

Nucleotide sequence of tomato ringspot virus RNA-2

M. E. Rott,¹ J. H. Tremaine² and D. M. Rochon^{2*}

¹Department of Plant Science, University of British Columbia, Vancouver, British Columbia V6H 2A2 and ²Agriculture Canada, Research Station, 6660 N.W. Marine Drive, Vancouver, British Columbia V6T 1X2, Canada

The sequence of tomato ringspot virus (TomRSV) RNA-2 has been determined. It is 7273 nucleotides in length excluding the 3' poly(A) tail and contains a single long open reading frame (ORF) of 5646 nucleotides in the positive sense beginning at position 78 and terminating at position 5723. A second in-frame AUG at position 441 is in a more favourable context for initiation of translation and may act as a site for initiation of translation. The TomRSV RNA-2 3' non-coding region is 1550 nucleotides in length. The coat protein is located in the C-terminal region of the large polypeptide and shows significant but limited

amino acid sequence similarity to the putative coat proteins of the nepoviruses tomato black ring (TBRV), Hungarian grapevine chrome mosaic (GCMV) and grapevine fanleaf (GFLV). Comparisons of the coding and non-coding regions of TomRSV RNA-2 and the RNA components of TBRV, GCMV, GFLV and the comovirus cowpea mosaic virus revealed significant similarity for over 300 amino acids between the coding region immediately to the N-terminal side of the putative coat proteins of TomRSV and GFLV; very little similarity could be detected among the non-coding regions of TomRSV and any of these viruses.

Introduction

Tomato ringspot virus (TomRSV), a member of the nepovirus group (Harrison & Murrant, 1977), is found mainly in North America around the Great Lakes and along the Pacific coast where populations of its nematode vectors *Xiphinema* sp. occur. The most serious disease problems caused by TomRSV occur in *Prunus* sp. such as peach and cherry, but TomRSV also causes diseases in apple, raspberry and grapevine. Chronically infected plants do not show obvious symptoms but are characterized by a general decline in productivity (Stace-Smith, 1984).

Like other nepoviruses, TomRSV consists of 28 nm isometric particles composed of 60 copies of a single coat protein species encapsidating each of the two components of the bipartite, single-stranded, positive-sense RNA genome (Harrison & Murrant, 1977; Schneider *et al.*, 1974). The 5' terminus of each viral RNA component is covalently attached to a small protein (VPg) (Mayo *et al.*, 1982) and their 3' termini are polyadenylated (Mayo *et al.*, 1979). Mature TomRSV proteins are probably released from two large polypeptide precursors corresponding to RNA-1 and -2 by proteolytic processing as shown for the nepovirus tomato black ring (TBRV) RNA-1 (Demangeat *et al.*, 1990).

It has been suggested that nepoviruses be included as

part of the picornavirus-like supergroup which includes the comoviruses, potyviruses and piconorviruses (Goldbach, 1987). Some common features within this group include genomic structure and organization, as well as regions of nucleotide and amino acid sequence similarity. Martelli (1975) suggested that members of the nepovirus group could be divided into three subgroups based on the M_r of their RNA-2 components. Subgroup I would include grapevine fanleaf (GFLV), raspberry ringspot, tobacco ringspot and arabis mosaic virus with RNA-2 components of M_r 1.4×10^6 to 1.5×10^6 ; subgroup II would include TBRV, Hungarian grapevine chrome mosaic (GCMV) and artichoke Italian latent virus with RNA-2 components of M_r 1.5×10^6 to 1.6×10^6 ; and subgroup III would include TomRSV, peach rosette mosaic, cherry leaf roll and myrobalan latent ringspot virus with RNA-2 components of M_r $> 1.6 \times 10^6$.

We report here the nucleotide sequence of TomRSV RNA-2. This is the first sequence of a nepovirus from the subgroup having a large RNA-2 component. Comparisons between this sequence and other nepovirus sequences are presented also.

Methods

cDNA clones. Viral RNA used for preparation of cDNA was obtained from a raspberry isolate of TomRSV. Two overlapping

cDNA clones derived from TomRSV RNA-2 were used to determine most of the nucleotide sequence of RNA-2; clone K6 which has been described previously (Rott *et al.*, 1988), and clone O35. Generation of O35 was essentially as described by D'Alessio *et al.* (1987), except that first-strand cDNA synthesis was primed from virion RNA using the phosphorylated synthetic oligonucleotide, oligonucleotide no. 2 (5' TTCTGGTTCCTCTTCC 3'), which was complementary to TomRSV RNA-2 at nucleotide positions 2201 to 2216. The amount of first-strand product was estimated using agarose gel electrophoresis. Solution containing first-strand cDNA was adjusted to 0.2 M-sodium acetate pH 5.8, and cDNA was precipitated with 2.5 volumes ethanol, and then washed with 70% ethanol before continuing with second-strand synthesis (D'Alessio *et al.*, 1987). Synthesized double-stranded cDNA was ligated into the *EcoRV* site of Bluescript (Stratagene) and used to transform competent *Escherichia coli* DH5 α cells (BRL). Northern-blotted TomRSV RNA (Vrati *et al.*, 1987) was used to confirm that clone O35 originated from TomRSV RNA-2. Clone O35 (approximately 2.2 kb) was analysed further by restriction enzyme mapping and dideoxynucleotide sequencing (see below).

Sequencing and sequence analysis. Subclones used to sequence clones K6 and O35 were generated by restriction enzyme digestion and/or exonuclease III-generated nested deletions (Henikoff, 1984). The dideoxynucleotide chain termination method of Sanger *et al.* (1977) was used to sequence double-stranded plasmid DNA templates using modified T7 DNA polymerase (Sequenase) as described by Toneguzzo *et al.* (1988). All portions of RNA-2 were sequenced completely in both directions except for the 5' extreme 28 nucleotides which were not present in any cDNA clone analysed. These 28 nucleotides were sequenced in one direction only using viral RNA as a template (see below). On average each nucleotide was sequenced over three times.

Two regions of TomRSV RNA-2 were sequenced using viral RNA as template with synthetic oligonucleotide primers no. 1 (5' GCCTTCGATGGAACC 3', complementary to positions 115 to 130) or no. 2 and Moloney murine leukaemia virus reverse transcriptase in the presence of dideoxynucleotides (Ahlquist *et al.*, 1981).

Sequence data were assembled and analysed using the Gene-Master (Bio-Rad) sequence analysis software package on a Compaq 386 computer. Further analysis was performed using programs and databases available through the Canadian National Research Council's Scientific Numeric Database Service on a DEC VAX. We would like to thank Bill Ronald and Jim Purves of Agriculture Canada for writing a program used to test the goodness of fit between the amino acid composition of the TomRSV coat protein (Tremaine & Stace-Smith, 1968) and the calculated amino acid composition of segments of the deduced polyprotein sequence encoded by RNA-2 (see Meyer *et al.*, 1986). The raspberry isolate of TomRSV used by Tremaine & Stace-Smith for the coat protein amino acid composition analysis and the raspberry isolate we have sequenced were both obtained from the British Columbia Fraser Valley growing area and are therefore probably very similar. The coat protein multiple sequence alignment was made by a progressive alignment procedure which uses the Needleman-Wunsch algorithm (Needleman & Wunsch, 1970) iteratively to generate a multiple sequence alignment (Feng & Doolittle, 1987). The predicted amino acid sequences were compared using the TFASTA program (Pearson & Lipman, 1988) with sequences in the protein sequence libraries of the NBRF, Swiss-Prot and PseqIp, and the nucleotide sequence libraries of the NBRF, GenBank and EMBL. The 5' region between two potential AUG initiation codons was analysed for a possible coding function using the program TESTSCORE (Fickett, 1982) and a window size of 67. Scores within the top 77 to 100% predict coding regions. This was followed with the absolute positional base preference method (Staden, 1984), using a window size of 50, to determine the coding frame.

Results and Discussion

Sequencing and primary structure of RNA-2

Two overlapping cDNA clones, designated K6 and O35, were used to determine most of the nucleotide sequence of TomRSV RNA-2. Clone K6 which has been previously described (Rott *et al.*, 1988), corresponds to 5498 nucleotides at the 3' end of RNA-2 (Fig. 5a). Clone O35, which corresponds to the 5'-terminal region of TomRSV RNA-2, was derived using the synthetic oligonucleotide primer oligonucleotide no. 2, which was complementary to RNA-2 at a region approximately 400 nucleotides from the 5' end of clone K6 (Fig. 5a). Two oligonucleotides (no. 1, complementary to TomRSV RNA-2 nucleotides 115 to 130 and no. 2, complementary to nucleotides 2201 to 2216) were used to confirm and/or extend the sequence obtained by sequencing cDNA clones. The sequencing reactions using oligonucleotide no. 1 yielded 28 nucleotides at the 5' terminus of RNA-2 not present in O35 and resulted in two strong stop points. These stop points correspond to nucleotides 1 and 2 of TomRSV RNA-2 and are each denoted N in Fig. 1. Oligonucleotide no. 2 was used to sequence part of the viral RNA to confirm the presence of three tandem repeats deduced from the nucleotide sequence within the overlapping portions of clones K6 and O35 (see below).

TomRSV RNA-2 determined as described above is 7273 nucleotides in length excluding the 3' poly(A) tail (Fig. 1). The calculated M_r of RNA-2 is 2.35×10^6 [excluding the 3' poly(A) tail and 5' VPg], which is similar to the 2.4×10^6 M_r value determined by denaturing gel electrophoresis (Murant *et al.*, 1981). The base composition of RNA-2 is 22.5% A, 23.3% C, 24.9% G and 29.3% U.

Open reading frames

A single long open reading frame (ORF) consisting of 5646 nucleotides is present in the virion sense orientation of RNA-2. This ORF, which accounts for 78% of the RNA-2 sequence, begins at the first AUG at position 78 (AUG 78) and terminates at a UAA stop codon at position 5723. The predicted translation product would have an M_r of 207K. AUG 78 (UUUGAUGUC) does not have either the optimal Kozak (CG/ACCAUGG) or Lütcke (AACAAUGGC) context for translation initiation in animals or plants, respectively (Kozak, 1986; Lütcke *et al.*, 1987). The next in-frame AUG, which is the fourth AUG from the 5' terminus (UGCAAUGGA) occurs at position 441 and is in a favourable Kozak context for the initiation of translation with a G in the -3 and +4 positions. This raises the possibility that AUG 441 may be an initiation site for translation. The

predicted translation product beginning at AUG 441 would have an M_r of 194K. However, there is further evidence that AUG 78 may act as an initiation site for translation. The 5' 800 nucleotides of TomRSV RNA-2 were analysed using the program TESTSCORE (Fickett, 1982). The region beginning from AUG 78 gave a score of between 50 and 77% for the first 100 nucleotides, and then rose to 100% through to the end of the sequence analysed. Scores between 77 and 100% indicate coding regions (Fickett, 1982). Further analysis using the absolute positional base preference method (Staden, 1984) indicates that there is only one coding frame in this region beginning at approximately nucleotide 180 to the end of the sequence analysed (the first 800 nucleotides). AUG 78 is the only potential in-frame initiation site upstream of the coding region indicated by TEST SCORE and the absolute positional base preference method. Translation initiation of TomRSV RNA-2 therefore, may be similar to that described for the comovirus cowpea mosaic virus (CPMV) M RNA. The latter has at least two in-frame sites for translation initiation, one at position 161 and the other at position 512 and/or 524 (van Wezenbeek *et al.*, 1983). Both sites appear to be functional (Holness *et al.*, 1989) and preliminary results suggest both leaky scanning and internal initiation as mechanisms for initiation at the 512 site (Wellink *et al.*, 1990). Initiation of translation at internal AUG codons has been established to occur in both poliovirus (Pelletier & Sonenberg, 1988) and encephalomyocarditis virus (Kaminski *et al.*, 1990) by a mechanism of internal ribosome binding. For TomRSV RNA-2, it is not known whether AUG 78 and/or AUG 441 act as the site for initiation of translation and further work is required to address this question.

All other potential ORFs in the positive-strand were less than 355 nucleotides in length. Two ORFs in the negative-strand orientation are present which are 603 and 582 nucleotides in length (positive-strand nucleotide positions 956 to 353 and 3103 to 2521, respectively). A search of the NBRF, Swiss-Prot, PseqIp, GenBank and EMBL databases failed to detect sequences with significant amino acid sequence similarity to either of these two ORFs.

Non-coding regions

The 5' non-coding region of TomRSV RNA-2 was analysed and compared to the corresponding regions of the two RNA components of TBRV (Meyer *et al.*, 1986; Greif *et al.*, 1988) and GCMV (Brault *et al.*, 1989; Le Gall *et al.*, 1989), RNA-2 of GFLV (Serghini *et al.*, 1990) and the B and M components of the comovirus CPMV (Lomonosoff & Shanks, 1983; van Wezenbeek *et al.*, 1983). It has been reported previously that TBRV,

GCMV, GFLV and CPMV share a conserved UGAAAAAU sequence downstream from the 5' terminus (Serghini *et al.*, 1990). TomRSV RNA-2 has a similar sequence (CGAAAAAU). This short octanucleotide was the longest region of sequence identity that could be detected between the 77 nucleotide 5' non-coding region of TomRSV RNA-2 and the 114 to 287 nucleotide 5' non-coding regions of any of these other viruses. The base composition of the 5' TomRSV RNA-2 non-coding region is similar to those of TBRV, GCMV, GFLV and CPMV in having a high U content (44.2%) and a low G + C content (35.1%).

The 3' non-coding region of TomRSV RNA-2 is 1550 nucleotides excluding the poly(A) tail. This is much longer than the sizes of the 3' non-coding regions of TBRV, GCMV, GFLV and CPMV, as well as those of most other positive polarity RNA viruses. The 3' non-coding regions for TBRV, GCMV, GFLV and CPMV are less than 305 nucleotides. Comparison of the 3' non-coding region of TomRSV RNA-2 with those of the RNA-2 components of GFLV, TBRV and GCMV, the M RNA from CPMV, and poliovirus RNA did not reveal any significant sequence similarity except for the sequence AAAAGC found immediately preceding the poly(A) tail in RNA-2 of TomRSV, TBRV and GCMV. The conserved blocks of sequence among the 3' non-coding regions of TBRV, GCMV, GFLV and CPMV previously reported by Serghini *et al.* (1990) could not be detected in the 3' non-coding region of TomRSV RNA-2.

The 3' non-coding regions of TBRV, GCMV, GFLV and CPMV have a high U content (40.5 to 48.0%) which is similar to that of their 5' non-coding regions. The entire 3' non-coding region of TomRSV RNA-2 has a lower U content of 31.2%. However, its extreme 3' end (approximately 110 nucleotides) has a U content of 44.2% which is similar to the high U content of the 3' non-coding regions of the other viruses.

The possible significance of the very long 3' non-coding region on TomRSV RNA-2 is unknown. In a previous paper (Rott *et al.*, 1988) we reported that the 3'-terminal regions of TomRSV RNA-1 and RNA-2 share an extended region of nucleotide sequence homology. Nucleotide sequence analysis of part of RNA-1 has confirmed that RNAs 1 and 2 share 3'-terminal sequence identity over a 1533 nucleotide region (unpublished results). The possible significance of the 3'-terminal identity of RNAs 1 and 2 in viral RNA replication will be discussed in that communication.

Analysis of the RNA-2-coding region

(i) *Location of the putative coat protein-coding region*
Tremaine & Stace-Smith (1968) reported the coat protein amino acid composition for a raspberry isolate of

M S S I C F A G G N H A R L P

[illegible]

Fig. 1. Nucleotide sequence of TomRSV RNA-2 and deduced amino acid sequence of the long polypeptide. The first two nucleotides which could not be read from the sequencing gel are each represented by an N. The second in-frame AUG proposed to be the initiation site for translation is indicated by three asterisks. The underlined regions are complementary to oligonucleotides no. 1 and no. 2. Nucleotides are numbered on the left beginning at the first N. Amino acids are numbered on the right and begin at the first in-frame M.

TomRSV 1320	<u>Q</u> G G S W <u>Q</u> E G T E A A Y L G K V T C A K D A K <u>G</u> T L L H T L D I I K E C K S Q N L L R Y K E W Q
TBRV 834	K A G G S Y A F G E T I E L P A T V T P G T V L A V F N I F D K I Q E T N T K V C S K W L
GCMV 810	R A G G E F A F I H T I D L P T A V T E G Q V L A K I D I F K K I Q D A K S M V C V Q W M
GFLV 680	R G L A G R G V I Y I P K D C Q A N R Y L G T L N I R D M I S D F K G V Q Y E K W I
TomRSV 1371	R Q G F L H G K L R L R L R C F I P T N I F C G H S M M C S L D A F G R Y D S N V L G A S F P V K L A S
TBRV 880	E Q G Y V S Q N L T A I S H L A P N A F S G I A I W Y I F D A Y G K I P G D V - T T T F E L E M A R
GCMV 856	S L C P H V H V L R D S K T S V W T I D F H K I C G V A I W Y I F D A Y G K I P S D V - T T S L E L E I A R
GFLV 723	T A G L V M P T F K I V I R L P A N A F T G L T W V M S F D A Y N R I T S R I - - T A S A D P V Y T
TomRSV 1421	L L P T E V I S L A D G P V V T W T F D I G R L C G H G L Y Y S E G A Y A R P K I Y F L V L S D N D
TBRV 929	S F D P H V Q V L R D V S T S T W V I D F H K I C G Q T L N F S G Q G Y C V P K I W V I A A S T F Q
GCMV 905	S L C P H V H V L R D S K T S V W T I D F H K I C G Q S L N F S G R G F S K P T L W V I A A S T A Q
GFLV 771	L S V P H W L I H H K L G T F C S E I D Y G E L C G H A M W F K S T T F E S P R L H F T C L T G N N
TomRSV 1471	V P A E A D W Q F T Y Q L L F E D H T F S N S F G A V P P I T L P H I F N R L - D I G Y W R G P T E
TBRV 979	L A R S T A T K F R L E F Y T R G E K L V R G L A E Q P - L S Y P I E A R H L T D L N L M L A P K Q
GCMV 955	L P W S A Q V T Y R L E A L A Q G D E I A H G L A T R S I V T Y P I S L E H L K D I E I M L P P R Q
GFLV 821	K E L A A D W Q A V V E L Y A E L E E A T S F L G K P T L V F D P G V F N G K F Q F - L T C P P I F
TomRSV 1520	I D L T S T P A P N A Y R L L F G L S T V I S G N M S T L N A N Q A L L R F F Q G S N G T L H G R I
TBRV 1028	I A V G T Y A M I T - F P V S L A A K L Q S T S G R T A Y S Y A A G L L S H F L G V G T G I H F V V
GCMV 1005	M A I G N A G S I N - F P L S F A V Q Q K S S S G R I A Y S Y A A G L L S H F L G I G T G I H F K I
GFLV 870	F D L T A V T A L R S A G L T L G Q V P M V G T T K V - Y N L N S T L V S C V L G M G T G V R G R V
TomRSV 1570	K K I G T A L T T C S L L L S L R H K D A S L T L E T A Y Q R P H Y I L A D G Q G A F S L P I S T P
TBRV 1077	R T T S S A F V T S K L R I A L W - - G T V P E T D Q L A Q M P H - V D V E V N V D A S L Q I Q S P
GCMV 1054	Q C T S S A F V T A R L R V A L W - - G D T I T L E Q L S Q M P H - V D C D V D V V S S L K I Q S P
GFLV 919	H I C A P I F Y S I V L W V V S E W N G T T M D W N E L F K Y P G - V Y V E E D G S F E V K I R S P
TomRSV 1620	- H A A - - - - T S F L E D M L R L E I F A I A G P F S P K D N K A K Y Q F M C Y F D H I E L V
TBRV 1124	F F S T - - - - - A N F G N S G S A F Y V S T L C A P M A P E T V E T G S E Y Y I Q I K G I E A N
GCMV 1101	F Y A T - - - - - A N F G D S G A R F W T P M S S P M A P E T M E S K L E Y Y I Q I L G I D A D
GFLV 968	Y H R T P A R L L A G Q S Q R D M S S L N F Y A I A G P I A P S G E T A Q L P I V V Q I D E I - V R
TomRSV 1663	E G V P R T I A G E Q Q F N W - C S F R - - N F K I D D W K F E W P A R L P D I L D D K S E V L L
TBRV 1168	P G L C R E I N Y K Q R F A W - C L L E C L D N S K A S P I K V K I P S R I G N L S S K H V K V T N
GCMV 1145	P P M C R Q I N Y D Q R F A W F T L L R P P D P K L S K I L K L T L P S R V C N I A Y K E A T V T N
GFLV 1017	P D L S L P S F E D D Y F V W - V D F S - - E F T L D K E E I E I G S R F F D F T S N T C R V S M
TomRSV 1709	R Q H P L S L L I S S T G F P T G R A I F V F Q W G L N T T A G N M K G S F S A K L A P G K G V E E
TBRV 1217	F V N A L A I L C A T T G M H H G N C T I H F S W L W H - - P A E L G K Q L - G R L K F V Q G M G I
GCMV 1195	Y V N A F A I M C A T T G M H A G K C I L H F S W T L N - - K G T S F K D L Q G H I S F Y S G M G D
GFLV 1063	G E N P F A A M I A C H G L H S G V L D L K L Q W S L N - - - T E F G K S S G S V T I T K L V G D
TomRSV 1759	I - - - E Q T S T V Q P L V G A C E A - R I P V E F K T Y T G - Y T T S G P P G S M E P Y I Y V R T
TBRV 1264	N - - - N E H I G D T M C Y N S L S N T H S V P F Q F G S F A G P I T S G G K A D E A E N W I E I Q S
GCMV 1243	S T I G E H H G E F H L G G P L S S L A V P F E F G S F A G P V T S G G T P F T S E N W L R V E T
GFLV 1109	K A - M G L D G P S H V F A I Q K L E G T E L L V G N F A G - A N P N T R F S L Y S R W M A I K L
TomRSV 1803	T Q A K L V D R L S V N V I L Q E G F S F Y G P S V K H F K K E V G T P S A T L G T N N P V G R P P
TBRV 1312	P D F S W V A S L H V S I E V H E G F K F Y G R S A G P L T I P A T V A D V S A V S G S
GCMV 1293	A H W D W L T S L T V D I Q V L P G F R F Y G R S A G P L T I P S
GFLV 1157	D Q A K S I K V L R V L C K P R P G F S F Y G R T S F P V
TomRSV 1853	E N V D T G G P G G Q Y A A A L Q A A Q Q A G K N P F G R G

Fig. 2. Alignment of the amino acid sequences of the TomRSV, TBRV, GCMV and GFLV putative coat proteins. Alignment was generated using the multiple sequence alignment program of Feng & Doolittle (1987). A double asterisk indicates amino acids common to all four sequences. A single asterisk indicates three of four amino acids at that position are identical. Underlined dipeptides are known or potential cleavage sites. Numbers to the left of each line of sequence refer to amino acid position on each nepovirus RNA-2 polyprotein.

```

      * * * * *
TomRSV 984 ARKRWLS-RKQEDSQEDNIKRY---ADKHGISFEEARAVYKAPKEGVPTQRSILPDVRDAYSARSAGARVRSIFG
GFLV 360 GRAQWISERRQALRRREQANSFEGLAQDTMTFEQARNAYLGAADMIEQGLPLPLPLRSAY-----APRGLWR

      * * * * *
TomRSV 1059 GSPTTRAQRTDFVLTPSAGDASSFSFYFNPVSEQEMAEQERGGNTMLSLDAVEVVIDPVGMPGDDTDLTVMVL
GFLV 435 G-PSTRANYTLDFRLNGIPTG-TNTLEILYNPVSEEMEYRDRGMSAVVIDALEIAINPFGMPGNPTDLTVVAT

      * * * * *
TomRSV 1134 WCQNSDDQRALIGAMSTFVGNGLARAVFYPLGLKLLYANCRVRDGRVLKVIVSSTNSTLTHGLPQAQVSIGTLRQH
GFLV 510 YGHERDMTRAFIGSASTFLGNGLARAIFFPGLQ--YSQEEPRRESIIRLYVASTNATVDTDSVLAIAISVGLTRQH

      * * * * *
TomRSV 1209 LGPGHRTISGALYASQQQGFNIRATEQGGAVTFAPQGGHVEGIPSANVQMGAGEHLIQAGPMQWRLQRSQSSRF
GFLV 585 VGSMTYRTVASTVHQAVQVQGTTLRATMMGNTVVVSEPGSLVTGTPEARVEIGGGSSIRMVGPLQWESVEEPQQT

      **
TomRSV 1284 VVSGHSRT
GFLV 660 SIRSRSRS

```

Fig. 3. Amino acid sequence alignment of the regions N-terminal to the TomRSV and GFLV coat proteins. Dashes indicate gaps introduced to maximize the alignment. Asterisks above the sequences indicate positions of amino acid identity. The numbers to the left of the sequences indicate amino acid positions on the long polyproteins of TomRSV RNA-2 and GFLV RNA-2.

Table 1. Amino acid composition analysis of the TomRSV coat protein

Amino acid	Relative mole ratio*	Scaled†	Determined from sequence‡
A	14.9	36	43
C	5.1	12	10
D+N	16.0	38	45
E+Q	18.1	43	49/47
F	14.5	35	37
G	18.1	43	53/51
H	5.2	12	13
I	12.7	30	28
K	9.7	23	26
L	24.1	58	60
M	2.7	7	7
P	11.5	28	32
R	10.5	25	26
S	16.2	39	38/37
T	14.9	36	40
V	9.9	24	26
W	5.2	12	9/8
Y	7.3	18	20

* From Tremaine & Stace-Smith (1968).

† Relative mole ratio values were rescaled for a coat protein with an M_r of approximately 58000 as determined by Allen & Dias (1977).

‡ Value to the left of the slash is determined from the potential Q-G cleavage site, the value to the right of the slash is determined for the potential Q-E cleavage site.

TomRSV assuming an M_r of 24000. Later, Allen & Dias (1977) reported that the TomRSV coat protein had an M_r of 58000. We have rescaled the coat protein amino acid composition determined by Tremaine & Stace-Smith for a protein with an M_r of 58K (Table 1) and compared this with the predicted amino acid composition encoded by the RNA-2 long ORF by chi-squared analysis. Values of chi-squared were greater than 27.6 for sequences C-terminal to residue 1037 and were less than 5.4 for

sequences between residues 1235 and 1882. This would suggest that the TomRSV coat protein is encoded at the C-terminal region of the RNA-2-encoded polyprotein. A comparison of the amino acid composition of the TomRSV coat protein and the C-terminal region of the RNA-2-encoded polyprotein is shown in Table 1. The coat protein-coding regions for the nepoviruses TBRV, GCMV and GFLV have also been localized to the C-terminal region of the RNA-2-encoded polyproteins (Meyer *et al.*, 1986; Brault *et al.*, 1989; Serghini *et al.*, 1990). An alignment of the amino acid sequence at the C-terminal region of the TomRSV RNA-2 polyprotein and the putative coat protein sequences of TBRV, GCMV and GFLV is shown in Fig. 2. The putative TomRSV coat protein sequence shared sequence identity of 21.4%, 22.9% and 23.4% with the putative coat protein regions of TBRV, GCMV and GFLV, respectively. The N-terminal region of the putative coat protein was analysed for the following potential protease cleavage sites: Q-S, Q-G, Q-M, E-S, E-G, R-A, R-G and K-A (Palmenberg, 1990; Wellink *et al.*, 1986; Serghini *et al.*, 1990; Brault *et al.*, 1989; Demangeat *et al.*, 1990). The sites Q-G and E-G, at amino acid positions 1320-1321 and 1325-1326, are tentatively identified as potential cleavage sites for the TomRSV putative coat protein based on their location and the previously determined size of the TomRSV coat protein (Fig. 2). The Q-G and E-G sites proposed for TomRSV are different from the R-A, R-G and K-A cleavage sites proposed for GCMV, GFLV and TBRV, respectively, but are similar to those of the como-, poty- and picornaviruses (Hellen *et al.*, 1989). A third potential cleavage site may occur at the K-G site at position 1343-1344 which would conform to a nepovirus consensus (K or R-G or A). The putative TomRSV coat protein would have a calculated size of 60K to 62K which is similar to

the 58K determined by SDS-PAGE (Allen & Dias, 1977). The TomRSV coat protein is larger than the coat proteins of TBRV, GCMV and GFLV. The larger size of the TomRSV coat protein is at least partially due to an additional 36 to 48 amino acids at the C terminus (Fig. 2).

(ii) *Coding region N-terminal to the putative coat protein*

The region to the N-terminal side of the putative coat protein has the capacity to encode a 132K to 145K polypeptide depending on which AUG acts as the translation initiation site. It is not known how many different protein products are encoded by this region or what their functions are. However this region can be divided into several domains based on similarities in amino acid sequence and genomic organization with other nepo- and comoviruses as well as some unique features of the TomRSV RNA-2 polyprotein itself.

The TBRV and GCMV RNA-2 polyproteins share amino acid sequence similarity throughout their entire length. However, the most highly conserved region is immediately N-terminal of the putative coat protein sequences (Brault *et al.*, 1989). We were unable to align this conserved region with the corresponding region in TomRSV. Interestingly however, we did detect sequence identity of 36.7% for over 300 amino acids between the regions immediately N-terminal of the TomRSV and GFLV putative coat proteins (Fig. 3). This was the highest match obtained in comparisons between the TomRSV RNA-2 polyprotein and those of other nepoviruses. Only a short region of identity could be detected in this region between TomRSV RNA-2 and

RNA-2 of TBRV or GCMV. It has been suggested, by analogy with the M RNA-encoded proteins of comovirus CPMV, that this region may encode a cell-to-cell transport function (Meyer *et al.*, 1986; Wellink & van Kammen, 1989). If so, the difference between the TomRSV/GFLV and TBRV/GCMV sequences may reflect slightly different mechanisms to potentiate virus movement throughout infected plants.

An internal region of TomRSV RNA-2 from nucleotides 1812 to 2244 consists of three tandem repeats. Two of the repeats were near perfect and 159 nucleotides in length whereas the third was degenerate and 114 nucleotides in length. This region fell almost completely within the overlap portion of clones K6 and O35, and the presence of the repeats was therefore confirmed by two independent clones. In addition, part of the repeated sequence was confirmed by dideoxynucleotide sequence analysis using oligonucleotide no. 2 as a primer and viral RNA as template. Sequence analysis indicated that this region encodes two identical amino acid repeats 53 amino acids in length, and one partial and degenerate

```
Repeat 1 554 SWSSLELFLFANFKVNRGACFLQVLQRVVLPDECDLLSLFEDQLPEGLPSF
*****
Repeat 2 607 SWSSLELFLFANFKVNRGACFLQVLQRVVLPDECDLLSLFEDQLPEGLPSF
*****
Repeat 3 660 SWSSLELFLFASFKVNRGACFLQVLQARKVVSDEFMDVLE
```

Fig. 4. Alignment of the amino acid sequence of the three tandem repeats near the N-terminal region of the TomRSV RNA-2 polyprotein. Asterisks between repeats indicate positions of amino acid identity. Proline residues and the L-P dipeptides are underlined. The numbers at the beginning of each line refer to amino acid position on the long TomRSV RNA-2 polyprotein.

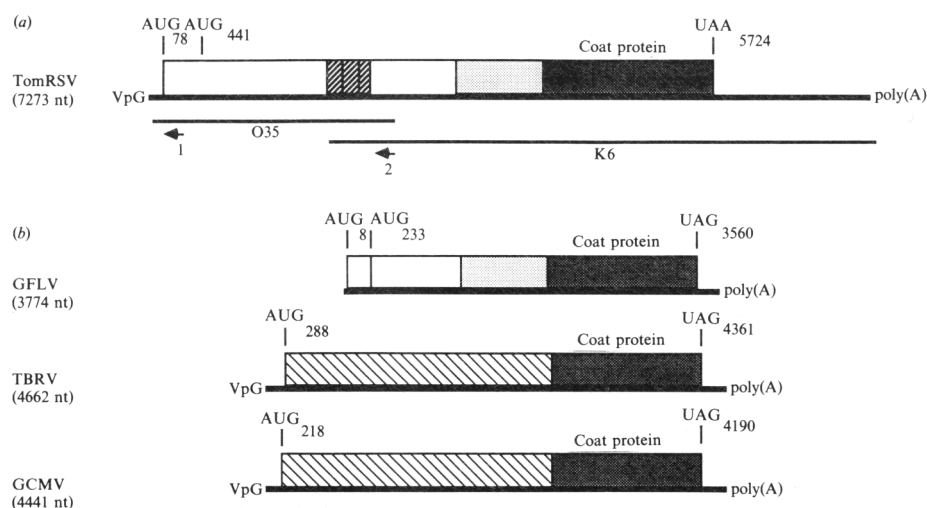


Fig. 5. (a) Location of the cDNA clones O35 and K6 relative to TomRSV RNA-2. Arrows labelled 1 and 2 refer to oligonucleotides no. 1 and no. 2, respectively. (b) Comparison of the TomRSV, TBRV, GCMV and GFLV RNA-2 components. Thick lines represent nucleotide sequence and the large boxed regions indicate the long ORFs. Amino acid sequence similarity is indicated by similar shading within boxed regions. The diagonal shading within the TomRSV sequence indicates the three tandem repeats. Boxed regions without shading do not share amino acid sequence similarity. Numbers refers to nucleotide (nt) positions on the respective RNAs.

repeat of 38 amino acids (Fig. 4). The sequence has a high proline content (12.1%) and the dipeptide L-P is repeated 13 times. The significance of this repeated region, the high frequency of P or the presence of the L-P dipeptide is unknown. A search of the NBRF, GenBank, EMBL, Swiss-Prot and PseqIp databases failed to detect significant amino acid sequence similarity with any other sequence.

In conclusion TomRSV RNA-2 has many similarities with the RNA-2 components of the nepoviruses TBRV, GCMV and GFLV, as well as with the M RNA of CPMV. These include a 5' VPg, 3' poly(A) tail, expression by cleavage of a polyprotein, location of the coat protein-coding region at the C terminus of the large polyprotein, and some sequence similarities within the N-terminal region of the long polyprotein. TomRSV RNA-2 also has some unique features which can account for the much larger size of its RNA in comparison to TBRV, GCMV and GFLV. These features include a very long 3' non-coding region of 1550 nucleotides, a larger putative coat protein-coding region and repeated elements within the N-terminal coding region. A comparison of the genomic organization of the RNA-2 components of TomRSV, TBRV, GCMV and GFLV is shown in Fig. 5. We believe that the unique features of TomRSV as compared to TBRV, GCMV and GFLV confirm the inclusion of TomRSV in a separate subgroup of nepoviruses (Martelli, 1975; Harrison & Murant, 1977).

A grant from NSERC of Canada to J. H. T. providing financial support for M. E. R. is gratefully acknowledged.

References

- AHLQUIST, P., LUCKOW, V. & KAESBERG, P. (1981). Complete nucleotide sequence of bromo mosaic virus RNA3. *Journal of Molecular Biology* **153**, 23-38.
- ALLEN, W. R. & DIAS, H. F. (1977). Properties of the single protein and two nucleic acids of tomato ring-spot virus. *Canadian Journal of Botany* **55**, 1028-1037.
- BRAULT, V., HIBRAND, L., CANDRESSE, T., LE GALL, O. & DUNEZ, J. (1989). Nucleotide sequence and genetic organization of Hungarian grapevine chrome mosaic nepovirus RNA2. *Nucleic Acids Research* **17**, 7809-7819.
- D'ALESSIO, J. M., NOON, M. C., LEY, H. L., III & GERAD, G. F. (1987). One tube double-stranded cDNA synthesis using cloned M-MLV reverse transcriptase. *Focus* **9** (1), 1-4. Gaithersburg: Bethesda Research Laboratories.
- DEMANGEAT, G., GREIF, C., HEMMER, O. & FRITSCH, C. (1990). Analysis of the *in vitro* cleavage products of the tomato black ring virus RNA-1-encoded 250K polyprotein. *Journal of General Virology* **71**, 1649-1654.
- FENG, D. & DOOLITTLE, R. F. (1987). Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *Journal of Molecular Evolution* **25**, 351-360.
- FICKETT, J. W. (1982). Recognition of protein coding regions in DNA sequences. *Nucleic Acids Research* **10**, 5303-5318.
- GOLDBACH, R. (1987). Genomic similarities between plant and animal RNA viruses. *Microbiological Sciences* **4**, 197-202.
- GREIF, C., HEMMER, O. & FRITSCH, C. (1988). Nucleotide sequence of tomato black ring virus RNA-1. *Journal of General Virology* **69**, 1517-1529.
- HARRISON, B. D. & MURANT, A. F. (1977). Nepovirus group. *CMI/AAB Descriptions of Plant Viruses*, no. 185.
- HELLEN, C. U. T., KRÄUSSLICH, H. & WIMMER, E. (1989). Proteolytic processing of polyproteins in the replication of RNA viruses. *Biochemistry* **28**, 9881-9890.
- HENIKOFF, S. (1984). Unidirectional digestion with exonuclease III creates targeted breakpoints for DNA sequencing. *Gene* **28**, 351-359.
- HOLNESS, C. L., LOMONOSOFF, G. P., EVANS, D. & MAULE, A. J. (1989). Identification of the initiation codons for translation of cowpea mosaic virus middle component RNA using site directed mutagenesis of an infectious cDNA clone. *Virology* **172**, 311-320.
- KAMINSKI, A., HOWELL, M. T. & JACKSON, R. J. (1990). Initiation of encephalomyocarditis virus RNA translation: the authentic site is not selected by a scanning mechanism. *EMBO Journal* **9**, 3753-3759.
- KOZAK, M. (1986). Point mutations define a sequence flanking the AUG initiation codon that modulates translation by eukaryotic ribosomes. *Cell* **44**, 283-292.
- LE GALL, O., CANDRESSE, T., BRAULT, V. & DUNEZ, J. (1989). Nucleotide sequence of Hungarian grapevine chrome mosaic nepovirus RNA1. *Nucleic Acids Research* **17**, 7795-7807.
- LOMONOSOFF, G. P. & SHANKS, M. (1983). The nucleotide sequence of cowpea mosaic virus B RNA. *EMBO Journal* **2**, 2253-2258.
- LÜTCKE, H. A., CHOW, K. C., MICKEL, F. C., MOSS, K. A., KERN, H. F. & SCHEELE, G. A. (1987). Selection of AUG initiation codons differs in plants and animals. *EMBO Journal* **6**, 43-48.
- MARTELLI, G. O. (1975). *Nematode Vectors of Plant Viruses*, pp. 223. Edited by F. Lamberti, C. E. Taylor & J. W. Seinhorst. London & New York: Plenum Press.
- MAYO, M. A., BARKER, H. & HARRISON, B. D. (1979). Polyadenylate in the RNA of five nepoviruses. *Journal of General Virology* **43**, 603-610.
- MAYO, M. A., BARKER, H. & HARRISON, B. D. (1982). Specificity and properties of the genome-linked proteins of nepoviruses. *Journal of General Virology* **59**, 149-162.
- MEYER, M., HEMMER, O., MAYO, M. A. & FRITSCH, C. (1986). The nucleotide sequence of tomato black ring virus RNA-2. *Journal of General Virology* **67**, 1257-1271.
- MURANT, A. F., TAYLOR, M., DUNCAN, G. H. & RACHKÉ, J. H. (1981). Improved estimates of molecular weight of plant virus RNA by agarose gel electrophoresis and electron microscopy after denaturation with glyoxal. *Journal of General Virology* **53**, 321-332.
- NEEDLEMAN, S. B. & WUNSCH, C. D. (1970). General method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* **48**, 443-453.
- PALMENBERG, A. C. (1990). Proteolytic processing of picornaviral polyprotein. *Annual Review of Microbiology* **44**, 603-623.
- PEARSON, W. R. & LIPMAN, D. J. (1988). Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences, U.S.A.* **85**, 2444-2448.
- PELLETIER, J. & SONENBERG, N. (1988). Internal initiation of translation of eukaryotic mRNA directed by a sequence derived from poliovirus RNA. *Nature, London* **334**, 320-325.
- ROTT, M. E., ROCHON, D. M. & TREMAINE, J. H. (1988). A 1.9 kilobase homology in the 3'-terminal regions of RNA-1 and RNA-2 of tomato ringspot virus. *Journal of General Virology* **69**, 745-750.
- SANGER, F., NICKLEN, S. & COULSON, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences, U.S.A.* **74**, 5463-5467.
- SCHNEIDER, I. R., WHITE, R. M. & CIVEROLO, E. L. (1974). Two nucleic acid-containing components of tomato ringspot virus. *Virology* **57**, 139-146.
- SERGHINI, M. A., FUCHS, M., PINCK, M., REINBOLT, J., WALTER, B. & PINCK, L. (1990). RNA2 of grapevine fanleaf virus: sequence analysis and coat protein cistron location. *Journal of General Virology* **71**, 1433-1441.
- STACE-SMITH, R. (1984). Tomato ringspot virus. *CMI/AAB Description of Plant Viruses*, no. 290.
- STADEN, R. (1984). Measurements of the effects that coding for a protein have on a DNA sequence and their use for finding genes. *Nucleic Acids Research* **12**, 551-567.

- TONEGUZZO, F., GLYNN, S., LEVI, E., MJOLSNES, S. & HAYDAY, A. (1988). Use of a chemically modified T7 DNA polymerase for manual and automated sequencing of supercoiled DNA. *Bio-Techniques* **6**, 460-469.
- TREMAINE, J. H. & STACE-SMITH, R. (1968). Chemical composition and biophysical properties of tomato ringspot virus. *Virology* **35**, 102-107.
- VAN WEZENBEEK, P., VERVER, J., HARMSSEN, J., VOS, P. & VAN KAMMEN, A. (1983). Primary structure and gene organization of the middle-component RNA of cowpea mosaic virus. *EMBO Journal* **2**, 941-946.
- VRATI, S., MANN, D. A. & REED, K. C. (1987). Alkaline northern blots: transfer of RNA from agarose gels to Zeta-Probe™ membrane in dilute NaOH. *Molecular Biology Reports* **1** (3), 1-4. Richmond: Bio-Rad.
- WELLINK, J. & VAN KAMMEN, A. (1989). Cell-to-cell transport of cowpea mosaic virus requires both the 58K/48K proteins and the capsid proteins. *Journal of General Virology* **70**, 2279-2286.
- WELLINK, J., REZELMAN, G., GOLDBACH, R. & BEYREUTHER, K. (1986). Determination of the proteolytic processing sites in the polyprotein encoded by the bottom-component of cowpea mosaic virus. *Journal of Virology* **59**, 50-58.
- WELLINK, J., VERVER, J., KROON, G. & VAN KAMMEN, A. (1990). Initiation of translation of cowpea mosaic virus M RNA. *VIIIth International Congress of Virology Abstracts, Berlin*, abstract W53-005.

(Received 14 December 1990; Accepted 25 March 1991)