Phytopathologia Mediterranea
The international journal of the
Mediterranean Phytopathological Union

Firenze University Press
www.fupress.com/pm

Research Papers

Determining intra-host genetic diversity of *Citrus tristeza virus*. What is the minimal sample size?

Silvija ČERNI[1,*], Zlatko ŠATOVIĆ[2,3], Jelena RUŠČIĆ[4], Gustavo NOLASCO[5], Dijana ŠKORIĆ[1]

[1] *University of Zagreb, Faculty of Science, Department of Biology, Division of Microbiology, Marulićev trg 9a, 10000 Zagreb, Croatia*
[2] *University of Zagreb, Faculty of Agriculture, Department for Seed Science and Technology, Svetošimunska 25, 10000 Zagreb, Croatia*
[3] *Centre of Excellence for Biodiversity and Molecular Plant Breeding (CoE CroP-BioDiv), Svetošimunska cesta 25, 10000 Zagreb, Croatia*
[4] *Autolus Therapeutics plc, Forest House, 58 Wood Lane, White City, London, W12 7RZ*
[5] *Universidade do Algarve, Campus de Gambelas, Building 8, DCBB, 8005-139 Faro, Portugal (retired)*
*Corresponding author: silvija.cerni@biol.pmf.hr

Summary. Intra-host populations of plant RNA viruses are genetically diverse. Due to frequent reinfections, these populations often include phylogenetically distant variants that may have different biological properties. Random selection of variants, which occurs during host-to-host virus transmission, may affect isolate pathogenicity. Accurate characterization of genetic variants in intra-host virus populations is therefore epidemiologically important. In routine molecular characterization of *Citrus tristeza virus* (CTV) isolates, common practice is to analyze only a few cDNA clones per isolate. In the present study, based on the characterization of CTV populations displaying different levels of genetic diversity, we evaluated if analyzing large numbers of clones increased diversity parameters. A sequential sampling approach, based on analysis of genetic richness. is proposed for determining the minimal sample size required to obtain reliable information on levels of CTV genetic diversity.

Keywords. Cloning, CTV, population structure, sequencing, SSCP.

OPEN ACCESS

Citation: S. Černi, Z. Šatović, J. Ruščić, G. Nolasco, D. Škorić (2020) Determining intra-host genetic diversity of *Citrus tristeza virus*. What is the minimal sample size?. *Phytopathologia Mediterranea* 59(2): 295-302. DOI: 10.14601/Phyto-11276

Accepted: June 4, 2020

Published: August 31, 2020

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Competing Interests: The Author(s) declare(s) no conflict of interest.

Editor: Anna Maria D'Onghia, CIHEAM/Mediterranean Agronomic Institute of Bari, Italy.

INTRODUCTION

High error rate of RNA dependent RNA polymerases is the main contributor in generating heterogenic populations of genetically related variants, attributed to all RNA viruses (Ojosnegros *et al.*, 2011; Domingo *et al.*, 2012). While this benefits virus populations, enabling adaptation to variable environmental conditions and different selective regimes (Domingo *et al.*, 1997; Schneider and Roossinck, 2001), the variation complicates study of virus isolate genetic structure and understanding of effects on the symptom expression. All evidence

suggests that mutant spectra can hide components that in isolation would display dissimilar biological properties (Domingo *et al.*, 2012). Due to frequent reinfections, some long-living hosts can also accumulate highly heterogenic virus variants, some of which are phylogenetically distant. As a result of repeated genetic bottlenecks occurring during virus host-to-host transmission events (Li and Roossinck, 2004; Ali *et al.*, 2006), virus populations are subjected to fitness reduction and biological alterations (Domingo *et al.*, 2012). Minor population variants may influence the symptom expression (Domingo *et al.*, 2006; Cerni *et al.*, 2008), so it is important to have an efficient tool for accurate characterization of virus isolates, both for gaining insights into isolate epidemiology, and for understanding virus pathogenicity.

For isolates of *Citrus tristeza virus* (CTV), the most important virus pathogen of citrus, the relationships between population structure of an isolate and its biological expression remain neither well documented nor understood. This is also the case for many other viruses. Based on coat protein (CP) gene sequences, CTV variants are grouped into seven phylogenetic clusters between which the representatives of at least four clusters clearly differ in pathogenic potential (Nolasco *et al.*, 2009; Hančević *et al.*, 2013). Several authors have not fully characterized the extent of CTV variants composing individual isolates. Consequently, there are inconsistencies between phylogenetic data and isolate biological characteristics. Much understanding of viral pathogenesis derives from studies of single viral clones, which may not reveal many of the most important aspects of natural infections (Lauring and Andino, 2011). Gao *et al.* (2005) proposed the hypothesis that the level of intra-host virus population diversity can easily be missed. They showed that for obtaining good population structure of *Hepatitis virus C* isolates, it was necessary to analyze up to 40 separated genomic variants (clones). As CTV infects long-living hosts subjected to frequent reinfections, with genetic bottlenecks often occurring during virus transmission by the aphid vectors or grafting (Nolasco *et al.*, 2008; Cerni *et al.*, 2008), CTV variants belonging to diverse phylogenetic lineages are expected to coexist.

The probability of sampling at least one sequence of genomic variant x is given by the formula $Pr = 1-(1-p)^n$, where p is the frequency of the genetic variant x, and n is the sample size. Thus, the minimal sample size can be calculated as: $n = \log(1-Pr)/\log(1-p)$ (Ott, 1992). Therefore, the sample size of n = 59 should be adequate to detect a genetic variant whose frequency in the population is greater than 5%, with the probability of 95%. Further, if k variants present at frequency p are to be detected with a probability of Pr, the equation becomes: $n = \log(1-Pr^{1/k})/\log(1-p)$, and, similarly, the sample size of n = 72 would be needed to detect two genetic variants with frequencies greater than 5%.

Bulk virus clone analysis is time consuming and expensive, and not always required. Our goal was to investigate to what extent analysis of increased numbers of clones, based on a range of diversity parameters, enhanced information quality on CTV genomic variants within a virus population. We also assessed if an approach based on analyses of genetic richness could determine optimum sample size.

MATERIALS AND METHODS

Citrus tristeza virus *isolates*

All samples used in this study were previously confirmed to be CTV positive, by DAS-ELISA (Loewe). Samples were taken randomly from seven citrus hosts including: the *Citrus unshiu* Mac. Mark. varieties Aoshima, Chahara and Zorica Rana (isolates AO, CH and ZR), *C. sinensis* (L.) Osbeck variety Fukumoto Navel (isolate FN), and three *C. wilsonii* Tanaka plants (isolates CwS1, CwS2, CwS3). Two additional samples were created by pooling tissue of isolates AO, CH and ZR (sample P1) and isolates CwS1, CwS2 and CwS3 (sample P2), to create samples having high genetic diversity. Total RNA was extracted from each sample of infected tissue collected from different plant sections (Cerni *et al.*, 2008), using the RNeasy Plant Mini Kit (Qiagen).

Separation and characterization of Citrus tristeza virus *variants*

Coat protein (CP) genes of CTV variants present in the sample populations were amplified by one-step RT-PCR (Cerni *et al.*, 2008), using primers corresponding to both ends of the CP gene (Gillings *et al.*, 1993). Resulting products (672 bp) were separated by electrophoresis in 1% agarose gel, and visualized in UV-light after ethidium bromide staining. To separate different CTV variants, amplicons were TA-cloned into the pTZ57R/T vector, followed by the transformation of competent *Escherichia coli* INVαF' cells, as described by the manufacturer (Fermentas). Transformed colonies were selected by α-complementation, and the presence of the insert was confirmed by PCR using the same primers and PCR reaction conditions as described above.

Identification of different genomic variants was performed by Single-Strand Conformation Polymorphism (SSCP) analysis (Rubio *et al.*, 1996). For each virus iso-

late, the PCR products from 40 transformed colonies were analyzed sequentially in groups of ten, and different patterns were visualized by silver staining (Beidler *et al.*, 1982). PCR products displaying different SSCP patterns were considered different genomic variants (Kong *et al.*, 2000), and these were chosen for sequencing. Plasmids containing different genomic variants of each isolate were purified using a PureLink™ Quick Plasmid Miniprep Kit (Invitrogen), and were sequenced in both directions (Macrogen Inc.) using a pair of M13-pUC universal primers. Different genomic variants and their phylogenetic clustering were determined by using MEGA 5.5 (Tamura *et al.*, 2011), applying the neighbour-joining method and Kimura 2-parameter evolutionary model. The sequences of reference isolates were the same as those described in a previous study (Cerni *et al.*, 2009).

Data analyses

Genetic diversity of CTV isolates was assessed at genomic variant and phylogenetic group levels.

For each isolate the following parameters were calculated: (1) the total numbers (N_{Total}) of genomic variants and phylogenetic groups detected; (2) the equivalent numbers (N_E) of genomic variants/phylogenetic groups; (3) genomic variant/phylogenetic group diversity, as measured by Shannon's information index (H) (Lewontin, 1972); (4) genomic variant/phylogenetic group diversity, expressed as a maximum diversity (H_{Max}) for a given number of genomic variants/phylogenetic groups achieved in the case of equal frequencies of all genomic variants/phylogenetic groups in a sample (isofrequency situation); (5) the proportion of genomic variant/phylogenetic group diversity within and among four sets of ten clones (H_{avg}/H_{Total}); and (6) the statistical significance (Fisher's exact test) for differences in number of clones per genomic variant/phylogenetic group among four sets of ten clones (P). All the parameters are explained below in terms of genomic variants. The same approach was used in the assessment of phylogenetic group diversity.

Equivalent number of genomic variants (N_E) represents the number of equally frequent genomic variants for a given level of diversity, thus allowing comparison of isolates where the numbers and distributions of genomic variants differ greatly. N_E was calculated as:

$$N_E = \frac{1}{\sum_{i=1}^{I} p_i^2}$$

where, p_i is the frequency of the ith genomic variant, and I is the total number of genomic variants of the isolate.

The proportion of genomic variant diversity within and among four sets of ten clones was calculated based on Shannon's information index. This was calculated as:

$$H = -\sum_{i=1}^{I} (p_i \log_2 p_i)$$

where, p_i is the frequency of the ith genomic variant, and I is the total number of genomic variants of the isolate.

Shannon's information index was used to measure the total genomic variant diversity (H_{Total}) of a complete set of 40 clones, as well as the average diversity (H_{Avg}), over four sets of ten clones. The proportions of diversity within (H_{Avg}/H_{Total}) and among sets [(H_{Total}–H_{Avg})/H_{Total}] were calculated.

Fisher's exact test in SAS (SAS Institute Inc., 2004) was used to test for differences in number of clones in each genomic variant among four sets of ten clones. The number of genomic variants detected in a first set of ten clones and the number of additional genomic variants detected in incremental sets of ten clones were determined for each permutation of four sets of clones (n! = 4! = 24 possible orders), and averaged.

An approach for determining the optimum sample size to obtain the maximum information on CTV genomic variants present in an isolate was based on the formula proposed by Hurlbert (1971) for the assessment of species richness. In this method, the expected number of species (species richness) is calculated for the collection to be compared after all collections are scaled down to the same number of individuals, namely that of the smallest collection. This scaling is necessary because large collections would have more species than small ones, even if they were drawn from the same community (Heck *et al.*, 1975). This approach, called rarefaction, has been adopted (Hughes *et al.*, 2001) for estimation and comparison of species richness among sites, treatments, or habitats that have been unequally sampled. The same formula was introduced by Petit *et al.* (2008) for the calculation of haplotype richness, as the measure of the number of haplotypes per population independent of sample size, while the modified version, described by El Mousadik and Petit (1996), estimates allelic richness, i.e. the number of alleles per population comprised of diploid organisms.

In the present case, Hurlbert's formula was used to calculate genomic variant richness (N_r) as:

$$N_r = \sum_{i=1}^{I} \left[1 - \frac{\binom{N - N_i}{n}}{\binom{N}{n}} \right]$$

where, N is the sample size, n is the subsample size, N_i is the count of genomic variant i, and I is the total number of genomic variants.

The approach used in the present study included sequential sampling coupled with calculation of genomic variant richness for increasing numbers of samples. The principle was to estimate the expected number of genomic variants in a subsample of n individuals (n < N), given that N individuals were sampled. As explained above, genomic variant richness is a measure of genomic variant numbers independent of sample size. This allows comparison of this quantity between different sample sizes. In the present case, the genomic variant richness of samples of N = 20, 30 and 40 was equal to the expected number of genomic variants that a sample would have had if the sample size had been n individuals instead of N, n being equal to N - 10. Genomic variant richness (N_r; n = N - 10) was compared to the observed number of genomic variants in a sample of size n, and the difference (ΔI) between these two parameters was calculated. The procedure was repeated for all possible sampling orders, and the differences (ΔI) between genomic variant richness (N_r) and the observed number of genomic variants in a sample of size n were averaged. This approach can be applied for any number of individuals given that n > 1.

RESULTS

Molecular characterization of Citrus tristeza virus isolates

RT-PCR gave strong amplification signals corresponding to 672 bp length products of the CTV coat protein gene in all isolates. After the separation of different genomic variants by TA-cloning, 40 PCR products, representing separated variants of each isolate, were subjected to SSCP analysis in groups of ten. For each isolate, clones displaying different SSCP patterns were sequenced in a number approximately proportional to its frequency in the population. It was verified that the sequenced clones of each isolate displaying the same SSCP pattern had no significant nucleotide differences, validating the assumption that each SSCP pattern corresponded to a single genomic variant as suggested by Kong *et al.* (2000). For each isolate, all clones displaying distinct SSCP patterns had different nucleotide sequences. The phylogenetic analysis (results not presented) of the obtained sequences showed that all analyzed variants of isolates CH and ZR clustered into one phylogenetic group, while for all the other isolates, a mixture of genomic variants were detected, belonging to different phylogenetic groups (Supplementary Table 1). Variants of isolates AO, CwS2 and in sample P1 clustered into two different phylogenetic groups, and variants of isolates FN, CwS1, CwS3 and in sample P2 clustered into three different phylogenetic groups.

Genomic variant diversity

Total number of genomic variants (N_{Total}) detected across isolates varied from three to ten, but the equivalent number (N_E) was much smaller, exceeding three variants per isolate in case of three out of nine isolates (Table 1). The most frequent genomic variant (f_{Max}) had a frequency greater than 75% in five out of nine isolates; in only two cases (isolates AO and sample P1) the frequency of the most frequent variant did not exceed 50%. Variant diversity (H_{Total}) was generally low; in seven out of nine isolates H_{Total} was less than 75% of the H_{Max}, and frequencies of the variants were far from being equal. Generally, the greatest diversity (H_{avg}/H_{Total}) was attributable to within-sample diversity. Different sets of ten clones were not significantly different ($P > 0.05$) in number of clones per variant, except in the case of sample P1.

Table 1. Genomic variant diversity of nine *Citrus tristeza virus* isolates, based on SSCP analyses.

Isolate	N_{Total}	N_E	N_{Range}	f_{Max}	H_{Total}	$\%H_{Max}$	H_{Avg}/H_{Total}	P
AO	4	3.00	3-4	42.5	1.679	83.94	0.951	0.971
CH	3	1.29	1-3	87.5	0.634	39.99	0.833	0.726
CwS1	7	1.74	2-5	75.0	1.442	51.36	0.749	0.500
CwS2	5	1.30	2-3	87.5	0.784	33.76	0.743	1.000
CwS3	6	1.73	2-5	75.0	1.342	51.91	0.842	0.958
FN	5	2.40	2-4	57.5	1.570	67.60	0.824	0.269
ZR	4	1.17	1-2	92.5	0.503	25.16	0.699	1.000
P1	7	3.60	3-5	40.0	2.156	76.81	0.652	<0.001
P2	10	3.45	3-7	50.0	2.454	73.88	0.761	0.196

N_{Total}, Total number of genomic variants detected.
N_E, Equivalent number of genomic variants.
N_{Range}, Range of genomic variants detected in sets of ten clones.
f_{Max}, Frequency of the most frequent variant.
H_{Total}, Genetic diversity, as measured by Shannon's information index.
$\%H_{Max}$, Genetic diversity expressed as percent of the maximum diversity for a given number of variants achieved in case of equal frequencies of all the variants in an isolate (isofrequency situation).
H_{avg}/H_{Total}, Proportion of genomic variant diversity within four sets of ten clones.
P, Probability (Fisher's exact test) of differences in number of clones per variant among four sets of ten clones.

The average number of genomic variants detected in the first set of ten clones ranged from 1.75 (ZR) to 4.75 (P2), as obtained by permutations (Figure 1). Thus, the proportion (%) of variants detected in the first set varied between 43.75% (isolate ZR) to 82.25% (isolate AO). The analysis of an additional set of ten clones (second) yielded on average more than a quarter of the total number of variants in only one case (isolate CwS1), while in both third and fourth sets of ten clones, the number of additional variants detected did not exceed a fifth of the total number of variants.

Analyses of ΔI values in relation to increasing numbers of sampled clones (Figure 2) showed that ten clones would represent the existing CTV diversity of isolates AO, CH, CwS2, CwS3 and ZR, so additional sets did not increase the genetic richness over the number of variants already detected in a first set of ten clones. Additional sets (second, third or fourth) of ten clones yielding higher genetic richness (N_r) over the number of variants (n) detected in the first set of ten clones (i.e. $\Delta I > 0$; Figure 2 indicated in red) were needed in case of four out of nine isolates. In the case of samples of isolates CwS1, FN and sample P1, at least three sets were needed, but for sample P2, the analysis of the fourth set still increased the genetic richness over the number of variants detected in previous three sets, so more sets should be analysed.

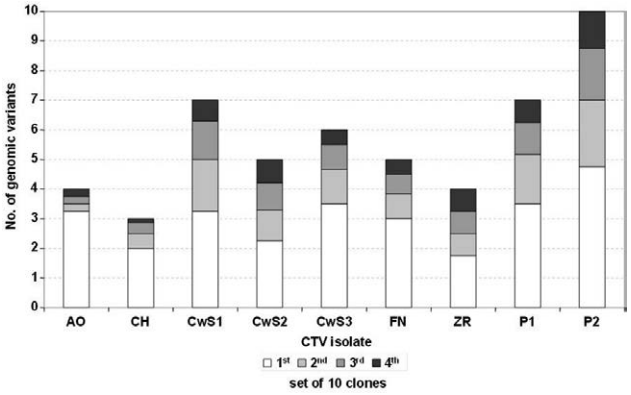

Figure 1. The number of genomic variants detected in a 1st set of 10 clones of nine *Citrus tristeza virus* isolates based on the results of the SSCP analysis and the number of additional variants detected in incremental sets of ten clones averaged over 24 possible sampling orders of four sets.

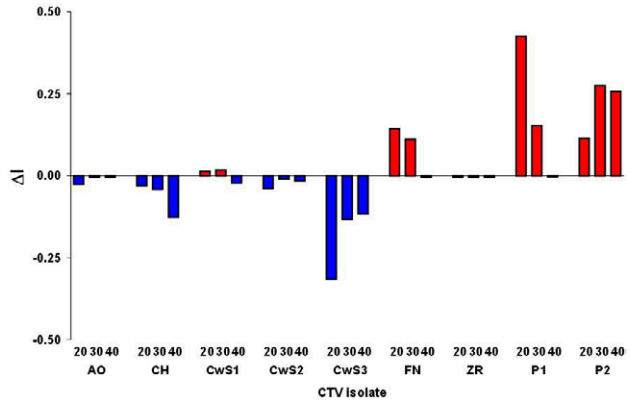

Figure 2. ΔI values of genomic variant richness in relation to increasing number of sampled clones averaged over 24 possible sampling orders of four sets based on the results of the SSCP analysis of nine *Citrus tristeza virus* isolates.

Phylogenetic group diversity

Total numbers of phylogenetic groups (N_{Total}) detected across isolates varied from 1 to 3, but the equivalent numbers (N_E) were below two in all the cases (Table 2). The most frequent (f_{Max}) phylogenetic group had a frequency greater than 75% in five out of seven isolates in

Table 2. Phylogenetic group diversity of nine *Citrus tristeza virus* isolates based on the results of the SSCP analysis and the classification, as proposed by Nolasco *et al.* (2009).

Isolate	N_{Total}	N_E	N_{Range}	f_{Max}	H_{Total}	$\%H_{Max}$	H_{Avg}/H_{Total}	P
AO	2	1.72	2-2	70.0	0.881	88.13	0.980	0.963
CH*	1	-	-	-	-	-	-	-
CwS1	3	1.23	1-2	90.0	0.569	35.90	0.729	0.408
CwS2	2	1.05	1-2	97.5	0.169	16.87	0.695	1.000
CwS3	3	1.36	2-3	85.0	0.748	47.17	0.930	1.000
FN	3	1.83	2-3	67.5	1.037	65.42	0.872	0.266
ZR*	1	-	-	-	-	-	-	-
P1	2	1.34	2-2	85.0	0.610	60.98	0.938	0.684
P2	3	1.59	2-3	77.5	0.976	61.57	0.879	0.456

N_{Total}, Total number of phylogenetic groups detected.
N_E, Equivalent number of phylogenetic groups.
N_{Range}, Range of phylogenetic groups detected in sets of ten clones.
f_{Max}, Frequency of the most frequent phylogenetic group.
H_{Total}, Genetic diversity, as measured by Shannon's information index.
$\%H_{Max}$, Genetic diversity, expressed as percent of the maximum diversity for a given number of phylogenetic groups achieved in the case of equal frequencies of all the groups in an isolate (isofrequency situation).
H_{avg}/H_{Total}, Proportion of phylogenetic group variant diversity within four sets of ten clones.
P, Probability (Fisher's exact test) for differences in number of clones in each phylogenetic group among four sets of ten clones.
* For samples CH and ZR, all the clones belonged to the same phylogenetic group.

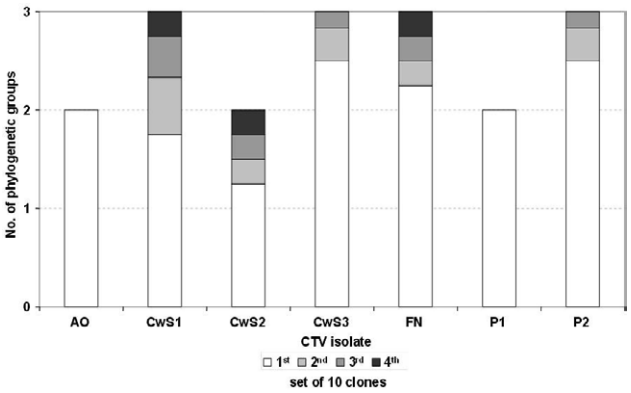

Figure 3. Number of phylogenetic groups detected in a first set of ten clones of seven *Citrus tristeza virus* isolates, based on the classification proposed by Nolasco *et al.* (2009), and number of additional variants detected in incremental sets of ten clones, averaged over 24 possible sampling orders of four sets.

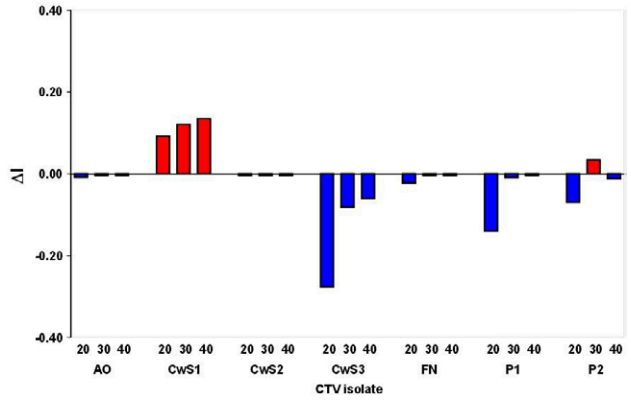

Figure 4. ΔI values of phylogenetic group richness in relation to increasing number of sampled clones averaged over 24 possible sampling orders of four sets based on the results of the classification proposed by Nolasco *et al.* (2009).

which more than one phylogenetic group was detected. Phylogenetic group diversity (H_{Total}) was generally low; in six out of seven isolates this was less than 75% of the H_{Max}; frequencies of the phylogenetic groups were far from being equal. Generally, the isolates tended to be comprised of genomic variants belonging predominantly to the same phylogenetic group. Most of the phylogenetic group diversity was attributable to within-sample diversity. The among-sample component was greater than 0.30 in just one case (isolate CwS2). Different sets of ten clones were not significantly different in number of clones per phylogenetic group.

The average number of phylogenetic groups (Figure 3) detected in the first set of ten clones ranged from 58.33% (isolate CwS1) to 100.00% (isolate AO and sample P1), as obtained by permutations. In the second, third and fourth sets of ten clones, the numbers of additional phylogenetic groups detected did not exceed a fifth of the total number of phylogenetic groups.

Analyses of ΔI values in relation to increasing number of sampled clones (Figure 4) showed that ten clones adequately represented the existing CTV diversity of isolates AO, CwS2, CwS3, and FN, and sample P1. Additional sets did not increase the genetic richness over the number of phylogenetic groups already detected in the first set of ten clones. Additional sets (up to four) of ten clones yielding greater phylogenetic group richness than the number of phylogenetic groups (n) detected in the first set (i.e. $\Delta I > 0$; Figure 4 marked in red), were required for isolate CwS1 and sample P2.

DISCUSSION

Accurate characterization and systemic monitoring of intra-host virus population diversity is important, as demonstrated from increasing numbers of studies (Domingo *et al.*, 2012). In epidemiological research, this information is essential for identification of risk factors and surveillance of diseases. In evolutionary studies characterization aids understanding of how evolutionary forces shape virus populations, and in the studies of virus pathogenesis it improves understanding of how genetically different variants contribute to particular viral phenotypes.

Cloning and sequencing are the standard molecular procedures used for the detection and quantification of genetic variants present in virus quasispecies. Although sampling theory gives strict guidelines for numbers of clones that should be analyzed to detect genetic variants having particular population frequencies, extensive clone analyses are often unnecessary. Cloning and sequencing are time consuming and expensive, and genetic diversity of virus isolates may range from very low to high. Therefore, our aim was to develop an approach that could be applied to individual samples, regardless of their diversity levels, and to provide the information of adequate sample size.

In the case of CTV (Nolasco *et al.*, 2009) and *Grapevine leafroll associated virus -3* (Gouveia *et al.*, 2011), PCR-ELISA based typing tool (APET) has been developed, which allows rapid identification of the phylogenetic groups present in a sample, without the need for cloning. Use of this tool for initial analyses would allow identifying *a priori* how many groups are present, and would provide indication of when to stop cloning and sequencing. Nevertheless, for quantification and identification of variants from the same phylogenetic groups, cloning and sequencing remains the only suitable tool.

There is an increasing number of reports using next-generation sequencing for virus quasispecies characterisation (Barbezange *et al*., 2018, Yu *et al*., 2018). Although promising, next-generation sequencing approaches that enable rapid detection of single-nucleotide polymorphisms within different haplotypes is still challenging for virus quasispecies reconstruction. This requires bioinformatics support, and tools are not always well adapted, especially when sequence divergence is low or rare haplotypes are present (Schirmer *et al*., 2012). Therefore, cloning-based approaches may serve as good alternatives, and also as biological controls in results obtained after next-generation sequencing.

The present study, based on comprehensive characterization of genetic diversity of nine intra-host CTV populations displaying different diversity levels, revealed the following at genomic variant and phylogenetic group levels: (i) the equivalent number of genetic variants was much smaller than the total number of genetic variants detected; (ii) the most frequent genetic variant had frequency greater than 50% in almost all cases; (iii) variant diversity was generally low and frequencies of the variants were far from being equal; (iv) most of the diversity was attributed to within-sample variation; and (v) there were no significant differences between sets of clones. These results are similar to those of Chare and Holmes (2004) and García-Arenal *et al*. (2001), who reported that pairwise nucleotide sequence diversity among plant virus populations is usually low. García-Arenal *et al*. (2001) also reported that the most common composition of plant virus populations was differing numbers of haplotypes, separated only by small genetic distances that exhibit L-shaped rank-abundance curves; that is, one or a few major haplotypes plus many minor haplotypes.

Although the most abundant variants of CTV would most probably be detected in the first set of ten clones, the question still remains of "How many clones should be analyzed to adequately provide new information on population diversity?". The answer is given by an approach based on the calculation of genetic richness. Sequential sampling based on genetic richness gives the optimal sample size, and enables termination of the analysis when sufficient information regarding genetic diversity of the population is reached. Sequential sampling, as proposed, is suitable for genomic variant and phylogenetic group analyses, and consists of the following steps: (i) analyze two sets of ten clones each; (ii) calculate allelic richness (N_r) of a sample of N = 20, equal to the expected number of genetic variants that a sample would have had if the sample size had been ten; (iii) compare the genetic variant richness (N_r) to the average number of genetic variants detected in two sets of ten clones, and calculate the difference (ΔI); (iv) if $\Delta I < 0$, the number of sequenced clones is sufficient to represent the genetic variant diversity of the isolate; and (v) if $\Delta I > 0$, there is need for sequencing an additional set of ten clones, and for the procedure to be repeated.

The proposed approach is advantageous especially in cases when multiple factors influencing the structure of plant virus populations are to be assessed, such as geographic locations, hosts or duration of infections. While in such cases hierarchical sampling and/or randomized complete block studies are required (D'Urso *et al*., 2003; Moury *et al*., 2006), sequential sampling could be useful. This can provide information about the unexplored diversity at each sampling stage, and can aid comparisons of diversity in populations (or strata) with unequal sampling effort.

The experimental system proposed here is readily applicable to other host/virus systems, independently of population diversity, and can be used to track virus variants.

LITERATURE CITED

Ali A., Li H., Schneider W.L., Sherman D.J., Gray S., Roossinck M.J., 2006. Analysis of genetic bottlenecks during horizontal transmission of Cucumber mosaic virus. *Journal of Virology* 80: 8345–8350.

Beidler J.L., Hilliard P.R., Rill R.L., 1982. Ultrasensitive staining of nucleic acids with silver. *Analytical Biochemistry* 126: 374–380.

Barbezange C., Jones L., Blanc H., Isakov O., Celniker G., van der Werf S., 2018. Seasonal Genetic Drift of Human Influenza A Virus Quasispecies Revealed by Deep Sequencing. *Frontiers in Microbiology* 9: 2596.

Cerni S., Ruscic J., Nolasco G., Gatin Z., Krajacic M., Skoric D., 2008. Stem pitting and seedling yellows symptoms of Citrus tristeza virus infection may be determined by minor sequence variants. *Virus Genes* 36: 241–249.

Cerni S., Skoric D., Ruscic J., Krajacic M., Papic T., Nolasco G., 2009. East Adriatic—a reservoir region of severe Citrus tristeza virus strains. *European Journal of Plant Pathology* 124: 701–706.

Chare E.R., Holmes E.C., 2004. Selection pressures in the capsid genes of plant RNA viruses reflect mode of transmission. *The Journal of General Virology* 85: 3149–3157.

D'Urso F., Sambade A., Moya A., Guerri J., Moreno P., 2003. Variation of haplotype distributions of two genomic regions of Citrus tristeza virus populations from eastern Spain. *Molecular Ecology* 12: 517–526.

Domingo E., Menéndez-Arias L., Holland J., 1997. RNA virus fitness. *Reviews in Medical Virology* 7: 87–96.

Domingo E., Martin V., Perales C., Grande-Pérez A., García-Arriaza J., Arias A, 2006. Viruses as quasispecies: biological implications. *Current Topics in Microbiology and Immunology* 299: 51–82.

Domingo E., Sheldon J., Perales C., 2012. Viral quasispecies evolution. *Microbiology and Molecular Biology Reviews: MMBR* 76: 159–216.

El Mousadik A., Petit R.J., 1996. High level of genetic differentiation for allelic richness among populations of the argan tree [*Argania spinosa* (L.) Skeels] endemic to Morocco. *TAG. Theoretical and Applied Genetics* 92: 832–839.

Gao G., Stuver S.O., Okayama A., Tsubouchi H., Mueller N.E., Tabor E., 2005. The minimum number of clones necessary to sequence in order to obtain the maximum information about hepatitis C virus quasispecies: a comparison of subjects with and without liver cancer. *Journal of Viral Hepatitis* 12: 46–50.

García-Arenal F., Fraile A., Malpica J.M., 2001. Variability and genetic structure of plant virus populations. *Annual Review of Phytopathology* 39: 157–186.

Gillings M., Broadbent P., Indsto J., Lee R., 1993. Characterisation of isolates and strains of citrus tristeza closterovirus using restriction analysis of the coat protein gene amplified by the polymerase chain reaction. *Journal of Virological Methods* 44: 305–317.

Gouveia P., Santos M.T., Eiras-Dias J.E., Nolasco G., 2011. Five phylogenetic groups identified in the coat protein gene of grapevine leafroll-associated virus 3 obtained from Portuguese grapevine varieties. *Archives of Virology* 156: 413–420.

Hančević K., Černi S., Nolasco G., Radić T., Djelouah K., Škorić D., 2013. Biological characterization of Citrus tristeza virus monophyletic isolates with respect to p25 gene. *Physiological and Molecular Plant Pathology* 81: 45–53.

Heck K.L., van Belle G., Simberloff D., 1975. Explicit Calculation of the Rarefaction Diversity Measurement and the Determination of Sufficient Sample Size. *Ecology* 56: 1459–1461.

Hughes J.B., Hellmann J.J., Ricketts T.H., Bohannan B.J., 2001. Counting the uncountable: statistical approaches to estimating microbial diversity. *Applied and Environmental Microbiology* 67: 4399–4406.

Hurlbert S.H., 1971. The nonconcept of species diversity: a critique and alternative parameters. *Ecology* 52: 577–586.

Kong P., Rubio L., Polek M., Falk B.W., 2000. Population structure and genetic diversity within California Citrus tristeza virus (CTV) isolates. *Virus Genes* 21: 139–145.

Lauring A.S., Andino R., 2011. Exploring the fitness landscape of an RNA virus by using a universal barcode microarray. *Journal of Virology* 85: 3780–3791.

Lewontin R.C., 1972. The apportionment of human diversity. *Evolutionary Biology* 6: 381–398.

Li H, Roossinck M.J., 2004. Genetic bottlenecks reduce population variation in an experimental RNA virus population. *Journal of Virology* 78: 10582–10587.

Moury B., Desbiez C., Jacquemond M., Lecoq H., 2006. Genetic diversity of plant virus populations: towards hypothesis testing in molecular epidemiology. *Advances in Virus Research* 67: 49–87.

Nolasco G., Fonseca F., Silva G., 2008. Occurrence of genetic bottlenecks during citrus tristeza virus acquisition by Toxoptera citricida under field conditions. *Archives of Virology* 153: 259–271.

Nolasco G., Santos C., Silva G., Fonseca F., 2009. Development of an asymmetric PCR-ELISA typing method for citrus tristeza virus based on the coat protein gene. *Journal of Virological Methods* 155: 97–108.

Ojosnegros S., Perales C., Mas A., Domingo E., 2011. Quasispecies as a matter of fact: viruses and beyond. *Virus Research* 162: 203–215.

Ott J. 1992 Strategies for characterizing highly polymorphic markers in human gene mapping. *American Journal of Human Genetics* 51: 283–290.

Petit R.J., El Mousadik A., Pons O., 2008. Identifying Populations for Conservation on the Basis of Genetic Markers. *Conservation Biology* 12: 844–855.

Rubio L., Ayllonl M.A., Guerri J., Pappu H., Niblett C., Moreno P., 1996. Differentiation of citrus tristeza closterovirus (CTV) isolates by single-strand conformation polymorphism analysis of the coat protein gene. *Annals of Applied Biology* 129: 479–489.

SAS Institute Inc. 2004. SAS/INSIGHT® 9.1 User's Guide, Volumes 1 and 2. Cary, NC

Schirmer M., Sloan W.T., Quince C., 2012. Benchmarking of viral haplotype reconstruction programmes: an overview of the capacities and limitations of currently available programmes. *Briefings in Bioinformatics*.

Schneider W.L., Roossinck M.J., 2001. Genetic diversity in RNA virus quasispecies is controlled by host-virus interactions. *Journal of Virology* 75: 6566–6571.

Tamura K., Peterson D., Peterson N., Stecher G., Nei M., Kumar S., 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution* 28: 2731–2739.

Yu F., Wen Y., Wang J., Gong Y., Feng K., Qiu M., 2018. The Transmission and Evolution of HIV-1 Quasispecies within One Couple: a Follow-up Study based on Next-Generation Sequencing. *Scientific Reports* 8: 1404.