JOURNAL OF VIROLOGY, Sept. 1983, p. 495-504 Vol. 47, No. 3
0022-538X/83/090495-10$02.00/0

Nucleotide Sequencing of an Apparent Proviral Copy of *env* mRNA Defines Determinants of Expression of the Mouse Mammary Tumor Virus *env* Gene

JOHN E. MAJORS* AND HAROLD E. VARMUS

Department of Microbiology and Immunology, University of California, San Francisco, California 94143

Received 21 March 1983/Accepted 7 June 1983

To extend our understanding of the organization and expression of the mouse mammary tumor virus genome, we determined the nucleotide sequence of large regions of a cloned mouse mammary tumor virus strain C3H provirus that appears to be a DNA copy of *env* mRNA. In conjunction with analysis of several additional clones of integrated and unintegrated mouse mammary tumor virus DNAs, we came to the following conclusions: (i) the mRNA for *env* is generated by splicing mechanisms that recognize conventional eucaryotic signals at donor and acceptor sites with a leader of at least 289 bases in length; (ii) the first of three possible initiation codons for translation of *env* follows the splice junction by a single nucleotide and produces a signal peptide of 98 amino acids; (iii) the amino terminal sequence of the major virion glycoprotein gp52^{env} is confirmed by nucleotide sequencing and is encoded by a sequence beginning 584 nucleotides from the 5′ end of *env* mRNA; (iv) the final 17 amino acids at the carboxyl terminus of the primary product of *env* are encoded within the long terminal repeat by the 51 bases at the 5′ end of the U3 domain; and (v) bases 2 through 4 at the 5′ end of the long terminal repeat constitute an initiation codon that commences an open reading frame capable of directing the synthesis of a 36-kilodalton protein.

The strategies used by retroviruses to express their polycistronic genomes are now well established. Viral genes are perpetuated in infected cells as a provirus integrated within a host chromosome, viral RNA is synthesized by host RNA polymerase II beginning at a site within the long terminal repeat (LTR) in proviral DNA, subgenomic mRNAs are generated by cellular splicing mechanisms, and mature viral proteins are produced from polyprotein precursors by proteolytic cleavage (see reference 29 for a review). Nevertheless, many important aspects of this scheme have yet to be fully elucidated.

We were interested in the mechanisms of gene expression employed by the mouse mammary tumor virus (MMTV), a retrovirus regulated at the transcriptional level by glucocorticoid hormones (23, 31) and capable of inducing mammary adenocarcinomas in susceptible mice. To explore the genetic content, organization, and expression of the MMTV genome, we used molecularly cloned viral DNAs as substrates for DNA sequence analysis. MMTV is known to encode at least three primary protein products, each of whose synthesis is probably directed by a separate mRNA: a 77-kilodalton (kd) polyprotein (Pr77^{gag}) that is cleaved to form the viral core proteins, a 180-kd polyprotein (Pr180$^{gag\text{-}pol}$) that is precursor to virion reverse transcriptase, and a protein of ca. 70 kd that is processed by cleavage and glycosylation to form the viral glycoproteins gp52^{env} and gp36^{env} (4, 24). A fourth gene product could be synthesized from an open reading frame situated within the LTR (5–8, 11, 13).

In this report, we present evidence for the structure of the *env* mRNA which is based in part upon the fortuitous finding of a provirus that constitutes a reverse transcript of *env* mRNA. We determined the leader sequence of this mRNA, the donor and acceptor splice sites used to generate it from genomic RNA, the sequence of the entire *env* gene, with the unusual finding that the 3′ terminus of the gene extends 51 nucleotides into the LTR, and the sequence of the long open reading frame from the U3 domain of the LTR, also present at the 3′ end of *env* mRNA.

MATERIALS AND METHODS

Cloning and fragment isolation. The proviral substrate was generated by low-multiplicity infection of rat XC cells with the C3H strain of MMTV. Single cell clones were screened for MMTV DNA, and those containing single proviruses were analyzed (17). Cell line 8 harbored a truncated provirus, recovered within a single *Eco*RI fragment. Molecular cloning of that proviral *Eco*RI fragment in Charon 4A has been described previously (16, 17). Subclones for sequence analysis were made by digesting recombinant bacteriophage DNA with either *Cla*I or *Pst*I, followed by ligation to plasmid pBR322 that was cleaved with the appropriate enzyme and treated with bacterial alkaline phosphatase (18). Fragments for sequencing were isolated either by electrophoresis through Seaplaque (Marine Colloids) agarose gels (18) or polyacrylamide gels. Fragments were extracted from Seaplaque agarose by suspending the gel slice in two volumes of 0.3 M NaCl–20 mM Tris-hydrochloride (pH 7.5)–1 mM EDTA and heating to 68°C for 5 min, followed by two extractions with phenol and precipitation with ethanol. Fragments were isolated from polyacrylamide by the "crush and soak" method of Maxam and Gilbert (20).

End labeling and sequencing. Fragments for sequencing were labeled with ^{32}P at their 3′ ends with avian myeloblastosis virus reverse transcriptase and ^{32}P-nucleotide triphosphates as described previously (17). Fragments were labeled wtih ^{32}P at their 5′ ends with T4 polynucleotide kinase and [^{32}P]ATP as described by Maxam and Gilbert (20). Sequencing was carried out by the chemical cleavage method of Maxam and Gilbert (19, 20). Sequencing gels were 8% acrylamide buffered with 90 mM Tris-borate (pH 8.3)–1 mM EDTA (20).

RESULTS AND DISCUSSION

Strategies for cloning and sequencing. The molecular cloning of wild-type genomes of milk-borne MMTV was confounded by our inability to clone sequences from a small region within or near the *gag* gene (Fig. 1). However, in a set of clonal rat cell lines containing single MMTV C3H proviruses, we were fortunate to find one line, designated line 8, that harbors a provirus lacking both the *gag-pol* region and the single *Eco*RI site in the wild-type genome. This provirus, which appears to have arisen by reverse transcription of *env* mRNA (see below), was cloned intact with flanking cellular DNA as an *Eco*RI fragment (17) and was used for determining the nucleotide sequences of regions indicated in Fig. 2A. These include the entire 5′ LTR, the sequences downstream from the LTR, including sequences from both sides of the missing *gag* and *pol* genes, the entire *env* gene, and sequences extending beyond *env* into the 3′ LTR. (In addition, we previously determined the sequences at the host-viral junctions in this clone [17].) We also determined the sequences of selected regions of other cloned, integrated, and unintegrated MMTV C3H and MMTV RIII DNAs to define the 5′ end of *env* and the gp52^{env} coding region, to determine sequences in wild-type DNA at the boundaries of the region absent in the line 8 provirus, and to compare *env* and LTR sequences between strains. One of the substrates for these sequencing exercises is diagrammed in Fig. 2B, and others will be described below. The sequence determined from the line 8 provirus is presented in Fig. 3 in the form of the sequence of the MMTV C3H *env* mRNA.

The information derived from this sequence and ancillary sequences is pertinent to three issues to be discussed separately: (i) the structure and function of the MMTV *env* gene product, (ii) the location of the splice donor and acceptor sites for *env* mRNA, and (iii) the extent of the long open reading frame in U3.

(i) Structure and function of *env* proteins. Cell line 8 contains apparently normal 24S *env* mRNA, *env* glycoprotein precursor, and the mature *env* products, gp52^{env} and gp36^{env} (un-

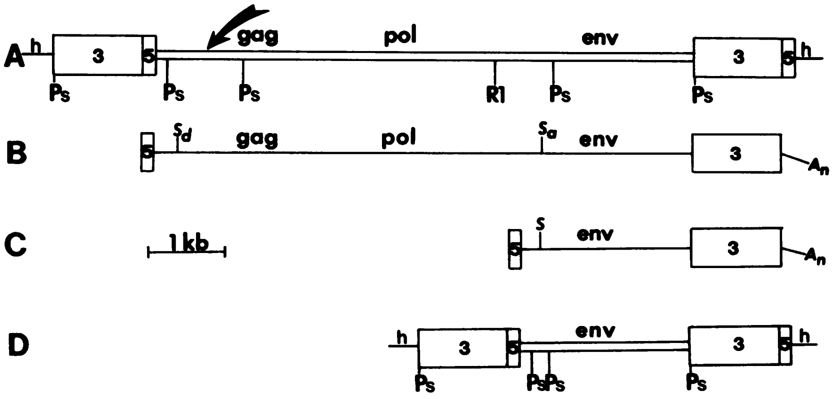

FIG. 1. Genesis and structure of the 24S MMTV *env* mRNA. (A) A wild-type MMTV provirus with sites for the restriction enzymes *Pst*I (Ps) and *Eco*RI (RI) and the approximate location of the poison sequence (large arrow). (B) 35S genomic RNA with the splice donor (S_d) and acceptor (S_a) sites. (C) The processed 24S *env* mRNA with *S* denoting the splice junction. (D) Structure of proviral copy of *env* mRNA as in line 8. Boxes labeled 3 and 5 represent U3 and U5 domains, respectively. Flanking host sequences are denoted by an h. Polyadenosine in mRNAs is denoted by an A_n.

published data; D. Robertson, personal communication); thus, we concluded that the single provirus from this line contains an intact *env* gene. As a landmark within the region of the genome previously shown to encode the *env* glycoproteins (9), we sought a nucleotide sequence that matched the recently determined amino terminus of gp52^{env} (2). This polypeptide, known to be a processed product of *env* (2–4, 24), begins with the sequence Glu-Ser-Tyr; the corresponding nucleotide sequence was located 584 bases downstream from the probable starting site for transcription in the 5′ LTR of the truncated line 8 provirus (amino acids 1, 2, and 3; Fig. 3) and lies within a continuous open reading frame of 2,064 base pairs (bp). Since gp52^{env} is thought to be generated by proteolytic removal of a signal sequence from the amino terminus of the primary product of *env* (2, 3; D. Robertson, personal communication), the nucleotide sequence on the 5′ side of the gp52^{env} coding domain was scanned for possible initiation sites for translation. Three AUG codons were present, each of which was in frame with the coding sequence for gp52^{env}, with no interrupting termination codons. Use of these initiation sites would produce signal peptides 98, 63, or 53 amino acids in length. Attempts to measure the length of the signal sequence by comparing the size of the in vitro translation product of *env* RNA with the unglycosylated, cleaved product of *env* in vivo have yielded ambiguous results, with estimates varying from 5 to 9 kd (2–4, 9; D. Robertson, personal communication). Hence, initiation at any of the three sites would produce a protein consistent with the available measurements. The amino terminal sequence of the *env* protein precursor must be determined directly to identify the initiation site unambiguously. Application of the rules of preferential use of translation initiation sites (based on the appearance of adenosine or guanosine at the −3 position and guanosine at the +4 position of known initiation sites [15]) provides little basis for choice. With the exception of a cytosine at the +4 position of the first AUG codon, all other positions are occupied by favored bases. Examination of the predicted amino acid sequence for signal sequence signatures is also not illuminating. Most signal sequences have hydrophilic amino terminal regions followed by a hydrophobic core which immediately precedes the cleavage site (10). As expected, we found such a hydrophobic region (amino acids −1 to −26). The preceding

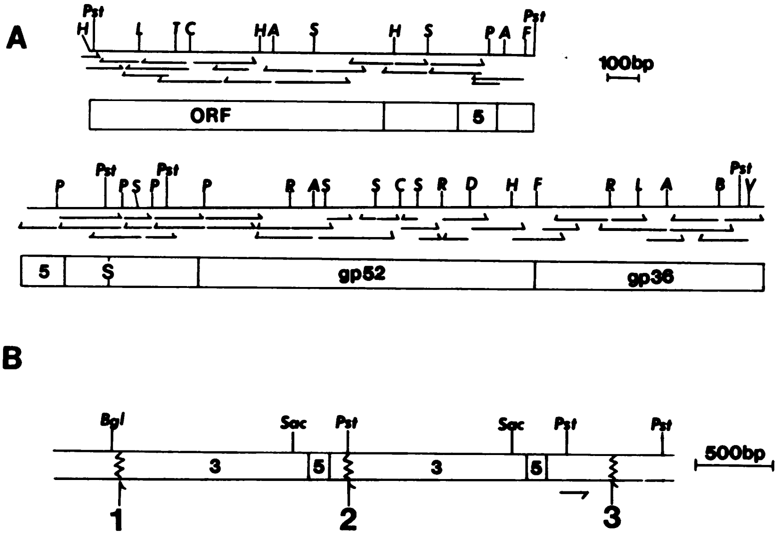

FIG. 2. (A) Strategy of sequencing line 8 proviral DNA. Sequenced regions are shown as arrows pointing away from the site of labeling. The upper segment of panel A shows the 5′ host-viral junction, with the open reading frame (ORF) and U5 domains (5); the lower segment shows a region of the line 8 provirus including U5 (5) and the coding domains for gp52 and gp36. The splice junction is denoted S. Sites indicated are: *H*, *Hae*III; *L*, *Alu*I; *T*, *Taq*I; *C*, *Clu*I; *A*, *Ava*II; *S*, *Sau*3a; *P*, *Hpa*II; *F*, *Hinf*; *R*, *Eco*RII; *D*, *Dde*(I); *B*, *Bgl*II; and *V*, *Ava*I. Only sites used in sequencing are shown; this is not a restriction site map. (B) Substrate for determining the splice donor site. The donor sequence was derived from a clone of unintegrated MMTV C3H circular DNA with rearrangements at the sites marked by arrows. The clone was an analog of a two-copy circle. Rearrangement (1) is a small deletion around the *Pst*I site at the left end of the LTR. Rearrangement (2) can be viewed as an aberrant circle junction. Rearrangement (3) is a small deletion, approximately 300 bp, in the 0.9-kb *Pst*I fragment which probably removes the poison sequence. The acceptor site sequence was derived from a 4-kb *Pst*I fragment cloned directly into plasmid pBR322 from unintegrated C3H circular DNA (data not shown).

```
|        R       | Begin U5
GCAACAGUUCCUAACAUUCACCUCUUGUGUGUUUGUGUCUGUUCGCCAUCCCGUCUCCGCUCGUCACUUAUCCUUCACUU   80

                                                  End U5    | Primer Binding Site
UCCUGCGGGUCCCCCGCAGACCCCGGCGACCUCAGGUCGGCCGACUGCGGCAGCUGGCGCCCGAACAGGGACCCCUCGG   160
       Ava II                                                           AvaII

AUAAGUGACCCUUGUCUCUAUUUCUACUAUUUGGUGUUUGUCUUGUAUUGUCUCUUUCUUGUCUUUCUAUCAUCACAAGA   240

                                                Splice Site
                                                    ▼ met pro asn his gln ser gly
GCGGAACGGACUCACCAUAGGGAGCUGCAGUCCCGCCUACGGAGAAGAGG AUG CCG AAU CAC CAA UCU GGG   311
                          PstI

    -90                                     -80
ser pro thr gly ser ser asp leu leu leu asp gly lys lys gln arg ala his leu ala
UCC CCG ACC GGU UCA UCC GAC CUU UUA CUA GAC GGA AAG AAG CAA CGC GCA CAC CUG GCA   371

    -70                                     -60
leu arg arg lys arg arg arg glu met arg lys ile asn arg lys val arg arg met asn
CUG CGG AGA AAA CGC CGC CGC GAG AUG AGA AAG AUC AAC AGG AAA GUC CGG AGG AUG AAU   431

    -50                                     -40
leu ala pro ile lys glu lys thr ala trp gln his leu gln ala leu ile phe glu ala
CUA GCC CCC AUC AAA GAG AAG ACG GCU UGG CAA CAU CUG CAG GCG UUA AUC UUC GAA GCG   491
                                                  Pstl

    -30                                     -20
glu glu val leu lys thr ser gln thr pro gln thr ser leu thr leu phe leu ala leu
GAG GAG GUU CUU AAA ACC UCA CAA ACU CCC CAA ACC UCU UUG ACU UUA UUU CUU GCU UUG   551
                                             A

    -10                                 -1 ┌ 1── Begin gp52
leu ser val leu gly pro pro pro val ser gly│glu ser tyr trp ala tyr leu pro lys
UUG UCU GUC CUG GGC CCC CCG CCU GUG UCC GGG GAA AGU UAU UGG GCU UAC CUA CCU AAA   611
             C

 10                                      20
pro pro ile leu his pro val gly trp gly asn thr asp pro ile arg val leu thr asn
CCA CCU AUU CUC CAU CCC GUG GGA UGG GGA AAU ACA GAC CCC AUU AGA GUU CUG ACC AAU   671

 30                                      40
gln thr ile tyr leu gly gly ser pro asp phe his gly phe arg asn met ser gly asn
CAA ACC AUA UAU UUG GGU GGG UCG CCU GAC UUU CAC GGG UUU AGA AAC AUG UCU GGC AAU   731

 50                                      60
val his phe glu glu lys ser asp thr leu pro ile cys phe ser phe ser phe ser thr
GUA CAU UUU GAG GAG AAG UCU GAU ACG CUC CCC AUU UGC UUU UCC UUC UCC UUU UCU ACC   791
             G

 70                                      80
pro thr gly cys phe gln val asp lys gln val phe leu ser asp thr pro thr val asp
CCC ACA GGC UGC UUU CAA GUA GAU AAG CAA GUA UUU CUU UCU GAU ACA CCC ACG GUU GAU   851
     G                                Δ                          Δ   Δ

 90                                     100
asn asn lys pro gly gly lys gly asp lys arg arg met trp glu leu trp leu thr thr
AAU AAU AAA CCU GGG GGA AAG GGU GAU AAA AGG CGU AUG UGG GAA CUU UGG UUG ACU ACU   911
                                                                        CA

110                                     120
leu gly asn ser gly ala asn thr lys leu val pro ile lys lys lys leu pro pro lys
UUG GGG AAC UCA GGG GCC AAU ACA AAA CUG GUC CCU AUA AAA AAG AAG UUG CCC CCC AAA   971
                                        AvaII

130                                     140
tyr pro his cys gln ile ala phe lys lys asp ala phe trp glu gly asp glu ser ala
UAU CCU CAC UGC CAG AUC GCC UUU AAG AAG GAC GCC UUC UGG GAG GGA GAC GAG UCU GCU   1031

150                                     160
pro pro arg trp leu pro cys ala phe pro asp gln gly val ser phe ser pro lys gly
CCU CCA CGG UGG UUG CCU UGC GCC UUC CCU GAC CAA GGG GUG AGU UUU UCU CCA AAA GGG   1091
                                             Δ      G

170                                     180
ala leu gly leu leu trp asp phe ser leu pro ser pro ser val asp gln ser asp gln
GCC CUU GGG UUA CUU UGG GAU UUC UCC CUU CCC UCG CCU AGU GUA GAU CAG UCA GAU CAG   1151

190                                     200
ile lys ser lys lys asp leu phe gly asn tyr thr pro pro val asn lys glu val his
AUU AAA AGC AAA AAG GAU CUA UUU GGA AAU UAU ACU CCC CCA GUC AAU AAA GAG GUU CAU   1211
                                                         U
```

FIG. 3

```
210                                     220
arg trp tyr glu ala gly trp val glu pro thr trp phe trp glu asn ser pro lys asp
CGA UGG UAU GAA GCA GGA UGG GUA GAA CCU ACA UGG UUC UGG GAA AAU UCU CCU AAG GAU   1271
ClaI                                                                      BamHI
230                                     240
pro asn asp arg asp phe thr ala leu val pro his thr glu leu phe arg leu val ala
CCC AAU GAU AGA GAU UUU ACU GCU CUA GUU CCC CAU ACA GAA UUG UUU CGC UUA GUU GCA   1331

250                                     260
ala ser arg tyr leu ile leu lys arg pro gly phe gln glu his glu met ile pro thr
GCC UCA AGA UAU CUU AUU CUC AAA AGG CCA GGA UUU CAG GAA CAU GAG AUG AUU CCU ACA   1391
            C                                                 A
270                                     280
ser ala cys val thr typ pro tyr val ile leu leu gly leu pro hln leu ile asp ile
UCU GCC UGU GUU ACU UAC CCU UAU GUC AUA UUA UUA GGA UUA CCU CAG CUA AUA GAU AUA   1451
                                 C
290                                     300
glu lys arg gly ser thr phe his ile ser cys ser ser cys arg leu thr asn cys leu
GAG AAA AGA GGA UCU ACU UUU CAU AUU UCC UGU UCU UCU UGU AGA UUG ACU AAU UGU UUA   1511

310                                     320
asp ser ser ala tyr asp thr ala ala ile ile val lys arg pro pro tyr val leu leu
GAU UCU UCU GCC UAC GAC UAU GCA GCG AUC AUA GUC AAG AGG CCG CCA UAC GUG CUG CUA   1571

330                                     340
pro val asp ile gly asp glu pro trp phe asp asp ser ala ile gln thr phe arg tyr
CCU GUA GAU AUU GGU GAU GAA CCA UGG UUU GAU GAU UCU GCC AUU CAA ACC UUU AGG UAU   1631

350                                     360         Begin gp36
ala thr asp leu ile arg ala lys arg phe val ala ala ile ile leu gly ile ser ala
GCC ACA GAU UUA AUU CGA GCC AAG CGA UUC GUC GCU GCC AUU AUU CUG GGC AUA UCU GCU   1691
                          U
370                                     380
leu ile ala ile ile thr ser phe ala val ala thr thr ala leu val lys glu met gln
UUA AUU GCU AUU AUC ACU UCC UUU GCU GUA GCU ACU ACU GCU UUA GUU AAG GAG AUG CAA   1751
                                                      C
390                                     400
thr ala thr phe val asn asn leu his arg asn val thr leu ala leu ser glu gln arg
ACU GCU ACG UUU GUU AAU AAU CUU CAU AGG AAU GUU ACA UUA GCC UUA UCU GAA CAA AGA   1811
                                      A                           C
410                                     420
ile ile asp leu lys leu glu ala arg leu asn ala leu glu glu val val leu asp leu
AUA AUA GAU UUA AAA UUA GAA GCU AGA CUU AAU GCU UUA GAA GAA GUA GUU UUA GAU UUG   1871

430                                     440
gly gln asp val ala asn leu lys thr arg met ser thr arg cys his ala asn tyr asp
GGA CAA GAU GUG GCA AAC UUA AAG ACC AGA AUG UCC ACC AGG UGU CAU GCA AAU UAU GAU   1931
                                  U                                           C
450                                     460
phe ile cys val thr pro leu pro tyr asn ala ser glu ser trp glu arg thr lys ala
UUU AUC UGC GUU ACA CCU UUA CCA UAU AAU GCU UCU GAG AGC UGG GAA AGA ACC AAA GCU   1991
                                            A   AG                       GG   C
470                                     480
his leu leu gly ile trp asn asp asn glu ile ser tyr asn ile gln glu leu thr asn
CAU UUA UUG GGC AUU UGG AAU GAC AAU GAG AUU UCA UAU AAC AUA CAA GAA UUA ACC AAC   2051
          A                   U                                           U
490                                     500
leu ile gly asp met ser lys gln his ile asp thr val asp leu ser gly leu ala gln
CUG AUU GGU GAU AUG AGC AAA CAA CAU AUU GAC ACA GUG GAC CUC AGU GGC UUG GCU CAG   2111
  A     A                         C         G        AvaII
510                                     520
ser phe ala asn gly val lys ala leu asn pro leu asp trp thr gln tyr phe ile phe
UCC UUU GCC AAU GGA GUG AAG GCU UUA AAU CCA UUA GAU UGG ACA CAA UAU UUC AUU UUU   2171
                                                          U
530                                     540
ile gly val gly ala leu leu leu val ile val leu met ile phe pro ile val phe gln
AUA GGU GUU GGA GCC CUG CUU UUA GUC AUA GUG CUU AUG AUU UUC CCC AUU GUU UUC CAG   2231
                                          A
550                                     560
cys leu ala lys ser leu asp gln val gln ser asp leu asn val leu leu leu
UGC CUU GCG AAG AGC CUU GAC CAA GUG CAG UCA GAU CUU AAC GUG CUU CUU UUA           2285
                                            BglII
```

FIG. 3—*Continued*

```
                     -1----  Begin O.R.F.       10           End gp36 ----      20
        570         [MetProArgLeuGlnGlnLysTrpLeuAsnSerArgGluCysProThrPro]ArgGlyGluAla
LysLysLysLysGlyGlyAsnAlaAlaProAlaAlaGluMetValGluLeuProArgValSerTyrThr***
AAAAAGAAAAAAGGGGGAAAUGCCGCGCCUGCAGCAGAAAUGGUUGAACUCCCGAGAGUGUCCUACACCUAGGGGAGAAGCA  2367
Poly-purine tract  | Begin U3    Pstl                 Aval       U

                                 30                                  40
ala lys gly leu phe pro thr lys asp asp pro ser ala his lys arg val ser pro ser
GCC AAG GGG UUG UUU CCC ACC AAG GAC GAC CCG UCU GCG CAC AAA CGG GUG AGC CCA UCA   2427

                                 50                                  60
asp lys asp ile phe ile leu cys cys lys leu gly ile ala leu leu cys leu gly leu
GAC AAA GAC AUA UUC AUU CUC UGC UGC AAA CUU GGC AUA GCU CUG CUU UGC CUG GGG CUA   2487
                C*                                        A           A
                                 70                                  80
leu gly glu val ala val arg ala arg arg ala leu thr leu asp ser phe asn ser ser
UUG GGG GAA GUU GCG GUU CGU GCU CGC AGG GCU CUC ACC CUU GAC UCU UUU AAU AGC UCU   2547
                                                    A C   U      G CAA   A
                                 90                                  100
ser val gln asp tyr asn leu asn asn ser glu asn ser thr phe leu leu arg gln gly
UCU GUG CAA GAU UAC AAU CUA AAC AAU UCG GAG AAC UCG ACC UUC CUC CUG AGG CAA GGA   2607
                              G   G                             U*  G*      AvaII
                                 110                                 120
pro gln pro thr ser ser tyr lys pro his arg phe cys pro ser glu ile glu ile arg
CCA CAG CCA ACU UCC UCU UAC AAG CCG CAU CGA UUU UGU CCU UCA GAA AUA GAA AUA AGA   2667
                                        Cla1 A              U
                                 130                                 140
met leu ala lys asn tyr ile phe thr asn lys thr asn pro ile gly arg leu leu val
AUG CUU GCU AAA AAU UAU AUU UUU ACC AAU AAG ACC AAU CCA AUA GGU AGA UUA UUA GUU   2727
                  G                                             C           C
                  A              150                                 160
thr met leu arg asn glu ser leu ser phe ser thr ile phe thr gln ile gln lys leu
ACU AUG UUA AGA AAU GAA UCA UUA UCU UUU AGU ACU AUU UUU ACU CAA AUU CAG AAG UUA   2787

                                 170                                 180
glu met gly ile glu asn arg lys arg arg ser thr ser ile glu glu gln val gln gly
GAA AUG GGA AUA GAA AAU AGA AAG AGA CGC UCA ACC UCA AUU GAA GAA CAG GUG CAA GGA   2847
                      G                     AA      G        G       A
                                 190                                 200
leu leu thr thr gly leu glu val lys lys gly lys lys ser val phe val lys ile gly
CUA UUG ACC ACA GGC CUA GAA GUA AAA AAG GGA AAA AAG AGU GUU UUU GUC AAA AUA GGA   2907
        G   U                                            G
                                 210                                 220
asp arg trp trp gln leu gly thr tyr arg gly pro tyr ile tyr arg pro thr asp ala
GAC AGG UGG UGG CAA CUA GGG ACU UAU AGG GGA CCU UAC AUC UAC AGA CCA ACA GAU GCC   2967
                      C                  AvaII
                                 230                                 240
pro leu pro tyr thr gly arg tyr asp leu asn trp asp arg trp val thr val asn gly
CCC UUA CCA UAU ACA GGA AGA UAU GAC UUA AAU UGG GAU AGG UGG GUU ACA GUC AAU GGC   3027
                                                             C     A   C
                                 250                                 260
tyr lys val leu tyr arg ser leu pro phe arg gly arg leu ala arg ala arg pro pro
UAU AAA GUG UUA UAU AGA UCC CUC CCU UUU CGU GAA AGA CUC GCC AGA GCU AGA CCU CCU   3087
                  C
                                 270                                 280
trp cys met leu ser gln glu glu lys asp asp met lys gln gln val his asp tyr ile
UGG UGU AUG UUG UCU CAA GAA GAA AAA GAC GAC AUG AAA CAA CAG GUA CAU GAU UAU AUU   3147
            C A A   G G A    G
                                 290                                 300
tyr leu gly thr gly met his phe trp gly lys ile phe his thr lys glu gly thr val
UAU CUA GGA ACA GGA AUG CAC UUU UGG GGA AAG AUU UUC CAU ACC AAG GAG GGG ACA GUG   3207
                                            G                    A     G
                                 310                       End O.R.F. ---
ala gly leu ile glu his tyr ser pro lys thr tyr gly met ser tyr tyr glu|***
GCU GGA CUA AUA GAA CAU UAU UCU CCA AAA ACU UAU GGC AUG AGU UAU UAU GAA UAG CCU   3267
        G                                       U                   U

UUAUUGGCCCAACCUUGCGGUUCCCAGGGCUUAAGUAAGUUUUUGGUUACAAACUGUUCUUAAAACGAGGAUGUGAGACA  3347
                                 U   A       A

AGUGGUUUCCUGACUUGGUUUGGUAUCAAAGGUUCUGAUCUGAGCUCUGAGUGUUCUAUUUUCCUAUGUUCUUUUGGAAU  3427
                              G           Sacl  U*
                                           A
                                           "TATA"               polyA  |    R        |
UUAUCCAAAUCUUAUGUAAAUGCUUAUGUAAACCAAGAUAUAAAAGAGUGCUGAUUUUUUUGAGUAAACUUGCAACAGUCCUAACA
CC        G*                        (UA)*         A
```

FIG. 3—*Continued*

amino acids were generally hydrophilic (extending to the first AUG codon). Because each of the initiation sites results in a signal sequence which has the typical signatures, although longer than normal, we were left with no reason for choosing one over the others.

Within the open reading frame we found two sequences in addition to the amino terminus of gp52^{env} which demonstrated it to be the correct one for the MMTV *env* gene. The DNA sequence predicts that the carboxyl terminal amino acid sequence of the primary gene product is Arg-Val-Ser-Tyr-Thr. S. Oroszlan (personal communication) has shown directly that the carboxyl terminal amino acid sequence of gp36^{env} is Arg-Val-Ser-Tyr/Thr-Thr/Tyr, confirming the DNA sequence and demonstrating that the carboxyl terminus is unprocessed. Oroszlan (personal communication) has also determined the NH_2 terminus of gp36 to be Phe-Val-Ala-Ala. We found the sequence encoding this oligopeptide at position 1669 in the nucleotide sequence, preceded by codons for Lys and Arg. The gp36 domain of *env* encodes 232 amino acids with a molecular weight of 25,500, and the gp52 domain encodes 357 amino acids with a molecular weight of 41,000. Several additional features of the *env* gene products are worth noting.

(i) gp36^{env} is thought to serve as a membrane anchor (4). Consistent with this function we found an extremely hydrophobic domain (amino acids 523 through 548), extending to within 24 residues of the carboxyl terminus of the protein, that may be involved in the anchor function.

(ii) The deduced amino acid sequence of the gp52:gp36 cleavage site is Lys-Arg:Phe-Val. The equivalent sites in the *env* genes of the Prague C strain of Rous sarcoma virus and the Moloney strain of murine leukemia virus are Lys-Arg:Ser-Val (25) and Lys-Arg:Glu-Pro (26), respectively, suggesting that the signal for cleavage includes the dipeptide Lys-Arg.

(iii) Both gp52 and gp36 are glycoproteins (4). An analysis of partially glycosylated *env* precursors suggests that the precursor gp73 contains five mannose-rich, asparagine-linked oligosaccharides (4).. Consistent with this, we find five Asn-X-Ser/Thr sites, three within gp52 sequences and two within gp36 sequences (Fig. 3).

(iv) Redmond and Dickson (22) recently determined the *env* gene sequence of the GR strain of MMTV. The existence of type-specific antigens which distinguish the gp52s and gp36s of the GR and C3H strains has been demonstrated (1, 4). Figure 3 shows points at which the *env* gene nucleotide sequences of the two strains diverge. We found 30 single base changes, some of which result in amino acid changes. Five of these are single base insertions and deletions, the most significant of which lie between amino acids 79 and 87 within the gp52 coding region. There, three extra base pairs in the MMTV C3H sequence result in divergence for a stretch of seven amino acids. This substitution may be sufficient to account for the type-specific differences between the gp52s of the two strains. The only other concerted change lies at amino acid positions 259 through 260, where a single base pair insertion followed 3 bp later by a single base pair deletion results in a single amino acid substitution.

(v) Finally, we observed that, unique among retroviruses, the termination site for the *env* gene lies within the LTR. Moreover, because MMTV is also unique in possessing a long open reading frame in its LTR, commencing close to the 5′ boundary (see below), the 3′ end of *env* overlaps another potential coding domain in another reading frame. Overlapping reading frames have been demonstrated in several retrovirus genomes, but MMTV is the only one for which the overlapping region includes the LTR. As a result, both the polypurine tract, which is presumably involved in (+) strand priming, and the 5′ terminus of the LTR, which is involved in integration, are included within the sequence coding for gp36.

Because the sequence encoding the hydrophobic anchor region precedes the LTR by 63 bp and lies upstream from the overlapping sequences, we speculate that that part of the gp36 amino acid sequence encoded by the LTR may not be of great functional significance.

FIG. 3. Deduced sequence of the MMTV C3H *env* mRNA. The bases are numbered with respect to the predicted start site of transcription. The translation product of *env* is numbered such that the first amino acid of mature gp52 is +1. The translation product of the LTR open reading frame is numbered from the first methionine residue. Several restriction enzyme cleavage sites and functional landmarks (primer binding site, splice sites, polypurine tract, and signals for transcriptional initiation [TATA] and polyadenylation) are included for convenience. Positions at which the MMTV C3H *env* sequence varies from that of MMTV GR (as determined by Redmond and Dickson [22]) are shown with the base from the GR sequence written below. Within the sequences derived from the LTR (R and U3), sites at which the MMTV GR and MMTV RIII sequences differ from the MMTV C3H sequence are similarly indicated; bases from the MMTV RIII sequence are in italics. Differences shared by MMTV GR and MMTV RIII are indicated by an asterisk. Deletions are indicated by Δ. The RIII sequence covers bases 2430 through 2830 and 3370 through 3513. At potential glycosylation sites, the Asn-X-Ser/Thr amino acid sequence is overlined.

(ii) Splice sites for *env* mRNA. Because the structure of the line 8 provirus suggested that it might be the product of DNA synthesis from a template of *env* mRNA, we pursued the possibility that we had been fortuitously provided with the necessary reagents for precisely determining a retroviral splice junction. This was accomplished by comparing the sequence from the provirus that lies between the 5′ LTR and the *env* coding domain with sequences upstream from *env* and downstream from an LTR in clones of unintegrated MMTV DNA. Like all of our clones of unintegrated circular DNA, the clone used to obtain the viral sequence downstream from the LTR exhibits a rearrangement (e.g., deletion) in the *gag* region (Fig. 2B) within the 0.9-kilobase (kb) *Pst*I fragment. We mapped and sequenced the 5′ end of this aberrant fragment and will argue shortly that the abnormality in the cloned DNA lies beyond the region pertinent to the present work.

A comparison of the relevant sequences is presented in Fig. 4. The sequences on the 3′ side of the 5′ LTR in line 8 DNA and in the unintegrated DNA are identical for 154 nucleotides; at the point of divergence, the unintegrated DNA exhibits a sequence closely related to the consensus sequence for a splice donor site (21). At the other end of the genome, the wild-type and line 8 proviral sequences diverge 1 bp on the 5′ side of the first of the three ATG codons that could serve as initiation sites for translation of *env* (see above; Fig. 4). The wild-type sequence 5′ to the point of divergence also shows a strong similarity to a consensus sequence for a splice acceptor site (21). Notably, the domain absent from the provirus does not contain direct repeats at its boundaries, suggesting that the region was probably not removed by homologous recombination between DNA sequences. Instead, it is likely that the sequences were removed by splicing of a primary transcript of genomic-sized RNA, forming an *env* mRNA that was then reverse transcribed during the round of infection that established the line. Evidence of a different sort for the synthesis of proviral DNA from *env* mRNA has been published by Stacey (27), who showed that Rous-associated virus *env* mRNA microinjected into cells infected with the *env*-deficient Bryan strain of Rous sarcoma virus not only would complement that deficiency but also could be packaged into virion particles, which were able to infect new cells and incorporate their genetic information into those cells in the form of an *env* provirus.

S1 nuclease mapping experiments, performed with labeled DNA from wild-type and line 8 proviral DNA, are consistent with the conclusion that the provirus is a copy of an *env* mRNA. DNA from a restriction fragment from the provirus which spans the putative splice site is fully protected by 24S *env* mRNA from cells bearing wild-type proviral DNA (D. Robertson, personal communication). Mapping with wild-type DNA annealed to *env* mRNA indicates a splice acceptor site approximately 180 bp on the 5′ side of the *Pst*I site near the 5′ end of *env* (D. Ucker, Ph.D. thesis, University of California, San Francisco; D. Robertson, personal communication). We conclude that the proviral DNA is a reverse transcript of a spliced *env* mRNA and that the splice sites conform to conventions established for eucaryotic genes, as expected from the use of host mechanisms for processing. Other retroviral splice sites estimated by the S1 mapping procedure also map near consensus sequences for donor or acceptor sites (12, 28).

Further inspection of the sequences adjacent to the splice site reveals information about the leader sequence of *env* mRNA and about the possible start site for *gag*. Transcription of MMTV RNA is thought to begin approximately 135 nucleotides from the 3′ end of the LTR, ca. 25 nucleotides downstream from the TATAAA sequence at positions 3466 through 3472 in the LTR (Fig. 3). The first 10 to 15 nucleotides in this sequence (R) are present at both ends of virion RNA (14); R is presumably present at the

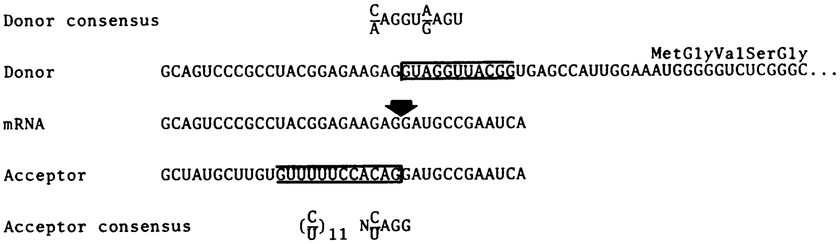
Donor consensus $\frac{C}{A}$AGGU$\frac{A}{G}$AGU
MetGlyValSerGly
Donor GCAGUCCCGCCUACGGAGAAGAGGUAGGUUACGGUGAGCCAUUGGAAAUGGGGGUCUCGGGC...
mRNA GCAGUCCCGCCUACGGAGAAGAGGAUGCCGAAUCA
Acceptor GCUAUGCUUGUGUUUUUCCACAGGAUGCCGAAUCA
Acceptor consensus $(\frac{C}{U})_{11}$ N$\frac{C}{U}$AGG

FIG. 4. Demonstration that the line 8 provirus is probably derived from a spliced *env* mRNA. The donor sequence was determined from the clone of unintegrated circular DNA described in Fig. 2B. The acceptor sequence was derived from a subclone of the 4-kb *Pst*I fragment which is also from unintegrated DNA. The consensus sequences are from Mount (21). Arrow, location of *env* mRNA splice.

ends of *env* mRNA as well, since the duplication is required for reverse transcriptiôn (29). Thus, 289 nucleotides from the 5′ end of the primary viral RNA transcript are found in *env* mRNA. This sequence is devoid of possible start codons, so the first AUG codon in *env* mRNA is 290 nucleotides from the 5′ end. As discussed earlier, it is uncertain whether this or one of the succeeding two AUG codons in the same reading frame initiates synthesis of the *env* polypeptide. The sequence of the fragment of unintegrated DNA containing the splice donor site is also free of translational start codons until position 314, where an AUG codon begins an open reading frame that extends for at least 50 nucleotides. Although this may represent the coding region for the amino terminus of $Pr77^{gag}$, the corresponding protein sequence has not been directly determined. Moreover, we cannot be certain that the nucleotide sequence is unaltered in this region, in view of the difficulties of generating clones from this region. In any case, the first AUG codon in the sequence of a retroviral mRNA is not necessarily an initiation site for translation; for example, the *gag* gene of Rous sarcoma virus is preceded by three unused AUG codons in viral RNA, and in *src* mRNA the gene is preceeded by four AUG codons, including the initiation codon from *gag* (12, 28).

(iii) Open reading frame in U3. We and others have previously identified sequences within and adjacent to the MMTV LTR likely to influence the initiation and polyadenylation of viral RNA and the priming and integration of viral DNA (8, 10, 13, 16). We have now determined the complete sequence of the 1,326-bp LTR from the 5′ LTR of our provirus (and selected regions of other MMTV LTRs) to facilitate the construction of deletion mutants used in studies of the hormonal responsiveness of MMTV DNA (to be reported elsewhere) and to examine an apparent paradox concerning the size of the open reading frame in the U3 domain.

The existence of a translatable region in or near the U3 sequence was first suggested by Dickson and Peters, who showed that fragmented virion RNA and RNA synthesized from the cloned 1.3-kb *Pst* D fragment (containing all but the 5′ 10 bp of the LTR) could direct synthesis in vitro of peptides 36, 24, 21, and 18 kd in size (5, 6). The MMTV C3H LTR sequence published by Donehower et al. (7) provided an open reading frame capable of encoding the 24-, 21-, and 18-kd proteins, but not the 36-kd polypeptide. By the addition of two extra bases at positions 2478 and 2486, our sequence extends the open reading frame to the left end of the LTR and allows the expression of the 36-kd protein. These additional bases were also found by Donehower et al. on reanalysis of their sequence data (7). The open reading frame thus begins with an AUG codon one base from the 5′ end of the LTR and extends for 319 codons. An open reading frame of similar or identical size has also been found in the LTR of MMTV GR (11) and endogenous provirus GR40 or unit II (7, 13).

The significance of the open reading frame in the MMTV U3 region has been widely discussed, but to date no protein products from this region have been encountered in infected cells. A few additional features of this putative viral gene should be mentioned here. First, since the candidate start codon begins with the second base in the LTR, the start codon is likely to be missing from the 5′ LTR due to the loss of 2 bp from the end of each LTR during integration (16). Thus, synthesis of a protein longer than 24 kd from the 5′ LTR would require an initiation codon in the flanking cellular sequence (as noted by Kennedy et al. [13]) in addition to signals for transcription of the 5′ LTR. A resolution to the question of the function of this open reading frame is promised by recent reports of viral RNAs that would appear to be appropriate mRNAs for its expression. Inspection of the nucleotide sequence immediately upstream from the start of the open reading frame reveals candidate splice acceptor sites. van Ooyen et al. (28a) and Wheeler et al. (30) have found in normal mammary tissue, from some but not all mouse strains, a 1.4-kb RNA species whose structure is consistent with a spliced RNA that employs one of these sites and would allow expression of the LTR open reading frame. However, this RNA is apparently transcribed from an unspecified endogenous MMTV provirus, and its function remains a mystery. Viral replication presumably requires only the three genes, *gag*, *pol*, and *env*, shared with other replication-competent retroviruses lacking unassigned open reading frames. The open reading frame cannot be required for the steroid responsiveness of MMTV, since LTRs from which the entire reading frame has been deleted are still competent to mediate steroidally regulated transcription (manuscript in preparation).

ACKNOWLEDGMENTS

We thank R. Nusse, G. Hager, L. Donehower, C. Dickson, and S. Oroszlan for communicating results before publication and Bertha Cook for secretarial assistance.

This work was supported by grants from the American Cancer Society and the National Institutes of Health.

LITERATURE CITED

1. **Arthur, L. O., B. W. Altrock, and G. Schochetman.** 1981. Type-specific determinants on proteins of an endogenous C3H mouse mammary tumor virus (MMTV) distinguish this virus from highly oncogenic exogenous MMTVs. Virology **110**:270–280.
2. **Arthur, L. O., T. D. Copeland, S. Oroszlan, and G.**

Schochetman. 1982. Processing and amino acid sequence analysis of the mouse mammary tumor virus *env* gene product. J. Virol. **41**:414–422.

3. **Dickson, C., and M. Atterwill.** 1980. Structure and processing of the mouse mammary tumor virus glycoprotein precursor Pr73^{env}. J. Virol. **35**:349–361.
4. **Dickson, C., R. Eisenman, H. Fan, E. Hunter, and N. Teich.** 1982. *In* R. Weiss, N. Teich, H. Varmus, and J. Coffin (ed.), Molecular biology of tumor viruses. Part III. RNA tumor viruses, p. 513–648. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
5. **Dickson, C., and G. Peters.** 1981. Protein-coding potential of mouse mammary tumor virus genome RNA as examined by in vitro translation. J. Virol. **37**:36–47.
6. **Dickson, C., R. Smith, and G. Peters.** 1981. *In vitro* synthesis of polypeptides encoded by the long terminal repeat region of mouse mammary tumor virus DNA. Nature (London) **291**:511–513.
7. **Donehower, L. A., B. Fleurdelys, and G. L. Hager.** 1983. Further evidence for the protein coding potential of the mouse mammary tumor virus long terminal repeat: nucleotide sequence of an endogenous proviral long terminal repeat. J. Virol. **45**:941–949.
8. **Donehower, L. A., A. L. Huang, and G. L. Hager.** 1981. Regulatory and coding potential of the mouse mammary tumor virus long terminal redundancy. J. Virol. **37**:226–238.
9. **Dudley, J. P., and H. E. Varmus.** 1981. Purification and translation of mouse mammary tumor virus mRNA's. J. Virol. **39**:207–218.
10. **Emr, S. D., and T. J. Silhavy.** 1982. Molecular components of the signal sequence that function in the initiation of protein export. J. Cell Biol. **95**:689–696.
11. **Fasel, N., K. Pearson, E. Buetti, and H. Diggelmann.** 1982. The region of mouse mammary tumor virus DNA containing the long terminal repeat includes a long coding sequence and signals for hormonally regulated transcription. EMBO J. **1**:3–7.
12. **Hackett, P. B., R. Swanstrom, H. E. Varmus, and J. M. Bishop.** 1982. The leader sequence of the subgenomic mRNA's of Rous sarcoma virus is approximately 390 nucleotides. J. Virol. **41**:527–534.
13. **Kennedy, N., B. Knedlitschek, N. E. Groner, P. Hynes, K. Herrlich, R. Michalides, and A. J. van Ooyen.** 1982. Long terminal repeats of endogenous mouse mammary tumor virus contain a long open reading frame which extends into adjacent sequences. Nature (London) **295**:622–624.
14. **Klememz, R., M. Reinhardt, and H. Diggelmann.** 1981. Sequence determination of the 3′ end of mouse mammary tumor virus RNA. Mol. Biol. Rep. **7**:123–126.
15. **Kozak, M.** 1981. Possible roles of flanking nucleotides in recognition of the AUG initiator codon by eukaryotic ribosomes. Nucleic Acids Res. **9**:5233–5252.
16. **Majors, J., and H. E. Varmus.** 1980. Learning about the replication of retroviruses from a single cloned provirus of mouse mammary tumor virus. ICN-UCLA Symp. Mol. Cell. Biol. **18**:241–253.
17. **Majors, J., and H. E. Varmus.** 1981. Nucleotide sequences at host-proviral junctions for mouse mammary tumor virus. Nature (London) **289**:253–258.
18. **Maniatis, T., E. F. Fritsch, and J. Sambrook.** 1982. Molecular cloning—a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
19. **Maxam, A., and W. Gilbert.** 1977. A new method for sequencing DNA. Proc. Natl. Acad. Sci. U.S.A. **74**:560–564.
20. **Maxam, A., and W. Gilbert.** 1980. Sequencing end-labeled DNA with base specific chemical cleavages. Methods Enzymol. **65**:497–559.
21. **Mount, S. M.** 1982. A catalogue of splice junction sequences. Nucleic Acids Res. **10**:459–472.
22. **Redmond, S. M. S., and C. Dickson.** 1983. Sequence and expression of the mouse mammary tumor virus *env* gene. EMBO J. **2**:125–131.
23. **Ringold, C. M., K. R. Yamamoto, J. M. Bishop, and H. E. Varmus.** 1977. Glucocorticoid-stimulated accumulation of mouse mammary tumor virus RNA: increased rate of synthesis of viral RNA. Proc. Natl. Acad. Sci. U.S.A. **74**:2879–2803.
24. **Schochetman, G. S., S. Oroszlan, L. Arthur, and D. Fine.** 1977. Gene order of the mouse mammary tumor virus glycoproteins. Virology **83**:72–83.
25. **Schwartz, D., R. Tizard, and W. Gilbert.** 1982. The complete nucleotide sequence of the Pr-C strain of Rous sarcoma virus, p. 1338–1348. *In* R. Weiss, N. Teich, H. E. Varmus, and J. Coffin (ed.), Molecular biology of tumor viruses. Part III. RNA tumor viruses. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
26. **Shinnick, T. M., R. A. Lerner, and J. G. Sutcliffe.** 1981. Nucleotide sequence of Moloney murine leukaemia virus. Nature (London) **293**:543–548.
27. **Stacey, D. W.** 1980. Expression of a subgenomic retroviral messenger RNA. Cell **21**:811–820.
28. **Swanstrom, R., H. E. Varmus, and J. M. Bishop.** 1982. Nucleotide sequence of the 5′ noncoding region and part of the *gag* gene of Rous sarcoma virus. J. Virol. **41**:535–541.

28a. **van Ooyen, A. J. J., R. J. A. M. Michalides, and R. Nusse.** 1983. Structural analysis of a 1.7-kilobase mouse mammary tumor virus-specific RNA. J. Virol. **46**:362–370.

29. **Varmus, H. E., and R. Swanstrom.** 1982. Replication of retroviruses in RNA tumor viruses, p. 369–512. *In* R. Weiss, N. Teich, H. E. Varmus, and J. Coffin (ed.), Molecular biology of tumor viruses. Part III. RNA tumor viruses. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
30. **Wheeler, D. A., J. S. Butel, D. Medina, R. D. Cardiff, and G. L. Hager.** 1983. Transcription of mouse mammary tumor virus: identification of a candidate mRNA for the long terminal repeat gene product. J. Virol. **46**:42–49.
31. **Young, H. H., T. Y. Shih, E. M. Scolnick, and W. P. Parks.** 1977. Steroid induction of mouse mammary tumor virus: effect upon synthesis and degradation of viral RNA. J. Virol. **21**:139–146.