Properties of *Commelina* yellow mottle virus's complete DNA sequence, genomic discontinuities and transcript suggest that it is a pararetrovirus

Scott L.Medberry, B.E.L.Lockhart[1] and Neil E.Olszewski
Departments of Plant Biology and [1]Plant Pathology, Plant Molecular Genetics Institute, University of Minnesota, St Paul, MN 55108, USA

Received May 21, 1990; Revised and Accepted August 20, 1990 EMBL accession no. X52938

ABSTRACT

The non-enveloped bacilliform viruses are the second group of plant viruses known to possess a genome consisting of circular double-stranded DNA. We have characterized the viral transcript and determined the complete sequence of the genome of *Commelina* yellow mottle virus (CoYMV), a member of this group. Analysis of the viral transcript indicates that the virus encodes a single terminally-redundant genome-length plus 120 nucleotide transcript. A fraction of the transcript is polyadenylated, although the majority of the transcript is not polyadenylated. Analysis of the genome sequence indicates that the genome is 7489 bp in size and that the transcribed strand contains three open reading frames capable of encoding proteins of 23, 15 and 216 kd. The function of the 25 and 15 kd proteins is unknown. Similarities between the 216 kd polypeptide and the cauliflower mosaic virus coat protein and protease/reverse transcriptase polyprotein suggest that the 216 kd polypeptide is a polyprotein that is proteolytically processed to yield the virion coat protein, a protease, and replicase (reverse transcriptase and ribonuclease H). Each strand of the CoYMV genome is interrupted by site-specific discontinuities. The locations of the 5'-ends of these discontinuities, and the presence and location of a region on the CoYMV transcript capable of annealing with the 3'-end of cytosolic initiator methionine tRNA are consistent with replication by reverse transcription. We have demonstrated that a construct containing 1.3 CoYMV genomes is infective when introduced into *Commelina diffusa*, the host for CoYMV, using *Agrobacterium*-mediated infection.

INTRODUCTION

Commelina yellow mottle virus (CoYMV) infects *Commelina diffusa* and is a member of the non-enveloped bacilliform virus group (1). Recently, CoYMV as well as group members infecting banana, canna, *Kalanchoe*, rice, *Schefflera* and sugarcane have been shown to possess genomes consisting of circular double-stranded DNA (2, 3, 4, Lockhart, unpublished; R. Hull and R. Beachy personal communications). The CoYMV genome is 7.5 kb and is not covalently closed because each strand of the genome is interrupted once by a site-specific discontinuity. The name badnaviruses has recently been proposed for group members possessing a dsDNA genome.

The other known group of dsDNA containing plant viruses are the caulimoviruses (for recent reviews see 5, 6). The caulimovirus genomes are approximately 8 kb in size and, like the CoYMV genome, both strands are interrupted by site-specific discontinuities. Unlike CoYMV, the caulimovirus genome is generally interrupted by three discontinuities although the genome of cauliflower mosaic virus (CaMV) isolate CM4-184 has only two discontinuities (7).

The caulimoviruses are believed to replicate their genome by reverse transcription of a greater-than-genome-length, terminally-redundant transcript. It is believed that a virally-encoded replicase (reverse transcriptase and ribonuclease H) is involved in this process and that cytosolic initiator methionine tRNA ($tRNA^{met}_i$) serves as the primer for minus-strand synthesis. The nomenclature pararetrovirus has been suggested for viruses possessing these properties (8).

Our analysis indicates that CoYMV encodes a single terminally-redundant genome-length plus 120 nt transcript. In addition, we have constructed two full-length CoYMV genomic clones and determined the sequence of one strand of each. Analysis of the sequence has identified three open reading frames (ORFs). We present and discuss evidence that the largest ORF encodes a polyprotein that is probably processed to yield the coat protein, protease, and viral replicase (reverse transcriptase and ribonuclease H). In addition, mapping of the site-specific discontinuities suggests that cytosolic initiator methionine tRNA ($tRNA^{met}_i$) and a purine-rich oligonucleotide prime reverse transcriptase-directed minus- and plus-strand synthesis, respectively. Finally, we demonstrate that the cloned CoYMV genome is infective following introduction into *C. diffusa*.

MATERIALS AND METHODS

Construction of genomic clones

CoYMV virions and virion DNA were purified as described previously (4). Genomic clones were constructed by ligating CoYMV DNA that had been partially digested with either *Cla*I or *Sac*I into pBluescript KS+ (Stratagene). Two clones containing

the entire CoYMV genome were identified by restriction mapping. Maps of pCoYMV89 and pCoYMV100 are shown in Figure 1C.

RNA isolation and Northern analysis

Total RNA was extracted from young infected and uninfected leaves of greenhouse grown C. *diffusa* using the phenol/SDS method and fractionated using the oligo(dT)-cellulose chromatography protocol in Ausubel et al. (9). The flow-through and first 1 ml of wash from the oligo(dT) column were collected as the non-polyadenylated fraction and the eluate as the polyadenylated fraction.

RNA was denatured and electrophoresed in a 1% agarose gel containing 2.2 M formaldehyde (10), transferred and UV cross-linked to GeneScreen (DuPont NEN Research Products), hybridized and washed according to recommendations of the membrane supplier. Hybridization with end-labeled polyuridine was performed at 50°C in 1 M NaCl, 1% SDS and the most stringent wash was performed at 50°C in 2× SSC and 1% SDS.

Strand-specific hybridization probes were made using pCoYMV89 and its approximately half-size derivatives pCoYMVR and pCoYMVL. The plasmid pCoYMVR contains CoYMV sequences from 6328 to 2736 and pCoYMVL contains sequences from 2918 to 6328. Radioactive single-stranded RNA probes that hybridize to either the plus- or minus-strand of the CoYMV genome were generated using either T7 RNA polymerase to transcribe pCoYMV89 and pCoYMVR or T3 RNA polymerase to transcribe pCoYMV89 and pCoYMVL, respectively. Transcription was performed according to the vector supplier. The plasmids pCoYMVR and pCoYMVL were included in the transcription reactions to ensure that the probe encompassed the entire genome. Polyuridine probes were prepared by end-labeling dephoshorylated poly(U) (Sigma; 11).

Mapping the transcript ends

To identify the transcript's 5′-end, the oligonucleotide 5′-CGAAACCTGGCTCTGATACCA-3′, that is similar in sequence to the 3′-end of wheat cytosolic $tRNA^{met}_i$ (12), was synthesized and used in primer extension analysis. The procedure used was an adaptation of the RNA sequencing protocol of Inoue and Cech (13). End-labeled primer was resuspended at 1×10^5 CPM/μl in 2× TK (1× TK is 100 mM Tris (pH 8.2), 100 mM KCl) and total RNA (20 μg) was resuspended by the addition of 5 μl primer and 2.5 μl H_2O. After annealing the primer-RNA mixture, 0.5 μl 120 mM $MgCl_2$, 0.5 μl RNAguard (Pharmacia), 1 μl 100 mM DTT, 1.0 μl of a mixture of dCTP, dATP, dGTP, and dTTP at 20 mM each, and 0.5 μl (20 units/μl) AMV reverse transcriptase (Life Sciences) were added. The reaction mix was incubated for 45 min at 42°C. The reaction was terminated by adding 6 μl of Stop Solution provided with the Sequenase Kit (US Biochemical). The primer extension product was run on a 6% denaturing acrylamide gel next to a pCoYMV89 sequence ladder generated using the primer extension oligonucleotide as a primer.

The location of the 3′-end of polyadenylated CoYMV transcript was determined by sequencing cDNA clones to map the junction between CoYMV sequences and the polyadenine tract. Double-stranded cDNA was made from 4 μg poly(A)$^+$ RNA by the method of Aruffo and Seed (14). Cloning into pBluescript KS- was carried out according to Olszewski et al. (15). CoYMV cDNA clones were identified by colony hybridization (11). The 1553 bp *Sac*I fragment of pCoYMV89 was used as a probe. Sequencing was performed as described below.

DNA sequencing

A series of sequencing templates (plasmids) were prepared from both pCoYMV89 and pCoYMV100 using the exonuclease III method of Henikoff (16). Minipreps of plasmid DNA were prepared using a method suggested by R. Pruitt (personal communication). Cells from 1.5 ml of a saturated overnight culture were collected by centrifugation in a microfuge for 30 sec, resuspended in 400 μl of resuspension buffer [0.2 M Tris-HCl (pH 8.0), 0.1 M EDTA, 1% (w/v) N-lauroyl sarcosinate and 75 μg/ml proteinase K] and incubated at 48°C for 15 min. Cellular debris and chromosomal DNA were pelleted by centrifugation in a microfuge for 10 min at 4°C. The pellet was removed from the bottom of the tube and the volume was adjusted to 400 μl by the addition of resuspension buffer. Nucleic acids were precipitated by the addition of 800 μl of ethanol containing 1 mM phenylmethylsulfonyl fluoride, collected by centrifugation in a microfuge for 5 min, washed with 70% ethanol, dried, and dissolved in 20 μl of TE [10 mM Tris-HCl (pH 8.0) and 1 mM EDTA]. This procedure generally yields 20 μg of plasmid DNA which is suitable for restriction analysis and sequencing.

Miniprep plasmid DNA was denatured and neutralized prior to sequencing essentially as described by Murphy and Kavanagh (17) except that Sepharose 4B was substituted for Sepharose CL-6B. Sequencing was carried out using 7 μl of neutralized DNA and a kit containing Sequenase (US Biochemical) as recommended by the manufacturer. In some cases, gaps in the sequence were filled in by sequencing from oligonucleotide primers. Computer analysis of the DNA sequence was performed using programs from the IntelliGenetics Suite.

Primer extension mapping of the 5′-end of the minus- and plus-strand discontinuities

Purified virion DNA (5−10 μg) was denatured and neutralized as described above for sequencing templates. Unless indicated, primer extension was performed essentially by the sequencing protocol using reagents provided with the Sequenase Kit. After the labeling reaction, instead of dispensing the labeling mix into tubes containing Termination Mix, 11 μl of a solution of 80 mM dATP, 80 mM dCTP, 80 mM dGTP, 80 mM dTTP and 50 mM NaCl was added. The reaction was incubated at 37°C for 5 min and terminated by the addition of 18 μl of the Stop Solution provided with the Sequenase Kit. Following denaturation by heating at 90°C for 3 min, the size of the reaction products was determined by electrophoresis on a sequencing gel. The sequences of the primers used to map the minus- and plus-strand 5′-ends were 5′-ATGCCGGTTCCCAAGC-3′ and 5′-CCTCATCTTT-TTCTCT-3′, respectively.

Construction of infective clones

Two constructs were prepared to test the infectivity of cloned DNA. The large *Sma*I/*Sph*I fragment of pCoYMV89, the small *Nde*I/*Sph*I fragment of pCoYMV89 and the small *Nde*I/*Sph*I fragment of pCoYMV100 were purified and ligated together to construct pCoinf. The other clone, pCoinf4 (Fig. 1C), was constructed by co-integration of pCoinf via its *Xho*I site into the *Sal*I site of the binary vector pOCA28 (Olszewski, unpublished). This construct was used for *Agrobacterium*-mediated infection experiments. Ligation products from the reaction that produced

pCoinf4 were introduced directly into *A. tumefaciens* strain A281 (18) by electroporation (19) and propagated in this host because pCoinf4 is unstable in *E. coli* hosts. This is probably due to the presence of two ColE1 origins of replication on the plasmid. A281 transformants containing pCoinf4 were selected on solid LB medium containing carbenicillin (100 μg/ml), spectinomycin (75 μg/ml) and streptomycin (300 μg/ml). A281 containing pCoinf4 was propagated in LB medium containing carbenicillin 100 μg/ml.

Infection of plants using molecular clones

Two approaches were employed to introduce molecular clones of the CoYMV genome into *C. diffusa*. Stem cuttings containing the shoot apex and 3–4 subtending nodes were rooted in soil for 2 weeks prior to use. Plant material was propagated in a growth chamber at 23°C under a 16:8 day:night cycle. In the first approach, pCoinf DNA was introduced into abraded *C. diffusa* leaves using the methods that have been described for the introduction of CaMV DNA into turnip leaves (20). In the second approach, pCoinf4 was introduced into wounded stems using *Agrobacterium*-mediated infection (21). The second internode below the apex was wounded and inoculated by stabbing it with a sharp toothpick that had been dipped in a saturated culture of *A. tumefaciens*. Alternatively, 5 μl of a saturated culture of *A. tumefaciens* that had been collected by centrifugation and suspended in an equal volume of sterile H_2O was injected into the second internode.

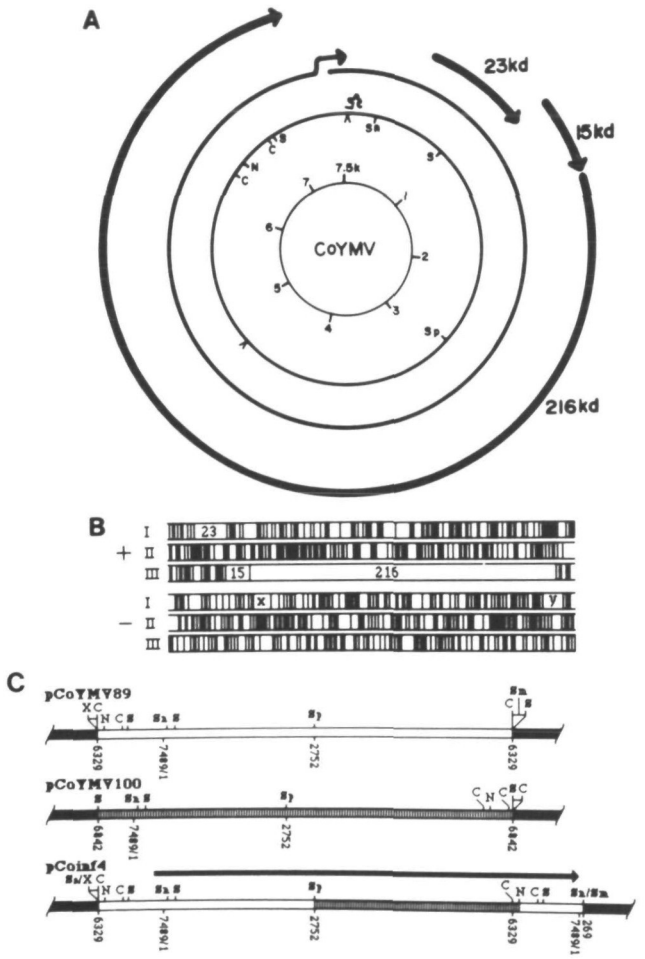

Figure 1. Organization of the CoYMV genome and construct maps. (**A**) The locations of the plus- and minus-strand discontinuities are indicated by the arrowheads located outside and inside the outer circle, respectively. The location of the $tRNA^{met}_i$ binding site is indicated by the cloverleaf. The location and extent of the ORFs are indicated by the heavy arrows. The thinner arrow indicates the direction and extent of the CoYMV transcript. (**B**) The ORFs contained in each reading frame of both the plus- (+) and minus-(−) strands are indicated below the map. Each frame was divided into groups of 10 codons and groups containing a stop codon are marked. (**C**) Open regions indicate pCoYMV89 viral DNA, the regions with vertical lines indicate pCoYMV100 viral DNA and the solid regions indicate vector DNA. The numbers below each map indicate the location on the CoYMV genome in bp. The arrow above the map of pCoinf4 indicates the location of the CoYMV transcript. Restriction sites are indicated as follows: C, *Cla*I; N, *Nde*I; S, *Sac*I; Sa, *Sal*I; Sm, *Sma*I; Sn, *Sna*BI; Sp, *Sph*I; X, *Xho*I.

RESULTS

Transcript characterization

When Northern blots containing RNA extracted from uninfected and infected *C. diffusa* are hybridized with strand-specific CoYMV probes, one probe does not hybridize to total RNA from uninfected tissues but does hybridize to a 7.6 kb transcript present in total RNA isolated from infected tissues (Fig. 2A, lanes 1 and 2). The other probe hybridizes only to a 1.1 kb transcript present in both infected and uninfected plants (Fig. 2A, lanes 3 and 4). This transcript is host-encoded since it is present in healthy tissue. The function, if any, of this transcript is unknown. These results indicates that only one strand of the genome is transcribed. Hereafter we will refer to the transcribed strand as theminus-strand and the non-transcribed strand as the plus-strand.

There is a considerable smear of hybridization below the 7.6 kb transcript (Fig 2A). This hybridization couild be due to the presence of many additional smaller viral transcripts or it could be due to degradation products from the larger transcript because

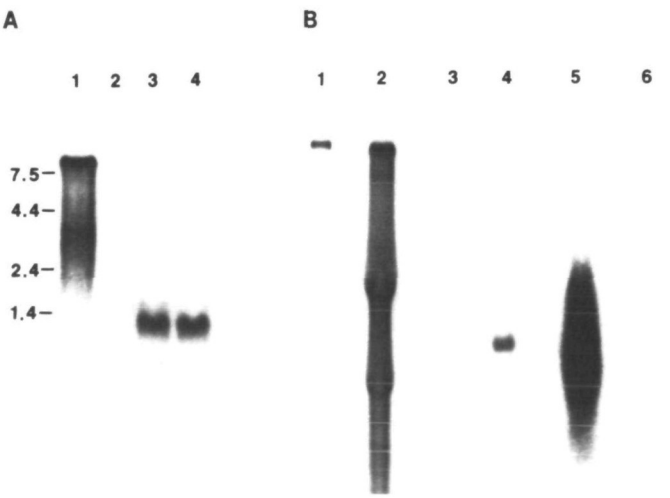

Figure 2. Northern blots of total RNA from CoYMV infected and uninfected *C. diffusa*. (**A**) Lanes 1 and 2 were loaded with 0.1 mg of total RNA. Lanes 3 and 4 were loaded with 10 mg of total RNA. Lanes 1 and 3 contain RNA from infected plants and lanes 2 and 4 contain RNA from uninfected plants. (**B**) All lanes contain RNA from infected plants. Lanes 1, 3 and 5 were loaded with 30 ng polyadenylated RNA and lanes 2, 4 and 6 were loaded with 3 μg non-polyadenylated RNA. In panels **A** and **B**, a probe that detects the plus-strand was used in lanes 1 and 2 and a probe that detects the minus-strand was used in lanes 3 and 4. Lanes 5 and 6 in panel **B** were hybridized with an end-labeled poly(U) probe.

it is a continuous smear extending from the 7.6 kb transcrpit rather than being composed of discretely sized RNAs. These degradation products probably arise *in vivo* rather than during RNA isolation because the phenomena is reproducible between independent RNA preparations and there is no apparent degradation of the 1.1 kb host-encoded transcript. The region of more intense hybridization at 2.8 kb (Fig. 2A, lane 1) is not observed in poly(A)$^+$ RNA (Fig. 2B, lane 1) and probably is an artifact produced due to trapping of viral transcripts by ribosomal RNA. Although these results suggest that CoYMV encodes only a single transcript, it is not possible to preclude the presence of less-abundant transcripts because their presence could be obscured by the 7.6 kb transcript degradation products.

To determine whether the viral and host-encoded transcripts are polyadenylated, total RNA from infected *C. diffusa* plants was fractionated by oligo(dT)-cellulose column chromatography. Figure 2B shows Northern blots of these RNAs. The mass of polyadenylated RNA loaded onto the gel was 100-fold less than the mass of non-polyadenylated RNA so that the ratio of hybridization to polyadenylated RNA and non-polyadenylated RNA should reflect the relative abundance of these forms *in vivo*. The hybridization pattern indicates that, while a portion of the full-length transcript is polyadenylated, the majority of the viral transcript and the host-encoded transcript appear to be non-polyadenylated or to have insufficient polyadenylation for retention by the column. Hybridization with end-labeled poly(U) RNA indicates that virtually all of the RNA with hybridizable poly(A) tracts is present in the poly(A)$^+$ RNA fraction (Fig. 2B, lanes 5 and 6).

We can not preclude the possibility that some portion of the CoYMV transcript present in the poly(A)$^+$ RNA fraction is not polyadenylated and is a contaminate present due to insufficient washing of the column or trapping. However, CoYMV cDNA clones prepared from the poly(A)$^+$ fraction are clearly derived from polyadenylated RNA (Fig. 3) indicating that some of the CoYMV transcript is polyadenylated.

Mapping the transcript ends

The 3′-end of the viral transcript was mapped by sequencing through the junction between the genomic sequence and the start of polyadenylation in seven CoYMV cDNA clones. These clones are independent since no two are identical. The 3′-ends of the RNAs giving rise to these cDNAs all map to a 23 nt region (nucleotides 7464−7486; Fig. 3). This regions is located between a putative polyadenylation signal sequence (AATAAA) (22) and the tRNA$^{met}_i$ sequence complementarity (see below).

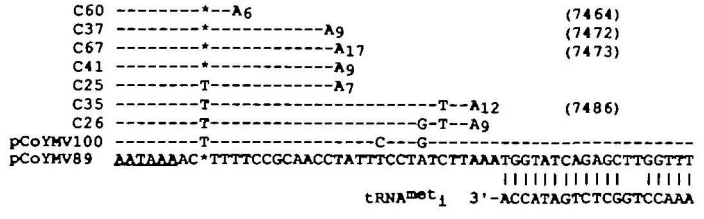

Figure 3. The location of the 3′-end of the transcript was determined by sequencing cDNA clones C25, C26, C35, C37, C41, C60 and C67. The sequence of this region of pCoYMV89 is shown below. Above this line is a representation of the sequences of pCoYMV100 and the seven cDNAs. A dash represents no change from the sequence of pCoYMV89 while a letter indicates the difference. The first adenosine of the polyadenosine tract of each cDNA is represented by a terminal A and the number of cloned adenosines in the tract is noted in subscript. The location of the nucleotide preceding the polyadenine tract is indicated in parenthesis. A T insertion in pCoYMV100 and three cDNAs is noted by an asterisk (*) in the sequences without the insertion. Below the sequence of pCoYMV89 is the sequence of the 3′-end of tRNA$^{met}_i$, showing its presumed hybridization to the genomic sequence. A sequence resembling a polyadenylation signal is underlined.

The heterogeneity in the sequences of the CoYMV cDNAs is not surprising because the virus inoculum used in these studies is probably heterogeneous. No local lesion host for CoYMV is available. Thus any variability present in the original isolate has probably been maintained. In addition, the virus has been maintained by serial propagation which would tend to promote the accumulation of additional variability.

To map the 5′-end of the viral transcript, a 21-mer oligonucleotide primer 5′-CGAAACCTGGCTCTGATACCA-3′ similar to the 3′-end of wheat cytosolic tRNA$^{met}_i$ was used in a primer extension reaction with RNA from infected *C. diffusa*. The results are shown in Fig 4. The extended primer is 156−157 nt long and terminates at two cytosine residues. No other major products are detected and no major primer extension product was seen using RNA from uninfected tissue (data not shown). Both primer extension with a primer complementary from position 7417 to 7434, 5′-CTTACTTCTCCGAAGAG-3′, and RNase protection experiments yield results consistent with those of the tRNA primer (data not shown). Based on this, the 5′-end of the transcript has been mapped to nucleotides 7354 and 7355. The mapping data indicate that the transcript is genome-length plus between 109 and 132 nt. The location of the transcript on the genome is illustrated in Figure 1A.

DNA sequence

The complete sequence of both strands of the CoYMV genome was obtained. To do this, the complete sequence of a single strand of each of two independent genomic clones, pCoYMV89 and pCoYMV100 (Fig. 1C), was determined.

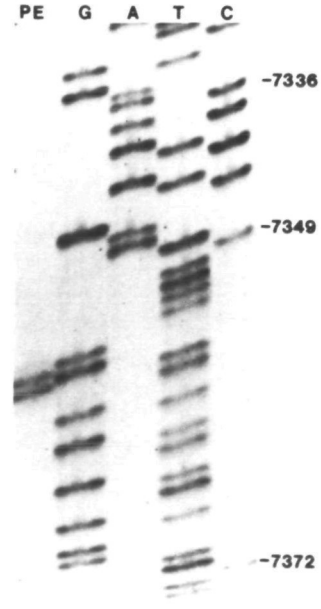

Figure 4. Primer extension analysis of the CoYMV transcript. The lane labeled PE indicates the lane containing the primer extension reaction products and the lanes labeled G, A, T and C represent single sequencing reactions to which the corresponding dideoxy nucleotide was added. The same primer was used for the primer extension and sequencing reactions. Nucleotide positions of three C residues are indicated.

The complete sequence of the plus-strand of pCoYMV89 is presented in Fig. 5. This CoYMV genome is 7489 bp in size and the G + C content is 39.6%. Numbering of the CoYMV sequence begins the 5'-end of the putative replication primer binding site (discussed below).

Comparisons of the sequences derived from pCoYMV89 and pCoYMV100 identified fifteen differences (Table 1). None of these differences introduce stop codons or cause frame shifts in the major open reading frames (ORFs). Due to the presumed heterogeneous nature of the virion DNA (see above), it is likely that the observed differences between these clones represents heterogeneity in the CoYMV virion DNA population rather than cloning artifacts.

Coding regions

Analysis of the CoYMV sequence identified five putative open reading frames (ORFs) capable of encoding proteins larger than 10 kd. Three of these ORFs are located on the plus-strand of the genome (Fig. 5) and potentially encode proteins of 23, 15 and 216 kd (Table 2). Two ORFs, designated x and y, capable of encoding proteins larger than 10 kd were identified on the minus-strand. Since our analysis of CoYMV transcripts has not detected any transcripts from the plus-strand (see above), it is likely that ORFs x and y are not expressed. It is possible that all of the plus-strand ORFs are expressed if this RNA serves as a polycistronic mRNA similar to what is proposed for CaMV (23, 24).

Computer analysis was performed to determine if similarities exist between the putative CoYMV-encoded proteins and any of the proteins contained in the PIR and Swiss protein data bases. This analysis identified no proteins with similarity to the 23 and 15 kd proteins. Even comparisons between these proteins and proteins encoded by the caulimoviruses, CaMV (25), CERV (26) and FMV (27) revealed no similarities. However, analysis of the 216 kd protein revealed similarity between portions of this protein and retroelement nucleocapsid protein (caulimovirus coat protein and the nucleocapsid portion of the *gag* protein), aspartic protease, reverse transcriptase and ribonuclease H (28, 29);

TABLE 1. Sequence differences between pCoYMV89 and pCoYMV100

Position[a]	pCoYMV89 Nucleotide	Amino[b] acid	pCoYMV100 Nucleotide	Amino acid
300–302	GAC	*	CGA	*
541	C	Pro	T	Ser
873	C	–[c]	T	–
1338–1340	GTA	Val	deletion	deletion
1970	G	–	A	–
2452	C	Thr	T	Ile
2977	T	Leu	C	Ser
4611	G	Ala	A	Thr
4794	C	Glu	A	Lys
6302	C	–	T	–
6434	T	–	C	–
7378	T	*d1	C	*
7462	T	*	TT	*
7478	T	*	C	*
7482	A	*	G	*

[a] Nucleotide positions refer to the position in the sequence of pCoYMV89.
[b] The amino acid encoded by the translated codon encompassing the nucleotide in question is indicated.
[c] – indicates no change.
[d] * indicates that the region is not translated.

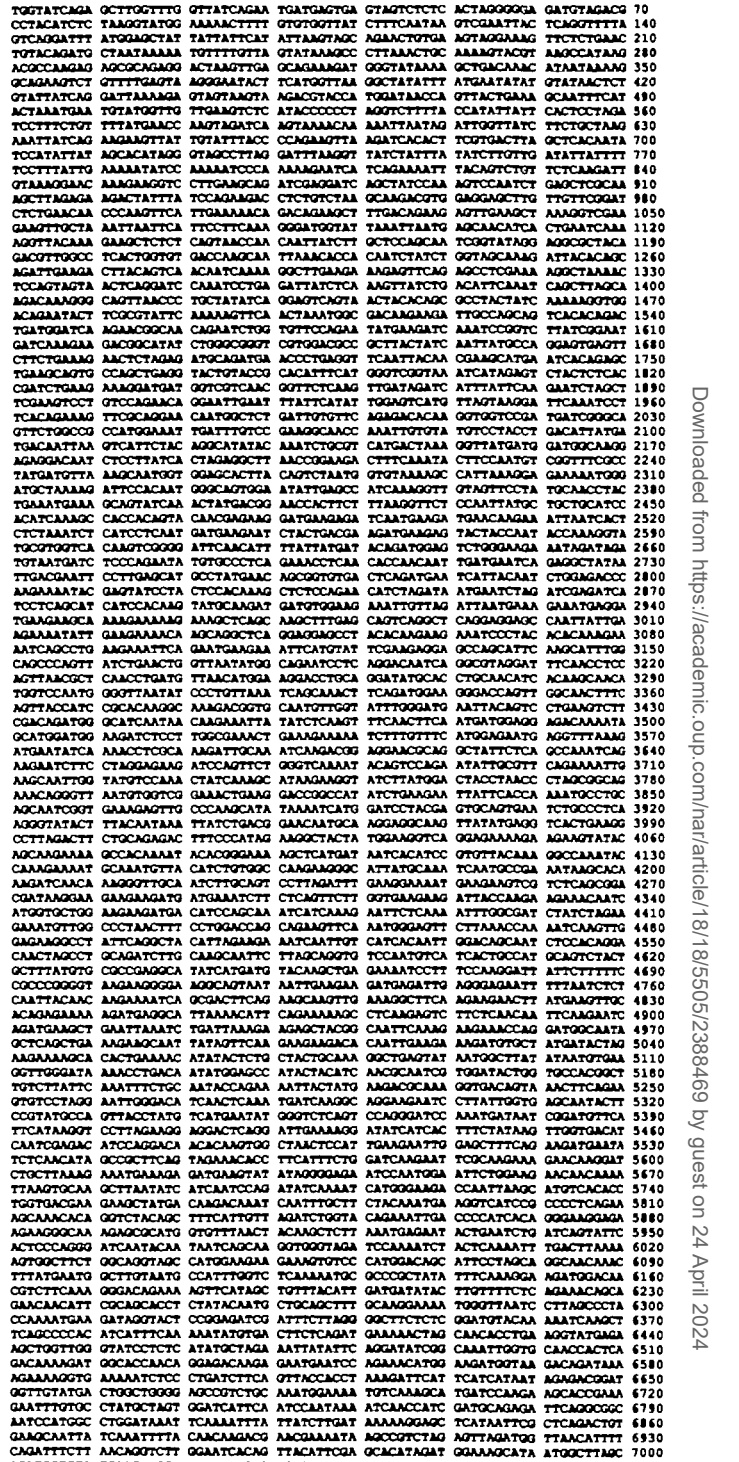

Figure 5. The nucleotide sequence of CoYMV DNA. The plus-strand sequence is presented and numbering begins at the 5'-end of the plus-strand discontinuity.

alignments of these regions from the 216 kd protein with the corresponding regions of the caulimovirus ORFs IV and V proteins are presented in Fig. 6.

Similarity to the retroelement nucleocapsid protein is limited to a zinc finger-like domain (30) that has the following sequence C-X_2-C-X_4-H-X_4-C. In the case of the caulimoviruses, it has been suggested that this motif plays a role in genome replication by binding the greater-than-genome-length transcript and sequestering it so that it can be acted upon by the viral replicases (31). The coat proteins of CaMV, CERV and FMV, are about 490 aa in size and the zinc finger-like motif is located about 425 aa from the amino terminus. In these proteins, the zinc finger-like sequence is preceded by a highly basic domain and followed by a terminal acidic domain. The zinc finger-like sequence of the 216 kd protein is preceded by a basic domain and followed by an acidic domain, but is located 878 aa from the amino terminus (Fig. 6). These observations suggest that at least (see discussion) a portion of the amino-half of the 216 kd protein becomes the virion coat protein.

When ORF V of CaMV, CERV, and FMV are aligned with the region of the 216 kd protein following the zinc finger-like domain, amino-acid identities of 33, 30 and 31% are observed, respectively (not shown). ORF V of the caulimoviruses is believed to encode a polyprotein containing domains with aspartic protease, reverse transcriptase and ribonuclease H activities which is post-translationally cleaved by the polyprotein protease to yield the 'mature' enzymes. Similarity is observed between the 216 kd protein and all of these domains (Fig.6). In addition, the spatial arrangement of these domains is similar. These similarities suggest that the 216 kd protein is a polyprotein consisting of the virus coat protein, an aspartic protease, and replicase (reverse transcriptase and ribonuclease H) that is post-translationally processed by the aspartic protease to yield these proteins. In addition, these observations suggest that CoYMV is a pararetrovirus that utilizes virally-encoded enzymes to replicate its genome by reverse transcription of the greater-than-genome-length transcript.

Table 2. Protein coding regions of the CoYMV genome.

ORF designation	Starting nucleotide	Ending nucleotide	Molecular weight (d)
23 kd	496	1095	23,335
15 kd	1098	1502	14,787
216 kd	1506	7163	215,673

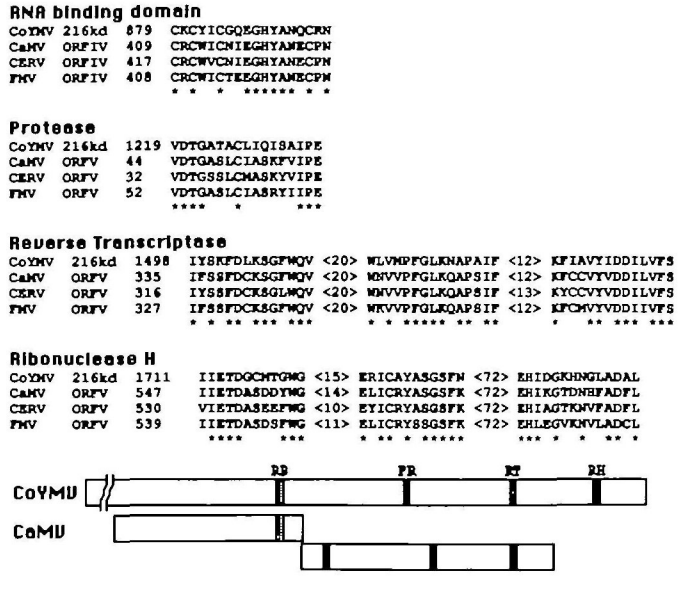

Figure 6. Comparisons between the CoYMV 216 kd ORF and CaMV (strain CM1841) (18), CERV (19) and FMV (20) proteins. Invariant amino acids are indicated with an asterisk. The ORF and starting amino acid are indicated before the sequence. In the case of the reverse transcriptase and ribonuclease H domains, the spacing between the motifs is indicated in brackets. The locations of these motifs in the 163 kd ORF of CoYMV, and ORFs IV and V of CaMV are shown below the alignments. The spatial arrangement of CaMV ORFs IV and V illustrates their arrangement on the CaMV genome.

Minus-strand replication primer

Cytosolic initiator methionine tRNA is believed to serve as the primer for caulimovirus minus-strand synthesis. Analysis of the CoYMV sequence suggests that CoYMV minus-strand synthesis is primed in a similar manner because a region that can potentially anneal with $tRNA^{met}_i$ is located at nucleotides 1−23 on the plus-strand of the genome (Fig. 7A). Nineteen out of 23 nucleotides located at the 3′-end of the wheat $tRNA^{met}_i$ (12) can potentially base pair with this region and thus the CoYMV transcript.

Mapping of the 5′-ends of the genomic discontinuities

If CoYMV utilizes $tRNA^{met}_i$ as a primer for minus-strand synthesis the 5′-end of the minus-strand discontinuity should map adjacent to the 3′-end of the $tRNA^{met}_i$ homology. To determine if this is the case, the location of 5′-end of the minus-strand discontinuity was determined by primer extension mapping. The

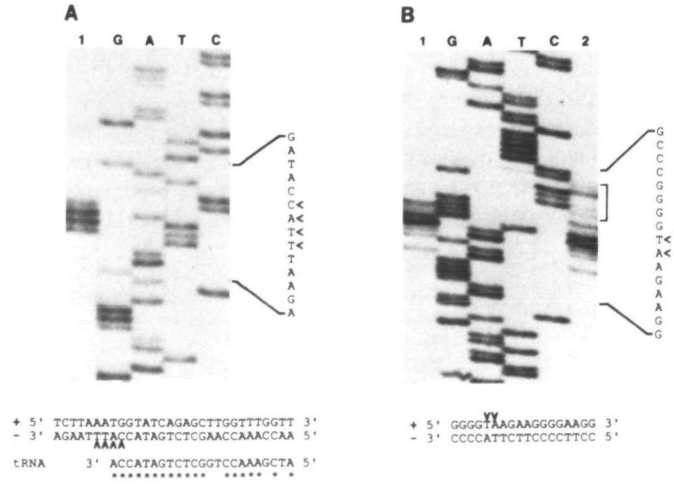

Figure 7. Primer extension mapping of the minus-strand (A) and plus-strand (B) discontinuities. Primer extension reaction products were subjected to electrophoresis adjacent to sequencing reactions that had been primed with the same primer. The sequencing gel is presented such that the complement to the extended strand can be read directly in a 5′- to 3′-direction by reading down the gel. This sequence is shown adjacent to the sequencing reactions. The sequence of both the plus- (+) and minus- (−) strand and the locations of the major 5′-ends of the discontinuities (indicated by the arrowheads) are shown below. (A) The sequence of the region from nucleotide 7483 to nucleotide 23 is shown below. The location and extent of sequence similarity to wheat $tRNA^{met}_i$ (12) is indicated. Similarities between the sequences are denoted by an asterisk. (B) The template for the primer extension reactions was either denatured CoYMV DNA (Lane 2) or denatured CoYMV DNA that was treated with RNase T1 and RNase A (Lane 1). Primer extension products that are absent following treatment with ribonuclease are indicated by bracket adjacent to Lane 2. The sequence below the figure encompasses the plus-strand polypurine-rich region (nucleotide 4696 to nucleotide 4713).

majority of 5′-ends of the discontinuity map to nucleotides 7488−2 (Fig. 7A). This location is adjacent to the 3′-end of the $tRNA^{met}_i$ homology supporting the suggestion that $tRNA^{met}_i$ serves as the primer for minus-strand synthesis.

The 5′-ends of this discontinuity appear to be heterogeneous, spanning four nucleotides. Some of this heterogeneity may be due to variation in the primary sequence of the viral DNA population. Approximately half of the DNA population used in this experiment has a one base pair deletion in the region through which the primer is extended (not shown). This accounts for some of the observed heterogeneity, but the remaining heterogeneity presumably reflects either heterogeneity of the 5′-ends of the encapsidated DNA population or heterogeneity that is generated during isolation of the DNA. Surprisingly, ends mapping to nucleotides 1 and 2 map inside of the tRNA homology. If intact $tRNA^{met}_i$ serves as a primer for minus-strand synthesis the 5′-end should map to nucleotide 4789. The reason for this result is unclear, but it is possible that the 3′-end of the tRNAs that serve as primers for minus-strand is heterogeneous.

Using primer extension mapping, the majority of the 5′-ends of the plus-strand have been mapped to nucleotides 4700 and 4701, in a polypurine-rich region (Fig. 7B). This is similar to what is observed with other retroelements where it is believed that the polypurine-rich region of the RNA transcript serves as a primer for plus-strand synthesis (32). The mechanism by which this primer is generated is unclear but possibilities include that the polypurine-rich region is either resistant to ribonuclease H digestion and persists or is a target for a specific cleavage event which generates the end that serves as a primer. Recent experiments suggest that retroelement-encoded ribonulease H activity possesses some specificity and that the primer may be generated by a specific cleavage (32 and references therein). Generally, plus-strand synthesis starts adjacent to the 3′-end of the polypurine-rich region. Surprisingly, in CoYMV, the 5′-end of the plus-strand maps inside the polypurine-rich region, with the majority of the ends mapping to the single pyrimidine that occurs in the region. This suggests that the plus-strand primer may be generated by a specific cleavage event. If primers were generated simply as a consequence of the resistance of polypurine-rich regions to digestion, then it would be predicted that the major plus-strand end would occur at the 3′-end (upstream) of the largest uninterrupted polypurine-rich region. However, the largest polypurine region is located directly 3′ (downstream) to the 5′-end of the plus-strand (Fig. 7B).

Primer extension mapping was performed using Sequenase 2. The properties of this enzyme have not been completely characterized. Since the 5′-ends of caulimovirus virion DNA have been shown to possess RNA tracts (presumably remnants of the RNA which served as a primer) and it was not clear whether Sequenase 2 possesses reverse transcriptase activity, we examined the effect of pretreatment of the denatured CoYMV DNA with ribonucleases T1 and A on the size of the primer-extended products. This treatment had no effect on either the major plus-strand (Fig.7B) or minus-strand (not shown) extension products. Thus these experiments have mapped the 5′-end of the DNA. However, minor larger extension products (Fig. 7B) did disappear following treatment suggesting that Sequenase 2 has some ability to use RNA as a template. It is not possible to determine the abundance of RNA tracts at the 5′-ends because, although the insensitivity of the major extension products to ribonuclease treatment may suggest that the majority of the 5′-ends lack an RNA tract, it is also possible that Sequenase 2 utilizes RNA templates inefficiently and thus underestimates the number of ends possessing RNA tracts.

Demonstration that the cloned CoYMV genome is infective

To demonstrate that the cloned CoYMV genome was infective we made and tested the infectivity of pCoinf and pCoinf4. These constructs contain about 1.3 CoYMV genomes and should be capable of producing a complete CoYMV transcript and cause an infection following introduction into *C. diffusa*.

Infection did not occur when pCoinf DNA was introduced into plants by either inoculation into abrasions on leaves or by injection (not shown). Similar results were also obtained when plants were treated with purified virion DNA, suggesting that intact DNA was not entering viable cells or that the DNA was degraded before a complete transcript could be produced.

Because *Agrobacterium*-mediated infection had been used successfully to introduce viral DNA constructs into monocot hosts and cause a systemic virus infection (33), we made and tested pCoinf4. Introduction of the pCoinf4 construct into *C. diffusa* by *Agrobacterium*-mediated infection causes the plants to become systemically infected (Table 3), demonstrating that the cloned genome is capable of producing an infection and that the DNA component of the virion contains the information necessary to cause an infection. The presence of CoYMV particles was confirmed by electron microscopy (not shown). When purified virions are inoculated into uninfected leaves symptoms appear 10−14 days after inoculation. When *Agrobacterium*-mediated infection is employed symptoms are first observed 16 days after inoculation. Other than developing more slowly, the symptoms observed on 'Agroinfected' plants are indistinguishable from those seen on plants infected using purified virus particles.

DISCUSSION

These characterizations of the CoYMV genome and its transcript have identified a number of features which strongly suggest that CoYMV is a pararetrovirus that replicates its genome by a mechanism similar to that employed by the caulimoviruses. These observations suggest that a virally-encoded reverse transcriptase is responsible for genome replication, that a genome-length plus 120 nt transcript is a template for minus-strand synthesis which is primed by $tRNA^{met}_i$, and that plus-strand synthesis is primed from a polypurine-rich region.

This analysis has identified differences between CoYMV and the CaMV, a member of the other group of plant pararetroviruses. Compared to CaMV, CoYMV encodes fewer transcripts and ORF's. CaMV encodes two abundant transcripts, the 35S and 19S transcripts. We have detected only one CoYMV transcript. The expressed strand of CaMV contains eight ORF's capable of encoding proteins >10 kd. The products of six of these ORF's have been identified in infected cells. CoYMV transcript contains only three ORF's capable of encoding proteins

TABLE 3. Agroinfection with pCoinf4

Strain	Plants inoculated	Plants infected at day 20
A281[a]	6	0
A281(pCoINF4)[b]	12	12

[a] A281 containing the Ti plasmid pTiBo542
[b] A281 containing both pTiBo542 and pCoinf4

> 10 kd. However, it is possible that CoYMV and CaMV may encode a similar number of mature proteins because the 216 kd ORF product is a polyprotein and the number of mature proteins that are derived from it is unknown.

Caulimovirus-infected cells are characterized by the presence of cytosolic inclusion bodies that contain virus particles (5, 6). The matrix of the inclusion body is composed of a protein which is encoded by the viral genome. In addition to a structural role, recent studies (34,35) suggest that the inclusion body protein is a trans-acting factor that facilitates translation of the polycistronic RNA. Badnavirus-infected cells do not contain discernable inclusion bodies. Although virus particles are observed in stacked arrays they do not appear to be surrounded by a matrix (Lockhart, unpublished) suggesting that the CoYMV genome may not encode a structural inclusion body protein. However, since the CoYMV proteins are presumably encoded by the polycistronic CoYMV RNA, it may encode a protein capable of facilitating the translation of this RNA. Attempts to identify an ORF that potentially encodes such a protein by comparisons with the caulimovirus inclusion body proteins have been unsuccessful.

CoYMV ORFs corresponding to caulimovirus ORFs I–III have not been identified. The caulimovirus ORF II product is believed to be an aphid transmissibility factor. Since mealybugs are the natural vector for CoYMV (3), it is not particularly surprising that an ORF corresponding to caulimovirus ORF II is not present. It has been suggested that the CaMV ORF I product is involved in cell-to-cell movement of the virus (reviewed by 5). Since cell-to-cell movement occurs during CoYMV infection, a CoYMV protein may be involved in this process. It is possible that one of the ORF products of unknown function is involved in this process. However, it is also possible that a protein such as the coat protein could perform this function. The CaMV ORF III product is a dsDNA binding protein (36). The role of this protein in the virus life cycle is unknown, thus it is unclear if CoYMV should be expected to encode a protein which performs an analogous function.

The caulimoviruses, retroviruses and some retrotransposons encode the *gag* or coat protein and replicase polyprotein in separate ORFs. In contrast, CoYMV encodes these in a single ORF. Retroviruses express the ORF containing the polyprotein either by occasionally shifting frame during the translation of *gag*, or by suppressing a stop codon that follows *gag*. This produces a polyprotein that is processed by the virally-encoded protease. As a consequence of this expression mechanism, production of the structural protein, *gag*, exceeds that of the replicase enzymes. Unless some mechanism exists to attenuate the translation of the downstream portion of the 216 kd ORF, CoYMV coat protein and replication enzymes are produced in equal amounts. Interestingly, the known plant retrotransposons (37,38,39) encode their proteins in a single ORF.

The carboxy-terminal half of the 216 kd polyprotein consists of the protease, reverse transcriptase and ribonuclease H. The nature of the amino-half of this protein is less clear. The presence of the zinc finger-like motif suggests that a portion of this protein is the virus coat protein. The major CoYMV virion coat proteins are 38 and 40 kd in size (Lockhart, unpublished). In CaMV, the 44 and 37 kd coat proteins are produced by proteolytic processing of the 57 kd ORF IV protein (40 and references therein). If the CoYMV coat proteins are produced by a similar strategy, the amino-half of the 216 kd polyprotein must contain an additional protein of unknown function.

The significance of the presence of both polyadenylated and non-polyadenylated forms of the CoYMV transcript in infected cells is unclear. Since the transcript presumably serves both as an mRNA and as a template for genome replication, it is possible that the different forms serve different functions Polyadenylation has been shown to enhance production of the encoded gene product 16 to 40 fold (41). Thus the polyadenylated form of the CoYMV transcript may function primarily as an mRNA. In contrast, the non-polyadenylated form might be more available for replication since it would not be sequestered by transcriptional machinery.

ACKNOWLEDGEMENTS

We thank M. Gopalraj for excellent technical assistance, R. Beachy for providing their RTBV sequence and Monsanto for providing some of the oligonucleotides used in this study. This work was supported by grant IN-13-30-11 from the American Cancer Society to N.O. Published as paper 18,110 of the contribution series of the Minnesota Agricultural Experiment Station based on research conducted on the Projects 22-79H and 72-14G.

REFERENCES

1. Migliori,A. and Lastra,R. (1978) *Ann. Phytopathology*, **10**, 467–477.
2. Lockhart,B.E.L , Bouhida,M. and Olszewski,N.E. (1988) *Phytopath*, **78**, 1559.
3. Lockhart,B.E.L. and Khaless,N. (1988) *Phytopath*, **79**, 1548.
4. Lockhart,B.E.L. (1990) *Phytopath*, **80**, 127–131.
5. Guilfoyle,T.J. (1987) In Kousuge,T. and Nester,E.W. (eds.), Plant-Microbe Interactions: Molecular and Genetic Perspectives. Macmillan Publishing, New York, Vol. II, pp. 327–358.
6. Mason,W.S., Taylor,J.M. and Hull,R. (1987) *Adv. Virus Res.*, **32**, 35–96.
7. Hull,R. and Howell,S.H. (1978) *Virology*, **86**, 482–493.
8. Temin,H.M. (1989) *Nature*, **339**, 254–255.
9. Ausubel,F.M., Brent,R., Kingston,R.E., Moore,D.D., Seidman,J.G., Smith,J.A. and Struhl,K. Current Protocols in Molecular Biology (1987) Greene Publishing Associates and Wiley-Interscience. New York.
10. Gerard,G.F. and Miller,K. (1986) *Focus*, **8**, 5–6.
11. Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982) Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Laboratory, Cold Spring Harbor.
12. Ghosh,H.P., Ghosh,K., Simsek,M. and RajBhandary,U.L. (1982) *Nucleic Acids Res.*, **10**, 3241–3247.
13. Inoue,T. and Cech,T.R. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 648–652.
14. Aruffo,A. and Seed,B. (1987) *Proc. Natl. Acad. Sci. USA*, **84**, 8573–8577.
15. Olszewski,N.E., Gast,R.T. and Ausubel,F.M. (1989) *Gene*, **77**, 155–162.
16. Henikoff,S. (1984) *Gene*, **28**, 351–359.
17. Murphy,G. and Kavanagh,T. (1988) *Nucleic Acids Res.*, **16**, 5198.
18. Montoya,A.L., Chilton,M.-D., Gordon,M.P., Sciaky,D. and Nester,E.W. (1977) *J. Bacteriol.*, **129**, 101–107.
19. Mersereau,M., Pazour,G.J. and Das,A. (1990) *Gene* **90**, 149–151.
20. Howell,S.H., Walker,L.L. and Dudley,R.K. (1980) *Science*, **208**, 1265–1267.
21. Grimsley,N.H., Hohn,B., Hohn,T. and Walden,R.M. (1986) *Proc. Nat. Acad. Sci. USA*, **83**, 3282–3286.
22. Proudfoot,N.J. and Brownlee,G.G. (1974) *Nature*, **252**, 359–362.
23. Sieg,K. and Gronenborn,B. (1982) In Abstracts, NATO Advanced Study Instituite on Structure and Function of Plant Genomes. Porto Portese, Italy, pp. 154.
24. Dixon,L.K. and Hohn,T. (1984) *EMBO J.*, **3**, 2731–2736.
25. Gardner,R.C., Howarth,A.J., Hahn,P., Brown-Luedi,M., Shepherd,R.J. and Messing,J. (1981) *Nucleic Acids Res.*, **9**, 2871–2888.
26. Hull,R., Sadler,J. and Longstaff,M. (1986) *EMBO J.*, **5**, 3083–3090.
27. Richins,R.D., Scholthof,H.B. and Shepherd,R.J. (1987) *Nucleic Acids Res.*, **15**, 8451–8466.
28. Toh,H., Kikuno,R., Hayashida,H., Miyata,T., Kugimiya,W., Inouye,S., Yuki,S. and Saigo,K. (1985) *EMBO J.*, **4**, 1267–1272.

29. Johnson,M.S., McClure,M.A., Feng,D.-F., Gray,J. and Doolittle,R.F. (1986) *Proc. Nat. Acad. Sci. USA*, **83**, 7648–7652.
30. Covey,S.N. (1986) *Nucleic Acids Res.*, **14**, 623–633.
31. Fuetterer,J. and Hohn,T. (1987) *Trends Biochem. Sci.*, **12**, 92–95.
32. Luo,G., Sharmeen,L. and Taylor,J. (1990) *J. Virol.*, **64**, 592–597.
33. Grimsley,N., Hohn,T., Davies,J.W. and Hohn,B. (1987) *Nature*, **325**, 177–179.
34. Bonneville,J.M., Sanfacon,H., Futterer,J. and Hohn,T. (1989) *Cell*, **59**, 1135–1143.
35. Gowda,S., Wu,F.C., Scholthof,H.B. and Shepherd,R.J. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 9203–9207.
36. Giband,M., Mesnard,J.M. and Lebeurier,G. (1986) *EMBO J.*, **5**, 2433–2438.
37. Voytas,D.F. and Ausubel,F.M. (1988) *Nature*, **336**, 242–244.
38. Grandbastien,M.-A., Spielmann,A. and Caboche,M. (1989) *Nature*, **337**, 376–380.
39. Smyth,D.R., Kalitsis,P., Joseph,J.L. and Sentry,J.W. (1989) *Proc. Nat. Acad. Sci. USA*, **86**, 5015–5019.
40. Torruella,M., Gordon,K. and Hohn,T. (1989) *EMBO J.*, **8**, 2819–2825.
41. Gallie,D.R., Lucas,W.J. and Walbot,V. (1989) *The Plant Cell*, **1**, 301–311.