Volume 11 Number 4 1983 Nucleic Acids Research

The complete DNA sequence of minute virus of mice, an autonomous parvovirus

Caroline R.Astell*, Marnie Thomson, Michael Merchlinsky+ and David C.Ward+

Department of Biochemistry, Faculty of Medicine, University of British Columbia, Vancouver, British Columbia, V6T 1W5, Canada and +Department of Human Genetics and Molecular Biophysics and Biochemistry, Yale University, 333 Cedar Street, New Haven, CT 06510, USA

Received 1 December 1982; Revised and Accepted 25 January 1983

ABSTRACT

We have determined the complete nucleotide sequence of the genome of Minute Virus of Mice, an autonomous parvovirus. This single-stranded DNA is 5081 nucleotides long. The 3'-and 5'-ends of the viral strand contain imperfect palindromic sequences which consist, respectively, of 115 and 206 nucleotides. The 3'-terminal palindrome is composed of a unique sequence, whereas the 5'-terminal palindrome contains two sequences in equimolar amounts; these are related in that one is the inverted complement of the other. The DNA strand complementary to that which is encapsidated into virions contains two large open reading frames which together span almost the entire genome. Transcriptional and translational signals within the sequence have been identified and related to the known map coordinates of the viral transcripts. In this report we summarize some of the salient structural and organizational features of the MVM genome.

INTRODUCTION

The Parvoviridae family contains two groups of viruses that infect mammalian hosts; the defective (helper-dependent) Adeno-associated viruses and the autonomous, helper-independent, parvoviruses (1). Minute Virus of Mice (MVM), a member of the autonomous subgroup, was first isolated and characterized by Crawford in 1966 (2). The virus was plaque purified by Tattersall in 1972 (3) and the resulting strain, now designated MVM(p) for prototype (4), has been used by many laboratories working on the molecular biology of this virus family. All autonomous parvoviruses examined to date exhibit similar structural and biological properties, thus information gained with MVM can be applied with some confidence to predict interesting features of its family relatives. Since several of these (for example, canine parvovirus, bovine parvovirus, porcine parvovirus, mink enteritis virus and feline panleukopenia virus) are of considerable economic importance, such information could be of utility in designing strategies for vaccine development.

Parvovirus genomes are single-stranded DNA molecules of approximately

$1.5X10^6$ daltons. While an early classification identified defective viruses as packaging both plus (+) and minus (-) strands (in separate particles) and the autonomous parvoviruses as packaging only minus strands (5), it has been shown recently that the autonomous viruses LuIII and bovine parvovirus also package both plus and minus strands (6, R. Bates, personal communication). Both the 3' and 5' ends of the DNA molecules contain palindromes (hairpins), which play an important role in the replication of the viral genome (7-9). Since only one strand of parvovirus DNA is transcriptionally active, these viruses exhibit a fairly simple transcriptional pattern (10,11) and synthesize only a small number of, possibly only four, primary gene products (12). To further define the genetic organization and the coding potential of MVM DNA, we have determined the complete nucleotide sequence of this viral genome and analyzed the distribution of transcriptional and translational signals within the sequence. This study allows us to predict the genomic origin of both structural and non-structural polypeptides and to identify initiation codons and splicing junctions that may function in MVM gene expression. A preliminary report of this work has been presented elsewhere (13).

MATERIALS AND METHODS.

MVM(p) was grown on the A-9 variant of mouse L cells. Cell cultures were maintained and virus stocks prepared as described (14). Plasmid clones containing MVM sequences were constructed using duplex replicative forms of MVM DNA, synthesized both *in vivo* and *in vitro*. The preparation and characterization of these clones will be described elsewhere (M. Merchlinsky et al, manuscript in preparation).

The procedures used for end-labeling DNA, restriction endonuclease digestions, acrylamide gel electrophoresis, and DNA sequencing have been described previously (15-17). *E. coli* DNA polymerase, Klenow fragment was purchased from Boehringer Mannheim and M13 mp7 RF DNA and the universal 17 mer primer were obtained from Bethesda Research Labs and Collaborative Research, respectively. Other sources of material were as described elsewhere (15).

RESULTS AND DISCUSSION.

The Sequencing Approach.

The 3'-and 5'-terminal hairpin sequences of the genome were determined as described previously (12,15), using viral DNA extracted from purified virions. The extreme left hand end of the genome from nucleotide 130 to beyond the first PstI site at nucleotide 411 was sequenced by a series of priming

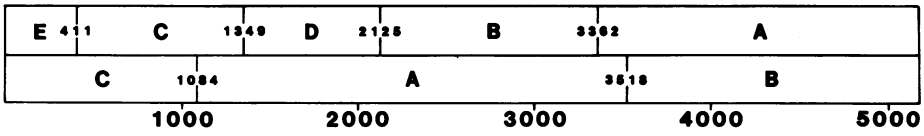
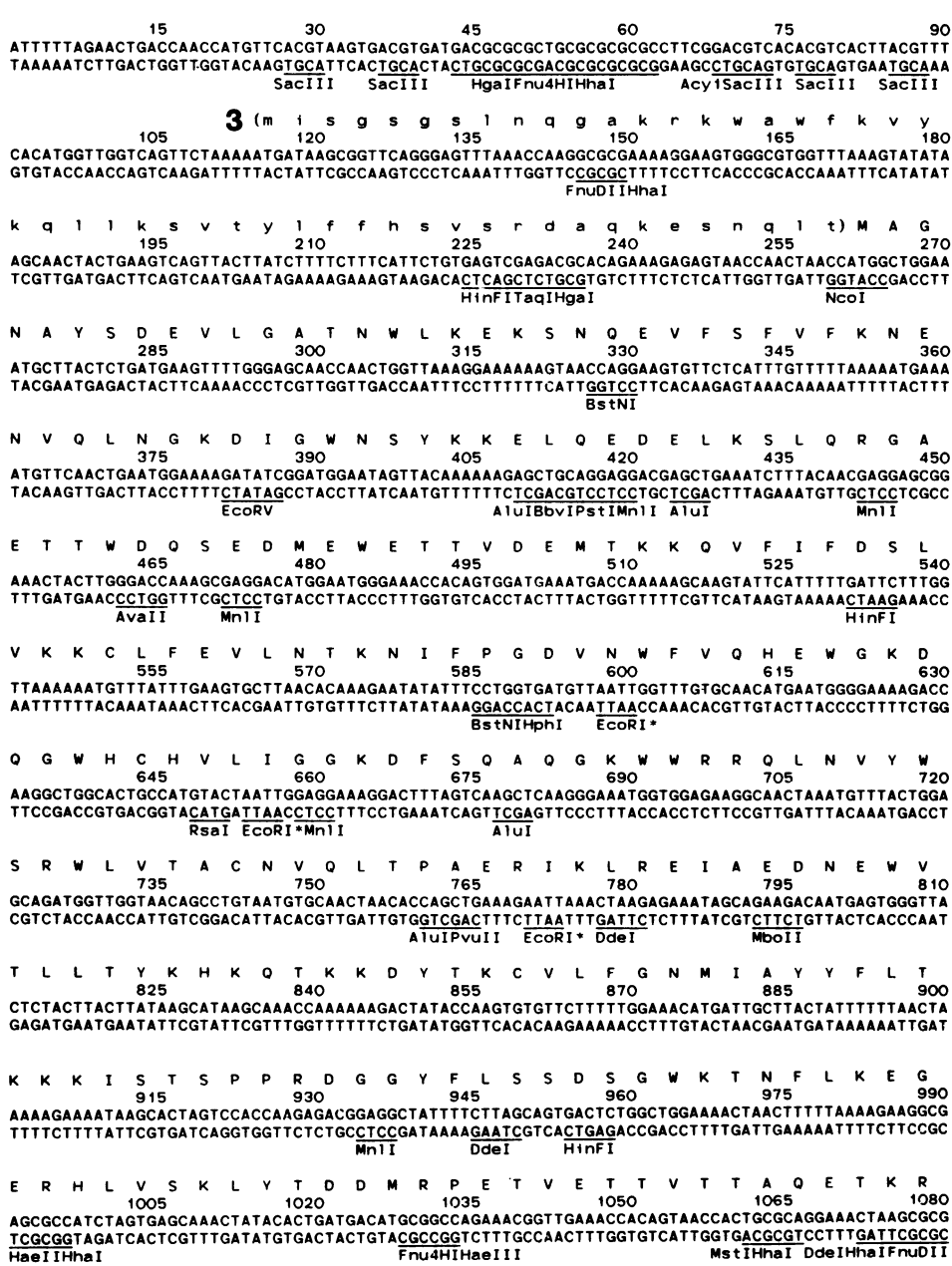


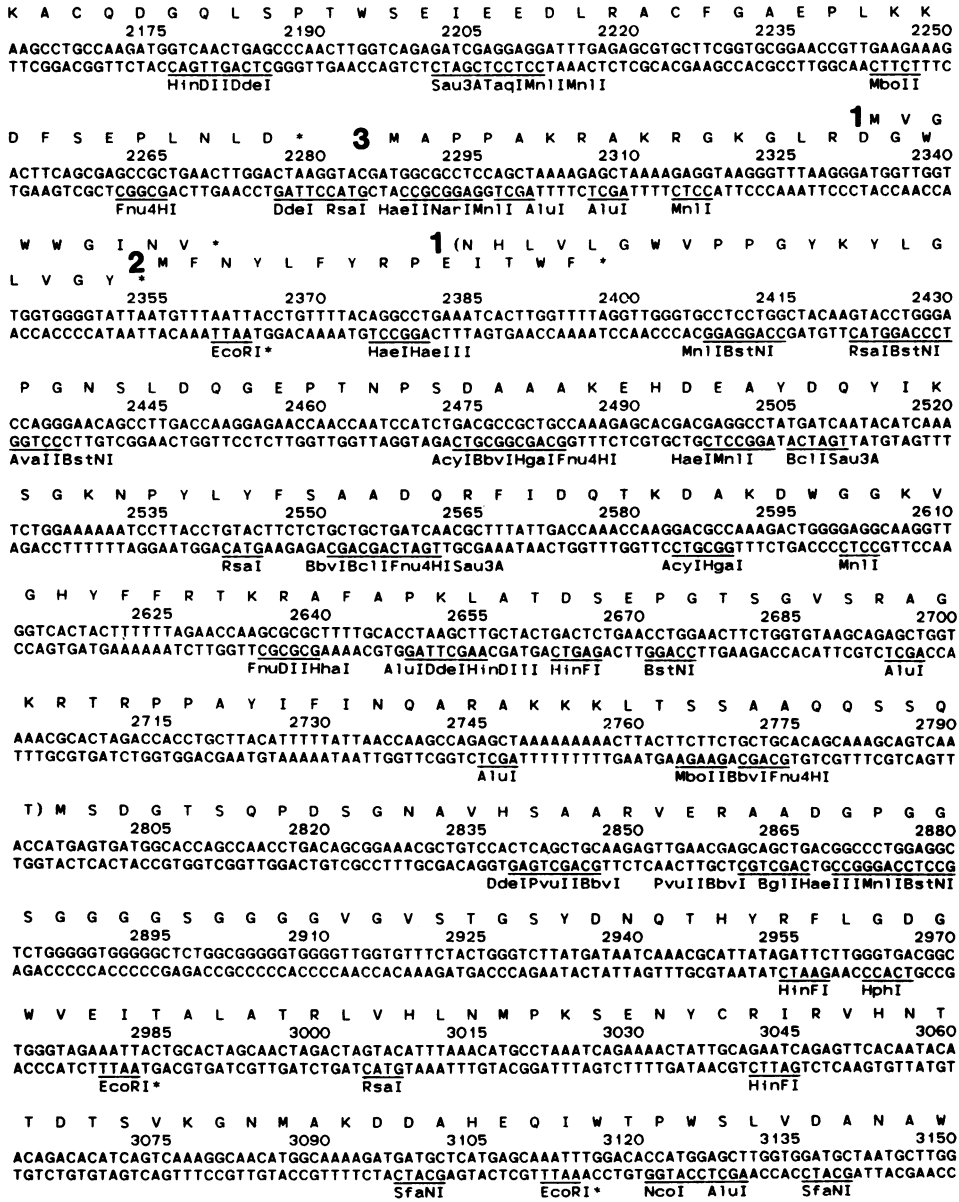
Figure 1. Cleavage sites within MVM(p) DNA for the restriction endonucleases PstI (upper map) and Eco RI (lower map). The 3' end of the viral strand is on the left side.

experiments using the dideoxyterminator method of Sanger (18). In addition to the data reported previously (15-16), the sequence in this region was extended to nucleotide 600 by priming with the synthetic oligonucleotide d-A_6GAG which is complementary to nucleotides 402 to 410. The major portion of the genome was sequenced by the method of Maxam and Gilbert (19), using exclusively 3'-end labeling techniques (17), and restriction fragments of MVM DNA cloned in pBR322. The EcoRI and PstI fragments that were used in this analysis are shown in Figure 1. Clonal isolates sequenced in their entirety included the PstI C clone extending from nucleotides 411 to 1349, an EcoRI A clone extending from nucleotides 1084 to 3518, and an EcoRI/BamHI clone which includes both the EcoRI A fragment (nucleotide 1084 to 3518) and an EcoRI B/BamHI fragment that extended from nucleotide 3518 to the end of the 5' hairpin (nucleotide 5081). More than 70% of the sequence was confirmed by sequencing both strands of the recombinant clones and overlaps between adjacent fragments were complete and unambiguous. In addition, we have cloned Sau3A fragments of the large EcoRI A fragment (Figure 1) and AluI and HaeIII fragments of RF DNA into appropriate sites of M13 mp7 and sequenced these using the dideoxynucleotide terminator method (18). This approach gave us more than 60% of the genome and confirmed the sequence obtained by the chemical method.

The complete nucleotide sequence of MVM(p) is given in Figure 2. The sequence is presented as a double stranded structure and according to convention (20) the viral or V-strand (3' → 5') is illustrated below the complementary or C-strand (5' → 3'). Two nucleotides remain uncertain. Nucleotide 3194 is a C in one clone and a G in another, while nucleotide 3725 is a G in one clone and an A in another. These differences may be due to nucleotide changes which have occurred during the cloning and amplification steps. In the sequence in Figure 2, we have listed nucleotide 3194 as a C and 3725 as a G because these nucleotides are the ones from clones that have been amplified less. Both of these sites lie within the presumed coding region for capsid proteins (see below) and in the first case a threonine or serine would be the


15 30 45 60 75 90
ATTTTTAGAACTGACCAACCATGTTCACGTAAGTGACGTGATGACGCGCGCTGCGCGCGCGCCTTCGGACGTCACACGTCACTTACGTTT
TAAAAATCTTGACTGGTTGGTACAAGTGCATTCACTGCACTACTGCGCGCGACGCGCGCGCGGAAGCCTGCAGTGTGCAGTGAATGCAAA
SacIII SacIII HgaIFnu4HIHhaI Acy1SacIII SacIII SacIII
3 (m i s g s g s l n q g a k r k w a w f k v y
105 120 135 150 165 180
CACATGGTTGGTCAGTTCTAAAAATGATAAGCGGTTCAGGGAGTTTAAACCAAGGCGCGAAAAGGAAGTGGGCGTGGTTTAAAGTATATA
GTGTACCAACCAGTCAAGATTTTTACTATTCGCCAAGTCCCTCAAATTTGGTTCCGCGCTTTTCCTTCACCCGCACCAAATTTCATATAT
FnuDIIHhaI
k q l l k s v t y l f f h s v s r d a q k e s n q l t) M A G
195 210 225 240 255 270
AGCAACTACTGAAGTCAGTTACTTATCTTTTCTTTCATTCTGTGAGTCGAGACGCACAGAAAGAGAGTAACCAACTAACCATGGCTGGAA
TCGTTGATGACTTCAGTCAATGAATAGAAAAGAAAGTAAGACACTCAGCTCTGCGTGTCTTTCTCTCATTGGTTGATTGGTACCGACCTT
HinFITaqIHgaI NcoI
N A Y S D E V L G A T N W L K E K S N Q E V F S F V F K N E
285 300 315 330 345 360
ATGCTTACTCTGATGAAGTTTTGGGAGCAACCAACTGGTTAAAGGAAAAAAGTAACCAGGAAGTGTTCTCATTTGTTTTTAAAAATGAAA
TACGAATGAGACTACTTCAAAACCCTCGTTGGTTGACCAATTTCCTTTTTTCATTGGTCCTTCACAAGAGTAAACAAAAATTTTTACTTT
BstNI
N V Q L N G K D I G W N S Y K K E L Q E D E L K S L Q R G A
375 390 405 420 435 450
ATGTTCAACTGAATGGAAAAGATATCGGATGGAATAGTTACAAAAAAGAGCTGCAGGAGGACGAGCTGAAATCTTTACAACGAGGAGCGG
TACAAGTTGACTTACCTTTTCTATAGCCTACCTTATCAATGTTTTTTCTCGACGTCCTCCTGCTCGACTTTAGAAATGTTGCTCCTCGCC
EcoRV AluIBbvIPstIMnlI AluI MnlI
E T T W D Q S E D M E W E T T V D E M T K K Q V F I F D S L
465 480 495 510 525 540
AAACTACTTGGGACCAAAGCGAGGACATGGAATGGGAAACCACAGTGGATGAAATGACCAAAAAGCAAGTATTCATTTTTGATTCTTTGG
TTTGATGAACCCTGGTTTCGCTCCTGTACCTTACCCTTTGGTGTCACCTACTTTACTGGTTTTTCGTTCATAAGTAAAAACTAAGAAACC
AvaII MnlI HinFI
V K K C L F E V L N T K N I F P G D V N W F V Q H E W G K D
555 570 585 600 615 630
TTAAAAAATGTTTATTTGAAGTGCTTAACACAAAGAATATATTTCCTGGTGATGTTAATTGGTTTGTGCAACATGAATGGGGAAAAGACC
AATTTTTTACAAATAAACTTCACGAATTGTGTTTCTTATATAAAGGACCACTACAATTAACCAAACACGTTGTACTTACCCCTTTTCTGG
BstNIHphI EcoRI*
Q G W H C H V L I G G K D F S Q A Q G K W W R R Q L N V Y W
645 660 675 690 705 720
AAGGCTGGCACTGCCATGTACTAATTGGAGGAAAGGACTTTAGTCAAGCTCAAGGGAAATGGTGGAGAAGGCAACTAAATGTTTACTGGA
TTCCGACCGTGACGGTACATGATTAACCTCCTTTCCTGAAATCAGTTCGAGTTCCCTTTACCACCTCTTCCGTTGATTTACAAATGACCT
RsaI EcoRI*MnlI AluI
S R W L V T A C N V Q L T P A E R I K L R E I A E D N E W V
735 750 765 780 795 810
GCAGATGGTTGGTAACAGCCTGTAATGTGCAACTAACACCAGCTGAAAGAATTAAACTAAGAGAAATAGCAGAAGACAATGAGTGGGTTA
CGTCTACCAACCATTGTCGGACATTACACGTTGATTGTGGTCGACTTTCTTAATTTGATTCTCTTTATCGTCTTCTGTTACTCACCCAAT
AluIPvuII EcoRI* DdeI MboII
T L L T Y K H K Q T K K D Y T K C V L F G N M I A Y Y F L T
825 840 855 870 885 900
CTCTACTTACTTATAAGCATAAGCAAACCAAAAAAGACTATACCAAGTGTGTTCTTTTTGGAAACATGATTGCTTACTATTTTTTAACTA
GAGATGAATGAATATTCGTATTCGTTTGGTTTTTTCTGATATGGTTCACACAAGAAAAACCTTTGTACTAACGAATGATAAAAAATTGAT
K K K I S T S P P R D G G Y F L S S D S G W K T N F L K E G
915 930 945 960 975 990
AAAAGAAAATAAGCACTAGTCCACCAAGAGACGGAGGCTATTTTCTTAGCAGTGACTCTGGCTGGAAAACTAACTTTTTAAAAGAAGGCG
TTTTCTTTTATTCGTGATCAGGTGGTTCTCTGCCTCCGATAAAAGAATCGTCACTGAGACCGACCTTTTGATTGAAAAATTTTCTTCCGC
MnlI DdeI HinFI
E R H L V S K L Y T D D M R P E T V E T T V T T A Q E T K R
1005 1020 1035 1050 1065 1080
AGCGCCATCTAGTGAGCAAACTATACACTGATGACATGCGGCCAGAAACGGTTGAAACCACAGTAACCACTGCGCAGGAAACTAAGCGCG
TCGCGGTAGATCACTCGTTTGATATGTGACTACTGTACGCCGGTCTTTGCCAACTTTGGTGTCATTGGTGACGCGTCCTTTGATTCGCGC
HaeIIHhaI Fnu4HIHaeIII MstIHhaI DdeIHhaIFnuDII

```
G  R  I  Q  T  K  K  E  V  S  I  K  T  T  L  K  E  L  V  H  K  R  V  T  S  P  E  D  W  M
          1095          1110          1125          1140          1155          1170
GCAGAATTCAAACTAAAAAAGAAGTTTCTATTAAAACTACACTTAAAGAGCTGGTGCATAAAAGAGTAACCTCACCAGAGGACTGGATGA
CGTCTTAAGTTTGATTTTTTCTTCAAAGATAATTTTGATGTGAATTTCTCGACCACGTATTTTCTCATTGGAGTGGTCTCCTGACCTACT
Fnu4HI EcoRI                                        AluI                 MnlIHphIMnlI

M  M  Q  P  D  S  Y  I  E  M  M  A  Q  P  G  G  E  N  L  L  K  N  T  L  E  I  C  T  L  T
          1185          1200          1215          1230          1245          1260
TGATGCAGCCAGACAGTTACATTGAAATGATGGCTCAACCAGGTGGAGAAAACCTGCTGAAAAATACGCTAGAGATTTGTACACTAACTC
ACTACGTCGGTCTGTCAATGTAACTTTACTACCGAGTTGGTCCACCTCTTTTGGACGACTTTTTATGCGATCTCTAAACATGTGATTGAG
BbvISfaNI                           BstNI                                       RsaI

L  A  R  T  K  T  A  F  D  L  I  L  E  K  A  E  T  S  K  L  T  N  F  S  L  P  D  T  R  T
          1275          1290          1305          1320          1335          1350
TAGCCAGAACCAAAACAGCATTTGACTTAATTTTAGAAAAAGCTGAAACCAGCAAACTAACCAACTTTTCACTGCCTGACACAAGAACCT
ATCGGTCTTGGTTTTGTCGTAAACTGAATTAAAATCTTTTTCGACTTTGGTCGTTTGATTGGTTGAAAAGTGACGGACTGTGTTCTTGGA
                    EcoRI*      AluI

C  R  I  F  A  F  H  G  W  N  Y  V  K  V  C  H  A  I  C  C  V  L  N  R  Q  G  G  K  R  N
          1365          1380          1395          1410          1425          1440
GCAGAATTTTTGCTTTTCATGGCTGGAACTATGTTAAAGTTTGCCATGCTATTTGCTGTGTTTTAAACAGACAAGGAGGCAAAAGAAATA
CGTCTTAAAAACGAAAAGTACCGACCTTGATACAATTTCAAACGGTACGATAAACGACACAAAATTTGTCTGTTCCTCCGTTTTCTTTAT
PstIEcoRI*                                                                  MnlI

T  V  L  F  H  G  P  A  S  T  G  K  S  I  I  A  Q  A  I  A  Q  A  V  G  N  V  G  C  Y  N
          1455          1470          1485          1500          1515          1530
CTGTTTTATTTCATGGACCAGCCAGCACAGGCAAATCTATTATTGCACAAGCCATAGCACAAGCAGTTGGCAATGTTGGTTGCTATAATG
GACAAAATAAAGTACCTGGTCGGTCGTGTCCGTTTAGATAATAACGTGTTCGGTATCGTGTTCGTCAACCGTTACAACCAACGATATTAC
               Sau96IAvaII

A  A  N  V  N  F  P  F  N  D  C  T  N  K  N  L  I  W  V  E  E  A  G  N  F  G  Q  Q  V  N
          1545          1560          1575          1590          1605          1620
CAGCCAATGTAAACTTTCCATTTAATGACTGTACCAACAAGAACTTGATTTGGGTAGAAGAAGCTGGTAACTTTGGACAGCAAGTAAACC
GTCGGTTACATTTGAAAGGTAAATTACTGACATGGTTGTTCTTGAACTAAACCCATCTTCTTCGACCATTGAAACCTGTCGTTCATTTGG
                         RsaI                                MboIIAluI

Q  F  K  A  I  C  S  G  Q  T  I  R  I  D  Q  K  G  K  G  S  K  Q  I  E  P  T  P  V  I  M
          1635          1650          1665          1680          1695          1710
AGTTTAAAGCCATTTGCTCTGGTCAAACTATTCGCATTGATCAAAAAGGAAAAGGCAGCAAACAGATTGAACCAACACCAGTCATCATGA
TCAAATTTCGGTAAACGAGACCAGTTTGATAAGCGTAACTAGTTTTTCCTTTTCCGTCGTTTGTCTAACTTGGTTGTGGTCAGTAGTACT
                                  BclISau3A       BbvIFnu4HI

T  T  N  E  N  I  T  V  V  R  I  G  C  E  E  R  P  E  H  T  Q  P  I  R  D  R  M  L  N  I
          1725          1740          1755          1770          1785          1800
CCACAAATGAGAACATTACAGTGGTCAGAATAGGCTGCGAAGAAAGACCAGAACACACTCAACCAATCAGAGACAGAATGCTTAACATTC
GGTGTTTACTCTTGTAATGTCACCAGTCTTATCCGACGCTTCTTTCTGGTCTTGTGTGAGTTGGTTAGTCTCTGTCTTACGAATTGTAAG
                              BbvIFnu4HIMboII

H  L  T  H  T  L  P  G  D  F  G  L  V  D  K  N  E  W  P  M  I  C  A  W  L  V  K  N  G  Y
          1815          1830          1845          1860          1875          1890
ATCTAACACATACCTTGCCTGGTGACTTTGGTTTGGTTGACAAAAATGAATGGCCCATGATTTGTGCTTGGTTGGTAAAGAATGGTTACC
TAGATTGTGTATGGAACGGACCACTGAAACCAAACCAACTGTTTTTACTTACCGGGTACTAAACACGAACCAACCATTTCTTACCAATGG
                  BstNIHphI          HinDII          Sau96IHaeIII                    BstEII

Q  S  T  M  A  S  Y  C  A  K  W  G  K  V  P  D  W  S  E  N  W  A  E  P  K  V  P  T  P  I
          1905          1920          1935          1950          1965          1980
AATCTACCATGGCAAGCTACTGTGCTAAATGGGGCAAAGTTCCTGATTGGTCAGAAAACTGGGCGGAGCCAAAGGTGCCAACTCCTATAA
TTAGATGGTACCGTTCGATGACACGATTTACCCCGTTTCAAGGACTAACCAGTCTTTTGACCCGCCTCGGTTTCCACGGTTGAGGATATT
      NcoI    AluI                                                                  EcoRI*

N  L  L  G  S  A  R  S  P  F  T  T  P  K  S  T  P  L  S  Q  N  Y  A  L  T  P  L  A  S  D
          1995          2010          2025          2040          2055          2070
ATTTACTAGGTTCGGCACGCTCACCATTCACGACACCGAAAAGTACGCCTCTCAGCCAGAACTATGCACTAACTCCACTTGCATCGGATC
TAAATGATCCAAGCCGTGCGAGTGGTAAGTGCTGTGGCTTTTCATGCGGAGAGTCGGTCTTGATACGTGATTGAGGTGAACGTAGCCTAG
                    HphI                RsaI MnlIDdeI                    Sau3ASfaNIXhoII

L  E  D  L  A  L  E  P  W  S  T  P  N  T  P  V  A  G  T  A  E  T  Q  N  T  G  E  A  G  S
          2085          2100          2115          2130          2145          2160
TCGAGGACCTGGCTTTAGAGCCTTGGAGCACACCAAATACTCCTGTTGCGGGCACTGCAGAAACCCAGAACACTGGGGAAGCTGGTTCCA
AGCTCCTGGACCGAAATCTCGGAACCTCGTGTGGTTTATGAGGACAACGCCCGTGACGTCTTTGGGTCTTGTGACCCCTTCGACCAAGGT
TaqIMnlIBstNIAvaII                                  PstI                          AluI
```



```
 G  V  W  L  Q  P  S  D  W  Q  Y  I  C  N  T  M  S  Q  L  N  L  V  S  L  D  Q  E  I  F  N
          3165          3180          3195          3210          3225          3240
GGAGTTTGGCTCCAGCCAAGTGACTGGCAATACATTTGCAACACCATGAGCCAGCTTAACTTGGTATCACTTGATCAAGAAATATTCAAT
CCTCAAACCGAGGTCGGTTCACTGACCGTTATGTAAACGTTGTGGTACTCGGTCGAATTGAACCATAGTGAACTAGTTCTTTATAAGTTA
                                                    AluI                  BclISau3A

 V  V  L  K  T  V  T  E  Q  D  L  G  G  Q  A  I  K  I  Y  N  N  D  L  T  A  C  M  M  V  A
          3255          3270          3285          3300          3315          3330
GTAGTGCTGAAAACTGTTACAGAGCAAGACTTAGGAGGTCAAGCTATAAAAATATACAACAATGACCTTACAGCTTGCATGATGGTTGCA
CATCACGACTTTTGACAATGTCTCGTTCTGAATCCTCCAGTTCGATATTTTTATATGTTGTTACTGGAATGTCGAACGTACTACCAACGT
                          DdeI MnlI   AluI                      AluI

 V  D  S  N  N  I  L  P  Y  T  P  A  A  N  S  M  E  T  L  G  F  Y  P  W  K  P  T  I  A  S
          3345          3360          3375          3390          3405          3420
GTAGACTCAAACAACATTTTGCCATACACACCTGCAGCAAACTCAATGGAAACACTTGGTTTCTACCCCTGGAAACCAACCATAGCATCA
CATCTGAGTTTGTTGTAAAACGGTATGTGTGGACGTCGTTTGAGTTACCTTTGTGAACCAAAGATGGGGACCTTTGGTTGGTATCGTAGT
AccIHinFI                 BbvIPstIFnu4HI                        BstNI         HphISfaNI

 P  Y  R  Y  Y  F  C  V  D  R  D  L  S  V  T  Y  E  N  Q  E  G  T  V  E  H  N  V  M  G  T
          3435          3450          3465          3480          3495          3510
CCATACAGGTACTATTTTTGCGTTGACAGAGATCTTTCAGTGACCTACGAAAATCAAGAAGGCACAGTTGAACATAATGTGATGGGAACA
GGTATGTCCATGATAAAAACGCAACTGTCTCTAGAAAGTCACTGGATGCTTTTAGTTCTTCCGTGTCAACTTGTATTACACTACCCTTGT
        RsaI          HinDII  BglII

 P  K  G  I  P  Q  F  F  T  I  E  N  T  Q  Q  I  T  L  L  R  T  G  D  E  F  A  T  G  T  Y
          3525          3540          3555          3570          3585          3600
CCAAAAGGAATTCCTCAATTTTTTACCATTGAGAACACACAACAAATCACATTGCTCAGAACAGGGGACGAATTTGCCACAGGTACTTAC
GGTTTTCCTTAAGGAGTTAAAAAATGGTAACTCTTGTGTGTTGTTTAGTGTAACGAGTCTTGTCCCCTGCTTAAACGGTGTCCATGAATG
       EcoRIMnlIEcoRI*                                 DdeI           EcoRI*      RsaI

 Y  F  D  T  N  S  V  K  L  T  H  T  W  Q  T  N  R  Q  L  G  Q  P  P  L  L  S  T  F  P  E
          3615          3630          3645          3660          3675          3690
TACTTTGACACAAATTCAGTTAAACTCACACACACGTGGCAAACCAACCGTCAACTTGGACAGCCTCCACTGCTGTCAACCTTTCCTGAA
ATGAAACTGTGTTTAAGTCAATTTGAGTGTGTGTGCACCGTTTGGTTGGCAGTTGAACCTGTCGGAGGTGACGACAGTTGGAAAGGACTT
          EcoRI*                SacIII          HinDII        MnlI        HinDII      AluI

 A  D  T  D  A  G  T  L  T  A  Q  G  S  R  H  G  T  T  Q  M  G  V  N  W  V  S  E  A  I  R
          3705          3720          3735          3750          3765          3780
GCTGACACTGATGCAGGTACACTTACTGCTCAAGGGAGCAGACATGGAACAACACAAATGGGGGTTAACTGGGTGAGTGAAGCAATCAGA
CGACTGTGACTACGTCCATGTGAATGACGAGTTCCCTCGTCTGTACCTTGTTGTGTTTACCCCCAATTGACCCACTCACTTCGTTAGTCT
        SfaNI  RsaI                                          HinDIIHpaIHphI

 T  R  P  A  Q  V  G  F  C  Q  P  H  N  D  F  E  A  S  R  A  G  P  F  A  A  P  K  V  P  A
          3795          3810          3825          3840          3855          3870
ACCAGACCTGCTCAAGTAGGATTTTGTCAACCACACAATGACTTTGAAGCCAGCAGAGCTGGACCATTTGCTGCCCCAAAAGTTCCAGCA
TGGTCTGGACGAGTTCATCCTAAAACAGTTGGTGTGTTACTGAAACTTCGGTCGTCTCGACCTGGTAAACGACGGGGGTTTTCAAGGTCGT
                          HinDII                AluIAvaII   BbvIFnu4HI

 D  I  T  Q  G  V  D  K  E  A  N  G  S  V  R  Y  S  Y  G  K  Q  H  G  E  N  W  A  S  H  G
          3885          3900          3915          3930          3945          3960
GATATTACTCAAGGAGTAGACAAAGAAGCCAATGGCAGTGTTAGATACAGTTATGGCAAACAGCATGGTGAAAATTGGGCTTCACATGGA
CTATAATGAGTTCCTCATCTGTTTCTTCGGTTACCGTCACAATCTATGTCAATACCGTTTGTCGTACCACTTTTAACCCGAAGTGTACCT
                  AccI                                          HphI  EcoRI*      AvaII

 P  A  P  E  R  Y  T  W  D  E  T  S  F  G  S  G  R  D  T  K  D  G  F  I  Q  S  A  P  L  V
          3975          3990          4005          4020          4035          4050
CCAGCACCAGAGCGCTACACATGGGATGAAACAAGCTTTGGTTCAGGTAGAGACACCAAAGATGGTTTTATTCAATCAGCACCACTAGTT
GGTCGTGGTCTCGCGATGTGTACCCTACTTTGTTCGAAACCAAGTCCATCTCTGTGGTTTCTACCAAAATAAGTTAGTCGTGGTGATCAA
          HaeIIHhaI             AluIHinDIII

 V  P  P  P  L  N  G  I  L  T  N  A  N  P  I  G  T  K  N  D  I  H  F  S  N  V  F  N  S  Y
          4065          4080          4095          4110          4125          4140
GTTCCACCACCACTAAATGGCATTCTTACAAATGCAAACCCTATTGGGACTAAAAATGACATTCATTTTTCAAATGTTTTTAACAGCTAT
CAAGGTGGTGGTGATTTACCGTAAGAATGTTTACGTTTGGGATAACCCTGATTTTTACTGTAAGTAAAAAGTTTACAAAAATTGTCGATA
                                                                                    AluI

 G  P  L  T  A  F  S  H  P  S  P  V  Y  P  Q  G  Q  I  W  D  K  E  L  D  L  E  H  K  P  R
          4155          4170          4185          4200          4215          4230
GGTCCACTAACTGCATTTTCACACCCAAGTCCTGTATACCCTCAAGGACAAATATGGGACAAAGAACTAGATCTTGAACACAAACCTAGA
CCAGGTGATTGACGTAAAAGTGTGGGTTCAGGACATATGGGAGTTCCTGTTTATACCCTGTTTCTTGATCTAGAACTTGTGTTTGGATCT
AvaIISau96I                   AccI  MnlI                          BglIISau3A
```

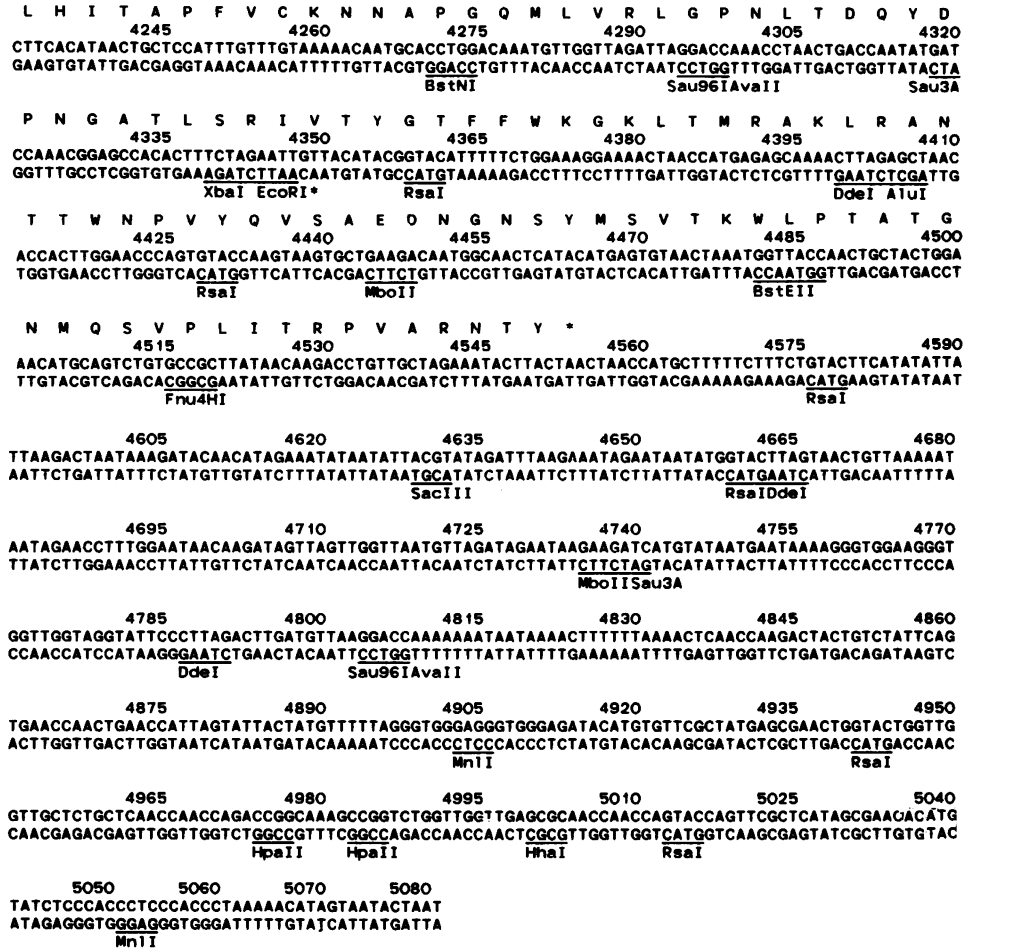


Figure 2. Complete nucleotide sequence of MVM(p) DNA. The upper strand is the complementary strand (5' → 3') and the lower strand is the viral strand (3' → 5'). All restriction endonuclease sites are underlined and indicated by their accepted abbreviations. The major open reading frames are translated above the sequence (single letter code). The major coding region in the left half of the genome begins at 261 (ATG=Met) and ends at 2277 (TAA=ochre). While the major left coding region (frame 3) begins with an ATG at 261, reading frame 3 is open back to another methionine codon at position 114. We have included this potential peptide sequence using lower case letters and it is bracketed. However, we have not detected transcripts in this region, and there is no TATA sequence upstream of this region (Pintel et al, preceeding paper). From arguments presented in the text, we believe the major coding region in the right half of the DNA begins at one of 2286, 2332 or 2354

and splicing around nucleotide 2300 shifts the reading frame at or somewhat downstream of 2383 to frame 1, which is open up to the TAA at 4552. The translation products starting at 2286, 2332, or 2354 are listed up to the first in phase termination codon. Translation of the large open reading frame located in the right half of the genome (frame 1, Figure 4) and beginning at 2383 (AAT, N) is also given. As noted in the text, the first ATG codon, in phase, occurs at 2794. Because we are not certain about the splice at map unit 46 we do not know how many of the codons between 2383 and 2794 are used. Presumably most are, however we have placed brackets around the amino acid in this region to emphasize this uncertainty. See text for further details.

amino acid coded for, while in the second case, it is a glycine or glutamic acid residue.

Base Composition

From the sequence presented here, the base composition of the viral strand is A, 24.4%; T, 33.4%; C, 21.7%; and G, 20.5%. This compares reasonably well with the early data of McGeoch et al. (21) (A=26.5%; T=32.7%; C=21.4%; G=19.5%). Analysis of the sequence also shows that the left and right halves of the genome, which encodes two distinct, but overlapping, transcription units (11), have approximately the same base composition. The nearest neighbor frequencies, based on the sequence data, are summarized in Table I.

Although there are a number of discrepancies between the nearest neighbor frequencies deduced from the sequence and those determined by McGeoch et al. (21) using in vitro cDNA synthesis, these variations are not surprising since the 3'-and 5'-terminal sequences would either be missing or markedly under-represented in the cDNA pool. For example, the relatively high abundance of

Table I

Nearest Neighbor base frequencies in MVM DNA C-strand[+].

Dinucleotide	Frequency	Dinucleotide	Frequency
CpC	39.75	ApC	82.05
CpT	63.36	ApT	58.84
CpA	85.20	ApA	122.20
CpG	16.92	ApG	70.64
TpC	33.06	GpC	50.37
TpT	76.54	GpT	44.86
TpA	60.02	GpA	66.31
TpG	73.99	GpG	55.88

[+] Frequencies are expressed as residues per thousand bases.

the dinucleotide CpG (Table I) is due in large part to a track of CpG residues in the 3'-hairpin structure (15).

Structural Features at the Genomic Termini.

Although illustrated in Figure 2 in a conventional duplex form, the 3'- and 5'-termini of the single-stranded genome contain palindromic sequences which adopt a "hairpin" configuration. The 3'-hairpin structure contains 115 nucleotides, 104 of which are base-paired (Figure 3a). Both virion (V) DNA and monomer replicative form (RF) DNA contain the same unique 3'-end sequence. In contrast, the 5'-terminal hairpin structure of V DNA, which is composed of 206 nucleotides, possesses two nucleotide sequences that are the inverted complement of each other (designated "Flip" and "Flop" sequences in Figure 3b). In both sequence orientations 200 of the 206 nucleotides are capable of base-pairing to form a stable duplex structure. The 5'-terminal sequence of the V strand in monomer RF DNA exhibits two sequence orientations as well, but it also contains an additional 18 nucleotides that are not present in V DNA encapsidated into mature virions. These additional nucleotides are complementary to a region of the genome between nucleotides 4858 and 4876, which lies immediately to the 3'-side of the 5'-hairpin structure itself. The additional nucleotides present in monomer RF DNA also exhibit homology with

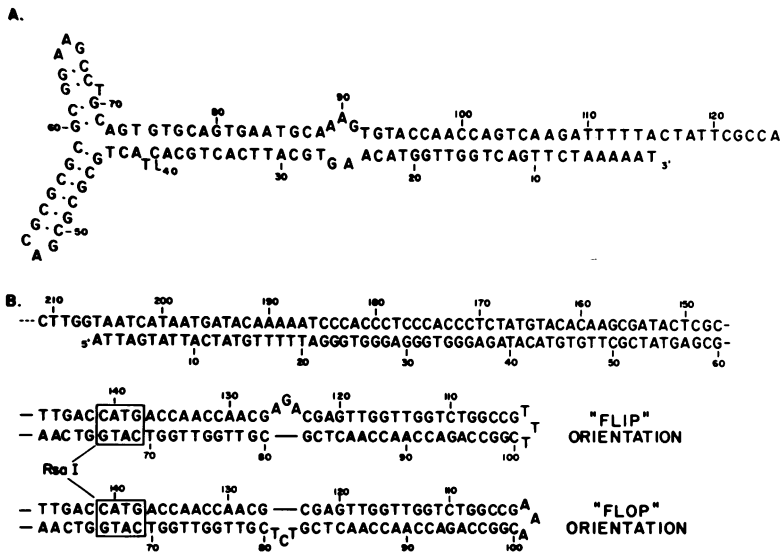


Figure 3. Nucleotide sequences at the 3' (A) and 5' (B) terminus of MVM virion DNA, illustrated in their hairpin conformation.

sequences near the boundary of the 3'-hairpin sequence, between nucleotides 96 and 104. The unique 3'-end sequence, the sequence heterogeneity at the 5'-end, and the additional nucleotides present at the 5'-end of RF DNA but not mature V DNA, are all manifestations of the mechanism by which the MVM genome is replicated. The significance of these observations with regard to the mechanism of MVM DNA replication is discussed elsewhere (12).

Transcription initiation signals

Studies on the transcripts encoded by the MVM genome have shown that there are two transcript starts, one about 4.5 map units (mu) and the other about 39 mu. (11,13). The transcripts, all derived from the complementary (C) strand of DNA, are spliced as illustrated at the top of Figure 4. RI (4.8 Kb) consists of almost full length transcripts with a small splice at 46 mu. R2 (3.3 Kb) consists of a short "leader" followed by a large splice from 10.7 mu to 38 mu, a second leader of about 300 nucleotides, followed by a small splice at 46 mu, and finally the main body of the RNA. The major transcript, R3 (3.0 Kb) consists of a short "leader" from 40 mu to 46 mu, which is joined by a small splice to the remainder of the transcript. All of these transcripts extend to approximately 95 mu and thus the genome contains two overlapping transcription units.

In eukaryotes, RNA polymerase II transcription control regions are often characterized by a TATA box which occurs approximately 30 nucleotides upstream of the actual site of initiation of the RNA chains. While the precise

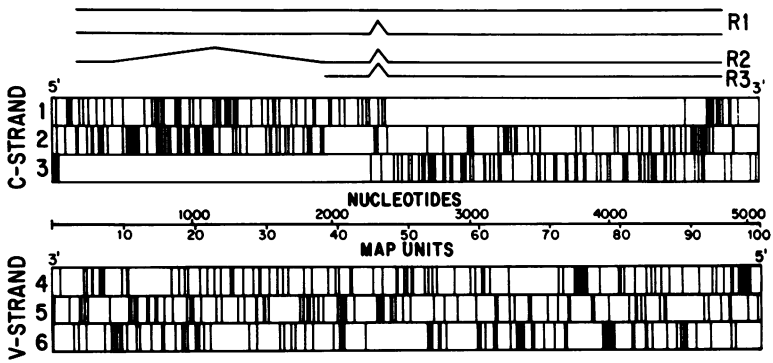


Figure 4. Open reading frames in the MVM(p) genome. The upper three lines correspond to the three reading frames in the C-strand. The lower three lines correspond to the three reading frames in the V-strand. A summary of the transcript splicing patterns is indicated above the figure (see Pintel et al., preceeding paper).

function of the TATA sequence is not clear, it is generally though to specify the start site for transcription (see reference 22 for a review). A search of the MVM sequence (C-strand) for TATA related sequences upstream of the apparent transcript start sites, 4 mu and 39 mu, indicates that there is a sequence TATAAG at position 178, and a sequence TATAAA at position 1976.

Other sequences in the -70 to -80 region have also been implicated in RNA polymerase II transcription initiation of both viral and non-viral genes (23-25). The consensus sequence, GGPyCAATCT (CCAAT), is believed to control the efficiency of transcription. A search for a similar block of nucleotides upstream of the transcript start at 4 mu failed to reveal a related sequence. Upstream of the second transcript start (at 39 mu) is the sequence, PyCAATCT (at nucleotide 1889). However, this would appear to be somewhat farther away from the transcript start than -80. It is interesting to note that transcript R3 (see above) is the major one *in vivo* and its presumed "promotor" at 39 mu is also the better promotor *in vitro* (11). Experiments to locate the precise 5' ends of the viral transcripts, which should coincide with capping sites (22,26), are in progress.

Polyadenylation signals

The most characteristic feature of the sequences at the 3' ends of eukaryote mRNAs is the sequence AAUAAA about 20 nucleotides before the PolyA track (27-28). A search of the C-strand of MVM shows there are only two polyadenylation signals in the entire molecule, and both are located at the extreme right hand end of the genome (position 4599 and 4817). Both occur just beyond the termination codon (at 4552) of the right hand open reading frame (Figure 4) and it is conceivable that either (or both) are functional polyadenylation sites.

We have also analyzed the sequence for direct and indirect repeats as well as potential secondary structure in the transcription control regions. The results of these studies will be published elsewhere.

In addition to nucleotide sequence signals in DNA, many studies have suggested the regulation of transcription by RNA polymerase II is influenced by chromatin structure (see reference 22 for a review). The genomes of DNA viruses such as SV40 (29) and Adenovirus (30) exist within the cell complexed with cellular histones in nucleosomal structures. Indeed, differences in the structure of adenovirus chromatin early and late in infection may be important in the transcriptional control of adenovirus gene expression (38). Although MVM DNA has also been reported to be associated with nucleosomes *in vivo* (31), attempts to confirm this observation with both MVM and H-1 have been unsucc-

essful (J.J. Leary, unpublished results, M. Goulian, personal communication). Thus the influence of chromatin structure on parvovirus transcription remains unclear.

Potential translation regions within the MVM genome

An analysis of the sequence for translation termination signals is illustrated in Figure 4. Reading frames 1,2, and 3 correspond to the C-strand which is equivalent to the mRNAs. There are no detectable transcripts in the leftward direction, equivalent to the V-strand (11), and it can be seen from Figure 4 that indeed there are no extensive open reading frames in the V-strand sequence (frames 4,5, and 6).

The extensive open reading frames (lines 1,2,3 in Figure 4) suggest that virtually the entire MVM genome can code for polypeptides. Preliminary experiments suggest the viral capsid proteins, VP1 or A (83,000 daltons), VP2 or B (64,000 daltons) and VP3 or C (61,000 daltons) are encoded by the right half of the genome while a non-capsid polypeptide of 85,000 daltons is encoded by the left half of the genome (S. Cotmore and P. Tattersall, personal communication; see discussion below). Studies by both Tal et al. (32) and Pintel et al. (11) indicate that the messages are spliced, and a summary of these spliced messenger RNAs is shown in Figure 4. Because of uncertainly about the precise splice positions and translation start signals used (see discussion below) it was only possible to tentatively identify which transcript is the messenger for which polypeptide(s) using sequence data alone.

Transcript R3; coding region for viral capsid protein VP2

R3 (approximately 3.0 Kb) is transcribed from the promotor at 39 mu and contains a small splice at 46 mu. It is not known where translation begins on this message. In a review by Kozak, (33) it is noted that the size of mRNA 5' non-coding regions can vary from less than 10 to somewhat more than 300 nucleotides, and the extremes are found predominantly in viral mRNAs.

With respect to the R3 messenger, the first ATG within the large open region of frame 1 (Figure 4) occurs at position 2794, more than 700 nucleotides from the 5' end of the mRNA and more than 400 nucleotides downstream from the TGA termination codon at 2380. Both of these observations suggest to us that the ATG at 2794 is not functional as an initiation codon, but rather an earlier translation start signal must be recognized by the ribosomes. Other ATGs at 2044 and 2173 (both frame 1) (see Figure 5) are followed closely by in frame termination codons, again suggesting these are also non-functional translation initiation signals. However, ATGs at 2286 (frame 3), 2332 (frame 1), and 2354 (frame 2) may be functional when the primary transcript is spli-

ced appropriately in the region of 46 mu.

An analysis of the region from nucleotide 2200 to 2500 for putative splicing consensus sequences (34,35) indicated that there are nine potential splicing patterns that would permit the reading of the extended coding sequence (frame 1) starting with the ATG at 2286 (frame 3), 2332 (frame 1) or 2355 (frame 2), all giving exons between 51 and 119 nucleotides, similar in size to that observed experimentally (11). Possible left-hand splice junctions were found for the consensus sequences AG↓GTA (site A), AG↓GT (site B), and G↓GT (sites C,D,E,F,G,H,I) (see Figure 5). Right-hand splice junctions were found for the consensus sequences TNCAG↓ (site a) and CAG↓ (sites b and c). Using a computer program to facilitate the analysis (36) all possible splices (e.g. A/a, A/b, etc.) were analyzed to determine if the reading frame was shifted to frame 1 (Figure 4) and open from one of the ATGs at 2282, 2332, or 2354. Surprisingly, it was found that nine splices could occur which would

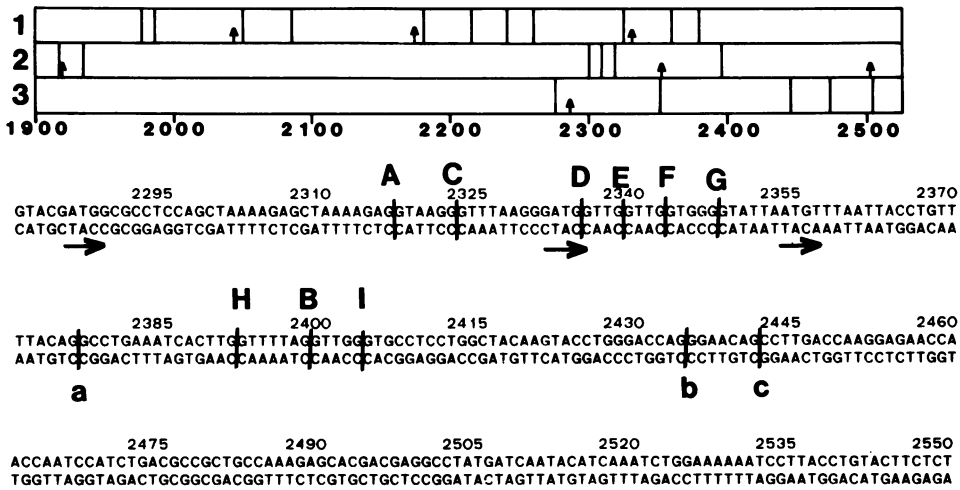


Figure 5. Summary of potential splices around map unit 46. The upper diagram illustrates the open reading frames in the region of nucleotides 1900 to 2500. Solid bars indicate termination codons (TAA, TAG, TGA) in reading frames 1, 2, and 3. Arrows indicate initiation codons (ATG). In the sequence below, capital letters A to I mark potential left hand splice junctions and lower case letter a to c mark potential right hand splice junctions. Possible splices (e.g., A/a, A/b, etc.) which eliminate in phase termination codons and which would results in translation of frame 1, the large open coding region in the right half of the genome, are summarized in Table II. See text for further details.

Table II

Summary of Splices which would permit reading of frame 1 starting at an AUG codon upstream of the splice, in transcript R3.

Spliced nucleotides	AUG start	Initial reading frame	Intron size
2315 to 2435 - A/b	2286	3	119
2322 to 2435 - C/b	2286	3	113
2334 to 2435 - D/b	2286	3	101
2338 to 2435 - E/b	2332	1	97
2338 to 2442 - E/c	2286	3	104
2342 to 2442 - F/c	2332	1	100
2347 to 2435 - G/b	2332	1	88
2347 to 2442 - G/c	2286	3	95
2391 to 2442 - H/c	2354	2	51

generate mRNAs that had a ATG (at either 2282, 2332 or 2354) followed by open reading frame up to the TAA termination signal at 4552. These results are summarized in Table II. The intriguing possibility exists that there is alternate splicing possible near map unit 46 and this may, in part, account for the viral capsid proteins containing many identical tryptic and chymotryptic peptides. These capsid polypeptides would contain different N-terminal regions followed by an identical C-terminal main body. Studies to analyze the splice position(s) near m.u. 45-46 in more detail are in progress. However, it should be noted that alternate splicing patterns that generate mRNA species sharing translated sequences have been observed previously with other viral systems, probably the best known examples are the two T-antigens of SV40 (T and t) (see reference 35 for a review). If, as we are suggesting, the ATGs at 2282, 2332 and/or 2354 are functional, then the first ATG (AUG in the mRNA) is not recognized by the ribosomes during initiation of protein synthesis (see Figure 5) possibly providing another example for a "relaxed" scanning mechanism by the ribosome (33). It should be noted that an open reading frame of approximately 2100 nucleotides (4554-2354, minus ~ 100, the size of the intron at 46 m.u.) would code for a polypeptide of approximately 700 amino acids (70,000 daltons). This is significantly smaller than VP1 (83,000) although somewhat larger than VP2 (64,000 daltons) (37). If the ATG

at 2794 were the true starting codon, this mRNA could only encode a polypeptide of approximately 58,600 daltons. This type of argument again suggests that the ATG at 2794 is not used in translation initiation.

Additional evidence indicates that transcript R3, the most abundant transcript, is the messenger for the viral capsid protein VP2. A comparison of the amino acid analysis of full and empty particles (14) with putative polypeptides encoded by the left and right halves of the complementary strand is summarized in Table III. It is clear that the amino acid composition of a polypeptide derived from the right half of the genome is much closer to that of both full and empty particles. Empty particles contain predominantly VP2, while full particles contain predominantly VP2 or VP3, depending on the preparation (14). However, it is clear that VP3 is a cleavage product of VP2, and it would contain 95% of the amino acid sequence of VP2 (37). The tentative assignment of the R-3 transcript as the messenger RNA for the VP2 capsid protein has recently been confirmed by in vitro translation of hybrid-selected RNA (S. Cotmore and P. Tattersall, manuscript in preparation).

Transcript R2; coding region for viral capsid protein VP1

Transcript R2 contains two "leader" sequences separated by introns of 1.5 Kb and less than 100 nucleotides (11; Figure 4). The main body of the transcript comes from the right half of the genome and hence, presumably encodes viral capsid protein. However, unlike R3, it contains additional nucleotides derived from the extreme left hand end of the genome. Altogether, if nucleotides between 1950 and 2300 (that correspond to the second exon) were also translated, this message could code for approximately 84,000 daltons of protein.

The possibility that this message encodes viral capsid protein VP1, starting with (Met)-Ala-Gly at nucleotide 261, was suggested initially by an analysis of "empty" viral particles, which revealed an N-terminus sequence of Ala-Gly (P. Paradisio and D. Ward, unpublished observations). However, VP1 is not a major protein of either empty or full particles and most of the polypeptides from viral proteins contain a blocked amino terminus (P. Paradisio, personal communication). Hence, the Ala-Gly sequence may have corresponded to VP1, or alternatively, to the N-terminus of the cleavage polypeptide (VP3).

A search of the entire C-strand sequence reveals that the nucleotides GCNGGN (Ala-Gly) occur 16 times. Several of these occur in regions where the reading for that phase is closed [at 413 (frame 2); 634 (1); 961 (1); 1064 (2); 1130 (2); 1372 (1); 2900 (2); 2696 (2); 4982 (2)]. Others, while they

Table III

The amino acid composition (mole percent) of putative polypeptides encoded by the left and right halves of the MVM (C-strand). For comparison the experimental values for full and empty particles are included (Tattersall et al., 1976).

	Left Half frame 3 (261-2279)	Right Half frame 1 (2332-4554)	Full Particles	Empty Particles
Asx	10.9	11.7	11.84	12.33
Thr	8.8	8.6	7.92	8.25
Ser	5.5	6.6	5.10	5.09
Glx	**12.2	8.6	8.47	8.60
Cys	2.5	0.8	1.22	1.22
Pro	*4.6	6.9	6.47	6.80
Gly	*6.2	9.0	12.20	9.00
Aly	5.8	8.0	7.38	7.42
Val	5.8	6.2	6.10	7.03
Met	2.1	1.9	1.64	2.03
Ile	**4.6	3.6	3.28	3.39
Leu	**8.0	6.2	5.92	6.25
Tyr	*1.2	3.1	3.73	3.87
Phe	3.9	3.5	3.19	3.36
His	1.8	2.4	2.64	2.77
Lys	**8.0	4.5	4.28	4.08
Arg	3.6	4.2	4.55	4.47
Trp	3.4	*2.6	4.06	4.06

* low
** High

occur in phase with open reading frames, are either too near a termination codon to be at the amino terminal end of proteins of 61,000 to 83,000 daltons (3703 and 3838 in frame 1) or are within the large intron that is spliced out of transcript R2 (1593, frame 3). This leaves four GCNGGN, potential Ala-Gly coding sequences at 264, 2118, 2151 and 2695. It is not easy to see how those at positions 2118 (3) and 2151 (3) would encode Ala-Gly as the N-terminal of any of the viral capsid proteins, either by their proximity to an initiation translation site or by proteolytic cleavage of a larger precursor. However, GCTGGA at 264 immediately follows a methionine codon and cleavage of a larger

polypeptide adjacent to the Ala-Gly that is encoded by GCTGGT (nucleotide 2695) would generate a truncated polypeptide of approximately 62,000 daltons. Other evidence suggests the N-terminal Ala-Gly sequence observed by Paradisio and Ward is unlikely to correspond to the region at 2695. The amino acid sequence in this region is ... (Gly-Val-Ser-Arg-Ala-Gly-Lys-Arg-). If cleavage of VP2 to VP3 occurs between Arg-Ala (generating an Ala-Gly N-terminus), then all of the tryptic peptides present in VP3 should occur in VP2. However, this is not the case, because an iododinated tryptic peptide unique to VP3 has been observed (37). Hence, cleavage of VP2 to VP3 cannot occur to the right of an arginine residue. Consequently, we assume that the Ala-Gly observed by Paradisio and Ward comes from the N-terminus of VP1 encoded adjacent to the methionine at position 261.

Additional information indicates that some of the extra tryptic peptides in the VP1 protein are basic and rich in arginine (37). Analysis of putative protein coded in the region from 1950 to 2300 (frame 2, Figure 4) shows that this region is indeed both basic and arginine rich (12 Arg; 5 Lys; 3 Asp; 9 Glu). The molar percentage of these amino acids in this region for reading frame 2 is 9.4% (Arg); 3.4% (Lys); 7.9% (Glu); 2.6% (Asp). [Similar values for frame 3 are 1.7% (Arg); 4.3% (Lys); 8.5% (Glu); 5.1% (Asp)]. Comparison values from the bulk of the viral capsid proteins (region 2350 to 4554, frame 1) are 4.2% (Arg); 4.5% (Lys); 3.5% (Glu); 5.5% (Asp).

Again, the tentative assignment of the R2 transcript as the mRNA for VP1 protein strictly on the grounds of sequence argument have recently been confirmed experimently by in vitro translation (S. Cotmore and P. Tattersall, personal communication). Their studies also indicate that the VP2 protein can be translated from the R2 transcript in vitro.

Transcript R1: Non-capsid viral protein

Transcript R1, (4.8 Kb), is a minor RNA component (5-10% of the total) which, unlike R1 or R2, includes the entire genome from about nucleotides 200 to 4700, excluding only a small region (less than 100 nucleotides) near map unit 46. The sequence of this region indicates that it includes a large open reading frame in both the left half and right half of the molecule and assuming an appropriate splice at 46 m.u. (see Figure 4 and Table I) it is conceivable that a very large peptide (approximately 140,000 daltons) could be made from this transcript. No viral protein of this size has been detected. However, recent studies indicate that there are two non-capsid viral protein of approximately 85,000 daltons and 25,000 daltons that are synthesized in MVM infected cells (S. Cotmore and P. Tattersall, personal communication).

Cotmore and Tattersall isolated the R1 transcript, using a plasmid clone of the PstI C fragment of MVM DNA which hybridizes uniquely to this RNA, and demonstrated that it promoted the translation of the 85,000 dalton non-structural protein in vitro.

A polypeptide of approximately 67,000 daltons could be translated from the ATG at 261 to the stop codon at 2277. However, if the splice at map unit 46 extended the open reading frame (in frame 2) to the TAA at 2687, a polypeptide of approximately 77,000 could be translated. Alternatively, with the appropriate splice a full size polypeptide (140,000 daltons) could be made; peptide cleavage could then generate the 85K non-capsid protein. The function of this protein in the MVM life cycle is currently being examined.

To summarize, the present results indicate that transcript R1 encodes an 85Kd non-capsid viral protein, R2 encodes the 83 Kd VP1 structural protein, and R3 encodes VP2, the major 64 Kd capsid protein. The transcript which encodes the newly discovered 25 Kd non-structural protein has not as yet been identified, although the possibility that it is translated from the very minor 1.8 Kb transcript R4 (11) is under active investigation.

ACKNOWLEDGEMENTS

This work was supported by grants from the United States Public Health Service (GM-20124 and CA-16038 to D.C.W.) and by the Medical Research Council of Canada and the British Columbia Health Care Research Fund (to C.R.A.). We thank Peter Tattersall, Jeffry Leary and Susan Cotmore for providing cloned stocks of MVM DNA and for many helpful comments and suggestions.

*Author to whom correspondence should be addressed.

REFERENCES

1. Berns, K.I. and Hauswirth, W.W. (1978) In Replication of Mammalian Parvoviruses, D.C. Ward and P.J. Tattersall, Eds. (New York:Cold Spring Harbor Laboratory), pp. 12-32.
2. Crawford, L.V. (1966) Virology 29, 605-612.
3. Tattersall, P. (1972) J. Virol. 10, 586-590.
4. Spalholz, B.A., Bratton, J., Ward, D.C. and Tattersall, P. (1982) In "Tumor Viruses and Differentiation", Cetus-UCLA Symposium on Molecular and Cellular Biology. Academic Press N.Y., in press.
5. Rose, J.A. (1974) In "Comprehensive Virology" (H. Fraenkel-Conrat and R.R. Wagner, eds.) pp. 1-61 vol. 3, New York, Plenum Press.
6. Majaniemi, I., Tratschin, J.D. and Siegl, G. (1981) Abstracts Vth Int. Congress of Virology, p. 366.
7. Tattersall, P. and Ward, D.C. (1976) Nature 263, 103-108,
8. Berns, K.I. and Hauswirth, W.W. (1979) Adv. Virus Res. 25, 407-449.

9. Ward, D.C. and Tattersall, P. (1982) in The Mouse in Biomedical Research, Vol. 2. H.L. Foster, J.D. Small and J.G. Fox, Eds. Academic Press New York, pp. 313-334.
10. Marcus, C.J., Laughlin, C.A. and Carter, B.J. (1981) Eur. J. Biochem. 121, 147-154.
11. Pintel, D., Dadachanji, D., Astell, C.R. and Ward, D.C. (1983) Nucl. Acids. Res., preceeding manuscript.
12. Astell, C.R., Thomson, M., Chow, M.B. and Ward, D.C. (1983) Cold Spring Harbor Symp. Quant. Biol. 47, in press.
13. Astell, C.R., Dadachanji, D., Merchlinsky, M., Pintel, D. and Ward, D.C. (1981b) Abstracts Vth Intl. Congress Virology p. 365.
14. Tattersall, P., Cawte, P.J., Shatkin, A.J., Ward, D.C. (1976) J. Virol. 20, 273-289.
15. Astell, C.R., Smith, M., Chow, M.B. and Ward, D.C. (1979) Cell 17, 691-703.
16. Astell, C.R., Smith, M., Chow, M.B. and Ward, D.C. (1979) Virology 94, 669-674.
17. Astell, C.R., Ahlstrom-Honasson, L., Smith, M., Tatchell, K., Nasmyth, K. and Hall, B.D. (1981) Cell 27, 15-23.
18. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. USA 74, 5463.
19. Maxam, A. and Gilbert, W. (1977) Proc. Natl. Acad. Sci. USA 74, 560-564.
20. Armentrout, R. et al. (1978) in Replication of Mammalian Parvoviruses, D.C. Ward and P. Tattersall, Eds. Cold Spring Harbor Press, Cold Spring Harbor, New York, pp. 523-526.
21. McGeoch, D.J., Crawford, L.V. and Follett, E.A.C. (1970) J. Gen. Virol. 63, 33-40.
22. Shenk, T. (1981) in Current Topics in Microbiol. and Immunol. 93, 25-46.
23. Benoist, C., O'Hare, K., Breathnach, R. and Chambon, P. (1980) Nucl. Acids Res. 8, 127-142.
24. Grosschedl, R. and Birnsteil, M.L. (1980) Proc. Natal. Acad. Sci. USA 77, 1432-1436.
25. Corden, J., Wasylyk, B., Buchwalder, A., Sassone-Corsi, P., Kedinger, C., and Chambon, P. (1980) Science 209, 1406-1414.
26. Ziff, E.B. (1980) Nature 287, 491.
27. Proudfoot, N.J. and Brownlee, G.G. (1976) Nature 263, 211-214.
28. Yelverton, E., Leung, D., Weck, P., Gray, P.W. and Goedel, D.V. (1981) Nucl. Acids Res. 9, 731-741.
29. Griffiths, J.D. (1976) Science 187, 1202-1203.
30. Sergeant, A., Tigges, M.A. and Raskas, H.J. (1979) J. Virol. 29, 888-898.
31. Ben-Asher, E., Bratosin, S. and Aloni, Y. (1982) J. Virol. 41, 1044-1054.
32. Tal, J., Ron, D., Tattersall, P., Bratosin, S., Aloni, Y. (1979) Nature 279, 649-651.
33. Kozak, M. (1981) Current Topics in Microbiol. and Immunol. 93, 81-123.
34. Breathnach, R., Benoist, C., O'Hare, K., Gannon, F., Chambon, P. (1978) Proc. Natl. Acad. Sci. USA 75, 4853-4857.
35. Flint, S.J. (1981) Current Topics in Microbiol. and Immunol. 93, 47-79.
36. Delaney, A.D. (1982) Nucl. Acids Res.
37. Tattersall, P., Shatkin, A.J. and Ward, D.C. (1977) J. Mol. Biol. III, 375-394.
38. Daniell, E., Groff, D.E. and Fedor, M.J. (1981) Mol. Cell Bio. 1, 1094-1105.