Physical structure and genetic organisation of the genome of maize streak virus (Kenyan isolate)

Stephen H. Howell*

Biology Department CO16, University of California San Diego, La Jolla, CA 92093, USA, and
Division of Plant Industry, CSIRO, Canberra City 2601, ACT, Australia

Received 16 July 1984; Accepted 31 August 1984

ABSTRACT

The structure of the maize streak virus genome (Kenyan isolate, MSV-K), as determined from the sequence of clones obtained from DNA isolated from virus particles, is composed of one major DNA component of about 2.6 kb. MSV virion DNA is partially double-stranded, composed of a full-length virion (V) strand and a short (70-80b) primer (P) strand. The primer strand has a fixed 5'-end capped with alkaline labile material, presumably 1-2 ribonucleotides. The MSV genome has two major coding regions oriented on opposite strands and flanked by two small intergenic regions. The coding region on the P strand is composed of two major open reading frames (ORFs), arranged in tandem and in the same reading frame. Because the predicted protein derived from a composite of these two ORFs closely corresponds to the product from a single ORF in the cassava latent virus genome, it is likely that this region encodes two proteins with common amino-termini, one a read-through product of the amber codon terminator in the first ORF. The intergenic regions contain potential transcription start and stop signals oriented in the direction of the two opposing coding regions. Considerable DNA sequence heterogeneity was observed, mostly silent or conservative third base substitutions in coding regions and base substitutions, small insertions and small, close-range transpositions in intergenic regions.

INTRODUCTION

Maize streak virus (MSV) is the type virus of the geminivirus group (1) and the only known DNA virus to infect corn (*Zea mays*). MSV causes one of the most important diseases of corn in Africa (2) and can only be propagated in the U.S. under strict pathogen containment conditions (3). Nonetheless, the study of the virus has great appeal to genetic engineers who are interested in adapting a disabled version of the viral genome as a vector for introducing foreign DNA into corn (4-6).

Geminiviruses are a recently recognized group of viruses noted for their unusual twinned particle morphology and their genomes composed of single-stranded DNA (7,8). Harrison *et al.* (9) showed that geminiviruses, including MSV, contain single-stranded DNA circles of about 2.5 kb. Haber *et al.* (10) found that the sequence complexity of the bean

golden mosaic virus (BGMV) genome, as revealed through restriction digest analysis, was about twice the size of a single strand DNA circle. They concluded that the BGMV genome must be divided, composed of two DNA circles of about 2.5 kb. The two component nature of geminivirus genomes has been borne out by recent restriction maps produced for the BGMV (11) and the tomato golden mosaic virus (TGMV) genome (12) and by the DNA sequence analysis of the cassava latent virus (CLV) genome (13). Unequivocal evidence for two components was obtained from infectivity studies of cloned CLV DNA by Stanley (14) and of TGMV by Hamilton *et al.* (15). The two groups found that both components (components 1 and 2 in the case of CLV and components A and B for TGMV), separately cloned, were required for infection.

The genetic organization of the CLV genome has been deduced by Stanley and Gay (13) from their sequencing data. They found 12 open reading frames (ORFs), capable of encoding proteins of molecular weight 10K or greater, oriented in both directions around the two components of the CLV genome. The bidirectional orientation of ORFs suggests that both strands of the virus genome, i.e. the virion strand and its complement, are expressed in infected plants. Unencapsidated double-stranded DNA forms have been detected intracellularly in BGMV-(16) and TGMV-(17,18) infected plants and presumably a similar form is found in CLV-infected plants. A property common to the two component geminiviruses, BGMV (11), TGMV (12) and CLV (13), is that the components share at least one short region of sequence homology. In CLV, the region of homology is about 200b long and is a non-coding region (13).

In this paper, I describe the DNA sequence and derived organization of the MSV-K genome. Unlike BGMV, TGMV and CLV, MSV appears to be composed of one DNA component of about 2.6 kb. This study was made possible by the collaboration with Dr. K.R. Bock of the Kenyan Agricultural Research Institute, Nairobi, Kenya who provided virus for this study. Isolated virus was imported to the U.S. with the permission of the U.S. and California Departments of Agriculture.

MATERIALS AND METHODS

Isolation of MSV DNA

MSV-K (Kenyan isolate) was purified as described by Bock *et al.* (19). 100-150 μ l of virus suspension was loaded onto a 12.5 ml sucrose gradient (10-40% in 10 mM NaPO₄, pH 7.2) and centrifuged at 35 krpm for

3 hrs at 15° in an SW41 rotor. Light scattering bands were removed by piercing the side of the tube with a needle. Fractions were diluted with an equal volume of H₂O, and virus was collected by centrifugation at 40 krpm for 3-5 hrs at 15° in a Ti50 rotor. The virus pellet was resuspended in 100 µl of H₂O with occasional vortex mixing over at least 1 hr.

To the virus suspension was added an equal vol of 100 mM NaCl, 50 mM EDTA, 20 mM Tris-HCl (pH 8.0). 5 µl of 4 mg/ml proteinase K (final conc. 100 µg/ml) was added and the suspension incubated at room temp for 10 min. 10 µl of 20% SDS was added (final conc. 1%) and the mixture incubated for 1 hr at 37°. DNA was extracted with phenol and precipitated from the aqueous phase with 0.1 vol of 1M NaOAc (pH 5.0) and 2 vol EtOH.

Construction of cloned MSV DNA libraries

Second strand virion DNA was synthesized in vitro using 1 µg of virion DNA as template. The reaction in 20 µl contained 10 mM Tris-HCl (pH 7.5), 10 mM MgCl₂, 1 mM dithiothreitol, 0.33 mM of each dATP, dCTP, dGTP and dTTP and 50 units/ml of E. coli DNA polymerase 1, large fragment (Boehringer Mannheim) or 100 units/ml of holoenzyme (Boehringer Mannheim). In labeling reactions, 40 µCi of [α -³²P]-dATP and/or -dCTP (600 Ci/mmol, New England Nuclear) were substituted for the corresponding unlabeled nucleotides. Reactions were incubated at 30°C for 30 min and labeling reactions were chased for 15 min with 0.33 mM of the corresponding unlabeled nucleotide. Reactions were terminated by phenol extraction and alcohol precipitation of DNA.

Double-stranded virion DNA was cut with various restriction enzymes, such as Alu I, Hpa II, Mbo II, Sau 3A, Taq I and Acc I, ligated into the appropriately cut RF forms of the DNA cloning vectors M13 mp8 and M13 mp9 (20) and used to transfect E. coli JM101.

DNA was sequenced by the dideoxynucleotide chain termination method (21) as modified according to Amersham.

RESULTS

Virus particles

When examined under the electron microscope, MSV preparations were composed of nearly an equal mixture of singlet and twinned or paired particles. As described by Bock et al. (19), singlet and paired particles have sedimentation coefficients of 54 and 76S, respectively, and can be

Fig. 1. Sedimentation of virus particles through sucrose density gradient. Virus particles from bands visualized by light scatter were negatively stained with uranyl acetate and examined under the electron microscope. Bar = 20 nm.

separated by sedimentation through sucrose density gradients as shown in Fig. 1. Several observations from other geminiviruses suggest that singlet particles are simply unpaired or are broken particles. When it has been tested as in the case of curly top virus (CTV), paired particles and not singlets are infective, indicating that paired particles contain all the genetic information necessary for virus multiplication (22). Based on electron microscope observations of chloris striate virus (CSMV), Hatta and Francki (23) proposed a structure relating singlet and paired geminivirus particles to each other. They proposed that paired particles were composed of two incomplete icosahedral singlet particles with a surface lattice of $T=1$ and with one capsomere removed. Singlet particles of CSMV appeared hexagonal on most faces and twinned particles appeared as paired pentagons joined at the face where the capsomere was missing. MSV preparations appeared much the same. Singlet particles were hexagonal, about 20 nm in diameter, while twinned particles appeared as paired pentagons (Fig. 1).

Fig. 2. Agarose gel (2.5%) of ethidium bromide stained DNA extracted from virus particles. Lane A, phage M13 mp8 single-stranded DNA marker; Lane B, DNA from MSV singlet particles; Lane C, DNA from paired MSV particles. Blot of MSV DNA from paired particles hybridized to ^{32}P -labeled cDNA (2×10^6 CPM) from the insert of ALU 19, Lane D, and ALU 21, Lane E (as shown in Fig. 3).

As with CTV, both singlet and paired MSV particles contained nucleic acid. However, in early efforts, only DNA was obtained from singlet particles using a detergent-phenol extraction procedure. To extract DNA from paired particles, it was necessary to digest particles with a protease prior to detergent-phenol extraction. This observation lends further support to the notion that singlet particles are broken and release their DNA more readily than do twinned particles.

The DNA isolated from paired particles migrated as one major band on non-denaturing agarose gels with a mobility relative to phage M13mp8 DNA and its derivatives roughly equivalent (estimated by extrapolation) to a circular single-stranded molecule of about 2.5 kb (Fig. 2, lane C). On the other hand, DNA from singlet particles appeared degraded, heterodispersed, generally migrating more rapidly than DNA from paired particles (Fig. 2, lane B).

Cloning and determining the virion strand of the MSV genome

The sequence of the MSV genome was determined from cloned DNA. To clone virion DNA, a second DNA strand was synthesized *in vitro*. It was surprising to find that a second strand from virion DNA could be

Fig. 3. Assemblage of sequenced clones derived from libraries of MSV DNA and ordered from the unique Bam H1 site into a single 2681 sequence by computer analysis. Sequences indicated by dotted line represent inserts "turned around" and cloned in the opposite orientation in M13 mp9 (or mp11).

synthesized *in vitro*, using either holoenzyme or the large fragment of *E. coli* DNA polymerase I, without adding exogenous primer. It was estimated that over 80% of the input single-stranded DNA from a gel-purified virion DNA preparation was converted to double-stranded form in the *in vitro* reaction. The self-priming activity was a property of DNA from both singlet and paired particle preparations. A further characterization of the primer is presented below.

The MSV genome was sequenced by a "shotgun" technique, that is, the virion DNA, double stranded *in vitro* was cut into small fragments with various restriction enzymes and cloned directly into DNA sequencing vectors. Clones which gave unique patterns on one lane sequencing runs using a single dideoxynucleotide terminator were subjected to complete sequence analysis. All clones obtained from the various libraries fit into the fragment overlap map shown in Fig. 3. Nearly the entire genome was covered by clones in both orientations, and some individual regions were represented in as many as six different clones. The overlap map was composed of one basic component, was circular and about 2.6 kb in length.

Because clones derived from virion DNA double-stranded *in vitro* were sequenced, the strand representing virion DNA had to be determined experimentally. To do so, two clones, ALU 19 and ALU 21 which represented opposite strands from the same region of the genome were used as hybridization probes (Fig. 3). Radioactive DNA copies from the virus DNA inserts in these clones were hybridized to virion DNA. Copies from ALU 19, alone (Fig. 2, lane D), hybridized to virion DNA establishing that virion DNA is in the same sense as the insert in ALU 19.

The sequence of the virion (V) strand of the MSV genome is shown in

Fig. 4. The virion strand as represented here is composed of 2681 nucleotides; however, because of sequence heterogeneity, as discussed below, this is not an absolute size for the genome. (Nonetheless, it will be used as the basis for the numbering system discussed in this paper).

Genetic organization of the MSV genome

The sequence of the genome (Fig. 4) reveals two major coding regions containing a total of three large open reading frames (ORFs). Two ORFs were located in one coding region, one in the sense of the V strand and, hence, derived from RNA transcripts from the opposite strand, referred to as the primer (P) strand. ORF P1a and P1b (P strand, reading frame 1) have the capacity to encode proteins of 272 and 110 amino acid residues, respectively and are oriented in tandem in the same reading frame. The other major ORF reads in the sense of P strand and, therefore, would be derived from RNA transcripts from the V strand. ORF V1 has the capacity to encode a protein of 244 AA.

The tandem, in-frame relationship of ORF P1a and P1b was a curious arrangement of ORFs, and so the proteins encoded by this region were compared to those predicted from the published sequence of the CLV genome (13) as shown in Fig. 5. The two proteins encoded by ORF P1a and P1b show nearly 40% amino acid sequence identity with the single product of the large ORF adjacent to the non-coding region derived from a transcript of the virion strand of component 1 of the CLV genome. This suggests that ORF P1a and P1b may be translated into a single composite protein by reading through the amber codon terminator in ORF P1a. In fact, it seems unlikely that ORF P1b may be translated as separate protein. The region around predicted translation start site for ORF P1b (AAATAATGT) does not closely correspond to the consensus sequence for the start of translation (CC_GACCATGG) according to Kozak (24); whereas, the predicted translation start site for ORF P1a (CGGCTATGG) is close to the consensus sequence (Fig. 4). The amber codon ending ORF P1a, however, appears to be a strong terminator codon with respect to context rules. The codon is followed by G, one of the more frequently used "next" nucleotide in eucaryotic and eucaryotic virus messages (25). Nonetheless, it is likely that ORF P1a and P1b encode two proteins which share a common amino-terminus, one a product of ORF P1a and the other a 391 AA read-through product of ORF P1a and P1b. The coding region on the opposite strand, ORF V1 (V strand, reading frame 1) has the capacity to encode a protein of 244 AA residues. The sequence

10	20	30	40	50	60	70	80	90	100
GATCCATGAA	TGAATCGCAC	TTGTAGGCA	ATTATAGTG	CCCGGGACCG	GTTTCTTTAG	AATGGGCTGC	TTTAGATAAA	TCATACAAG	CAGGACCAAG
(V3) 110	120	130	140	150	160	170	180	190	200
CTTTACATCA	AATCGCTCGG	GCCTACCGC	AGGAAAAGAA	GGCGCGGTAA	TATTAGTGCG	CGCCTTCTTT	TCTTGTAGC	CGGCCCGGT	CGGCCCATTT
210	220	230	240	250	260	270	280	290	300
ACAACCACTA	TAAGACAGGG	GAGCGGGCGA	TTTCGAAGGC	GCACAGGTAT	ACTATAGAGT	TGGAGCTCC	CATTCTGCCA	TCCTCAGGCC	TCAGCCTCCG
310	320	330	340	350	360	370	380	390	400
TCGTGGGCTA	TGGCTCTCTC	CTCATCCAAC	CGTCAGTTCT	CACACAGGAA	CGCTAACACG	TTCTTAACCT	ACCCAAGTG	TCCAGAAAAT	CCTGAAATCG
410	420	430	440	450	460	470	480	490	500
CCTGTCCAGT	GATCTGGGAG	CTCGTTGTTT	GTTGGATTCC	CAATATATTT	CTATGTGCLC	GAGAGGCACA	TAAAGATGGA	AGTTTGCATT	TACATGCATT
510	520	530	540	550	560	570	580	590	600
GCTTCAGACA	GAGAAGCCGG	TAAGGATATC	TGACTCAAGG	TTCTTTGATA	TAAATGGGTT	TCACCCAAT	ATTCAGAGTG	CCAAGTCAGT	TAACAGGGTG
610	620	630	640	650	660	670	680	690	700
AGGGATTACA	TTCTCAAGGA	ACCTCTGGCT	GTGTTTGAGA	GAGGTACTTT	CATTCTTAGG	AAGTCCCCCT	TCCTAGGAAA	ATCTGATTCA	GAAGTAAAG
710	720	730	740	750	760	770	780	790	800
AAAAAAGGCC	TTCCAAGATG	GAAATATATC	GAGACATTAT	TTCAACACTCT	ACTTCCAGAG	AAGAGTACTC	CTCCATGATC	CAGAAAGAGC	TTCCCTTTGA
810	820	830	840	850	860	870	880	890	900
TTGGTCCACA	AAATGTCAGT	ATTTTGAATA	CTCTGCAAA	AAGCTTTTCC	CTGAGATTCA	GGAAGAGTTC	ACCAATCTCT	ATCCACCCTC	CTACCTGAT
910	920	930	940	950	960	970	980	990	1000
TTACTTTGTA	ATGAGTCAAT	CAATGATTGG	CTCCAGCTGC	ACATCTTCCA	GGTTAGTCCC	GAAGCTTACA	TGCTCCTTCA	ACCTGCTCTT	TATACCTCTG
1010	1020	1030	1040	1050	1060	1070	1080	1090	1100
ATGATGCAAT	TTCTGACCTC	CAATGGATGG	ATTCTGTATC	CAGTCATCAG	ATGAAGATGC	AAGAAGACAG	AGCCTCTACA	TCGTGGGCC	AACAAGAACCA
1110	1120	1130	1140	1150	1160	1170	1180	1190	1200
GGAAATCTA	CTTGGGCCAG	AAGCTTAGGG	GTTTCATAAT	ACTGGCAAAA	TAATGTTGAT	TGGTCTTCAT	ACAACGAGAG	ACGCAATCTA	CAACATCGTA
1210	1220	1230	1240	1250	1260	1270	1280	1290	1300
GATGATATTC	CGTTTAAATT	CTGTCTTTGT	TGGAACACAG	TAGTTGGCTG	TCAGAGGGAT	TTCAATTGTA	ACCCCAATA	TGATAAGAGA	AAGAAGGTGC
1310	1320	1330	1340	1350	1360	1370	1380	1390	1400
ACAAGAGATC	TAAGCTTACA	ATAATCTCTG	CCAACCTCGA	TGAAGATTGG	ATGAAGGAAA	TGACTCCAGG	GCAGCTGGAG	TATTTCCAGG	CAAACTGCAT
1410	1420	1430	1440	1450	1460	1470	1480	1490	1500
CATTTCATTT	ATGTGCGCCG	GGGAGAGAGT	GTATTTCTCC	CTGAGCTGAC	CTCCTACGGA	GGCAGTACAT	TCAGATAGAT	CTTGATTTTT	CGATGTTTTG
1510	1520	1530	1540	1550	1560	1570	1580	1590	1600
CCCGCCGAGC	ACCACCCTTA	GCCGCCCGCC	TACTACCCCC	CTGTATTGAT	TGCGTGTGTT	TTTTCTGTCC	ATCGACGAGC	ATATTAATGT	AAGCTTTTCA
1610	1620	1630	1640	1650	1660	1670	1680	1690	1700
CATTATCAG	ATATAATAAA	ACGGGAGTTT	TTATTCATTA	CTGGTTGCCA	ACACTCTTAA	AGTACAGAGC	GGTCTGCCCA	TGGGCAGTAA	ATGTAAGACC
1710	1720	1730	1740	1750	1760	1770	1780	1790	1800
ATTGCCGGGG	GCAATGACCA	TGTACAGGAC	TCCTCTCTGG	ATGGCACCAA	CTCCTCCGTC	CGTTACATTC	TTCCACTGCG	TTCTCACTCC	CAACCCACTC
1810	1820	1830	1840	1850	1860	1870	1880	1890	1900
GTGAACCTGT	GGAAGTAGAT	GTTCGCTTGG	CAAGGCTTCC	AACCTGCAAT	CGATGGAGGA	ATATCCGAAC	CAATCCGACC	GTCGCTCTCC	ATGTTGAACA
1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
ACCATCGGCC	TTTCACCAAG	AAGCGATTAC	ACAGCTCCCG	GCTCACTTTC	CATGTGGCCG	GCCACGCTTT	TAGCGTGCTA	GGGTAGGCAA	ATATAGTTTG
2010	2020	2030	2040	2050	2060	2070	2080	2090	2100
CGGGGTCCGA	GCTTGTCCCG	CGGGAGTGGT	GTCATACACC	AGCCACATTA	CACCGGTTCC	GGTGTGGAG	TACGGGCAGG	CTGCCCGCTC	GGCAACGAAG
2110	2120	2130	2140	2150	2160	2170	2180	2190	2200
TGGTAGTCGA	CGGCAATCTT	GTACGTGAGA	GTCTCGTGG	TGTGGCGGTT	GCCCTCGTCA	GATCCTCGGG	CATAGGTGTT	GATGAGGTCA	CATACTCTCT
2210	2220	2230	2240	2250	2260	2270	2280	2290	2300
CGGAGGGGAC	CGTTATCATG	GTGGTCCGAG	CATGCTGGAG	TGCTGGGATT	TGGAGGGATG	GCCTATCGGC	CTTGCTTCCA	GCCCTCTTCA	GCCAGCTGA
2310	2320	2330	2340	2350	2360	2370	2380	2390	2400
AGAAGGCTTC	TTCTTAGTCA	CCCGCTTACT	CCAATTGCGA	TCATCTCCCC	GCTTCTCTTT	GGACGTGGAC	ATGGCTGATT	GCTTATCCCG	TGCTTGGAAC
2410	2420	2430	2440	2450	2460	2470	2480	2490	2500
AAATGGCCCG	GGATTGCCCT	GACTTGGTGG	TGCTGGTATA	TTAGGGATAG	GGTTGCTCTT	ATCCACAGCT	TGTCCTCCAA	ATATCAGCTC	CTCCGTGGAT
2510	2520	2530	2540	2550	2560	2570	2580	2590	2600
CTGCTTGTGC	GAGCCTTCAG	AACTAAGATA	AGGTCTCTCA	GCACCCAAAG	GTAAAGCAGG	TAAAGCAAAA	TCAATGCAC	AAAGCTCAAA	ATAGCTACCT
2610	2620	2630	2640	2650	2660	2670	2680	2690	2700
CGCTACGCGC	ACTCCACGGC	ACTCTCCGGC	ATGTCGGAGC	TGCTGTGGGT	ACCCGCGGCT	GATAATACAG	GCGTCTCTGT	G	

Fig. 4. DNA sequence of MSV-K virion (V) strand DNA ordered from the unique Bam H1 site. ORFs and direction of reading indicated by single overlined arrow. First ATG in each ORF indicated by underlined arrow. Heterogeneity in DNA sequence is shown as base substitutions (e.g. $\rightarrow$), single base deletions ($\downarrow$) or transpositions ($\blacktriangleleft$ ----- $\blacktriangleright$). Amino acid changes resulting from DNA sequence heterogeneity indicated above the line, such as L $\rightarrow$ L. ORF V3 is an extended form of ORF V1 in a sequence variant. Position and direction of primer (P) DNA strand indicated by double underlined arrow.

surrounding the start codon (CAGCCATGT) has moderate correspondence to the consensus translation start sequence (24). A larger form of the ORF on the V-strand, ORF V3 (V strand, reading frame 3), results from a sequence variant discussed in a later section.

Thus, the MSV genome has two major coding regions composed of a single ORF derived from transcripts from the V strand and a composite ORF encoded by the P strand. The general organization of ORFs in the MSV genome is illustrated in Fig. 6. The two coding regions diverge away from each other on either strand from a central non-coding or intergenic region, called the starting intergenic region (SIR) and converge toward each other at a terminal intergenic region (TIR). The arrangement of ORFs suggests bidirectional transcription of the genome and inspection of the sequence in the SIR reveals potential transcription start signals oriented away from each other and in the TIR potential polyadenylation signals (AATAAA) oriented towards each other. TATAA boxes are located 120 bases upstream from the first ATG of ORF P1a and 344 bases upstream from the translation start codon of ORF V1. These potential transcription start signals are shown with respect to the secondary structure of the SIR in Fig. 7. Central to this region and between the potential transcription start signals lie two large stem loop structures, one containing the sequence GGAAAAG closely corresponding to the enhancer "core" consensus sequence, (G)TGG^{AAA}_{TTT}G, for animal viruses (26).

Sequence heterogeneity

Because certain regions of the genome were represented in a number of different clones, it was possible to assess the level of sequence heterogeneity in the virus DNA population. Much sequence variation was found in this limited sample suggesting that considerable heterogeneity lies in the total virus DNA population.

In coding regions the observed heterogeneity involved primarily single base substitutions in the third position of codons which did not alter the predicted amino acid sequence (Fig. 4). Two DNA sequence variants did lead to significant amino acid sequence changes. A two base substitution at nucleotide position 892-893 changed a serine to a glutamine, and single base transposition of a C from position 2154 to 2159 produced a short frame-shift leading to a change from a leucine-valine pair to a serine-serine pair. The single base deletion at position 2410 had a profound effect on the length of the ORF V1. The deletion extended ORF V1 (denoted ORF V3 in Fig. 4) by 105 amino acid residues, such that the

```

MSV  . . . . .
CLV  . . . . .
MAS--SSSN----RGFSHRN-A-NTFLTYPKCPENP-EIACQMIWELVVRWIPKYIL-CA
M-SFIDLVMRTPR-F--RIGAKNVFLTYPKCSI-PKEHLLSFIGTSLGNSNPKFIKIC-
. . . . .
REAHKDGSLHLHALLGTE-KPVRISDSRFFD-I--N---GFHPNIGSAKSVNRVRYIL-
RELHGNGEPLHLALIGFEGK-MTITNNRLFDCVHPSCSTSFHPNIGGAKSSSDVKSY-LDK
. . . . .
KEPLAVFERGTFTIP-RKSPFLGK-SDSEVKEKKPSKDEIMRDIISHSTSKEEYLSMIQK
KDGDTV-EWGGFGIDGR-SARGGGGSANDAYAKA-N-L-----SGSKSEALNVIR-
. . . . .
EL-FFDWSTKLQYFEY--SANKL---FPEIQEEFTNPHPPSSPDLLCNESINDWLQFNIF
ELVFKDF--VLQ-FNMLNS-N-LDRIFQEPPAPYVSPFPCSSFDQVPVE-IEEWVADNVR
. . . . .
QVSPDLLC--NE-SINDWLQPNIFQVSPEAYMLLQFACYTLDDAISDLQWMDSVSSHQMK
D-SAARPWRPN--SI---V---I-----E-----GDS-RT-G-K
. . . . .
DQESRASTSSAQGEENLGPEA*GFIITGKIMLIGLHTTKHAIYNIV-DDI-PF--K-FC
--TIWA--RS-----LGPHN--Y-LCGHLDLSPKVFNNAAWYN-VIDDVDPHYLKHF-
. . . . .
PCWKQLVGCQDFIVNPKYGGKKKVQ-KKSKPTIILANS-DEDWMKEMTPG--QLEYFQ-
---KEFMGSGRDWQSDPHYGKP--VQIKGGIATIFLCNPGTSSYKEFLAEKQ-EALKA
. . . . .
-A--NCIIYI-MS-PGEKWYSPPELPFT--EAVHSDRS
WALKNA-IFITLTEFLYSG-SNGSHSQTSQEASHPA

```

Fig. 5. Comparison of predicted amino acid sequence of MSV ORF P1a and P1b (upper line) and predicted CLV protein (13) from the large ORF starting from the region of homology on the complement of CLV component 1 (lower line). Dots above the line indicate identity of amino acid sequences. Sequences have been spaced to maximize identity. Arrow and asterik indicate amber codon termination of ORF P1a.

Fig. 6. Organization of the MSV genome deduced from the sequence of MSV-K virion DNA. ORFs and direction of reading indicated by black arrows. Composite ORF P1a + P1b proposed by reading through amber terminator codon. ORF V3 is an extended form (extension indicated by cross-hatching) of ORF V1 (and in a different reading frame) resulting from DNA sequence variant. Stippled areas define two intergenic regions, the starting intergenic region (SIR) and the terminal intergenic region (TIR).

Greater variation was found in the intergenic regions than in the coding regions including single base substitutions, a two adjacent nucleotide substitution (at nucleotide position 1622-1622), single base deletions and a two base transposition from nucleotide position 126-127 to position 118. This transposition involves bases in the stem loop structure shown in Fig. 7.

The self-priming of second strand MSV DNA *in vitro* suggested the

existence of a primer fragment base-paired to the virion strand. The notion of a primer, was strengthened by the observation that the self-priming ability of the DNA was lost following heat denaturation and quick-cooling of the DNA (data not shown). In an effort to determine the size and location of the primer on the genome, primer extension reactions were carried out in the presence of dideoxynucleotides. In general, the sequence pattern resulting from several primer extension runs was unsatisfactory, comparable to the pattern shown in Fig. 8 (-) lane, in this case, when dideoxyguanosine triphosphate was used as a chain terminator. On sequencing gels, the extension product was blurred over 2-3 nucleotide positions, suggesting some heterogeneity at the 5'-end of the primer. The pattern sharpened considerably by treating the product with alkali before gel analysis (Fig. 8, (+) lane). The pattern sharpening indicated that alkaline labile material, presumably 1-2 ribonucleotides, was bound to the 5'-end of the primer. The sequence determined downstream from the primer placed the primer at a unique site on the MSV genome as shown in Fig. 6. The positions of the G doublets (C doublets on the template strand) were key in locating the sequence downstream from the primer, because single G's did not always give strong stops. The pattern of G stops derived from Fig. 8 is shown against the sequence in Fig. 9. Extrapolation back to the 5'-end of the primer places the first deoxyribonucleotide of the primer at position 435.

The primer extension technique could not be used to determine whether the 3'-end of the primer was also unique and where it was located on the genome, nonetheless, G stops were picked up starting at about + 83 nucleotides from the 5'-end (Figs. 8 and 9). (The stop at about +71 was not regarded as a bonafide stop because it topped a ladder of non-specific stops). Since labeled dCTP and dATP were both used in the primer extension reactions, the 3'-end of the primer would probably lie somewhere before +82. Hence, the primer is probably 70-80 nucleotides long.

With respect to other landmarks on the MSV genome, the primer or P strand is located at the amino-terminal coding region of ORF P1a with the 3'-end of the primer pointing toward the SIR (Fig. 6). The primer was preceded upstream by an AT-rich region (-6 to -14b) and was flanked on either end by small stem loop structures (Fig. 9). The stem loop at the 3' end of the primer may have led to some non-specific chain termination in the primer extension reactions and possibly was the source of sequence ambiguity in the +88 to +93b region of the primer extension sequence.

Fig. 8. Primer extension from endogenous primer on MSV virion DNA in the presence of dideoxyguanosine triphosphate. Gel purified MSV virion DNA from paired virus particles as in Fig. 2 was used in primer extension reaction. Extended primer was labeled with [α - 32 P]-dATP and -dCTP in usual sequencing reaction (21) and analyzed on an 8% polyacrylamide-urea sequencing gel. Nucleotide positions were determined from sequence of MBO 6 clone in adjacent lanes. Lane (-), untreated product; Lane (+), product treated with 0.3 N NaOH for 30 min at 37°, then neutralized. Both samples were alcohol precipitated prior to electrophoresis.

strand is efficient and nearly complete since it is estimated that over 80% of input virion DNA is second-stranded. Addition of exogenous random primers (oligodeoxynucleotides derived from calf thymus DNA) does not stimulate further synthesis (unpublished observations). These data argue that most all of the virion DNA is involved as template in the synthesis of the second DNA strand in vitro, and that the cloned DNAs represent in toto the DNA which has been double-stranded. Therefore, the one component nature of the MSV genome sequence reflects the total complexity of most, if not all, of the MSV virion DNA. However, as stated above, the possibility cannot be excluded that another DNA component, of lesser abundance, is an indispensable second component of the MSV genome.

Preliminary data from Marriott and Symons (28) on the structure of the chloris striate mosaic virus genome (CSMV) also suggest one component. These authors have speculated that the two different geminivirus subgroups, distinguished by the insect vector which transmits them, may differ in the composition of their genomes. CLV, TGMV and BGMV are transmitted by whiteflies and have two component genomes while MSV and CSMV are leafhopper-transmitted and may have one component genomes.

The unexpected finding of a short primer strand requires some rethinking about whether single-strandedness is a unique and identifying feature of geminivirus genomes. Strictly speaking, the MSV genome is not single-stranded, but is partially double-stranded. In this respect, the MSV genome shares certain properties in common with the virion DNA of the hepatitis B-viruses (HBV). HBV virion DNA is partially double-stranded with an incomplete second strand of variable length covering from half to all the genome (see review, 29). Like MSV, the 5'-end of the second HBV DNA strand is fixed, but the 3'-end varies. Unlike MSV, the first HBV DNA strand is discontinuous with a break in region base-paired to the second strand. Virion DNA of the double-strand plant DNA viruses, such as CaMV, is much the same. Both DNA strands are discontinuous, but cover the entire CaMV genome. Nonetheless, the DNA strands are variable in length with fixed 5'-ends and 3'-ends which overlap the 5'-ends for an indeterminate distance, by usually 8-20 nucleotides (30). Like MSV, CaMV virion DNA strands have ribonucleotides on their 5'-ends which are thought to be remnants of RNA primers used in synthesis of the DNA strands (31). Whether the presumed ribonucleotides on the 5'-end of the MSV P strand are remnants

of an RNA primer will be interesting to investigate.

With the ease by which a second strand of virion DNA can be synthesized in vitro, it is surprising that virion DNA is not entirely double-stranded. Perhaps, the truncated form of the P strand is the signal for packaging and packaging blocks further P strand elongation. Alternatively, some property of the V-strand DNA template may prevent further elongation of the P strand. The small stem loop at the 3'-end of the primer seems an unlikely impediment to the growing end of the P strand. However, the stem loop might be a site to which a protein binds that blocks further elongation. Such a protein would be lost upon extraction of the DNA. Alternatively, at the time when the short P-strand segment is synthesized, the template (or V-strand) might not be continuous. Template discontinuities in the synthesis of CaMV DNA strands lead to the accumulation of replicating forms with truncated strands (32, 33).

The bidirectional organization of ORFs in the MSV genome argues for the presence within the cell of transcriptionally active double-stranded (or opposite-stranded) forms. Double-stranded forms of geminivirus genomes have been found in BGMV (16) and (17, 18) TGMV infected plants. Presumably, synthesis of a second DNA strand would be an early step in the infection process following the uncoating of the viral genome.

The virus coat protein is the only product known which may be encoded by the MSV genome. The coat protein is reported to have a molecular weight of 28K (34). It is not clear which coding region, if either, encodes the coat protein.

ACKNOWLEDGEMENTS

This work was supported by the NSF (PCM-814-1755) and the Science and Education Administration, USDA (5901-04010-8-0178-0). Special thanks to Dr. Jim Peacock in whose laboratory where much of this work was carried out, and to Linda Walker and Fiona Clay for their valuable technical assistance.

*Please address all correspondence to Dr. Stephen H. Howell CO16, University of California San Diego, La Jolla, CA 92093, USA

REFERENCES

1. Matthews, R.E.F. (1979) Intervirology 12, 129-296.
2. Bock, K.R. (1982) Plant Disease 66, 266-270.
3. Damsteegt, V.D. (1983) Plant Disease 67, 734-737.

4. Howell, S.H. (1982) *Ann. Rev. Plant Physiol.* **33**, 609-650.
5. Hull, R. and Davies, J.W. (1983). *Adv. Virus Res.* **28**, 1-33.
6. Buck, K.W. and Coutts, R.H.A. (1983) *Plant Mol. Biol.* **2**, 351-354.
7. Goodman, R.M. (1981) *J. Gen. Virol.* **54**, 9-21.
8. Goodman, R.M. (1981) in *Handbook of Plant Virus Infections and Comparative Diagnosis*, ed. Kurstak, E. (Elsevier, Amsterdam), pp. 879-910.
9. Harrison, B.D., Barker, H., Bock, K.R., Guthrie, E.J., Meredith, G. and Atkinson, M. (1977) *Nature* **270**, 760-762.
10. Haber, S., Ikegami, M., Bajet, N.B. and Goodman, R.M. (1981) *Nature* **289**, 324-326.
11. Haber, S., Howarth, A.J. and Goodman, R.M. (1983) *Virology* **129**, 469-473.
12. Bisaro, D.W., Hamilton, W.D.O., Coutts, R.H.A. and Buck, K.W. (1982) *Nucl. Acids Res.* **10**, 4913-4922.
13. Stanley, J. and Gay, M.R. (1983) *Nature* **301**, 260-262.
14. Stanley, J. (1983) *Nature* **305**, 643-645.
15. Hamilton, W.D.O., Bisaro, D.M., Coutts, R.H.A. and Buck, K.W. (1983) *Nucl. Acids Res.* **11**, 7387-7396.
16. Ikegami, M., Haber, S. and Goodman, R.M. (1981) *Proc. Nat. Acad. Sci. U.S.A.* **78**, 4102-4106.
17. Hamilton, W.D.O., Bisaro, D.M. and Buck, K.W. (1982) *Nucl. Acids Res.* **10**, 4901-4912.
18. Sunter, G., Coutts, R.H.A. and Buck, K.W. (1984) *Biochem. Biophys. Res. Comm.* **118**, 747-752.
19. Bock, K.R., Guthrie, E.J. and Woods, R.D. (1974) *Ann. Appl. Biol.* **77**, 289-296.
20. Messing, J. and Vieira, J. (1982) *Gene* **19**, 269-276.
21. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) *Proc. Nat. Acad. Sci. U.S.A.* **74**, 5463-5467.
22. Lansen, R.C. and Duffus, J.E. (1984) *Phytopathology* **74**, 114-118.
23. Hatta, T. and Francki, R.I.B. (1979) *Virology* **92**, 428-435.
24. Kozak, M. (1984) *Nature* **308**, 241-246.
25. Kohli, J. and Grosjean, H. (1981) *Mol. Gen. Genet* **182**: 430-439.
26. Weiher, H., König, M. and Gruss, P. (1983) *Science* **219**, 626-631.
27. Goelet, P., Lomonossoff, G.P., Butler, P.J.G., Akam, M.E., Gait, M.J. and Karn, J. (1982). *Proc. Nat. Acad. Sci. U.S.A.* **79**, 5818-5822.
28. Marriott, T.W. and Symons R.H. (1983) in *Plant Infectious Agents*, eds. Robertson, H.D., Howell, S.H., Zaitlin, M and Malmberg, R.L. (Cold Spring Harbor).
29. Tiollais, P., Charnay, P. and Vyas, G.N. (1981) *Science* **213**, 406-411.
30. Richards, K.E., Guilley, H. and Jonard, G. (1981) *FEBS Letters* **134**, 67-70.
31. Guilley, H., Richards, K.E. and Jonard, G. (1983) *EMBO J.*, **2**, 277-282.
32. Hull, R. and Covey, S.N. (1983) *Nucl. Acids Res.* **11**, 1881-1895.
33. Marco, Y. and Howell, S.H. (1984) *Nucl. Acids Res.* **12**, 1517-1528.
34. Bock, K.R., Guthrie, E.J., Meredith, G. and Barker, H. (1977) *Ann. Appl. Biol.* **85**, 305-308.