Contribution of recombination and selection to molecular evolution of *Citrus tristeza virus*

Susana Martín [1, 2†], Adrián Sambade [1‡], Luis Rubio [1], María C. Vives [1], Patricia Moya [1], José Guerri [1], Santiago F. Elena [2], Pedro Moreno [1*]

Authors Susana Martín and Adrián Sambade contributed equally to this work.

[1] *Centro de Protección Vegetal y Biotecnología, Instituto Valenciano de Investigaciones Agrarias, Moncada, 46113 Valencia, Spain*

[2] *Instituto de Biología Molecular y Celular de Plantas (CSIC-UPV), 46022 Valencia, Spain*

Running title: *Citrus tristeza virus molecular evolution*

[*] Corresponding author: Instituto Valenciano de Investigaciones Agrarias, Ctra. Moncada-Náquera Km. 4.5, Moncada, 46113 Valencia, Spain. Phone: 34-96-3424069, Fax: 34-96-3424001. *E-mail address*: pmoreno@ivia.es

[†] Present address: Instituto de Biología Molecular y Celular de Plantas (CSIC-UPV), Campus UPV CPI 8E, Ingeniero Fausto Elio s/n, 46022 Valencia, Spain

‡ Present address: Department of Cell and Developmental Biology, John Innes Centre, Norwich NR4 7UH, UK

SUMMARY: 212 words

MAIN TEXT AND FIGURE LEGENDS: 5496 words

NUMBER OF TABLES: 2

NUMBER OF FIGURES: 3

Summary

The genetic variation of *Citrus tristeza virus* (CTV) was analyzed comparing the predominant sequence variants in seven genomic regions (p33, p65, p61, p18, p13, p20, and p23) of 18 pathogenically distinct isolates from seven different countries. Analyses of the selective constraints acting on each codon suggest that most regions were under purifying selection. Phylogenetic analysis show diverse patterns of molecular evolution for different genomic regions. A first clade composed by isolates genetically close to the reference mild isolates T385 or T30 was inferred from all genomic regions. A second clade, mostly comprising virulent isolates, was defined from regions p33, p65, p13, p20, and p23. For regions p65, p61, p18, p13, and p23 a third clade that mostly included South American isolates could not be related with any reference genotype. Phylogenetic relationships among isolates did not reflect their geographical origin, suggesting significant gene flow between geographically distant areas. Incongruent phylogenetic trees for different genomic regions suggested recombination events, an extreme that was supported by several recombination-detecting methods. A phylogenetic network incorporating the effect of recombination showed an explosive radiation pattern for the evolution of some isolates and grouped isolates by virulence. Taken together, the above results suggest that negative selection, gene flow, sequence recombination, and virulence may be important factors driving CTV evolution.

Keywords: CTV; genetic variability; maximum likelihood; recombination; selective constraints; virus evolution

Introduction

Citrus tristeza virus (CTV) is a closterovirus, family *Closteroviridae*, with two capsid proteins of 25 and 27 kDa, coating ~97 and ~3% of the virion length, respectively (Febres *et al.*, 1996; Satyanarayana *et al.*, 2004). The single-stranded and positive-sense CTV genomic RNA (gRNA) is about 20 kb in size and contains 12 open reading frames (ORFs) potentially encoding at least 19 proteins (Karasev *et al.*, 1995). ORFs 1a and 1b, encoding replication-related proteins, are translated from the gRNA, whereas the 10 3'-proximal ORFs, encoding proteins p33, p6, p65, p61, p27, p25, p18, p13, p20, and p23, are expressed via 3' co-terminal subgenomic RNAs (Hilf *et al.*, 1995). Protein p6 may operate as a membrane anchor (Satyanarayana *et al.*, 2000); proteins p65 (a homologue of the HSP70 heat shock proteins), p61 and the two coat proteins are involved in virion assembly (Satyanarayana *et al.*, 2000); p20 accumulates in amorphous inclusion bodies (Gowda *et al.*, 2000), and p23, an RNA-binding protein (López *et al.*, 2000), controls asymmetrical accumulation of plus and minus strands during RNA replication (Satyanarayana *et al.*, 2002) and is involved in symptom expression (Ghorbel *et al.*, 2001, Fagoaga *et al.*, 2005). Proteins p23, p20 and p25 act as RNA silencing suppressors (Lu *et al.*, 2004). The functions of p33, p13 and p18 remain unknown.

CTV is primarily dispersed by propagation of infected buds, and then it is locally spread aphids. CTV-induced symptoms include *i*) decline of citrus species propagated on sour orange (*Citrus aurantium* L.) rootstock, *ii*) yellowing and growth cessation of sour orange, lemon (*C. limon* (L.) Burn. f.) or grapefruit (*C. paradisi* Macf.) seedlings (seedling yellows), or *iii*) stunting, stem pitting, and poor yield of different citrus varieties regardless the rootstock used (Moreno *et al.*, 2008). The molecular mechanisms involved in symptom expression are still unknown.

As for other RNA viruses, genetic variation has been observed in CTV isolates resulting from the error-prone nature of RNA polymerases and selection pressures (Domingo & Holland, 1994), superinfection of field trees with divergent CTV variants (Rubio *et al.*, 2001), genetic drift after transmission to new hosts (Albiach-Martí *et al.*, 2000a; D'Urso *et al.*, 2000; Ayllón *et al.*, 2006), or recombination (Rubio *et al.*, 2001; Vives *et al.*, 2005). Characterization of the

genetic structure of viral populations and factors contributing to their evolution may help understanding important features like the outbreak of new epidemics or virulence changes in current isolates (Fernández-Cuartero *et al.*, 1994; Escriu *et al.*, 2000). These studies have practical implications in virus control, since durability of host resistance largely depends on genetic variability of the virus (García-Arenal & McDonald, 2003).

Previously we compared the predominant sequence variants of gene *p23* from 18 CTV isolates of different origins and pathogenicity characteristics (Sambade *et al.*, 2003). Phylogenetic analyses showed that sequence variants predominant in mild isolates (causing mild to moderate symptoms in Mexican lime (*C. aurantifolia* (Christ.) Swing.) and sometimes decline of sweet orange (*C. sinensis* (L.) Osb.) propagated on sour orange rootstock) and those predominant in virulent isolates (additionally inducing seedling yellows and stem pitting in grapefruit or sweet orange) clustered separately. To gain further insight into the mechanisms of CTV evolution, we analyzed the genetic variation and phylogenetic relationships in seven gRNA regions of these isolates and sought for recombination between divergent sequence variants. Our analyses showed variable selection pressures along the gRNA and frequent recombination events. Apparently, CTV variants cluster within at least two evolutionarily divergent lineages.

Materials and Methods

Virus isolates

The CTV isolates used in this study were from Argentina (C-268-2, C-269-6 and C-270-3), Brazil (Barão B, Cald-CB, Galego 50 and Val-CB), France (K), Florida (T36 and T55), Israel (VT), Japan (T388), and Spain (T32, T300, T305, T312, T346, and T385), and their pathogenicity characteristics have been described (Sambade *et al.*, 2002). These isolates were classified into 5 biogroups according to symptoms induced in a panel of indicator hosts (Garnsey *et al*., 2005). In short, the K isolate is asymptomatic in all hosts (biogroup 0); isolates T32, T55 and T385, induce symptoms in Mexican lime (biogroup 1), and T300, T312, and T346 also cause decline of sweet orange

grafted on sour orange rootstock (biogroup 2); T36, Galego 50 and C-268-2 additionally induce seedling yellows (biogroup 3); while the remaining isolates, in addition to the latter symptoms, cause stem pitting on grapefruit (Barão B, C-269-6, C-270-3, and VT; biogroup 4) or on grapefruit and sweet orange (T305, T388, Cald-CB, and Val-CB; biogroup 5).

cDNA synthesis, cloning and sequencing

cDNA of regions located in genes *p33*, *p65*, *p61*, *p18*, *p13*, *p20*, and *p23* of the CTV gRNA was synthesized by reverse transcription (RT) and PCR amplification using double-stranded RNA (dsRNA)-rich preparations (Moreno *et al.*, 1990) as template and appropriate primers (Supplementary data 1). Primers amplify 41%-44% of *p33*, *p65* and *p61* genes, 76% of *p18* and 87-99% of *p13*, *p20* and *p23*. RT-PCR was performed in a 25 µL reaction mix containing: 20 mM Tris-HCl, pH 8.4, 50 mM KCl, 500 µg/mL bovine serum albumin, 3 mM $MgCl_2$, 4 mM each of dATP, dCTP, dGTP, and dTTP, 1 µM of each primer, 20 U of SuperScript™ II reverse transcriptase, 1 U of RNaseOut and 1 U of Taq DNA polymerase (Invitrogen, USA). The reaction proceeded in an air thermal cycler (Idaho Technologies, USA) using 30 min at 46 ºC for RT, 2 min at 94 ºC, 40 cycles of 5 s at 94 ºC, 5 s at 55 ºC and 30 s at 72 ºC, and a final step of 2 min at 72 ºC. The resulting RT-PCR products were cloned in the pGEM-T vector (Promega, USA) (Sambrook *et al.*, 1989).

The sequence variants predominant in each CTV isolate were selected by single-strand conformation polymorphism (SSCP) analysis (Rubio *et al.*, 2001). For this purpose, ten clones from each cDNA product were PCR-amplified as described above, and the DNA synthesized was SSCP-analyzed in the same gel as the RT-PCR product from which the clones were obtained (Sambade *et al.*, 2003). Clones whose DNA strands co-migrated with the most intense DNA bands of the starting RT-PCR product were sequenced. The nucleotide sequence of the cDNA clones selected was determined in both directions with an ABI PRISM 3100 DNA sequence analyzer (Applied Biosystems). The sequence of the virulent CTV isolate NUagA from Japan (AB046398) was used for comparisons.

The nucleotide sequences obtained have been deposited in GenBank under accession numbers FM955890- FM956002.

Sequence analyses

Nucleotide sequences were translated to proteins using GENEDOC and multiple protein alignments were performed with MUSCLE program (Karl & Hugh, 1997; Edgar, 2004). Nucleotide alignments were then obtained by concatenating codons with the amino acid alignment as guide. Sites containing insertions were removed from all subsequent analyses.

Nucleotide substitution models for different CTV regions were inferred using the model selection tool available at the DATAMONKEY server (http://www.datamonkey.org) of the HYPHY package (Kosakovsky-Pond & Frost, 2005a). Genetic distances and substitution parameters were calculated by maximum likelihood with the TREE-PUZZLE 5.2 program (Schmidt *et al.*, 2002) assuming that sites had heterogeneous substitution rates described by a gamma distribution with eight classes. The best amino acid substitution model (lowest *AIC* value among competing models) was inferred with PROTTEST (Abascal *et al.*, 2005), available at http://darwin.uvigo.es/software/prottest.html. Detection of codons under selection was done using the fixed effects maximum likelihood (FEL) method of the HYPHY package (Kosakovsky-Pond & Frost, 2005b). Recombination was detected with the GARD program available at the DATAMONKEY server using the HKY85 substitution model and a beta-gamma distribution with four classes for rate variation. Further confirmation of recombination events and identification of parental sequences were performed with the RDP3 package (Martin *et al.*, 2005a) that incorporates the recombination-detecting algorithms GENECONV (Padidam *et al.*, 1999), BOOTSCAN (Salminem *et al.,* 1995; Martin *et al.*, 2005b), MAXCHI (Smith, 1992; Posada & Crandall, 2001), CHIMAERA (Posada & Crandall, 2001), SISCAN (Gibbs *et al.*, 2000), 3SEQ (Boni *et al.*, 2007), and RDP (Martin & Rybicki, 2000), using their default parameter values. Average nucleotide distances of CTV isolates were calculated using MEGA 4.0 software (Tamura *et al*., 2007), after testing homogeneity of pattern substitution among lineages. Evolutionary distances were estimated by the composite maximum likelihood

method assuming that substitution rates among sites fitted a gamma distribution.

Phylogenetic analysis

Protein maximum likelihood trees were inferred with PROTTEST (Abascal *et al.*, 2005); and significance for the nodes was estimated with 1000 bootstrap replicates using PHYML program (Guindon & Gascuel, 2003), available at http://phylemon.bioinfo.cipf.es/ (Tárraga *et al.,* 2007). Nucleotide maximum likelihood trees were constructed by the sequential addition method using HYPHY (Kosakovsky-Pond *et al.*, 2005), with the HKY85 substitution model and a gamma distribution with six classes for rate heterogeneity. Topology comparisons were performed with Shimodaira & Hasegawa (1999) test implemented in the DNAML program of the PHYLIP 3.67 package (Felsenstein, 2005). Phylogenetic trees were drawn using MEGA 4.0 (Tamura *et al.*, 2007). The ratio of non-synonymous to synonymous substitution rates in different branches of maximum likelihood trees was estimated using the codon-based genetic algorithm implemented in the GA-BRANCH program (Kosakovsky-Pond & Frost, 2005c) available at the DATAMONKEY server. Split-decomposition analysis of concatenated CTV sequences was performed using the SPLITSTREE program with default parameters (Huson, 1998).

Results

Phylogenetic relationships between CTV isolates

A predominant sequence variant was observed for each isolate and gRNA region, except for isolates Galego 50, that showed two p13 variants, and C-268-2 that had three variants in p33 and p61. Deduced amino acid sequences were used to infer phylogenetic trees, including the NUagA sequence in these regions for comparison (Fig. 1). At least two clades were observed in most regions: one of them (clade I), comprising the sequence variants of mild isolates T32, T55, T300, T312, and T385, and T346 in p20, was supported by bootstrap values of 86.3% (p33), 67.5% (p65), 96.1% (p61), 65.5% (p18), 75.9% (p13),

99.5% (p20) and 94.4% (p23). A second clade (clade II) enclosing biogroup 5 plus a variable set of isolates of biogroup 4, was supported by bootstrap values of 96.8% (p33), 87.2% (p23) 76.9% (p13) and 56.2% (p65), albeit in regions p33 and p65 a biogroup 3 isolate (Galego 50) was in the same cluster . Although support for clade II was less robust than for clade I, the virulent isolates T388, T305 and NUagA (NUagA type group) were closely related to each other and distantly related to clade I in all regions, forming a stable nucleus within clade II. Thus, phylogenetic relationships reflected to some degree pathogenecity characteristics of the isolates. In p18 all South American isolates, except Galego 50 grouped together into a distinct clade (clade III).

While isolates defining clade I and NUagA type isolates clustered together regardless the genomic region, other isolates showed incongruent phylogenetic relationships for different regions (Fig. 1). The virulent isolates Cald-CB, Barão B and Val-CB clustered together, and closely related to NUagA type isolates, only in p33, p65 and p23. In p33, p65 and p61, isolate K was genetically closer to clade I than to NUagA, but in other regions it was located between them. The two major p13 variants found in Galego 50 were divergent, with the variant Galego 50A grouping with isolates C-270-3 and Barão B and the variant Galego 50B being closer to NUaGA group. Similarly, for isolate C-268-2, one of the three p33 variants was close to NUaGA and the other two were separated from clade I and NUaGA, and two of the three p61 variants were close to clade I, and the third was separated from both groups (Fig. 1).

The incongruent phylogenetic relationships observed for some isolates in different genomic regions suggested that their gRNA might have originated from recombination events between diverged sequences.

Genetic variation and selective pressures in different genomic regions

The average nucleotide distance for p33 (0.1641±0.0196) (Table 1) was significantly higher than those of regions p65, p18, p13 p20, and p23 (ranging from 0.0741±0.0240 to 0.1145±0.0188), and average distance for p13 (0.0741±0.0240) was lower than those of p33 and p61 (Model II ANOVA: $F_{6,131}$=48.4279, p<0.0001 and Tukey-Kramer *post hoc* test at 95% confidence level). To evaluate the selective constraints operating in each region, codons

under selection were detected using the FEL method (Supplementary data 2). Since d_N and d_S estimates are sensitive to the effect of recombination, we preliminarily screened the different genomic regions with the GARD tool and found significant recombination signals in *p61*, *p20* and *p23* genes (positions 283, 252 and 239, respectively). Alignments for these three genes were split in the corresponding non-recombinant regions. The number of negatively selected sites and the mean normalized d_N–d_S values for each genomic region, after correcting the significance *p* values for multiple comparisons of the same null hypothesis using a false discovery rate (FDR) of 5%, are in Table 1. A total of 150 codons were subjected to significant purifying selection, with ratios of negatively selected codons ranging from 10% for p20 to 15.87% for p33. Although these ratios were not significantly different (χ^2=5.145, 6 d.f., p=0.525), the strength of negative selection estimated by the mean normalized d_N–d_S value did differ among genomic regions (Kruskal-Wallis test: H=53.4090, 6 d.f., p<0.0001), this difference being entirely driven by the less negative d_N–d_S estimated for p33 (–2.1098±0.2505) relative to the average value for the other six regions (–7.6334±0.4393) (Dunn's *post hoc* test p<0.05). Only codons 203 (d_N–d_S=3.6789) and 244 (d_N–d_S=6.6412) in p61 showed a significant signature of positive selection.

Branch-specific analysis of the ratio of non-synonymous to synonymous substitution rates in CTV phylogenies

To get deeper insights into the selective pressures acting at the protein level during the CTV evolution, branch-specific ratios of d_N to d_S rates (ω) were estimated using a genetic algorithm that identified models with variable numbers of ω categories per lineage that fitted better to data than the single-ratio (all lineages evolving with equal ω) or the fully saturated (each lineage evolves at different ω) models (Table 2). Although ω values and the proportion of associated branches varied in different regions and periods of CTV diversification, most classes had ω<1 and most branches were assigned to these classes (100% in p61 and over 74% in the other regions). Figure 2b illustrates a simplified comparison of branch-specific ω values in phylogenetic

trees grouping these values in four selection categories: *i*) strong negative selection (ω<0.1); *ii*) moderate negative (0.1<ω<0.4); *iii*) weak negative (0.4<ω<1); and *iv*) positive selection (ω>1). In most cases, internal branches connecting the groups of mild and virulent isolates, particularly clade I and NUagA group were associated to ω<1, indicating that divergence of these genotypes occurred under negative selection pressure. In p65 purifying selection was strong for all internal branches, whereas in p33, p18 and p13, periods of strong and moderate selection alternated. While in p33 negative selection was moderate for lineages leading to mild and virulent groups and strong in the internal branch leading to the out-groups of C-270-3 and the ancestor of T36 and C-268-2B, the opposite was true in p18, with strong selection for lineages leading to the mild and virulent groups and moderate in the lineage leading to C-270-3 and other South American isolates. In p61, p20 and p23 selection was moderate for lineages leading to virulent or mild groups and to C-270-3, and in p23 it was weak or even positive after diversification of the cluster formed by K and clade I. Positive selection occurred during limited periods in all regions but p61, and except for p13, it was observed only in terminal branches, with highest frequency being observed in p33. For some isolates signature of positive or weak negative selection was detected in most regions, i.e., T385 (p33, p61, p13, p20, and p23), T312 (p33, p65, p61, p18, and p23) or VT (p33, p65, p61, and p18).

Frequent recombination events in CTV genomes

Sequences were first examined for recombination using the GARD tool that identifies recombination breakpoints when the likelihood of phylogenetic trees inferred for the partitioned alignments is significantly higher than that obtained for the non-partitioned alignment. Due to computational limits and to avoid arbitrary assembling of C-268-2 sequence variants, alignments of pair-wise concatenated regions corresponding to adjacent genes were used as input. A total of nine recombination breakpoints were detected: six of them located in the boundaries of different regions, suggesting recombination events somewhere between the analyzed fragments, and three located within p61, p20

and p23 regions (positions 283, 252 and 239 of the corresponding region). Since GARD does not require different topologies, a model containing partitions could outperform a non partitioned one if both share the same topology due to best fit of branch lengths. To test for topological differences, maximum likelihood trees for the non recombinant regions defined by GARD (HKY85 substitution model and sequential addition method for topology inference) were compared using the Shimodaira & Hasegawa (1999) test that compares the goodness of a set of competing phylogenetic trees to describe the evolution of a given alignment. Maximum likelihood estimates of the substitution parameters (transversion-transition rates ratio; shape parameter of the gamma distribution of substitution rates per site, and relative substitution rates) were inferred using TREE-PUZZLE. This analysis confirmed the differences in tree topology inferred for different genomic regions, although in p61, p20 and p23 the trees inferred from the entire region were not significantly worse than those inferred from each partition defined by GARD, suggesting that partitions established by GARD were mostly due to differences in branch length ($p<0.05$).

To assess the frequency and extension of recombination during CTV diversification, the seven genomic regions were concatenated and recombination events were identified using the RDP3 program that implements several recombination-detecting methods (Martin *et al.*, 2005a), using default setting parameters for the subset of fast detection methods GENECONV, BOOTSCAN, MAXCHI, CHIMAERA, SISCAN, 3SEQ and RDP. Isolate C-268-2 was excluded to avoid arbitrary assembling of its p33 and p61 variants. A total of 14 recombination events were detected by at least one of the methods, but only those predicted by at least four different methods were accepted (Fig. 2c) and assignment of parental and daughter sequences was confirmed by constructing maximum likelihood trees (Fig. 2b). For example, isolate C-269-6 grouped with NUagA group in regions p33, p65, p23 and p20, but with C-270-3 in p61 and p18 and as outgroup in p13 (Fig. 2b). A recombination involving a NUagA ancestor as major parental, and the p61 and p18 regions from a C-270-3 ancestor, was identified by the seven methods used (Fig. 2c). Recombination events involving the same parentals were also detected for Cald-CB and Val-CB in p18, Barão B in p18 and p13 and Galego 50A in p13. Isolates Barão B and Val-CB, that grouped together and distant from other isolates in p20,

presented in this region an additional recombination between a NUagA-type and an unknown ancestor not represented in the alignment. Galego 50 was closely related to VT in most regions (p33, p65, p61, p18, and p20), but not in p23 and p13. In this latter region the variant Galego 50B grouped with VT and the variant Galego 50A with C-270-3, as a result of a recombination between VT and C-270-3 ancestors (Fig. 2b and 2c). In p23 Galego 50 was close to isolates C-270-3 and C-268-2, and RDP3 predicted a recombination between a Cald-CB ancestor as major parental, and an unknown isolate providing the 3' end of p20 and p23. The non assignment of C-270-3 as minor parental was due to the use of UPGMA trees in RDP3 default analysis (not shown), but inspection of maximun likelihood trees indicated that likely VT and C-270-3 are the major and minor parentals, respectively (Fig. 2b and 2c).

Isolate T346 was close to clade I in regions p33, p61 and p23, to C-270-3 in p65, and in intermediate positions in p20, p18 and p13, an incongruence that was compatible with a recombination event in the p65 region between T312 and C-270-3 ancestors as major and minor parental, respectively. The phylogenetic relationships of isolates T346 and T36 widely varied among regions: while in p33, p65 and p23 both isolates were divergent (genetic distances from 0.1180 to 0.2447), in p61, p18 and p13 they were closely related (genetic distances from 0.0090 to 0.0535); and two recombination events involving regions p61 and p13 were detected between ancestors of these isolates. In p13 and p18 T346 and K clustered in the same group. These results and the genetic distances observed in these regions are compatible with a recombination between K and T36 in p18 and p13 and later recombination between T346 and T36 in p13, and with other possibly older recombination between T346 and T36 in p61.

Finally, isolate C-268-2 contained three major diverged variants in p33 and p61 but it was monomorphic for the other regions. Furthermore, it clustered with C-270-3 in p65, p18, p20 and p23 regions, but close to clade I in p13, suggesting that this isolate may be the result of multiple recombination events.

A phylogenetic network for CTV isolates

Due to the recombinant nature of CTV genomes, bifurcating phylogenetic trees do not properly reflect the actual evolutionary history of different isolates, since one isolate may be directly linked to more than one ancestral sequence. To provide a more accurate representation of those relationships a phylogenetic network was constructed from the concatenated alignment of the seven regions by the split-decomposition method implemented in SPLITSTREE (Huson, 1998). Again isolate C-268-2 was excluded for the reasons given above (Fig. 3).

The largest splits divided CTV isolates into two groups: one formed by isolates of biogroups 0, 1 and 2 (with the exception of T346) and the other by isolates of biogroups 3, 4 and 5 (with the exception of T36). Within the mild group, isolates T55 from Florida, and T312, T300 and the ancestor of T32 and T385 from Spain showed a radiation pattern. In the second group including isolates from South America, Japan, Israel and Spain, Cald-CB and the ancestors of Barão B and Val-CB, and NUagA group, diverged from a common ancestor in a star-like manner, whereas the other isolates had a more complex phylogeny. Isolate C-269-6 was connected to the node joining isolates of biogroup 5 and to the ancestor of C-270-3, consistent with its recombinant nature (between NUagA and C-270-3) revealed by RDP3 analysis. Galego 50A was connected to Galego 50B and to the common ancestor of both variants and VT, also in agreement with recombination analysis. C-270-3 was connected to the T346 and C-269-6 ancestors giving further support to previous finding that these isolates likely arose from a recombination between a mild (T346) or a severe (C-269-6) major ancestor and C-270-3. Finally, the ancestors of isolates K, T36 and T346, that showed variable phylogenetic relationships (Fig. 2b), were interconnected in the network in a complex pattern (Fig. 3), supporting recombination between them detected by RDP3 (Fig. 2c).

DISCUSSION

The genetic variation and evolutionary factors shaping CTV populations were studied comparing the predominant sequence variants in seven genomic regions of 18 isolates from different geographical origin and pathogenicity characteristics. It was assumed that pathogenicity would be largely associated to the major sequence variant since: *i*) virions obtained from a cDNA clone of

the major component of isolate T36 induced the symptoms characteristics of this isolate (Satyanarayana *et al.*, 1999, 2001), and *ii*) in citrus plants successively co-inoculated with a mild and a virulent CTV isolate, symptom onset was associated with predominance of the sequence variant characteristic of the virulent isolate (Sambade *et al.*, 2002, 2007).

Analysis of selective pressures acting on different codons suggests that all regions examined are mostly subjected to purifying selection with only two codons in p61 being positively selected. Purifying selection measured as normalized d_N–d_S showed similar proportion of selected sites and selection intensity among regions, except for p33 that had less intense selection. Less negative d_N–d_S value of p33 indicates that in this region more nonsynonymous substitutions are allowed, coherently this region also had higher evolutionary distances. Although net selection pressure was similar in the other regions, branch-specific analysis showed variable strength of selection depending on the genomic region and period of CTV diversification, i.e., this pressure was strong in the diversification of clade I and NUagA type isolates in p65 but moderate or weak in p23.

Data available on the functional domains of CTV proteins are still limited, and the function of proteins p33, p13 and p18, that are dispensable for CTV infection and movement (Tatineni *et al.*, 2008), is unknown. Albeit selective pressures are less intense in p33, the fraction of selected sites is similar in all regions, indicating selective constraints to amino acid changes and providing candidate positions to test in functional studies. Negative selection was expected in p65, p61, p20, and p23, considering their role in the CTV biology. Genes *p65* and *p61* are part of a conserved five-gene block encoding proteins involved in virion assembly and movement (Satyanarayana *et al.*, 2000; Dolja *et al.*, 2006). The p65 region analyzed here encodes 5 of the 8 motifs conserved among HSP70 proteins (Pappu *et al.*, 1994). Proteins p20 and p23 act as silencing suppressors (Lu *et al.*, 2004). Within the p20 region, amino acids I38, Y113, R130, L137, S141, and L159 are strictly conserved among silencing suppressors of closteroviruses (Reed *et al.*, 2003). Gene *p23*, that was completely sequenced, contains the RNA-binding domain and putative zinc finger required for asymmetrical accumulation of positive and negative RNA strands, with conserved residues C68, C71, H75, and C85 being involved in this

463 activity (López *et al.*, 2000; Satyanarayana *et al.*, 2002). These amino acids
464 were encoded by invariable codons in all CTV isolates, with the exception of
465 L137 in p20 and C71 in p23, that were found under significant purifying
466 selection using cut-off values of p=0.5 and p=0.1, respectively. The observation
467 that protein p23 is a pathogenicity determinant in citrus (Ghorbel *et al.*, 2001;
468 Fagoaga *et al.*, 2005) is consistent with separation of the mild and virulent CTV
469 isolates in phylogenetic analysis.
470 Phylogenetic analysis showed that the mild isolates T32, T55, T300, T312,
471 and T385 form a clade (clade I), supported by bootstrap values >70% in six
472 regions, thus defining a CTV lineage which also includes isolate T30 from
473 Florida and probably others from Colombia and Taiwan (Ruiz-Ruiz *et al.*, 2006;
474 Albiach-Martí *et al.*, 2000b). A second clade (clade II) including a variable set of
475 virulent isolates was supported by bootstrap values >75% in three regions,
476 suggesting a higher recombination frequency for those isolates. Within clade II,
477 isolates T305, T388 and NUagA were closely related to each other and distantly
478 related to clade I in all regions, and together with other virulent isolates showed
479 a star-like evolution pattern from a common ancestor in the phylogenetic
480 network. These data suggest that virulent isolates could represent a second
481 CTV lineage distantly related to clade I that would also include isolate T318A
482 from Spain (Ruiz-Ruiz *et al.*, 2006). A third clade including Brazilian and
483 Argentinean isolates genetically related with isolate C-270-3 was observed in
484 several genomic regions. Sequence comparisons showed that several clones of
485 two Colombian isolates released on GeneBank were closely related to C-270-3,
486 while others were closely related to severe isolates in *p23*. However, the latter
487 were related to clade I or to severe isolates but not to C-270-3 in *p33* (not
488 shown). Although no complete gRNA sequence from South American isolates is
489 currently available, these findings are compatible with the diversification of a
490 third CTV lineage and frequent recombinations in this area. All lineages
491 included closely related CTV variants from distant locations, a circumstance
492 that, together with the radiation pattern observed for diversification of some CTV
493 isolates, provide further support to previous suggestion that genetic flow has
494 likely occurred (Rubio *et al.*, 2001).
495 Most isolates showed incongruent phylogenetic relationships in different
496 regions, suggesting frequent recombination events, a possibility that was

supported by recombination-detecting methods and by a split-decomposition phylogenetic network. Nine of the 19 isolates compared were recombinant, particularly, most isolates from Brazil (between ancestors of NUagA or VT and C-270-3) and Argentina (C-268-2 appears as a mosaic of mild, severe and C-270-3-type sequence variants), in agreement with results obtained for these and other Argentinean isolates (Iglesias *et al.*, 2008). Recombination involving p18 was so frequent that all South American isolates except Galego 50 formed in this region a monophyletic group diverged from the other isolates. A similar grouping was described for genes *p25* and *p27*, located upstream *p18*, for Argentinean isolates (Iglesias *et al*., 2008). The variable position of K, T346 and T36 in phylogenetic trees, their lack of association with other isolates and the network topology suggest that they might represent CTV lineages with a more complex history involving recombination events among their ancestors and possibly with genotypes unrelated with those analyzed here.

The phylogenetic network grouped CTV isolates in two major clusters separated by long splits: one of them comprising isolates of biogroups 0, 1 and 2 (except for isolate T346) that induce mild to moderate symptoms in the most sensitive hosts and cause symptomless infections in grapefruit and sweet orange seedlings, and the other, including isolates of biogroups 3, 4 and 5 (except for isolate T36) that incite severe symptoms in sensitive hosts and are also pathogenic on grapefruit or sweet orange. Sequence separation between mild and virulent CTV isolates is consistent with a different host response after infection, as indicated by specific changes induced in the citrus transcriptome by both types of isolates (Gandía et al., 2007). This separation suggests that virulence might be an important evolutionary factor shaping CTV populations.

Homologous recombination seems a common process in some plant RNA viruses, particularly potyviruses (Chare & Holmes, 2005) and bromoviruses (Codoñer & Elena, 2008). It has been postulated that recombination might prevent accumulation of deleterious mutations in small populations and/or allow a faster adaptation to changing environments (Lai, 1992; Roossinck, 1997; García-Arenal *et al.*, 2001). On the other hand, simulation studies showed that recombination in RNA viruses is more likely to create combinations of deleterious mutations than purge them from genomes, thus causing fitness reductions (Holmes, 2003). Our results and previous analyses (Hilf *et al*., 2009)

suggest that RNA recombination is a major factor in CTV variation and likely plays a role in its evolution. Potential factors contributing to frequent recombination in CTV include: *i*) dispersal of diverged CTV genotypes in the same area by movement of infected buds, *ii*) the long life of citrus trees providing many opportunities for repeated infections with diverged sequence variants (Rubio *et al.*, 2001; Weng *et al.*, 2007; Gomes *et al.*, 2008), *iii*) the presence in infected cells of multiple viral RNA species produced during CTV replication that likely facilitates recombination events (Hilf *et al.*, 1995; Yang *et al.*, 1997; Ayllón *et al.*, 1999; Gowda *et al.*, 2003), and *iv*) the large size of CTV genome that may accumulate 2-3 mutations per genome and replication round. Recombination might help maintaining functional genomes even if many nonfunctional recombinants were produced (Allison *et al.*, 1990).

Summarizing, analyses of genetic variation in the 3' half of CTV genome suggest that at least two different lineages might have evolved, and that selective pressure against amino acid changes, gene flow, homologous recombination and perhaps virulence, may be important factors in CTV evolution and in shaping CTV populations.

Acknowledgements

We are indebted to S. M. Garnsey (University of Florida-C.R.E.C., Lake Alfred) for kindly providing freeze-dried citrus tissue infected with isolates T36, T55, K, and VT from the international collection of exotic citrus pathogens maintained at the quarantine facilities of the USDA in the Beltsville Agricultural Research Center (Maryland, USA), and to S. Gago-Zachert (Universidad de La Plata, Argentina) and M. A. Machado (Centro APTA Citros "Sylvio Moreira", Cordeirópolis, SP, Brazil) for providing dsRNA-rich preparations of the Argentinean and Brazilian isolates, respectively. We are also thankful to M. E. Martínez and M. Boil for technical assistance in the laboratory and to J. Piquer for excellent care of plants. S. Martín and A. Sambade were recipient of fellowships from the Spanish Ministerio de Ciencia e Innovación and Generalitat Valenciana, respectively. This work was supported in part by grants AGL2004-05099/AGR and AGL2007-61885/AGR (work at IVIA) and BFU2006-14819-C02-01/BMC (work at IBMCP) from the Ministerio de Ciencia e Innovación.

Reference List

Abascal , F., Zardoya R. & Posada, D. (2005). ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21,** 2104-2105.

Albiach-Martí, M. R., Guerri J., Hermoso de Mendoza, A., Laigret F., Ballester-Olmos, J. F. & Moreno P. (2000a). Aphid transmission alters the genomic and defective RNA populations of *Citrus tristeza virus* isolates. *Phytopathology* **90,** 134-138.

Albiach-Martí, M. R., Mawassi, M., Gowda, S., Satyanarayana, T., Hilf, S. Shanker, M. E., Almira, E. C., Vives, M. C., López, C. & other authors (2000b). Sequences of *Citrus tristeza virus* separated in time and space are essentially identical*. J Virol* **74,** 6856-6865.

Allison, R., Thompson, C. & Ahlquist, P. (1990). Regeneration of a functional RNA virus genome by recombination between deletion mutants and requirement for *Cowpea chlorotic mottle virus* 3a and coat genes for systemic infection. *Proc Natl Acad Sci U S A* **87,** 1820-1824.

Ayllón, M. A., López, C., Navas-Castillo, J., Mawassi, M., Dawson, W. O., Guerri, J., Flores, R. & Moreno, P. (1999). New defective RNAs from *Citrus tristeza virus*: evidence for a replicase-driven template switching mechanism in their generation. *J Gen Virol* **80,** 817-821.

Ayllón, M. A., Rubio, L., Sentandreu, V., Moya, A., Guerri, J. & Moreno, P. (2006). Variations in two gene sequences of *Citrus tristeza virus* after host passage. *Virus Genes* **32,** 119-128.

Boni, M. F., Posada, D. & Feldman, M. W. (2007). An exact nonparametric method for inferring mosaic structure in sequence triplets. *Genetics* **176,** 1035-1047.

Chare, E. R. & E. C. Holmes. (2005). A phylogenetic survey of recombination frequency in plant RNA viruses. *Arch Virol* **151,** 933-946.

Codoñer, F. M. & Elena, S. F. (2008). The promiscuous evolutionary history of the *Bromoviridae* family. *J Gen Virol* **89,** 1739-1747.

Dolja, V. V., Kreuze, J. F. & Valkonen, J.P.T. (2006). Comparative and functional genomics of closteroviruses. *Virus Res* **117,** 38–51.

Domingo, E. & Holland, J. J. (1994). Mutation rates and rapid evolution of RNA viruses. In *The evolutionary biology of viruses*, pp. 161-184. Edited by S. S. Morse. Raven Press. USA : New York.

D'Urso, F., Ayllón, M. A., Rubio, L., Sambade, A., Hermoso de Mendoza, A., Guerri, J. & Moreno, P. (2000). Contribution of uneven distribution of genomic RNA variants of *Citrus tristeza virus* (CTV) within the plant to changes in the viral population following aphid transmission. *Plant Pathol* **49,** 288-294.

Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *Bioinformatics* **5,** 113.

Escriu, F., Fraile, A. & García-Arenal, F. (2000). Evolution of virulence in natural populations of the satellite RNA of *Cucumber mosaic virus*. *Phytopathology* **90,** 480-485.

Fagoaga, C., López, C., Moreno, P., Navarro, L., Flores, R. & Peña, L. (2005). Viral-like symptoms induced by the ectopic expression of the p23 gene of *Citrus tristeza virus* are citrus specific and do not correlate with the pathogenicity of the virus strain. *Mol Plant-Microbe Interact* **18,** 435-445.

Febres, V. J., Ashoulin, L., Mawassi, M., Frank, A., Bar-Joseph, M., Manjunath, K. L., Lee, R. F. & Niblett, C. L. (1996). The p27 protein is present at one end of *Citrus tristeza virus* particles. *Phytopathology* **86,** 1331-1335.

Felsenstein, J., 2005. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Fernández-Cuartero, B., Burgyan, J., Aranda, M. A., Salanki, K., Moriones, E. & García-Arenal, F. (1994). Increase in the relative fitness of a plant virus RNA associated with its recombinant nature. *Virology* **203,** 373-377.

Gandía, M., Conesa, A., Ancillo, G., Gadea, J., Forment, J., Pallás, V., Flores, R., Duran-Vila, N., Moreno, P. & Guerri, J. (2007). Transcriptional response of *Citrus aurantifolia* to infection by *Citrus tristeza virus*. *Virology* **367,** 298–306.

García-Arenal, F., Fraile, A. & Malpica, J. M. (2001). Variability and genetic structure of plant virus populations. *Annu Rev Phytopathol* **39,** 157-186.

García-Arenal, F. & McDonald, B. A. (2003). An analysis of the durability of resistance to plant viruses. *Phytopathology* **93,** 941-952.

Garnsey, S. M., Civerolo, E. L., Gumpf, D. J., Paul, C., Hilf, M. E., Lee, R. F., Brlansky, R. H., Yokomi, R. K., & Hartung, J. S. (2005). Biological characterization of an international collection of *Citrus tristeza virus* (CTV) isolates. In *Proceedings of the 16th Conference of the International Organization of Citrus Virologists*, pp. 75-93, Edited by M. E. Hilf, N. Duran-Vila & M. A. Rocha-Peña. Riverside, CA: IOCV. http://www.ivia.es/iocv/

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

Gomes, C. P., Nagata, T., de Jesus, W. C., Neto, C. R., Pappas, G. J. & Martin, D.P. (2008). Genetic variation and recombination of RdRp and HSP 70h genes of *Citrus tristeza virus* isolates from orange trees showing symptoms of citrus sudden death disease. *Virol J* **16,** 5-9.

Gibbs, M. J., Armstrong, J. S. & Gibbs, A. J. (2000). Sister-Scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. *Bioinformatics* **16,** 573-582.

Gowda, S., Satyanarayana, T., Davis, C. L., Navas-Castillo, J., Albiach-Martí, M. R., Mawassi, M., Valkov, N., Bar-Joseph, M., Moreno, P. & Dawson, W. O. (2000). The p20 gene product of *Citrus tristeza virus* accumulates in the amorphous inclusion bodies. *Virology* **274,** 246-254.

Gowda, S., Ayllón, M. A., Satyanarayana, T., Bar-Joseph, M. & Dawson W. O. (2003). Transcription strategy in a Closterovirus: a novel 5'-proximal controller element of *Citrus tristeza virus* produces 5'- and 3'- terminal subgenomic RNAs and differs from 3' open reading frame controller elements. *J Virol* **77,** 340-352.

Guindon, S. & Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52,** 696-704.

Hilf, M. E. (2009). Two distinct evolutionary pathways for *Citrus tristeza virus*: Recombination defines two gene modules and provides for increased genetic diversity in a narrow host range plant virus. In *Proceedings of the 17th Conference of the International Organization of Citrus Virologists* (in press).

Hilf, M. E., Karasev, A. V., Pappu, H. R., Gumpf, D. J., Niblett, C. L., Garnsey, S. M. (1995). Characterization of *Citrus tristeza virus* subgenomic RNAs in infected tissue. *Virology* **208,** 576-582.

Holmes, E. C. (2003). Error thresholds and the constraints to RNA virus evolution. *Trends Microbiol* **11,** 543-546.

Huson, D. H. (1998). Splits Tree: analyzing and visualizing evolutionary data. *Bioinformatics* **14,** 68-73.

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

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. & other authors (1995). Complete sequence of the *Citrus tristeza virus* RNA genome. *Virology* **208,** 511-520.

Karl, B. N. & Hugh B. N. (1997). GeneDoc: a tool for editing and annotating multiple sequence alignments. www.nrbsc.org/gfx/genedoc/ebinet.htm.

Kosakovsky-Pond & S. L., Frost, S. D. W. (2005a). Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics* **21,** 2531-2533.

Kosakovsky-Pond, S. L. & Frost, S. D. W. (2005b). Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol Biol Evol* **22,** 1208-1222.

Kosakovsky-Pond, S. L. & Frost, S. D. W. (2005c). A genetic algorithm approach to detecting lineage-specific variation in selection pressure. *Mol Biol Evol* **22,** 478-485.

Kosakovsky-Pond, S. L., Frost, S. D. W. & Muse S. V. (2005). HyPhy: hypothesis testing using phylogenies. *Bioinformatics* **21,** 676-679.

Lai, M. M. (1992). RNA recombination in animal and plant viruses. *Microbiol Rev* **56,** 61-79.

López, C., Navas-Castillo, J., Gowda, S., Moreno, P. & Flores, R. (2000). The 23-kDa protein coded by the 3'-terminal gene of *Citrus tristeza virus* is an RNA-binding protein. *Virology* **269,** 462-470.

Lu, R., Folimonov, A., Shintaku, M., Li, W. X., Falk, B. W., Dawson, W. O. & Ding, S. W. (2004). Three distinct suppressors of RNA silencing encoded by a 20-kb viral RNA genome. *Proc Natl Acad Sci U S A* **101,** 15742-15747.

Martin, D. & Rybicki, E. (2000). RDP: detection of recombination amongst aligned sequences. *Bioinformatics* **16,** 562-563.

Martin, D. P., Williamson, C. & Posada, D. (2005a). RDP2: recombination detection and analysis from sequence alignments. *Bioinformatics* **21,** 260-262.

Martin, D. P., Posada, D., Crandall, K. A. & Williamson, C. (2005b). A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. *AIDS Res Hum Retroviruses* **21,** 98-102.

Moreno, P., Guerri, J. & Muñoz, N. (1990). Identification of Spanish strains of *Citrus tristeza virus* by analysis of double-stranded RNA. *Phytopathology* **80,** 477-482.

Moreno, P., Ambrós, S., Albiach-Martí, M. R., Guerri, J. & Peña L. (2008). Plant diseases that changed the world - *Citrus tristeza virus*: a pathogen that changed the course of the citrus industry. *Mol Plant Pathol* **9,** 251-268.

Padidam, M., Sawyer, S. & Fauquet, C. M. (1999). Possible emergence of new geminiviruses by frequent recombination. *Virology* **265,** 218-225.

Pappu, H. R., Karasev, A. V., Anderson, E. J., Pappu, S. S., Hilf, M. E., Febres, V. J., Eckloff, R. M., McCaffery, M., Boyko, V. & Gowda, S. (1994). Nucleotide sequence and organization of eight 3' open reading frames of the *Citrus tristeza closterovirus* genome.*Virology* **199,** 35-46.

Posada, D. & Crandall, K. A. (2001). Evaluation of methods for detecting recombination from DNA sequences: Computer simulations. *Proc Natl Acad Sci U S A* **98,** 13757-13762.

Reed, J. C., Kasschau, K. D., Prokhnevsky, A. I., Gopinath, K., Pogue, G. P., Carrington, J. C. & Dolja, V. V. (2003). Suppressor of RNA silencing encoded by *Beet yellows virus*. *Virology* **306,** 203-209.

Roossinck, M. J. (1997). Mechanisms of plant virus evolution. Annu. Rev. *Phytopathology* **35,** 191-209.

Rubio, L., Ayllón, M. A., Kong, P., Fernández, A., Polek, M., Guerri, J., Moreno, P. & 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.

Ruiz-Ruiz, S., Moreno, P., Guerri, J. & Ambrós, S. (2006). The complete nucleotide sequence of a severe stem pitting isolate of *Citrus tristeza virus* from Spain, comparison with isolates from different origins. *Arch Virol* **151,** 387-398.

Salminen M. O., Carr J. K., Burke D. S., McCutchan, F. E. (1995). Identification of breakpoints in intergenotypic recombinants of HIV type 1 by Bootscanning. *AIDS Res Hum Retroviruses* **11,** 1423-1425.

Sambade, A., Rubio, L., Garnsey, S. M., Costa, N., Müller, G. W., Peyrou, M., Guerri, J. & Moreno, P. (2002). Comparison of viral RNA populations of pathogenically distinct isolates of *Citrus tristeza virus*, application to monitoring cross-protection. *Plant Pathol* **51,** 257-265.

Sambade, A., López, C., Rubio, L., Flores, R., Guerri, J. & Moreno, P. (2003). Polymorphism of a specific region in gene p23 of *Citrus tristeza virus* allows discrimination between mild and severe isolates. *Arch Virol* **148,** 2325-2340.

Sambade, A., Ambrós, S., López, C., Ruiz-Ruiz, S., Hermoso de Mendoza, A., Flores, R., Guerri J. & Moreno, P. (2007). Preferential accumulation of severe variants of *Citrus tristeza virus* in plants co-inoculated with mild and severe variants. *Arch Virol* **152,** 1115-1126.

Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989). Molecular Cloning: a Laboratory Manual, 2nd ed. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory.

Satyanarayana, T., Gowda, S., Boyko, V. P., Albiach-Martí, M. R., Mawassi, M., Navas-Castillo, J., Karasev, A. V., Dolja, V., Hilf, M. E. & other authors (1999). An engineered closterovirus RNA replicon and analysis of heterologous terminal sequences for replication. *Proc Natl Acad Sci U S A* **96,** 7433-7438.

Satyanarayana, T., Gowda, S., Mawassi, M., Albiach-Martí, M. R., Ayllón, M. A., Robertson, C., Garnsey, S. M. & Dawson, W. O. (2000). Closterovirus encoded HSP70 homolog and p61 in addition to both coat proteins function in efficient virion assembly. *Virology* **278,** 253-265.

Satyanarayana, T., Bar-Joseph, M., Mawassi, M., Albiach-Martí, M. R., Ayllón, M. A., Gowda, S., Hilf, M. E., Moreno, P., Garnsey, S. M. & Dawson, W. O. (2001). Amplification of *Citrus tristeza virus* from a cDNA clone and infection of citrus trees. *Virology* **280,** 87-96.

Satyanarayana, T., Gowda, S., Ayllón, M. A., Albiach-Martí, M. R., Rabindran, S. & Dawson, W. O. (2002). The p23 protein of *Citrus tristeza virus* controls asymmetrical RNA accumulation. *J Virol* **76,** 473-483.

765 **Satyanayanana, T., Gowda, S., Ayllón, M. A. & Dawson, W. O. (2004).**
766 Closterovirus bipolar virion: evidence for initiation of assembly by minor coat
767 protein and its restriction to the genomic RNA 5′ region. *Proc Natl Acad Sci USA*
768 **101,** 799–804.
769 **Schmidt, H .A., Strimmer, K., Vingron, M. & von Haeseler A. (2002).** TREE-
770 PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel
771 computing. *Bioinformatics* **18,** 502-504.
772 **Shimodaira, H. & Hasegawa, M. (1999).** Multiple comparisons of log-
773 likelihoods with applications to phylogenetic inference. *Mol Biol Evol* **16,** 1114-
774 1116.
775 **Smith, J. M. (1992).** Analyzing the mosaic structure of genes. *J Mol Evol* **34,**
776 126-129.
777 **Tamura, K., Dudley, J., Nei, M. & Kumar, S. (2007).** MEGA4, Molecular
778 Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* **24,**
779 1596-1599.
780 **Tárraga, J., Medina, I., Arbiza, L., Huerta-Cepas, J., Gabaldón, T., Dopazo,**
781 **J. & Dopazo, H. (2007).** Phylemon: a suite of web tools for molecular evolution,
782 phylogenetics and phylogenomics. *Nucl Acids Res* **35,** W38-42.
783 **Tatineni, S., Robertson, C. J., Garnsey, S. M., Bar-Joseph, M., Gowda, S. &**
784 **Dawson, W. O. (2008).** Three genes of *Citrus tristeza virus* are dispensable for
785 infection and movement throughout some varieties of citrus trees. *Virology* **376,**
786 297-307.
787 **Vives, M. C., Rubio, L., Sambade, A., Mirkov, T. E., Moreno P. & Guerri, J.**
788 **(2005).** Evidence of multiple recombination events between two RNA sequence
789 variants within a *Citrus tristeza virus* isolate. *Virology* **331,** 232-237.
790 **Weng, Z., Barthelson, R., Gowda, S., Hilf, M. E., Dawson, W.O., Galbraith,**
791 **D. W. & Xiong, Z. (2007).** Persistent infection and promiscuous recombination
792 of multiple genotypes of an RNA virus within a single host generate extensive
793 diversity. *PLoS ONE* **9,** e917.
794 **Worobey, M. & Holmes E. C. (1999).** Evolutionary aspects of recombination in
795 RNA viruses. *J Gen Virol* **80,** 2535-2543.
796 **Yang, G. A., Mawassi, M., Gofman, R., Gafny, R. & Bar-Joseph, M. (1997).**
797 Involvement of a subgenomic mRNA in the generation of a variable population
798 of defective *Citrus tristeza virus* molecules. *J Virol* **71,** 9800-9802.

Figure legends

Figure 1. Unrooted protein maximun likelihood phylogenetic trees of CTV genomic regions p33, p65, p61, p18, p13, p20 and p23. Isolates with congruent phylogenetic relationships are in bold italics. Circles beside the isolate names indicate their pathogenicity characteristics (biogroups 1-2, 4 and 5), with biogroups 0 (isolate K) and 3 left without circle. Geographical origin is indicated by double daggers (Florida), underlined names (South America), rectangular boxes (Spain), ovals (Japan) or no mark for isolates K (France) and VT (Israel). Scale bars indicate number of changes per position for a unit branch length. Bootstrap values for significance of nodes are indicated by asterisks (*** 90-100%, ** 70-89%, * 50-69%).

Figure 2. Recombination analysis of concatenated CTV sequences. a) Layout of the CTV genome with the regions analyzed indicated as black boxes. ORFs are represented by empty boxes with indication of the encoded protein. Motifs protease (Pro), methyl transferase (MT) and helicase (HEL) of the p349 protein are also indicated. b) Unrooted nucleotide maximum likelihood phylogenetic trees of seven CTV genomic regions (see Fig. 1). Isolates with congruent phylogenetic relationships are in bold italics. In each region, isolates detected as recombinant by RDP3 programs are highlighted with solid ovals and others showing incongruent phylogenetic relationships with dotted ovals. Values for branch-specific non-synonymous to synonymous class substitution rates (ω) are indicated with colors. c) Recombination hypotheses generated by at least four algorithms of the RDP3 program and further refined by inspection of maximum likelihood trees. Concatenated alignments are outlined at the top; long colored boxes represent CTV concatenated sequences (isolate code above the box) and internal pale colored segments indicate recombinant regions; the major parental for each recombinant sequence is indicated below the isolate code, and the minor, by short boxes below the pale colored segments.

Figure 3: Phylogenetic network of CTV isolates constructed by split-decomposition analysis. Circles beside the isolate names as in Figure 1.

Table 1. Nucleotide distances and frequency and strength of negative selection at protein level in different CTV genomic regions.

Genomic regions	Average nucleotide distance ± SD	Number of negatively selected codons	d_N–d_S ± SEM [†]
p33	0.1641±0.0196[**]	20 (0.1587)	–2.1098±0.2505[**]
p65	0.1091±0.0160	30 (0.1224)	–6.7162±0.6392
p61	0.1250±0.0160	35 (0.1471)	–7.2201±1.1737
p18	0.0955±0.0170	12 (0.1034)	–8.6499±1.2031
p13	0.0741±0.0240[*]	16 (0.1403)	–6.3229±1.0417
p20	0.1145±0.0188	16 (0.1000)	–7.8506±0.9940
p23	0.1018±0.0143	21 (0.1005)	–9.0407±1.7731

SD: Standard deviation for nucleotide distance computed by the bootstrap method (500 replicates).

[†] d_N–d_S: Average of normalized values of the difference between non-synonymous and synonymous substitutions of selected codons estimated by the fixed effects maximum likelihood method (FEL) and their standard errors (SEM).

[**] Statistically different to the rest of values (p<0.05).

[*] Statistically different to p33 and p61 values (p<0.05).

Table 2. Lineage-specific analysis of selective pressures in seven genomic regions of CTV.

	Genomic regions						
	p33	p65	p61	p18	p13	p20	p23
ω^*	1.029	1.327		21.785	∞	∞	∞
	(25%)	(14%)		(10%)	(12%)	(1%)	(2%)
	0.253	0.211	0.641	0.265			0.461
	(38%)	(22%)	(21%)	(64%)			(40%)
			0.122		0.208	0.155	0.143
			(67%)		(71%)	(54%)	(57%)
	0.087	0.043		0.013		0.030	
	(36%)	(50%)		(27%)		(44%)	
		0.000	0.000		0.000		
		(14%)	(12%)		(17%)		

* ω: d_N/d_S class ratios for non-synonymous to synonymous substitution rates; the proportion of branches assigned to each class is shown in parenthesis.

Figure 1

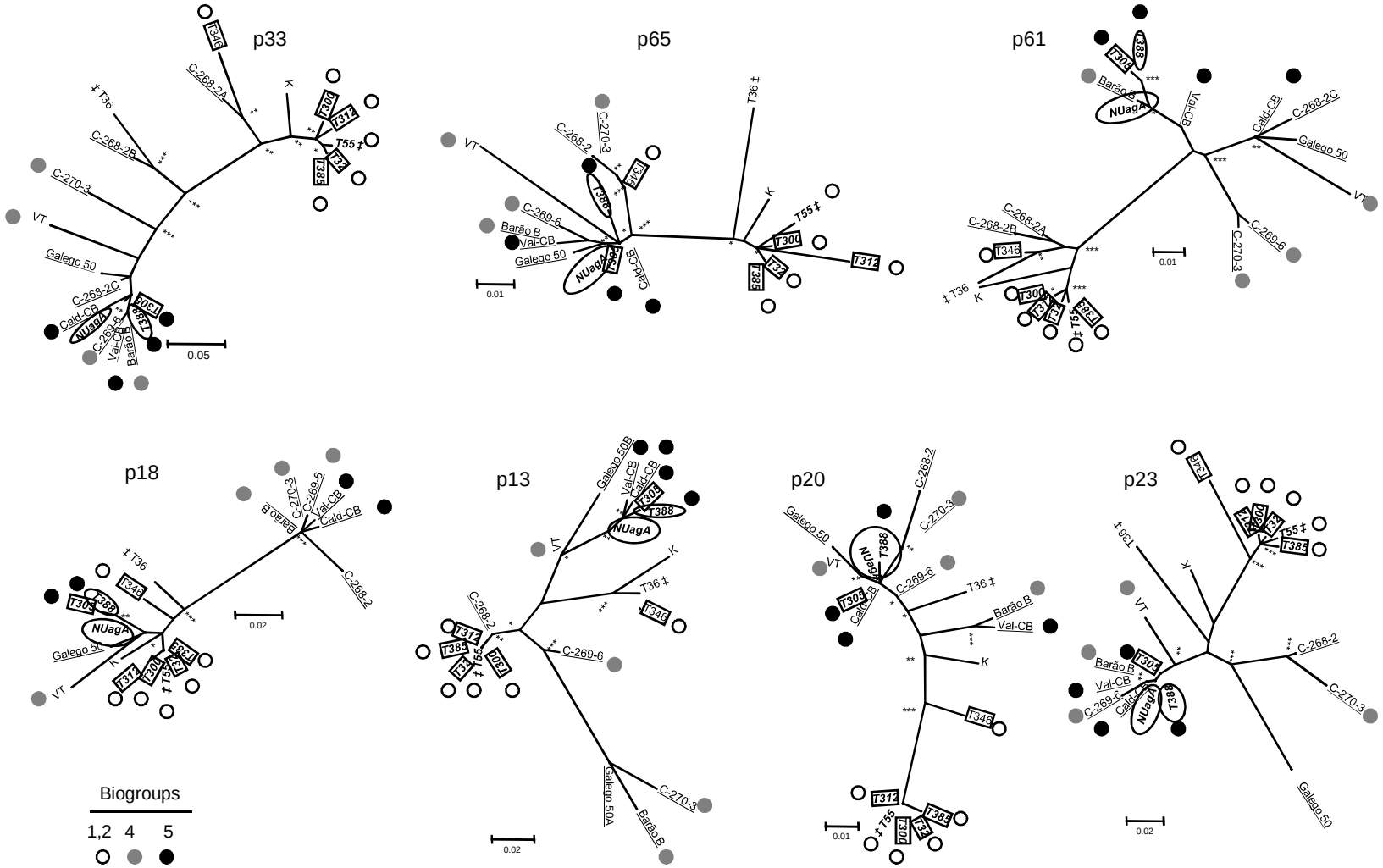


Figure 2

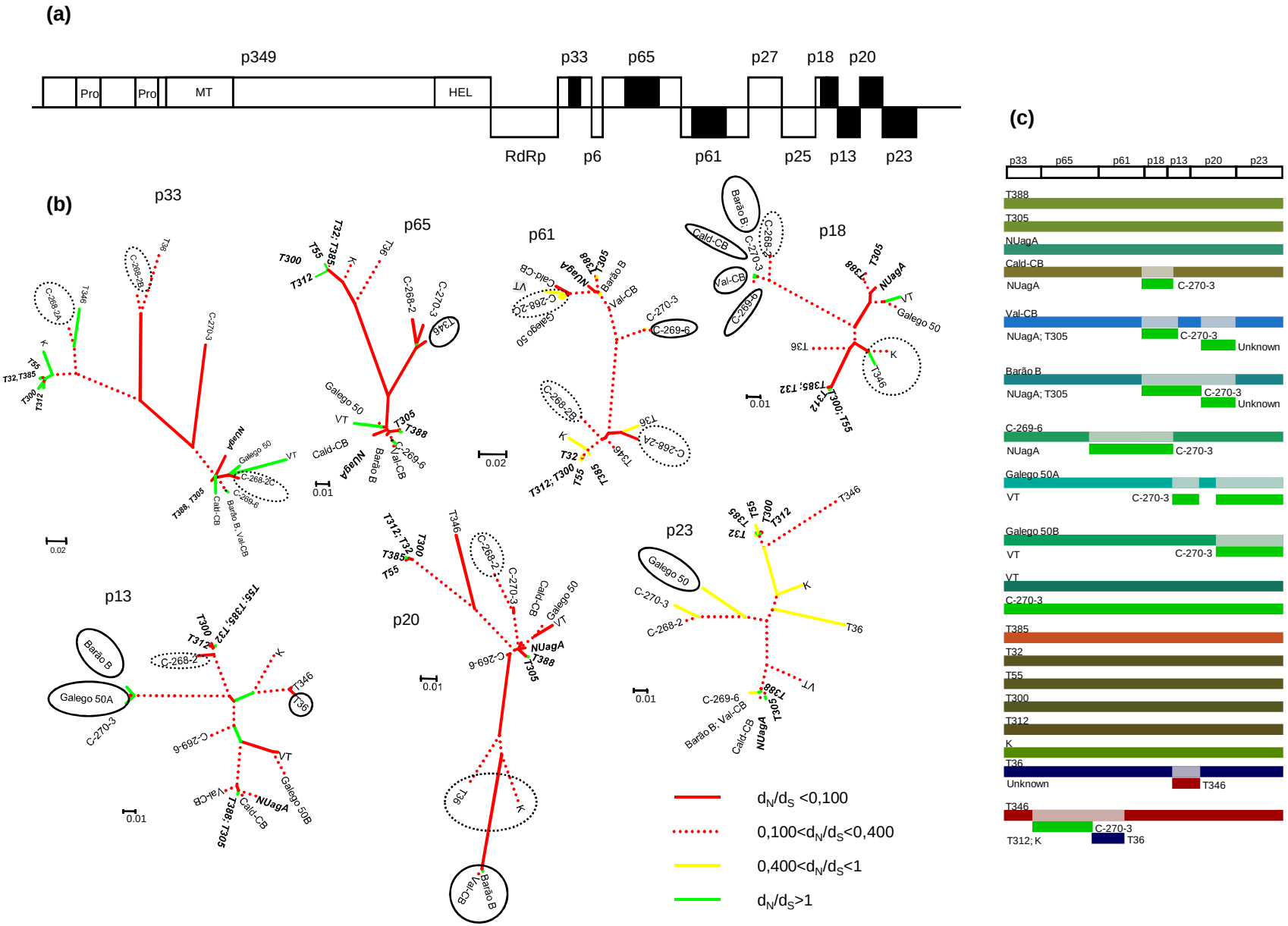
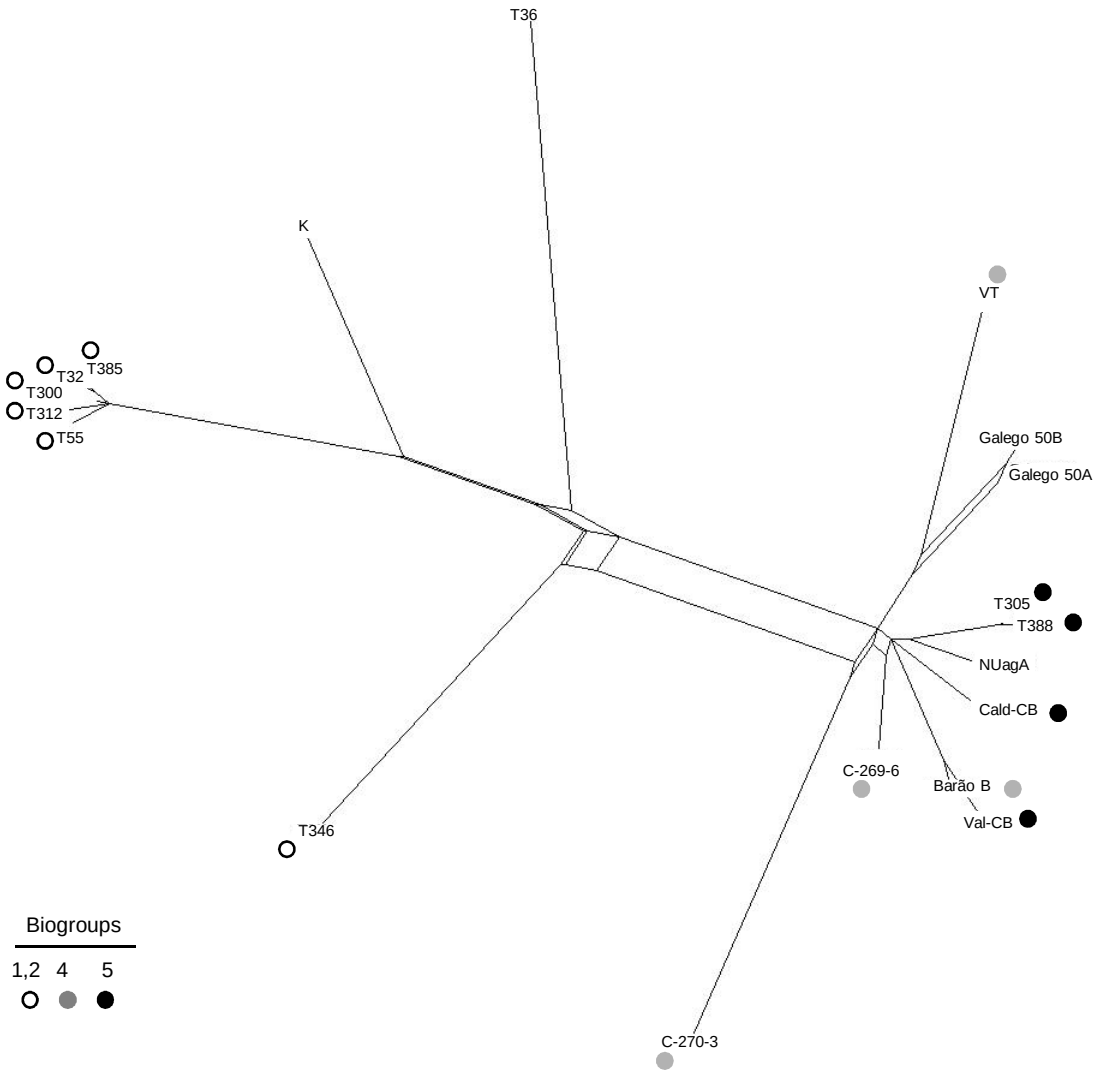

(a)
p349
Pro
Pro
MT
HEL
p33
p65
p27
p18
p20
RdRp
p6
p61
p25
p13
p23
(b)
p33
p65
p61
p18
p13
p20
p23
0.02
0.01
0.02
0.01
0.01
0.01
dN/dS <0,100
0,100<dN/dS<0,400
0,400<dN/dS<1
dN/dS>1
(c)
p33
p65
p61
p18
p13
p20
p23
T388
T305
NUagA
Cald-CB
NUagA
C-270-3
Val-CB
NUagA; T305
C-270-3
Unknown
Barão B
NUagA; T305
C-270-3
Unknown
C-269-6
NUagA
C-270-3
Galego 50A
VT
C-270-3
Galego 50B
VT
C-270-3
VT
C-270-3
T385
T32
T55
T300
T312
K
T36
Unknown
T346
T346
C-270-3
T312; K
T36

Figure 3


T36
K
VT
T385
T32
T300
T312
T55
Galego 50B
Galego 50A
T305
T388
NUagA
Cald-CB
C-269-6
Barão B
Val-CB
T346
C-270-3
Biogroups
1,2 4 5