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

LAWRENCE A. DONEHOWER, BARBARA FLEURDELYS, AND GORDON L. HAGER*
Laboratory of Tumor Virus Genetics, National Cancer Institute, Bethesda, Maryland 20205

Received 9 July 1981/Accepted 23 November 1982

The 3' half of an endogenous mouse mammary tumor virus from a C3H mouse was cloned in the Charon 4A vector phage. A comparison of the proviral clone with previously published endogenous mouse mammary tumor virus restriction maps identified it as endogenous unit II (J. Cohen and H. Varmus, *Nature [London]* **278**:418-423, 1979), which is present in all inbred mouse strains derived from the original Bagg albino $\times$ DBA cross. The nucleotide sequence of the unit II long terminal redundancy (LTR) was determined and compared with the sequence previously determined for the exogenous C3H virus LTR (Donehower et al., *J. Virol.* **37**:226-238, 1981). Virtually all sequence differences between the two LTRs were base substitutions. The total amount of sequence divergence was 6.6%. The large open reading frame reported previously in the exogenous LTR was preserved in the endogenous LTR. In addition, the pattern of sequence divergence was highly nonrandom with respect to the putative amino acid codons of the two open reading frames. Most of the base substitutions in this region resulted in silent or conservative amino acid codon changes. The nonrandom divergence pattern indicates that selective forces are operating on this segment of DNA and argues that the putative protein is functional in the life cycle of mouse mammary tumor virus. Possible roles for the protein and its mode of expression are discussed.

Mouse mammary tumor virus (MMTV) is a type B retrovirus that causes mammary carcinomas in susceptible mice (3, 26). This virus is transmitted horizontally to newborn mice through nursing (5). In addition, it has been shown that MMTV is transmitted genetically in both wild and inbred mice in the form of endogenous proviruses (25). Inbred mouse strains generally contain three or more copies of discrete proviruses (25). Inbred mouse strains generally contain three or more copies of discrete proviruses per haploid genome (11). Individual endogenous proviruses can be distinguished and identified with respect to their position within the host cell genome by restriction endonuclease mapping techniques (10, 11, 32). These techniques have been used to show that (i) wild mice display marked heterogeneity with respect to copy number and integration sites of endogenous MMTV (11), (ii) endogenous MMTV segregates as multiple, stable, independent genetic elements in inbred mouse strains (1, 20), and (iii) each endogenous MMTV unit is structurally distinguishable from other endogenous units on the basis of integration sites and internal se-

quence differences. For example, mice of the C3H/HeN inbred strain contain at least three endogenous MMTV units, which were identified as units I, II, and V by Cohen and Varmus (11) in their study of segregation patterns of endogenous MMTV in the inbred mouse strains derived from the original Bagg albino $\times$ DBA cross of 60 years ago (34). Unit II is the only complete endogenous provirus present in all strains descended from this cross.

Endogenous MMTV has been shown to be correlated with both early (in GR mice) and late (in C3Hf mice) occurring mammary carcinomas (2, 39). In addition, genetically transmitted viruses with oncogenic potential have been recovered from these strains, indicating that one or more of the endogenous units are expressed to produce a biologically active virus. Recent evidence indicates that mammary tumors caused by endogenous MMTV are linked to a single defined genetic locus (*Mtv-2* in the GR strain and *Mtv-1* in the C3Hf strain) (1, 23, 39, 40). *Mtv-1* appears to correspond to endogenous unit V (24). Other endogenous loci may not be expressed or, if they are expressed, may have little

or no oncogenic activity. This variability in expression or oncogenicity may arise from a number of factors, such as virus-specific sequence changes (37), proviral position within the genome (29), viral DNA modification and methylation (8), chromatin conformation at proviral sequences, and host cell genes and environment.

Changes in the primary sequence of a retrovirus may have drastic effects on its growth rate and level of expression, as well as its oncogenicity (7). Furthermore, comparisons of endogenous and exogenous retroviral sequences may provide insights into some of the mechanisms of retrovirus evolution. With this in mind, we have determined the nucleotide sequence of an endogenous MMTV long terminal redundancy (LTR) and compared it with the previously determined sequence of the exogenous MMTV LTR (17). The importance of the LTR in viral expression and integration has been noted by many investigators (4, 13, 33, 35, 36, 38, 41). In MMTV, the LTR has an added significance, in that it contains a large translational open reading frame, which may encode an MMTV nonstructural protein (14, 17, 31). The endogenous LTR analyzed in this study was derived from endogenous unit II of a C3H mouse. Although this endogenous provirus is present in a great number of inbred mouse strains, preliminary evidence suggests that it is not oncogenic in C3H mice (24).

In this report we compare the DNA sequence of the endogenous unit II MMTV LTR with that of the exogenous MMTV LTR. The sequence divergence between the two LTRs was 6.6% and was primarily the result of base substitutions. The nature of sequence changes within the open reading frame was highly nonrandom with respect to the amino acid codons. The codon changes primarily resulted in silent or conservative changes, providing a strong argument that an evolutionarily conserved protein is encoded within the MMTV LTR. We also speculate on the role of the LTR-encoded protein and suggest a model for its expression during MMTV infection.

MATERIALS AND METHODS

Mammalian cells. A murine mammary tumor cell line producing high levels of MMTV (C3H) was used. This cell line was derived from the original C3H MT cell line established by Owens and Hackett (27) from a mammary adenocarcinoma of a C3H mouse. The 34I cl 101 subclone has been described by Parks et al. (28). Restriction mapping of the proviral DNA of these cells indicated that they contained 20 to 30 copies of newly acquired exogenous C3H virus in addition to the three endogenous MMTV units present in the genomes of all C3H mice (L. A. Donehower, Ph.D. thesis, George Washington University, Washington, D.C., 1981).

Bacteria, bacteriophage, and plasmids. HV2-certi-

fied lambda phage vector Charon 4A was supplied by F. Blattner. HV2 *Escherichia coli* host strain DP50 (*supE supF m_k⁻ r_k⁻*) was obtained from D. Tiemeier. *E. coli* strain LE392 (*thyA supE supF m_k⁺ r_k⁺*) was supplied by L. Enquist. *E. coli* strain NS428 (N100 [λ Aam11 b2 red3 c1857 Sam7]) was kindly provided by Nat Sternberg, and *E. coli* strain λ dg805 (W3350 [λ dgal805 c1857 Sam7]) was received from F. Blattner. Plasmid pBR322, an HV2-certified vector, was obtained from Bethesda Research Laboratories. The HV2 *E. coli* host χ 1776 was generously supplied by Roy Curtiss.

Molecular cloning of MMTV proviruses in phages and plasmids. Molecular cloning experiments were performed as described previously (17).

Restriction mapping and hybridization analysis. Restriction endonucleases were obtained from Bethesda Research Laboratories, New England Biolabs, and Boehringer-Mannheim Biochemical and used according to the specifications of the manufacturers. Restriction mapping and Southern hybridizations were performed as described previously (17).

Preparation of DNA fragments for sequencing. DNA fragments from the unit II-containing plasmids were purified and prepared for sequencing as described previously (17).

DNA sequencing. DNA sequencing was performed as described by Maxam and Gilbert (22) without modification. DNA fragments were 5'-end-labeled by using previously described protocols (22). The sequencing reactions used were the G, G+A, C+T, C, and A+C reactions. Thin gels (0.4 mm) (30) 80 and 40 cm long were used in the gel electrophoresis steps.

RESULTS

Cloning and restriction mapping of endogenous unit II. Cellular DNA from 34i cells contained approximately 20 to 30 copies of newly integrated exogenous MMTV, as well as endogenous MMTV units, I, II, and V (Donehower, Ph.D thesis). In previously described molecular cloning experiments in which this DNA was used, we generated 86 MMTV-positive recombinants (16). Restriction mapping of an initial set of 10 recombinants (selected by maximum hybridization to MMTV cDNA_{rep}) indicated that 8 of these contained the 3' half of integrated exogenous MMTV, 1 contained full-length endogenous unit V MMTV, and 1 contained the 5' half of unit V MMTV (L. Johnson, unpublished data).

During subsequent restriction mapping of other MMTV recombinants, a clone (clone 598) corresponding to the 3' half of endogenous unit II was identified. This was determined by comparing the clone restriction map with a detailed restriction map of MMTV unit II published by Cohen et al. (9). Figure 1 shows this comparison, with the clone map aligned with the 3' half of the MMTV unit II restriction map derived from BALB/c mice. Not only do the MMTV-containing regions on the two maps have identical restriction sites, but the adjacent cellular

FIG. 1. Restriction map of endogenous unit II MMTV clone 598. The position of each restriction endonuclease site was determined as described previously (16). (A) Restriction map of endogenous MMTV unit II, as determined with Southern blotting techniques by Cohen et al. (12). (B) Restriction map derived from recombinant clone 598. Dotted lines are drawn between restriction sites in the two maps to emphasize the identity of the two DNA sequences. R, *EcoRI*; H *HpaI*; B *BglIII*; P, *PstI*. The serrated lines indicated Charon 4A DNA, and the straight lines indicate mouse cellular DNA. Kb, Kilobase.

sequences have virtually the same map, indicating that this clone represents the original endogenous proviral locus rather than an activated endogenous virus which has reintegrated in another place within the genome.

Preparation of unit II clone for DNA sequencing. Phage clone 598 containing the 3' half of endogenous unit II MMTV was restricted with *EcoRI*, and the 7-kilobase MMTV-containing *EcoRI* fragment (containing approximately 4 kilobases of MMTV information and 3 kilobases of 3' adjacent cellular DNA) was ligated into the *EcoRI* site of pBR322. MMTV-containing plasmids were selected by colony hybridization and amplified to obtain milligram quantities of DNA in preparation for sequencing. These plasmids were restricted with *HpaI*, which generated two fragments (1,000 and 1,450 base pairs) spanning the LTR region. These two fragments were purified from the larger plasmid fragment for sequencing by techniques described previously (17).

DNA sequencing of the MMTV endogenous unit II LTR. The strategy used for determining the DNA sequence of the unit II LTR is shown in Fig. 2. Approximately 83% of the endogenous LTR was sequenced twice either along the same strand or on opposite strands. The entire open reading frame region was sequenced at least twice. Ambiguities were further reduced by using five sequencing reactions (an additional A>C reaction) instead of the customary four and through the use of long (80-cm), thin (0.4-mm) gels (30). The nucleotide sequence of the MMTV endogenous unit II LTR is shown in Fig. 3.

DISCUSSION

Unit II MMTV. Unit II MMTV is a representative endogenous virus that is present in the genomes of many current inbred mouse strains. The structural similarity of this virus to the exogenous milk-borne form of MMTV is obvious from the restriction map similarities (32) and the fact that probes from the exogenous virus hybridize readily to various segments of unit II DNA. In collaboration with C. Garon, we have also found that the 3' halves of cloned exogenous and endogenous unit II DNAs form heteroduplexes in which the MMTV regions are perfectly homologous (at electron microscopic resolution), indicating that there are no large-scale sequence divergences between the two proviruses. These similarities indicate that the two proviruses are closely related evolutionarily.

Unit II MMTV has not been correlated with tumorigenicity in C3H mice and may remain in an unexpressed form in the host genome. This apparent lack of biological activity could be a result of viral sequence changes, as has been observed in one avian retrovirus system. In this system, viral growth and expression has been correlated with the U3 region, where the viral transcriptional promoter is located (6, 7). Sequence changes in the transcriptional control regions of the MMTV LTR of unit II might provide one possible explanation for the low oncogenic potential of this virus. Alternatively, sequence changes in viral structural gene regions might also result in reduced biological activity.

----->

-1350 -1340 -1330| -1320 -1310 -1300 -1290 -1280
TCCTTTAAAAAAGAAAAAGGGGAAATGCCGCGCTGCAGCAGAAATGGTTGAACCTCCGAGAGTGCCTACACTTAGG
MetProArgLeuGlnGlnLysTrpLeuAsnSerArgGluCysProThrLeuArg
-1270 -1260 -1250 -1240 -1230 -1220 -1210 -1200
AGAGAAGCAGCCAAGGGGTGTTCTCCACCAAGGACGACCCGTCTGCGTCAGCCGGATGAGCCCATCAGACAAAGACAT
ArgGluAlaAlaLysGluPheProThrLysAspAspProSerAlaCysThrArgMetSerProSerAspLysAspIle
-1190 -1180 -1170 -1160 -1150 -1140 -1130 -1120
ACTCATCTCTGCTGCAAACTGGGATCGACTCTGCTTCGCGGGGTATTGGGGAGAGTTGCGGGTCTGCTCGCAGAGG
LeuIleLeuCysCysLeuGlyIleAlaLeuLeuCysLeuGluGluGlyValAlaValAlaArgAlaArgArg
-1110 -1100 -1090 -1080 -1070 -1060 -1050 -1040
CTCTCACCCCTGATTCTTTAAATCACTCTCTGTCAGAATTAACTTAACAATTCGGAGAATCCGACCTTCTCCTCTG
AlaLeuThrLeuAspSerPheAsnAsnSerSerValGlnAspTyrAsnLeuAsnAsnSerGluuAsnSerThrPheLeuLeu
-1030 -1020 -1010 -1000 -990 -980 -970 -960
GGGCAAGGACCAAGCAACTTCTCTTCAAGCCACACCGACTTGTCTTCAGAAATAAGAAATGCTTGCTGAA
GlyGlnGlyProGlnProThrSerSerTyrLysProHisArgLeuCysProSerGluIleGluIleArgMetLeuAlaLys
-950 -940 -930 -920 -910 -900 -890 -880
AAATTATATTTTTACCAATAAGACCAATCCAATAGGTCGATTATTAATCATGATGTTAAGAAATGAATCTTGTCTTTTA
AsnTyrIlePheThrAsnLysThrAsnProIleGlyArgLeuLeuIleMetMetLeuArgAsnGluSerLeuSerPhe
-870 -860 -850 -840 -830 -820 -810 -800
GCACATATTTATCTCAAAATCAAAGGTAGAAATGGGAATGAAATAGAAAGAGCAGCGTCAACCTCAGTTGGAAGAACAG
SerThrIlePheThrGlnIleGlnArgLeuLeuMetGlyIleGluAsnArgLysArgArgSerThrSerValGluGluGln
-790 -780 -770 -760 -750 -740 -730 -720
GTGCAAGGACTAAGGGCGTACGGCCTAGAAGTAAAGGGAAGAGGAGTACCGCTGTGCAAAATAGGAGACAGGTGGTG
ValGlnGlyLeuArgAlaSerGlyLeuGluValLysArgGlyLysArgSerThrLeuValLysIleGlyAspArgTrpTrp
-710 -700 -690 -680 -670 -660 -650 -640
GCAACGAGGACTTATAGGGGACCTTACATCTACAGACCAACAGACGCCCGCTACCATATACAGGAAGATACGATTTAA
GlnProGlyThrTyrArgGlyProTyrIleTyrArgProThrAspAlaProLeuProTyrThrGlnIleArgTyrAspLeu
-630 -620 -610 -600 -590 -580 -570 -560
ATTTTGATAGGTGGGTCACAGTCAACGGCTATAAAGTGTGTGACAGATCCCTCCCTTTCGTGAAAGACTCGCCAGAGCT
AsnPheAspArgTyrValThrValAsnGlyTyrLysValIleTyrArgSerLeuProPheArgGluArgLeuAlaArgAla
-550 -540 -530 -520 -510 -500 -490 -480
AGACCTCTCTGGTGTGTGTTAACTCAGGAAGAAAAGACGACATAAACAAGGATACATGATTATTTATCTAGGAAG
ArgProProTyrCysValLeuThrGlnGluLysAspAspIleLeuGlnGlnValHisAspTyrIleTyrLeuGlyThr
-470 -460 -450 -440 -430 -420 -410 -400
TGGAATGAACGTTTGGGGAAGATTTTTTCATTATAACCAAGGAGGGGCGAGTGGCTAGACAAATAGAACACATTTCTGCAG
GlyMetAsnValTrpGlyLysIlePheHisTyrThrLysGluGlyAlaValAlaArgGlnLeuGluHisIleSerAla
-390 -380 -370 -360 -350 -340 -330 -320
ATACCTTTGGCATGAGCTATAATGGATAACCTTTATGAGCCAACCTTGGGTTTCCCAAGGTTTAAGTAAGTTCAGGGTCT
AspThrPheGlyMetSerTyrAsnGly
-310 -300 -290 -280 -270 -260 -250 -240
ACAACTGTTCTTAAACAGGATGTGAGACAAGTGGTTTCTGACTTGGTTGGTATCAAGTGTTTGATCTAAGCTCT
-230 -220 -210 -200 -190 -180 -170 -160
GAGTGTCTATTCTCTATGTTCTTTTGGAACTTATCCAAGTCTTATGTAATGCTTATGTAAACCATGATATAAAGAG
-150 -140 -130 -120 -110 -100 -90 -80
TGCTGATTTTTGAGTAAACTTGAACAGCTCTCAATCTCTGGTGTGTGTTGTGTCTGTGGCCATCCGCTCCCG
-70 -60 -50 -40 -30 -20 -10
CTCGTCACTTATCCTTCACTTTCAGAGGGTCCCCCGACAGCCCGGTGACCTCAGGTGACGGCACTGCGCGATTGTA

FIG. 3. Nucleotide sequence of the endogenous unit II MMTV LTR. The sequence is numbered in a negative direction from the putative virus-host joint. The six-base inverted repeats at the ends of the LTR are indicated by dashed arrows. The amino acid codons of the putative protein, pLTR, are indicated below the nucleotide sequence. Proposed transcriptional and translational regulatory signals are underlined. The boundaries of the LTR are indicated by vertical lines.

Molecular cloning, restriction mapping, and partial sequencing of endogenous MMTV unit II indicate that this MMTV has integrated at the same site with respect to the viral genome as exogenous MMTV (17). These data support the hypothesis that endogenous proviruses have been established by multiple independent infections of germ cells rather than by somatic mutation of ancestral proviruses or cellular genes (11). Clone restriction maps identical to previously published maps confirm that unit II has remained stable during the neoplastic state and consequent development and passage in cell cultures.

Comparison of the unit II endogenous LTR sequence with the exogenous LTR sequence. Upon further analysis, the previously published exogenous C3H-S MMTV LTR sequence (17)

has been shown to contain two additional bases within the open reading frame, corresponding in Fig. 3 to positions -1,148 (thymine) and -1,155 (cytosine). These nucleotides are also present in unit II and extend the open reading frames in both exogenous and endogenous sequences to the ATG codon at the 5' end (position -1,329) of the LTR.

Figure 4 shows a comparison of the exogenous and endogenous unit II LTR sequences. Only the sequence differences are shown. These differences are primarily single base substitutions and at first glance appear to be arrayed randomly throughout the LTR. Of the 1,367 nucleotides compared in the exogenous and endogenous MMTV sequences, there were 91 nucleotide differences, including 90 base substitutions and 1 three-base insertion in unit II. The

FIG. 4. Nucleotide sequence comparison between the exogenous and endogenous unit II LTRs. The standard numbering system of Fig. 3 is used. Each dash represents a base which is not changed between the two LTRs. Pairs of letters appear at the sites of difference between the two LTRs; the top letter of the pair indicates the exogenous base, and the bottom letter indicates the unit II base. The total number of changes for each 80-nucleotide segment is indicated on the right. The boundaries of the LTRs are indicated by vertical lines.

LTR sequence variation was 6.6%. The open reading frame has a greater percentage of differences (7.3%) than the noncoding regions of the LTR (5.2%). It is likely that at least some of the noncoding regions are highly conserved because they are involved in viral transcriptional regulation, replication (19), and integration. Comparison of endogenous and exogenous avian retrovirus LTRs has shown that the U5 region and promoter regions are relatively well conserved, whereas much of the remainder of the LTRs have diverged considerably (7, 21).

The transcriptional regulatory region of the endogenous LTR, which is centered around the TATAAAA box at position -165 (Fig. 3), appears to be very well conserved and to contain no significant base changes. Sequence-specific transcriptional modification is less likely to be the basis for the low oncogenicity of unit II MMTV. Post-integrational sequence changes in the 5' LTR promoter could have occurred, however.

Amino acid codon comparison between the endogenous and exogenous open reading frames. The codon changes generated by the nucleotide differences within the open reading frames of the two LTRs are shown in Table 1. The putative open reading frame-encoded protein, designated pLTR (also *orf* by Dickson and Peters [14], has been arbitrarily assigned as beginning at the first ATG initiation codon within the open reading frame at position -1,329 and as ending at the termination codon at position -370 (Fig. 3) in the endogenous sequence. This open reading frame encodes a protein of about 37,000 daltons. The open reading frame previously reported in the exogenous LTR (after correction for the two additional nucleotides) is preserved in the endogenous unit II LTR. No new terminator codons have been introduced by base changes, and no frameshifts which would close the reading frame have been observed. In fact, a 3-base pair insertion occurs in the endogenous open reading frame at position -444, which preserves the downstream amino acid codons in relation to the exogenous codons.

There are 61 codon changes within the pLTR coding sequence. Of these changes, 26 are silent, resulting in no change in the encoded amino acid. Based on the criteria assembled by Dayhoff (12), 21 codon changes are conservative, resulting in no significant changes in the biochemical properties of the encoded protein (e.g., arginine to lysine). Four codon changes are considered moderate, and 10 codon changes are considered radical, possibly resulting in significant changes in the biochemical properties of the protein. Thus, more than 75% of the amino acid codon changes are insignificant with respect to the properties of the putative protein.

Random changes in the nucleotide sequence of a protein coding region would be expected to generate the different categories of codon changes (silent, conservative, moderate, and radical) with certain probabilities (18). Doolittle (18) has calculated the probabilities of codon changes in randomly mutating coding sequences, and these are compared in Table 2 with the observed frequencies of codon changes in the two LTRs. It is obvious from Table 2 that the number of silent and conservative changes in the pLTR codons are significantly greater than might be expected by a random pattern of sequence divergence. A chi-square goodness-of-fit test indicated that the observed sequence changes could not be a result of random sequence divergence even at a level of significance of 0.005. This nonrandom pattern of sequence divergence in the pLTR coding region implies that there have been selective pressures acting on this segment of DNA during the recent evolution of MMTV. It further argues that the open reading frame encodes a biologically functional protein active at some stage in the life cycle of MMTV.

Other evidence for an LTR-encoded protein. Dickson and Peters (14) and Sen et al. (31) have shown by in vitro translation of MMTV genomic RNA that an open reading frame is present in the 3' region of the viral RNA. Translation of polyadenylic acid-containing partially degraded viral RNA generated proteins of 36,000, 24,000, 21,000, and 18,000 daltons, which did not correspond to any of the known MMTV structural proteins, as determined by tryptic mapping. These four proteins contained common tryptic peptides and were apparently closely related. In addition, insertion of cloned MMTV LTR DNA within the β -lactamase gene of plasmid pAT153 by Dickson and colleagues has resulted in in vitro synthesized MMTV derived proteins of 36,000, 24,000, 21,000, and 18,000 daltons, which correspond precisely to the proteins generated by in vitro translation of genomic MMTV RNA (15). Partial sequencing of some of the peptides from the 24,000-dalton protein indicates that the amino-terminal end of this protein corresponds precisely to the methionine codon initiating the putative pLTR discussed above (C. Dickson, personal communication). Analysis of the exogenous sequence indicates that the 24,000-, 21,000-, and 18,000-dalton proteins could easily have been generated by downstream translational initiation at methionine codons corresponding to positions -965, -903, and -843, respectively, in the endogenous sequence.

Further evidence that the LTR encodes a protein has been obtained in our laboratory by genetic engineering techniques. The putative pLTR gene was fused to the *E. coli lacZ* gene,

TABLE 1. Summary of changes in pLTR codons between exogenous and unit II LTRs

Codon change	Position	Exogenous codon	Endogenous codon	Position change(s)	Exogenous amino acid	Endogenous amino acid	Type of change ^a
1	-1,280	CCT	CTT	2	Pro	Leu	R
2	-1,275	GGA	AGA	1	Gly	Arg	R
3	-1,227	CAC	TGC	1,2	His	Cys	R
4	-1,223	AAA	ACG	2,3	Lys	Thr	M
5	-1,194	TTC	CTC	1	Phe	Leu	C
6	-1,102	GAC	GAT	3	Asp	Asp	—
7	-1,091	AGC	AAC	2	Ser	Asn	C
8	-1,035	AGG	GGG	1	Arg	Gly	R
9	-1,000	CCG	CCA	3	Pro	Pro	—
10	-997	CAT	CAC	3	His	His	—
11	-993	TTT	CTT	1	Phe	Leu	C
12	-918	AGA	CGA	1	Arg	Arg	—
13	-909	GTT	ATC	1,3	Val	Ile	C
14	-905	ACT	ATG	2,3	Thr	Met	M
15	-886	TCA	TCT	3	Ser	Ser	—
16	-883	TTA	TTG	3	Leu	Leu	—
17	-874	AGT	AGC	3	Ser	Ser	—
18	-868	ATT	ATA	3	Ile	Ile	—
19	-853	CAG	CAA	3	Gln	Gln	—
20	-851	AAG	AGG	2	Lys	Arg	C
21	-805	ATT	GTT	1	Ile	Val	C
22	-783	TTG	AGG	1,2	Leu	Arg	R
23	-780	ACC	GCC	1	Thr	Ala	C
24	-777	ACA	TCA	1	Thr	Ser	C
25	-758	AAG	AGG	2	Lys	Arg	C
26	-751	AAA	AAG	3	Lys	Lys	—
27	-748	AAG	AGG	2	Lys	Arg	C
28	-744	GTT	ACG	1-3	Val	Thr	C
29	-741	TTT	CTT	1	Phe	Leu	C
30	-670	GAT	GAC	3	Asp	Asp	—
31	-664	CCC	CCT	3	Pro	Pro	—
32	-663	TTA	CTA	1	Leu	Leu	—
33	-643	TAT	TAC	3	Tyr	Tyr	—
34	-640	GAC	GAT	3	Asp	Asp	—
35	-632	TGG	TTT	2,3	Trp	Phe	C
36	-619	GTT	GTC	3	Val	Val	—
37	-610	AAT	AAC	3	Asn	Asn	—
38	-595	TTA	TTG	3	Leu	Leu	—
39	-592	TAT	TAC	3	Tyr	Tyr	—
40	-580	CCT	CCC	3	Pro	Pro	—
41	-540	ATG	GTG	1	Met	Val	C
42	-535	TTG	TTA	3	Leu	Leu	—
43	-534	TCT	ACT	1	Ser	Thr	C
44	-529	CAA	CAG	3	Gln	Gln	—
45	-511	ATG	ATA	3	Met	Ile	C
46	-475	ACA	ACT	3	Thr	Thr	—
47	-468	CAC	AAC	1	His	Asn	C
48	-465	TTT	GTT	1	Phe	Val	R
49	-448	TTC	TTT	3	Phe	Phe	—
50	-444	—	TAT	Ins ^b	—	Tyr	M
51	-429	ACA	GCA	1	Thr	Ala	C
52	-420	GGA	AGA	1	Gly	Arg	R
53	-416	CTA	CAA	2	Leu	Gln	R
54	-414	ATA	TTA	1	Ile	Leu	C
55	-406	CAT	CAA	3	His	His	—
56	-405	TAT	ATT	1,2	Tyr	Ile	R
57	-396	AAA	CAC	1,3	Lys	Asp	M
58	-389	TAT	TTT	2	Tyr	Phe	C
59	-379	AGT	AGC	3	Ser	Ser	—
60	-375	TAT	AAT	1	Tyr	Asn	R
61	-371	GAA	GGA	2	Glu	Gly	C

^a —, Silent change; C, conservative change; M, moderate change; R, radical change.^b Ins, Insertion.

TABLE 2. Comparison of LTR codon changes with random coding sequence changes

Codon change ^a	Random expected frequency ^b	Observed LTR frequency ^c
Silent	0.25 (134/549) ^d	0.43 (26/61)
Conservative	0.24 (132/549)	0.34 (21/61)
Moderate	0.18 (100/549)	0.07 (4/61)
Radical	0.29 (160/549)	0.16 (10/61)
To terminator	0.04 (23/549)	0.00 (0/61)

^a Categories of codon changes are classified by the method of Dayhoff (12).

^b Fraction of codon changes expected in a randomly mutating protein coding sequence, assuming 50% guanine-cytosine nucleotide sequence (18).

^c Fraction of codon changes between LTR coding sequences.

^d There are 549 different codon changes possible in a nucleotide sequence coding for a protein (18).

and stable LTR-derived fusion proteins of the expected size were synthesized in *E. coli* (Donehower and Hager, manuscript in preparation). We are in the process of preparing antibodies to this LTR-derived protein to assay for its presence in MMTV-infected cells.

Speculation on the role and expression of an LTR-encoded protein in MMTV-infected cells. A biological role for the putative LTR-encoded protein in MMTV-infected cells can only be the subject of speculation at this time. It has been hypothesized that this region of the viral genome encodes the function involved in mammary carcinogenesis (3). Since MMTV is unique among retroviruses in being transcriptionally regulated by glucocorticoids, an alternative hypothesis is that an additional MMTV gene may be involved in the hormonal regulatory response. However, experiments in progress (M. Kessel et al., unpublished data) indicate that hormone responsiveness is retained in fusions between selectable genes and LTR mutants in which the pLTR coding sequences have been deleted. Involvement in viral integration or transcriptional processing are other possible roles for the putative protein.

The existence of a gene in the two LTRs of a provirus raises the novel possibility of the expression of this gene from either or both ends of the viral genome. Expression of pLTR from the 3' LTR could be accomplished by splicing of the normal transcriptional leader into the 3' pLTR region or by transcription from its own promoter just upstream from the 3' LTR. Expression of pLTR from the 5' LTR would require transcriptional promotion from adjacent upstream cellular information. We have recently established (D. Wheeler et al., *J. Virol.*, in press) that the pLTR coding sequences can be expressed as a spliced message with the normal

viral leader in preneoplastic BALB/c mammary tissue.

The DNA sequence data and the in vitro translation experiments of other workers strongly support hypothesis that a novel additional gene is present in MMTV. However, demonstration of a functional protein product in vivo still is the critical step in confirming the existence of pLTR. Attempts are currently underway to detect this protein in MMTV-infected cells.

ACKNOWLEDGMENTS

We thank Clive Dickson, John Majors, and Heidi Diggelmann for generously communicating data before publication; Diana Berard, Ron Wolford, and Patricia Morgan for excellent technical assistance; Linda Kunz for growing plasmids; and David Wheeler and Philip Tschlis for helpful discussions.

ADDENDUM IN PROOF

After this manuscript was submitted, Kennedy et al. (*Nature [London]* 295:622-624, 1982) reported the sequence for the LTR of an MMTV provirus endogenous to GR mice. This provirus is apparently identical to the unit II locus of C3H mice, since the sequence of the LTR differs only at three positions from the sequence reported here, and the 3' cellular flanking sequences are approximately 80% homologous.

LITERATURE CITED

1. Bentvelzen, P. 1968. Genetic control of the vertical transmission of the Muhlbock mammary tumor virus in the GR mouse strain. Hollandia, Amsterdam.
2. Bentvelzen, P., and J. H. Daams. 1969. Hereditary infections with MTV in mice. *J. Natl. Cancer Inst.* 43:1025-1035.
3. Bentvelzen, P., and J. Hilgers. 1980. Murine mammary tumor virus, p. 311-355. *In* G. Klein (ed), *Viral oncology*. Raven Press, New York.
4. Benz, E. W., R. M. Wydro, B. Nadal-Ginard, and D. Dina. 1980. Moloney murine sarcoma proviral DNA is a transcriptional unit. *Nature (London)* 288:665-669.
5. Bittner, J. J. 1936. Some possible effects of nursing on the mammary gland tumor incidence in mice. *Science* 84:162-164.
6. Coffin, J. M., M. Champion, and F. Chabot. 1978. Nucleotide sequence relationships between the genomes of an endogenous and an exogenous avian tumor virus. *J. Virol.* 28:972-991.
7. Coffin, J. M., P. N. Tschlis, C. S. Barker, and S. Voynow. 1980. Variation in avian retrovirus genomes. *Ann. N. Y. Acad. Sci.* 354:410-425.
8. Cohen, J. C. 1980. Methylation of milk-borne and genetically transmitted mouse mammary tumor virus proviral DNA. *Cell* 19:653-662.
9. Cohen, J. C., J. E. Majors, and H. E. Varmus. 1979. Organization of mouse mammary tumor virus-specific DNA endogenous to BALB/c mice. *J. Virol.* 32:483-496.
10. Cohen, J. C., P. R. Shank, V. L. Morris, R. Cardiff, and H. E. Varmus. 1979. Integration of the DNA of mouse mammary tumor virus in virus infected normal and neoplastic tissue of the mouse. *Cell* 16:333-345.
11. Cohen, J. C., and H. E. Varmus. 1979. Endogenous mammary tumor virus DNA varies among wild mice and segregates during inbreeding. *Nature (London)* 278:418-423.
12. Dayhoff, M. O. 1969. Atlas of protein sequence and structure, vol. 4. National Biomedical Research Foundation, Silver Spring, Md.
13. Dhar, R., W. L. McClements, L. W. Enquist, and G. F. Vande Woude. 1980. Nucleotide sequences of integrated

- Moloney sarcoma provirus long terminal repeats and their host and viral junctions. *Proc. Natl. Acad. Sci. U.S.A.* 77:3937-3941.
14. 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.
 15. Dickson, C., R. Smith, and G. Peters. 1981. In vitro synthesis of polypeptides encoded by the long terminal repeat region of mouse mammary tumour virus DNA. *Nature (London)* 291:511-513.
 16. Donehower, L. A., J. Andre, D. S. Berard, R. G. Wolford, and G. L. Hager. 1980. Construction and characterization of molecular clones containing integrated mouse mammary tumor virus sequences. *Cold Spring Harbor Symp. Quant. Biol.* 44:1153-1159.
 17. 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.
 18. Doolittle, F. 1979. Protein evolution, p. 1-118. *In* H. Neurath and R. Hill (ed.), *The proteins*, vol. 4, 3rd ed. Academic Press, Inc., London.
 19. Efstratiadis, A., J. W. Posakony, T. Maniatis, R. M. Lawn, C. O'Connell, R. A. Spritz, J. K. DeRiel, B. G. Forget, S. M. Weissman, J. L. Slightom, A. E. Blechl, O. Smithies, F. E. Baralle, C. C. Shoulders, and N. J. Proudfoot. 1980. The structure and evolution of the human β -globin gene family. *Cell* 21:653-668.
 20. Groner, B., and N. E. Hynes. 1980. Number and location of mouse mammary tumor virus proviral DNA in mouse DNA of normal tissue and of mammary tumors. *J. Virol.* 33:1013-1025.
 21. Hishinuma, F., P. J. Debona, S. Astrin, and A. M. Skalka. 1981. Nucleotide sequence of acceptor site and termini of integrated avian endogenous provirus ev1: integration creates a 6 bp repeat of host DNA. *Cell* 23:155-164.
 22. Maxam, A., and W. Gilbert. 1980. Sequencing end-labelled DNA with base specific chemical cleavages. *Methods Enzymol.* 65:499-560.
 23. Michalides, R., L. Van Deemter, R. Nusse, and R. Van Nie. 1978. Identification of the Mtv-2 gene responsible for the early appearance of mammary tumors in the GR mouse by nucleic acid hybridization. *Proc. Natl. Acad. Sci. U.S.A.* 75:2368-2372.
 24. Michalides, R., R. van Nie, R. Nusse, N. E. Hynes, and B. Groner. 1981. Mammary tumor induction loci in GR and DBA mice contain one provirus of the mouse mammary tumor virus. *Cell* 23:165-173.
 25. Morris, V. L., E. Medeiros, G. M. Ringold, J. M. Bishop, and H. E. Varmus. 1977. Comparison of mouse mammary tumor virus-specific DNA in inbred, wild and Asian mice, and in tumors and normal organs from inbred mice. *J. Mol. Biol.* 114:73-91.
 26. Nandi, S., and C. M. McGrath. 1973. Mammary neoplasia in mice. *Adv. Cancer Res.* 17:353-414.
 27. Owens, R. B., and A. J. Hackett. 1972. Tissue culture studies of mouse mammary tumor cells and associated viruses. *J. Natl. Cancer Inst.* 49:1321-1328.
 28. Parks, W. P., E. S. Hubbell, R. J. Goldberg, F. J. O'Neill, and E. M. Scolnick. 1976. High frequency variation in mammary tumor virus expression in cell culture. *Cell* 8:87-93.
 29. Ringold, G. M., P. R. Shank, H. E. Varmus, J. Ring, and K. R. Yamamoto. 1979. Integration and transcription of mouse mammary tumor virus DNA in rat hepatoma cells. *Proc. Natl. Acad. Sci. U.S.A.* 76:665-669.
 30. Sanger, F., and A. R. Coulson. 1978. The use of thin acrylamide gels for DNA sequencing. *FEBS Lett.* 87:107-110.
 31. Sen, G. C., J. Racevskis, and N. H. Sarkar. 1981. Synthesis of murine mammary tumor viral proteins in vitro. *J. Virol.* 37:963-975.
 32. Shank, P. R., J. C. Cohen, H. E. Varmus, K. R. Yamamoto, and G. M. Ringold. 1978. Mapping of linear and circular forms of mouse mammary tumor virus DNA with restriction endonucleases: evidence for a large specific deletion occurring at high frequency during circularization. *Proc. Natl. Acad. Sci. U.S.A.* 75:2112-2116.
 33. Shimotohno, K., S. Mijitani, and H. M. Temin. 1980. Sequence of retrovirus provirus resembles that of bacterial transposable elements. *Nature (London)* 285:550-554.
 34. Staats, J. 1976. Standardized nomenclature for inbred strains of mice: sixth listing. *Cancer Res.* 36:4333-4377.
 35. Sutcliffe, J. G., T. M. Shinnick, I. M. Verma, and R. A. Lerner. 1980. Nucleotide sequence of Moloney leukemia virus: 3' end reveals details of replication, analogy to bacterial transposons, and an unexpected gene. *Proc. Natl. Acad. Sci. U.S.A.* 77:3302-3306.
 36. Swanstrom, R., W. J. DeLorbe, J. M. Bishop, and H. E. Varmus. 1981. Sequencing of cloned unintegrated avian sarcoma virus DNA: viral DNA contains direct and inverted repeats similar to those in transposable elements. *Proc. Natl. Acad. Sci. U.S.A.* 78:124-128.
 37. Tschlis, P., and J. M. Coffin. 1979. Recombination between the defective component of an acute leukemia virus and Rous associated virus 0, an endogenous virus of chickens. *Proc. Natl. Acad. Sci. U.S.A.* 76:3001-3005.
 38. Van Beveren, C., J. G. Goddard, A. Berns, and I. M. Verma. 1980. Structure of Moloney murine leukemia viral DNA: nucleotide sequence of the 5' long terminal repeat and adjacent cellular sequences. *Proc. Natl. Acad. Sci. U.S.A.* 77:3307-3311.
 39. Van Nie, R., and A. A. Verstraeten. 1975. Studies of genetic transmission of mammary tumor virus by C3Hf mice. *Int. J. Cancer* 16:922-931.
 40. Van Nie, R., A. A. Verstraeten, and J. De Moes. 1977. Genetic transmission of mammary tumor virus by GR mice. *Int. J. Cancer* 19:383-390.
 41. Yamamoto, K. R., M. R. Stallcup, J. Ring, and G. M. Ringold. 1978. Mammary tumor virus DNA: a glucocorticoid-responsive transposable element. *Cold Spring Harbor Symp. Quant. Biol.* 42:625-638.