JOURNAL OF VIROLOGY, June 1985, p. 764–772
0022-538X/85/060764-09$02.00/0
Copyright © 1985, American Society for Microbiology

Vol. 54, No. 3

Nucleotide Sequence of a Full-Length Human Endogenous Retroviral Segment

ROY REPASKE,* PAUL E. STEELE, RAYMOND R. O'NEILL, ARNOLD B. RABSON, AND MALCOLM A. MARTIN

Laboratory of Molecular Microbiology, National Institute of Allergy and Infectious Diseases, Bethesda, Maryland 20205

Received 20 December 1984/Accepted 27 February 1985

The nucleotide sequence of a full-length (8.8-kilobase) endogenous C-type human retroviral DNA (clone 4-1) is presented and compared with that of Moloney murine leukemia virus (MoMuLV) DNA. Colinearity of deduced amino acids of clone 4-1 with MoMuLV in the *gag* and *pol* regions was clearly evident, and overall amino acid homology in these regions was about 40%. Identification of the putative N terminus of *gag* and p30, the *gag-pol* junction, and the C terminus of *pol* could be established on the basis of sequence homology with MoMuLV. Unique characteristics of the endogenous human retroviral DNA included a tRNA^{Glu} primer binding site separated from the 5′ long terminal repeat by a pentanucleotide and a putative *env* sequence which does not appear to overlap the C terminus of *pol* and has virtually no homology with the *env* gene of known infectious retroviruses. Clone 4-1 represents a defective prototype of a human C-type retrovirus which integrated into the germ line some time in the distant past.

We have previously described the isolation of a human genomic clone containing retrovirus-like sequences by employing an African green monkey endogenous retroviral probe under low-stringency hybridization conditions (9). Additional human retroviral clones (17) were subsequently selected under high-stringency hybridization conditions, using a *pol*-containing restriction fragment derived from the first human clone, 51-1. Of the 13 new retroviral clones obtained, all were closely related to 51-1 in the *pol* and the 3′ *gag* region, although four were distinct in having common restriction sites in the putative *env* region (15). This latter group of clones proved to represent full-length retrovirus-like structures about 8.8 kilobases (kb) in length. They contained an associated terminal repeating element that possessed many of the features of a retroviral long terminal repeat (LTR). Nucleotide sequencing of these human endogenous LTRs revealed the presence of a TATA box, a polyadenylation signal, and a putative CCAAT sequence, as well as a tRNA binding site and a polypurine tract adjacent to the 3′ LTR (23). In this paper we present the complete nucleotide sequence of a prototype full-length human endogenous retroviral clone, 4-1, and discuss its structural and evolutionary relationships to other known type C proviral DNAs.

MATERIALS AND METHODS

Human DNA clones. Clone 4-1 (15) was obtained from a human genomic DNA library which was screened with a *pol* fragment derived from the first isolated human clone (9). Clone 5-1 was an additional isolate picked during the screening of clone 4-1. Both clones had similar restriction maps in the putative *env* regions. Clone 4-14 was obtained from the human genomic DNA library by using the 3′ LTR segment of clone 4-1 as a probe (23). Twelve additional human clones, reactive with an *env* probe derived from clone 4-1, were obtained from a genomic library (8) of human leukemic spleen DNA. They were cloned into λBF101, and one of them, pE40, was subcloned into pBR322.

DNA sequence and data analysis. Restricted DNA fragments were 5′ end labeled ([λ-^{32}P]ATP; 3,000 Ci/mmol; Amersham) with T4 polynucleotide kinase (P-L Biochemicals). After restriction, single end-labeled fragments were separated and sequenced by the partial degradation method of Maxam and Gilbert (10). The computer program of Queen and Korn (14) was used for translation to amino acids, identification of restriction sites, and determination of sequence homology. Preliminary alignment of sequences was performed by using NUCALN and PRTLAN (25). The DOTMATRIX comparison program of Blomquist et al. (1) was used to show homology and colinearity of the human sequence with Moloney MuLV sequence.

RESULTS AND DICSUSSION

Human genomic DNA contains full-length (8.8 kb) copies of retroviral sequences (15), "truncated" retroviral segments that consist of a closely related 4.1-kb stretch of *gag-pol* sequence but no associated *env* (17) or LTR regions (23) and isolated LTR elements that are not affiliated with retroviral sequences (23). We previously reported the partial nucleotide sequence of the *gag* and *pol* regions of one of the cloned human retroviral segments and showed that the deduced amino acid sequence could be colinearly aligned with analogous regions of Moloney (Mo) murine leukemia virus (MuLV) proviral DNA (17).

During the past year we have focused our attention on the full-length family of human endogenous retroviral clones and have reported (i) the nucleotide sequence of five different human LTR elements (23) and (ii) the presence of discrete species of polyadenylated RNA in a variety of human cells that hybridize to human LTR or *env* DNA probes (15). In this paper we present the complete nucleotide sequence of a prototype (clone 4-1) full-length human retroviral sequence. In addition, the partial nucleotide sequences of several closely related 8.8-kb retroviral segments were also determined and included when appropriate in comparisons with other proviral DNAs.

Nucleotide sequence of 4-1 DNA: general considerations. The 8,806 base pairs (bp) comprising the retroviral sequences in clone 4-1 are shown in Fig. 1; the sequencing strategy employed and the relevant restriction sites are

* Corresponding author.

```
LTR                                                                                                                        120
TATGGTATGAGGTCACCACTTCTCCTGTTGTCCTTCTCAGTTCCTCCCCAACCTCCCCTTTTCCCCAGTTTATAAGACAGGAGAAAAGGGAGAAAGCAAAAAGTTGAAAAGAAACAGAAG
                                                                                                                       240
TAAGATAAATAGCTAGATGACCTTGGCACCACCACCTGGCCCTGGTGGCTAAAATATAATATTATTAACCCCTGACCAAAACTGTTGGTGTTATCTGTAAATTCCAGATATTGTATGAGA
                                                                                                                       360
AAGTACTGTAAAACTTTTTATTCTGTTAGCTGATGTAGGTAGCCCCCAGTCATGTTTCTCACGCTTACTTGACCTATTATGACTTTTTCATGTAGACCCCTTAGAGTTGTAAGCCCTTAA
                                                                                                                       480
AAGGGCTAGGAATTTCTTTTTTGGGGAGCTCGGCTCTTAAGATACGAGTCTGCCAATGCTCCCGGCCAAATAAAAAACCTCTTCCTTCTTTAATCTGGTGTCTGAGGAGTTTTGTCTGTG
               \/   t RNA Binding Site                                                                                 600
ACTCGTCCTGCTACATTTCTTGGTTCCCTGGCCAGGAAGCAAGGTAATTGAAGGACAGTCGAGGCAGCCCCTTAGGTGGCTTAGGCCTGCCCTGTGGAGCATCCCTGCAGGGGACTCTGG
                                   Splice Donor                                                                        720
CCAGCTTGAGTGACGCGGATCCTGAGAGCGCTCCCAGGTAGGCAATTACCCCGGTGGAAAGCCTCGTCAGAGAGTGCGTGGCAGGCCCCTGTGGAGGATCAATGCAGTGGCTGAACACTG
                                                                                                                       840
GGAAGGAACAGGCACTTGGAGTCCAGACATTTGAAACTTGGTAAGACTGGTCTTCGGAACTTGCCCACTCCATTTGAGTGGAAGCGTGGCCTGATCAACCACGGCATGCCTGTACTGGCA
                                                                                                                       960
CTTTGGTTTTTGTTTTTGACTTGACTTGAATTGCTTGATACTTTGGTTTTGGTTTGACCTGGCTTGGATTTCTGGATACTCTGATTTTGGTTTTGATTCTGGTTTGGTGAAAACTGAAAA
                                                                                  \/ gag    (p15/p12)                 1080
AGTGTGTGTGTGCACTTTTTACCCATTCTTTGTTTTGTGGTGTGCATGTGGTGTGAGCTTGGTGTTTTGTCTTGAGGAAACATGGATCAGACACAAAATAAGCCTACTCCTCTAGGAACT
                                                                             ***GlyAsnMetAspGlnThrGlnAsnLysProThrProLeuGlyThr
                                                                                                                      1200
ATGTTGAAAAATTTTAAGAAGGGATTTAATGGAGACTATGGGGTTACTATGACACCAGGGAAACTTAGAACTTTGTGTGAAATAGATTGGCCAACATTAGAAGTGGGTTGGCCATCAGAA
MetLeuLysAsnPheLysLysGlyPheAsnGlyAspTyrGlyValThrMetThrProGlyLysLeuArgThrLeuCysGluIleAspTrpProThrLeuGluValGlyTrpProSerGlu
                                                                                                                      1320
GGGAGCCTGGACGGGTCCCTTGTTTCTAAGGTATGGCACAAGGTAACTAGTAAGTCAGGACACTCAGACCAGTTTCCATACATAGACACTTGGTTACAGCTGGTGCTAGACCCCCCACAG
GLySerLeuAspGLySerLeuValSerLysValTrpHisLysValThrSerLysSerGlyHisSerAspGlnPheProTyrIleAspThrTrpLeuGlnLeuValLeuAspProProGln
                                                                                                                      1440
TGGCTAAGAGGGCAGGCAGCAGCAGTGCTAGTAGCAAAGGGACAGATAGTCAAGGAAGGATTCTGCTCCACCCGCTGAGGGAAATCAACTCCTGAAGTTCTGTTCGACCAAACATCAGAA
TrpLeuArgGlyGlnAlaAlaAlaValLeuValAlaLysGlyGlnIleValLysGluGlyPheCysSerThrArg***GlyLysSerThrProGluValLeuPheAspGlnThrSerGlu
                                                                                                                      1560
GATCCATTGCAGGAGATGGCACCAGTGATCCCAGTGTTGCCCTCCCCTTATCAGGGAGAGAGGCTCCCCACTTTTGAGTCCACAGTGCTTGCGCCTCTGCCAGACAAATGTATCCCTAGG
AspProLeuGlnGluMetAlaProValIleProValLeuProSerProTyrGlnGlyGluArgLeuProThrPheGluSerThrValLeuAlaProLeuProAspLysCysIleProArg
                                                                     ↓p30                                              1680
CCACTCAGAGTAGACAAGAGAGGAGGTGAAGCCTCGGGAGAAACCCCTCCCTTGGCAGCTCATTTAAGACCCAAAACAGGGATACAAATGCCCCTGAGAGAGCAGCAGTATACTGGAATA
ProLeuArgValAspLysArgGlyGlyGluAlaSerGlyGluThrProProLeuAlaAlaHisLeuArgProLysThrGlyIleGlnMetProLeuArgGluGlnGlnTyrThrGlyIle
                                                                                                                      1800
GATGAGGATGGGCACATGGTGGAGAGTCGTGTTTTTGTGTACCAGCCCTTCACCTCTGCCGACCTTCTCAACTGGAAAAACAATACCCCGTCCTATACTGAAAAGCCGCAAGCTCTAATT
AspGluAspGlyHisMetValGluSerArgValPheValTyrGlnProPheThrSerAlaAspLeuLeuAsnTrpLysAsnAsnThrProSerTyrThrGluLysProGlnAlaLeuIle
                                                                                                                      1920
GATTTGCTCCAAACTATTATCCAGACCCATAACCCCACTTGGGCTGATTGCCACCAGTTGCTCATGTTCCTCTTTAAAACAGATGAAAGGTGAAGGGTGCTTCAAGCAGCAACTAAGTGG
AspLeuLeuGlnThrIleIleGlnThrHisAsnProThrTrpAlaAspCysHisGlnLeuLeuMetPheLeuPheLysThrAspGluArg***ArgValLeuGlnAlaAlaThrLysTrp
                                                                                                                      2040
CTAGAGGAACATGCACTGGCTGATTACCAAAACCCCCAAGAGTATGTAAGGACACAGTTACCAGGAACCGACCCCCAGTGGGACCCAAATTAAAGAGAGGATATGCAAAGGCTAAACCGA
LeuGluGluHisAlaLeuAlaAspTyrGlnAsnProGlnGluTyrValArgThrGlnLeuProGlyThrAspProGlnTrpAspProAsn***ArgGluAspMetGlnArgLeuAsnArg
                                                                                                                      2160
TACAGGAAAGCTCTCTTAGAAGGTTTAAAGAGGAGAGCCCAGAAGGCCACAAACATTAACAAGGTCTCTGAGGTCATTCAGGGAAAAGAAGAAAGTCCAGCAAAATTCCACGAGAGACTG
TyrArgLysAlaLeuLeuGluGlyLeuLysArgArgAlaGlnLysAlaThrAsnIleAsnLysValSerGluValIleGlnGlyLysGluGluSerProAlaLysPheHisGluArgLeu
                                                                                                                      2280
TGTGAGGCTTATTGTATGTATACTCCCTTTGATCCCGATAGCCCTGAAAATCAACGCATGATTAACATGGCTTTAGTTAGTCAAAGCACAGAAGACATTAGAAGAAAACTGCAGAAAAAG
CysGluAlaTyrCysMetTyrThrProPheAspProAspSerProGluAsnGlnArgMetIleAsnMetAlaLeuValSerGlnSerThrGluAspIleArgArgLysLeuGlnLysLys
                                                                                                                      2400
GCTGGGTTTGCAGGGATGAACACATCACAGTTATTAGAAATAGCCAACCAGGTGTTTGTAAACAGGGATGCAGCAAGCCGTAAGGAAACCACATAGAGAATGAACGTCAGGCCCGGCGAA
AlaGlyPheAlaGlyMetAsnThrSerGlnLeuLeuGluIleAlaAsnGlnValPheValAsnArgAspAlaAlaSerArgLysGluThrThr***ArgMetAsnValArgProGlyGlu
                                                                                                                      2520
ACGCGCCTGTTAGCTGCAGCAATTAGAGGGGTCCCCCAAAAGAGGCAAGGCAAAAGGGGGGCCCTGGGAAAGAAACTCAGCCTGGCTGTCAGAGCTTGCAGTGTAATCAGTGTGCTTAT
ThrArgLeuLeuAlaAlaAlaIleArgGlyValProProLysGluAlaArgGlnLysGlyGlyProGlyLysGluThrGlnProGlyCysGlnSerLeuGlnCysAsnGlnCysAlaTyr
                                                                                                                      2640
CGTAAAGAAATAGGATATTGGAAGAACAAATGCCCTCAGCTAAAAGGAAAACAAGGTGACTCGGAGCAGGAGGCTCCAGACAAGGAGGAAGGGGCCCTGCTCAACCTAGCAGAAGGGTTA
ArgLysGluIleGlyTyrTrpLysAsnLysCysProGlnLeuLysGlyLysGlnGlyAspSerGluGlnGluAlaProAspLysGluGluGlyAlaLeuLeuAsnLeuAlaGluGlyLeu
      \/ pol                                                                                                          2760
TTGGACTGAGGGGGACTGGGCTCAAGGACCTCCAAAGAGCCTATGGTCAGGATGACAGTTGGGGGTAAAGACATTGATTTTCTTGTAGATACCAGTGCTGAACATTCGGTAGTAACTGCC
LeuAsp***GlyGlyLeuGLySerArgThrSerLysGluProMetValArgMetThrValGlyGlyLysAspIleAspPheLeuValAspThrSerAlaGluHisSerValValThrAla
                                                                                                                      2880
TCAGTCGCCCCCTTATCCAAAAAGACTATTGACATCATCGGAGCCATGGGAGTTTCAGCAAAACAAGCTTTCTGCTTGCCCCAGACTTGTACTATAGGAGGACATAAAGTGATTCATCAG
SerValAlaProLeuSerLysLysThrIleAspIleIleGlyAlaMetGlyValSerAlaLysGlnAlaPheCysLeuProGlnThrCysThrIleGlyGlyHisLysValIleHisGln
                                                                                                                      3000
TTTTTGTACATGCCTGATTGTCCCTTGCCCTTGTTGGGAAGAGACTTGCTTAGCAAACTGAGAGCCACTATCTCTTTTACAGAGCACGGCTCTTTGCTGCTAAAGTTACCCGGAACAGGA
PheLeuTyrMetProAspCysProLeuProLeuLeuGlyArgAspLeuLeuSerLysLeuArgAlaThrIleSerPheThrGluHisGlySerLeuLeuLeuLysLeuProGlyThrGly
                                                                                                                      3120
GTCATTATGACCCTTATGCTCCCCCGAGAGGAGGAATGGAGACTTTTCTTAACTGAGCCGGGCCAAGAGATAAGACCAGCTCTGGCTAAGCGGTGGCCAAGAGTGTGGGCGGAAGCGAAC
ValIleMetThrLeuMetLeuProArgGluGluGluTrpArgLeuPheLeuThrGluProGlyGlnGluIleArgProAlaLeuAlaLysArgTrpProArgValTrpAlaGluAlaAsn

                                                                                                                      3240
CCTCCAGGGTTGGCAGTCAACCAAGCCCCCGTGCTTATAGAAGTTAAGCCTGGGGTCCAGCCGGTTAGGCAAAAACAGTACCCGGTCCTCAGAGAAGCTCTTGAAGGTATCCAGGTCCAT
ProProGlyLeuAlaValAsnGlnAlaProValLeuIleGluValLysProGlyValGlnProValArgGlnLysGlnTyrProValLeuArgGluAlaLeuGluGlyIleGlnValHis
                                                                                                                      3360
CTCAAGTGCCTAAGAACCTTTAGAATTATAGTTCCTTGTCAGTCTCCATGGAACACTCCCCTCCTGCCTGTTCCCAAGCCTGGGACCAAGGACTACAGGCCGGTACAGGATTTGCGCTTG
LeuLysCysLeuArgThrPheArgIleIleValProCysGlnSerProTrpAsnThrProLeuLeuProValProLysProGlyThrLysAspTyrArgProValGlnAspLeuArgLeu
                                                                                                                      3480
GTTAATCAGGCTACAGTGACTTTACATCCAACAGTACCTAACCTGTACACATTGCTGGGGTTGCTGCCAGCTGAGGACAGCTGGTTCACCTGCTTGGACCTGAAAGATGCTTTCTTTAGC
ValAsnGlnAlaThrValThrLeuHisProThrValProAsnLeuTyrThrLeuLeuGlyLeuLeuProAlaGluAspSerTrpPheThrCysLeuAspLeuLysAspAlaPhePheSer
                                                                                                                      3600
ATCAGATTAGCCCCTGAGAGACAGAAGCTGTTTGCCTTTCAGTGGGAAGATCCAGAGTCAGGTGTCACTACTCAATACACTTGGACCCAGCTTCCCCAAAGGTTCAAGAACTCCCCCACC
IleArgLeuAlaProGluArgGlnLysLeuPheAlaPheGlnTrpGluAspProGluSerGlyValThrThrGlnTyrThrTrpThrGlnLeuProGlnArgPheLysAsnSerProThr
                                                                                                                      3720
ATCTTTGGGGAGGCGTTGGCTCGAGACCTCCAGAAGTTTCCCACCAGAGACCTAGGCTGCGTGTTGCTCCAGTACGTTGATGACCTTTTGCTGGGACACCCCACGGCAGTCGGGTGGCCA
IlePheGlyGluAlaLeuAlaArgAspLeuGlnLysPheProThrArgAspLeuGlyCysValLeuLeuGlnTyrValAspAspLeuLeuLeuGlyHisProThrAlaValGlyTrpPro
                                                                                                                      3840
AGGGAACAGATGCTCTACTCCGGCACCTGGAGGACTGTGGGTATAAGGTGTCCAAGAAAAAAAGCTCAGATCTGCCGACAGCAGGTATGTTACTTGGGATTTACTATCCAACAGGGGGAG
ArgGluGlnMetLeuTyrSerGlyThrTrpArgThrValGlyIleArgCysProArgLysLysAlaGlnIleCysArgGlnGlnValCysTyrLeuGlyPheThrIleGlnGlnGlyGly
```

Continued on following page

```
3960
CACAGCCTAGGATCAGAAAGAAAGCAGGTCATTTGTAATCTACCGGAGCCTAAGACCAGAAGGCAGGTGAGAGAATTCTTAGGGGCTGTGGGTTTTTGCAGACTGTGGATCCCAAACTTT
HisSerLeuGlySerGluArgLysGlnValIleCysAsnLeuProGluProLysThrArgArgGlnValArgGluPheLeuGlyAlaValGlyPheCysArgLeuTrpIleProAsnPhe
4080
GCAGTATTAGCTAAGCCTTTGTATGAGGTCACAAAGGCGGGGGACCAGGAACCTTTTGAATGGGGATCCCAGCAACAGCAAGCCTTTCATGAGTTAAAGGAAAGACTTATGTCAGTCCCA
AlaValLeuAlaLysProLeuTyrGluValThrLysAlaGlyAspGlnGluProPheGluTrpGLySerGlnGlnGlnGlnAlaPheHisGluLeuLysGluArgLeuMetSerValPro
4200
GCCCTGGGGCTACCTGATCTGACAAAGCCTTTTACATTGTATGTGTCAGAGAGTGAAAAGATGGCAGTTGGAGTTTTAACCCAAACTGTGGGGCCCTGGCCGAGGCCGGTGACCTACCTC
AlaLeuGlyLeuProAspLeuThrLysProPheThrLeuTyrValSerGluSerGluLysMetAlaValGlyValLeuThrGlnThrValGlyProTrpProArgProValThrTyrLeu
4320
TCTAAACAACTAGACGGGGTTTCTAAAGGATGGCCCCCGTGTTTGAGGGCCTTGGCAGCAACTGCCCTGCTAGTACAAGAAGCAGATAAGCTGATTCTTGGGCAAAACCTGAACATAAAG
SerLysGlnLeuAspGlyValSerLysGlyTrpProProCysLeuArgAlaLeuAlaAlaThrAlaLeuLeuValGlnGluAlaAspLysLeuIleLeuGlyGlnAsnLeuAsnIleLys
4440
GACCCCCATGCTGTGGTGACTTTAATGAATACTAGAGGACATCATTGGCTAACGAATGCTAGACTTACTAAGTACCAAAGTTTGCTTTGTGAAAATCCCCATATAACCATTGAAGTTTGT
AspProHisAlaValValThrLeuMetAsnThrArgGlyHisHisTrpLeuThrAsnAlaArgLeuThrLysTyrGlnSerLeuLeuCysGluAsnProHisIleThrIleGluValCys
4560
AACACCCTGAACCCCGCTACCTTGCTCCCAGTATTAGAGATCCCTGTCGAGCATGACTGTGTAGAAGTGTTGGACTCAGTTTACTCTGGGCAXTCAGTAGACTGGGAACTATACGTGGAT
AsnThrLeuAsnProAlaThrLeuLeuProValLeuGluIleProValGluHisAspCysValGluValLeuAspSerValTyrSerGLy   SerValAspTrpGluLeuTyrValAsp
4680
GGGAGCAGCTTTGTCAACCCACAAGAAGAGAGATGTGCAGGGTATGCGGTGGTAACTCTGGACACTGTTGCTGAAGCCAGATCGTTTCCCCAGGGCACTTCAACTCAGAAAGCTGAACTC
GLySerSerPheValAsnProGlnGluGluArgCysAlaGlyTyrAlaValValThrLeuAspThrValAlaGluAlaArgSerPheProGlnGlyThrSerThrGlnLysAlaGluLeu
4800
ATTGCTTTAATTCGGGCCTTAGAACTCAGTGAAGGTAAGACTGTAAACATTTACACTGACTCTTGATATGTCTTTTTAACCCTTCAAGTGCATGGAGCATTATGTAAAGAAAAGGGCCTA
IleAlaLeuIleArgAlaLeuGluLeuSerGluGlyLysThrValAsnIleTyrThrAspSer***TyrValPheLeuThrLeuGlnValHisGlyAlaLeuCysLysGluLysGlyLEU
4920
TTGAACTCTGGGGGAAAAGACATAAAATATCAACAAGAAATCTTGCAATTATTAGAAGCAGTATGGAAACCCCACAAGGTGGCTGTTATACATTGCGGAGGACACCAGTGAGCTTCCACC
LeuAsnSerGlyGlyLysAspIleLysTyrGlnGlnGluIleLeuGlnLeuLeuGluAlaValTrpLysProHisLysValAlaValIleHisCysGlyGlyHisGln***AlaSerThr
5040
TTGGTGGGTTTGGGGAATTCCTGCACTGACTTAGAGGCTCAAAAAGCAGCATCTGCCCTTCCGGGCATCAGTGACAGCCCCCCTGCTCCCTCAAGCACCTGATCTTGTACCTACTTATTC
LeuValGlyLeuGlyAsnSerCysThrAspLeuGluAlaGlnLysAlaAlaSerAlaLeuProGlyIleSerAspSerProProAlaProSerSerThr***SerCysThrTyrLeuPhe
5160
TAAAGAAGAAAAGGACTTTCTCCAGGCAGAGGGAGGACAAGTGATGGAGGAAGGATGGATTTGGTTACCAGATGGGAGAGTAXXGCTGTGCCACAGCTGCTAGGAGCTGCAGTTGTACTG
***ArgArgLysGlyLeuSerProGlyArgGlyArgThrSerAspGlyGlyArgMetAspLeuValThrArgTrpGluSer   AlaValProGlnLeuLeuGlyAlaAlaValValLeu
5280
GCTGTGCATAAAACCACCCATCTAGGTCAGGAATCACTTGAAAAGTTGTTAGGCTGGTATTTCTACATCTCGCATTTGTCAGCCCTTGCCAAAACAGTGACGCAGCGGTGTGTTACCTGC
AlaValHisLysThrThrHisLeuGlyGlnGluSerLeuGluLysLeuLeuGlyTrpTyrPheTyrIleSerHisLeuSerAlaLeuAlaLysThrValThrGlnArgCysValThrCys
5400
CGACAGCATAATGCGAGACAAGGTCCAGCTGTTCCCCCTGGCATACAAGCTTATGGAGCAGCCCCCTTTGAAGATCTCCAGGTGGACTTCACAGAGATGCCAAAGTGTGGAGGTAACAAG
ArgGlnHisAsnAlaArgGlnGlyProAlaValProProGlyIleGlnAlaTyrGlyAlaAlaProPheGluAspLeuGlnValAspPheThrGluMetProLysCysGlyGlyAsnLys
5520
TATTTACTAGTTCTTGTGTGTACCTACTCTGGGCAGGTGGAGGCTTATCCAACACGAACTGAGAAAGCTCATGAAGTAACTCGTGTGCTTCTTCGAGATCTTATTCCTAGATTTGGACTG
TyrLeuLeuValLeuValCysThrTyrSerGlyGlnValGluAlaTyrProThrArgThrGluLysAlaHisGluValThrArgValLeuLeuArgAspLeuIleProArgPheGlyLeu
5640
CCCTTACGGATTGGCTCAGATAATGGGCTGGTGTTTGTGGCTGACTTGGTACAGAAGACGGCAAAGGTATTGGGGATCACATGGAAACTGCATGCTGCCTACCAGCCTCAGAGTTCCGGA
ProLeuArgIleGLySerAspAsnGlyLeuValPheValAlaAspLeuValGlnLysThrAlaLysValLeuGlyIleThrTrpLysLeuHisAlaAlaTyrGlnProGlnSerSerGly
5760
AAGGTAGAGCGGATGAATCGGACTATCAAAAATAGTTTAGGGAAAGTATGTCAAGAAACAGGATTAAAATGGATACAGGCTCTTCCTATGGTATTATTTAAAATTAGATGTACCCCTTCT
LysValGluArgMetAsnArgThrIleLysAsnSerLeuGlyLysValCysGlnGluThrGlyLeuLysTrpIleGlnAlaLeuProMetValLeuPheLysIleArgCysThrProSer
Splice Acceptor 5880
AAAAGAACAGGATATTCCCCTTATGAAATATTATATCATAGGCCCCCTCCTATATTGCGGGGACTTCCAGGCACTCCCCGAGAGTTAGGTGAAATTGAGTTACAGCGATAGCTACAGGCT
LysArgThrGlyTyrSerProTyrGluIleLeuTyrHisArgProProProIleLeuArgGlyLeuProGlyThrProArgGluLeuGlyGluIleGluLeuGlnArg***LeuGlnAla
6000
TCAGGAAAAATTACACAAACAATCTCGGCCTGGGTAAATGAGAGATGCCCTGTTAACTTATTCTCCCCAGTTCACCCTTTCTCCCCAGGTGATCTAGTGTGGATCAAGGACTGAAACGTA
SerGlyLysIleThrGlnThrIleSerAlaTrpValAsnGluArgCysProValAsnLeuPheSerProValHisProPheSerProGlyAspLeuValTrpIleLysAsp***AsnVal
6120
GCCTGTTTGTGTCCACGGTGGAAAGGACCCCAGACTGTCATCCTGAGCACTCCCACCGCTGTGAAGGTAGAGGGAATCCCAACCTGGATCCACCACAGCCGTGTAAAACCTGCAGTGCCT
AlaCysLeuCysProArgTrpLysGlyProGlnThrValIleLeuSerThrProThrAlaValLysValGluGlyIleProThrTrpIleHisHisSerArgValLysProAlaValPro
                                                                                                  \/                      \/ ENV
GAAACCTGGGAGGCAAGACCAAGCCCAGAAAACCCCTGCAGAGTGACCCCGAAGAAGACAACAAGCCCTGCTCCAGTCACACCCGGAAGCTGACTGGTCCACGCACGGCCGAAGCATGCAG
GluThrTrpGluAlaArgProSerProGluAsnProCysArgValThrProLysLysThrThrSerProAlaProValThrProGLySer***                        MetGln
6360
AAGCTCATCATGGGATTCATTTTTCTTAAATTTTGGACTTATACAGTAAGGGCTTCAACTGATCTTACTCAAACTGGGGACTGTTCCCAGTGTATTCATCAGGTCACCGAGGTAGGACAG
LysLeuIleMetGlyPheIlePheLeuLysPheTrpThrTyrThrValArgAlaSerThrAspLeuThrGlnThrGlyAspCysSerGlnCysIleHisGlnValThrGluValGlyGln
6480
CAAATTAAAACAATGTTTCTGTTCTATAGTTATTATAAATGTATAGGAACATTAAAAGAAACTTGTTTGTATAATGCTACTCAGTACAATGTATGTAGCCCAGGAAATGACCGACCTGAT
GlnIleLysThrMetPheLeuPheTyrSerTyrTyrLysCysIleGlyThrLeuLysGluThrCysLeuTyrAsnAlaThrGlnTyrAsnValCysSerProGlyAsnAspArgProAsp
6600
GTGTGTTATAACCCATCTGAGCCTCCTGCAACCACCATTTTTGAAATAAGAATAAGAACTGGCCTTTTCCTAGGTGATACAAGTAAAATAATAACTAGAACAGAAGAAAAAGAAATCCCC
ValCysTyrAsnProSerGluProProAlaThrThrIlePheGluIleArgIleArgThrGlyLeuPheLeuGlyAspThrSerLysIleIleThrArgThrGluGluLysGluIlePro
6720
AAACAAATAACTTTAAGATTTGATGCTTGTGCAGCCATTAATAGTAAAAAGCTAGGAATAGGATGTGATTCTCTTAACTGGGAAAGGAGCTACAGAATAAAAAATAAATATGTTTGTCAT
LysGlnIleThrLeuArgPheAspAlaCysAlaAlaIleAsnSerLysLysLeuGlyIleGlyCysAspSerLeuAsnTrpGluArgSerTyrArgIleLysAsnLysTyrValCysHis
6840
GAGTCAGGGGTTTGTGAAAATTGTGCCTATTGGCCATGTGTTATTTGGGCTACTTGGAAAAAGAACAAAAAGGACCCGGTTTATCTTCAGAAGGGGGAAGCCAACCCCTCCTGTGCTGCT
GluSerGlyValCysGluAsnCysAlaTyrTrpProCysValIleTrpAlaThrTrpLysLysAsnLysLysAspProValTyrLeuGlnLysGlyGluAlaAsnProSerCysAlaAla
6960
GGTCACTGTAACCCACTAGAACTAATAATTACCAATCCCCTAGATCCCCATTGGAAAAAGGGAGAACGTGTAACCCTGGGGATTGATGGGACAGGGTTAAACCCCCAAGTTGCCATTTTA
GlyHisCysAsnProLeuGluLeuIleIleThrAsnProLeuAspProHisTrpLysLysGlyGluArgValThrLeuGlyIleAspGlyThrGlyLeuAsnProGlnValAlaIleLeu
7080
ATTAGAGGGGAGGTCCACAAGTGCTCTCCCAAACCAGTATTTCAAACCTTTTATAAGGAGCTGAATCTGCCAGCACCAGAATTTCCAAAAAAGACAAAAAATTTGTTTCTCCAATTAGCA
IleArgGlyGluValHisLysCysSerProLysProValPheGlnThrPheTyrLysGluLeuAsnLeuProAlaProGluPheProLysLysThrLysAsnLeuPheLeuGlnLeuAla
7200
GAAAATGTAGCTCATTCCCTTAATGTTACTTCTTGTTATGTATGCGGGGGAACCACTATCGGAGACCGATGGCCTTGGGAAGCCCGAGAGTTGGTGCCTACTGATCCAGCTCCTGATATA
GluAsnValAlaHisSerLeuAsnValThrSerCysTyrValCysGlyGlyThrThrIleGlyAspArgTrpProTrpGluAlaArgGluLeuValProThrAspProAlaProAspIle
7320
ATTCCAGTTCAGAAAACCCAAGCTAGCAACTTCTGGGTCCTAAAAACCTCAATTATTGGACAATACTGTATAGCTAGAGAAGGGAAAGACTTTATCATCCCTGTAGGAAAGCTTAATTGT
IleProValGlnLysThrGlnAlaSerAsnPheTrpValLeuLysThrSerIleIleGlyGlnTyrCysIleAlaArgGluGlyLysAspPheIleIleProValGlyLysLeuAsnCys
7440
ATAGGACAGAAGTTGTATAACAGTACAACAAAGACAATTACTTGGTGGGGCATAAACCACACTGAAAAGAATCCATTTAGTAAATTTTCAAAATTAAAAACTGCTTGGGCTCATCCAGAA
IleGlyGlnLysLeuTyrAsnSerThrThrLysThrIleThrTrpTrpGlyIleAsnHisThrGluLysAsnProPheSerLysPheSerLysLeuLysThrAlaTrpAlaHisProGlu
```

Continued on following page

```
                                                                                                                    7560
TCTCATCAGGACTGGATGGCTCCCGCTGGACTATACTGGATATGTGGGCACAGAGCCTACATTCGGTTACCTAATAAATAGGCAGGCAGTTGTGTTATTGGCACTATTAAGTCGTCCTTT
SerHisGlnAspTrpMetAlaProAlaGlyLeuTyrTrpIleCysGlyHisArgAlaTyrIleArgLeuProAsnLys***AlaGLySerCysValIleGlyThrIleLysSerSerPhe
                                                                           ↓p15 E                                   7680
TTCTTATTACCCATAAAAACAGGTGAGACCCTAGGTTTCCCTGTCTATGCCTCCCGAGAAAAGAGAGGCATAGTTATAGGAAACTGGAAAGATAATGAGTGGCGCCCTGAAAGGATCATA
PheLeuLeuProIleLysThrGlyGluThrLeuGlyPheProValTyrAlaSerArgGluLysArgGlyIleValIleGlyAsnTrpLysAspAsnGluTrpArgProGluArgIleIle
                                                                                                                    7800
CAGTATTATGGGCCTGCCACATGGGCACAAGACGGCTCATGGGGATACCGAACCCCCATTTACATGCTCAATCGGATCATACGGTTGCAGGCCATCTTAGAAATAATTACTAATGAAACT
GlnTyrTyrGlyProAlaThrTrpAlaGlnAspGLySerTrpGlyTyrArgThrProIleTyrMetLeuAsnArgIleIleArgLeuGlnAlaIleLeuGluIleIleThrAsnGluThr
                                                                                                                    7920
GGCAGAGCTTTGACTGTTTTAGCTCGGCAGGAAACCCAAACGAGGAATGCTATCTATCAGAATAGACTGGCCTTGGACTACTTGCTAGCAGCTGAAGGAGGAGTTTGTGGAAAATTTAAC
GlyArgAlaLeuThrValLeuAlaArgGlnGluThrGlnThrArgAsnAlaIleTyrGlnAsnArgLeuAlaLeuAspTyrLeuLeuAlaAlaGluGlyGlyValCysGlyLysPheAsn
                                                                                                                    8040
TTAACCAATTACTGCCTACAAATAGATGATCAAGGACAGGTGGTTGAAAACATAGTCAGGGACATGGCAAAGGTGGCACATGTGCCTGTACAGGTTTGGCACAAGTTTAATCCTGAGTCT
LeuThrAsnTyrCysLeuGlnIleAspAspGlnGlyGlnValValGluAsnIleValArgAspMetAlaLysValAlaHisValProValGlnValTrpHisLysPheAsnProGluSer
                                                                                                                    8160
TTATTTGGAAAATGGTTTCCAGCTATAGGAGGATTTAAAACCCTCATTGTAGGTGTATTGCTAGTGATAGGAACTTGCTTGCTGCTCCCCTGTGTATTACCCTTGCTTTTTCAAATGATA
LeuPheGlyLysTrpPheProAlaIleGlyGlyPheLysThrLeuIleValGlyValLeuLeuValIleGlyThrCysLeuLeuLeuProCysValLeuProLeuLeuPheGlnMetIle
                                                                                                                    8280
AAATATTTTGTTGTTACTTTAGTTCATCAGAAAACTTCAGCACATGTGTATTATACAAATCACTATCGCTCTATCTCACAAAGAGACTAAAAAAGTGAGGACGAGAGTAAGAACTCCCAC
LysTyrPheValValThrLeuValHisGlnLysThrSerAlaHisValTyrTyrThrAsnHisTyrArgSerIleSerGlnArgAsp***LysSerGluAspGluSerLysAsnSerHis
                                 \/ LTR                                                                             8400
TAAAAGTGAAAATTCTCAAAGGGGGGGAAATATGGTATGAGGTCGCCACTTCTCCTGTTGTCCTTCTCAGTTTCTCCCCAACCTCCCCTTTTCCCTAGTTTATAAGACAGGAGAAAAGGG
***Lys***LysPheSerLysGlyGly
                                                                                                                    8520
AGAAAGCAAAAAGTTGAAAAGAAACAGAAGTAAGATAAATAGCTGGACGACCTTGGCACCACCACCTGGCCCTGGTGGCTAAAATAATAATAATATTATTAACCCCTGACCAAAACTATT
                                                                                                                    8640
GGTGTTATCTGTAAATTCCAGACACTGTATGAGAAAATACTGTAAAACTTTTTGTTCTGTTAGCTGATGTATGTAGCCCCCAGTCATGTTTCTCACGCTTACTTGATCTATTATGACTTT
                                                                                                                    8760
TTCATGTAGACCCCTTAGAGTTCTAAGCCCTTAAAAGGGCAAGAATTTCTTTTTCGGGGAGCTCGGCTCTTAAGACACGAGTCTGCCAATGATCCCGGCCGAATAAAAAACCTCTTCCTT
                                            8809
CTTTAATCTGGCGTCTGAGGAGTTTTGTCTGCGACTCATCCTGCTACA
```

FIG. 1. Complete nucleotide sequence of an endogenous human retroviral DNA. The sequence of clone 4-1 (15), of which the LTR has been published (23), is shown in proviral form. Putative N and C termini indicated for *gag* and *pol* genes and the N terminus of p30 were identified by deduced amino acid homology of aligned sequences with Mo (22) and AKV ecotropic (4, 7) MuLVs. Splice donor and acceptor signals occurred in the same relative positions and were homologous with the MuLV sequence. Tentative placement of the gp70-p15E junction is discussed. In the *pol* gene, three inserted spaces (X), equivalent to deleted nucleotides, alter the true nucleotide numbering. These insertions were required to maintain the open reading frame and continue amino acid sequence homology with MoMuLV. All potential glycosylation sites in the putative *env* gene are boxed.

shown in Fig. 2. The 5′ LTR encompasses the first 495 bp, and the 3′ LTR sequence begins at nucleotide 8312 (23). The positioning of putative *gag* and *pol* gene sequences is based on the alignment of the deduced amino acid sequence with those of Mo (22) and AKV (4) MuLVs; the presumptive N termini of the *gag* and *pol* genes are located at nucleotides 1042 and 2650, respectively. The absence of significant polynucleotide or deduced amino acid sequence homology of the putative human retroviral *env* region with analogs of other known infectious mammalian type C proviral DNAs precluded the positioning of its 5′ terminus (see below).

None of the full-length human endogenous retroviral segments examined represents a potentially infectious proviral DNA; termination codons or point deletions or both have rendered clone 4-1 defective for replication. However, long open reading frames are present (e.g., the first 1,881 and 1,284 nucleotides in the putative *pol* and *env* regions, respectively). Adjustments for nucleotide deletions were made to maintain a reading frame possessing colinear amino acid homology with MoMuLV; these are indicated by three "X" insertions included in the 4-1 sequence shown in Fig. 1. Alignment of clone 4-1 and MoMuLV-deduced amino acid sequences occasionally created insertion-deletions in one sequence relative to the other to accommodate for an unequal number of intervening codons. A comparison of the deduced amino acid sequence of the 4-1 retroviral DNA with that of MoMuLV is shown in the computer-generated dot matrix analysis (1) shown in Fig. 3. An uninterrupted diagonal line would indicate colinearity of amino acid sequences as well as sequence homology. The deduced amino acids of clone 4-1 are obviously colinear with those of MoMuLV in the *gag* and *pol* regions, and sequence homology is clearly evident. The overall homology in the *gag* and *pol* regions is approximately 40%.

Analysis of the 5′ leader sequence and the *gag* region. The RNA primer binding site (pbs) abuts the 5′ LTR of most type

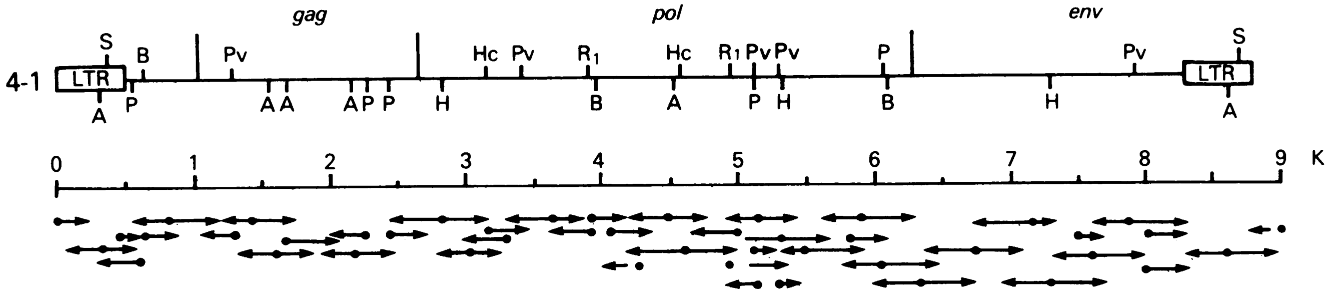

FIG. 2. Physical map of clone 4-1 retroviral DNA and nucleotide sequencing strategy. Primary restriction sites within the 8.8-kb retroviral sequence of clone 4-1 are indicated. Assignment of putative *gag*, *pol*, and *env* regions is based upon alignment of deduced amino acids homologous with those of MoMuLV as discussed in the text. Sequencing strategy is shown by arrows which indicate the direction and extent of individual sequence runs. Symbols: A, *AccI*; B, *BamHI*; H, *HindIII*; Hc, *HincII*; P, *PstI*; Pv, *PvuII*; R, *EcoRI*; S, *SacI*.

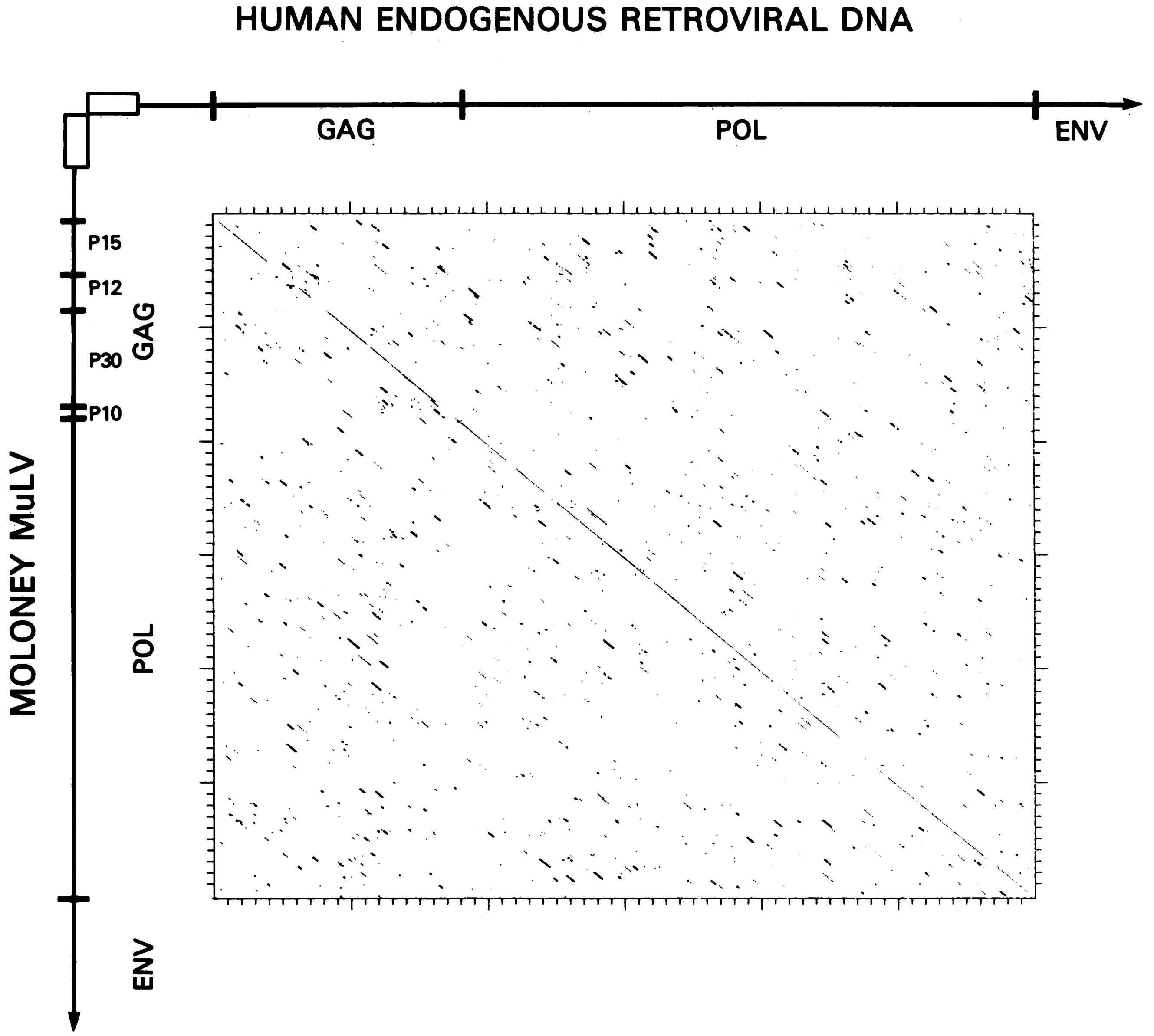

FIG. 3. Dot matrix analysis of clone 4-1 relative to MoMuLV amino acid sequences. The computer-generated section shows the extent of homology and colinearity of amino acid sequences of human clone 4-1 with sequences of MoMuLV (3) in the *gag* and *pol* regions. No extensive homology was observed beyond the regions shown. *x* axis = 1,723 amino acids; *y* axis = 1,738 amino acids; window size = 25; compression = 3; homology = 40%; interval = 30.

C retroviral DNAs. Exceptions are the human T-cell leukemia viruses I and II which contain a dinucleotide between the LTR and pbs (18, 21). Clone 4-1 and a second partially sequenced full-length human endogenous retroviral segment (clone 4-14) contain the same pentanucleotide, TTTCT (Fig. 1, position 496 to 500), that separates the LTR from a putative pbs. Furthermore, as previously noted (23), the 18-bp putative pbs beginning at position 501 (Fig. 1) does not match the complement of $tRNA^{Pro}$ but is a 16-of-18 match for the 3′ end of rat glutamic acid tRNA (19). The pbs in human clone 4-14 is a 17-of-18 complementary match for the same rat $tRNA^{Glu}$. No significant polynucleotide sequence homology of the 5′ *gag* leader regions of clone 4-1 (nucleotides 496 to 1041) and MuLV proviruses could be demonstrated. No open reading frame within the leader sequence capable of encoding a putative *gag* precursor polypeptide (4, 22) could be identified. A potential splice donor sequence (AGGTAGG) that is a 6-of-7 nucleotide match with MuLV splice donors was present, however, at positions 636 to 642 (Fig. 1).

The first ATG codon (Fig. 1, nucleotides 1042 to 1044) within a long open reading frame that would correspond to the N terminus of a putative human retroviral *gag* region is present within a sequence exhibiting a 4-of-5 amino acid identity (Asn-Met-X-Gln-Thr) in the region demarcating the 5′ end of the MoMuLV *gag* gene (22). This open reading frame extends 354 nucleotides (118 deduced amino acids) to a region corresponding to the C terminus of MoMuLV p15, where a termination codon (Fig. 1, position 1396 to 1398) is encountered. Three other stop codons not present in MuLV

A

```
        p30                                                                                                                     1801
MoMuLV  CCCCTCCGCGCAGGAGGAAACGGACAGCTTCAA                                    TACTGGCCGTTCTCCTCTTCTGACCTTTACAACTGGAAAAATAATAACCCTTCT
        ProLeuArgAlaGlyGlyAsnGlyGlnLeuGln                                    TyrTrpProPheSerSerSerAspLeuTyrAsnTrpLysAsnAsnAsnProSer

4-1     CCCCTGAGAGAGCAGCAGTATACTGGAATAGATGAGGATGGGCACATGGTGGAGAGTCGTGTTTTTGTGTACCAGCCCTTCACCTCTGCCGACCTTCTCAACTGGAAAAACAATACCCCGTCC
        ProLeuArgGluGlnGlnTyrThrGlyIleAspGluAspGlyHisMetValGluSerArgValPheValTyrGlnProPheThrSerAlaAspLeuLeuAsnTrpLysAsnAsnThrProSer
```

B

```
            \/pol      2661
MoMuLV  GAC TAG GGA GGT CAG GGT
        Asp End Gly Gly Gln Gly

4-1     GAC TGA GGG GGA CTG GGC
        Asp End Gly Gly Leu Gly
```

C

```
       7861                                                                                                     7936
4-1     Gln Asn Arg Leu Ala Leu Asp Tyr Leu Leu Ala Ala Glu Gly Gly Val Cys Gly Lys Phe Asn Leu Thr Asn Tyr Cys
         :   :   :           :   :       :               :   :   :       :
MoMuLV  Gln Asn Arg Arg Gly Leu Asp Leu Leu Phe Leu Lys Glu Gly Gly Leu Cys Ala Ala Leu Lys Glu Glu Cys Cys Phe
         :   :   :   :   :   :   :   :   :   :               :   :   :   :       :   :       :       :       :
HTLV-I  Gln Asn Arg Arg Gly Leu Asp Leu Leu Phe Thr Glu Gln Gly Gly Leu Cys Lys Ala Leu Gln Glu Gln Cys Arg Phe
```

FIG. 4. Conservation of C-type retroviral consensus sequences. (A) Fourteen highly conserved amino acid residues were identified (11, 12) with the N terminus of C-type retroviral DNAs (underlined, MoMuLV). Conservation of these residues in the human retroviral sequence is shown. Note that the insertion of 12 amino acids in clone 4-1 extends the usual distance between the first and second group of conserved amino acids. (B) Homology of deduced amino acids at the putative *gag-pol* junction of clone 4-1 with those of MoMuLV. (C) Correspondence of conserved amino acid sequences in the p15E region (3) of MoMuLV and in the p21 region of HTLV-I with sequences in the putative p15E region in human clone 4-1. Reference nucleotide numbers for MoMuLV or clone 4-1 refer to the proviral form of the sequences.

genomes occur in the putative human *gag* region (all localized to a segment corresponding to p30 (Fig. 1, positions 1630 to 2451). Nucleotide sequence data in the region of one of the four termination codons are available from another human retroviral clone, and the stop codon (Fig. 1, position 2011 to 2013) is replaced with one specifying arginine. The dot matrix analysis of sequence homology between clone 4-1 and MoMuLV in the *gag* region (Fig. 3) shows that polynucleotide sequence homology is greatest in the segment corresponding to p30 of MuLVs and least in the portion aligned with the MuLV p12 *gag* coding sequence.

Oroszlan et al. (11, 12) have analyzed the N-terminal sequences of mammalian retroviral p30 *gag* polypeptides and have identified 14 conserved amino acid residues characteristic of C-type retroviruses. Eleven of the 14 conserved amino acid residues are present in clone 4-1 (Fig. 4A). The first three residues (Pro-Leu-Arg) are a perfect match with the conserved N-terminal tripeptide and identify the putative N terminus of p30 in clone 4-1. An insertion of 12 amino acids in clone 4-1 relative to MoMuLV proviral DNA displaces the position of the remaining conserved residues which show an 8-of-11 colinear match with MoMuLV. An insertion at this position of p30 has been seen in several other proviral DNAs (12). The nucleotide sequence of a second human retroviral clone (4-14) was determined in the same region (data not shown), and both the 12-amino-acid insertion and the same 11-of-14 deduced amino acids were conserved.

The *pol* gene of clone 4-1. The equivalent of the *gag-pol* junction in clone 4-1 was identified by the presence of identical deduced amino acids (Asp-End-Gly-Gly) (Fig. 4B) in both the MoMuLV genome and the human retroviral sequences. Within the *pol* reading frame of clone 4-1, six termination codons and three nucleotide deletions occur. The sequence of several other human retroviral clones was determined in regions encompassing four of the stop codons and one of the nucleotide deletions. In each instance, the nucleotide sequences of a related retroviral segment "supplanted" a termination signal or contained an extra nucleotide that had been deleted from the *pol* region of clone 4-1. However, a termination codon located at position 6211 to 6213 (Fig. 1) was conserved in four different human clones (data not shown). This stop codon was aligned one amino residue 5′ to the stop codon for *pol* in MoMuLV.

A more detailed comparison of the amino acids over the entire 3,561-bp human *pol* region with those of MoMuLV is shown in Fig. 5. Overall, the putative human *pol* region shares 44% deduced amino acid homology with the MoMuLV *pol* gene. An interesting feature of this analysis is the presence of a well-demarcated conserved stretch of amino acids (Fig. 5, between arrows) within which are three regions, each having at least 70% homology to MoMuLV. Recent reports have suggested that retroviral *pol* genes encode a series of functions required for viral replication. For example, the N and C termini of reverse transcriptase have been positioned within the region of nucleotides 3047 to 5059 of MoMuLV proviral DNA (T. D. Copeland, G. F. Gerard, C. G. Hixon, and S. Oroszlan, personal communication). This corresponds to a 1,994-bp segment in clone 4-1 which is situated between nucleotides 3007 and 5001 (Fig. 1). The highly conserved amino acids between clone 4-1 and MoMuLV (shown between the arrows in Fig. 5) occur in a sequence of 1,869 nucleotides encompassing 3,120 to 4,989 bp (Fig. 1), essentially representing the putative reverse transcriptase coding region of clone 4-1. These three highly conserved regions could represent potential functional domains of the enzyme.

At the position corresponding to the *env* splice acceptor site in MoMuLV, clone 4-1 has the sequence TAGCTACAG (position 5869 to 5877, Fig. 1), which represents a 6-of-9 nucleotide identity with the MoMuLV splice acceptor sequence. In human clone 4-14, a C/T substitution for the first base results in a 7-of-9 identity with Moloney splice acceptor sequence and, incidentally, corrects one of the "premature" termination codons in clone 4-1. However, both of these sequences deviated from the consensus splice acceptor sequence, Y-YYY-CAG (20), by a single base.

Characteristics of the envelope region. The nucleotide and

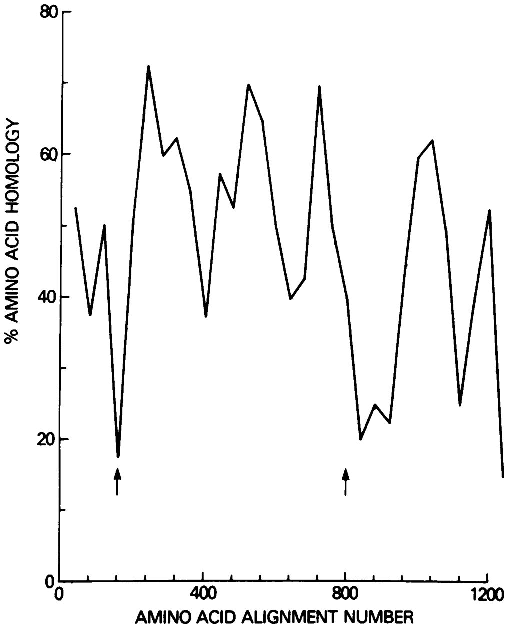

FIG. 5. Percentage of amino acid homology between the human retroviral sequence and MoMuLV in the *pol* region. The percent homology of successive groups of 40 deduced amino acids in an aligned sequence of clone 4-1 versus MoMuLV in the *pol* region is shown. Amino acid alignment in clone 4-1 was begun at nucleotide position 2651 (Fig. 1), and each increment of 40 amino acids represents a 120-bp segment. Arrows indicate the N and C terminus of the reverse transcriptase coding region (see the text).

deduced amino acid sequences of the putative *env* region of clone 4-1, unlike *gag* and *pol*, do not match the corresponding sequences in MoMuLV (22), so the correct reading frame cannot be ascertained by alignment. However, two of the three reading frames contain numerous termination codons and, therefore, have been excluded. The remaining *env* reading frame is open for 1,283 nucleotides (bases 6236 to 7519, Fig. 1); four other clones sequenced in this region have a base substitution at position 7521 which converts the TAG stop codon to TGG, a codon for tryptophan. With this substitution, the consensus putative human *env* region extends to 2,012 bp and ends at position 8248 (Fig. 1).

By analogy to MoMuLV, the human *env* RNA should be spliced from the *gag* leader to the 3′ portion of *pol* and read in a different frame than *pol*. As pointed out in the previous section, a potential splice acceptor site (Fig. 1, position 5869 to 5877) was identified in the *pol* region of clone 4-1. However, the sequence following the splice acceptor encodes a methionine followed by stop codons in the long open *env* reading frame. This ambiguity involving the 5′ terminus of a putative human *env* gene is further compounded by analogy with MoMuLV proviral DNA in which the *env* initiation methionine is 5′ to the *pol* termination, resulting in overlapping reading frames for *pol* and *env*. No other methionine codon was present in clone 4-1 or in three other human retroviral clones sequenced in the region of a potential *pol-env* overlap. In the *env* reading frame of clone 4-1 and three other human clones for which sequence data are available (Fig. 6), the first methionine not subsequently followed by a termination codon is situated 3′ to the end of *pol*. A second methionine, located 13 codons 3′ to the *pol* terminus, is conserved in three of four human clones examined; in the fourth, a single base substitution converts a valine to a methionine (Fig. 6). This consensus ATG, 13 amino acid residues 3′ to the *pol* terminus, is the only methionine codon in this region fulfilling secondary ATG context criteria proposed by Kozak (6) having the form $^{A}_{G}$NNAUGG.

Eight potential glycosylation sites of the form Asn-X-Thr/Ser (Fig. 1, boxed amino acids) exist within the *env* region, and six are located in that portion of *env* which would correspond to gp70 of MoMuLV (22). Sequence data from other human endogenous clones showed overall *env* amino acid homology to clone 4-1 of 70 to 90%. Notwithstanding the variation in homology among these clones, potential glycosylation sites and their positions in clone 4-1 were maintained in all other clones. This finding suggests that there may be conservation of glycosylation sites despite substantial drift of amino acid sequences among some of the human endogenous retroviral structures.

Cianciolo et al. (3) have compared the amino acid sequence of the p15E transmembrane protein from several different mammalian retroviruses and found a stretch of 26 residues exhibiting extensive conservation. Many of these amino acids are present in clone 4-1, mapping between nucleotides 7859 and 7936 (Fig. 4C). This area of homology is positioned 453 bp upstream from the 3′ LTR in clone 4-1; in MoMuLV proviral DNA it is located 425 bp from the 3′ LTR within the p15E region. The gp70 of Mo (22) or AKV (4, 7) MuLVs terminates in a basic amino acid pair, Lys-Arg (7,627 to 7,632 bp in Mo proviral DNA [22]). This dipeptide situated in the same relative location in clone 4-1 DNA (Fig. 1, nucleotides 7622 to 7627) would position the N terminus of a putative p15E region at nucleotide 7628. The highly conserved p15E amino acid sequence identified by Cianciolo et al. (3) is located 69 amino acids downstream from the N terminus of MoMuLV p15E and 77 amino acids downstream from the beginning of the putative human p15E coding sequence.

In AKV MuLV, a hydrophobic 29-amino-acid sequence near the carboxy terminus of p15E is followed by a basic amino acid and a C-terminal hydrophilic region. It was proposed that the hydrophobic sequence was that portion of

```
       end pol \/                              6241                                       6271
4-1    GAA GCT GAC TGG TCC ACG CAC GGC CGA AGC ATG CAG AAG CTC ATC ATG GGA TTC ATT TTT CTT AAA TTT TGG ACT TAT ACA GTA AGG
       Glu Ala Asp Trp Ser Thr His Gly Arg Ser Met Gln Lys Leu Ile Met Gly Phe Ile Phe Leu Lys Phe Trp Thr Tyr Thr Val Arg

5-1    GAA GCT GAC TGG TCC ACG CAC GGC CGA AGC CTG AGG AAG CTC ATC ATG AGA TTC ATT TTT CTT AAA TTT TGG ACT TAT ACA GTA AGG
       Glu Ala Asp Trp Ser Thr His Gly Arg Ser Leu Arg Lys Leu Ile Met Arg Phe Ile Phe Leu Lys Phe Trp Thr Tyr Thr Val Arg

pE40   GAA GCT GAC TGG TCT ATG TGC GGC CAA AGC ATG AGG AAA CTC ATC ATG GGA CTT ATT CTC CTT AAA ATT TGG ACT TGT AGA ATA AGG
       Glu Ala Asp Trp Ser Met Cys Gly Gln Ser Met Arg Lys Leu Ile Met Gly Leu Ile Leu Leu Lys Ile Trp Thr Cys Arg Ile Arg

4-14   GAA GCT GAC TGG TCT ATG CAC AGC CAA NGC ATA AGG AAA CTC ATC GTG GGA CTC ATT TTT CTC AAA ATT TGG ACT TGT TCA ATA AGA
       Glu Ala Asp Trp Ser Met His Ser Gln     Ile Arg Lys Leu Ile Val Gly Leu Ile Phe Leu Lys Ile Trp Thr Cys Ser Ile Arg
```

FIG. 6. Potential *env* initiation codons in four human retroviral clones. The first ATG codons in the open *env* reading frame of four human clones are shown. The C-terminal end of *pol* in a different reading frame is indicated. Nucleotide numbers refer to the clone 4-1 sequence (Fig. 1). A single base (N) in clone 4-1 could not be identified.

p15E which traversed the viral membrane, thereby anchoring p15E to the virus surface (7). An analogous hydrophobic (5) sequence of 36 amino acids is found in the C-terminal p15E region of clone 4-1 (Fig. 1, position 8054 to 8161), followed by a lysine residue and a subsequent hydrophilic region.

LTRs of clone 4-1. The nucleotide sequence of human LTRs including clone 4-1 has been previously published (23). They contain putative TATA boxes and polyadenylation signals as well as imperfect inverted terminal repeats. The 5′ LTR of clone 4-1 is 495 bp, 4 bp shorter than the 3′ LTR due to a single 4-bp deletion. The LTRs are 95% homologous to each other and contain 21 single nucleotide differences.

Origin of human retroviral sequences. The nucleotide sequence shown in Fig. 1 and its alignment with the *gag* and *pol* genes of MoMuLV shown in Fig. 3 provide strong evidence that full-length retrovirus-like DNA segments are present in human chromosomal DNA. The 35 to 50 copies (23) of these sequences present in the human genome are bounded by terminally repeating elements containing most of the structural features of retroviral LTRs. As mentioned above, the nearly 8,000 bp lying between the two LTRs is organizationally similar to other C-type retroviral DNAs, retaining such landmarks as (i) a putative splice donor and acceptor at positions appropriate for the processing of a potential *env* mRNA; (ii) 11 of the 14 p30 *gag* consensus amino acids (Fig. 4A); and (iii) a highly conserved C-terminal *gag*/N-terminal *pol* junction (Fig. 4B).

On the other hand, the human retroviral sequences we have studied possess several unusual features that set them apart from known infectious C-type viruses. These segments undoubtedly entered the human germ line at some point in the distant past as a result of exogenous infection. Unlike the C-type mammalian retroviruses encountered today, all of which contain a pbs for tRNA^{Pro}, the "prehistoric" progenitor viruses that gave rise to the human endogenous retroviral segments possessed a tRNA^{Glu}-like pbs. Since a variety of tRNA pbs's have been reported for different classes of retroviruses (2, 13), it is not clear whether the tRNA^{Glu} pbs's associated with the human segments represent a major structural-functional alteration associated with these retroviral sequences or is merely the "signature" of an extinct virus.

One perplexing feature of the full-length human endogenous retroviral segments has been the lack of polynucleotide and deduced amino acid sequence homology of the putative *env* region with analogous genes of known infectious murine, feline, simian, and human retroviruses. Divergence of *env* genes even among C-type viruses isolated from the same species (16) has been previously reported and certainly could be anticipated in view of the role of the retroviral envelope glycoprotein in binding to cell receptors during the initial stages of viral infection. During investigations regarding the origin of the human endogenous retroviral sequences, we recently carried out Southern blot-hybridization experiments of mammalian genomic DNAs by employing human *env* DNA probes. Quite unexpectedly, the results of these experiments indicate that all primates tested contain multiple copies of the human *env* segment linked to sequences that anneal to human *gag* and *pol* probes (R. Repaske, A. B. Rabson, T. Bryan, and M. A. Martin, manuscript in preparation). These monkey endogenous retroviral segments hybridize to the human *env* probe even under high-stringency conditions and, after digestion with some restriction enzymes, generate fragments that comigrate with cleavage products of human genomic DNA (Repaske et al., in preparation). Nucleotide sequencing of these simian DNA segments should reveal their relationship to the full-length human endogenous sequences and indicate whether they contain an associated tRNA^{Glu} pbs.

Clone 4-1 contains several nucleotide substitutions or deletions or both that introduce stop codons or changes in the reading frame (Fig. 1), thereby precluding its expression in the form of infectious virus. Partial nucleotide sequencing of at least four other human endogenous retroviral segments has revealed the presence of similar alterations rendering them defective. Despite these findings, some of the human retroviral sequences are expressed since *env*-reactive, polyadenylated RNAs have been detected in human placentas and colon carcinoma cells and *gag-pol* RNA is present in T-cell leukemia cells (15). Recent experiments, involving rabbit antiserum raised against synthetic oligopeptides whose structure is based on the nucleotide sequence of 4-1 DNA and related clones, indicate the presence of specific retroviral-related proteins in human cells (A. Adachi, W. L. Maloy, A. B. Rabson, A. M. Lew, J. E. Coligan, and M. A. Martin, manuscript in preparation). It is of interest that antibodies raised against a synthetic undecapeptide corresponding to a deduced *gag* sequence from another human endogenous retroviral clone (*erv*-1) specifically reacted with a M_r 75,000 polypeptide expressed in normal first-trimester placenta and cultured chorio-carcinoma cells (24). The detection of viral RNA and protein in the absence of virus production is reminiscent of the situation in certain inbred strains of mice (e.g., 129) which do not harbor inducible MuLVs but do express endogenous retroviral genetic information. Thus humans, like mice, have acquired multiple copies of retroviral DNA during their evolutionary history. A majority of these germ-line retroviral sequences have undergone multiple alterations that prevent their expression as infectious viruses; however, individual subgenomic retroviral segments, if associated with an intact LTR or promoter element, could be expressed as RNA and proteins. With the passage of time these endogenous retroviral genes could evolve in their eucaryotic chromosomal environment to encode important cellular products.

ACKNOWLEDGMENTS

We thank John Owens and Jacob V. Maizel for performing the dot matrix analysis for this publication and other analyses made during the course of this investigation.

LITERATURE CITED

1. **Blomquist, M. C., L. T. Hunt, and W. C. Barker.** 1984. Vaccinia virus 19-kilodalton protein: relationship to several mammalian proteins, including two growth factors, Proc. Natl. Acad. Sci. U.S.A. **81**:7363–7367.
2. **Chen, H. R., and W. C. Barker.** 1984. Nucleotide sequences of the retroviral long terminal repeats and their adjacent regions. Nucleic Acids Res. **12**:1767–1778.
3. **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.
4. **Herr, W.** 1984. Nucleotide sequence of AKV murine leukemia virus. J. Virol. **49**:471–478.
5. **Hopp, T. P., and K. R. Woods.** 1981. Prediction of protein antigenic determinants from amino acid sequences. Proc. Natl. Acad. Sci. U.S.A. **78**:2173–2192.
6. **Kozak, M.** 1983. Comparison of initiation of protein synthesis in procaryotes, eucaryotes, and organelles. Microbiol. Rev.

47:1–45.

7. **Lenz, J., R. Crowther, A. Straceski, and W. Haseltine.** 1982. Nucleotide sequence of the AKV *env* gene. J. Virol. **42:**519–529.

8. **Maniatis, T., E. F. Fritsch, and J. Sambrook.** 1982. Molecular cloning, p. 265–285. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.

9. **Martin, M. A., T. B. Bryan, S. Rasheed, and A. S. Khan.** 1981. Identification and cloning of endogenous retroviral sequences present in human DNA. Proc. Natl. Acad. Sci. U.S.A. **78:**4892–4896.

10. **Maxam, A. M., and W. Gilbert.** 1980. Sequencing endlabeled DNA with base-specific chemical cleavages. Methods Enzymol. **65:**499–560.

11. **Oroszlan, S., T. D. Copeland, R. V. Gilden, and G. J. Todaro.** 1981. Structural homology of the major internal proteins of endogenous type C viruses of two distantly related species of old world monkeys. Virology **115:**262–271.

12. **Oroszlan, S., T. Copeland, G. Smythers, M. R. Summers, and R. V. Gilden.** 1977. Comparative primary structure analysis of the p30 protein of woolly monkey and gibbon type C viruses. Virology **77:**413–417.

13. **Ou, C. Y., L. R. Boone, and W. K. Yang.** 1983. A novel sequence segment and other nucleotide structural features in the long terminal repeat of a BALB/c mouse genomic leukemia virus-related DNA clone. Nucleic Acids Res. **11:**5603–5620.

14. **Queen, C. L., and L. J. Korn.** 1980. Computer analysis of nucleic acids of proteins. Methods Enzymol. **65:**595–609.

15. **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.

16. **Repaske, R., R. R. O'Neill, A. S. Kahn, and M. A. Martin.** 1983. Nucleotide sequence of the *env*-specific segment of NFS-Th-1 xenotropic murine leukemia virus. J. Virol. **46:**204–211.

17. **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. U.S.A. **80:**678–682.

18. **Seiki, M., S. Hattori, and M. Yoshida.** 1982. Human adult T-cell leukemia virus: molecular cloning of the provirus DNA and the unique terminal structure. Proc. Natl. Acad. Sci. U.S.A. **79:**6899–6902.

19. **Sekiya, T., Y. Kuchino, and S. Nishimura.** 1981. Mammalian tRNA genes: nucleotide sequence of rat genes from $tRNA^{Asp}$, $tRNA^{Gly}$ and $tRNA^{Glu}$. Nucleic Acids Res. **9:**2239–2250.

20. **Sharp, P.** 1981. Speculations on RNA splicing. Cell **23:**643–646.

21. **Shimotohno, K., D. W. Golde, M. Miwa, T. Sugimura, and I. S. Y. Chen.** 1984. Nucleotide sequence analysis of the long terminal repeat of human T-cell leukemia virus type II. Proc. Natl. Acad. Sci. U.S.A. **81:**1079–1083.

22. **Shinnick, T. M., R. A. Lerner, and J. G. Sutcliff.** 1981. Nucleotide sequence of Moloney murine leukemia virus. Nature (London) **293:**543–548.

23. **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.

24. **Suni, J., A. Narvanen, T. Wahlstrom, M. Aho, R. Pakkanen, A. Vaheri, T. Copeland, M. Cohen, and S. Oroszlan.** 1984. Human placental syncytiotrophoblastic M_r 75,000 polypeptide defined by antibodies to a synthetic peptide based on a cloned human endogenous retroviral DNA sequence. Proc. Natl. Acad. Sci. U.S.A. **81:**6197–6201.

25. **Wilbur, W. J. and D. J. Lipman.** 1983. Rapid similarity searches of nucleic acid and protein data banks. Proc. Natl. Acad. Sci. U.S.A. **80:**726–730.