VIROLOGY 208, 511–520 (1995)

Complete Sequence of the Citrus Tristeza Virus RNA Genome[1]

A. V. KARASEV, V. P. BOYKO, S. GOWDA, O. V. NIKOLAEVA, M. E. HILF,* E. V. KOONIN,† C. L. NIBLETT,‡ K. CLINE,§ D. J. GUMPF,‖ R. F. LEE, S. M. GARNSEY,* D. J. LEWANDOWSKI, and W. O. DAWSON[2]

*Citrus Research and Education Center, University of Florida, Lake Alfred, Florida 33850-2299; *USDA-ARS, Horticultural Research Laboratory, 2120 Camden Road, Orlando, Florida 32803; †National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894; ‡Plant Pathology and §Horticultural Sciences Departments, University of Florida, Gainesville, Florida 32611; and ‖Department of Plant Pathology, University of California, Riverside, California 92521-0122*

Received October 26, 1994; accepted December 5, 1994

The sequence of the entire genome of citrus tristeza virus (CTV), Florida isolate T36, was completed. The 19,296-nt CTV genome encodes 12 open reading frames (ORFs) potentially coding for at least 17 protein products. The 5′-proximal ORF 1a starts at nucleotide 108 and encodes a large polyprotein with calculated MW of 349 kDa containing domains characteristic of (from 5′ to 3′) two papain-like proteases (P-PRO), a methyltransferase (MT), and a helicase (HEL). Alignment of the putative P-PRO sequences of CTV with the related proteases of beet yellows closterovirus (BYV) and potyviruses allowed the prediction of catalytic cysteine and histidine residues as well as two cleavage sites, namely Val–Gly/Gly for the 5′ proximal P-PRO domain and Met–Gly/Gly for the 5′ distal P-PRO domain. The autoproteolytic cleavage of the polyprotein at these sites would release two N-terminal leader proteins of 54 and 55 kDa, respectively, and a 240-kDa C-terminal fragment containing MT and HEL domains. The apparent duplication of the leader domain distinguishes CTV from BYV and accounts for most of the size increase in the ORF 1a product of CTV. The downstream ORF 1b encodes a 57-kDa putative RNA-dependent RNA polymerase (RdRp), which is probably expressed via a +1 ribosomal frameshift. Sequence analysis of the frameshift region suggests that this +1 frameshift probably occurs at a rare arginine codon CGG and that elements of the RNA secondary structure are unlikely to be involved in this process. The complete polyprotein resulting from this frameshift event has a calculated MW of 401 kDa and after cleavage of the two N-terminal leaders would yield a 292-kDa protein containing the MT, HEL, and RdRp domains. Phylogenetic analysis of the three replication-associated domains, MT, HEL, and RdRp, indicates that CTV and BYV form a separate closterovirus lineage within the alpha-like supergroup of positive-strand RNA viruses. Two gene blocks or modules can be easily identified in the CTV genome. The first includes the replicative MT, HEL, and RdRp genes and is conserved throughout the entire alpha-like superfamily. The second block consists of five ORFs, 3 to 7, conserved among closteroviruses, including genes for the CTV homolog of HSP70 proteins and a duplicate of the coat protein gene. The 3′-terminal ORFs 8 to 11 encode a putative RNA-binding protein (ORF 11), and three proteins with unknown functions; this gene array is poorly conserved among closteroviruses. The genomic double-stranded CTV RNA had an extra G at the 3′ terminus of the minus strand and an extra U at the 3′ terminus of the plus strand.

INTRODUCTION

Citrus tristeza virus (CTV) is a filamentous plant virus with flexible virions composed of one molecule of single-stranded (ss) RNA of positive (+) polarity and one species of coat protein (CP) with MW 25 kDa (Bar-Joseph and Lee, 1989; Sekiya *et al.*, 1991). With the genome estimated to be ca. 20 kb (Bar-Joseph *et al.*, 1985), CTV is the largest known plant virus. The virus is aphid transmitted semipersistently and has a very narrow host range confined to species of *Rutaceae*, in which it is limited to phloem-associated cells (Bar-Joseph *et al.*, 1979; Bar-Joseph and Lee, 1989). CTV is a major pathogen of *Citrus* spp., often causing quick decline and death, or stem pitting and reduced vigor and yield in susceptible varieties, and is considered a serious threat to the citrus industry worldwide. The 3′-terminal 9288 nt of the CTV genome was shown to encompass at least 11 open reading frames (ORFs) (Karasev *et al.*, 1994; Pappu *et al.*, 1994). Based on the sequence analysis of the incomplete RNA-dependent RNA polymerase (RdRp) gene product, CTV was shown to belong to a distinct closterovirus lineage within the alpha-like superfamily of positive-strand RNA viruses (Karasev *et al.*, 1994; Dolja *et al.*, 1994). The CTV genome encodes a homolog of an HSP70 family of heat shock proteins and a duplication of a CP gene (Pappu *et al.*, 1994), thus resembling the beet yellows virus (BYV) genome (Agranovsky *et al.*, 1991a; Boyko *et al.*, 1992). However, the size of the CTV genome was estimated to be ca. 4.5 kb larger than the BYV genome, and the number of genes in the 3′-terminal half of the genome and their organization are different from those of the BYV (Karasev *et al.*, 1994; Pappu *et al.*, 1994).

The sequence reported has been deposited with the GenBank data library under Accession No. U16304.

[1] Florida Agricultural Experiment Station Journal Series No. R-04159.

[2] To whom correspondence and reprint requests should be addressed. Fax: (813) 956-4631.

0042-6822/95 $6.00

Nine of the 11 3′-terminal ORFs were demonstrated to have corresponding 3′-coterminal subgenomic RNAs, suggesting expression of the 3′-proximal genes via formation of the nested set of subgenomic RNAs (Hilf *et al.*, 1995). The CTV RdRp was suggested to be expressed within a large protein product presumably encoded by the 5′-terminal half of the CTV genome (Karasev *et al.*, 1994). The expression strategy of CTV seems to be similar to that of BYV, a type representative of the closterovirus group. At least five subgenomic RNAs were found in BYV-infected plants (Dolja *et al.*, 1990), while the 5′-proximal gene was demonstrated to program synthesis of a 250-kDa protein *in vitro* (Karasev *et al.*, 1989).

CTV has long been classified into a diverse closterovirus group (Bar-Joseph *et al.*, 1979; Lister and Bar-Joseph, 1981; Bar-Joseph and Murant, 1982) in which BYV was the type representative. Comparison of sequences of the entire 15.4-kb BYV genome (Agranovsky *et al.*, 1994) and the 3′ region of the CTV genome taxonomically separated CTV from BYV based on a larger genome size and substantial differences in the 3′-terminal gene organization (Karasev *et al.*, 1994; Dolja *et al.*, 1994; Candresse and Martelli, 1995). According to the proposed revision, the closterovirus group should be upgraded to the family level, and both CTV and BYV would represent two related virus groups within the family *Closteroviridae* (Dolja *et al.*, 1994).

Here we present the complete 19,296-nt sequence of the CTV genome that potentially encodes at least 12 ORFs. The 5′-terminal half of the CTV genome codes for two putative papain-like proteases and the replication-associated complex including the RdRp, helicase (HEL), and methyltransferase (MT) domains. The RdRp domain may be fused to HEL and MT domains in a giant polyprotein via a +1 translational frameshift.

MATERIALS AND METHODS

Virus source and purification

The Florida decline isolate of CTV, T36 (Rosner *et al.*, 1986), was maintained in Mexican lime [*Citrus aurantifolia* (Christm.) Swing.] under greenhouse conditions, and the virus was isolated according to the previously described procedure (Lee *et al.*, 1987).

RNA isolation and cDNA synthesis

Virion RNA was isolated by conventional phenol/chloroform extraction after disruption of CTV virus particles with SDS at 65° (Hilf *et al.*, 1994). Double-stranded RNA (dsRNA) preparations were isolated by two cycles of CF-11 cellulose column chromatography according to the procedure of Valverde *et al.* (1990). To facilitate cloning of the exact 5′ terminus of the CTV genome, dsRNAs were polyadenylated as described previously (Pappu *et al.*, 1994). The first-strand cDNA synthesis was performed on dsRNA templates after denaturation with 20 m*M* methylmercuric hydroxide using random primers and Moloney murine leukemia virus reverse transcriptase (USB) as described previously (Karasev *et al.*, 1994).

PCR amplification and cloning strategy

Polymerase chain reaction (PCR) amplifications utilizing two basic approaches were used for the second-strand cDNA synthesis. We used degenerate primer-mediated PCR in the HEL-encoding region and a step-by-step 3′ to 5′ walking upstream of the HEL region. Treatment of the synthesized PCR products, fractionation, and cloning into *Sma*I-cut pBluescript (SK) (Stratagene) were the same as described previously (Karasev *et al.*, 1994).

Sequencing

Cloned cDNAs were sequenced directly in double-stranded plasmids using a Sequenase 2.0 kit (USB) according to the manufacturer's instructions. Universal T3 and T7 primers were used to sequence termini of cDNA inserts, and virus-specific primers to sequence the rest of the inserts. Both strands were sequenced at least twice.

Primer extension

The 5′ terminus of the primer C75 (Table 1) was labeled with [γ-^{32}P]ATP (New England Nuclear, 6000 mCi/mmol) using T4 polynucleotide kinase (Promega) according to the manufacturer's instructions. Labeled primers were added to the CTV ss- or dsRNAs denatured with 20 m*M* methylmercury hydroxide, and after annealing in 16 m*M* Tris–HCl buffer containing 16 m*M* KCl, 3 m*M* $MgCl_2$, 3 m*M* DTT, and 0.2 m*M* spermidine at 60° for 20 min, the run-off transcripts were synthesized using avian myeloblastosis virus reverse transcriptase (Promega) at 42° for 1 hr (Sambrook *et al.*, 1989). The resulting products were analyzed in 6% urea/polyacrylamide sequencing gels.

Sequence analysis

The sequence obtained was translated with the PROTMAKE program (Trifonov, 1987; Brodsky *et al.*, 1990). Putative translation products were compared with the nonredundant sequence database at the National Center for Biotechnology Information (NIH) using programs based on the BLAST algorithm (Altschul *et al.*, 1990, 1994). The program BLASTP was used to search the amino acid sequence database and the program TBLASTN was used to search the conceptually translated nucleotide sequence database. Pairwise comparisons were performed with the ALIGN-2 program (Brodsky *et al.*, 1990). Multiple sequence alignments were generated by the MACAW program (Schuler *et al.*, 1991). RNA

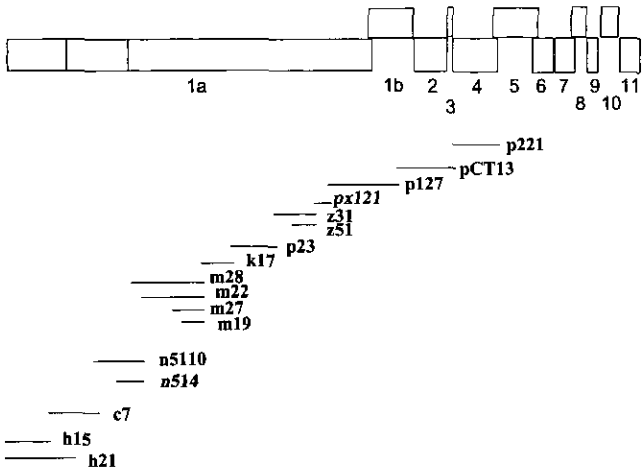

FIG. 1. Schematic representation of the CTV genome and the cloning strategy used in this work. Rectangles represent ORFs with the corresponding number below, and vertical lines within the ORF 1a represent the proposed protease cleavage sites. Horizontal bars represent clones used for sequencing, with the respective designations to the right.

secondary structure was predicted using the algorithm of Zucker and Stigler (1981).

RESULTS AND DISCUSSION

Cloning and sequencing strategy

The 5′-terminal ca. 12 kb of the CTV genome has been cloned in 12 steps (Fig. 1) utilizing PCR to generate the CTV-specific second-strand cDNAs. Based on the assumption that CTV would have a HEL domain similar to helicases of tricorna and tobamoviruses, we designed the degenerate primer D2165, targeted to the N-terminal "GKT" conserved HEL motif. This assumption was substantiated by the fact that RdRp of CTV as well as other closteroviruses is most closely related to the RdRp's in the tobamoviral lineage of alpha-like positive-strand RNA viruses (Karasev *et al.*, 1994; Agranovsky *et al.*, 1991a). The amplification of the DNA fragment that subsequently generated clone p127 was performed with the use of this degenerate primer and the D2205-specific primer from the already-sequenced RdRp region essentially as described previously (Karasev *et al.*, 1994). Attempts to exploit this degenerate primer-mediated cloning strategy further, targeting the conserved MET domains, were unsuccessful. Thus, the ca. 7.8 kb upstream of the HEL was amplified by PCR using a step-by-step walking procedure, in which specific primers from the sequenced region were combined with a random one. The exact 5′ terminus of the CTV genome was amplified from poly(A)-tailed dsRNA with the use of an oligo(dT)-containing primer M111 and two different specific primers (C66 and C67) from the downstream sequenced region, producing clones h15 and h21, respectively. All primers used for the amplifications are listed in Table 1. Seventeen CTV-specific clones representing independent cloning experiments were chosen for sequencing (Fig. 1). Clones pCT13 and p221 were described elsewhere (Karasev *et al.*, 1994; Pappu *et al.*, 1994, respectively) and are presented here for illustrative purposes. The complete 19,296-nt sequence of the CTV RNA genome has been deposited with the GenBank data library under Accession No. U16304 and is not presented here as a separate figure.

The 5′-terminal 10,008 nt reported here together with 9288 nt determined previously (Karasev *et al.*, 1994; Pappu *et al.*, 1994) complete the sequence of the CTV genomic RNA. The CTV genome includes 12 ORFs designated 1 to 11 (Fig. 1), plus the 5′-terminal untranslated region (UTR) of 107 nt and the 3′-terminal UTR of 275 nt.

Noncoding regions

The 5′-terminal nucleotide of the CTV genome is an A, differing from BYV in which the first nucleotide was reported to be a G (Agranovsky *et al.*, 1994). Interestingly, the full-genomic CTV dsRNA had an extra G at the 3′ terminus of its minus strand. Primer extension and sequencing experiments demonstrated an exact correspondence of the 5′-terminal 94 nt of the virion genomic RNA and the plus strand of the dsRNA genomic equivalent (Fig. 2, lanes 1 and 2). But the same region cloned after polyadenylation of the 3′ terminus of the ds- minus strand had an additional G residue (Fig. 2). The run-off transcript synthesized on the virion genomic RNA produced on sequencing gels a strong band corresponding to the 5′-terminal A residue and probably a faint upper band (Fig. 2, lane 2). It is generally believed that the presence of double bands in primer extension experiments reflects the presence of a cap structure at the RNA 5′ terminus. Indeed, the capped odontoglossum ringspot virus genomic RNA produces a clearly visible double band (Fig. 2, lane 3). The 5′-terminal UTR probably consists of 107 nt that is rich in C (29) and A (37) residues and has a low G content (15 residues), thus resembling other Sindbis-like viruses. The 5′-UTR sequence of CTV had no statistically significant similarity to the respective 5′-terminal UTR of the BYV genome (data not shown).

The 3′-terminal UTR of 275 nt also has no statistically significant similarity to the respective 3′-terminal UTR of BYV. The exact 3′ terminus of the plus strand in the CTV dsRNA was found to contain an extra U, absent from the 3′ terminus of the genomic ssRNA. The difference was revealed by comparison of two sets of 3′-terminal clones, one obtained on the polyadenylated virion ssRNA and another one obtained on the polyadenylated dsRNA (data not shown). Thus, the actual 3′-terminal UTR appears to be 1 nt shorter than was reported previously (Pappu *et al.*, 1994), ending in CCA, the canonical tRNA 3′ terminus, and thus resembling 3′-terminal UTRs of alpha-like vi-

TABLE 1

Primers Used for Cloning and the Primer Extension

Primer designation	Primer sequence	Clone(s) generated
Sense-strand primers		
D2056	GTGATGGTNGTNTTYGGNTTNGAYTTYGGNA	p221
D2134	GGTGAGAYCTNTCNAARTTYGAYAARTCNCA	pCT13
D2165	GGTCTCGAGGNGTNSCNGGNTGYGGNAARA	p127
M111	GGTCTCGAGTTTTTTTTTTTTTTTTT	h15, h21
Complementary-strand primers		
D2058	CCAGAATTCGAGTTGATTCAGAGAGGT	p221
D2146	CACGTAGGTATGTACGATGA	pCT13
D2205	GGTGTCGACGCTGACCATCAACTGA	p127
D2247	GCGCTGAGATATCTCGGTTTGA	px121
D2374	CTATCGAAACCTGCGACCTTCA	z31, z51
C12	CTTCCCTCTGCTTCCGTGAT	p23
M34	CCTTACACCGCTCCAGAGAA	k17
M48	CCTGGTCAGGTTGATCGCTA	m19, m22, m27, m28
D2466	CACCGGTCTACAAACGTGAA	n514, n5110
C44	CGGCATACTGATTCTCATACCT	c7
C66	CCACGCATAGGAACCTTAAGA	h15
C67	CCTGAGTGGAGAAATCAGTT	h21
Run-off primer, complementary strand		
C75	GTTGTGGGAATATTAGAGCT	

Note. Redundancy code: R = (AG); S = (CG), Y = (CT); N = (ACGT).

ruses possessing tRNA-like structures. However, computer-assisted analysis demonstrated that the 3′-terminal UTR can form several hairpins (not shown) which do not resemble tRNA-like structures characteristic of several groups of Sindbis-like viruses of plants (Mans *et al.*, 1991).

Double-stranded RNAs of several plant viruses were demonstrated to have nontemplate nucleotides at their termini. An extra G has been reported at the 3′ terminus of the minus strand of the genomic dsRNA for cucumber mosaic virus and its satellite RNA (Collmer and Kaper, 1985), barley stripe mosaic virus ds-equivalents of RNAs 1, 2, and 3 (Dolja and Atabekov, 1987), and genomic-size dsRNA of potato virus X (Dolja *et al.*, 1987). Based on the absence of such an extra G at the 3′ terminus of the minus strands of subgenomic dsRNAs, it is generally believed that nontemplate nucleotides are features characteristic of virus replicative forms, (Collmer and Kaper, 1985; Dolja and Atabekov, 1987). CTV, with an extra G at the 3′ terminus of the minus strand of its genome-size dsRNA, fits well among Sindbis-like viruses, suggesting basic similarity in the mechanism of the minus-strand synthesis. The structure of the exact 5′ terminus of the CTV RNA is at present unclear, but the presence of a faint upper band in the gels with primer extension products (Fig. 2) may indicate that the 5′ terminus of the CTV virion RNA is capped. Synthetic cap analog, m^7Gpp, completely inhibited cell-free translation of the BYV virion RNA, suggesting the presence of the 5′-terminal cap structure (Karasev *et al.*, 1989).

Open reading frames

The first AUG codon is located 108 nt downstream of the 5′ terminus. It has a favorable context (Kozak, 1989) and is likely to be an initiator for the first ORF, hereafter referred to as ORF 1a (Fig. 1). The ORF 1a starting at nucleotide 108 and ending at nucleotide 9480 potentially encodes a large protein product of 3124 amino acids with the calculated MW of 348,954 Da. The downstream ORF 1b starting at nucleotide 9355 and ending at nucleotide 10,855 overlaps ORF 1a by 123 nt and encodes a putative protein product of 500 amino acids with the calculated MW of 56,889 kDa. Overall organization of ORFs 1a and 1b in the CTV genome is similar to that in BYV (Agranovsky *et al.*, 1994), suggesting by analogy that ORF 1b may be expressed via a +1 ribosomal frameshift resulting in a giant 3582-amino-acid fusion polyprotein with the calculated MW of 401,097 Da. The downstream ORFs 2 to 11 (Fig. 1) encoding 10 protein products ranging from 6 to 65 kDa were described in detail previously (Karasev *et al.*, 1994; Pappu *et al.*, 1994).

Duplication of the putative leader papain-like protease and predicted polyprotein processing

Analysis of the amino acid sequence of the N-terminal portion of the CTV ORF 1a-encoded product revealed two putative papain-like thiol protease domains. The sequences of these domains showed statistically highly significant similarity to each other and to the papain-

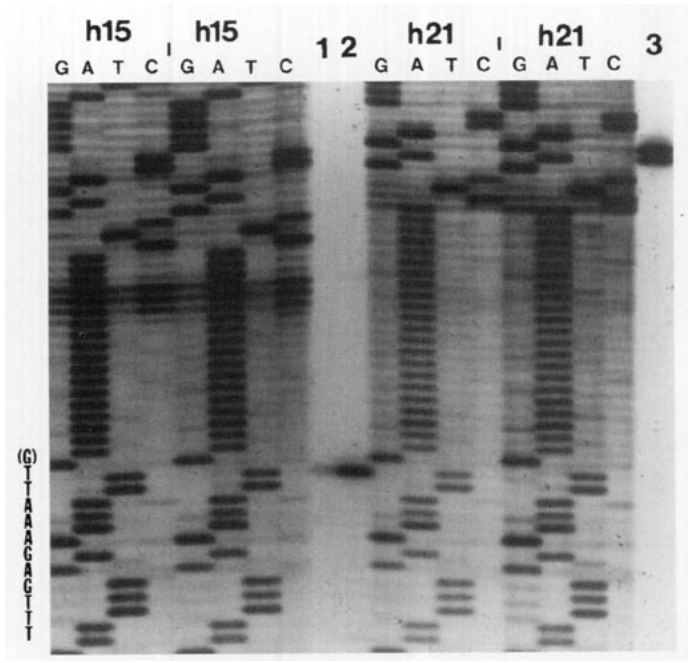

FIG. 2. Analysis of the 5′-terminal CTV-specific primer extension products in a 6% sequencing polyacrylamide gel. ^{32}P-labeled primers C75 (lanes 1 and 2) and D377 (lane 3) were used for run-off transcription. Lane 1, total dsRNA from CTV-infected plants; lane 2, virion CTV ssRNA; lane 3, virion ORSV ssRNA. Sequencing ladders obtained on CTV-specific clones h15 and h21 with the use of the C75 primer are indicated. The 3′-terminal sequence of the genome-size dsRNA minus strand is presented at the left; the 5′ to 3′ direction is from bottom to top; the nontemplate 3′-terminal G residue is indicated by parentheses.

like leader protease of BYV, allowing an unambiguous prediction of catalytic cysteine and histidine residues as well as the cleavage sites for both of the putative CTV proteases (Fig. 3; see the figure legend for statistical estimates). The predicted catalytic residues are Cys-403 and His-464 for the first, N-proximal protease, and Cys-896 and His-956 for the second, distal protease. The cleavage is predicted to occur between Gly-484 and Gly-485 in the upstream domain and between Gly-976 and Gly-977 in the downstream domain (Fig. 3). Characteristically, both of the cleavage sites are preceded by a bulky, hydrophobic amino acid, Val and Met, respectively. Cleavage at these sites would release two leader polypeptides of nearly identical size, namely 484 and 492 amino acid residues for the proximal and distal products, respectively, and a C-terminal fragment of the ORF 1a-encoded product of 2148 amino acids with the calculated MW of 240,038 Da. The near identity of the size of both leaders as well as the fact that the two putative protease domains of CTV showed somewhat higher similarity to one another than to the BYV protease (Fig. 3) may suggest that these two CTV leader proteases evolved by tandem duplication.

The regions of these putative CTV leader proteins upstream of the papain-like protease (P-PRO) domains showed no significant similarity to one another, to the corresponding region of the BYV leader protein, or to any other protein in the available database.

Replication-associated proteins

The region of the polyprotein immediately downstream of the suggested P-PRO cleavage site was identified as a putative MT domain (Fig. 4A). This sequence showed high similarity to the homologous domain of BYV (47.8% identical amino acid residues in a 268-residue alignment) and encompassed all the motifs characteristic of positive-strand RNA viral type I MTs (Rozanov *et al.*, 1992; Koonin and Dolja, 1993).

The C-terminal portion of the ORF 1a-encoded product of CTV showed high similarity (48% identity in a 394-residue alignment) to the putative HEL domain of BYV (Fig. 4B) and contained the seven conserved motifs typical of superfamily I helicases (Gorbalenya and Koonin, 1993).

The ORF 1b-encoded product of CTV contained the conserved motifs typical of positive-strand RNA viral RdRp's (Koonin and Dolja, 1993) and was closely related to the putative RdRp of BYV (60% identity in a 463-residue

```
BYV-L   447-589 DDFEIEVHPLRGGKLSVLLILPKGEAYCVVTAAT-PQYHA-ALTIARGDR
                        :    *:  *  :* :::  *      :     * * * * :::   *
CTV-L1  339-485 THVLAITDPVLCG--TVSVLMSDGVNTARLDVVA-PAYYAYAESLGARVR
                : ::*  :**  *   * * :   *  * : : *   * * :*    *
CTV-L2  832-977 SRFVACIEPVESG--LVKVHFRSDVFRASFPFDGQPIHPATALTLEVVVS
consensus       ..&.....PU..G...V.U.&........&.....P...A.A.$U.....

                                  1
BYVL            PRVGELLQYRPG--EGLCYLAHAALCCALQKRTFREEDFF-VGMYPTKFV
                                :  :* **: *       *      :*****
CTVL1           FLMESSDVDDAKVRDGQCYVRHVFDVALYFGRRADLSVRRALGMYPTVGA
                : *  : :::  ::** **:**: :*:* :**          ** :* *
CTVL2           WDWMVTAPQNP-LKDGYCYLRHFAEVSLSMGRIFFRRDVD-LGPFPYVFE
consensus       ..&..........BG.CYU.H.........R.........UG.@P....

                                  1
BYVL            FAKRLTEKLGPSALKHPVRGRQVSRSLFHCD-VASAFSSPFYSLPRFIG/G
                :   *  :*  :** *:**     *:***  :   :   :    ::* *
CTVL1           LKAYLVREYGRDSLKVPMRGIYTFGSVFHCLSLNLQVDLRSIPNHHLVG/G
                :   * * **: :*:  :** *:      ***    :   : &:   *  :* *
CTVL2           VQHRLERLYGKAALRYGVRGQYSAPRCFHCCYNDSPRPMASFNGYHKMG/G
consensus       &...L...&G...LJ..URG.......FHC..........&.....UG/G
```

FIG. 3. Alignment of the predicted amino acid sequences of the two putative papain-like proteases of CTV with the homologous protease domain of BYV. CTV-L1 and CTV-L2 designate the proximal and distal leader protease domains of CTV, respectively, and BYV-L designates the leader protease domain of BYV. Asterisks indicate identical amino acid residues and colons indicate similar residues between CTV-L1 and CTV-L2, and between CTV-L1 and BYV-L. Percentage identity: CTV-L1/CTV-L2 — 28.8% of 146 aligned residues; CTV-L1/BYV-L — 22% of 145 aligned residues; CTV-L2/BYV-L — 24% of 146 aligned residues. The exclamation marks indicate the predicted catalytic Cys and His. The slashes indicate the predicted cleavage sites. The consensus shows amino acid residues conserved in the three aligned sequences; U, a bulky aliphatic residue (I, L, V, or M); @, an aromatic residue (F, Y, or W); &, any bulky hydrophobic residue; $, S or T; B, a negatively charged residue (D or E); J, a positively charged residue (K or R); dot, any residue. The segments of statistically significant similarity identified using the MACAW program are overlined; the probability of obtaining the alignment of the three sequences by chance was below 10^{-15} for the region around the putative catalytic histidine, and below 10^{-4} for the region around the putative catalytic cysteine.

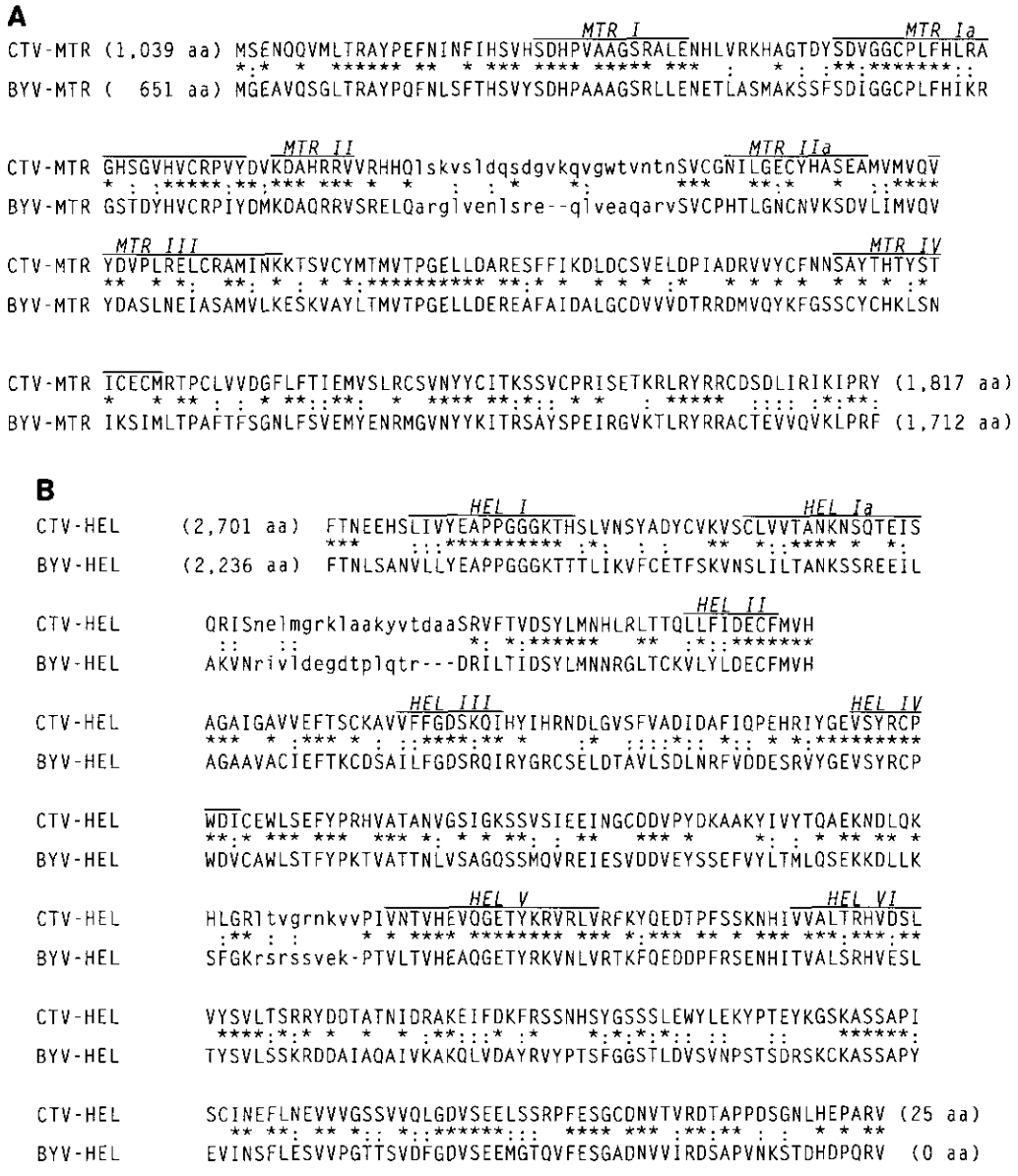

FIG. 4. Alignment of the predicted amino acid sequences of the conserved, replication-associated domain of CTV and BYV. (A) The methyltransferase domain. Identical (asterisks) and similar (colons) residues are indicated. The conserved motifs typical of positive-strand RNA virus type I methyltransferases are overlined and designated after Rozanov *et al.* (1992). The distances from the ORF 1a ends are indicated. (B) The helicase domain. The conserved motif for superfamily I viral and cellular helicases are from Gorbalenya and Koonin (1993). (C) The RNA-dependent RNA polymerase domain. The conserved motifs are from Koonin and Dolja (1993). A complete alignment of ORF 1b is shown.

alignment; Fig. 4C) and beet yellow stunt virus (Karasev *et al.*, 1994; data not shown).

Phylogenetic analysis grouped the closterovirus RdRp's into a distinct lineage within the "tobamovirus subdivision" of the Sindbis-like supergroup of positive-stranded RNA viruses (Koonin and Dolja, 1993; Dolja *et al.*, 1994). Similar results were obtained for the MT and HEL domains (data not shown), suggesting that all three replication-associated genes probably evolved as a single replication module.

Absence of significant similarity between 5′- and 3′-UTRs of CTV and BYV is surprising as replication-associated proteins of both viruses demonstrate striking conservation at the amino acid sequence level (Figs. 4A–4C). This might suggest that elements of the RNA secondary structure rather than primary sequence are responsible for interaction between the closterovirus replicase complex and the 3′- and 5′-terminal UTRs or that protein domains involved in those interactions lay outside of the most conserved replicase regions. Interestingly, 5′- and 3′-terminal UTRs in two genome components of another closterovirus, lettuce infectious yellows virus (LIYV), also lack similarity (Klaassen *et al.*, 1994a).

The vast region of the CTV ORF 1a-encoded polyprotein between easily identifiable MT and HEL domains showed no significant similarity to any protein in the

C

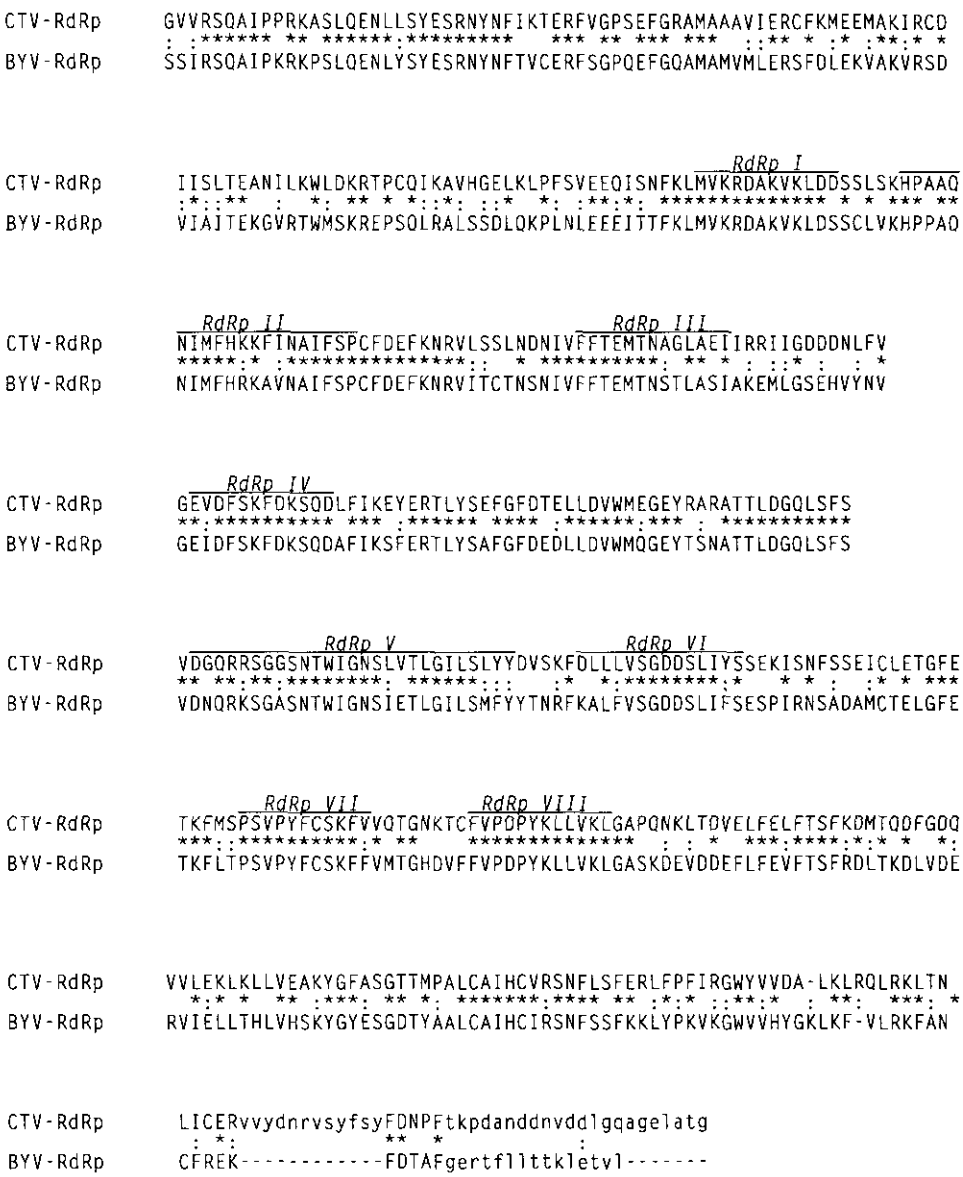

FIG. 4—*Continued*

available databases. Somewhat surprisingly, it has barely detectable similarity even to the corresponding region in the ORF 1a-encoded product of BYV (data not shown). Thus, the closely related replication-associated MT, HEL, and RdRp domains of CTV and BYV are apparently contained within substantially diverged genetic backgrounds.

ORF 1a/ORF 1b frameshift site

It has been proposed previously that the BYV RdRp encoded in ORF 1b is expressed via a +1 ribosomal frameshift (Agranovsky *et al.,* 1994). Direct comparison of the CTV and BYV nucleotide sequences within the ORF 1a/1b overlap revealed apparent and statistically significant similarity (Fig. 5) which, at least in part, could be a reflection of a very high level of amino acid sequence conservation between HEL and RdRp domains of both viruses (Figs. 4B and 4C). Inspection of the C-terminal portion of the helicase alignment and the N-terminal portion of the RdRp alignment between CTV and BYV allowed precise mapping of the putative frameshift within the CTV ORF 1a/1b overlap, namely at C-9405 of the CTV sequence (Fig. 5). The frameshift is predicted to occur in homologous regions of the genomes of the two closteroviruses. However, the CTV region around the predicted frameshift point had important differences from the corresponding region of BYV. First, there is no stop codon corresponding to the UAG stop codon of the BYV. In CTV, an arginine (CGG) codon aligns with the BYV

```
CTV-HEL  P  D  S  G  N  L  H  E  P  A  R  V
                                  CTV-RdRp  G  V  V  R  S  Q  A  I  P  P  R  K  A
CTV(nt) ccT gAC tcG ggt AaC tta CAC GAa CCG gct CGc GTTc gGC gta gTa aGg TCa CAa GCa ATT CCT cca AGa AAA gCG
BYV(nt) gtT aAC aaG tcg AgC gat CAC GAc CCG cag CGg GTTt aGC tcg aTt cGc TCg CAg GCg ATT CCT aag AGg AAA cCG
BYV-HEL  V  N  K  S  T  D  H  D  P  Q  R  V
                                  BYV-RdRp  S  S  I  R  S  Q  A  I  P  K  R  K  P
```

FIG. 5. Alignment of the CTV and BYV nucleotide sequences (presented as DNA) in the vicinity of the proposed frameshift, nt 9380–9455 in CTV and nt 7368–7443 in BYV. Identical nucleotides are typed in uppercase letters, the BYV TAG terminator and the corresponding CTV Arg codon (CGG) are underlined. The encoded C-terminus of HEL and N-terminus of RdRp are presented above (CTV) and below (BYV) the nucleotide alignment.

UAG terminator. Second, this region lacks the so-called "slippery" GGGUUU sequence believed to be an important element of the BYV frameshift signal. Third, a stem-and-loop structure postulated to be involved in the BYV frameshift (Agranovsky *et al.*, 1994) seems to be absent from the CTV frameshift region.

Several groups of RNA viruses have a −1 translational frameshift mechanism of RdRp expression, including plant luteo- and dianthoviruses. A −1 frameshift signal has a very complex structure, which includes extensive stem–loop and pseudoknot elements (cf. ten Dam *et al.*, 1990). Both RdRp's of CTV and BYV are presumably expressed via a +1 ribosomal frameshift resulting in a huge fusion product of ORF 1a plus ORF 1b (Agranovsky *et al.*, 1994; Dolja *et al.*, 1994; Fig. 5). This type of the frameshift, which so far is unique among positive-strand RNA viruses, has been reported to occur in some yeast retroelements (cf. Farabaugh *et al.*, 1993), in a dsRNA virus of *Leishmania* (Scheffter *et al.*, 1994), and in an artificially mutated thymidine kinase gene of herpes simplex virus (Hwang *et al.*, 1994). Information gathered on the best-studied yeast retrotransposons indicates that the +1 frameshift signal is relatively simple and might contain only seven bases. To shift an ORF into a +1 direction, a ribosome needs to be stalled or paused, which may be achieved at a termination codon, a slippery sequence, a hungry or rare codon, or even due to strong competition for the abundant tRNA species by neighboring codons (Farabaugh, 1993; Farabaugh *et al.*, 1993). Alignment of nucleotide and amino acid sequences in the vicinity of the proposed frameshift for CTV and BYV (Fig. 5) suggested that in both viruses the frameshift event occurs at the same point, even though mechanisms mediating this event might differ. A rare arginine codon (CGG) in the CTV sequence aligns with the UAG terminator in BYV, suggesting different pausing mechanisms. A specific RNA secondary structure and a slippery sequence which were postulated to be involved in mediating the +1 frameshift of BYV (Agranovsky *et al.*, 1994) are not conserved in CTV, and their significance in the shift mechanism remains uncertain.

Genome organization of CTV and relationships with other closteroviruses

We have completed the 19,296-nt sequence of CTV genome, the largest among known plant viruses (Francki *et al.*, 1991). The single-stranded positive-sense RNA of CTV is ca. 20% longer than the RNA of BYV, a related closterovirus. Along with predictable, significant similarities in amino acid sequence and genome organization, CTV demonstrated considerable difference from BYV (Fig. 6). Perhaps the most conspicuous of these differences is the apparent duplication of the leader protease domain in CTV which accounts for ca. 25% of the difference in the genome size between CTV and BYV. Duplication of a leader papain-like protease is a common theme that is thought to have emerged independently in the evolution of three other groups of positive-strand and double-stranded RNA viruses: coronaviruses (Lee *et al.*, 1991), arteriviruses (Godeny *et al.*, 1993), and fungal dsRNA hypovirulence virus of *Cryphonectria parasitica* (Koonin *et al.*, 1991). Such a persistent trend toward an increase of the genome via leader protease duplication in spite of the severe constraints on the RNA genome size in general suggests that there might be some selective advantage for a virus to have two leader proteases instead of only one. Understanding of the nature of this advantage may come with a better functional characterization of this leader protease domain. The function of the autocatalytically cleaved leader product in the closteroviruses is unknown.

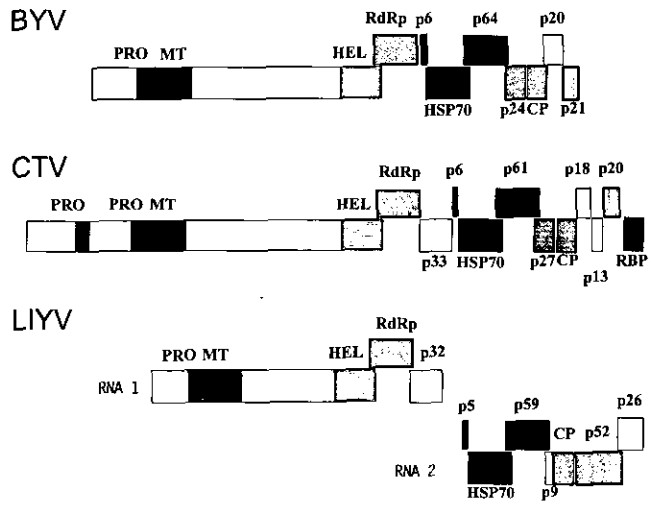

FIG. 6. Schematic representation of the genomes of BYV (after Agranovsky *et al.*, 1994), CTV, and LIYV (after Klaassen *et al.*, 1994b). Rectangles represent ORFs; homologous genes are dashed similarly; open boxes indicate genes with no statistical similarity to other proteins in existing databases. PRO, papain-like protease; MT, methyltransferase; HEL, helicase; RdRp, polymerase; HSP70, homolog of HSP70 proteins; CP, coat protein; RBP, putative RNA-binding protein.

The ca. 4-kb difference in CTV and BYV genomes comprises, in addition to the leader protease duplication, a ca. 1-kb insert encoding an additional 33-kDa protein between the RdRp and HSP70 genes, a 3′-terminal, 0.7-kb insertion encoding a 23-kDa protein that probably binds RNA, and small inserts in the polyprotein between the MT and HEL domains (Fig. 6).

Both CTV and BYV have apparent similarities in genome organization, suggesting a common evolutionary origin for all closteroviruses (Dolja *et al.,* 1994). However, the CTV genome organization is clearly distinct from that of BYV as well as that of the two-component LIYV (Fig. 6). CTV, BYV, and LIYV have very similar replication-associated gene modules or blocks including MT, HEL, and RdRp domains (which is conserved in Sindbis-like viruses), preceded by the similar papain-like protease upstream of this replicative "module" (Figs. 3, 4, and 6). Apparent similarity of both CTV P-PROs to papain-like proteases of potyvirus helper components (data not shown) could suggest involvement of one or even both N-terminal leader proteins in aphid transmission of the virus. CTV, BYV, and LIYV have a five-gene module which seems to be a unique feature of closteroviruses. This five-gene block encodes (1) a small hydrophobic, probably membrane-associated protein; (2) a homolog of the HSP70 proteins; (3) a 58- to 64-kDa protein with unknown function; (4) a diverged copy of the CP; and (5) a coat protein. It is possible that some of the common biological features of closteroviruses, i.e., association with the phloem and unique cytopathology might be determined by this "closterovirus-specific" five-gene module.

Outside of these two modules or gene blocks, other genes in CTV, BYV, and LIYV differ (Agranovsky *et al.,* 1994; Klaassen *et al.,* 1994b; Karasev *et al.,* 1994; Pappu *et al.,* 1994). Besides an additional copy of the leader thiol protease, CTV has four unique genes that have no counterparts in genomes of BYV and LIYV, i.e., genes encoding p33, p18, p13, and p23 (RBP) (Fig. 6). BYV has one unique gene encoding p20 (Agranovsky *et al.,* 1991a), and LIYV has apparently four unique genes (Klaassen *et al.,* 1994b). The functional significance of all these "unique" genes remains to be elucidated.

CONCLUSION

Positive-strand RNA viruses represent more than three-quarters of all plant viruses known to date (cf. Mathews, 1991; Francki *et al.,* 1991). While the genome organization and gene repertoir of these plant viruses demonstrate great variability (Dolja and Carrington, 1992; Koonin and Dolja, 1993), they seem to have a certain limit imposed on the length of their genome, i.e., very few viruses exceed the 10-kb size of a single genome component. Closteroviruses are a notable exception to this general rule, with CTV having a single genome component roughly twice as large as, for example, potyviruses. Mechanisms underlying and genes involved in stable maintenance of a genome of this size remain to be determined. An unusually large replicase and an HSP70 homolog are immediate candidates which might determine this "jump" in the genome size (Dolja *et al.,* 1994). Whatever mechanisms may be involved in the genome expansion, closteroviruses will undoubtedly provide important clues to understanding interactions between complex viral genomes and the plant cell.

ACKNOWLEDGMENTS

The authors are grateful to Vicki Klaassen and Bryce Falk for providing data prior to publication, and to Valery Dolja for helpful discussions. This work was supported in part by the USDA-ARS Cooperative Agreements 58-43YK-0-0008 (to IFAS, University of Florida) and 58-5310-1-205 (to University of California, Riverside), a box-tax grant from the Florida Citrus Growers Association, and an endowment in honor of J. R. and Addie S. Graves.

REFERENCES

Agranovsky, A. A., Boyko, V. P., Karasev, A. V., Koonin, E. V., and Dolja, V. V. (1991a). Nucleotide sequence of the 3′ terminal half of beet yellows closterovirus RNA genome: Unique arrangement of eight virus genes. *J. Gen. Virol.* **72,** 15–23.

Agranovsky, A. A., Boyko, V. P., Karasev, A. V., Koonin, E. V., and Dolja, V. V. (1991b). Putative 65-kDa protein of beet yellows closterovirus is a homologue of HSP70 heat shock proteins. *J. Mol. Biol.* **217,** 603–610.

Agranovsky, A. A., Koonin, E. V., Boyko, V. P., Maiss, E., Frotschl, R., Lunina, N. A., and Atabekov, J. G. (1994). Beet yellows closterovirus: *Complete genome structure and identification of a leader papain-like* thiol protease. *Virology* **198,** 311–324.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410.

Altschul, S. F., Boguski, M. S., Gish, W., and Wootton, J. C. (1994). Issues in searching molecular sequence databases. *Nat. Genet.* **6,** 119–129.

Bar-Joseph, M., Garnsey, S. M., and Gonsalves, D. (1979). The closteroviruses: A distinct group of elongated plant viruses. *Adv. Virus Res.* **25,** 93–168.

Bar-Joseph, M., Gumpf, D. J., Dodds, J. A., Rosner, A., and Ginzberg, I. (1985). A simple purification method for citrus tristeza virus and estimation of its genome size. *Phytopathology* **75,** 195–198.

Bar-Joseph, M., and Lee, R. F. (1989). Citrus tristeza virus. *In* "CMI/AAB Descriptions of Plant Viruses," No. 353. Assoc. Appl. Biol., Wellesbourne, Warwick, U.K.

Bar-Joseph, M., and Murant, A. F. (1982). Closterovirus group. *In* "CMI/AAB Descriptions of Plant Viruses," No. 260. Assoc. Appl. Biol., Wellesbourne, Warwick, U.K.

Boyko, V. P., Karasev, A. V., Agranovsky, A. A., Koonin, E. V., and Dolja, V. V. (1992). Coat protein gene duplication in a filamentous virus of plants. *Proc. Natl. Acad. Sci. USA* **89,** 9156–9160.

Brodsky, L. I., Drachev, A. L., Tatuzov, R. L., and Chumakov, K. M. (1990). GENBEE: A package of computer programs for biopolymer sequence analysis. *Biopolim. Kletka* **7,** 10–14.

Candresse, T., and Martelli, G. P. (1995). Beet yellows virus group. *In* "Sixth Report of the International Committee on Taxonomy of Viruses." *Arch. Virol.,* in press.

Collmer, C. W., and Kaper, J. M. (1985). Double-stranded RNAs of cucumber mosaic virus and its satellite contain an unpaired terminal guanosine: Implications for replication. *Virology* **145,** 249–259.

Dolja, V. V., and Atabekov, J. G. (1987). The structure of barley stripe mosaic virus double-stranded RNAs. *FEBS Lett.* **214,** 313–316.

Dolja, V. V., and Carrington, J. C. (1992). Evolution of positive-strand RNA viruses. *Sem. Virol.* **3,** 315–326.

Dolja, V. V., Grama, D. P., Morozov, S. Yu., and Atabekov, J. G. (1987). Potato virus X-related single- and double-stranded RNAs: Characterization and identification of terminal structures. *FEBS Lett.* **214,** 308–312.

Dolja, V. V., Karasev, A. V., and Agranovsky, A. A. (1990). Organization of the beet yellows closterovirus genome. *In* "New Aspects of Positive-Strand RNA Viruses" (M. Brinton and F. Heinz, Eds.), pp. 31–35. Am. Soc. Microbiol., Washington, DC.

Dolja, V. V., Karasev, A. V., and Koonin, E. V. (1994). Molecular biology and evolution of closteroviruses: Sophisticated build-up of large RNA genomes. *Annu. Rev. Phytopathol.* **32,** 261–285.

Farabaugh, P. J. (1993). Alternative readings of the genetic code — Meeting review. *Cell* **74,** 591–596.

Farabaugh, P. J., Zhao, H., and Vimaladithan, A. (1993). A novel programmed frameshift expresses the *POL3* gene of retrotransposon Ty3 of yeast: Frameshifting without tRNA slippage. *Cell* **74,** 93–103.

Francki, R. I. B., Fauquet, C., Knudson, D. L., and Brown, F. (1991). Classification and nomenclature of viruses. *In* "Fifth Report of the International Committee on Taxonomy of Viruses." *Arch. Virol.* (Suppl. 2).

Godeny, E. K., Chen, L., Kumar, S. N., Methven, S. L., Koonin, E. V., and Brinton, M. A. (1993). Complete sequence and phylogenetic analysis of the lactate dehydrogenase-elevating virus. *Virology* **194,** 585–596.

Gorbalenya, A. E., and Koonin, E. V. (1993). Helicases: Amino acid sequence comparisons and structure-function relationship. *Curr. Opin. Struct. Biol.* **3,** 419–429.

Hilf, M. E., Karasev, A. V., Pappu, H. R., Gumpf, D. J., Niblett, C. L., and Garnsey, S. M. (1995). Northern-blot analysis of the citrus tristeza virus-specific RNAs from infected plants. *Virology,* in press.

Hwang, C. B. C., Horsburgh, B., Pelosi, E., Roberts, S., Digard, P., and Coen, D. M. (1994). A net +1 frameshift permits synthesis of thymidine kinase from a drug-resistant herpes simplex virus mutant. *Proc. Natl. Acad. Sci. USA* **91,** 5461–5465.

Karasev, A. V., Agranovsky, A. A., Rogov, V. V., Miroshnichenko, N. A., Dolja, V. V., and Atabekov, J. G. (1989). Virion RNA of beet yellows closterovirus: Cell-free translation and some properties. *J. Gen. Virol.* **70,** 241–245.

Karasev, A. V., Nikolaeva, O. V., Koonin, E. V., Gumpf, D. J., and Garnsey, S. M. (1994). Screening of the closterovirus genome by degenerate primer-mediated polymerase chain reaction. *J. Gen. Virol.* **75,** 1415–1422.

Klaassen, V. A., Boeshore, M., Dolja, V. V., and Falk, B. W. (1994a). Partial characterization of the lettuce infectious yellows virus genomic RNAs, identification of the coat protein gene and comparison of its amino acid sequence with those of other filamentous RNA plant viruses. *J. Gen. Virol.* **75,** 1525–1533.

Klaassen, V. A., Boeshore, M., Koonin, E. V., Tian, T., and Falk, B. W. (1994b). Genome structure and phylogenetic analysis of lettuce infectious virus, a whitefly-transmitted, bipartite closterovirus. *Virology,* in press.

Koonin, E. V., Choi, G. H., Nuss, D. L., Shapira, R., and Carrington, J. C. (1991). Evidence for a common ancestry of a chestnut blight hypovirulence-associated double-stranded RNA and a group of positive-strand RNA plant viruses. *Proc. Natl. Acad. Sci. USA* **88,** 10647–10651.

Koonin, E. V., and Dolja, V. V. (1993). Evolution and taxonomy of positive-strand RNA viruses: Implications of comparative analysis of amino acid sequences. *Crit. Rev. Biochem. Mol. Biol.* **28,** 375–430.

Kozak, M. (1989). The scanning model for translation — An update. *J. Cell Biol.* **108,** 229–241.

Lee, C.-J., Shieh, C.-K., Gorbalenya, A. E., Koonin, E. V., La Monica, N., Tuler, J., Bagdzhadzhyan, A., and Lai, M. M. C. (1991). The complete sequence (22 kilobases) of murine coronavirus gene 1 encoding the putative proteases and RNA polymerase. *Virology* **180,** 567–582.

Lee, R. F., Garnsey, S. M., Brlansky, R. H., and Goheen, A. C. (1987). A purification procedure for enhancement of citrus tristeza virus yields and its application to other phloem-limited viruses. *Phytopathology* **77,** 543–549.

Lister, R. M., and Bar-Joseph, M. (1981). Closteroviruses. *In* "Handbook of Plant Virus Infections and Comparative Diagnosis" (E. Kurstak, Ed.), pp. 809–844. Elsevier, Amsterdam.

Mans, R. M. W., Pleij, C. W. A., and Bosch, L. (1991). tRNA-like structures. Structure, function and evolutionary significance. *Eur. J. Biochem.* **201,** 303–324.

Mathews, R. F. E. (1991). "Plant Virology," 3rd ed. Academic Press, San Diego.

Pappu, H. R., Karasev, A. V., Anderson, E. J., Pappu, S. S., Hilf, M. E., Febres, V., Eckloff, R. M. G., McCaffery, M., Boyko, V., Gowda, S., Dolja, V. V., Koonin, E. V., Gumpf, D. J., Cline, K., Garnsey, S. M., Dawson, W. O., Lee, R. F., and Niblett, C. L. (1994). Nucleotide sequence and organization of eight 3′ open reading frames of the citrus tristeza closterovirus genome. *Virology* **199,** 35–46.

Rosner, A., Lee, R. F., and Bar-Joseph, M. (1986). Detection of a specific Florida isolate of citrus tristeza virus by differential hybridization with cloned cDNA sequences. *Phytopathology* **76,** 820–824.

Rozanov, M. N., Koonin, E. V., and Gorbalenya, A. E. (1992). Conservation of the putative methyltransferase domain: A hallmark of the "Sindbis-like" supergroup of positive-strand RNA viruses. *J. Gen. Virol.* **73,** 2129–2134.

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

Scheffter, S., Widmer, G., and Patterson, J. L. (1994). Complete sequence of Leishmania RNA virus 1–4 and identification of conserved sequences. *Virology* **199,** 479–483.

Schuler, G. D., Altschul, S. F., and Lipman, D. J. (1991). A workbench for multiple alignment construction analysis. *Proteins: Struct. Funct. Genet.* **9,** 180–190.

Sekiya, M. E., Lawrence, S. D., McCaffery, M., and Cline, K. (1991). Molecular cloning and nucleotide sequencing of the coat protein gene of citrus tristeza virus. *J. Gen. Virol.* **72,** 1013–1020.

ten Dam, E. B., Pleij, C. W., and Bosch, L. (1990). RNA pseudoknots: Translational frameshifting and readthrough on viral RNAs. *Virus Genes* **4,** 121–136.

Trifonov, E. N. (1987). Translation framing code and frame-monitoring mechanism as suggested by the analysis of mRNA and 16S rRNA nucleotide sequences. *J. Mol. Biol.* **194,** 643–652.

Valverde, R. A., Nameth, S. T., and Jordan, R. L. (1990). Analysis of double-stranded RNA for plant virus diagnosis. *Plant Dis.* **74,** 255–258.

Zucker, M., and Stigler, P. (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.* **9,** 133–148.