies, PCR primer-associated bias is considered to be one of the most significant, especially during the amplification of mixed templates (Ihrmark et al, 2012). Many population diversity studies have made use of primers from previous experiments that are sometimes combined arbitrarily without any considerations of the effects that potential bias may have in skewing the population data (Waud et al, 2014). Three of the CTV specific primer sets targeting conserved sites, namely the A and F (Rubio et al, 2001) and the p23 gene specific primers (Sambade et al, 2003), were designed more than a decade ago. It is therefore reasonable to assume that significant “consensus decay” (Zheng et al, 2008) has occurred due to the sequencing of additional CTV genomes. The A and F primers were designed based on five reference genomes. When considering the additional 41 genomes (used as references in this study), an additional seven and nine nucleotide differences are collectively added to the binding

sites of the A and F regions respectively (data not shown). Zheng et al, (2008) quantified the potential for the consensus decay of conserved sites among members of *Potyvirus*. However, the loss of consensus among primer binding sites is often not considered when estimating viral population diversity.

Degenerate primer design is one approach to achieving more universal detection of a targeted virus population or group of viruses. This approach has been used to target diverse populations of plant viruses (Elbeaino et al, 2013; Ramos-González et al, 2016; Varanda et al, 2014). However, most studies using degenerate primers to detect plant viruses are purely qualitative and degeneracies may not cover all possible variations within primer binding sites. Inosine bases can be used to add polyvalent degeneracies in primers targeting potentially variable positions within conserved sites (Foissac et al, 2005; Teycheney et al, 2007), although this approach appears to have had limited use in the field of plant virology. An additional disadvantage to using both degenerate primers and polyvalent inosine-containing primers is the reduction in sensitivity of the PCR (Nam et al, 2016; Teycheney et al, 2007).

The reliable detection and identification of CTV genotypes circulating within citrus production areas and in candidate cross protecting sources are essential prerequisites for successful mild-strain cross protection of citrus (Folimonova, 2013). The genotype-specific nature of CTV cross-protection (Folimonova et al., 2010), requires pre-immunization of mild-protecting sources of the same genotype to sustainably protect citrus trees in areas where severe CTV strains occur endemically (Folimonova, 2013). Information regarding the relative abundance of each genotype within the viral population within a host will yield important information as to genotype selection occurring and competition amongst genotypes. It is therefore important to, not only determine the genotypes present within the CTV population, as achieved by marker analysis (Hilf & Garnsey, 2000; Roy et al., 2005) but also to determine the abundance the known as well as unknown genotypes (Waud et al, 2014). The results of this study confirms that an increased number of nucleotide mismatches between a primer binding site and its conspecific primer leads to decreased ability of primers to amplify the template, especially within a multi-template PCR reaction (Ibarbalz et al, 2014). A number of exceptions however were observed; in the A-region analysis, the B165 sequence was represented at close to the expected percentage of 14% of the total clone sequences despite the primer binding sites

having a combined eight nucleotide mismatches with the forward and reverse primers. The Illumina MiSeq data corroborates the unexpectedly high representation of B165. The apparent lack of bias against this genotype may be due to the fact that only one of the eight nucleotide mismatches occurred within the critical 5 bp of the 3' end of these primers, while the others are likely to have only moderate effects on amplification efficiencies (Lefever et al, 2013; Waud et al, 2014). This was further supported by the data obtained to the F-region, where all of the forward primers, except that specific to T30, contained a nucleotide that differed from the consensus, 4 bp from the 3' end of each primer, probably resulting in the preferential amplification of T30.

The under-representation of T36 in the p33 gene clone sequence dataset was most likely due to chance due to the relatively low total number of clones that were sampled. The greater number of sequences (reads) analysed with the Illumina MiSeq datasets probably provides a much more accurate representation of the sequences within a PCR amplification product. However, a range of 12-24% of reads observed from five genotypes with identical p33 gene primer binding sites is probably due to minor errors in preparing the artificial template. At the same time, the observed under-representation of CTZA sequences in amplicons of the p33 gene primer pair can be attributed to the single nucleotide difference of the genotype at the terminal 3' binding position to that of the forward primer. Ihrmark et al, (2012) showed that a mismatch up to 14 bp away from the 3' terminus of a binding primer can reduce amplicon abundance by up to two orders of magnitude when amplifying from a heterogeneous template. Therefore, the single mismatch at the midpoint of the Taiwan-Pum/SP/T1 and RB forward primer binding sites could have been responsible for the reduction in amplification efficiencies of these templates. The generally accepted tolerances for primer mismatches during amplification (Lefever et al, 2013) are clearly complex and should be tested empirically.

Unexpected distortions in the levels of genotypes obtained from artificial templates could be reduced in future studies of this kind, through the use of technical replicates. This should allow for the normalisation of errors introduced during the amplification of the artificial template, as well as random PCR distortion (Ihrmark et al, 2012). However, in primer validation studies such as these, the practice of performing technical replicates should be taken a step back to the point of producing the artificial templates, where the components of the artificial template can be

quantified and then assembled multiple times and the resulting replicates pooled in order to normalise the effect that individual errors may have during the production of an artificial heterogeneous template.

Due to the increased demands on primers to accurately represent microbial population structures, the current focus should be the production of reliable PCR primers able to amplify genetic markers with minimal bias (Ihrmark et al, 2012). While most authors acknowledge that the goal of attaining a totally bias free protocol for community characterisation might not be possible, Ibarbalz et al, (2014) validates the use of amplicon sequencing by suggesting that while significant biases may exist during amplification, confidence can be placed in the relative abundances of members from samples that have been subject to identical biases during their analyses.

This study represents a significant step forward for the improvement of the quantification of components within CTV populations. While the primer pair targeting the p23 gene appears to have less associated bias than any of the others, the p23 genes location within the 3' half of the genome means that it has inherently less capacity for resolving genotypes when compared with regions located within the variable 5' half of the genome. The primer pair targeting the p33 gene, proposed for routine use in this study, has significantly reduced associated bias, when compared with primers targeting regions closer to the 5' end and the ability to resolve the major CTV genotypes at a phylogenetic level, for use in the reliable assessment of CTV population. This study should also bring the potential effects that unverified primer pairs can have on population datasets, to the attention of plant virologists and to the field of virology as a whole.

We show the value of using Illumina MiSeq for virus population composition determination. Significant levels of bias were observed for both cloned sequences and MiSeq data in all regions except that of the p23 gene, which did not show significant levels of bias within the MiSeq dataset. These results suggest that great care should be taken when designing polyspecific primers targeting potentially heterogeneous binding sites, not only for CTV population composition determination but for any organism. Primers used for the routine amplification and subsequent estimation of population compositions should be under constant revision as new sequences of variants become available to ensure that primer-directed selection does not cause a significant bias during the PCR amplification of sequences.

5. ACKNOWLEDGEMENTS

We wish to acknowledge Citrus Research International (CRI) and the NRF THRIP for funding and Miss Glynnis Cook, CRI, for pointing out that the possibility that amplification bias may exist in existing primers.

6. REFERENCES

Anderson I.C., Campbell C.D. and Prosser J.I. (2003). Potential bias of fungal 18S rDNA and internal transcribed spacer polymerase chain reaction primers for estimating fungal biodiversity in soil. *Environ. Microbiol.* 5, 36-47

Bar-Joseph M., Gumpf D.J., Dodds J.A., Rosner A. and Ginzberg I. (1985). A Simple Purification Method for Citrus Tristeza Virus and Estimation of its Genome Size. *Phytopathology.* 75, 195-198

del Portillo P., Murillo L.A. and Patarroyo M.E. (1991). Amplification of a Species-Specific DNA Fragment of *Mycobacterium tuberculosis* and Its Possible Use in Diagnosis. *J Clin Microbiol.* 29, 2163-2168

Elbeaino T., Whitfield A., Sharma M. and Digiario M. (2013). Emaravirus-specific degenerate PCR primers allowed the identification of partial RNA-dependent RNA polymerase sequences of Maize red stripe virus and Pigeonpea sterility mosaic virus. *J Virol Methods.* 188, 37-40

Farris M.H. and Olsen J.B. (2007). Detection of Actinobacteria cultivated from environmental samples reveals bias in universal primers. *Lett Appl Microbiol.* 45, 376-381

Foissac X., Svanella-Dumas L., Gentit P., Dulucq M.J., Marais A. and Candresse T. (2005). Polyvalent Degenerate Oligonucleotides Reverse Transcription-Polymerase Chain Reaction: A Polyvalent Detection and Characterization Tool for Trichoviruses, Capilloviruses, and Foveaviruses. *Phytopathology.* 95, 617-625

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

Folimonova S.Y. (2013). Developing an understanding of cross-protection by Citrus tristeza virus. *Front Microbiol.* 4, 76

Gal-On A. and Shibolet Y.M. (2005). Cross-protection, pp261-288. In G. Loebenstein and J.P. Carr (ed.), *Natural resistance mechanisms of plants to viruses*. Springer, Dordrecht, the Netherlands

Ghorbel R., López C., Fagoaga C., Moreno P., Navarro L., Flores R. and Peña L. (2001). Transgenic citrus plants expressing the citrus tristeza virus p23 protein exhibit viral-like symptoms. *Mol Plant Pathol.* 2, 27-36

Hall, T.A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp Ser.* 41, 95-98

Harper S.J. (2013). Citrus tristeza virus: evolution of complex and varied genotypic groups. *Front Microbiol.* 4, 93

Harper, S.J., Cowell, S. J., and Dawson, W. O. (2015). Finding balance: Virus populations reach equilibrium during the infection process: *Virology* 485, 205-212.

Hilf M.E. and Garnsey S.M. (2000). Characterization and Classification of Citrus Tristeza Virus Isolates by Amplification of Multiple Molecular Markers. In: *Proc. 14th Conf. IOCV*, 18-27 IOCV, Riverside, California

Ibarbalz F.M., Pérez M.V., Figuerola E.L.M. and Erijman L. (2014). The Bias Associated with Amplicon Sequencing Does Not Affect the Quantitative Assessment of Bacterial Community Dynamics. *PLOS One* vol. 9, 99722

Iglesias N.G., Gago-Zachert S.P., Robledo G., Costa N., Plata M.I., Vera O., Grau, O. and Semorile L.C. (2008). Population structure of Citrus tristeza virus from field Argentinean isolates. *Virus Genes.* 36, 199-207

Ihrmark K., Bödeker I.T.M., Cruz-Martinez K., Friberg H., Kubartova A., Schenck J., Strid Y., Stenlid J., Brandström-Durling M., Clemmensen K.E. and Lindahl B.D. (2012). New primers to amplify the fungal ITS2 region – evaluation by 454-sequencing of artificial and natural communities. *FEMS Microbiol Ecol.* 82, 666-667

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

Lee R.F and Keremane M.L. (2013). Mild-strain cross protection of tristeza: a review of research to protect against decline on sour orange in Florida. *Front Microbiol.* 4, 259

Lefever S., Pattyn F., Hellermans J. and Vandesompele J. (2013). Single-Nucleotide Polymorphisms and Other Mismatches Reduce Performance of Quantitative PCR Assays. *Clin Chem.* 59, 1470-1480

Logan G., Freimans G.L., King D.J., Valdazo-González B., Bachanek-Bankowska K., Sanderson N.D., Knowles N.J., King P.J. and Cottam E.M. (2014). A universal protocol to generate consensus level genome sequences for foot-and-mouth disease virus and other positive-sense polyadenylated RNA viruses using the Illumina MiSeq. *BMC Genomics.* 15, 828

Moreno P., Ambros S., Albiach-Marti M.R., Guerri J. and Pena L. (2008). 'Citrus tristeza virus: A pathogen that changed the course of the citrus industry'. *Mol Plant Pathol.* 9, 251-268

Mullan B., Kenny-Walsh E., Collins J.K., Shanahan F. and Fanning L.J. (2001). Inferred Hepatitis C Virus Quasispecies Diversity Is Influenced by Choice of DNA Polymerase in Reverse Transcriptase-Polymerase Chain Reactions. *Anal Biochem.* 289, 137-146

Mullan B., Sheehy P., Shanahan F. and Fanning L. (2004). Do Taq-generated RT-PCR products from RNA viruses accurately reflect viral genetic heterogeneity?. *J Viral Hepatitis.* 11, 108-114

Nam Y.R., Lee U., Choi H.D., Lee K.J., Kim N., Jang Y.J. and Joo C.H. (2016). Degenerate PCR primer design for the specific identification of rhinovirus C. *J Virol Methods.* 214, 15-24

Ramos-González P.L., Sarubbi-Orue H., Gonzales-Segnana L., Chabi-Jesus C., Freitas-Astúa J. and Kitajima E.W. (2016). Orchid Fleck Virus Infecting Orchids in Paraguay: First Report and Use of Degenerate Primers for its Detection. *J Phytopathol.* 164, 342-347

- Read D. 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., Fayad A., Barthe G. and Brlansky R.H. (2005). A multiplex polymerase chain reaction method for reliable, sensitive and simultaneous detection of multiple viruses in citrus trees. *J Virol Methods.* 129, 47-55
- Rubio L., Ayllón 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
- Sambade A., Lopez C., Rubio L., Flores R., Guerri J. and Moreno P. (2003). Polymorphism of a specific region in gene p23 of Citrus tristeza virus allows for the discrimination between mild and severe isolates. *Arch Virol.* 148, 2325-2340
- Scott K.A., Hlela Q., Zablocki O., Read D., van Vuuren S. and 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
- Sekiya M.E., Lawrence S.D., McCaffery M. and Cline K. (1991). Molecular cloning and nucleotide sequencing of the coat protein gene of citrus tristeza virus. *J Gen Virol.* 72, 1013-1020
- Tamura K., Dudley J., Nei M. & Kumar S. (2007). MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol.* 24, 1596-1599.
- Teycheney P.Y., Acina I., Lockhart B.E.L. and Candresse T. (2007). Detection of Banana mild mosaic virus and Banana virus X by polyvalent degenerate oligonucleotide RT-PCR (PDO-RT-PCR). *J. Virol. Methods.* 142, 41-49
- van Dijk E.L., Auger H., Jaszczyszyn Y. and Thermes C. (2014). Ten years of next-generation sequencing technology. *Trends Genet.* 30, 418-426
- van Kuppeveld F.J.M., van der Logt J.T.M., Angulo A.F., van Zoest M.J., Quint W.G.V., Niesters H.G.M., Galama J.M.G. and Melchers W.J.G. (1992). Genus- and Species-Specific Identification of Mycoplasmas by 16S rRNA Amplification. *Appl Environ Microb.* 58, 2606-2615

Varanda C.M.R., Cardoso J.M.S., Oliveira M.D.M., Oliveira S., Clara M.I.E. and Félix M.R.F. (2014). A degenerate pair of primers for simultaneous detection of four alpha-and betanecroviruses. *J Virol Methods*. 208, 63-35

Watanabe K., Kodama Y. and Harayama S. (2001). Design and evaluation of PCR primers to amplify bacterial 16S ribosomal DNA fragments used for community fingerprinting. *J Microbiol Meth*. 44, 253-262

Waud M., Busschaert P., Ruyters S., Jacquemyn H. and Lievens B. (2014). Impact of primer choice on characterization of orchid mycorrhizal communities using 454 pyrosequencing. *Mol Ecol Resour*. 14, 679-699

Zheng L., Wayper P.J., Gibbs A.J., Fourment M., Rodoni B.C. and Gibbs M.J. (2008). Accumulating variation at conserved sites in Potyvirus genomes is driven by species discovery and affects degenerate primer design. *PLOS One*. 3, e1586