
The complete nucleotide sequence of the potexvirus white clover mosaic virus

Richard L.S.Forster*, Michael W.Bevan⁺, Sally-Ann Harbison¹ and Richard C.Gardner¹

Plant Diseases Division, DSIR, Private Bag, Auckland and ¹Department of Cellular and Molecular Biology, University of Auckland, Private Bag, Auckland, New Zealand

Received November 3, 1987; Revised and Accepted December 8, 1987

ABSTRACT

The complete nucleotide sequence (5845 nucleotides) of the genomic RNA of the potexvirus white clover mosaic virus (WC1MV) has been determined from a set of overlapping cDNA clones. Forty of the most 5'-terminal nucleotides of WC1MV showed homology to the 5' sequences of other potexviruses. The genome contained five open reading frames which coded for proteins of Mr 147,417, Mr 26,356, Mr 12,989, Mr 7,219 and Mr 20,684 (the coat protein). The Mr 147,417 protein had domains of amino acid sequence homology with putative polymerases of other RNA viruses. The Mr 26,356 and Mr 12,989 proteins had homology with proteins of the hordeivirus barley stripe mosaic virus RNA β and the furovirus beet necrotic yellow vein virus (BNYVV) RNA-2. A portion of the Mr 26,356 protein was also conserved in the cylindrical inclusion proteins of two potyviruses. The Mr 7,219 protein had homology with the 25K putative fungal transmission factor of BNYVV RNA-3.

INTRODUCTION

White clover mosaic virus (WC1MV) is a member of the potexvirus group, an agronomically very important group of viruses with flexuous filamentous particles. Potexviruses have one positive-sense genomic RNA that is 6-7 kb long, capped, and polyadenylated (1-4). The genomic RNA directs synthesis in vitro of a non-structural protein of Mr 150,000 (150K) to 180K (5-10). The coat protein is translated from a polyadenylated subgenomic RNA of 0.8-1 kb that is co-linear with the 3' terminus of the genomic RNA (8-11). This subgenomic RNA is efficiently encapsidated by some, but not all, potexviruses (5-12). Other putative subgenomic RNAs, less abundant than the coat protein subgenomic RNA, have been reported in tissues infected with potexviruses (8,10,12).

The nucleotide sequences of the 3' regions of the genomic RNAs of the potexviruses potato virus X (PVX), potato aucuba mosaic virus (PAMV) and WC1MV have been reported recently (13-15). Each virus has an open reading

frame (ORF) coding for a protein of Mr 7,219 to Mr 7,667 located 5' to a coat protein gene (13-15). To further elucidate the genetic organisation of potexviruses, we have determined the complete nucleotide sequence of the genomic RNA of WC1MV.

MATERIALS AND METHODS

cDNA cloning

Double-stranded cDNA corresponding to the 5'-terminal region of the WC1MV genomic RNA was synthesized using oligo (dT)₁₂₋₁₈ as a primer for first strand synthesis, and a synthetic 16-mer corresponding to the 5'-terminal 16 nucleotides for second strand synthesis (16). cDNA clones to other regions of the genome were synthesized using oligo (dT)₁₂₋₁₈ or oligo (dG)₁₂₋₁₈ as primers for first strand synthesis (16) and DNA polymerase I and ribonuclease H (BRL) for second strand synthesis (17). The double-stranded cDNA was dC-tailed, annealed to dG-tailed, PstI-cut pBR322, and transformed to E. coli strain RR1 (9). cDNA inserts were excised from recombinant plasmids using PstI, ligated to PstI-cut pUC19, and transformed to E. coli strain MC1022.

RNA sequencing

The 5'-terminal sequence was obtained by enzymatic digestion (18) of WC1MV RNA that had been terminally labelled (19) with guanylyltransferase (BRL) following treatment with aniline to remove a putative cap structure (20).

DNA sequencing

cDNA clones p8A, pI90, pI43, pI106, pI4B and pM1 were sequenced in pUC19 (21) using an overlapping set of deletions produced by sequential digestions with exonuclease III and S1 nuclease (22). Sequence was obtained from one direction for all clones, leading to at least two independent cDNAs being sequenced for every region except for the 5' most 600 bp. This area, and approximately 90% of the remainder of the virus, was sequenced in both directions.

Nucleic acid and amino acid sequences were analysed using the University of Wisconsin Genetics Computing Group programs mounted on a VAX 11/750 computer. Nucleic acid secondary structures were analysed using the program FOLD. Amino acid homologies were determined with the programs COMPARE and DOTPLOT (using a 30 amino acid window and a stringency of 8). Sequences were aligned manually or with the program BESTFIT (gap weight 1-5, length weight 0.3).

RESULTS**Sequence analysis of WC1MV RNA**

Eight clones containing cDNA inserts which collectively spanned the genomic RNA of WC1MV were selected for sequence analysis (Fig. 1). The nucleotide sequence of clones p5-12 and p14D which correspond to the 3'-terminal region has been presented elsewhere (15). In addition, 38 nucleotides from the 5' terminus of the viral genome were determined by direct RNA sequencing using terminally-labelled RNA. The same 38 nucleotides were found at one end of pA8, the 5'-most cDNA clone.

The nucleotide sequence of the genomic RNA, including the 3'-terminal region, is presented in Fig. 2. The sequence contained 5845 nucleotides in addition to a 3' tract of poly (A) of up to 300 nucleotides (9). This value was close to the length of 6.2 kb estimated previously using RNA denatured with glyoxal and dimethylsulphoxide (9). The base composition estimated from the nucleotide sequence was 55.92% A+U and 44.08% G+C. These were close to the values of 57.5% and 42.5% determined chromatographically (23).

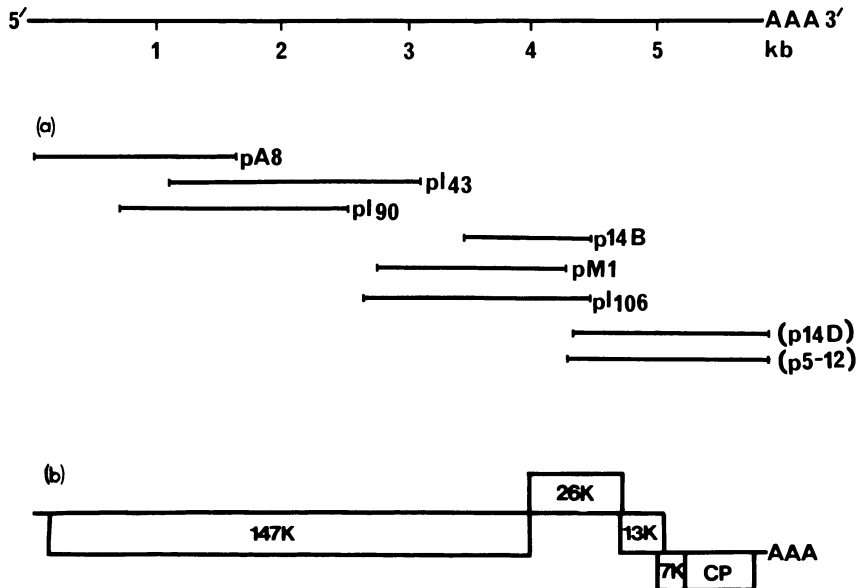


Fig. 1. The genome of white clover mosaic virus (WC1MV). (a) The location of the cDNA clones used for sequencing the genome. The 3'-terminal clones in parentheses have been described elsewhere (15). (b) The location of five major open reading frames on the WC1MV genomic RNA.

Nucleic Acids Research

GAANAACAGACGACGAACTAAACAGAACGAGGCATCCGAGAAAATAAACCACTCCGGTTTTCTTGAACATAACCAACACGTAGTTGACAAAGGCTGCCATGGCTTAAGTTC
10 20 30 40 50 60 70 80 90 100 110 120 M A K V R
A A L D R I T D P S V K A V L N E E A Y S H I R P V L R E S L T N N P Y A I A P
GTGCGCTCTGATGAATCACTGATCCCTGGTAAAGCTGTACTCAACGAAGAGGCTACAGGCACATCCGACCGTCTCTGATGATCCCTGACTTAATCAATGATGCGCATCGCAC
130 140 150 160 170 180 190 200 210 220 230 240
D A A D T L E K Y G I A T N P F A V K V H S H G A V K S E N T L L E F V G T N
CCGATGCCCGACAGCTTAGAAATTTGGAAATTCCTACTAATCCATCCGACGAGAAAGTACACTCCATCGGGGGCGTTAAAGATTCGAAACACCTTACTCGAAGAGTGGTTTA
250 260 270 280 290 300 310 320 330 340 350 360
L P K E P C I F L F L K R S K L R Y L R R G P S N K K D I F I N L A I E R P R D L Q
ACTTCCGAAAGAACATGCATTTCTCTCCCAAAGAAGTAAAGCTTACCTCAGCGTGGACCAAGTAAACAAAGACATTTTCAATAACTAGCAGTAAGGACCCCGGACCTTC
370 380 390 400 410 420 430 440 450 460 470 480
R Y E E D T L V E S W T R I T T R Y A Y I S D T L H F F T R K M L A D L F F H N
AAAGGTAGAAAGACACTCTAGTTGAGAGTTGGACTCGATACACACTAGGTATGCATATATTAGTGACACTTACTGCTCTTCTACTAGGAAGATGCTGGCTGACTTCTTCTCA
490 500 510 520 530 540 550 560 570 580 590 600
P A L D V L P V L A C C T L V L P E A L H K H P S I E P D L Y T I N Y N F N G F O Y I
CTAGTCTAGATGTATGCACTAGCCCTAGTCTCTCCCGCAGAGCCCTTCAACAAACCTAGCATAGAACCTGACTTATATACATTAACTACACTTCAAGTTTTCATGAC
610 620 630 640 650 660 670 680 690 700 710 720
P G N H G G G S Y S H E F K O L E W L K V G H L K S P E L S D L T F O M I E S I G
TCCAGGTAATCATGGTGGCGTCTTACTCCCATGAATTAAACACTGGAATGGCTCAAAGTTGGACTCTCAAATCCCGAGCTGAGTCTCACTTCAGATGTGAATCTTATG
730 740 750 760 770 780 790 800 810 820 830 840
A N H L F M I T R G I K I T P R V R T F T K D S Y V L F P O I F H P R N L N P S
GTGCTAACCACTCTCATGATCACCCTGGCATTAATAACCCAGAGTTCGAACACTTCAATAAGACTCTCATGTCTCTTCCCTCAATCTTCCACCTCGAAACCTCAATCTC
850 860 870 880 890 900 910 920 930 940 950 960
K A A P F P K A A V K A M O L F T V Y V K S V K N P T E R D I Y A K I R O L I K T S E L S
CAAAACCTTTCCAAAGCTTAAAGCTTCTACTGTGAAATCTGTCAATGACTCAACTGACAGGAGAGAGCCCTGACTCGACTCGACTCGACTCGACTCGACTCGACTCGACTCG
970 980 990 1000 1010 1020 1030 1040 1050 1060 1070 1080 1090 1100 1110 1120 1130 1140 1150 1160 1170 1180 1190 1200
D Y H P D E I V H I V N V F V F I S K L D S I N S D I L S P I M S K A L L
CTGATATCACTCAGAGAAATGTCACATAGTGAATTACTTTGTTCATCTCAAATGGATAGCATTAATCTTATSTGACATCTCTCTCTTCTTGTGKCAAAAGCTT
1090 1100 1110 1120 1130 1140 1150 1160 1170 1180 1190 1200
P I K T K I T Q L W E K L T G A R A F N O L L D A L Q W K T F T Y S L E V D F
TGCCATCAAAACAAATACACACACTTGGAAAGCTCACCGGAGCAGAGCTTCAACCACTCTAGATGACACTTCAATGAAACTTCCAGTATCTCTTAGAGATGTATG
1210 1220 1230 1240 1250 1260 1270 1280 1290 1300 1310 1320
T S A P S O R D C F M E D E R L E T D T L E D E V S O N A N N K T S L O N I
TTCCACCGCTCTCTGAGAGACTTTTCATGGAAGTAGAGACTGAAACTGACACACTGAAATGAAGTCTCACAAAATGCAATAAATAAACCAACAGCTCGAAATA
1330 1340 1350 1360 1370 1380 1390 1400 1410 1420 1430 1440
E A V K N N P D L P W A P W L L I L A H N A C T O K Q Y D P E N N L I L P
TTGAGGCGCTCAGAAAGCTCAGACCTCCCTCCCTGCTCATCTGTAATCTCACTGACAGGAGAGAGCCCTGACTCGACTCGACTCGACTCGACTCGACTCGACTCGACTCG
1450 1460 1470 1480 1490 1500 1510 1520 1530 1540 1550 1560
I O E I N T L P K H O H P D I P T D L L T L L K H R E P T T V P L D N H R A
CCATCAGGAGATCAACACTCTCCCAACCAACCAACCTGACATCCCACTGACCTTCACTCACTTCAACCAACTACACAGAGAGCAACCACTTCCACTTGAACACCAAG
1570 1580 1590 1600 1610 1620 1630 1640 1650 1660 1670 1680
R A Y G S D V K N L R I G A L L K K O S K D W L S F A L K T E N I E R A O V L M
CTCGACTTATGATGATTAATAAATCTACGAATAGGCGCACTTCTCAAGAAGCAAGTAAAGCTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGGTGG
1690 1700 1710 1720 1730 1740 1750 1760 1770 1780 1790 1800
S V H G A G G S G K S H A I Q T W M B S L N R R D R H V T I I L P T T D L R N
TGCTGCATCCATGGGCTGGGCTGCAATGCTGCAACTTGGAGCTGCAACTTGGAGCTGCAACTTGGAGCTGCAACTTGGAGCTGCAACTTGGAGCTGCAACTTGGAGCTG
1810 1820 1830 1840 1850 1860 1870 1880 1890 1900 1910 1920
D W T T K V P N L E Q A N F K T F E K A L C Q P C G K I I V F D D Y S K L P Q G
ATGATGGACTACCAAGTACCAACCTGGAACAGCACTTCAAATCTCGAAAAGCACTCTGCAACCTCTGGAAGATCATAGTGTGTGATCACTTAAACTTCCCGCA
1930 1940 1950 1960 1970 1980 1990 2000 2010 2020 2030 2040
Y I E A F A I I N O N V I L A I L T G D S K O S F H H E S E N D A Y T A T L E P
GCTACATGAGCATTCTCTGCCATAAACCAAGCTGTATAGCACTCTCACTGGAGTCTTAAACAGAGTTCATCATGAATCAATGAGGATGCTTATACAGCACTCTGGAAC
2050 2060 2070 2080 2090 2100 2110 2120 2130 2140 2150 2160
S I N T Y O P F C R Y I N I T H R N K P D L A N K L G V Y S C S C S G T G T S F T
CCGATTAACACTTACAGCCCTTCTGCTATATTTGAACTCACTCCGAAACCAACAGCACTTGGCCAGAAATTTGGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGT
2170 2180 2190 2200 2210 2220 2230 2240 2250 2260 2270 2280
M S S O A L K G M P I L S P S I M K K T A L G E M G Q K S M T Y A G C O G L T T
CTATGCTACAGGCCCTCAAAAGCTGCATCTTCTCCAGCATATGAAGAAAATGCTCTTGTGAAATGGCCGAGAAAGCATGACATACCTGGATGCGCAAGGTCTACAA
2290 2300 2310 2320 2330 2340 2350 2360 2370 2380 2390 2400
K A V O I L D T N T P L C S S N V I Y A L S R A V D H I H F T A T T A C A C T G G S N S T
CTAAGGCTCCAAATCTTTGGAACAACCACTTGTGCGAGCTCAATGTCATATACAGAGCCCTCAGTGGCTGTGTGATCATATTCATTAACACTGACCCCAACAA
2410 2420 2430 2440 2450 2460 2470 2480 2490 2500 2510 2520
D F W E K L D S T P Y L K T F L D C V R E E R M N E I V A V E E P P A P V P P
CAGCTCTGGGAGAACTGATTCACACCTTACTCAAACTTCTAGACTGTTCGAGAGGAAAGATGATGAATTTGCGGAGTAGAAGACCCCGCTCGACTTACCTTAC
2530 2540 2550 2560 2570 2580 2590 2600 2610 2620 2630 2640
T H F P K V N P T T V I E S Y V H D L P E K H G R E I F S E T H G H S N A I O
CTACCACTTCCAAAGTCAACCTTACCAAGTGAATCATATGATGATCTCCCGAAAACATGGTAGAGATCTTTTCTGAAACTCAGGCACTCAATGCAATC
2650 2660 2670 2680 2690 2700 2710 2720 2730 2740 2750 2760
T D N P V O L F P H O O A K D E T L Y W A T I E A R V D L O C T S S E A N L K E F
AACTGCAACCTGCTGCTCTTCTCATGAGGCTTAAAGTGAACCTTTACTGGGCACTTGAAGCCAGATTAACATGACTTCTTACCAAGAAACCACTCAAAAGAT
2770 2780 2790 2800 2810 2820 2830 2840 2850 2860 2870 2880
H K H D T G D I L F L N V K Q A H N L P O D P I P F N P D L W T L C K O E I E
TTCATCTCAACATGATGATGATTTCTCTCTCACTCAAAAGCTGATGATCTCTCAGAGCCCTTCAACAGCACTCTGAGCCCTTCAACAGCACTCTGAGCCCTTCAACAGCACTCT
2890 2900 2910 2920 2930 2940 2950 2960 2970 2980 2990 3000
N Y T L K K S A A L V N A A T R O S P D F D S H A I A L F L K S Q W V K K T E
AGAACAATCTCAGAAAGTGTCTGCTTGTAAAGCGGCACTGCAATCACCTGATTTGATTCCTGCAATAGGCTCTTCTGAAATCAGTGTGTCAGAAAGCCG
3010 3020 3030 3040 3050 3060 3070 3080 3090 3100 3110 3120
K I G C L K I K A G G O T I A A F M O A O T V M I G Y G T M A R Y G M R K R F R N O Y C P
AGAAGTGTGCTTAAATAAGAGCCAGACCTGCTGCTTATGCAAACTGCTGCTTATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATG
3130 3140 3150 3160 3170 3180 3190 3200 3210 3220 3230 3240
R A A I F V N C E T T P A D F N S F I L D E W H F R T C F S N D F T A F D O S Q
CAAGAAKCTTTGTAATTTGAAACCACTTCAACTCACTTCACTTCACTTCACTTCACTTCACTTCACTTCACTTCACTTCACTTCACTTCACTTCACTTCACTTCACTTCA
3250 3260 3270 3280 3290 3300 3310 3320 3330 3340 3350 3360
D G S I L O F E V I K A K F H N I P E D I T E G V I O I K T H A K I F L G T L S
AAGAGGCTTCTCATCTCAATTTGAAGTCAAAAGCTTAATTCATAACATCCGAGGACCATTTGAGAGCTCAATTAATCAAAACATGCAAAATTTTCTGGGCACTCTCA
3370 3380 3390 3400 3410 3420 3430 3440 3450 3460 3470 3480
I R L S G E G P T T D A N T E A N I A Y T H T K F N I P C D A A A F O V Y A G D D
GCATATGAGACTCTCGAAGAGCCCTG
3490 3500 3510 3520 3530 3540 3550 3560 3570 3580 3590 3600
M S I D V A S V K P S F N M I E H L M K L K G K P V F N T O F O G G D F A E F C
ATATGCTCATCTG
3610 3620 3630 3640 3650 3660 3670 3680 3690 3700 3710 3720
G W T I S P K G I I K K P E K M N M S I E L O K N I N K F H E V K R S Y A L D H
GTGGCTGGCAATCTCAAAAGGATTAATCAAGAACTGAGAAATGAATATGAGCACTTGAACCTTAAAGAAATCAATAAATTCATGAAGTAAAGAAATGATGCTTCTGAC
3730 3740 3750 3760 3770 3780 3790 3800 3810 3820 3830 3840

```

A F A Y O L G D E L H E L Y N E S E A E H H O L A T R S L I L A G O A T A L D I
ATGCCCTTCGATACCACACTTGGTGGATGAATTCGATGAACTGTACAGTGTAGAGCAGAGCATCCACACTTGGCACAAGGTCCTCATCTAGTGTGTCAGGCCACAGCCCTAGACI
3850 3860 3870 3880 3890 3900 3910 3920 3930 3940 3950 3960

L D Y G L R D L K * M D H I H L L S A H G F T R T R L A K S K P I V V H A I
TTCCTGACTACGGGTAAAGAGACCTTAGTAGCGATGGATCATATTCACCTCCTCCTCAGCGCCACGGCTTTACCGCACCAGACTCGCCAAATCCAAACCCATGTGCTTCATCGAT
3970 3980 3990 4000 4010 4020 4030 4040 4050 4060 4070 4080

A G S G K S T V I R K I L S D L P T P K A Y T L G K P D P Y S L S N P T I K A F
AGCAGGCTCTGGAAAGCTTACCGTGTACAGGAAATTCCTCTCAGACCTACCCACACCTAAAGCTTAGGCGAAACAGAGACCCTTCTTATCAAAACCCCAACCAAGGCTT
4090 4100 4110 4120 4130 4140 4150 4160 4170 4180 4190 4200

A D F K R G T L D I L D E V G O L P L T D L D S S F F E F I F T D P V O A P T D N
CGCCCAATCAAAAGAGTTCGCTCGACATCTTAGAGATATACGGCCACTCCCATGAGCCGATTTAGACATCTTTTGAATTCATCTTCCAGACCTTACCAAGCCCACTGACCA
4210 4220 4230 4240 4250 4260 4270 4280 4290 4300 4310 4320

L F E P H Y T L E T T Y R F G P N T C N L L N O A F O S N I T S L V T K D N I S
TCTCTGACCCCACTACACACTAGAAACCACTATAGATTTGGCCCAAACTTGCACCTTCTCAATCAGCTTTCCAACTAATATTACAGCTTCTCACCAGCAAGATAACTTTC
4330 4340 4350 4360 4370 4380 4390 4400 4410 4420 4430 4440

F G S P Y L V D P V G T I L A F Q P D T Y L I L C L H Q A S F F K V S D V I G Y
ATTGTGTCACCTATTGGTTGACCACTGGTACTTTCCTTCAACAGACACCTACCTTATCTTTGCTTACAGCCTTCTTTCAGCTCTCAGACCTGATTTGGTTA
4450 4460 4470 4480 4490 4500 4510 4520 4530 4540 4550 4560

Q W P T V T L Y L A C K I S E I P E E E R H L L F I G L T R H T E S L L I L G P
CCAGTGGCCACCGTAGCTTATCCTGAACTTCAAAKISSEIPEEEERHLLFLIFGLTRHTESESLILGLGP
4570 4580 4590 4600 4610 4620 4630 4640 4650 4660 4670 4680

D A F D S S P *
M P L T P P P N P O K T Y O I A I L A L G L V L L A F V L I S D H S P K V G D H
AGATGCTTACCTCCCTCCCACTTCCAGAAAATTCACAAATGACCATCTGATAGGATTAGTACTCTCGCCCTTCTCATTTCTGACACTCCCAAACTTTGGTGTG
4690 4700 4710 4720 4730 4740 4750 4760 4770 4780 4790 4800

L H N L P F G G E Y K D G T K S I K Y F O R P N O H S L S K T L A K S H N T T I
ATTTACACATCTCCCAAGGAGGTGATACAAAGAGACTACTAAATCTCAGATATTTCCAAAGACTAAACCATCTTCCAAAGCTTACCAAGCTTACCAAGCACTGACCA
4810 4820 4830 4840 4850 4860 4870 4880 4890 4900 4910 4920

M D F T L I I G V Y L L V F I V F A K I N T S V C
F L L I L G L I V T L H G L H Y F N N N R R V S S S L H C V L C O N K H *
TTTCTCTCATCTAGGCTTGATGTGACCTGACTTCACTACTTAAATAAGGCGTATCTCTACTCTCATGTGTACTTGGCAAAATAAACACTAGTGTGT
4930 4940 4950 4960 4970 4980 4990 5000 5010 5020 5030 5040

T I S I S G A S I E I S G C D N P T L F E I L P K L R P P F N H G L S L P S N *
ACCATTTGATATCCGGAGCTTCCATTTAAATCTCAAGTTGCGAACCTTCAACACTTTTGGAAATCTCTCCAAACCTCAGACCCCTTACCAAGCTTAAATTTGAAA
5050 5060 5070 5080 5090 5100 5110 5120 5130 5140 5150 5160

N A T T T A T T P P S L T D I R A L K Y T S S T V S V A S P A E I E A I T K T
CCATGGCAGCACCAGGACCACTCCACATCTTGGCGAACCTCGTGTCTGAATACACCTCTCCACCGTTCCTGTGCTTCCACCGCTGAGATTGAGATTCACCAAAACC
5170 5180 5190 5200 5210 5220 5230 5240 5250 5260 5270 5280

W A E T F K I P N D V L P L A C W D L A R A F A D V G A S S K S E L T G D S A A
TGGGCTGAACTTCAAAATCCGAATCCGCTCTGCTGCTTCTGCTGATGTGGCGCTTCTTAACTTGAATCTGATGCTGCTGCT
5290 5300 5310 5320 5330 5340 5350 5360 5370 5380 5390 5400

L A G V S R K O L A O A I K I H C T I R O F C M Y F A N I V W N I M L D T R K T P
CTTGGGCTGTTACAGGAAACACTTCCCGCCATCAAAATCCATTCGACCATTCGCAATCTCCCAATCTCTTGAAGCAATATGCTAGACCAACCAAAACCA
5410 5420 5430 5440 5450 5460 5470 5480 5490 5500 5510 5520

P A S W S K L G Y K E S K F A G F D F D G V N H P A A L M P A D G L I R G P
CCAGCATCTGGTCTAGCTGGCTACAGAGAGAGCAAAATTCGCGCTTGCAGTCTCTCGATGGTCAACCAACCCCGCTGCTTATGCTCCGCTGACCGCCCTCATTCGTGTCT
5530 5540 5550 5560 5570 5580 5590 5600 5610 5620 5630 5640

S D A E I L A H Q T A K O V A L H R D A K P T W H K R C O L C *
TCCAGCGGGAAATCTAGCACCAACCACTGACAGAGTGGCCCTCCACCGTACGCCAAACCGACTGGCACAAACGTTGTACTCTGTGTAACTACTAATGGTCCCTCGACCC
5650 5660 5670 5680 5690 5700 5710 5720 5730 5740 5750 5760

T A T T G C C C T T A T T A C T A T C C C A G T A A T T G C C T T A C T C A C T T A A T A T A T A T A T T T C A G ( A ) n
5770 5780 5790 5800 5810 5820 5830 5840

```

Fig. 2. Nucleotide sequence of WC1MV. The complete DNA sequence of 5845 nucleotides derived from the clones shown in Fig. 1 is presented. The predicted amino acid sequences of the five open reading frames are shown above the DNA sequence in single letter code. In additional clones, G residues were found at positions 1640 and 3546. The alteration at position 1640 is silent; the alteration at position 3546 would change the corresponding amino acid from Thr to Ala.

Coding capacity of WC1MV genomic RNA

Computer analysis of the WC1MV sequence revealed five ORFs (Fig. 1b), coding for proteins of Mr 147,417, Mr 26,356 and Mr 12,989, in addition to the 7K and coat protein ORFs reported previously for WC1MV (15). The amino acid sequences of all five proteins are shown in Fig. 2.

The 147K ORF began 108 nucleotides from the 5' terminus and terminated with two in-frame UAG amber codons at nucleotides 3990 and 4080 and a UGA codon at nucleotide 4104. The two possible read-through proteins terminating at the second and third termination codons respectively had estimated Mr values of 150,528 and 151,266. No evidence for readthrough *in vivo* or *in*

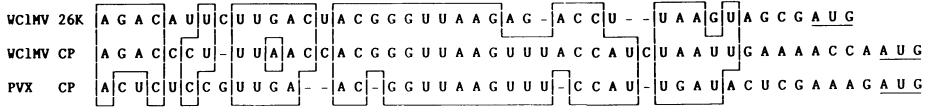


Fig. 3. Conserved RNA sequences 5' of potexvirus open reading frames. The nucleotide sequence is shown for the regions of the genomic RNA preceding the initiation codons (underlined) of the WC1MV 26K protein, the WC1MV coat protein (CP), and the potato virus X (PVX) coat protein. Boxes indicate identical aligned nucleotides. Gaps (-) have been introduced for maximum alignment of identical nucleotides.

(a)

```

WC1MV GAGSGKSHAIQTWMSLNRRDRHVTIILPTDLRNDWTTKVPNLEQANFKTFEKALCQP. 621
TMV   GVPGCKTKEILSRVNFEDLILVPGQAEMIRRRANSSGII VATKNVKTVDSFPMNFG 893
      * * * * *
WC1MV ...CGKII VFDDYSKLPQGYIEAFLAINQNVLAILTGDSSQSF...HHESNEDAYTAT 682
TMV   KSTRCQFKRLFDI DEGLMLHTGCVNFLVAMSLCEIAYVYGD TQOIPYINRVSGPPYPAHFAK 954
      | : | | : | : | : | : | : | : | : | : | : | : | : | : | : | : |
WC1MV LE. PSINTYQPFRCRYLNI THRNKPDLANLKGVYSCSSGTSFTMSSQALKGMPILSPSIM 742
TMV   LEVDEVETRRITLRCPADVTHYLNRRYEGFVMSTSSVKKSVSQEMVGGAAVINPISKPLHG 1015
      | : | | : | : | : | : | : | : | : | : | : | : | : | : | : | :
WC1MV K.....KTALGEMG.QKSMTYAGCQGLTTRAVQILLDTNTP...LCSSNVIYALSR 790
TMV   KILTFQSDKKEALLSRGYSDVHTVHEVOGETYSDVSLVRLTPTTVSIIAGDSPHVLVLSR 1076
      * * * * *
    
```

(b)

```

WC1MV PADNSFILDEWNNFRNFCFSN.DFTAFDQSDGSILOFEVIRAKFHNIPEDIIEGYIQIK 1114
TMV   PAQIEDFFGDLDSHPMDVLELDISKYDKSQNEFHCVEYEIWRRLGFEDFLGEVWKQGH 1423
      + ++ ** **
WC1MV THAKIFLGTLSI.....MRLSGEGPTFDANTEANIAYHTTKFNIPCDAAQVYAGDDMSI 1168
TMV   RKTTLKDYTAGIKTCIYWQRKSGDVTFIGNVTVIIAACLASHLPMKEIIRKGAFCGDDSL 1483
      + * * * * *
    
```

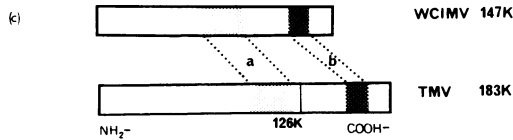


Fig. 4. Amino acid homology between the WC1MV 147K protein and the tobacco mosaic virus (TMV) 183K protein. Identical aligned amino acids are indicated by vertical lines. Aligned residues with similar biochemical properties (32) are indicated by double dots. In Fig. 4(a), asterisks below the sequences indicate positions of residues conserved between the TMV protein and non-structural proteins of three other RNA viruses (24). In Fig. 4(b), asterisks denote positions of strictly conserved residues in the putative RNA-dependent RNA polymerases of a larger sample of RNA viruses (26,27). Plus signs denote positions of biochemically similar amino acids in these putative polymerase enzymes (26,27). (c) Location of the homologous amino acids. Shaded regions of homology in the proteins correspond to parts of the figure above.

in vitro exists however. The 26K ORF began at nucleotide 3995 and extended into the first 26 nucleotides of the 13K ORF. The 13K ORF began at nucleotide 4683 and extended into the first 77 nucleotides of the 7K ORF.

Computer analysis of the nucleotide sequence revealed two other ORFs coding for proteins greater than 10K. One ORF encoded a protein of Mr 10,081, beginning at nucleotide 1360. It was entirely contained within the 147K ORF but was in a different reading frame. The other ORF encoded a protein of Mr 20,625 on the negative strand.

The alignment of nucleotide sequences 5' to the initiation codons of the 26K and coat protein ORFs revealed a region of significant nucleotide homology (Fig. 3). Twenty-six nucleotides were able to be aligned correctly if four single gaps were introduced into a region of 38 positions. Similar stretches of nucleotides were found upstream of the coat protein ORFs of PVX (Fig. 3) and PAMV (13). This nucleotide sequence was not found preceding the WC1MV 13K and 7K ORFs.

Homology between WC1MV proteins and proteins of other viruses

The 147K protein. Computer analysis of the 147K protein revealed two domains of homology with the 126K and 183K proteins of tobacco mosaic virus (TMV) (Fig. 4) and the corresponding proteins of other RNA viruses. The domain shown in Fig. 4(a), located between amino acids 571 and 790 of the WC1MV 147K protein, contained 17 of the 27 amino acids that were conserved between the corresponding domain of the TMV 126K/183K proteins and those of three other RNA viruses (24). This domain of the TMV 126K/183K proteins, and the corresponding protein domain of other RNA viruses, contained a putative RNA-binding motif (25). A similar motif was found in a similar position in the WC1MV 147K protein. A second domain on the WC1MV 147K protein was homologous to the read-through region of the TMV 183K protein (Fig. 4(b)). This domain contained all seven of the identical amino acids and 13 out of the 19 biochemically similar amino acids, that were conserved between the putative RNA-dependent RNA polymerases of RNA viruses (26,27). Included in the homology was the invariant amino acid motif, GDD, found in all these RNA polymerases.

The 26K protein. The 26K protein shared homology with a portion of the 58K protein of barley stripe mosaic virus (BSMV; 28) RNA β , as shown in Fig. 5(a). Homology has previously been noted between the same region of the BSMV 58K protein and the 42K protein of beet necrotic yellow vein virus (BNYVV) RNA-2 (29).

Fig. 5(b) shows a region conserved in the N-terminal portion of the

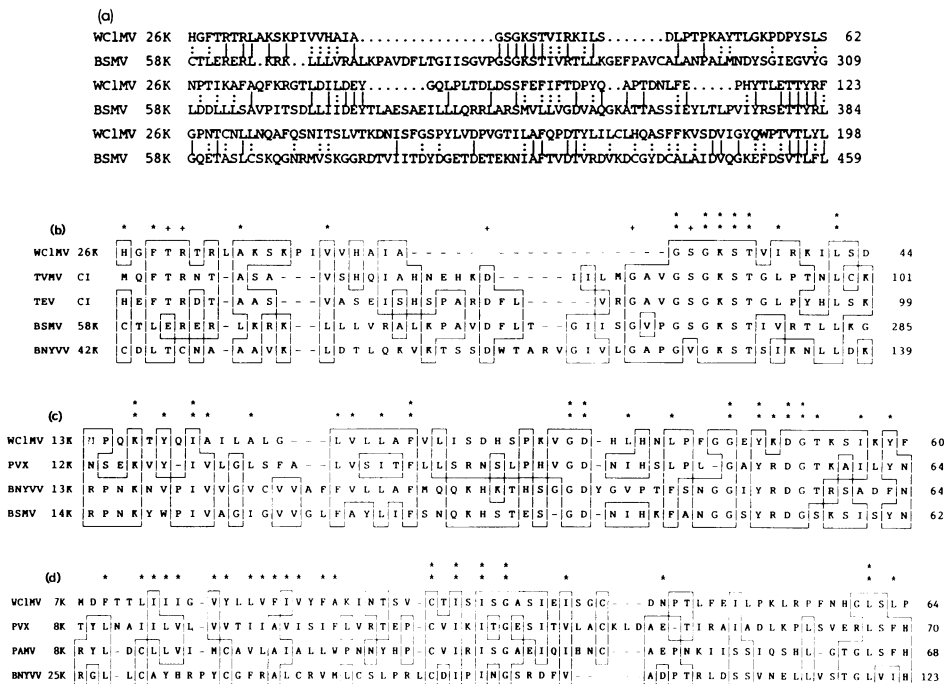


Fig. 5. Homology between the 26K, 13K and 7K proteins of WC1MV and proteins of other RNA viruses. (a) BESTFIT alignment of the WC1MV 26K protein and the corresponding region of the 58K protein of barley stripe mosaic virus (BSMV) RNA β . Symbols for amino acid homology are as in Fig. 4. (b) Homology between the N-terminal portion of the WC1MV 26K protein, the corresponding regions of the BSMV 58K protein and beet necrotic yellow vein virus (BNYVV) 42K protein, and portions of the cylindrical inclusion proteins (30,31) of tobacco vein mottling virus (TVMV) and tobacco etch virus (TEV). (c) Homology between a portion of the WC1MV 13K protein, the PVX 12K protein, the 14K protein of BSMV RNA β , and the 13K protein of BNYVV RNA-2. (d) Homology between portions of the WC1MV 7K protein, PVX and potato aucuba mosaic virus (PAMV) 8K proteins and the 25K protein of BNYVV RNA-3. In (b), (c) and (d), boxes indicate identical aligned amino acids. Double asterisks denote strictly conserved residues. Single asterisks denote aligned biochemically similar residues (32) in all proteins. Plus signs in (b) denote identical residues in four of the five proteins.

WC1MV 26K protein, the corresponding regions of the BSMV 58K protein and BNYVV 42K protein, and in a portion of the polyproteins (probably the cylindrical inclusion proteins) (30,31), of two potyviruses, tobacco vein mottling virus (TVMV, 30) and tobacco etch virus (TEV; 32). A possible mononucleotide-binding motif (33) was located in a similar position in the conserved region of each protein. The conserved stretch of amino acids in

the N-terminal hydrophobic area and part of the stretch of uncharged or hydrophilic amino acids.

The 7K protein. Homology has previously been noted between the WC1MV 7K protein and the 8K proteins of PVX and PAMV (15). In addition, these three potexvirus proteins were all similar to a portion of the 25K protein of BNYVV RNA-3 (35), as shown in Fig. 5(d).

The 5' and 3' termini

Similarity between the 5'-terminal regions of PMV and WC1MV was evident in the possible secondary structures which can be formed. Using the UWGCG FOLD program the terminal 132 nucleotides of WC1MV were aligned to give a stable structure (-41.3 Kcal, see Fig. 6(a)) with two hairpin-like structures similar to those reported for PMV (36). The 5'-terminal region has been implicated as the origin of assembly of the potexvirus papaya mosaic virus (PMV;36). The sequence of the 40 nucleotides at the 5' terminus of these two viruses, and of PVX (37), were very similar, as shown in Fig. 6(b). However, the WC1MV and PVX 5' terminal sequences lacked the consecutive repeating pentamers present in this region of PMV RNA (36).

The 60 nucleotides adjacent to the 3'-terminal poly (A) tract of WC1MV RNA and the predicted 60 nucleotides at the 3' terminus of the negative strand contained respectively 43% and 48% of U residues. Both strands contained the sequence UUCUGUUUA, separated by 12 nucleotides from the 3'-terminal poly (A) tract of the genomic strand and by 21 nucleotides from the 3' terminus of the negative strand. A small amount of additional homology also occurred 5' to this motif. This homology may be important for viral RNA replication.

DISCUSSION

This paper describes the complete nucleotide sequence of a potexvirus for the first time. The genomic RNA of WC1MV contained 5845 nucleotides and a tract of poly (A) residues. ORFs coding for proteins of 147K, 26K, 13K, 7K and the coat protein extended from nucleotide 108 to 109 nucleotides from the 3' poly (A) tract.

The 147K ORF probably corresponds to an in vitro translation product of 160K which is synthesized from the genomic RNA (9). In addition, hybrid-arrested translation experiments (results not presented) supported this conclusion. The homology between the 147K amino acid sequence and the TMV 126K and 183K proteins and other putative RNA-dependent RNA polymerases suggested a similar function for the WC1MV 147K protein.

There was no direct experimental evidence for the existence of protein products from the 26K, 13K and 7K ORFs. However, the homology observed between the predicted amino acid sequences of these proteins and other viral proteins suggested that these ORFs do code for functional proteins. Available evidence suggested that the smaller proteins of WC1MV were translated from subgenomic RNAs. The sequence of highly conserved nucleotides upstream of the ORFs for the coat protein and 26K ORFs of WC1MV, and the coat protein ORFs of other potexviruses, may be part of a subgenomic promoter.

The function of the 26K protein is unknown. However, the potyvirus cylindrical inclusion proteins, which had homology with the 26K protein, were suggested to have had a role in cell-to-cell movement on the basis of their association with plasmodesmata early in the infection process (38).

No functions have yet been ascribed to the 12K to 14K putative membrane-bound proteins of PVX, BSMV and BNYVV. Therefore, we cannot infer a function for the 13K ORF of WC1MV. However, the 25K protein of BNYVV RNA-3, which had homology with the 7K and 8K proteins of potexviruses, was thought to be directly involved in the natural infection of BNYVV by the fungal vector Polymyxa betae (39). On the basis of the homology between the 7K protein of WC1MV and both the 25K protein of BNYVV and the 8K protein of PVX, another virus for which a fungal vector has been reported (40), we predict that WC1MV may also have a fungal vector. However, the role of Synchytrium endobioticum as a PVX vector is no longer certain (41). Although WC1MV spreads naturally in the field (P.R. Fry, pers. comm.), no vector has yet been identified.

As might be expected, all of the protein sequences available from four potexviruses showed a degree of similarity. Homology has been noted between the coat proteins (14,15), 7K and 8K proteins (14,15) and the 12K and 13K proteins. In addition to the protein homology, the 5' termini of the three potexviruses reported to date also shared homology and were able to form similar secondary structures.

The inter-viral comparisons between the smaller potexvirus proteins described in this paper and elsewhere (14,15) suggested that the potexvirus group may be similar to the potyviruses, the hordeiviruses and the furoviruses. For example, the 26K protein of WC1MV showed homology between all these groups and the 12K and 13K proteins of potexviruses had homology to the hordeiviruses and furoviruses. The coat proteins of the potexviruses also had homology to the coat proteins of the potyvirus group (14,15) but not to the coat proteins of the other two groups, while the 7K and 8K proteins of

the potexviruses may be related to the 25K protein of the furovirus BNYVV. However, taken together, the genome size, the number and nature of the ORFs, and the RNA terminal structures clearly distinguish potexviruses from all other RNA viruses.

ACKNOWLEDGEMENTS

We thank P.J. Guilford for helpful discussions and K.A. Boxen and D.V. Faulds for excellent technical assistance. M.W.B. was supported by a N.Z. National Research Advisory Council Fellowship. S-A.H. was supported by a Post Doctoral Research Fellowship awarded by the N.Z. University Grants Committee.

*To whom correspondence should be addressed

+Present address: Department of Molecular Genetics, IPSR Cambridge Laboratory, Maris Lane, Trumpington, Cambridge CB2 2LQ, UK

REFERENCES

1. Koenig, R. (1971) *J. Gen. Virol.* 10, 111-114.
2. AbouHaidar, M. and Bancroft, J.B. (1978) *J. Gen. Virol.* 39, 559-563.
3. Sonenberg, N., Shatkin, A.J., Ricciardi, R.P., Rubin, M. and Goodman, R.M. (1978) *Nucleic Acids Res.* 5, 2501-2512.
4. AbouHaidar, M.G. (1983) *Can. J. Micro.* 29, 151-156.
5. Wodnar-Filipowicz, A., Skrzeczkowski, L.J. and Filipowicz, W. (1980) *FEBS Lett.* 109, 151-155.
6. Bendena, W.G., AbouHaidar, M. and Mackie, G.A. (1985) *Virology* 140, 257-268.
7. Bendena, W.G. and Mackie, G.A. (1986) *Virology* 153, 220-229.
8. Guilford, P.J. and Forster, R.L.S. (1986) *J. Gen. Virol.* 67, 83-90.
9. Forster, R.L.S., Guilford, P.J. and Faulds, D.V. (1987) *J. Gen. Virol.* 68, 181-190.
10. Bendena, W.G., Bancroft, J.B. and Mackie, G.A. (1987) *Virology* 157, 276-284.
11. Short, M.N. and Davies, J.W. (1983) *Biosci. Rep.* 3, 837-846.
12. Dolja, V.V., Grama, D.P., Morozov, S.Yu. and Atabekov, J.G. (1987) *FEBS Lett.* 214, 308-312.
13. Bundin, V.S., Vishnyakova, O.A., Zakharyev, V.M., Morozov, S.Yu., Atabekov, J.G. and Skryabin, K.G. (1986) *Dokl. Acad. Nauk. SSSR* 290, 728-733.
14. Morozov, S.Yu., Lukasheva, L.I., Chernov, B.K., Skryabin, K.G. and Atabekov, J.G. (1987). *FEBS Lett.* 213, 438-442.
15. Harbison, S-A., Forster, R.L.S., Guilford, P.J. and Gardner, R.C. (1987) *Virology*, submitted.
16. Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual*. New York: Cold Spring Harbor Laboratory.
17. Gubler, U. and Hoffman, B.J. (1983) *Gene* 25, 263-269.
18. D'Alessio, J.M. (1985) *RNA Sequencing*. In 'Gel Electrophoresis of Nucleic Acids: A Practical Approach'. (Rickwood, D., Hames, B.D., Eds.) IRL Press Ltd., Oxford and Washington DC.

19. Ahlquist,P., Dasgupta,R., Shih,D.S., Zimmern,D. and Kaesberg,P. (1979) *Nature* 281, 277-282.
20. Cowie,A., Tyndall,C. and Kamen,R. (1981) *Nucleic Acids Res.* 9, 6305-6322.
21. Chen,E.J. and Seeburg,P.H. (1985) *DNA* 4, 165-170.
22. Henikoff,S. (1984) *Gene* 28, 351-359.
23. Varma,A., Gibbs,A.J. and Woods,R.D. (1970) *J. Gen. Virol.* 8, 21-32.
24. Ahlquist,P., Strauss,E.G., Rice,C.M., Strauss,J.H., Haseloff,J. and Zimmern,D. (1985) *J. Virol.* 53, 536-542.
25. Goldbach,R. (1987) *Microbiol. Sci.* 4, 197-202.
26. Kamer,G. and Argos,P. (1984) *Nucleic Acids Res.* 12, 7269-7282.
27. Guilley,H., Carrington,J.C., Balazs,E., Jonard,G., Richards,K. and Morris,T.J. (1985) *Nucleic Acids Res.* 13, 6663-6677.
28. Gustafson,G. and Armour,S.L. (1986) *Nucleic Acids Res.* 14, 3895-3909.
29. Bouzoubaa,S., Quillet,L., Guilley,H., Jonard,G. and Richards,K. (1987) *J. Gen. Virol.* 68, 615-626.
30. Domier,L.L., Franklin,K.M., Shahabuddin,M., Hellmann,G.M., Overmeyer,J.H., Hiremath,S.T., Siaw,M.F.E., Lomonosoff,G.P., Shaw,J.G. and Rhoads,R.E. (1986) *Nucleic Acids Res.* 14, 5417-5430.
31. Domier,L.L., Shaw,J.G. and Rhoads,R.E. (1987) *Virology* 158, 20-27.
32. Allison,R., Johnston,R.E. and Dougherty,W.G. (1986) *Virology* 154, 9-20.
33. Higgins,C.F., Hiles,I.D., Salmond,G.P.D., Gill,D.R., Downie,J.A., Evans,I.J., Holland,I.B., Gray,L., Buckel,S.D., Bell,A.W. and Hermodson,M.A. (1986) *Nature* 323, 448-450.
34. Bouzoubaa,S., Ziegler,V., Beck,D., Guilley,H., Richards,K. and Jonard,G. (1986) *J. Gen. Virol.* 67, 1689-1700.
35. Bouzoubaa,S., Guilley,H., Jonard,G., Richards,K. and Putz,C. (1985) *J. Gen. Virol.* 66, 1553-1564.
36. Lok,S. and AbouHaidar,M.G. (1986) *Virology* 153, 289-296.
37. Morozov,S.Yu., Zakharyev,V.M. Chernov,B.K., Prasolov,V.S., Kozlov,Yu.V., Atabekov,J.G. and Skryabin,K.G. (1983) *Dokl. Acad. Nauk. SSSR* 271, 211-215.
38. Andrews,J.H. and Shalla,T.A. (1974) *Phytopathology* 64, 1234-1243.
39. Lemaire,O. and Merdinoglu,D. (1987) Abstracts of the Association of Applied Biologists Meeting on Viruses with Fungal Vectors (St. Andrews, Scotland) p16-17.
40. Nienhaus,F. and Stille,B. (1965) *Phytopath. Z.* 54, 335-337.
41. Lange,L. and Olson,L.W. (1979) *Phytopath. Z.* 95, 217-227.