JOURNAL OF VIROLOGY, Feb. 1986, p. 422–432
0022-538X/86/020422-11$02.00/0

Vol. 57, No. 2

Murine Mammary Tumor Virus *pol*-Related Sequences in Human DNA: Characterization and Sequence Comparison with the Complete Murine Mammary Tumor Virus *pol* Gene

KEITH C. DEEN* AND RAYMOND W. SWEET

Smith Kline & French Laboratories, Philadelphia, Pennsylvania 19101

Received 17 June 1985/Accepted 10 October 1985

Sequences in the human genome with homology to the murine mammary tumor virus (MMTV) *pol* gene were isolated from a human phage library. Ten clones with extensive *pol* homology were shown to define five separate loci. These loci share common sequences immediately adjacent to the *pol*-like segments and, in addition, contain a related repeat element which bounds this region. This organization is suggestive of a proviral structure. We estimate that the human genome contains 30 to 40 copies of these *pol*-related sequences. The *pol* region of one of the cloned segments (HM16) and the complete MMTV *pol* gene were sequenced and compared. The nucleotide homology between these *pol* sequences is 52% and is concentrated in the terminal regions. The MMTV *pol* gene contains a single long open reading frame encoding 899 amino acids and is demarcated from the partially overlapping putative *gag* gene by termination codons and a shift in translational reading frame. The *pol* sequence of HM16 is multiply terminated but does contain open reading frames which encode 370, 105, and 112 amino acid residues in separate reading frames. We deduced a composite *pol* protein sequence for HM16 by aligning it to the MMTV *pol* gene and then compared these sequences with other retroviral *pol* protein sequences. Conserved sequences occur in both the amino and carboxyl regions which lie within the polymerase and endonuclease domains of *pol*, respectively.

The human genome carries multiple copies of retrovirus-related sequences which were identified through their homology with mammalian type C (3, 28–31, 36, 37, 44) or type B (6, 27) retroviruses. Several of the type C sequences have a proviral structure (31, 37, 48), although other organizations are observed (48), and a comparison of partial sequence information indicates that they are all related to each other (31). Distinct from this group is the family of sequences originally detected by their homology to the murine type B virus, murine mammary tumor virus (MMTV) (6, 27). We refer to these MMTV-related sequences as the HM family. The single characterized member of this family also has a provirus-like structure and displays a mosaic homology to type A, B, and D retroviruses (5).

The functional significance of these human endogenous proviral elements is not known. Those that have been characterized apparently did not result from a recent retroviral infection, since the 5′ and 3′ long terminal repeat (LTR) regions show sequence heterogeneity with respect to each other (30, 31, 48), and in at least one instance, the genomic location of the provirus was conserved in the chimpanzee genome (3). Furthermore, these proviral sequences could not encode a fully competent retrovirus since limited sequence determination within the structural genes reveals inappropriate termination signals (3, 31). However, some of the retrovirus-related sequences are transcriptively active. Rabson et al. (33) observed subgenomic RNA transcripts in various human tissues, although the sense of these transcripts was not established.

We were intrigued by the possibility of a functional polymerase (reverse transcriptase) activity derived from a human endogenous provirus. It has been suggested that such an activity could be responsible for the generation of pseudogenes and the expansion of repeat DNA families such as *Alu* and *Kpn* (44). Also, there is recent evidence of reverse transcription of transposable elements in both yeasts and drosophila (reviewed in reference 1). We chose to characterize the *pol* region of the MMTV-related family because of the relatively strong homology these sequences demonstrated to the MMTV *pol* gene. We describe here the general organization of several members of this family in which the *pol*-related sequence is colinear with the MMTV gene and present the complete sequence of one such member. For appropriate comparison, we sequenced the entire MMTV *pol* gene, only a portion of which was previously reported (7, 35), and further discuss its organization within the viral genome.

MATERIALS AND METHODS

Plasmids and DNA fragment probes. Plasmids pMTVP-1, pMTVP-2, pMTVP-3, and pMTVP-4 (Fig. 1), *Pst*I subclones of the MMTV GR40 provirus (16), were kindly provided by B. Groner and N. Hynes (Ludwig Institute for Cancer Research, Bern, Switzerland). The 1.3-kilobase (kb) *Bgl*II-*Eco*RI and 1.1-kb *Eco*RI-*Pst*I fragments spanning the *pol* region in pMTVP-3 were subcloned into pBR322 to yield plasmids pMM3-2 and pMM3-1, respectively (Fig. 1). For the pMM3-2 construction, pBR322 was first modified by the insertion of a *Bgl*II linker at the unique *Nru*I site (J. Young, Smith Kline & French Laboratories). The inserts from these plasmids were subsequently cleaved with *Ava*I or *Hin*dIII to yield four fragments spanning the *pol* region (designated a to d in Fig. 1) which were purified by polyacrylamide gel electrophoresis and electroelution. Plasmids and DNA fragments were ^{32}P labeled by nick translation (38) to a specific activity of 1 × 10^8 to 3 × 10^8 dpm/μg for use as probes.

Hybridizations and plaque screening. DNA fragments in agarose gels were transferred to nitrocellulose by a depurination modification (51) of the procedure described by Southern (47). Filters were baked for 4 h at 80°C under vacuum

* Corresponding author.

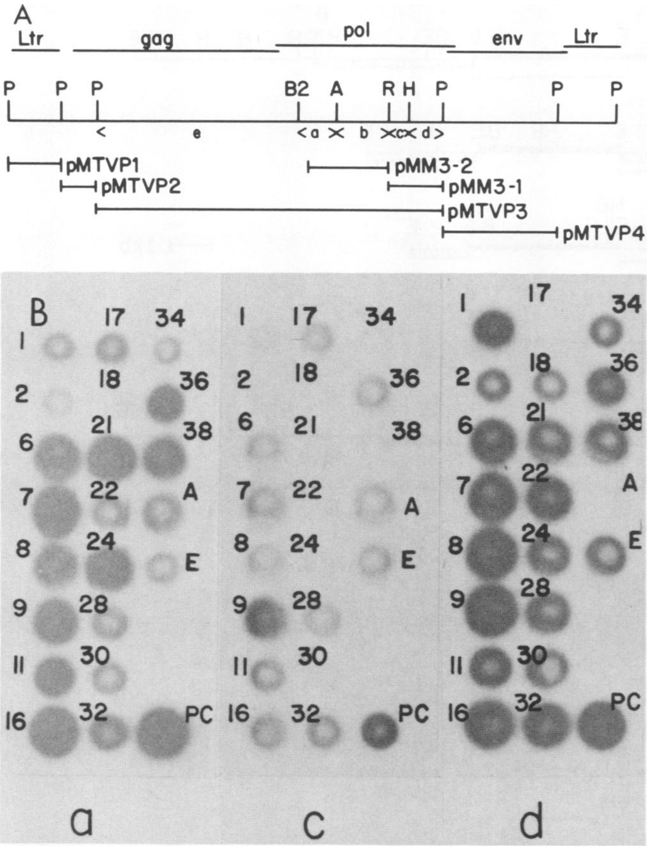


FIG. 1. (A) Restriction endonuclease map of the endogenous MMTV provirus GR40 (16). The approximate position of the genes for *gag*, *pol*, *env*, and LTR are indicated above the map, and regions subcloned into pBR322 are shown below. Fragments purified from plasmids pMM3-2 (a and b), pMM3-1 (c and d), and pMTVP-3 (e) are indicated by brackets. Restriction sites are *Ava*I (A), *Bgl*II (B2), *Hin*dIII (H), *Eco*RI (E), and *Pst*I (P). The *Ava*I and *Bgl*II sites are not unique in the genome and are shown to indicate regions covered by probes a, b, and e. (B) Identification of clones containing sequences related to the MMTV polymerase-coding region. DNAs (100 ng) from recombinant clones (1 to 38, A, E) of a Charon 4A library of human fetal liver DNA were spotted onto nitrocellulose filters and probed with MMTV fragments a through d under lowered stringency conditions (see Materials and Methods). Results are shown for fragments a, c, and d; no hybridization was observed with fragment b. The 5.2-kb *Pst*I *gag-pol* fragment (500 pg) of MMTV (isolated from plasmid pMTVP-3) was used as a positive control (PC). Films were exposed for 16 h at −70°C with an intensifying screen. Based on the signal intensities relative to PC, clones 6, 7, 8, 11, 16, 21, 24, 32, 36, and 38 were selected for further characterization.

and blocked for 4 to 6 h at 65°C in 6× SSPE (0.9 M NaCl, 60 mM $NaPO_4$ [pH 7.4], 6 mM EDTA) containing 0.04% polyvinylpyrrolidone 360, 0.08% Ficoll 400, and 0.04% bovine serum albumin. Hybridizations were carried out in 6× SSPE for 16 h at 65°C (stringent conditions) or 53°C (lowered stringency conditions) at probe concentrations of approximately 10 ng/ml. Filters were washed in 2× SSPE containing 0.05% sodium dodecyl sulfate at 65°C (stringent conditions) or 53°C (lowered stringency conditions).

Clones of a Charon 4A library of human fetal liver DNA (21) were screened for MMTV-related polymerase sequences by plaque hybridization (2) under lowered stringency conditions with the 5.2-kb *Pst*I *gag-pol* fragment of clone pMTVP-3 described above. Of the 9×10^5 plaques screened, 39 strongly hybridized to the *gag-pol* probe, with an additional 119 plaques yielding weak hybridization signals. All of the 39 strongly hybridizing plaques and 5 of the weakly hybridizing plaques were chosen for further study.

DNA sequencing. Fragments for sequencing were 5′ end labeled with [γ-^{32}P]ATP and polynucleotide kinase and sequenced by the chemical cleavage method of Maxam and Gilbert (26).

Cell lines and tissue specimens. Cell lines used were HeLa, HL-60 (8), GM1056 (Epstein-Barr virus-transformed B cell; B. Bloom, Columbia University), and the breast tumor lines MCF7 (46) and 47-D (18). Surgical breast tumor specimen J-1 was kindly provided by T. Ohno (Tokyo Jikei University, Tokyo, Japan), and placental tissues were obtained from C. Crum (Columbia University). DNA was extracted from tissues (after pulverizing in liquid N_2) and cell lines by lysis in 1% sodium dodecyl sulfate, overnight incubation at 37°C with 0.1 mg of pronase per ml, phenol extraction, and ethanol precipitation.

Quantitation of homology among retroviral polymerases. The *pol* protein regions of Rous sarcoma virus (RSV) (42), Moloney murine leukemia virus (MMLV) (45), human adult T-cell leukemia virus (HTLV I) (43), lymphoadenopathy-associated virus (LAV) (52), MMTV, and clone 16 of the HM family were aligned with each other according to the program of Kanehisa (17). Regions of conserved sequence with few gaps were noted among these *pol* proteins at the amino and carboxyl termini. Based on these alignments, we arbitrarily divided the *pol* protein sequence into conserved amino and carboxyl domains separated by a less-conserved middle domain. These regions are listed with respect to nucleotide sequence number in Table 1. The corresponding residues in the MMTV protein sequence are 21 to 234 and 611 to 751 (see Fig. 6) for the amino and carboxyl regions, respectively. The distance scores obtained by alignment of these regions with each other were divided by the score obtained from self-alignment to obtain the percentage values shown in the table.

RESULTS

Isolation of human *pol* gene sequences. The organization of the MMTV GR40 provirus (16) is shown schematically in Fig. 1A. The position of the LTR, *gag*, *pol*, and *env* genes is based on partial sequence data for this and related MMTV genomes (12, 14, 23, 35) and on sequence data presented herein. Under lowered stringency hybridization conditions, fragments cross-reactive with the MMTV genome are observed in Southern blots of human DNA (6, 27). This homology is limited almost exclusively to the viral *pol* gene (6, 27; see below). To isolate these *pol*-related human sequences, we screened a human phage library (21) with an MMTV *gag-pol* probe (insert of plasmid pMTVP-3; Fig. 1A) as described above. To identify positive phage clones which contained potentially complete *pol* genes, we screened each of the 39 strongly hybridizing clones and 5 of the weakly hybridizing clones by DNA dot-blot analysis for homology to the amino and carboxyl domains of the MMTV *pol* gene (plasmids pMM3-2 and pMM3-1, respectively). Of these phage clones, 21 hybridized with both probes (data not shown). The extent of *pol* homology in these clones was further assessed by dot-blot hybridization with four nonoverlapping fragments (a through d; Fig. 1A) of the MMTV *pol* gene. Most of the clones annealed to a variable extent with three of the fragments (a, c, and d), but none hybridized with fragment b (Fig. 1B). A similar result was

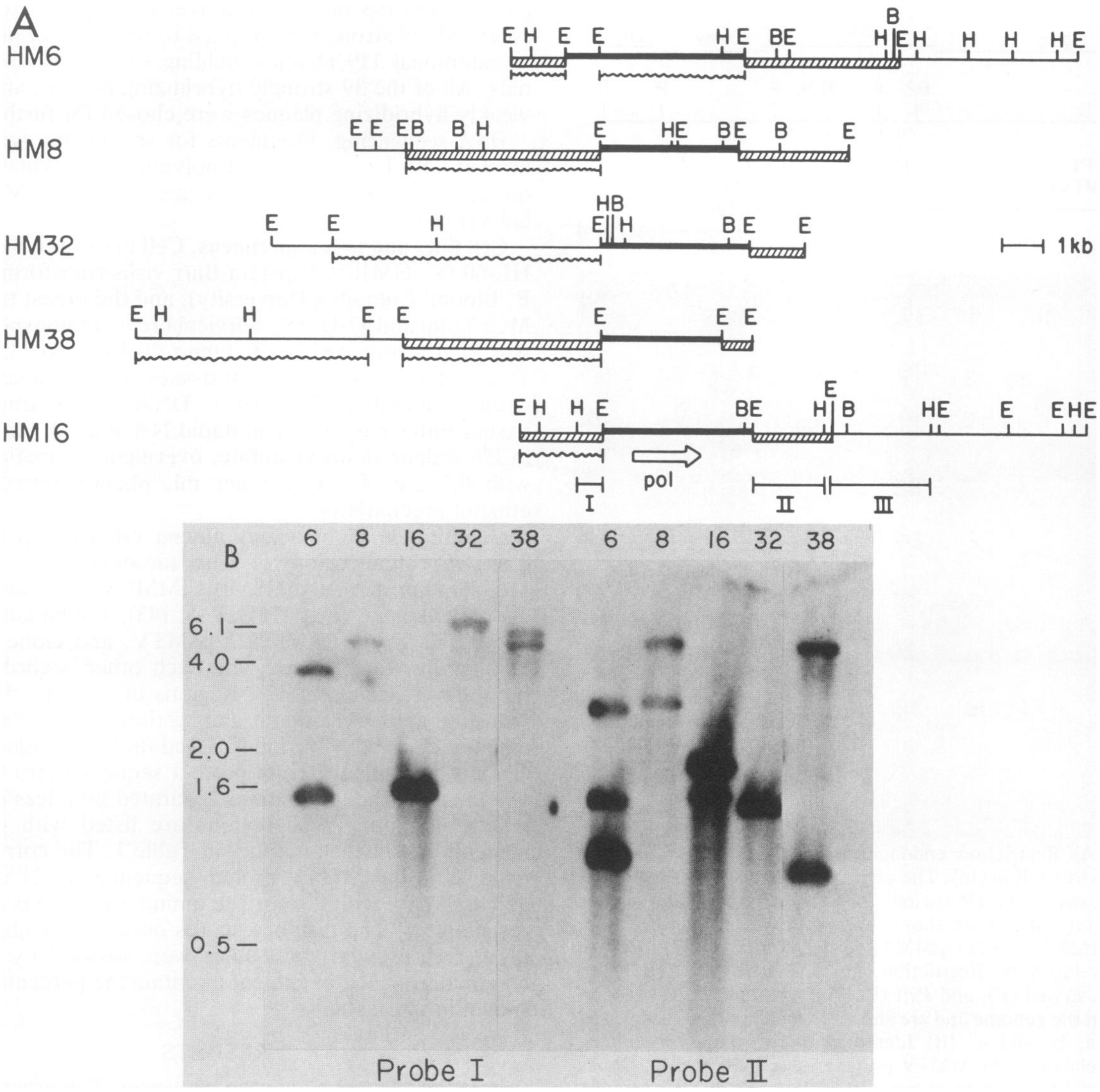


FIG. 2. Restriction enzyme maps and orientation of five unique *pol*-related clones. (A) Restriction enzyme maps were obtained by single and combination digests with the indicated enzymes: *Bam*HI (B), *Eco*RI (E), and *Hin*dIII (H). *Eco*RI fragments which hybridized with MMTV *pol* probes a, c, and d are indicated as solid bars. In HM6, the 0.7-kb *Eco*RI fragment annealed with probe a, and the 3.7-kb fragment annealed with probes c and d. In clone HM8, the 1.9-kb fragment (containing the *Hin*dIII site) annealed with a and c, whereas the 1.6-kb fragment annealed only with d. (B) *Eco*RI digests of each clone were probed under stringent conditions with the 0.9-kb (I) and 1.8-kb (II) *Eco*RI-*Hin*dIII fragments bordering the *pol* region in HM16. The film was exposed for 16 h. Fragments which annealed with probe I or II are indicated in panel A by wavy lines and hatched bars, respectively. The clones were oriented with respect to each other on the basis of which fragment annealed with probe I. This orientation was verified for HM6 and HM8 by the pattern of hybridization of the *pol*-related *Eco*RI fragments to MMTV probes a, c, and d (above) and for HM16 by sequence comparison with MMTV (see Fig. 4). Migration of molecular size markers (in kilodaltons) is indicated to the left of the figure.

obtained by using the same b-region fragment from the C3H exogenous strain of MMTV (13), indicating that this lack of homology was not an artifact associated with the GR40 endogenous provirus (data not shown). We discuss the absence of b-fragment homology in the later section on nucleic acid sequence comparison. Ten clones, enumerated in the legend to Fig. 1, whose relative homology to probes a, c, and d paralleled that of a fragment containing the complete MMTV *pol* gene (Fig. 1B) were selected for further analysis.

Characterization of human *pol* clones. Restriction endonuclease maps were derived for each of the 10 selected clones. Comparison of the maps revealed that these clones were derived from at least five distinct loci. Representative members of each of these loci are shown in Fig. 2. As determined by hybridization with the MMTV *pol* fragments a, c, and d, the region of *pol* homology in each of these clones was limited to a single or two adjacent *Eco*RI fragments spanning 2.9 to 3.8 kb (Fig. 2). A previously described member (HLM-2) of this repeat family, distinct on the basis of restriction maps from the clones in Fig. 2, showed weak homology to MMTV *gag* and *env* probes (6). Thus, we tested the 21 phage clones from the intermediate *pol* screen for homology to the LTR, *env*, and *gag* regions of MMTV GR40 (plasmids pMTVP-1 and pMTVP-4 and fragment e [Fig. 1], respectively). By DNA dot-blot analysis under lowered stringency conditions, about half the clones hybridized with the *gag* fragment, but the intensity was less than 5% of that observed with the *pol* probes (data not shown). None of the

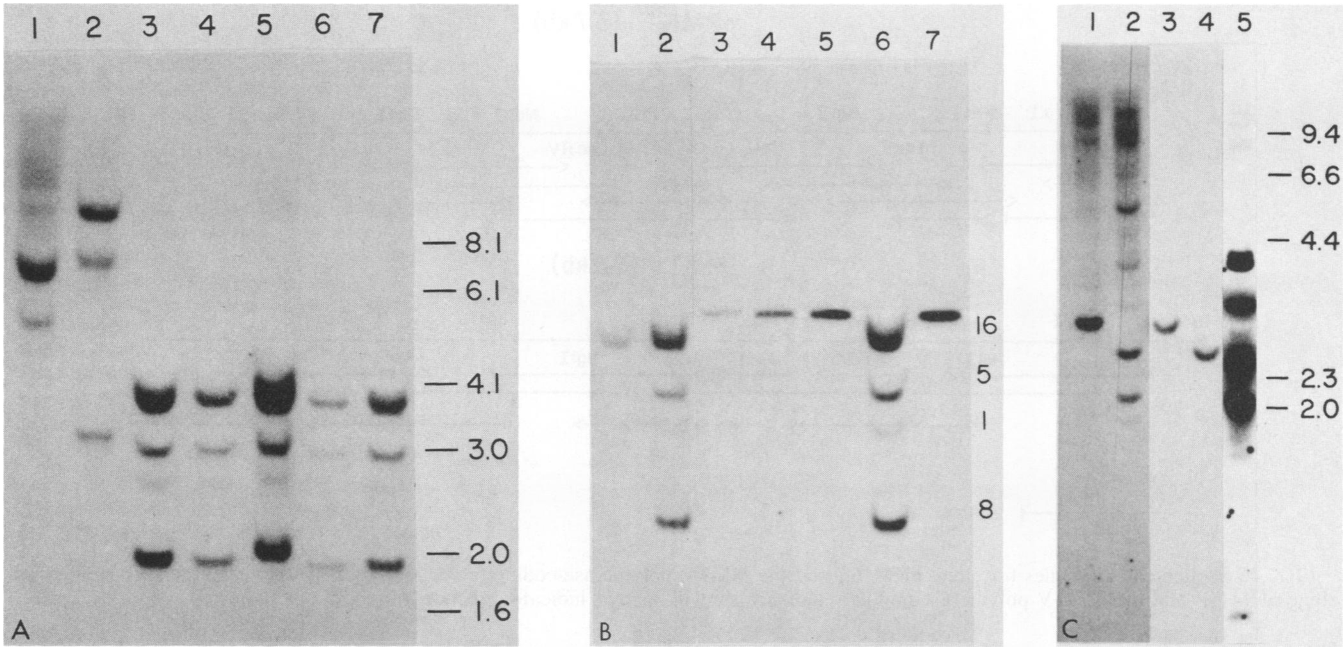


FIG. 3. Identification and quantitation of sequences in human DNA hybridizing with the *pol* and 3′ border fragments of HM16. (A) DNA (10 to 15 μg) from J-1 (lanes 1 to 3) and placental (lane 4) tissues and from the 47-D (lane 5), GM1056 (lane 6), and HeLa (lane 7) cell lines was digested with *Eco*RI (lanes 3 to 7), *Hin*dIII (lane 1) or *Bam*HI (lane 2), electrophoresed on a 0.6% agarose gel, transferred to nitrocellulose, and hybridized under stringent conditions to the 1.5-kb *Eco*RI-*Eco*RV *pol* fragment of pHM16E. The migration of molecular size markers (kilodaltons) is indicated to the right of the figure. (B) DNA from J-1 (20 μg, lanes 2 and 6) and pHM16E (95, 190, 380, and 480 pg in lanes 3, 4, 5, and 7, respectively) was digested with *Eco*RI, electrophoresed on a 0.8% agarose gel, and processed as in panel A. In the presence of 10 μg of *Eco*RI-digested chick carrier DNA, the 3.7-kb *Eco*RI fragment of pHM16E (95 pg) comigrates with the 3.7 to 3.8-kb band of J-1 (lane 1). Assuming a human haploid size of 3.2 × 10^9 bp, the amounts of pHM16E in lanes 3, 4, 5, and 7 correspond to 2, 4, 8, and 10 copies of *pol* sequence per 20 μg of cellular DNA. The estimated copy number of each of the major *pol*-related *Eco*RI fragments of J-1 is indicated on the right margin. From the relative intensities of the 3.7- and 3.8-kb fragments (panel C, lane 5), we estimate that the 3.7/3.8-kb doublet represents about 10 and 6 copies, respectively. Film exposure was 2 days in both panels A and B. (C) DNA (10 μg) from cell line GM1056 (lanes 1, 2, and 5) and clone HM16 (lanes 3 and 4) was digested with *Eco*RI (lanes 1, 3, and 5) or *Hin*dIII (lanes 2 and 4), electrophoresed on a 0.8% agarose gel, and transferred to nitrocellulose. Lanes 1 to 4 were hybridized under stringent conditions to the 2.3-kb *Eco*RI-*Hin*dIII fragment bordering the 3′ repeat region of HM16 (probe III, Fig. 2A). Lane 5 was hybridized under stringent conditions to the 1.5-kb *Eco*RI-*Eco*RV *pol* fragment of pHM16E. Film exposure was 3 days for lanes 1, 2, and 5; 4 h for lanes 3 and 4. Molecular size markers (kilodaltons) are shown on the right.

clones hybridized with the *env* or LTR probes. The HLM-2 clone did not show homology with the LTR of MMTV, but is bounded by two short repeat sequences of about 1 kb (6) which hybridize with the squirrel monkey virus genome (5). Subsequent sequence characterization of these repeats has revealed an LTR-like organization (R. Callahan, personal communication). We searched for such repeat structures among the clones shown in Fig. 2 by probing Southern blots of these clones with the 0.9- and 1.8-kb *Eco*RI-*Hin*dIII fragments bordering the *pol* region of HM16 (probes I and II, Fig. 2). In all but clone HM32, the 1.8-kb probe II annealed to fragments on either side of the *pol* region, indicating the presence of repeat sequences. The spacing of these repeats within 6 to 8 kb around the *pol* region and their lack of hybridization with the common human repeat elements *Alu* and *Kpn* (data not shown) was suggestive of LTR-like sequences in a proviral context. The presence of LTR-like elements within these regions is further suggested by their hybridization with the repeat elements of HLM-2 (R. Callahan, personal communication). Probe I, 5′ to the *pol* region of HM16, annealed with similarly positioned fragments in all five clones (Fig. 2A). The hybridization to two noncontiguous fragments in clones HM6 and HM38 may have resulted from rearrangements in this region. A rearrangement in HM6 is also suggested from a comparison of more detailed restriction maps of the HM6, -16, and -38 *pol* regions. When orientated as described below, these clones share several similarly positioned sites throughout the *pol* region except in the 3′ segment of HM6, which diverges from HM16 and HM38 about 1 kb beyond the internal *Eco*RI site (data not shown).

The orientation of the *pol* region in HM16 was determined through sequence comparison with the MMTV *pol* gene (Fig. 2A). In this alignment, the 0.9-kb *Eco*RI-*Hin*dIII fragment (probe I) of HM16 lies within the *gag* region of the putative provirus. Clones HM6, -8, -32, and -38 were then aligned with HM16 on the basis of: (i) the *Eco*RI fragment bordering the *pol* region which hybridized with probe I (Fig. 2) and (ii) the homology of the MMTV *pol* gene subregion probes (a, c, and d) to *Eco*RI fragments within the *pol* region (clones HM6 and -8; see the legend to Fig. 2).

To determine whether the homology among these clones extended beyond the repeat regions, we used as a probe the 2.3-kb *Eco*RI-*Hin*dIII fragment immediately 3′ to probe II (probe III, Fig. 2). This probe did not anneal with HM6, HM36 (a clone overlapping HM6), or HM8, clones which contain or appear to contain sequences 3′ or 5′ to the repeated elements. Thus, the proviral-like element of HM16 is embedded in different genomic sequences than those of HM6 and -8.

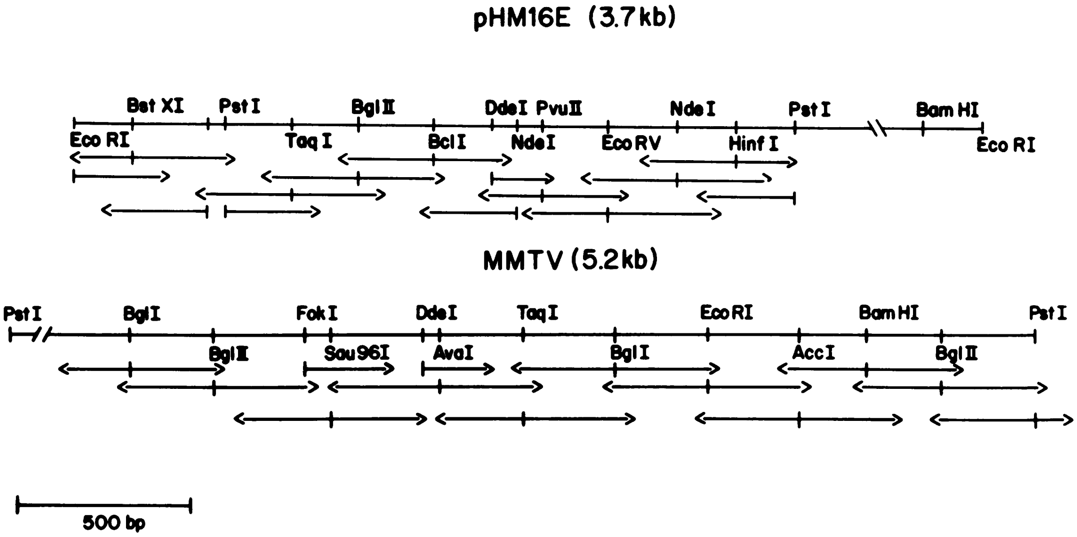


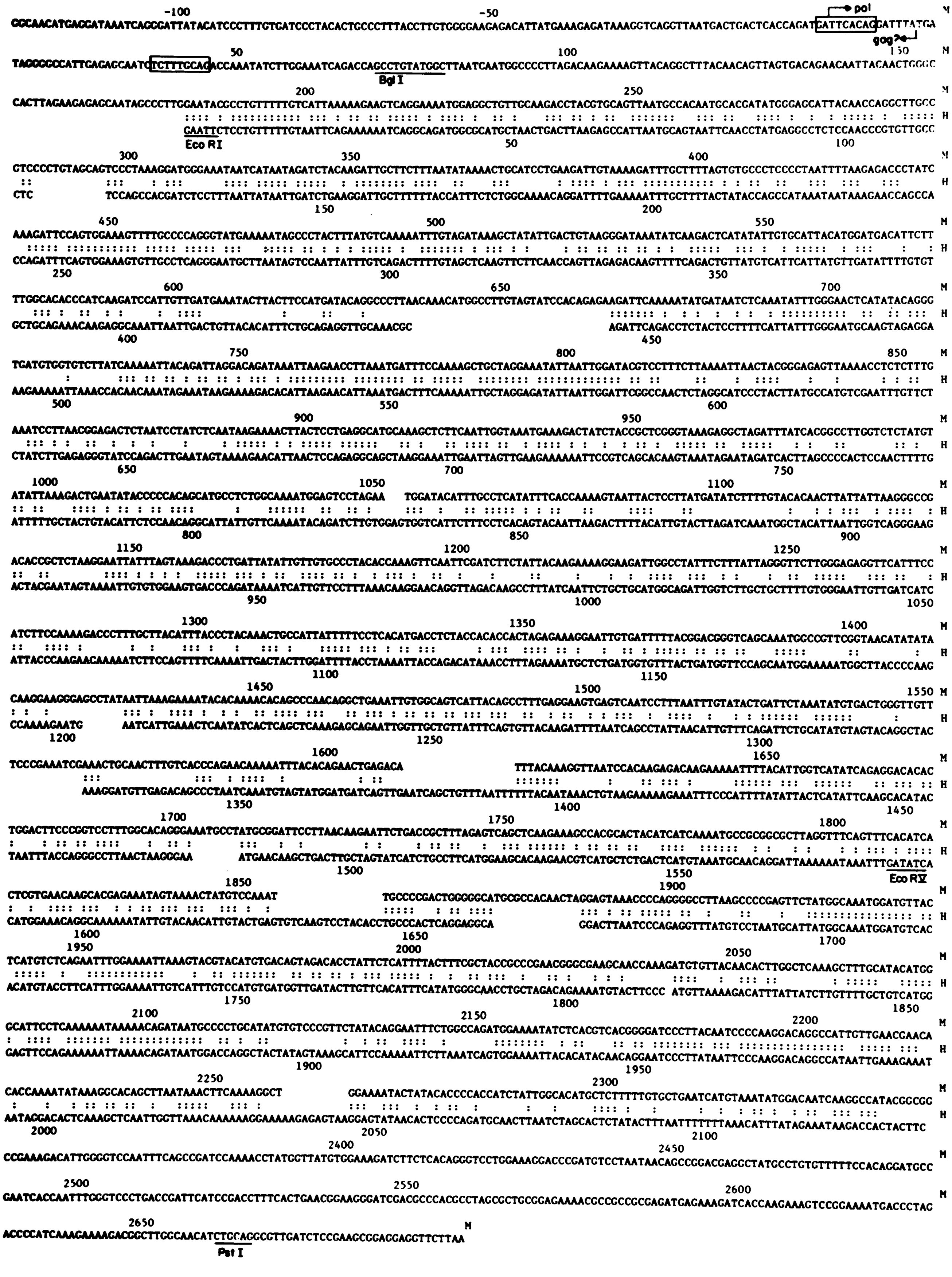
FIG. 4. Sequencing strategies for clone pHM16E and the MMTV polymerase-coding region. Restriction sites used for sequencing both clone pHM16E and the MMTV polymerase-coding region are shown. Arrows indicate direction and extent of sequencing from each site.

Multiplicity of the HM family. To examine the heterogeneity and copy number of the MMTV *pol*-related sequences in human DNA, we utilized a 1.5-kb *Eco*RI-*Eco*RV fragment from the *pol* region of HM16 (see Fig. 4) as a probe. This probe spans most of the *pol* region but lacks the 3′ sequences defined by MMTV fragment d (Fig. 1) and does not detect fragments such as the 3′ 1.6-kb *Eco*RI segment of HM8. Major *pol* fragments of 3.7 to 3.8, 2.9, 2.6, and 1.9 kb as well as several weak bands are apparent in *Eco*RI digests of human genomic DNA (Fig. 3A). Overexposure of this and other DNA blots revealed at least 15 distinct fragments (Fig. 3C, lane 5). This same pattern was observed in each of the five human DNA samples we examined (Fig. 3A). Similar fragments are detected with MMTV genomic or *gag-pol*-region probes under conditions of reduced stringency (6, 27; unpublished observations). Most of these major *Eco*RI fragments are represented in the clones described in Fig. 2; for example, the 3.7-, 3.8-, and 1.9-kb bands comigrate with the *pol* fragments of HM16 (Fig. 3B), HM6, and HM8 (data not shown), respectively. Thus, we presume that each contains a relatively complete *pol* region. By quantitation relative to the cloned HM16 *pol* region, these fragments each correspond to between 1 and 16 copies of *pol* sequence, as indicated in the margin of Fig. 3B. Similar results were obtained for digests of genomic DNA with other restriction enzymes. A family of *pol*-related fragments of various intensities was seen in both *Hin*dIII and *Bam*HI digests (Fig. 3A). Many of the major *Hin*dIII fragments were also represented in the *pol* regions in our phage clones (e.g., the 6-kb band comigrates with the *pol* fragment of HM16). The placement of *Bam*HI sites in our clones precludes a similar comparison with the genomic *Bam*HI digests. From these results we estimate that the HM family contains about 30 to 40 copies with relatively complete *pol* regions and that they fall within several groups based on common restriction patterns.

To further investigate the genomic organization of the *pol* family represented by HM16, we probed genomic digests with probe III (Fig. 2A) which borders the 3′ repeat region in HM16. Digestion with *Eco*RI resulted in a prominent 2.7-kb band which comigrated with the corresponding fragment of HM16 (Fig. 3C). By quantitation relative to a plasmid clone, we estimate that there are about eight copies of this fragment per haploid genome (data not shown), a value similar to that for the 3.7-kb *pol* segment of HM16 (see the legend to Fig. 3B). Similarly, digestion with *Bam*HI plus *Eco*RI gave rise to a prominent 2.4-kb fragment as predicted from the map of HM16 (data not shown). In contrast, digestion with *Hin*dIII gave the predicted 2.5-kb fragment, but in addition, fragments of 4.7, 3.6, and 2.1 kb as well as several weakly hybridizing fragments were also evident (Fig. 3C). The intensities of each of the major *Hin*dIII bands were about one-third that of the 2.7-kb *Eco*RI band. The multiple *Hin*dIII fragments probably do not reflect allelic polymorphism since identical patterns were observed with two other human DNA samples. These results indicate that the multiplicity of the 3.7-kb *Eco*RI family stems in part from duplication of the HM16 provirus-like element together with the flanking sequence. The members of this family are not identical, as indicated by heterogeneity in the *Hin*dIII sites in the region flanking the 3′ repeat. Finally, we note that probe III displayed weak hybridization to multiple fragments in each of the above genomic digests, suggesting that this region contains a moderately repeated sequence. This sequence is not related to the *Alu* or *Kpn* families since probes to these families do not hybridize to HM16.

Sequencing and comparison of HM16 and MMTV *pol* genes.

FIG. 5. Nucleotide sequence and alignment of clone pHM16E and the MMTV polymerase-coding region. Sequences of clone pHM16E and the polymerase-coding region of MMTV are shown aligned by the program of Wilbur and Lipman (53), using a *k*-tuple size of 3, a window size of 20, and a gap penalty of 7. M, MMTV; H, human sequence. Breaks in the sequences represent gaps created by the alignment program. The ORF of the MMTV *pol* gene begins at nucleotide 1 and continues to the end of the sequence. Boxed regions indicate potential splice acceptor sites at the 5′ end of the MMTV sequence.


pol
gag
Bgl I
Eco RI
Eco RV
Pst I
M
H

To address the organization and potential functionality of the human *pol* sequence, we sequenced the *pol* region of HM16 and, for comparison, the complete *pol* gene of the MMTV endogenous GR40 provirus. The 3.7-kb *Eco*RI *pol* fragment of HM16 was subcloned into the *Eco*RI site of pBR322 to yield plasmid pHM16E. Restriction nuclease mapping and Southern blot analysis with the MMTV *pol* fragments a, c, and d further localized the region of *pol* homology in pHM16E to a 2.1-kb *Eco*RI-*Pst*I fragment (Fig. 4). The restriction maps and sequencing strategy for the 2.1-kb human and 2.7-kb MMTV *pol* genes are shown in Fig. 4. The complete sequences are compared in Fig. 5.

Alignment of the human and MMTV polymerase sequences by the program of Wilbur and Lipman (53) revealed an overall nucleotide match of 52%. Several regions of strong homology were apparent in both the 5′ and 3′ thirds of the human sequence, whereas none occurred in the middle portion. This localization of homology is consistent with the failure of MMTV probe b to hybridize with the phage clones shown in Fig. 1. In each of the regions aligning with probes a, c, and d, the overall nucleotide homology was 54 to 56%, whereas in the region aligning with probe b, the homology was only 46%.

Comparison of the predicted human and MMTV *pol* proteins. More relevant to the potential function of these genes is a comparison of their encoded protein products. The MMTV *pol* region has a single open reading frame (ORF) encoding 899 amino acids extending between termination codons at −3 and 2698 in the sequence in Fig. 5. The 121-base-pair (bp) sequence upstream of the 5′ termination site is open in only one reading frame, and this frame differs from that of the *pol* gene. This presumed *gag* reading frame overlaps 13 bp of the *pol* region and terminates with a tandem (TGATAG) stop signal. Thus, in MMTV, the *gag* and *pol* coding sequences are in different reading frames and are further interrupted by termination signals.

In contrast, the 2.1-kb human sequence contains 47, 42, and 23 termination signals in reading frames 1, 2, and 3 respectively. However, each frame contains a single major ORF: bp 93 to 1202, 370 amino acids in frame 3; bp 1460 to 1795, 112 amino acids in frame 2; and bp 1765 to 2079, 105 amino acids in frame 1. These three major ORFs are largely nonoverlapping and together span most of the human sequence. Comparison of the human sequence, translated in all three reading frames, with the predicted MMTV *pol* protein identified regions of homology which largely coincided with these ORFs. These results suggested that the putative human *pol* gene was extensively disrupted by frameshift mutation. To facilitate protein comparisons, nucleotide insertions or deletions were introduced into the *pol* region of HM16 at three positions (see the legend to Fig. 6) to produce an essentially contiguous protein sequence. The resultant composite human protein sequence and its comparison to the predicted MMTV and published RSV polymerases are shown in Fig. 6.

For HM16 and MMTV, the overall amino acid homology is 52%. Consistent with the nucleic acid alignment, the amino acid homology is markedly clustered in the amino and carboxyl termini. The homology (including conservative substitutions) in these regions, as well as in the less-conserved middle region, is quantitated in Table 1. Segments of greater than 70% identity occur at residues 1 through 26, 46 through 55, 72 through 93, and 174 through 196 in the amino domain and at residues 557 through 593, 615 through 628, and 644 through 664 in the carboxyl domain of HM16.

Comparison with other retroviral *pol* protein sequences. The predicted MMTV and composite HM16 *pol* protein sequences were compared with those of mammalian (MMLV) are avian (RSV) type C and HTLV I and LAV retroviruses. As summarized in Table 1, conservation of the terminal domains is common to all these *pol* genes, but both HM16 and MMTV have the greatest homology with RSV. The protein sequence alignment of RSV *pol* with HM16 and MMTV is shown in Fig. 6. Within the terminal domains, the pattern of conserved residues closely resembles that between HM16 and MMTV. Others have pointed out blocks of conserved sequences in comparisons of retroviral *pol* genes with each other (7, 32, 40, 49) and with other polymerases which transcribe RNA to DNA (41, 49). However, alignment of all the *pol* genes listed in Table 1 reveals only a few invariant amino acids. The largest block of consecutive identical or closely conserved amino acids is four, and this occurs twice, at positions 153 through 156 and 700 through 703 of the MMTV sequence (Fig. 6).

DISCUSSION

Phage clones hybridizing to the MMTV *pol* gene were isolated from a human phage library. These clones contained sequences homologous to the amino or carboxyl regions or both of MMTV *pol*, although none annealed with a probe from the middle of the MMTV gene. Ten clones with extensive homology to the MMTV *pol* region were further characterized. These clones defined five distinct genomic loci and contained most of the major MMTV *pol*-related *Eco*RI fragments identified in the human genome. The *pol* region at one locus, contained in phage HM16, and the complete MMTV *pol* gene were sequenced and compared. MMTV *pol* contained a single ORF encoding 899 amino acids and was in a different reading frame from the upstream, putative *gag* region. The human *pol* sequence was multiply interrupted in all three reading frames, but a composite protein sequence was derived by alignment of the ORF with MMTV *pol*. Features of the human and MMTV *pol* genes are discussed below.

Human HM *pol* gene family. Recent characterization of one member of this family, HLM-2, revealed a proviral structure which displayed a mosaic pattern of homology to different regions of type A, B, and D retroviruses (5). The *pol* sequences at four of the loci we describe in Fig. 2 are bounded within 6 to 8 kb by sequences homologous to the LTR element of HLM-2, and hence, we presume that they are also proviral in nature. The fifth locus appears to be rearranged in this respect. From comparison of restriction patterns, HLM-2 is not included among our five representative clones but does share several similarly positioned sites with HM16. The provirus-like sequences at the five loci have dissimilar restriction maps, but they are closely related as judged by the homology among their repeat regions and within and 5′ to the *pol* region. However, sequences outside the 3′ repeat in HM16 do not occur in HM6 or -8. The different *Eco*RI fragments encompassing the *pol* regions of

FIG. 6. Comparison of MMTV, RSV, and clone pHM16E (composite) amino acid sequences. For alignment purposes, a composite protein sequence for pHM16E was generated by deleting G at nucleotide 122 and inserting a T after nucleotides 1473 and 1793 of the sequence shown

```
    1                                                                                   *               * RSV
     TVALHL AIPLKWKPDHTPVWIDQWPLPEGKLVALTQLVEKELQLGHIEPSLSCWNTPVFVIRKASGSYRLLHDLRAVNAKLVPFGAVQQGAPVLSALPR
   7      * *    ** *  ***  ****   ** ** ***   ***** * * * ******** * **  *** *******     ** * * * * *   MTV
    GAIESNLFADQISWKSDQ PVWLNQWPLRQEKLQALQQLVTEQLQLGHLEESNSPWNTPVFVIKKKSGKWRLLQDLRAVNATMHDMGALQPGLPSPVAVPK
                                                           1* ***** **** ** * **** **    *  *** ****   *   HUM
                                                            NSPVFVIQKKSGRWRMLTDLRAINAVIQPMRPLQPVLPSP   PR

  100  **     *        * *  *    *     *   *                 *      *** *                     *     *      RSV
     GWPLMVLDLKDCFFSIPLAEQDREAFAFTLPSVNNQAPARRFQWKVLPQGMTCSPTICQLVVGQVLEPLRLKHPSLCMLHYMDDLLLAASSHDGLEAAGE
  107**     ** **** * *   *   ***  ** *   *  ***********  *** **   *       * *       ***** ***   *          MTV
     GWEIIIIDLQDCFFNIKLHPEDCKRFAFSVPSPNFKRPYQRFQWKVLPQGMKNSPTLCQKFVDKAILTVRDKYQDSYIVHYMDDILLAHPSRSIVDEILT
   43    ***** **** * *   *   ***  *  * * *  *********** *** **  **      **** * *   ** * ** *     *       * HUM
     S PLIIIDLKDCFFTISLAKQDFEKFAFTIPAINNKEPATRFQWKVLPQGMLNSPIICQTFVAQVLQPVRDKFSDCYVIHYVD ILCAAETRGKLIDCYT

  200                                      *                              ***   *       *   * *             RSV
     EVISTLERAGFTISPDKVQREPGVQYLGYKLGSTYVAPVGL VAEPRIATLWDVQKLVGSLQWLRPALGIPPRLMGPFYEQLRG SDPNEAREWNLDMKM
  207  *  *   *   *  * *      ***       *     *        ** * *** *   * ** *        *  * * * * *   *          MTV
     SMIQALNKHGLVVSTEKIQKYDNLKYLGTHIQGDVVSYQKLQIRTDKLRTLNDFQKLLGNINWIRPFLKLTTGELKPLFEILNGDSNPISIRKLTPEACK
  141   * *        *          ***          *   ** * ************ ****** *   *     ** ** *     * * ***** * HUM
      FLQRLQTQ  I QTST P F H  YLGMQVEERKIKPQQIEIRKDTLRTLNDFQKLLGDINWIRPTLGIPTYAMSNLFSILRGYPDLNSKRTLTPEAAK

  298                        *            *                        *   *               *   *                RSV
     AWREIV RLSTTAALERWDPALPLEGAVARCEQGAIGVLGQGLSTHPRPCLWLFSTQPTKAFTAWLEVL TLLITKLRASAVRTFGKEVDILLLP ACFR
  307*      **** *   * *   *        *      * *           *    * *  *        * ** * *      * *  *    *        MTV
     ALQLVNERLST ARVKRLDLSRPWSLCILKTEYTPTACLWQN GVLEW  IHLPHISP KVITPY DIFCTQLIIKGRHRSKELFSKDPDYIVVPYTKVQ
  232   ** *     * * * *   *  * *  *   **     **    **      **    *  * * *   * **  ** *   *   *** * **  * * HUM
     EIELVEEKIPS AQVNRIDHLAPLQLLIFATVHSPTGIIVQNTDLVEW  SFFPHSTI KTFTLYLDQMAT LIGQGRLRIVKLCGSDPDKIIVPLNKEQ

  395                               **         *              *     *   *                                   RSV
     EDLPLPE GIL LALKGFAGKIRSS DTPSIFDIARPLHVSLKVRVTDHPVP GPTVFTDASSSTHKGVVVWREGPRWEIKEIADLGASVQQLEARAVAM
  401 ** * *       * ** *                             *   *   *** *      *     *    ***        ** *  **    MTV
     FDLLLQEKEDWPISLLGFLGEVHFHLPKDPLLTFTLQTAIIFPHMTSTTPLEKGIVIFTDGSANG RSVT YIQGREPIIKENTQ NTA QQAEIVAVIT
  327           * *  * * * *  * *      *   *  * *  *   ***     ***** **      *    * **        * * ** ****  HUM
     VRQAFINSAAWQIGLAAFVGIVDHHYPRTKIFQFSKLTTWILPKITRHKPLENALMVFTDGSSNG K MA YPKPKE■IIETQYH S A QRAELVAVIS

  491 *      *   *   *                *                                      *                  *           RSV
     ALLLWPTPTNVVTDSAFVAKMLLKMGQEGVP STAAAFI L E  DALSQ RSAMAAVLHVRSHSEVPGFFTEGNDVADSQATFQAYPLREAKDLHTAL
  497*         *  ***  *                      *      *   *       * * *   **    **  ***        *  *   *     MTV
     AFEE VSQSFNLYTDSKYVTGLFPEIETATLS PRTKIYTELRH LQRLIHKRQEKFYIGHIRGHTGLPGPLAQGNAYADSLTRILT ALESAQESHALH
  421       *  *   ** **       ***             *   **    **   *** **  ** *****  ** **  * *      *   *** ***  HUM
     VLQD FNQPINIVSDSAYVVQATKDVETALIKCSMDDQLNQLFNFLQ■TVRKRNFPFYITHIQAHTNLPGPLTKGNEQAD LL VSS AFMEAQERHALT

  585                      *              *                    *     *            *      *                  RSV
     HIGPRALSKACNISMQQAREVLQTCPHC NSAPALE AGVNPRGLGPLQIWQTDFTLEPRMAPRSWLAVTVDTASSAIVVTQHGRVTSVAVQHHWATAIA
  593*    **     *   ****    ** *     *    ****** *   ** * *             ***** *     *         *  * *   * MTV
     HQNAAALRFQFHITREQAREIVKLCPNCPDWGHAPQ LGVNPRGLKPRVLWQMDVTHVSEFGKLKYVHVTVDTYSHFTFATARTGEATKDVLQHLAQSFA
  517* **  *   * **  **  **  *  *       *  * ***** *  *********  ****  *** *** ***  **      *   *  **   ** HUM
     HVNATGLKNKFDITWKQAKNIVQHCTECQVLHLPTQEAGLNPRGLCPNALWQMDVTHVPSFGKLSFVHVMVDTCSHFIWATLLDRKCTSHVKRHLLSCFA

  683*           *     *             *      *         ** *                                                  RSV
     VLGRPKAIKTDNGSCFTSKSTREWLARWGIAHTTGIPGNSQGQAMVERANRLLKDRI R VLAEGDGFMKRIPTSKQGELLAKAMYALNHFERGENTKTP
  692  * * *****     * *  * **** * * **** * **** ***     *       *          *      *** *   ***        *   MTV
     YMGIPQKIKTDNAPAYVSRSIQEFLARWKISHVTGIPYNPQGQAIVERTHQNIKAQL N KLQKAGKYYT  P   H HLLAHALFVLNHVNMDNQGHTA
  617 ** * ****** * * *   * **  *** * ****** ***** **     ****    *  *     *  *      *  *                   HUM
     VMGVPEKIKTDNGPGYYSKAFQKFLNQWKITHTTGIPYNSQGQAIIERNNRTLKAQLVKQKRKKRVRSITL P   RCNLI■HSIL■FF■TFIEIRPLL

  781                                                                                                       RSV
     IQKHWRPTVLTEGPPVKIR IETGEWEKGWNVLVW GRGYAAVKNRDTDKVIWVPSRKVKPDITQKDEVTKKDEASPLFAGISDWIPWEDEQEGLQGETA
  784   ** *      * *      ** * **  **   ***** *   *    **** *   *  *     *    *             *              MTV
     AERHWGPISADPKPMVMWKDLLTGSW KGPDVLITAGRGYACVFPQDAESPIWVPDRFIRP FTERKGSTPTPSAAEKTPPRDEKDHQESPENDPRPHQR

  879                    RSV
     SNKQERPGEDTLAANES■
  882           *        MTV
     KDGLATSAGVDLRSGGGS■
```

in Fig. 4. The corresponding positions in the protein sequence are circled in this figure. This composite sequence and the predicted protein sequences of the polymerase-coding region of MMTV and RSV (42) were aligned by using the program of Kanehisa (17). Breaks in the sequences represent gaps introduced by the alignment program. Exact amino acid matches between sequences are indicated by an asterisk. Matches between RSV and the human sequence which do not also match the MMTV sequence are indicated by an asterisk above the RSV sequence. Black boxes represent stop codons.

TABLE 1. Percent amino acid sequence conservation among retroviral polymerase-coding regions[a]

Region	Nucleotides	Virus	% Sequence conservation with:				
			MMTV	RSV	HTLV I	MMLV	LAV
Amino	61–702	MMTV		55	38	29	30
	2–478	HM16	51	45	38	18	31
	2539–3138	RSV	55		40	33	32
	2584–3201	HTLV I	41	44		32	30
	2757–3416	MMLV	29	32	29		23
	2153–2788	LAV	30	30	28	23	
Middle	703–1830	MMTV		12	11	5	8
	479–1603	HM16	40	14	10	4	8
	3184–4308	RSV	12		8	6	4
	3202–4377	HTLV I	11	8		5	5
	3417–4871	MMLV	4	4	4		4
	2789–3874	LAV	8	4	6	5	
Carboxyl	1831–2253	MMTV		38	27	26	20
	1604–2026	HM16	56	36	24	25	20
	4309–4722	RSV	40		20	17	15
	4378–4746	HTLV I	31	23		29	14
	4872–5309	MMLV	25	16	24		13
	3875–4270	LAV	22	15	13	14	

[a] See Materials and Methods. The comparisons are presented horizontally with the identity and position of the sequence listed in the left column.

these clones account for four of the five major *pol*-related *Eco*RI fragments observed in digests of human DNA. However, most of these fragments are present in multiple copies and in toto represent 30 to 40 copies of relatively intact *pol* genes per haploid genome. Utilizing the 3′ border fragment of HM16, we observed that this multiplicity extended beyond the provirus-like element into flanking sequence. It thus appears that the HM family was generated by a combination of independent integrations or transpositions of a provirus-like element, subsequent genomic duplication of some regions encompassing these elements, and other rearrangements. Given the multiplicity of the HM family, the isolation of identical or apparently overlapping clones in our library screen is puzzling. At present, we presume this resulted from the preferential growth of some phage clones during amplification of the library.

It seems unlikely that any of the *pol* genes of the HM family encode a functional protein. The *pol* sequence of HM16 is multiply interrupted, and in many other clones the *pol* gene regions are rearranged (K. Deen, unpublished observations). Also, the recently sequenced *pol* region of HLM-2 contains termination signals in all three reading frames (R. Callahan, personal communication). Comparison of HM16 with the published sequence of the 3′ region of HLM-2 (5) revealed a 90% nucleotide homology which, however, translates into only a 66% amino acid match. The multiple termination signals, rearrangements, and protein sequence misalignments are consistent with the mutational decay of nonconserved provirus-like elements. This extensive mutation, together with the common restriction patterns observed for *pol* and flanking sequences in all human samples, implies that these elements are ancient residents of the genome, perhaps predating the evolution of the human species.

MMTV *pol* gene. The sequence of the 3′ half of the MMTV polymerase-coding region was previously reported (7, 35). The sequence determined here closely agrees with that of Redmond and Dickson (35) with the exception of insertions of a CC doublet at position 1914, a G at position 2011, and a second CC doublet at position 2435. The first two insertions, which are in agreement with the sequence of Chiu et al. (7), result in a frameshift, involving 32 amino acids, from that predicted by the Redmond and Dickson sequence. The third insertion creates a frameshift which extends the polymerase ORF 164 nucleotides beyond the previously reported termination codon at position 2535 (35). This frameshift does not affect the predicted ORF of the MMTV envelope gene which begins at position 2486 (23). With this additional 164 nucleotides, the total ORF of the MMTV polymerase-coding region is 2697 nucleotides, which could encode an 899-amino acid protein with a calculated molecular size of 102 kilodaltons (kDa).

The reverse transcriptase isolated from MMTV is reported to be a single polypeptide of 100 kDa (11) or a complex of a 90 kDa subunit and a 50 kDa subunit (19, 24). Polymerase and RNase H activities of the enzyme have been characterized, but they have not been mapped within the MMTV *pol* gene. However, the conserved homology among MMTV and other retroviral *pol* genes (Table 1) permits tentative deductions about the functional organization of MMTV *pol*. On the basis of the relatively close homology to the RSV *pol* gene, we suggest that the amino terminus of the MMTV reverse transcriptase is positioned about 3 bases downstream of the *gag* reading frame tandem termination codon. The RNase H active site probably lies within the strongly conserved initial 200 to 250 residues, whereas the polymerase site probably encompasses or is strongly influenced by additional downstream sequences (20, 22). An endonuclease activity has not been reported for MMTV; however, a region homologous to the endonuclease domain of RSV (15) begins at about residue 585 and includes the strongly conserved 3′ regions of the *pol* gene. Because the RSV and MMTV sequences are not conserved at the amino terminus of the RSV endonuclease (proline at position 573 in Fig. 6), we cannot predict where or whether the carboxyl domain is cleaved in MMTV. However, processing as in RSV would yield a 66-kDa amino-terminal fragment which together with the predicted full-length 102-kDa polypeptide is similar to reported subunits of MMTV reverse transcriptase (19, 24).

MMTV *gag-pol* junction. In retroviruses, the *pol* protein is translated as a *gag-pol* precursor from genomic-length RNA (9). However, an emerging common feature among retroviruses and provirus-like transposable elements is a translation block between the major *gag* and *pol* reading frames (reviewed in reference 50). The position and nature of this block vary in different retroviruses. Translational suppression or cryptic RNA splicing have been proposed as solutions to the dilemma of these internal termination signals (42, 45). In either case, the process must be relatively inefficient because the *gag* precursor is synthesized at much higher levels than the *gag-pol* polyprotein (9). Translation of the putative *gag-pol* junction in MMTV requires suppression of at least one stop codon as well as a shift in reading frame. The single in-frame stop codon intervening between *gag* and *pol* in MMLV is suppressed in vivo (54). Also, inefficient in vitro translation of a *gag-pol* fusion product has been reported with viral RNA from MMTV (10) and RSV (T. Jacks and H. Varmus, personal communication). Since in RSV, as in MMTV, *gag* and *pol* are in different reading frames, these results suggest that the *gag-pol* polyprotein is synthesized via frameshift suppression, for which precedent exists in bacteria and yeasts (reviewed in reference 39). Alternatively, the translational block could be removed by splicing out a short RNA segment. For example, a minor RNA species in which a 282-bp deletion within the *gag* region created a new ORF was recently discovered in RSV-infected cells (4). A search for consensus splice signals in our MMTV sequence revealed possible acceptor sites at the 5′ end of *pol*, but there are no appropriately positioned donor sites within the 121 bp that we sequenced upstream of the *pol* ORF (Fig. 5).

Retroviral *pol* gene phylogeny. The *pol* gene is the most conserved sequence among retroviruses, and this homology is particularly concentrated in the amino- and carboxyl-terminal regions (Table 1). Previous comparisons of the conserved carboxyl sequences suggested a divergence of the mammalian type C viruses from other retroviruses (7, 40) and led to a proposed phylogenetic tree (40). The data summarized in Table 1 extend these comparisons to the more conserved amino-terminal region and include the recently reported LAV (HTLV III) sequence (34, 52). The patterned relationship suggested by the carboxyl (endonuclease) region (40) is generally sustained by the relative homologies at the amino terminus. MMTV *pol* is most similar to the deduced composite sequence for HM16 as emphasized by the homology within the middle region. Among other retroviruses, the closest relatives of MMTV are the avian type C viruses (RSV) and the mammalian type D squirrel monkey virus (5). HTLV I is more distant, followed by the mammalian type C viruses (MMLV) and LAV, which appear as distant from each other as either is from MMTV. These results, together with published data (7, 40), suggest that there are at least three divergent groups of retroviruses: (i) mammalian types A, B (MMTV), and D (squirrel monkey virus), avian type C, and HTLV I plus its close relative, bovine leukemia virus (40); (ii) mammalian type C; and (iii) LAV (HTLV III). However, this conclusion must be tempered by the relatively low homology within the conserved domains, the limited size of these domains, and the potentially chimeric nature of retroviruses.

ACKNOWLEDGMENTS

We sincerely thank R. Callahan for his helpful discussions and communication of results before publication, J. Majors for providing the C3H MMTV *gag-pol* clone, and T. Ohno and R. Nayak for their participation in the early stages of this work. We are especially grateful to M. Rosenberg for his support of this project, and for his and J. Young's critical reading of this manuscript.

LITERATURE CITED

1. **Baltimore, D.** 1985. Retroviruses and retrotransposons: the role of reverse transcription in shaping the eukaryotic genome. Cell **40**:481–482.
2. **Benton, W. D., and R. W. Davis.** 1977. Screening λgt recombinant clones by hybridization to single plaques *in situ*. Science **196**:180–182.
3. **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.
4. **Broome, S., and W. Gilbert.** 1985. Rous sarcoma virus encodes a transcriptional activator. Cell **40**:537–546.
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 **79**:5503–5507.
7. **Chiu, I.-M., R. Callahan, S. R. Tronick, J. Schlom, and S. A. Aaronson.** 1984. Major *pol* gene progenitors in the evolution of oncoviruses. Science **223**:364–370.
8. **Collins, S. J., R. C. Gallo, and R. E. Gallagher.** 1977. Continuous growth and differentiation of human myeloid leukaemic cells in suspension culture. Nature (London) **270**:347–349.
9. **Dickson, C., R. Eisenman, H. Fan, E. Hunter, and N. Teich.** 1982. Protein biosynthesis and assembly, p. 541–601. *In* R. A. Weiss, N. Teich, H. Varmus, and J. Coffin (ed.), RNA tumor viruses: molecular biology of tumor viruses. Cold Spring H arbor Laboratory, Cold Spring Harbor, N.Y.
10. **Dickson, C., and G. Peters.** 1981. Protein coding potential of mouse mammary tumor virus genomic RNA as examined by in vitro translation. J. Virol. **37**:36–47.
11. **Dion, A., A. B. Vaidya, G. S. Fout, and D. H. Moore.** 1974. Isolation and characterization of RNA-directed DNA polymerase from a B-type RNA tumor virus. J. Virol. **14**:40–46.
12. **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.
13. **Dudley, J. P., and H. E. Varmus.** 1981. Purification and translation of murine mammary tumor virus mRNA's. J. Virol. **39**:207–218.
14. **Fasel, N., E. Buetti, J. Firzlaff, K. Pearson, and H. Diggelmann.** 1983. Nucleotide sequence of the 5′ noncoding region and part of the *gag* gene of mouse mammary tumor virus; identification of the 5′ splicing site for subgenomic mRNAs. Nucleic Acids Res. **11**:6943–6955.
15. **Hippenmeyer, P. J., and D. P. Grandgenett.** 1984. Requirement of the avian retrovirus pp32 DNA binding protein domain for replication. Virology **137**:358–370.
16. **Hynes, N. E., N. Kennedy, U. Rahnsdor, and B. Groner.** 1981. Hormone-responsive expression of an endogenous proviral gene of mouse mammary tumor virus after molecular cloning and gene transfer into cultured cells. Proc. Natl. Acad. Sci. USA **78**:2038–2042.
17. **Kanehisa, M. I.** 1982. Los Alamos sequence analysis package for nucleic acids and proteins. Nucleic Acids Res. **10**:183–196.
18. **Keydar, I., L. Chen, S. Karby, F. R. Weiss, J. Delarea, M. Radu, S. Chaitcik, and H. J. Brenner.** 1979. Establishment and characterization of a cell line of human breast carcinoma origin. Eur. J. Cancer **15**:659–670.
19. **Kohno, M., and A. Ishihama.** 1979. Purification and properties of RNA-dependent DNA polymerase from cytoplasmic A-type particles of murine mammary tumor virus. Eur. J. Biochem. **97**:257–266.
20. **Lai, M.-H. T., and I. M. Vera.** 1978. Reverse transcriptase of RNA tumor viruses. V. In vitro proteolysis of reverse transcriptase from avian myeloblastosis virus and isolation of a polypeptide manifesting only RNase H activity. J. Virol. **25**:652–663.
21. **Lawn, R. M., E. F. Fritsch, R. C. Parker, G. Blake, and T. Maniatis.** 1978. The isolation and characterization of α- and β-globin gene from a cloned library of human DNA. Cell **15**:1157–1174.
22. **Levin, J. G., S. C. Hu, A. Rein, L. I. Messer, and B. I. Gerwin.** 1984. Murine leukemia virus with a frameshift in the reverse transcriptase coding region: implication for *pol* gene structure. J. Virol. **51**:470–478.
23. **Majors, J. E., and H. E. Varmus.** 1983. Nucleotide sequencing of an apparent proviral copy of *env* mRNA defines determinants of expression of the mouse mammary tumor virus *env* gene. J. Virol. **47**:495–504.
24. **Marcus, S. L., N. H. Sarkar, and M. J. Modak.** 1976. Purification and properties of murine mammary tumor virus DNA polymerase. Virology **71**:242–254.
25. **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.
26. **Maxam, A. M., and W. Gilbert.** 1980. Sequencing end-labeled DNA with base-specific chemical cleavages. Methods Enzymol. **65**:499–560.
27. **May, F. E. B., B. R. Westley, H. Rochefort, E. Buetti, and H. Diggelmann.** 1983. Mouse mammary tumor virus related se-

quences are present in human DNA. Nucleic Acids Res. **11**:4127–4139.

28. **Noda, M., M. Kurihara, and T. Takano.** 1982. Retrovirus-related sequences in human DNA: detection and cloning of sequences which hybridize with the long terminal repeat of baboon endogenous virus. Nucleic Acids Res. **10**:2865–2878.
29. **O'Brien, S. J., T. I. Bonner, M. Cohen, C. O'Connell, and W. G. Nash.** 1983. Mapping of an endogenous retroviral sequence to human chromosome 18. Nature (London) **303**:74–77.
30. **O'Connell, C. D., and M. Cohen.** 1984. The long terminal repeat sequences of a novel human endogenous retrovirus. Science **226**:1204–1206.
31. **O'Connell, C. D., 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.
32. **Patarca, R., and W. A. Haseltine.** 1984. Matters arising. Nature (London) **309**:288.
33. **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.
34. **Ratner, L., W. Haseltine, R. Patarca, K. J. Livak, B. Starcich, S. F. Josephs, E. R. Doran, J.A. Rafalski, E. A. Whitehorn, K. Baumeister, L. Ivanoff, S. R. Petteway, Jr., M. L. Pearson, J. A. Lautenberger, T. S. Papas, J. Ghrayeb, N. T. Chang, R. C. Gallo, and F. Wong-Stall.** 1985. Complete nucleotide sequence of the AIDS virus, HTLV-III. Nature (London) **313**:277–284.
35. **Redmond, S. M. S., and C. Dickson.** 1983. Sequence and expression of the mouse mammary tumor virus env gene. EMBO J. **2**:125–131.
36. **Repaske, R., R. R. O'Neill, P. E. Steele, and M. A. Martin.** 1983. Characterization and partial nucleotide sequence of endogenous type C retrovirus segments in human chromosomal DNA. Proc. Natl. Acad. Sci. USA **80**:678–682.
37. **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.
38. **Rigby, P. W. J., M. Dieckmann, C. Rhodes, and P. Berg.** Labeling deoxyribonucleic acid to high specific activity *in vitro* by nick translation with DNA polymerase I. J. Mol. Biol. **113**:237–251.
39. **Roth, J. R.** 1981. Frameshift suppression. Cell **24**:601–602.
40. **Sagata, N., T. Yasunaga, J. Tsuzuku-kawamura, K. Ohishi, Y. Ogawa, and Y. Ikawa.** 1985. Complete nucleotide sequence of the genome of bovine leukemia virus: its evolutionary relationship to other retroviruses. Proc. Natl. Acad. Sci. USA **82**:677–681.
41. **Saigo, K., W. Kugimiya, Y. Matsuo, S. Inouye, K. Yoshioka, and S. Yuki.** 1984. Identification of the coding sequence for a reverse transcriptase-like enzyme in *Drosophila melanogaster*. Nature (London) **312**:659–661.
42. **Schwartz, D. E., R. Tizard, and W. Gilbert.** 1983. Nucleotide sequence of Rous sarcoma virus. Cell **32**:853–869.
43. **Seiki, M., S. Hattori, Y. Hirayama, and M. Yoshida.** 1983. Human adult T-cell leukemia virus: complete nucleotide sequence of the provirus genome integrated in leukemia cell DNA. Proc. Natl. Acad. Sci. USA **80**:3618–3622.
44. **Sharp, P. A.** 1983. Conversion of RNA to DNA in mammals: Alu-like elements and pseudogenes. Nature (London) **301**: 471–472.
45. **Shinnick, T. M., R. A. Lerner, and J. G. Sutcliffe.** 1981. Nucleotide sequence of Moloney murine leukaemia virus. Nature (London) **293**:543–548.
46. **Soule, H. D., J. Vazquez, A. Long, S. Albert, and M. Brennan.** 1973. A human cell line from a pleural effusion derived from a breast carcinoma. J. Natl. Cancer Inst. **51**:1409–1416.
47. **Southern, E. M.** 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol. **98**:503–517.
48. **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.
49. **Toh, H., H. Hayashida, and T. Miyata.** 1983. Sequence homology between retroviral reverse transcriptase and putative polymerases of hepatitis B virus and cauliflower mosaic virus. Nature (London) **305**:827–829.
50. **Varmus, H.** 1985. Reverse transcriptase rides again. Nature (London) **314**:583–584.
51. **Wahl, G. M., M. Stern, and G. R. Stark.** 1979. Efficient transfer of large DNA fragments from agarose gels to diazobenzloxymethyl-paper and rapid hybridization by using dextran sulfate. Proc. Natl. Acad. Sci. USA **76**:3683–3687.
52. **Wain-Hobson, S., P. Sonigo, O. Danos, S. Cole, and M. Alizon.** 1985. Nucleotide sequence of the AIDS virus, LAV. Cell **40**:9–17.
53. **Wilbur, W. J., and D. J. Lipman.** 1983. Rapid similarity searches of nucleic acid and protein data banks. Proc. Natl. Acad. Sci. USA **80**:726–730.
54. **Yoshinaka, Y., I. Katoh, T. D. Copeland, and S. Oroszlan.** 1985. Murine leukemia virus protease is encoded by the *gag-pol* gene and is synthesized through suppression of an amber termination codon. Proc. Natl. Acad. Sci. USA **82**:1618–1622.