CONTRIBUTION OF RECOMBINATION AND SELECTION TO MOLECULAR EVOLUTION OF CITRUS TRISTEA VIRUS

Susana Martín, Adrián Sambade, Luis Rubio, María C. Vives, Patricia Moya, José Guerri, Santiago F. Elena and Pedro Moreno

The genetic variation of Citrus tristeza virus (CTV) was analysed by 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 shows diverse patterns of molecular evolution for different genomic regions. A first clade composed of isolates that are 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 and p23. For regions p65, p61, p18, p13 and p23, a third clade that mostly included South American isolates could not be related to 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 also 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.

INTRODUCTION

Citrus tristeza virus (CTV) is a closterovirus, in the 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, positive-sense CTV genomic RNA (gRNA) is about 20 kb 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 is then locally spread by 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.) Burm. f.] or grapefruit (C. paradisi Macf.) seedlings (seedling yellows) or (iii) stunting, stem pitting and poor yield of different citrus varieties regardless of 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’Ursó 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 in the understanding of 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 the genetic variability of the virus (García-Arenal & Martínez, 2003).

We previously compared the predominant sequence variants of gene p23 from 18 CTV isolates with different origins and pathogenicity characteristics (Sambade et al., 2003). Phylogenetic analyses showed that sequence variants that are 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 (also inducing seedling yellows and stem pitting in grapefruit or sweet orange) clustered separately. To gain further insight into the mechanisms of CTV evolution, we analysed the genetic variation and phylogenetic relationships in seven gRNA regions of these isolates and investigated recombination between divergent sequence variants. Our analyses showed variable selection pressures along the gRNA and frequent recombination events. They also suggest that CTV variants cluster within at least two evolutionarily divergent lineages.

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); their pathogenic characteristics have been described previously (Sambade et al., 2002). These isolates were classified into five biogroups according to the symptoms induced in a panel of indicator hosts (Garnsey et al., 1990) as well as 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 RT-PCR amplification using double-stranded RNA (dsRNA)-rich preparations (Moreno et al., 1990) as template and appropriate primers (Supplementary Table S1, available in JGV Online). Primers amplified 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 BSA ml⁻¹, 3 mM MgCl₂, 4 mM each of dATP, dCTP, dGTP and dTTP, 1 µM each primer, 20 U SuperScript II reverse transcriptase, 1 U RNaseOut and 1 U Taq DNA polymerase (Invitrogen). The reaction was carried out in an air thermal cycler (Idaho Technologies) using the following conditions: 30 min at 46 °C for RT, 2 min at 94 °C, 40 cycles of 5 s at 94 °C, 5 s at 55 °C and 30 s at 72 °C, and a final step of 2 min at 72 °C. The resulting RT-PCR products were cloned in the pGEM-T vector (Promega) (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, 10 clones from each cDNA product were PCR-amplified as described above and the synthesized DNA was SSCP-analysed 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 sequences of the selected cDNA clones were determined in both directions with an ABI PRISM 3100 DNA sequence analyser (Applied Biosystems). The sequence of the virulent CTV isolate NUagA from Japan (GenBank accession no. AB046398) was used for comparisons.

**Sequence analyses.** Nucleotide sequences were translated to proteins using GenDoc and multiple protein alignments were performed with the MUSCLE program (Karl & Hugh, 1997; Edgar, 2004). Nucleotide alignments were then obtained by concatenating codons with the amino acid alignment as a 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). Codons under selection were detected 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. Recombination events and identification of parental sequences were confirmed further with the RDP3 package (Martin et al., 2005a) that incorporates the recombination-detecting algorithms GENECONV (Padidam et al., 1999), BOOTSCAN (Salminen et al., 1995; Martin et al., 2005b), MAXCHI (Smith, 1992; Posada & Crandall, 2001), CHIMAERA (Posada & Crandall, 2001), SISCAN (Gibbs et al., 2000), 3SEQ (Bonito 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); significance for the nodes was estimated with 1000 bootstrap replicates using the PHYLML program (Guindon & Gascuel, 2003), available at http://phylogen.-
RESULTS

Phylogenetic relationships between CTV isolates

A phylogenetic sequence variant was observed for each isolate and gRNA region, except for isolates Galego 50 (two p13 variants) and C-268-2 (three variants in p33 and p61). Deduced amino acid sequences were used to infer phylogenetic trees; the NUagA sequence was included in these regions for comparison (Fig. 1). At least two clades were observed in most regions. One of these (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), comprising 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), although, in regions p33 and p65, a biogroup 3 isolate (Galego 50) was in the same cluster. Although support was less robust for clade II 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 genomic regions, forming a stable nucleus within clade II. Thus, phylogenetic relationships reflect, to some degree, pathogenicity 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 of the genomic region, other isolates showed contrasting phylogenetic relationships for different regions (Fig. 1). The virulent isolates Cald-CB, Barão B and Val-CB clustered together and were 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 it was located between them in other regions. 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 the 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, two of the three p61 variants were close to clade I and the third was separate from both groups (Fig. 1).

The different phylogenetic relationships observed for some isolates in different genomic regions suggested that their gRNA might have originated from recombination events between divergent 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 that of the other regions (ranging from 0.0741 ± 0.0240 to 0.1250 ± 0.0160); the average distance for p13 (0.0741 ± 0.0240) was significantly lower than that of p33 and p61 (model II ANOVA: F_{6,131}=48.4279, P<0.0001; 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 Table S2, available in JGV Online). Since the rate of non-synonymous substitutions (dN) and the rate of synonymous substitutions (dS) 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. A total of 150 codons were subjected to significant purifying selection (Table 1), with ratios of negatively selected codons ranging from 10% for p20 to 15.87% for p33. Although these ratios were not significantly different (χ²=5.145, 6 d.f., P=0.525), the strength of negative selection estimated by the mean normalized dS–dN values differed among genomic regions (Kruskal–Wallis test: H=53.4090, 6 d.f., P<0.0001), this difference being entirely driven by the less negative dS–dN value 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 for amino acids 203 (dS−dN=3.6789) and 244 (dS−dN=6.6412) in p61 showed a significant level of positive selection.

Branch-specific analysis of the dN/dS ratio in CTV phylogenies

To gain deeper insights into the selective pressures acting at the protein level during the CTV evolution, branch-specific dN/dS ratios for dN to dS substitution 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 categories had ω<1 and most branches were assigned to
Fig. 1. Unrooted protein maximum-likelihood phylogenetic trees of CTV genomic regions p33, p65, p61, p18, p13, p20 and p23. Isolates with congruent phylogenetic relationships are in bold italics. Biogroups 1/2, 4 and 5 are indicated by circles beside the isolate names (empty, shaded and filled, respectively); biogroups 0 (isolate K) and 3 are left without a circle. Geographical origin: double dagger (†), Florida; underlined, South America; boxes, Spain; ovals, Japan; no mark, 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%).

S. Martín and others
Journal of General Virology 90

Downloaded from www.microbiologyresearch.org by
IP: 54.70.40.11
On: Fri, 23 Nov 2018 17:44:43
these categories (100% in p61 and >74% in the other regions). Fig. 2(b) illustrates a simplified comparison of branch-specific \( \omega \) values in phylogenetic trees, grouping these values within four selection intervals: (i) strong negative selection \((\omega<0.1)\), (ii) moderate negative \((0.1<\omega<0.4)\), (iii) weak negative \((0.4<\omega<1)\) and (iv) positive selection \((\omega>1)\). In most cases, internal branches connecting the groups of mild and virulent isolates, particularly clade I and the NUagA group, were associated with \( \omega<1 \), 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. In p33, negative 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 the highest frequency being observed in p33. For some isolates, a characteristic 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 con-

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

The number of negatively selected sites and the mean normalized \( d_N-d_S^* \) values for each genomic region, after correcting the \( P \)-values for multiple comparisons of the same null hypothesis using a false discovery rate (FDR) of 5%, are given. Nucleotide distances were determined by using the bootstrap method (500 replicates).

<table>
<thead>
<tr>
<th>Genomic region</th>
<th>Average nucleotide distance ± SD</th>
<th>No. negatively selected codons</th>
<th>( d_N-d_S^* ) ± SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>p33</td>
<td>0.1641 ± 0.0196†</td>
<td>20 (0.1587)</td>
<td>−2.1098 ± 0.2505†</td>
</tr>
<tr>
<td>p65</td>
<td>0.1091 ± 0.0160</td>
<td>30 (0.1224)</td>
<td>−6.7162 ± 0.6392</td>
</tr>
<tr>
<td>p61</td>
<td>0.1250 ± 0.0160</td>
<td>35 (0.1471)</td>
<td>−7.2201 ± 1.1737</td>
</tr>
<tr>
<td>p18</td>
<td>0.0955 ± 0.0170</td>
<td>12 (0.1034)</td>
<td>−8.6499 ± 1.2031</td>
</tr>
<tr>
<td>p13</td>
<td>0.0741 ± 0.0240‡</td>
<td>16 (0.1403)</td>
<td>−6.3229 ± 1.0417</td>
</tr>
<tr>
<td>p20</td>
<td>0.1145 ± 0.0188</td>
<td>16 (0.1000)</td>
<td>−7.8506 ± 0.9940</td>
</tr>
<tr>
<td>p23</td>
<td>0.1018 ± 0.0143</td>
<td>21 (0.1005)</td>
<td>−9.0407 ± 1.7731</td>
</tr>
</tbody>
</table>

* \( d_N-d_S^* \) average of normalized values of the difference between non-synonymous and synonymous substitutions of selected codons estimated by the FEL method.
† Statistically different from the rest of the values \((P<0.05)\).
‡ Statistically different from p33 and p61 values \((P<0.05)\).

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

For each genomic region, \( \omega \) values are shown. \( \omega \times d_N/d_S \) category ratios for \( d_N \) to \( d_S \) substitution rates; the proportion of branches assigned to each class is shown in parentheses (%).

<table>
<thead>
<tr>
<th>( \omega ) Category interval</th>
<th>Genomic region</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p33</td>
</tr>
<tr>
<td>&gt;1</td>
<td>1.029 (25)</td>
</tr>
<tr>
<td>0.2–1</td>
<td>0.253 (38)</td>
</tr>
<tr>
<td>0.1–0.199</td>
<td>0.122 (67)</td>
</tr>
<tr>
<td>0.01–0.099</td>
<td>0.013 (27)</td>
</tr>
<tr>
<td>0–0.009</td>
<td>0.000 (12)</td>
</tr>
</tbody>
</table>

http://vir.sgmjournals.org
concatenated regions corresponding to adjacent genes were used as input. A total of nine recombination breakpoints were detected: six of them were located in the boundaries of different regions, suggesting recombination events somewhere between the analysed fragments, and three were 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 shared 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 (ratio of transversion to transition rates, shape parameter of the gamma distribution of substitution rates per site and relative substitution rates) were inferred using TREEPUZZLE. 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, Sister-Scan, 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-type isolates in regions p33, p65, p23 and p20, but with C-270-3 in p61 and p18 and as an 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 parental strains 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, which grouped together and distant from other isolates in p20, presented an additional recombination in this region 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 c). 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 a major parental and an unknown isolate providing the 3’ end of p20 and p23. The non-assignment of C-270-3 as a minor parental was due to the use of UPGMA trees in RDP3 default analysis (not shown), but inspection of maximum-likelihood trees indicated that VT and C-270-3 are likely to be the major and minor parental strains, respectively (Fig. 2b and c).

Isolate T346 was close to clade I in regions p33, p61 and p23, to C-270-3 in p65 and was 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 parents, 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 events between T346 and T36 in p61.

**Fig. 2.** Recombination analysis of concatenated CTV sequences. (a) Schematic of the CTV genome with the regions analysed indicated as black boxes. Open reading frames are represented by empty boxes. Motifs of the p349 protein are also indicated: Pro, protease; MT, methyl transferase; HEL, helicase. (b) Unrooted nucleotide maximum-likelihood phylogenetic trees of seven CTV genomic regions. Isolates with congruent phylogenetic relationships are in bold italics. In each region, recombinant isolates (detected by the RDP3 program) are highlighted with solid ovals; those showing incongruent phylogenetic relationships are highlighted with dotted ovals. Branch-specific $d_{j} / d_{g}$ category substitution rates ($w$) are indicated by coloured branches. (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 coloured boxes represent CTV concatenated sequences (isolate code above the box) and internal pale coloured segments indicate recombinant regions; the major parental for each recombinant sequence is indicated below the isolate code and the minor parentals are indicated by short boxes below the coloured segments.
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 several 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) (Fig. 3). Isolate C-268-2 was excluded for the reasons given above.

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, isolate T55 from Florida, and isolates 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, the ancestors of Barão B and Val-CB, and NUagaA-type isolates 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 NUagaA 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 a previous finding that these isolates probably arose from a recombination between a mild (T346) or a virulent (C-269-6) major ancestor and C-270-3. Finally, the ancestors of isolates K, T36 and T346, which showed variable phylogenetic relationships (Fig. 2b), were interconnected in the network in a complex pattern (Fig. 3), supporting recombination between them [as detected by RDP3 (Fig. 2c)].

DISCUSSION

The genetic variation and evolutionary factors shaping CTV populations were studied by comparing the predominant sequence variants in seven genomic regions of 18 isolates from different geographical origins and with different pathogenicity characteristics. It was assumed that pathogenicity would be largely associated with the major sequence variants since: (i) virions obtained from a cDNA clone of the major component of isolate T36 induced the
symptoms characteristic 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 characteristics 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 a similar proportion of selected sites and selection intensity among regions, except for p33 which had less intense selection. p33 had less negative $d_N - d_S$ values, indicating that more non-synonymous substitutions are allowed in this region; this region also had higher evolutionary distances. Although net selection pressure was similar in the other regions, branch-specific analysis showed that the strength of selection varied 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 was 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, which are dispensable for CTV infection and movement (Tatineni et al., 2008), is unknown. Although 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 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 analysed here encodes five of the eight 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, which has been 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 activity (López et al., 2000; Satyanarayana et al., 2002). These amino acids were encoded by invariable codons in all CTV isolates, with the exception of L137 in p20 and C71 in p23, which were found to be under significant purifying selection using cut-off values of $P = 0.5$ and $P = 0.1$, respectively. The observation that p23 is a pathogenicity determinant in Citrus spp. (Ghorbel et al., 2001; Fagoaga et al., 2005) is consistent with separation of the mild and virulent CTV isolates in phylogenetic analysis.

Phylogenetic analysis showed that the mild isolates T32, T55, T300, T312 and T385 form a clade (clade I) supported by bootstrap values $>70\%$ in six regions, thus defining a CTV lineage which also includes isolate T318A from Spain (Ruiz-Ruiz et al., 2006; Albiach-Martí et al., 2000b). A second clade (clade II) including a variable set of virulent isolates was supported by bootstrap values $>75\%$ in three regions, suggesting a higher recombination frequency for those isolates. Within clade II, isolates T305, T388 and NUagA were closely related to each other and distantly related to clade I in all regions and, together with other virulent isolates, showed a star-like evolution pattern from a common ancestor in the phylogenetic network. These data suggest that virulent isolates could represent a second CTV lineage distinctly related to clade I that would also include isolate T318A from Spain (Ruiz-Ruiz et al., 2006). A third clade, including Brazilian and Argentinean isolates genetically related to isolate C-270-3, was observed in several genomic regions. Sequence comparisons showed that several clones of two Colombian isolates released in GenBank were closely related to C-270-3, while others were closely related to virulent isolates in p23. However, the latter were related to clade I or to virulent isolates but not to C-270-3 in p33 (not shown). Although a complete gRNA sequence from South American isolates is currently not available, these findings are compatible with the diversification of a third CTV lineage and frequent recombinations in this area. All lineages included closely related CTV variants from distant locations, a finding that, together with the radiation pattern observed for diversification of some CTV isolates, provide further support to a previous suggestion that genetic flow has occurred (Rubio et al., 2001).

Most isolates showed incongruent phylogenetic relationships in different 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, including 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 the results obtained for these and other Argentinean isolates (Iglesias et al., 2008). Recombination involving p18 was so frequent that, in this region, all South American isolates except Galego 50 formed a monophyletic group diverged from the other isolates. A similar grouping was described for genes p25 and p27, located upstream of 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 to those analysed here.

The phylogenetic network grouped CTV isolates in two major clusters separated by long splits: one of them comprises 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 includes isolates of biogroups 3, 4 and 5 (except for isolate T36) that induce 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 & Elena, 2008) and bromoviruses (Codoñer & Elena, 2005). 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, 2009) suggest that RNA recombination is a major factor in CTV variation and is likely to play 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 divergent sequence variants (Rubio et al., 2001; Weng et al., 2007; Gomes et al., 2008), (iii) the presence in infected cells of several viral RNA species produced during CTV replication that probably 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 the CTV genome that may accumulate 2–3 mutations per genome and replication round. Recombination might help in maintaining functional genomes, even if many non-functional recombinants were produced (Allison et al., 1990).

In summary, analyses of genetic variation in the 3' half of the 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, T35, 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. Martinez and M. Boil for technical assistance in the laboratory and to J. Piquer for excellent care of plants. S. M. and A. S. were recipients 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.

REFERENCES


