JOURNAL OF VIROLOGY, Oct. 1985, p. 176–182
0022-538X/85/100176-07$02.00/0

Vol. 56, No. 1

Characterization of Human Endogenous Retroviral Envelope RNA Transcripts

ARNOLD B. RABSON,[1*] YASUTARO HAMAGISHI,[1†] PAUL E. STEELE,[1] MARK TYKOCINSKI,[2] AND MALCOLM A. MARTIN[1]

Laboratory of Molecular Microbiology, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland 20205,[1] and Institute of Pathology, Case Western Reserve University, Cleveland, Ohio 44106[2]

Received 1 May 1985/Accepted 9 July 1985

We characterized the structure of human endogenous retroviral *env* RNA transcripts by Northern blot hybridization and cDNA cloning. Polyadenylated 3.0- and 1.7-kilobase *env* RNAs can be identified in placenta, colon carcinoma, and breast carcinoma cells. We have obtained partial cDNA clones of both size classes of RNAs. Both *env* RNAs contained putative gp70 coding sequences; the 1.7-kilobase species, however, lacked sequences in the 3′ *env* region which could specify a p15E analog. Both cDNA clones contained in-frame termination codons; thus, neither could encode full-length *env* proteins.

Mammalian cells contain multiple copies of endogenous retroviral sequences (7, 22, 37, 39). In some species, such as the mouse (8, 31, 32) or the baboon (2), a few of these endogenous proviruses may be expressed as inducible infectious retroviral particles. In other species, such as African green monkeys or humans, no infectious endogenous retroviruses have been isolated, although their genomes contain abundant retroviral sequences (22, 23). Even in animals not producing intact retroviral particles, the endogenous retroviral DNA may be actively expressed in the form of mRNA and proteins. The expression of retroviral mRNA and protein in the absence of virus production has been observed in strains of mice that do not contain inducible murine leukemia viruses (MuLVs) (10, 11, 17, 25, 40). In this regard MuLV envelope protein gp70 has been detected on the surface of cells of many mouse tissues, particularly lymphocytes (25, 40) and secretory cells such as the epididymis and colon in virus-negative strains (10, 17). Secreted forms of the gp70 molecule have been reported in mouse sera and epididymal fluid (10, 11). Studies of MuLV envelope (*env*) mRNA in normal mouse tissues have confirmed the presence of a 3.0-kilobase (kb) mRNA that comigrates with the 21S spliced long terminal repeat (LTR)-*env* mRNA transcribed in MuLV-infected cells (18). In addition to this endogenous 21S *env* mRNA, several other MuLV *env*-containing mRNAs have been identified in normal mouse tissues, including a smaller 1.8- to 2.0-kb LTR- and *env*-hybridizing mRNA (3, 18).

We have previously obtained and characterized human endogenous retroviral DNA segments related to type C MuLVs and endogenous primate type C retroviruses (23, 30). These endogenous human proviruses are present at a frequency of 50 to 100 copies per haploid genome and are thus distinct from ERV-1 and ERV-3 (4, 26), endogenous type C-related proviruses present as a single copy in human DNA. They are also distinct from recently identified multicopy retrovirus-like elements (19) and human type B- and D-related retroviral segments (5, 6). We have examined the expression of the *env* region of our cloned human proviruses in several human tissues (29) and have detected RNA species that hybridized to labeled human LTR and *env* DNA probes. In addition to a 3.0-kb LTR-*env* RNA, placenta cells contained a 1.7-kb LTR-*env*-hybridizing RNA that appeared to be similar in size to the 1.8- to 2.0-kb species reported in the mouse. In this paper, we describe the further characterization of human endogenous *env* RNAs by Northern blot hybridization and cDNA cloning.

MATERIALS AND METHODS

Tissues, cells, and media. Full-term human placenta was obtained at Caesarean section, quick-frozen in liquid nitrogen, and stored at −70°C. The human colon carcinoma cell line SW 1116 (16) was obtained from the American Type Culture Collection and grown in Leibovitz-15 medium (16) supplemented with 10% fetal calf serum (Quality Biologics), penicillin, streptomycin, and glutamine (Microbiological Associates). The human breast carcinoma cell line T47D (14), was a generous gift of J. Zavada. The cell line was grown at 37°C in Dulbecco minimal essential medium with 10% fetal calf serum.

RNA preparation and analysis. Total cellular RNA was prepared by homogenization of tissue or cells in 8 M guanidine hydrochloride (38) (International Biotechnologies, Inc.), followed by precipitation with 0.1 M sodium acetate and 1/2 volume of ethanol. Pellets were suspended in 6 M guanidine hydrochloride and precipitated with 0.1 M sodium acetate and 1/2 volume of ethanol two additional times. The third pellet was suspended in 10 mM EDTA (pH 7.0), extracted with chloroform-butanol (4:1), and finally precipitated in 3 M sodium acetate at 4°C overnight. Polyadenylated RNA was obtained by one cycle of chromatography on oligo(dT)-cellulose columns (Collaborative Research, Inc.) (1). Polyadenylated RNA was analyzed by blot hybridization. Polyadenylated RNA (5 μg) was subjected to electrophoresis in 1% agarose gels containing 6% formaldehyde as a denaturing agent (15) and transferred to nitrocellulose membranes in 20× SSC (1× SSC is 0.15 M NaCl plus 0.015 M sodium citrate). Membranes were baked at 80°C for 2 h and prehybridized at 45°C for 5 to 16 h in hybridization buffer. Hybridization was carried out for 16 to 20 h at 45°C in 50% formamide (Fluka)–5× SSC–5× Denhardt solution–0.1 M Tris (pH 7.5)–10% dextran sulfate (Pharmacia Fine Chem-

* Corresponding author.
† Current address: Central Research Laboratories, Sanraku-Ocean Co., Ltd., Fujisawa, Japan.

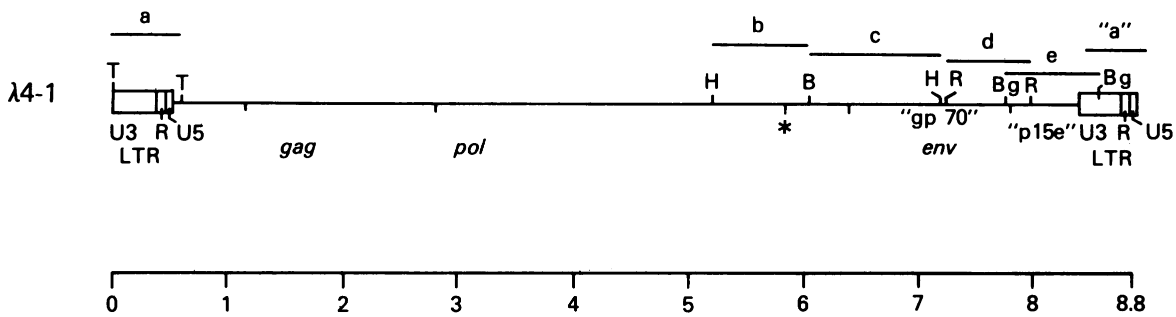

FIG. 1. Map location of human endogenous retroviral probes. The map positions of probes a through e are shown superimposed on a schematic diagram of λ 4-1, a full-length human endogenous provirus. The precise nucleotide coordinates of each probe are described in the text. Line a is shown to indicate that the LTR probe, a, will hybridize to sequences derived from both the 5′ and 3′ retroviral LTR. A partial restriction endonuclease map of λ 4-1 indicating restriction sites used in the generation of subgenomic probes is shown by lettering above the λ 4-1 line. Abbreviations: T, *Taq*I; R, *Rsa*I; H, *Hin*dIII; B, *Bam*HI, Bg, *Bgl*I. The genetic organization of λ 4-1, including the position of the subcomponents of the LTR (U3, R, U5) *gag*, and *pol* and the two putative *env* genes ("gp70" and "p15E") are indicated below the λ 4-1 line. The position of a putative human *env* mRNA splice acceptor at bp 5869 is indicated by an asterisk. The scale is in kilobase pairs.

icals)–300 μg of yeast RNA (Type 3, Sigma Chemical Co.) per ml. ^{32}P-labeled DNA was used at 5×10^5 cpm/ml of hybridization mix. After hybridization, membranes were washed two times for 10 min at 50°C in 2× SSC–0.1% sodium dodecyl sulfate, two times for 15 min in 0.1× SSC–0.1% sodium dodecyl sulfate, and two times for 5 min in 2× SSC. Membranes were air dried and subjected to autoradiography with Kodak X-AR film at −70°C with intensifying screens. DNA segments used as hybridization probes were purified from agarose gels and were radiolabeled with ^{32}P by nick translation (21) with DNase and T4 polymerase (Amersham Corp.).

The positions of the DNA sequences used as probes are shown in Fig. 1; map positions numbered as described by Repaske et al. (30): probe a, a *Taq*I fragment, base pairs (bp) 0 to 540, encompassing the entire 5′ LTR and 40 bases of adjacent pre-*gag* sequence; probe b, a *Hin*dIII-*Bam*HI fragment, bp 5329 to 6088, containing 3′ *pol* sequences, including sequences 3′ to a putative *env* mRNA splice acceptor site at bp 5869 (indicated by asterisk); probe c, a *Bam*HI-*Hin*dIII fragment, bp 6088 to 7312, encompassing 5′ *env* sequences (putative gp70); probe d, a *Rsa*I fragment, bp 7346 to 8010, containing sequences analogous to 3′ gp70 and 5′ p15E of MuLV; probe e, a *Bgl*I fragment, bp 7702 to 8176, encompassing the 3′ putative p15E sequences and approximately 50% of the 3′ U3 LTR (30, 36).

Size fractionation of RNA. Polyadenylated placental RNA (200 μg) in 0.01 M Tris (pH 7.6)–0.1 mM EDTA was heated at 68°C for 5 min, chilled on ice, and then loaded on a 5 to 20% sucrose gradient. Gradients were centrifuged at 35,000 rpm in the SW 41 rotor (Beckman Instruments, Inc.) for 17 h at 4°C. After centrifugation, 0.33-ml fractions were collected, sodium acetate was added to a 0.3 M final concentration, and the fractions were precipitated with 2.5 volumes of cold ethanol at −20°C for 16 h. RNA pellets were suspended in H_2O and analyzed by Northern blot hybridization.

cDNA cloning procedures. cDNA cloning was performed by a modification (13) of the Okayama and Berg method (27). The primer plasmid was tailed as previously described (41) and annealed to approximately 8 μg of size-selected polyadenylated RNA. Reverse transcription was carried out in a total volume of 50 μl with 21 U of reverse transcriptase (Life Sciences, Inc.), 30 U of RNasin (Promega Biotech) and 50 mg of actinomycin D (Sigma Chemical Co.) per ml and incubated at 37°C for 30 min. The RNA-DNA hybrid was tailed with deoxycytidine residues by incubation with 10 U of terminal transferase (P-L Biochemicals, Inc.) at 37°C for 5 min. Cyclization of linker and primer was carried out in the presence of T4 ligase (Bethesda Research Laboratories, Inc.) and replacement of RNA by DNA was performed as described previously (13), using T4 ligase, RNase H (P-L Biochemicals), and DNA polymerase Klenow fragment (Boehringer Mannheim Biochemicals). Recombinant plasmid molecules were introduced into *Escherichia coli* K-12 RRI rendered competent by pretreatment with $CaCl_2$ (20). Transformed bacteria were plated on LB agar plates containing 50 μg of ampicillin per ml. Bacterial colonies were transferred to nitrocellulose, lysed in situ (12), and hybridized with ^{32}P-labeled human endogenous retroviral segments.

Analysis of cDNA clones. Restriction endonucleases (Bethesda Research Laboratories, Boehringer Mannheim Biochemicals, New England Biolabs, Inc.) were used as described in the recommendations of the manufacturers. Nucleotide sequences were determined by the chemical degradation method of Maxam and Gilbert (24), using isolated restriction enzyme fragments end labeled ([γ-^{32}P]ATP, 3,000 Ci/mmol; Amersham) with T4 polynucleotide kinase (P-L Biochemicals). The cDNA insert of pPL1 was sequenced on both strands. The computer program of Queen and Korn (28) was used for translation to amino acids and determination of sequence homology.

RESULTS

Hybridization analysis of human endogenous *env* transcripts. Our previous study of polyadenylated RNA from two human placentas and a human colon carcinoma cell line has shown that these tissues contain several discrete RNA species that hybridized to human endogenous *env* DNA segments (29). The most prominent species was a 3.0-kb LTR- and *env*-related RNA that comigrated with the 3.0-kb 21S spliced LTR-*env* mRNA present in MuLV-infected cells. A second major RNA species of 1.7 kb as well as minor species of 2.2 and 3.6 kb was also detected. To characterize the content of thse *env* RNAs in greater detail, a series of subgenomic segments spanning the *env* and LTR regions of the prototypic full-length human endogenous retroviral clone λ 4-1 (29, 30) was isolated and used as hybridization probes in further Northern analyses. The positions of thse probes (described in Materials and Methods) are shown in Fig. 1.

Figure 2 shows Northern blot hybridizations of poly-

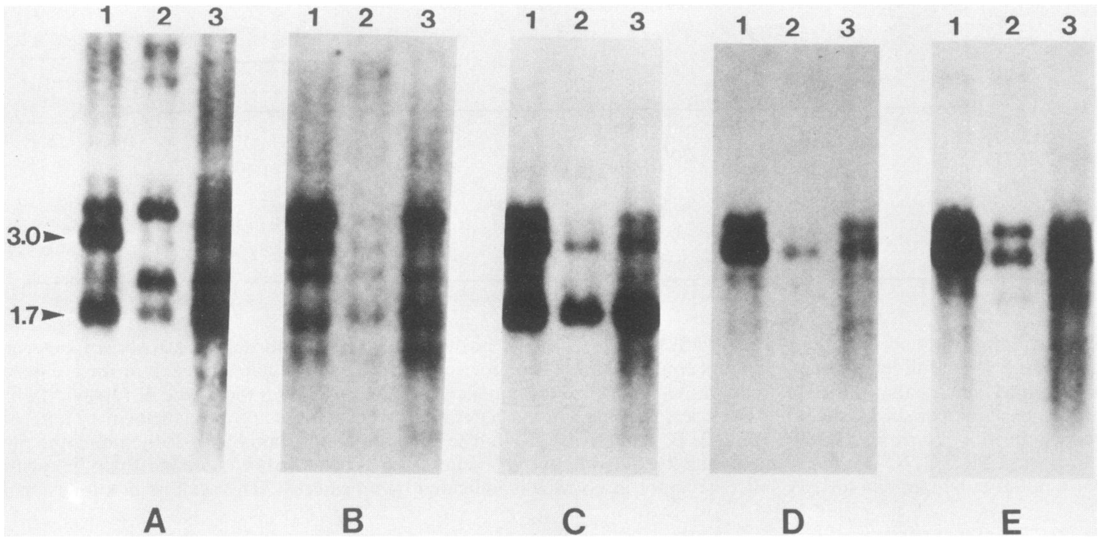

FIG. 2. Northern blot hybridization analysis of human endogenous retroviral RNA in placenta (lanes 1), colon carcinoma (lanes 2), and breast carcinoma (lanes 3) cells. Polyadenylated RNA (5 μg per lane) was electrophoresed through formaldehyde-agarose gels, transferred to nitrocellulose membranes, and hybridized with the five ^{32}P-labeled DNA probes shown in Fig. 1. Panels A through E show hybridization results obtained with probes a through e, respectively. Sizes in kb were determined by comparison with the 8.2- and 3.0-kb mRNAs of MuLV electrophoresed on the same gel and identified by hybridization to MuLV probes.

adenylated RNA derived from human placenta (lanes 1), colon carcinoma cell line SW1116 (lanes 2), and breast carcinoma cell line T47 (lanes 3). Each of the RNA preparations was hybridized with the human retroviral probes a through e described above and shown in panels A through E, respectively.

All three tissues contain several similar LTR- and *env*-hybridizing RNA species including prominent 3.0- and 1.7-kb species as well as additional RNAs of 3.6, 2.2, and 1.3 kb. The 3.0-kb RNA present in all three tissues has been previously shown to comigrate with 21S 3.0-kb spliced LTR-*env* mRNA present in MuLV-infected cells (29). If this transcript represents a human retroviral 21S envelope mRNA, it would be expected to hybridize to probes derived from the LTR as well as sequences 3′ to the putative *env* mRNA splice acceptor. The 3.0-kb RNA hybridized to all five probes (Fig. 2A through E) and failed to react with human *gag* or 5′ *pol* probes (data not shown), thus exhibiting the hybridization properties predicted for a spliced retroviral *env* mRNA. Alternatively, a 3.0-kb RNA with similar hybridization properties could be transcribed from a defective provirus lacking *gag* and *pol* sequences.

A second prominent *env*-containing RNA, 1.7 kb in length, was also present in the three cell types examined. In the colon and breast cell lines, the 1.7-kb RNA was the predominant *env*-reactive RNA species (Fig. 2C). This RNA annealed to probes spanning the LTR, 3′ *pol* (including the potential *env* mRNA splice acceptor), and putative gp70 coding regions (Fig. 2A and C). However, the 1.7-kb RNA failed to hybridize to probes encompassing the 3′ "gp70" sequences, putative p15E sequences, or the 5′ 150 bp of the 3′ LTR (Fig. 2D and E). To further characterize the endogenous *env* RNAs present in human cells, cDNA cloning was carried out with size-selected polyadenylated RNA derived from human placenta.

cDNA cloning of human placental endogenous *env* RNA. Polyadenylated RNA (200 μg) was isolated from a human placenta expressing *env* RNAs and size fractionated by centrifugation through a 5 to 20% sucrose gradient. Individual fractions were analyzed by Northern hybridization to identify the presence of the various species of retroviral *env* RNAs. The hybridization pattern of unfractionated placental polyadenylated RNA (Fig. 3, lane a) was compared with that seen in sucrose gradient fractions 8 and 11 (lanes b and c).

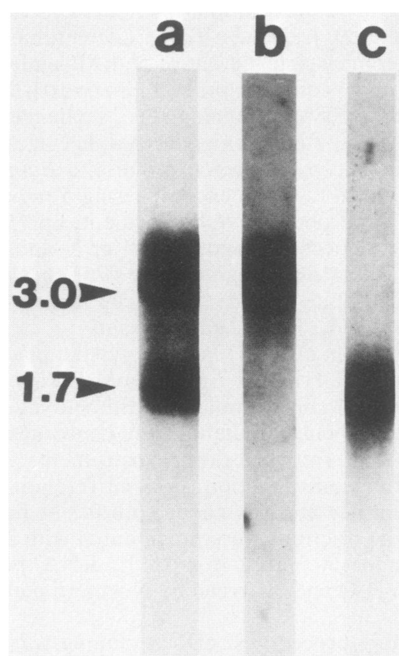

FIG. 3. Sucrose gradient separation of the 3.0- and 1.7-kb *env* RNAs of human placenta. Aliquots of fractions collected from a 5 to 20% sucrose gradient after a rate-zonal centrifugation were analyzed by Northern blot hybridization with a human endogenous retroviral *env* probe (probe c from Fig.1). Lane a, unfractionated placental polyadenylated RNA; lane b, sucrose gradient fraction 8; lane c, sucrose gradient fraction 11.

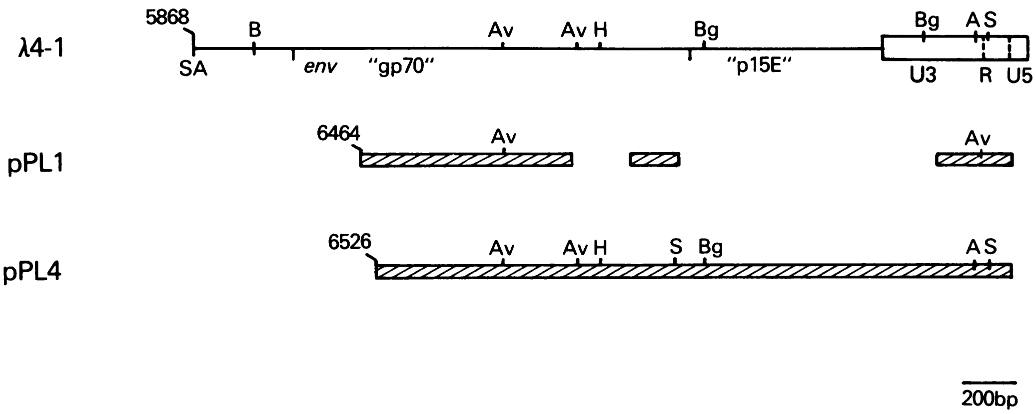

FIG. 4. Schematic representation of human endogenous retroviral DNA sequences present in two cDNA clones. The structure of cDNA clones pPL1 and pPL4 is compared with the full-length endogenous provirus λ 4-1. A potential *env* splice acceptor (SA) for λ 4-1 is shown at bp 5869, and the boundaries of the putative major *env* glycoprotein and transmembrane protein genes ("gp70" and "p15E") and the U3, R, and U5 regions of the 3′ LTR are shown. Restriction endonuclease sites are drawn above each clone. Abbreviations: A, *Acc*I; Av, *Ava*II; B, *Bam*HI; Bg, *Bgl*I; H, *Hin*dIII; S, *Sac*I.

Fractions 8 (lane b) and 9 (not shown) contained *env* RNA species of 3.0 to 3.6 kb; fraction 11 (lane c) contained only the 1.7-kb *env* RNA. Separate cDNA libraries were prepared by the method of Okayama and Berg (27) from the RNA present in gradient fraction 9 (containing the 3.0-kb RNA) and gradient fraction 11 (containing the 1.7-kb RNA). Plasmids containing cDNA inserts were introduced into *E. coli* RR1 and screened by colony hybridization with a mixture of *env* and LTR probes. Sixteen thousand colonies from both the fraction 9 and the fraction 11 libraries were screened, and one *env*-hybridizing clone was identified from each library.

Characterization of *env*-cDNA clones. cDNA clones pPL1, derived from the fraction 11 library, and pPL4, derived from the fraction 9 library, were initially characterized by restriction endonuclease digestion and nucleic acid hybridization with probes derived from the LTR and *env* regions of λ 4-1 (29). The deduced restriction endonuclease maps and arrangement of retroviral sequences for these clones (additionally defined by nucleotide sequencing studies presented below) were compared with the *env* region of the prototype full-length human endogenous clone, λ 4-1 (29) (Fig. 4). pPL1 contained a 1.1-kb cDNA insert that hybridized to probes a and c, including LTR and putative gp70 sequences, respectively (Fig. 1). pPL1 weakly hybridized to probe d and did not hybridize to probe b or to probe e, the latter encompassing putative p15E sequences (data not shown). In addition, clone pPL1 was missing many restriction enzyme sites found in genomic clones of human endogenous retroviral sequences such as the *Hin*dIII site at bp 7312 and the *Acc*I and *Sac*I sites located within the LTR (29). The 5′ end of the pPL1 insert mapped 3′ to the putative *env* mRNA splice acceptor site at bp 5869 in λ 4-1 (30). Thus, pPL1, which was derived from the gradient fraction containing the *env*-reactive 1.7-kb transcript, appeared to be a partial cDNA clone of the 1.7-kb RNA. It possessed hybridization properties similar to those of the 1.7-kb RNA (i.e., it contained LTR and gp70 regions but no p15E region sequences).

By contrast, clone pPL4, which was derived from a gradient fraction containing the 3.0-kb *env* RNA, hybridized to probes encompassing the putative gp70 and p15E region as well as LTR sequences (data not shown). This clone contained a 2.2-kb cDNA insert and possessed several restriction enzyme sites characteristic of human endogenous retroviral sequences (29), including the *Hin*dIII site at bp 7312 and the *Acc*I and *Sac*I sites present in the LTR. The retroviral sequences present in pPL4 did not extend to the putative *env* mRNA splice acceptor; thus, pPL4 was also a partial cDNA clone of one of the *env* RNA species present in fraction 9 (either the 3.6- or 3.0-kb RNA). This clone contained putative gp70 and p15E regions as well as 3′ LTR sequences.

To further characterize the structure of the 1.7-kb *env* RNA, the complete DNA sequence of the 1.1-kb cDNA insert of pPL1 was determined by the method of Maxam and Gilbert (24) and compared with corresponding segments of λ 4-1 (30) (Fig. 5). The nucleotide and deduced amino acid sequences of the *env* region are presented in Fig. 5A. The first base of pPL1 corresponds to position 6464 of the 8,809-bp DNA sequence of λ 4-1 (30). This location is 595 bp 3′ to a putative *env* mRNA splice acceptor site present in λ 4-1 (at a position analagous to the *env* mRNA splice acceptor site of Moloney MuLV) and is within the putative gp70 region of the human endogenous sequence. With the exception of a 222-bp in-frame deletion in pPL1 with respect to λ 4-1 (λ 4-1 map positions 7205 to 7427; Fig. 4 and 5A), pPL1 exhibits colinear nucleic acid sequence homology with λ 4-1 through most of the "gp70" coding region extending to position 7582. Over this region, pPL1 exhibits 94% nucleic acid identity and 87% deduced amino acid identity with the genomic sequences. pPL1 contains three termination codons within this region at positions 6473, 6767, and 7166. Each of these stop codons differs by a single nucleotide from the coding sequence present in λ 4-1 at these positions. Conversely, λ 4-1 contains a termination codon at position 7520 which encodes tryptophan in pPL1.

pPL1 is missing *env* sequences 3′ to position 7582 (Fig. 4 and 5A). The RNA corresponding to pPL1 would therefore lack 45 bp of putative carboxy-terminal gp70 sequences as well as the entire p15E region (beginning at position 7582; Fig. 4). In fact, the deletion in pPL1, relative to λ 4-1, extends into the U3 region of the 3′ LTR (Fig. 4 and 5B). LTR sequences in pPL1 begin at position 8519, 207 bp into the U3 region (30). The nucleic acid sequence of pPL1 over the remainder of the U3 and R portions of the LTR is 90% homologous to that of λ 4-1 (36) and terminates in a polyadenylation sequence at position 8766. Thus, as is shown schematically in Fig. 4, pPL1 is missing a large

A

```
                                                                                                                                  6559
λ4-1  GGA AAT GAC CGA CCT GAT GTG TGT TAT AAC CCA TCT GAG CCT CCT GCA ACC ACC ATT TTT GAA ATA AGA ATA AGA ACT GGC CTT TTC CTA GGT GAT
      Gly Asn Asp Arg Pro Asp Val Cys Tyr Asn Pro Ser Glu Pro Pro Ala Thr Thr Ile Phe Glu Ile Arg Ile Arg Thr Gly Leu Phe Leu Gly Asp
pPL1  GGA AAT GAC TGA CCT GAT GTG TGT TAT AAC CCA TCT GAG CCC CCT ACA ACC ACC AGT TTT GAA ATA AGA TTA AGA ACT GGC CTT TTC CTA GGT GAT
       *   *   *  End  *   *   *   *   *   *   *   *   *   *   *  Thr  *   *  Ser  *   *   *   *  Leu  *   *   *   *   *   *   *   *

                                                                                                                                  6655
λ4-1  ACA AGT AAA ATA ATA ACT AGA ACA GAA GAA AAA GAA ATC CCC AAA CAA ATA ACT TTA AGA TTT GAT GCT TGT GCA GCC ATT AAT AGT AAA AAG CTA
      Thr Ser Lys Ile Ile Thr Arg Thr Glu Glu Lys Glu Ile Pro Lys Gln Ile Thr Leu Arg Phe Asp Ala Cys Ala Ala Ile Asn Ser Lys Lys Leu
pPL1  ACA AGT GAA ATA ATA ACT AGA ACA GAA GAA AAA GGA ATC CCC AAA CAA GTA ACT TTA AGA TTT GAC GCT TGT GCA GCC ATT AAT AGT AAC AAG CTA
       *   *  Glu  *   *   *   *   *   *   *   *   *   *   *   *   *  Val  *   *   *   *   *   *   *   *   *   *   *   *  Asn  *   *

                                                                                                                                  6751
λ4-1  GGA ATA GGA TGT GAT TCT CTT AAC TGG GAA AGG AGC TAC AGA ATA AAA AAT AAA TAT GTT TGT CAT GAG TCA GGG GTT TGT GAA AAT TGT GCC TAT
      Gly Ile Gly Cys Asp Ser Leu Asn Trp Glu Arg Ser Tyr Arg Ile Lys Asn Lys Tyr Val Cys His Glu Ser Gly Val Cys Glu Asn Cys Ala Tyr
pPL1  GGA ACA GGA TGT GGT TCT CTT AAC TGG GAA AGG AGC TAC AGA GTA GAA AAT AAA TAT GTT TGT CAT GAG TCA GGG GTT TGT GAA AAT TGT GCC TTT
       *  Thr  *   *  Gly  *   *   *   *   *   *   *   *   *  Val Glu  *   *   *   *   *   *   *   *   *   *   *   *   *   *   *  Phe

                                                                                                                                  6847
λ4-1  TGG CCA TGT GTT ATT TGG GCT ACT TGG AAA AAG AAC AAA AAG GAC CCG GTT TAT CTT CAG AAG GGG GAA GCC AAC CCC TCC TGT GCT GCT GGT CAC
      Trp Pro Cys Val Ile Trp Ala Thr Trp Lys Lys Asn Lys Lys Asp Pro Val Tyr Leu Gln Lys Gly Glu Ala Asn Pro Ser Cys Ala Ala Gly His
pPL1  TGG CCA TGT GTT ATT TAG GCT ACT TGG AAA AAG AAC AAA AAG GAC TTG GTT CAT CTT CAG AAA GGG GAA GCC AAC CCC TCC TGT GCT GCC AGT CAC
       *   *   *   *   *  End  *   *   *   *   *   *   *   *   *  Leu  *  His  *   *   *   *   *   *   *   *   *   *   *   *  Ser  *

                                                                                                                                  6943
λ4-1  TGT AAC CCA CTA GAA CTA ATA ATT ACC AAT CCC CTA GAT CCC CAT TGG AAA AAG GGA GAA CGT GTA ACC CTG GGG ATT GAT GGG ACA GGG TTA AAC
      Cys Asn Pro Leu Glu Leu Ile Ile Thr Asn Pro Leu Asp Pro His Trp Lys Lys Gly Glu Arg Val Thr Leu Gly Ile Asp Gly Thr Gly Leu Asn
pPL1  TGT AAC CCA CTA GAA CTA ATA ATT ACC AAT CCC CTA GAT CCC CAT TGG AAA AAG GGA GAA TGT GTA ACC CTG GGG ACC AAA GGG ACA GGG TTA AAC
       *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *  Cys  *   *   *   *  Thr Lys  *   *   *   *   *

                                                                                                                                  7039
λ4-1  CCC CAA GTT GCC ATT TTA ATT AGA GGG GAG GTC CAC AAG TGC TCT CCC AAA CCA GTA TTT CAA ACC TTT TAT AAG GAG CTG AAT CTG CCA GCA CCA
      Pro Gln Val Ala Ile Leu Ile Arg Gly Glu Val His Lys Cys Ser Pro Lys Pro Val Phe Gln Thr Phe Tyr Lys Glu Leu Asn Leu Pro Ala Pro
pPL1  CCC CAA GTT GCC ATT TTA GTT CAA GGG GAG GTC CAC AAG CAC TCT CCC AAA CCA GTG TTT CAA ACC TTT TAT GAG GAG TTA AAT CTG CCA GCA CCA
       *   *   *   *   *   *  Val Gln  *   *   *   *   *  His  *   *   *   *   *   *   *   *   *   *  Glu  *   *   *   *   *   *   *

                                                                                                                                  7135
λ4-1  GAA TTT CCA AAA AAG ACA AAA AAT TTG TTT CTC CAA TTA GCA GAA AAT GTA GCT CAT TCC CTT AAT GTT ACT TCT TGT TAT GTA TGC GGG GGA ACC
      Glu Phe Pro Lys Lys Thr Lys Asn Leu Phe Leu Gln Leu Ala Glu Asn Val Ala His Ser Leu Asn Val Thr Ser Cys Tyr Val Cys Gly Gly Thr
pPL1  GAA CTT CTG AAA AAG ATA AAA AAT TTG TTT CTC CAA TTA GCA GAA AAT GTA GCT CAT TCC CTT AAT GTT ACT TCT TGT TAT ATA TGC GGG GGA ACC
       *  Leu Leu  *   *  Ile  *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *  Ile  *   *   *   *

                                                                                                                                  7444
λ4-1  ACT ATC GGA GAC CGA TGG CCT TGG GAA GCC CGA GAG TTG GTG CCT ACT GAT CCA GCT CCT GAT ATA ATT  (222 BP)  TGG GCT CAT CCA GAA TCT
      Thr Ile Gly Asp Arg Trp Pro Trp Glu Ala Arg Glu Leu Val Pro Thr Asp Pro Ala Pro Asp Ile Ile            Trp Ala His Pro Glu Ser
pPL1  ACT ATC AGA GAC CGA TGG CCT TGG GAA GCC TGA GAG TTG GTG CCC ACT GAT CCA GCT CCT GAT ATA ATG    ////    GGG GCT TGT CCA GGA TCT
       *   *  Arg  *   *   *   *   *   *   *  End  *   *   *   *   *   *   *   *   *   *   *  Met            Gly  *  Cys  *  Gly  *

                                                                                                                                  7540
λ4-1  CAT CAG GAC TGG ATG GCT CCC GCT GGA CTA TAC TGG ATA TGT GGG CAC AGA GCC TAC ATT CGG TTA CCT AAT AAA TAG GCA GGC AGT TGT GTT ATT
      His Gln Asp Trp Met Ala Pro Ala Gly Leu Tyr Trp Ile Cys Gly His Arg Ala Tyr Ile Arg Leu Pro Asn Lys End Ala Gly Ser Cys Val Ile
pPL1  CAT CAG GAC TGG ATG GCT CTC GCT GGA CTA TAC TGG ATA TGT GGG CAG AGA GCC TAC ATT CAG TTA CCT AAT GAA TGG GCA GAC AGT TGT GTT ATT
       *   *   *   *   *   *  Leu  *   *   *   *   *   *   *   *  Gln  *   *   *   *  Gln  *   *   *  Glu Trp  *  Asp  *   *   *   *

λ4-1  GGC ACT ATT AAG TCG TCC TTT TTC TTA TTA CCC ATA AAA ACA  (727 BP)
      Gly Thr Ile Lys Ser Ser Phe Phe Leu Leu Pro Ile Lys Thr
pPL1  GGC ACT ATT AAG CCA TCC TTT TTC TTA TTA CCG ATA AAA ACT    ////
       *   *   *   *  Pro  *   *   *   *   *   *   *   *   *
```

B

```
                                                                                                                           8437
λ4-1  TATGGTATGAGGTCGCCACTTCTCCTGTTGTCCTTCTCAGTTTCTCCCCAACCTCCCCTTTTCCCTAGTTTATAAGACAGGAGAAAAGGGAGAAAGCAAAAAGTTGAAAAGAAACAGAAGTAAGAT
pPL1  ------------------------------------------------------------------------------------------------------------------------------
                                                                                                                           8563
λ4-1  AAATAGCTGGACGACCTTGGCACCACCACCTGGCCCTGGTGGCTAAAATAATAATAATATTATTAACCCCTGACCAAAACTATTGGTGTTATCTGTAAATTCCAGACACTGTATGAGAAAATACTG
pPL1  ---------------------------------------------------------------------------------ACTGGTACTATCTGTAAATTCCAGACATTGTATGAGAAAGCACTG
                                                                                                                           8688
λ4-1  TAAAACTTTTTGTTCTGTTAGCTGATGTATGTAGCCCCCAGTCATGTTTCTCACGCTTACTTGATCTATTATGACTTTTTCATGTAGACCCCTTAGAGTTCTAAGCCCTTAAAAGGGCAAG AATT
pPL1  TAAAACTTTTTGTTCTGTTAGCTGATATATGTAGCCTCCAGTCACATTCCTCATGCTTACTTGATCTATCATGACCCTTTCACGTGGACCCCTTAGAGTTGTAAGCCCTTAAAAGCGCTAGGAATT
                                                                                                                    8809
λ4-1  TCTTTTTCGGGGAGCTCGGCTCTTAAGACACGAGTCTGCCAATGATCCCGGCCGAATAAAAAACCTCTTCCTTCTTTAATCTGGCGTCTGAGGAGTTTTGTCTGCGACTCATCCTGCTACA
pPL1  TCTTTTTGGGGGAGCTTGGCTCTTAAGACATGAGTCTGCCAATGCTACCGGCCAAATAAAA ACCTCTTCCTTCTTTAAAAA(A)n
```

FIG. 5. Nucleotide and deduced amino acid sequences of cDNA clone pPL1, compared with corresponding sequences of the integrated provirus, λ 4-1. (A) Comparison of the *env* region of pPL1 and λ 4-1. Amino acid identities are identified by asterisks. Two deletions are present, a 222-bp deletion at bp 7,205 to 7,7427 and a deletion of the last 727 bp of *env* that extends into the LTR. (B) Comparison of U3 and R nucleotide sequences of λ 4-1 and pPL1. The first 207 bp of U3 are deleted in pPL1 relative to λ 4-1.

segment of retroviral sequence, including the carboxy-terminal coding sequence of the putative human gp70 protein, the entire coding region for the potential human p15E equivalent, and 5′ U3 LTR sequences. These DNA sequence data correlate with the Northern blot hybridization data, thus confirming that the 1.7-kb *env* RNA is deleted in 3′ *env* sequences.

Partial DNA sequence data obtained from pPL4 (not shown) has confirmed that it contains putative gp70 and p15E sequences as well as the complete U3 and R regions of the 3′ LTR as shown schematically in Fig. 4. The nucleotide and deduced amino acid sequences of pPL4 are very similar to those of λ 4-1 and pPL1. However, pPL4 contains at least one termination codon (position 6893) not present in λ 4-1.

	5'	3'
λ 4-1	TA AAAACAG GTGAG	CCAAAACTA TTGGTGT
pPL4	TA AAAACGG GTGAG	CCAAAACTA CTGGTGT
pPL1	TA AAAAC	____TA CTGGTAC
CONSENSUS SPLICE	AG'GTRAG	Y-YYY-CAG'

FIG. 6. Comparison of nucleotide sequences forming the boundaries of the *env*-U3 deletion of pPL1 with corresponding sequences of λ 4-1 and pPL4, a cDNA clone with a nondeleted *env* and U3 region. A potential target duplication for homologous recombination, AAAAC, present in two copies in λ 4-1 and pPL4 but only one copy in pPL1, is underlined. The consensus mRNA splice donor and acceptor sequences (33) are shown on the bottom line in their best alignment with the retroviral sequences.

Conversely, one termination codon present in λ 4-1 encodes a tryptophan in pPL4.

DISCUSSION

Human endogenous retroviral DNA sequences are transcribed into polyadenylated RNA in a number of human tissues. The transcription of human LTR and *env* sequences is particularly prominent in placenta, colon carcinoma, and breast carcinoma cells. In addition to a 3.0-kb LTR-*env*-hybridizing RNA that comigrates with the spliced LTR-*env* mRNA expressed in MuLV-infected cells, these human cells contain a shorter, 1.7-kb LTR-*env* RNA species. We have structurally characterized the human *env* RNAs by Northern blot hybridization and cDNA cloning experiments. DNA sequence analysis of a partial cDNA clone, pPL1, of the 1.7-kb RNA has revealed that in comparison with a cloned, full-length, endogenous provirus, λ 4-1, pPL1 contains a 938-bp deletion involving 3′ *env* region sequences as well as a portion of the U3 LTR. A second, 222-bp deletion is present entirely within the *env* region. The position of the large deletion is in complete agreement with the Northern blot analyses that showed the failure of the 1.7-kb RNA to hybridize to probes derived from 3′ *env* and U3 sequences. Thus, the smaller size of the 1.7-kb RNA relative to the 3.0-kb LTR-*env* RNA can be explained by the deletions present in the cDNA clone; additional small deletions in the RNA 5′ to sequences cloned in pPL1 cannot be excluded.

The large deletion involving 3′ *env* and U3 LTR sequences in pPL1 could have arisen either as a result of alternative splicing of a 3.0-kb LTR-*env* RNA or transcription of a deleted DNA provirus. An analysis of thse two possibilities is presented in the DNA sequence comparison shown in Fig. 6. The DNA sequences surrounding the 938-bp deletion in pPL1 have been compared with those present in the same regions of λ 4-1 and pPL4. These sequences are aligned with each other and with the consensus nucleic acid sequence of mRNA splice donor and acceptor sites (33). If the 1.7-kb RNA arises as a result of RNA splicing of a larger *env* RNA precursor, the sequences adjacent to the putative splice junction in a potential RNA precursor such as pPL4 or in a putative genomic template for such a precursor such as λ 4-1 should resemble deduced consensus splice donor and acceptor sequences. Alignment of λ 4-1 or pPL4 with pPL1 (Fig. 6) reveals that both λ 4-1 and pPL4 contain consensus splice donor sequences adjacent to the 5′ (*env*) breakpoint in pPL1. However the 3′ LTRs of λ 4-1 and pPL4 contain no consensus splice acceptor sequences, both having purine-rich regions preceding TA codons instead of pyrimidine-rich regions preceding AG codons.

An alternative explanation for the generation of the 1.7-kb mRNA is that it is transcribed from a provirus containing an extensive deletion in the *env* and U3 regions. λ 4-1 and pPL4 both contain a duplicated 5-bp sequence, AAAAC, which flanks the segment deleted from pPL1 (Fig. 6). pPL1 contains a single copy of the 5-bp sequence, possibly the result of homologous recombination between the repeated units present in a primordial proviral DNA. Such a recombinational event would generate an altered provirus that contains a deletion of the sequences between the duplication. Additional evidence for the existence of such a deleted provirus has been obtained by Southern blot hybridization of colon carcinoma DNA (SW1116 cells) with probe D from Fig. 1 (data not shown). This experiment indicated the existence of an *Ava*II fragment in the colon cells that comigrates with the unique 518-bp *Ava*II fragment present in pPL1 (Fig. 4) encompassing the *env* deletion.

The sequences deleted in the 1.7-kb RNA include the 3′ *env* region and a portion of the U3 LTR. Although the sequences of the putative human retroviral *env* gene do not match the corresponding sequence of MuLVs, this region contains a long open reading frame lying 3′ to a potential *env* mRNA splice acceptor analogous to the structure of murine leukemia viruses. In the absence of definitive identification of human retroviral *env* proteins, it is not possible to assign gene products to the *env* segment. However, this region of the human retroviral clones that would correspond to gp70 of Moloney MuLV (34) contains multiple glycosylation sites (30). A potential "gp70-p15E" proteolytic cleavage site can be identified in the human clone which would give rise to a "p15E" gene region approximately equal in size to that of Moloney MuLV (34) (as shown for λ 4-1 in Fig. 4). The putative human p15E contains 40% deduced amino acid homology to the 26-amino acid region of the Moloney MuLV p15E conserved in a wide range of retroviruses (9) and contains a hydrophobic region that could serve as a transmembrane anchor sequence. Thus, by analogy with MuLV, the 1.7-kb human *env* RNA is missing approximately 100 bp of the 3′ gp70-analogous sequences as well as the entire p15E-analogous region, including potential transmembrane anchor sequences.

The DNA sequence of pPL1 contains three termination codons within the *env* reading frame. pPL4 contains at least one in-frame termination codon. The presence of these termination codons has been confirmed by DNA sequencing of each strand and cannot be due to ambiguities of nucleotide sequencing. Errors of reverse transcription or second strand synthesis could be responsible for artifactual conversion of coding triplets to termination codons; however, it is more likely that the polyadenylated RNA species from which the cDNA clones were derived did, in fact, contain these stop signals precluding their translation into *env* proteins. Clearly, the two proviruses from which these nonfunctional RNAs were transcribed had active promoter elements. Multiple additional copies of *env*-related genes exist in the human genome (36), any one of which may be transcribed. If one or more of these contain a completely open reading frame, its transcription and translation could produce a human *env* protein. It is of interest that a 1.8- to 2.0-kb MuLV *env* mRNA has been identified in normal mouse tissue (3, 18). This RNA apparently possesses a structure similar to that observed for the human 1.7-kb RNA: it hybridizes to 5′ gp70 and LTR probes but appears to be deleted in 3′ gp70 and transmembrane-anchor p15E sequences. This truncated MuLV *env* mRNA has been hypothesized to encode a secreted form of gp70. Secreted forms of the gp70 molecule have in fact been identified in mouse serum (11) and epididymal fluid (10). We are currently attempting to identify membrane-bound and secreted human *env* peptides by using antisera against synthetic oligopeptides predicted from the human *env* DNA sequence (35).

ACKNOWLEDGMENTS

We thank Kim Boulukos for assistance in screening cDNA clones, R. Willey for assistance in the analysis of pPL4, J. Zavada for T47D breast carcinoma cells, and S. Rosenfeld and J. Barnhart for editorial assistance.

LITERATURE CITED

1. **Aviv, H., and P. Leder.** 1972. Purification of biologically active globin messenger RNA by chromatography on oligothymidylic acid cellulose. Proc. Natl. Acad. Sci. USA **69:**1408–1412.
2. **Benveniste, R. E., and G. J. Todaro.** 1974. Multiple divergent copies of endogenous C-type virogenes in mammalian cells. Nature (London) **252:**170–173.
3. **Boccara, M., M. Souyri, C. Magarian, E. Stavnezer, and E. Fleissner.** 1983. Evidence for a new form of retroviral *env* transcript in leukemic and normal mouse lymphoid cells. J. Virol. **48:**102–109.
4. **Bonner, T. I., C. O'Connell, and M. Cohen.** 1982. Cloned endogenous retroviral sequences from human DNA. Proc. Natl. Acad. Sci. USA **79:**4709–4713.
5. **Callahan, R., I. M. Chiu, J. F. H. Wong, S. R. Tronick, B. A. Roe, S. A. Aaronson, and J. Schlom.** 1985. A new class of endogenous human retroviral genomes. Science **228:**1208–1211.
6. **Callahan, R., W. Drohan, S. Tronick, and J. Schlom.** 1982. Detection and cloning of human DNA sequences related to the mouse mammary tumor virus genome. Proc. Natl. Acad. Sci. USA **29:**5503–5507.
7. **Chattopadhyay, S. K., D R. Lowy, N. M. Teich, A. S. Levine, and W. P. Rowe.** 1974. Qualitative and quantitative studies of AKR-type murine leukemia virus sequence in mouse DNA. Cold Spring Harbor Symp. Quant. Biol. **39:**1085–1101.
8. **Chattopadhyay, S. K., D. R. Lowy, N. M. Teich, A. S. Levine, and W. P. Rowe.** 1974. Evidence that the AKR murine-leukemia-virus genome is complete in DNA of the high-virus AKV mouse and incomplete in the DNA of the "virus negative" NIH mouse. Proc. Natl. Acad. Sci. USA **71:**167–171.
9. **Cianciolo, G. J., R. J. Kipnis, and R. Synderman.** 1984. Similarity between p15E of murine and feline leukemia viruses and p21 of HTLV. Nature (London) **311:**515.
10. **Del Villano, B. C., and R. A. Lerner.** 1976. Relationship between the oncornavirus gene product gp70 and major protein secretion of the mouse genital tract. Nature (London) **259:**497–499.
11. **Elder, J. H., F. C. Jensen, M. L. Bryant, and R. A. Lerner.** 1977. Polymorphism of the major envelope glycoprotein (gp70) of murine C-type viruses: virion associated and differentiation antigens encoded by a multi-gene family. Nature (London) **267:**23–28.
12. **Grunstein, M., and D. Hogness.** 1975. Colony hybridization: a method for the isolation of cloned DNAs that contain a specific gene. Proc. Natl. Acad. Sci. USA **72:**3961–3965.
13. **Gunning, P., P. Ponte, H. Okayama, J. Engel, H. Blau, and L. Kedes.** 1983. Isolation and characterization of full-length cDNA clones for human α-, β- and γ-actin mRNAs: skeletal but not cytoplasmic actins have an amino-terminal cysteine that is subsequently removed. Mol. Cell. Biol. **3:**787–795.
14. **Keydar, I., L. Chen, S. Karby, F. R. Weiss, J. Delarea, M. Radu, S. Chaiteik, and H. J. Brenner.** 1979. Establishment and characterization of a cell line of human breast carcinoma origin. Eur. J. Cancer. **15:**659–670.
15. **Lehrach, H., D. Diamond, J. M. Wozney, and H. Boedtker.** 1977. RNA molecular weight determination by gel electrophoresis under denaturing conditions, a critical reexamination. Biochemistry **16:**4743–4751.
16. **Leibovitz, A., J. C. Stinson, W. B. McCombs, C. E. McCoy, K. C. Mazur, and N. P. Mabry.** 1976. Classification of human colorectal adenocarcinoma cell lines. Cancer Res. **36:**4562–4569.
17. **Lerner, R. A., C. B. Wilson, B. C. Del Villano, P. J. McConahey, and F. J. Dixon.** 1976. Endogenous oncoviral gene expression in adult and fetal mice: quantitative, histologic, and physiologic studies of the major viral glycoprotein, gp70. J. Exp. Med. **143:**151–166.
18. **Levy, D. E., R. A. Lerner, and M. C. Wilson.** 1982. A genetic locus regulates the expression of tissue-specific mRNAs from multiple transcription units. Proc. Natl. Acad. Sci. USA **79:**5823–5827.
19. **Mager, D. L., and P. S. Henthorn.** 1984. Identification of a retrovirus-like repetitive element in human DNA. Proc. Natl. Acad. Sci. USA **79:**4709–4713.
20. **Mandel, M., and A. Higa.** 1970. Calcium dependent bacteriophage DNA infection. J. Mol. Biol. **53:**159–162.
21. **Maniatis, T., A. Jeffrey, and D. G. Kleid.** 1975. Nucleotide sequence of the rightward operator of phage λ. Proc. Natl. Acad. Sci. USA **72:**1184–1188.
22. **Martin, M. A., T. Bryan, T. F. McCutchan, and H. W. Chan.** 1981. Detection and cloning of murine leukemia virus-related sequences from African green monkey liver DNA. J. Virol. **39:**835–844.
23. **Martin, M. A., T. Bryan, S. Rasheed, and A. S. Khan.** 1981. Identification and cloning of endogenous retroviral sequences present in human DNA. Proc. Natl. Acad. Sci. USA **78:**4892–4896.
24. **Maxam, A. M., and W. Gilbert.** 1980. Sequencing end-labeled DNA with base-specific chemical cleavages. Methods Enzymol. **65:**499–560.
25. **Morse, H. C., III, T. M. Chused, M. Boehm-Truitt, B. J. Mathieson, S. O. Sharrow, and J. W. Hartley.** 1979. XenCSA: cell surface antigens related to the major glycoproteins (gp70) of xenotropic murine leukemia viruses. J. Immunol. **122:**443–454.
26. **O'Connell, C. S. O'Brien, W. G. Nash, and M. Cohen.** 1984. ERV3, a full-length human endogenous provirus: chromosomal localization and evolutionary relationships. Virology **138:**225–235.
27. **Okayama, H., and P. Berg.** 1982. High efficiency cloning of full-length cDNA. Mol. Cell. Biol. **2:**161–170.
28. **Queen, C. L., and J. L. Korn.** 1980. Computer analysis of nucleic acids and proteins. Methods Enzymol. **65:**595–609.
29. **Rabson, A. B., P. E. Steele, C. F. Garon, and M. A. Martin.** 1983. mRNA transcripts related to full-length endogenous retroviral DNA in human cells. Nature (London) **306:**604–607.
30. **Repaske, R., P. E. Steele, R. R. O'Neill, A. B. Rabson, and M. A. Martin.** 1985. Nucleotide sequence of a full-length human endogenous retroviral segment. J. Virol. **54:**764–772.
31. **Risser, R., and J. M. Horowitz.** 1983. Endogenous mouse leukemia viruses. Annu. Rev. Genet. **17:**85–121.
32. **Rowe, W. P.** 1978. Leukemia virus genomes in the chromosomal DNA of the mouse. Harvey Lect. **71:**173–192.
33. **Sharp, P.** Speculations on RNA splicing. 1981. Cell **23:**643–646.
34. **Shinnick, T. M., R. A. Lerner, and J. G. Sutcliff.** 1981. Nucleotide sequences of Moloney murine leukemia virus. Nature (London) **293:**543–548.
35. **Shinnick, T. M., J. G. Sutcliff, N. Green, and R. A. Lerner.** 1983. Synthetic peptide immunogens as vaccines. Annu. Rev. Microbiol. **37:**425–446.
36. **Steele, P. E., A. B. Rabson, T. Bryan, and M. A. Martin.** 1984. Distinctive termini characterize two families of human endogenous retroviral sequences. Science **225:**943–947.
37. **Steffen, D. L., and R. Weinberg.** 1978. The integrated genome of murine leukemia virus. Cell **15:**1003–1010.
38. **Strohman, R. C., P. S. Moss, S. Micou-Eastwood, D. Spector, A. Przbyla, and B. Paterson.** 1977. Messenger RNA for myosin polypeptides: isolation from single myogenic cell cultures. Cell **10:**265–273.
39. **Todaro, G. J., R. E. Benveniste, R. Callahan, M. M. Lieber, and C. J. Sherr.** 1974. Endogenous primate and feline type C viruses. Cold Spring Harbor Symp. Quant. Biol. **39:**1159–1168.
40. **Tung, S. S., E. S. Vitetta, E. Fleissner, and E. A. Boyse.** 1975. Biochemical evidence linking G_{IX} thymocyte surface antigen to the gp 69/71 envelope glycoprotein of murine leukemia virus. J. Exp. Med. **141:**198–205.
41. **Tykocinski, M. L., P. N. Marche, E. E. Max, and T. J. Kindt.** 1984. Rabbit class I MHC genes: cDNA clones define full-length transcripts of an expressed gene and a putative pseudogene. J. Immunology **133:**2261–2269.