

## The Nucleotide Sequence of the *env* Gene from the Human Provirus ERV3 and Isolation and Characterization of an ERV3-Specific cDNA<sup>1</sup>

MAURICE COHEN,<sup>2</sup> MARILYN POWERS, CATHERINE O'CONNELL,  
AND NOBUYUKI KATO

*Laboratory of Molecular Virology and Carcinogenesis, LBI-Basic Research Program, National Cancer Institute-Frederick Cancer Research Facility, Frederick, Maryland 21701*

*Received July 5, 1985; accepted September 2, 1985*

The nucleotide sequence of the *env* gene of a previously described human provirus (ERV3) has been determined beginning near the 3'-end of the *pol* gene and continuing through the 3'-LTR. Analysis of the nucleotide sequence revealed the presence of a long open reading frame of 1944 nucleotides that is capable of encoding a polypeptide that has characteristics of other retroviral glycoproteins and transmembrane proteins. These include the presence of seven potential glycosylation sites, a typical glycoprotein-transmembrane protein cleavage sequence, and amino acid homologies to the glycoproteins and transmembrane proteins of other retroviruses. Further, we have isolated an ERV3-specific cDNA clone from a library prepared from liver RNA of a 20-week human fetus. DNA sequence analysis of this clone revealed that it is identical to the ERV3 genomic clone in the 1110 nucleotides that were sequenced. © 1985 Academic Press, Inc.

### INTRODUCTION

In the last few years, it has become increasingly clear that the DNA of man, like that of other vertebrates, contains many integrated retroviral genomes. Although endogenous retroviruses are not normally associated with any pathologic state, the awareness that exogenous retrovirus infection poses a significant health threat to humans has caused many to consider the etiologic potential of human proviruses as well. This potential is demonstrated in AKR mice where endogenous xenotropic-like retroviral *env* sequences are involved in the formation of MCF recombinants, which are necessary intermediates in the induction of thymic lymphomas (Chattopadhyay *et al.*, 1982). Thus, nondefective or partially nondefective human proviruses may provide a source of new human retroviral recombinants.

<sup>1</sup> The U. S. Government's right to retain a nonexclusive royalty-free license in and to the copyright covering this paper, for governmental purposes, is acknowledged.

<sup>2</sup> Author to whom requests for reprints should be addressed.

The presence of retroviral sequences in the human genome was first detected by low stringency hybridization of BaEV or R-MuLV cDNA to human cell DNA (Benveniste and Todaro, 1974, 1976; Kominami *et al.*, 1980). Definitive evidence that human DNA indeed contains retroviral sequences was obtained from DNA sequence analyses of cloned DNA from human recombinant libraries using retroviral probes in non-stringent hybridization conditions (Bonner *et al.*, 1982; Repaske, 1983b). One multicopy family of human sequences (Martin *et al.*, 1981; Repaske *et al.*, 1983b, 1985) is composed of both full-length and truncated defective proviruses. Clones related to the type B retrovirus, mouse mammary tumor virus (MMTV), have also been isolated (Callahan *et al.*, 1982, 1985; Westley and May, 1984) and hybridization studies indicate MMTV-related sequences are moderately repetitive in human DNA. Further, Noda *et al.* (1982) isolated from a human library several clones related to baboon endogenous virus (BaEV) LTR.

This group previously described two human proviral clones, ERV1 (Bonner *et al.*, 1982; O'Brien *et al.*, 1983) and ERV3 (O'-

Connell and Cohen, 1984; O'Connell *et al.*, 1984) that by DNA sequence analyses are significantly related to Moloney murine leukemia virus (M-MuLV), BaEV, and the human proviral genomes described by Martin *et al.* (1981). Unlike the latter family of human sequences, ERV1 and ERV3 are apparently single copy in humans and have been mapped, respectively, to chromosomes 18 and 7. Both proviruses are the result of ancient integrations in the primate lineage and both are replication defective: ERV1 has no 5' LTR and ERV3 contains in-frame terminator codons in the *gag* and *pol* genes. In this report we further characterize the ERV3 provirus and show that its *env* glycoprotein gene is capable of encoding a polypeptide product. In addition, we report the isolation and characterization of an ERV3-specific cDNA clone isolated from a human fetal cDNA library.

#### MATERIALS AND METHODS

**Human retroviral clones.** Isolation of a clone containing the ERV3 provirus from a  $\lambda$ -library of human cellular DNA and derivation of pBR322 subclones, pRI4.8 and pHP1.7, were previously described (O'Connell *et al.*, 1984). Two *BalI*-*PstI* fragments from the ERV3 *env* region were subcloned in the *SmaI*-*PstI* sites of pUC8 (Vieira and Messing, 1982): BP6 (nucleotides -190 to 1499) and BP7 (nucleotides 1597 to 2273) (see Fig. 1 for nt coordinates). Another subclone, termed DD2, was derived from pRI4.8 by cleavage with *KpnI*, followed by *Bal*-31 digestion, recleavage with *EcoRI*, and subcloning of the approximately 1.86-kb fragment in the *EcoRI*-*SmaI* sites of M13mp8. The subclone extends from nt 2545 to the *EcoRI* site in the 3'-flanking sequence.

A human 20-week fetal liver cDNA library cloned in the *EcoRI* site of  $\lambda$ -gt10 was the kind gift of Dr. E. F. Fritsch. The 1726 nt *env* fragment from the ERV3 subclone, pHP1.7, was isolated and used as a probe in hybridization to filters containing DNA from  $9 \times 10^5$  individual plaques (Benton and Davis, 1977). The final hybridization wash was at 0.08 M Na<sup>+</sup>, 68°. DNA from one positive plaque ( $\lambda$ HF9) was di-

gested with *EcoRI* and the 2851 nt insert was subcloned in the *EcoRI* site of pUC8. Also, an *SspI*-*EcoRI* 1.2-kb fragment at the 3'-end of the cDNA clone was subcloned in the *SmaI*-*EcoRI* sites of the Riboprobe vector, pSP64 (Promega Biotec).

**DNA sequence analysis.** Restriction fragments of ERV3 genomic and cDNA subclones were labeled at their 3'-ends using DNA polymerase I (Klenow fragment, Bethesda Research Labs) and [<sup>32</sup>P]dNTPs (Amersham). All ERV3 DNA sequences were obtained by the chemical degradation method of Maxam and Gilbert (1980). Both strands of DNA were sequenced where necessary to clarify uncertainties. Occasional compressions in the DNA sequence were resolved by performing the electrophoresis at 65°.

#### RESULTS AND DISCUSSION

**The envelope gene.** A restriction map of the 3' half of the ERV3 genome is shown in Fig. 1 beginning with the *EcoRI* site in the *pol* gene that was previously shown to share significant homologies with M-MuLV near M-MuLV nucleotide (nt) 4980 (O'Connell *et al.*, 1984). Continuous DNA sequence spanning 3387 nucleotides at the 3' end of the ERV3 provirus was determined (Fig. 2). The nucleotide sequence was translated into the three reading frames. Figure 3 depicts the initiator and terminator codons in each reading frame. The translated sequence in frame 3 contains amino acid homologies to the M-MuLV *pol* gene near the M-MuLV COOH terminus (underlined in Fig. 2). These homologies place ERV3 nt 1 (Fig. 2) at M-MuLV nt 5389 (Shinnick *et al.*, 1981). If ERV3 contains a *pol* gene of M-MuLV length, and if the ERV3 *pol* and *env* gene reading frames overlap as in M-MuLV, the *env* glycoprotein initiator codon would be expected to occur near nt 380 (Fig. 1). Frame 1 in the ERV3 sequence contains a long open reading frame of 1944 nucleotides beginning at nt 559 (Figs. 2, 3). The first two potential initiator codons in this open frame occur at nucleotides 769 and 808. This would place the ERV3 *env* initiator 389 or 428 nucleotides 3' of the ap-

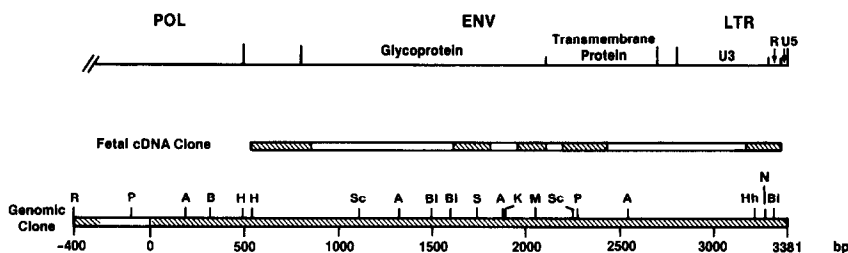


FIG. 1. Restriction map of the ERV3 genome 3' end. Top: the ERV3 genetic map. Gaps between *pol-env* and *env*-LTR represent apparent noncoding regions. Middle: bar represents extent of cDNA clone (see text). Crosshatching indicates sequenced regions (discussed in text). Bottom: restriction map of genomic clone. R, *EcoRI*; H, *HindIII*; P, *PstI*; A, *AhaIII*; Nd, *NdeI*; B, *BamHI*; BI, *BalI*; S, *StuI*; K, *KpnI*; Hh, *HhaI*; N, *NciI*; Sc, *ScaI*; M, *MstII*. Crosshatching indicates region that was sequenced.

proximate site of M-MuLV glycoprotein initiation. Thus the *pol* and *env* reading frames of ERV3, like those of the AIDS retrovirus isolates (Muesing *et al.*, 1985; Ratner *et al.*, 1985; Sanchez-Pescador *et al.*, 1985; Wain-Hobson *et al.*, 1985), are not overlapping. A potential *env* splice acceptor sequence, -CCTTCTCCCCCAGG-, was identified in the ERV3 sequence beginning 112 nt upstream of the 5' ATG codon. This sequence is a perfect fit to that of a consensus splice acceptor (Mount, 1982).

The encoded amino acid sequence of the long open reading frame (Fig. 2) has certain characteristics in common with typical retroviral glycoproteins. First, it contains 8 potential glycosylation signals (7 within the putative glycoprotein region), sequences of the form Asn-X-Ser/Thr (Marshall, 1974) (Fig. 2). By comparison, the glycoproteins of M-MuLV (Shinnick *et al.*, 1981), bovine leukemia virus (BLV) (Rice *et al.*, 1984), and feline leukemia virus (FeLV) (Elder and Mullins, 1983) contain 7, 8, and 12 such signals, respectively. Also present in the open reading frame is a sequence of basic amino acids, -Lys-Arg-Lys-Ser-Lys-Arg- (nt 2083-2100), that may represent the proteolytic cleavage site between the glycoprotein and transmembrane protein (Fig. 2). Typically this occurs following the final arginine.

The ERV3 envelope protein amino acid sequence and that of other retroviral envelope glycoproteins was compared using the computer program, ALIGN (Dayhoff,

1976). The ERV3 glycoprotein sequence revealed limited but significant homology to the encoded amino acid sequence of two xenotropic MuLV genomes, NFS-Th-1 (Repaske *et al.*, 1983a) and NZB (O'Neill *et al.*, 1985) and these homologies are underlined in Fig. 2. Figure 4 shows the homologies between ERV3 and this region in several other type C retroviruses that could be aligned by their homologies with the xenotropic MuLV genomes. These include AKV (Lenz *et al.*, 1982), M-MuLV (Shinnick *et al.*, 1981), M-MCF (Bosselman *et al.*, 1982), F-MCF (Adachi *et al.*, 1984), and FeLV (Elder and Mullins, 1983). In each case, the sequence, at least through the common cysteine residue, lies within a hydrophobic domain as calculated by the program of Kyte and Doolittle (1982). The highly conserved nature of this region is also indicated by its location 13 residues downstream of a potential glycosylation site in M-MCF, FeLV, NZB, and NFS-Th-1, and 12 residues downstream of a potential site in ERV3. In each case the glycosylation site contains threonine rather than serine. The conserved region in the several type C retroviruses is 73-74 amino acids downstream of the methionine that is the initiator of glycoprotein synthesis. Thus, the envelope initiator in ERV3 may be the methionine (nt 769) located 92 residues upstream of the conserved region or the second methionine (nt 808) that is 79 residues distant. The nucleotide sequence around the two ATGs would suggest that

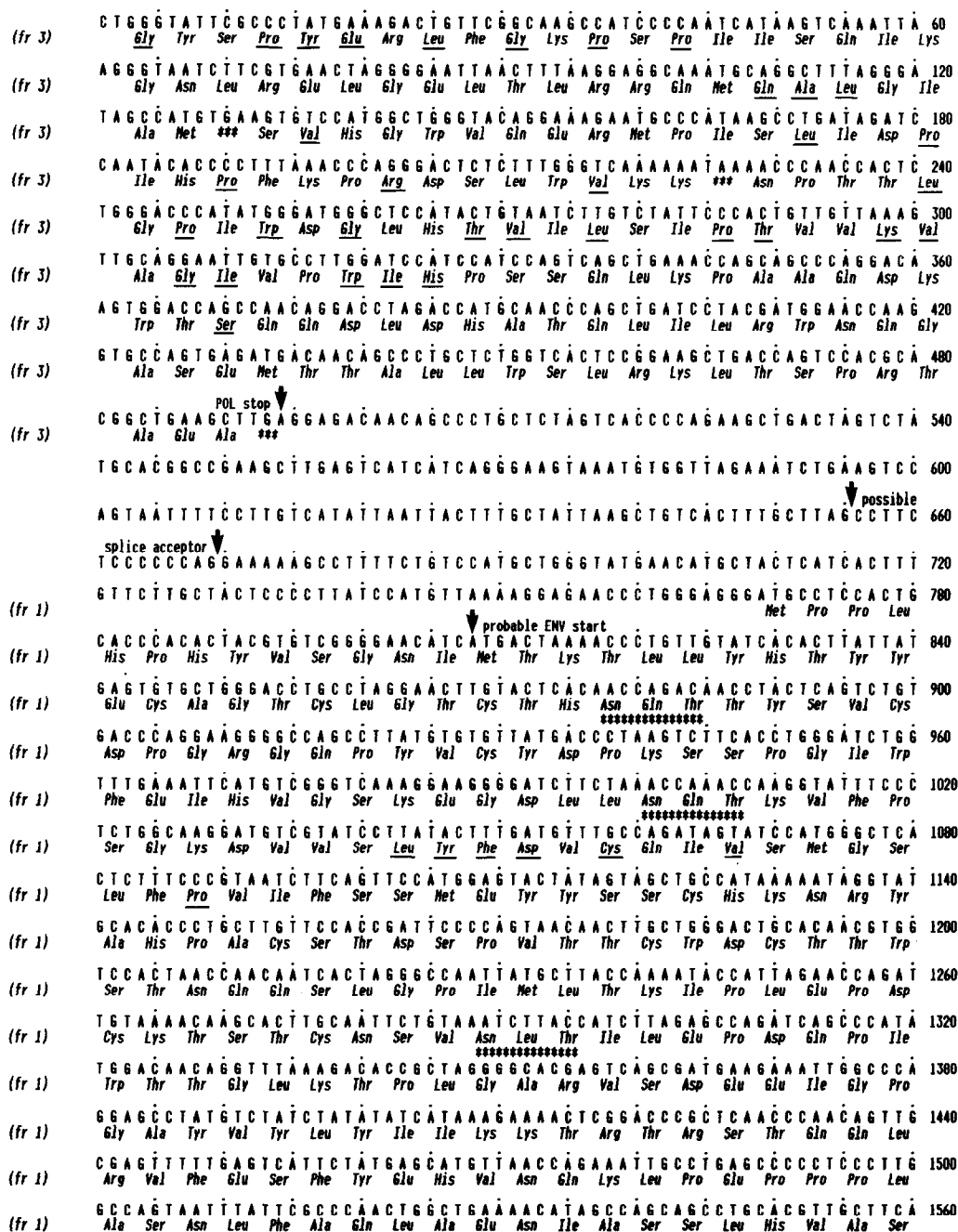


FIG. 2. DNA sequence of the ERV 3'-end. Nucleotide number 1 begins in the *pol* gene corresponding to M-MuLV nucleotide 5153 (Shinnick *et al.*, 1981). Encoded amino acid homologies to M-MuLV in *pol* gene, to NFS-Th-1 and NZB in glycoprotein region of gene *env* (Repaske *et al.*, 1983a; O'Neill *et al.*, 1985), and to M-MuLV in *env* p15E, are underlined. Three asterisks indicate terminator codons while underlining with asterisks represents potential glycosylation sites. The ERV 3'-LTR begins at nucleotide 2799 (O'Connell and Cohen, 1984).

(fr 1)	TGTTATGCTCTGTGGGGGAATGAACATGGGAGACCAATGGCCATGGGAAGCAAGGGAACTA	1620
	Cys Tyr Val Cys Gly Gly Met Asn Met Gly Asp Gln Trp Pro Trp Glu Ala Arg Glu Leu	
(fr 1)	ATGCCCAAGATAAATTTCACTAACCAGCTCTTCCCTCGAACCTGCACCATCAAGTCAG	1680
	Met Pro Gln Asp Asn Phe Thr Leu Thr Ala Ser Ser Leu Glu Pro Ala Pro Ser Ser Gln	
(fr 1)	AGCACTCTGGTCTTAAACCTCCATTATGGAAATTTCTGATTGCTCGCTGGGGAAAG	1740
	Ser Ile Trp Phe Leu Lys Thr Ser Ile Ile Gly Lys Phe Cys Ile Ala Arg Trp Gly Lys	
(fr 1)	GCCTTTACAGACCCAGTAGGAGAGATTAACCTGCTAGGACAACAATATTACAACGAGACA	1800
	Ala Phe Thr Asp Pro Val Gly Glu Leu Thr Cys Leu Gly Gln Gln Tyr Tyr Asn Glu Thr	
(fr 1)	CTAGGAAAGACTTTATGGAGGGGCAAAAGCAATAATTCTGAATCACCACACCCCAAGCCCA	1860
	Leu Gly Lys Thr Leu Trp Arg Gly Lys Ser Asn Asn Ser Glu Ser Pro His Pro Ser Pro	
(fr 1)	TTCTCTCGTTTCCCATCTTTAAACCATTTCTGGTACCAACTTGAAGCTCCAATAACCTGG	1920
	Phe Ser Arg Phe Pro Ser Leu Asn His Ser Trp Tyr Gln Leu Glu Ala Pro Asn Thr Trp	
(fr 1)	CAGGCACCTCTGGGCTCTACTGGATCTGTGGGCCACAAGCATATCGACAACTGCGCAGCT	1980
	Gln Ala Pro Ser Gly Leu Tyr Trp Ile Cys Gly Pro Gln Ala Tyr Arg Gln Leu Pro Ala	
(fr 1)	AAATGGTCAGGGGCGCTGTGTACTGGGGACAATTAGGCCGTCCTTCTTCTTAATGCCCTA	2040
	Lys Trp Ser Gly Ala Cys Val Leu Gly Thr Ile Arg Pro Ser Phe Phe Leu Met Pro Leu	
(fr 1)	AAACAGGGAGAAAGCCTTAGGATACCCATCTATGATGAAACTAAAGGAAAAGCAAAAGA	2100
	Lys Gln Gly Glu Ala Leu Gly Tyr Pro Ile Tyr Asp Glu Thr Lys Arg Lys Ser Lys Arg	
(fr 1)	GGCACTAACTATAGGAGATTGGAGGACAGTGAATGGCTCCTGAAAGAAATATTAATAT	2160
	Gly Ile Thr Lys Asp Trp Lys Asp Ser Glu Trp Pro Gln Trp Ala Ile Ile Thr Trp	
(fr 1)	TATGGCCCAAGCCACCTGGGCGAGAGATGGAAATGTGGGGAATACCGCACCCCACTTTACATG	2220
	Tyr Gly Pro Ala Thr Trp Ala Glu Asp Gly Met Trp Gly Tyr Arg Thr Pro Val Tyr Met	
(fr 1)	CTTAACCGCATTTATAGATTGCAAGGACGATCTAGAAATCATTACCAATGAAGCTGAGGGG	2280
	Leu Asn Arg Ile Ile Arg Leu Gln Ala Val Leu Glu Ile Ile Thr Asn Glu Thr Ala Gly	
(fr 1)	GCCTTGAATCTGCTTGGCCAGCAAGCCACAAAATGAGAAATGTCTATTATCAAAATAGA	2340
	Ala Leu Asn Leu Ala Gln Ala Thr Lys Met Arg Asn Val Ile Tyr Gln Asn Arg	
(fr 1)	CTGGCCTTAGACTACTCTCTAGCCAGGAAGAGGGAGATGCGGAAAGTTCAGCCTTACT	2400
	Leu Ala Leu Asp Tyr Leu Leu Ala Gln Glu Glu Gly Val Cys Gly Lys Phe Ser Leu Thr	
(fr 1)	AATGCTGCTGCTGGAATTTGATGACGAAGGAAGGTTATCAAGAATAACTGCTAAAAATC	2460
	Asn Cys Cys Leu Glu Leu Asp Glu Lys Val Ile Lys Glu Ile Thr Ala Lys Ile	
(fr 1)	CAAAAGTTAGCTCACATCCAGTTGAGACTTGGAAAGSAAGTCTCCAGATTCCTTTTCT	2520
	Gln Lys Leu Ala His Ile Pro Val Gln Thr Trp Lys Gly *** Ser Pro Asp Ser Leu Phe	
(fr 1)	AGAGGTTGGTCTTATCCCTTGGAGGATTAAACCTTAGTACAATAGTCTTACGCCATAT	2580
	Arg Gly Trp Phe Leu Ser Leu Gly Phe Lys Thr Leu Val Gln Ile Val Leu Ala Ile	
(fr 1)	TTGGGAGTTTGCCTTATACCTTCTCTTACCCCTCATTTGTCAAAAATATCCAAACA	2640
	Leu Gly Val Cys Leu Ile Leu Pro Cys Leu Leu Pro Leu Ile Val Lys Asn Ile Gln Thr	
(fr 1)	GCCATAGAGGCTCTGTGGACAGACGGACTACCAACGACTAATGGCCCTAAGTAAGTAT	2700
	Ala Ile Glu Ala Leu Val Asp Arg Arg Thr Thr Thr Arg Leu Met Ala Leu Thr Lys Tyr	
	TAAACCCCTGCAAGAAAGAGCTACTTCCCTCTTGAAGTAATGAAGATAGTGCTTTC	2760
	***	
	begin LTR	
	TCTTAACTTTACTTATAAAAGCATCAAAGGGGGGAATGAAGCAGGAAATATAAAGGA	2820
	AAAAACAAGTAAAGGGGAAACAAGTCCCTTCTGACCAAGTCTGACTCACTCCAAAGTCTCTG	2880
	CTGGAGCTATGATAAATATCTGCAAGGCCAGGCAGGGGGCTCCGAGGAGGGCTCCAGGAG	2940
	CAGGGATGAGAAACAAGATTTCTCTTATCAGTTTCCCTGTTGAATATTCTCTCCCAATA	3000
	ACATTATTCTTTTGTCTGCTCTCACAACTATTTTGTAACTATTCTGCAAGCTGTGTAAG	3060
	GATTTTGTAAAGTCTTGTCTTTCTTTCTGTAGCATGGCAAGGTCACAAAGCATGTTTAAAG	3120
	TAAAGTAGGCTCATGTTGCAAACTCTGTTGTAAACCTGTACAGGATGATTAACCTGCT	3180
	TTGTTCTGCTTCTGTAAAGACTGCTTTCTACCTCGCAGGTTTTCGCCCAAAAACCCGACT	3240
	TGCCCCCTGCTGATGCATGTATAAAGTCAAGCCCGTCTTGTTCGGGCTCAGCCCTTG	3300
	poly (A) site	
	GATGTAATCCGCTGGGCCAGTGGCCACCTAAATAAACCTTCTCTGTGACCCAGTGAT	3360
	end LTR	
	CTCTCGGGCTCTCTGATACCCACAACA	3387

FIG. 2—Continued.

either could serve as a site of translation initiation since each has a purine in position -3; that is, three nucleotides upstream of the ATG. But the second ATG, having the sequence -ACATCATGG-, is closer to

the consensus eukaryotic initiation sequence, -CC<sub>3</sub>CCATGG- (Kozak, 1984), and is more likely to be the ERV3 *env* initiator. Alternatively, the initiator ATG in ERV3 *env* mRNA could be spliced on from a lo-

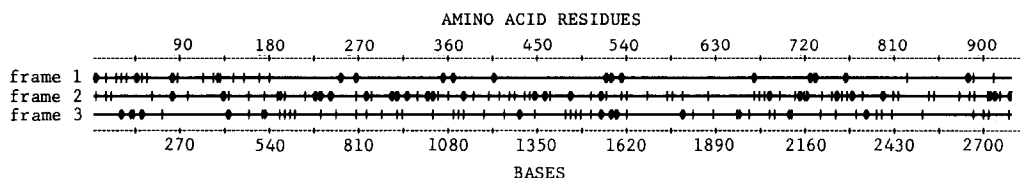


FIG. 3. Initiation and termination codons in the ERV3 3' region. Solid ovals depict methionine codons in each reading frame; vertical lines depict termination codons. Numbering of bases is the same as that in Fig. 1.

cation upstream in the provirus as it is in Rous sarcoma virus (RSV) (Hackett *et al.*, 1982; Ficht *et al.*, 1984).

Retroviral glycoprotein precursors are similar to other membrane-bound or secreted proteins in the presence of a hydrophobic leader peptide at their N-terminus (Blobel and Dobberstein, 1975; Lenz *et al.*, 1982). During maturation the leader peptide is proteolytically removed. In ERV3, neither potential initiator methionine is followed by a sequence sufficiently hydrophobic to penetrate the cellular membrane, according to the hydrophobicity index of Segrest and Feldman (1974). However, this index may not be an unequivocal indicator of signal peptide functionality; the bovine leukemia virus signal peptide isolated from an infectious clone also fails to meet the Segrest and Feldman criterion (Rice *et al.*, 1984).

The ERV3 *env* transmembrane protein should begin with the glycine residue (at nt 2101) following the cluster of basic amino acids. ERV3, like other retroviruses, contains a highly conserved region in the transmembrane protein described by

Cianciolo *et al.* (1984) between nucleotides 2332 and 2409. Homologies in this region between ERV3 and RSV (Schwartz *et al.*, 1983), M-MuLV (Shinnick *et al.*, 1981), BaEV (Stephens and Cohen, unpublished), reticuloendotheliosis virus (REV-A) (Wilhelmsen *et al.*, 1984), and the human T-cell leukemia viruses, HTLV I and II (Sodroski *et al.*, 1984) are shown in Fig. 5. The homologies to M-MuLV are underlined in Fig. 2. The transmembrane protein of typical type C retroviruses as well as MMTV (Redmond and Dickson, 1983) and RSV (Schwartz *et al.*, 1983) contains hydrophobic domains near each end which, by penetrating the cell membrane, help in anchoring the envelope protein in the membrane. The nucleotide sequence, however, indicates that ERV3 could encode only a truncated version of a transmembrane protein that lacks both a long hydrophobic region near its amino terminus, and an entire second hydrophobic domain because of the termination codon at nt 2502. Because the transmembrane protein would lack these apparently important features, the ERV3 envelope proteins would probably

ERV3	leu	tyr	phe	asp	val	cys	gln	ile	val	ser	met	gly	ser	leu	phe	pro
NFSxeno, NZBxeno	leu	tyr	phe	asp	leu	cys	asp	leu	val	gly	asp	tyr	trp	asp	asp	pro
AKV	leu	thr	pro	asp	leu	cys	met	leu	ala	leu	his	gly	pro	ser	tyr	trp
M-MuLV	leu	thr	pro	asp	leu	cys	met	leu	ala	his	his	gly	pro	ser	tyr	trp
M-MCF	leu	tyr	phe	asp	leu	cys	asp	leu	ile	gly	asp	asp	trp	asp	glu	thr
F-MCF	leu	tyr	phe	asp	leu	cys	asp	leu	met	gly	asp	asp	trp	asp	glu	thr
FeLV-B, GA	met	tyr	phe	asp	leu	cys	asp	ile	ile	gly	asn	thr	trp	asn	pro	ser
MMTV	pro	lys	tyr	pro	his	cys	gln	ile	ala	phe	lys	lys	asp	ala	phe	trp

FIG. 4. Amino acid homologies between ERV3 and other retroviruses in the *env* glycoprotein. Using the computer program ALIGN (Dayhoff, 1976), the ERV3 amino acid sequence deduced from the DNA sequence was aligned with those of the endogenous murine xenotropic retroviruses, NFS and NZB, and AKV, M-MuLV, M-MCF, F-MCF, and FeLV. Amino acid homologies with ERV3 beginning at ERV3 nt 1042 are enclosed in boxes.

ERV3	gln	asn	arg	leu	ala	leu	asp	tyr	leu	leu	ala	gln	glu	glu	gly
RSV	gln	asn	arg	ala	ala	ile	asp	phe	leu	leu	leu	ala	his	gly	his
M-MuLV	gln	asn	arg	arg	gly	leu	asp	leu	leu	phe	leu	lys	glu	gly	gly
BaEV	gln	asn	arg	arg	gly	leu	asp	leu	leu	thr	ala	glu	gln	gly	gly
REV-A	gln	asn	arg	arg	gly	leu	asp	leu	leu	thr	ala	gln	gln	gly	gly
HTLV-I	gln	asn	arg	arg	gly	leu	asp	leu	leu	phe	trp	glu	gln	gly	gly
HTLV-II	gln	asn	arg	arg	gly	leu	asp	leu	leu	phe	trp	glu	gln	gly	gly

val	cys	gly	lys	phe	ser	leu	thr	asn	cys	cys
gly	cys	glu	asp	val	ala	gly	—	met	cys	cys
leu	cys	ala	ala	leu	lys	glu	—	gly	cys	cys
ile	cys	leu	ala	leu	gln	glu	—	lys	cys	cys
ile	cys	leu	ala	leu	gln	glu	—	lys	cys	cys
leu	cys	lys	ala	ile	gln	glu	—	gln	cys	cys
leu	cys	lys	ala	ile	gln	glu	—	gln	cys	cys

FIG. 5. Amino acid homologies between ERV3 and other retroviruses in the transmembrane protein. The ERV3 amino acid sequence inferred from the DNA sequence was aligned with those of RSV, M-MuLV, BaEV, REV-A, and HTLV-I and II. Amino acid homologies with ERV3 beginning at ERV3 nt 2332 are enclosed in boxes.

not be bound to the cell membrane nor serve a normal retroviral function. It should be noted that C terminal to the stop codon, the amino acid sequence in the same reading frame is typical of retroviral transmembrane proteins and contains the long hydrophobic domain followed by a hydrophilic tail.

The ERV3 *env* gene nucleotide and encoded amino acid sequences were also compared to those of the human provirus, clone 4-1 (Repaske *et al.*, 1985). Beginning with ERV3 nt 769, the two proviruses share identities of 42.8% (amino acid) and 47.4% (nucleotide) in their glycoprotein sequences, and 62% (amino acid) and 63.7% (nucleotide) in their transmembrane protein sequences including the open frame C terminal to the stop codon at nt 2502. ALIGN program comparisons (Dayhoff, 1976) resulted in exceedingly low probabilities that the two *env* genes are related by chance. In a previous analysis, we showed that ERV3 is significantly related in the *gag* p30 sequence to another of the 4-1 family of human endogenous retroviruses, clone 51-1 (O'Connell *et al.*, 1984). However, the similarity in restriction maps of the family of human proviruses typified by clone 4-1 (Repaske *et al.*, 1983b, 1985) and their primer binding site homology to tRNA glutamic acid (Steele *et al.*, 1984)

contrasts with ERV3, which is single copy in human DNA, has a substantially different restriction map, and has primer binding site homology to tRNA arginine. This suggests that during evolution, a progenitor of the ERV3 provirus separated from the lineage that gave rise to the clone 4-1 family of proviruses.

**Fetal cDNA clone.** Because of the known expression of endogenous retroviral glycoprotein in fetal and adult tissues of normal mice (Lerner *et al.*, 1976) and the expression of retroviral *gag* antigens in normal human first trimester and term placentas (Suni *et al.*, 1984; Jerabek *et al.*, 1984), we decided to screen a human fetal cDNA library using the ERV3 genomic clone as probe. The  $\lambda$ -gt10 cDNA library (a gift of E. F. Fritsch) was prepared from mRNA isolated from the liver of a 20-week-old human fetus. After high stringency hybridization with a 1726 nt *env* fragment isolated from the ERV3 pHP1.7 genomic subclone (nucleotides 552-2277), two positive plaques were isolated and were found to contain the identical 2.85-kb insert. Cleavage with several restriction enzymes revealed only those fragments expected from the ERV3 genomic clone map (Fig. 1) (not shown).

Several regions in the cDNA clone were sequenced including those at both ends

(Fig. 1). The 2851-nt cDNA insert extends from nt 531 in the genomic clone sequence (Fig. 2) through the adenosine at nt 3355, plus another 26 adenosine residues. This identification of the polyadenylation site revealed that the ERV3 LTR region R is 4 nucleotides longer than previously postulated (O'Connell and Cohen, 1984), or 67 nucleotides, while U5 is only 32 nucleotides. The DNA sequence of the cDNA clone is identical to that of the ERV3 genomic clone in the 1110 nucleotides sequenced. Isolation of an ERV3 envelope-containing cDNA clone, then, represents direct evidence that the ERV3 proviral locus is transcribed in human tissues.

If the 15-nucleotide sequence beginning at nt 656 (Fig. 2) is the correct *env* splice acceptor in ERV3, then the isolated cDNA clone (which begins upstream at nt 531) must be derived from an mRNA molecule that was not spliced at this acceptor. In this regard, we have identified from human first trimester fetuses two ERV3 *env*-containing transcripts, one greater than genome length of approximately 13 kb, and one of 3.7 kb (Pfeifer *et al.*, in preparation). While it is clear that ERV3 encodes a defective human endogenous retrovirus (O'Connell *et al.*, 1984), the ERV3 envelope gene may nevertheless encode a viral protein product. This putative glycoprotein may furnish the human cell with another, as yet unknown, function which may or may not be virally related. To address this question, we have prepared antibody against the ERV3 envelope protein produced in a bacterial expression vector system and found that it is reactive with several normal human tissues (Cohen *et al.*, in preparation).

#### ACKNOWLEDGMENTS

We thank Jeannie Clarke for preparing the manuscript. Research was sponsored by the National Cancer Institute, DHHS, under Contract NO1-CO-23909 with Litton Bionetics, Inc. The contents of this publication do not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U. S. government.

#### REFERENCES

- ADACHI, A., SAKAI, K., KITAMURA, N., NAKANISHI, S., NIWA, O., MATSUYAMA, M., and ISHIMOTO, A. (1984). Characterization of the *env* gene and long terminal repeat of molecularly cloned Friend mink cell focus-inducing virus DNA. *J. Virol.* **50**, 813-821.
- BENTON, W. D., and DAVIS, R. W. (1977). Screening  $\lambda$ gt recombinant clones by hybridization to single plaques *in situ*. *Science (Washington, D. C.)* **196**, 180-182.
- BENVENISTE, R. E., and TODARO, G. J. (1974). Evolution of type C viral genes: I. Nucleic acid from baboon type C virus as a measure of divergence among primate species. *Proc. Natl. Acad. Sci. USA* **71**, 4513-4518.
- BENVENISTE, R. E., and TODARO, G. J. (1976). Evolution of type C viral genes: Evidence for an Asian origin of man. *Nature (London)* **261**, 101-108.
- BLOBEL, G., and DOBBERSTEIN, B. (1975). Transfer of proteins across membranes. *J. Cell Biol.* **67**, 835-851.
- BONNER, T. I., O'CONNELL, C., and COHEN, M. (1982). Cloned endogenous retroviral sequences from human DNA. *Proc. Natl. Acad. Sci. USA* **79**, 4709-4713.
- ROSSELMAN, R. A., VAN STRAATEN, F., VAN BEVEREN, C., VERMA, I. M., and VOGT, M. (1982). Analysis of the *env* gene of a molecularly cloned and biologically active Moloney mink cell focus-forming proviral DNA. *J. Virol.* **44**, 19-31.
- CALLAHAN, R., CHIU, I.-M., WONG, J. F. H., TRONICK, S. R., ROE, B. A., AARONSON, S. A., and SCHLOM, J. (1985). A new class of endogenous human retroviral genomes. *Science (Washington, D. C.)* **228**, 1208-1211.
- CALLAHAN, R., DROHAN, W., TRONICK, S., and SCHLOM, J. (1982). Detection and cloning of human DNA sequences related to the mouse mammary tumor virus genome. *Proc. Natl. Acad. Sci. USA* **79**, 5503-5507.
- CHATTOPADHYAY, S. K., CLOYD, M. W., LINEMEYER, D. L., LANDER, M. R., RANDS, E., and LOWY, D. R. (1982). Cellular origin and role of mink cell focus-forming viruses in murine thymic lymphomas. *Nature (London)* **295**, 25-31.
- CIANCIOLO, G. J., KIPNIS, R. J., and SYNDERMAN, R. (1984). Similarity between p15E of murine and feline leukemia viruses and p21 of HTLV. *Nature (London)* **311**, 515.
- DAYHOFF, M. O. (1976). In "Atlas of Protein Sequence and Structure" (M. O. Dayhoff, ed.), Vol. 5, Suppl. 2, pp. 1-8. National Biomedical Research Foundation, Washington, D. C.
- ELDER, J. H., and MULLINS, J. I. (1983). Nucleotide sequence of the envelope gene of Gardner-Arnstein Feline Leukemia Virus B reveals unique sequence homologies with a murine mink cell focus-forming virus. *J. Virol.* **46**, 871-880.
- FICHT, T. A., CHANG, L.-J., and STOLTZFUS, C. M. (1984). Avian sarcoma virus *gag* and *env* gene structural



- protein precursors contain a common amino-terminal sequenced. *Proc. Natl. Acad. Sci. USA* **81**, 362-366.
- HACKETT, P. B., SWANSTROM, R., VARMUS, H. E., and BISHOP, J. M. (1982). The leader sequence of the subgenomic mRNAs of Rouse sarcoma virus is approximately 390 nucleotides. *J. Virol.* **41**, 527-534.
- JERABEK, L. B., MELLORS, R. C., ELKON, K. B., and MELLORS, J. W. (1984). Detection and immunochemical characterization of a primate type C retrovirus-related p30 protein in normal human placentas. *Proc. Natl. Acad. Sci. USA* **81**, 6501-6505.
- KOMINAMI, R., TOMITA, Y., CONNORS, E. C., and HATANAKA, M. (1980). Conserved sequence related to the 3'-terminal region of retrovirus RNAs in normal cellular DNAs. *J. Virol.* **34**, 684-692.
- KOZAK, M. (1984). Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. *Nucleic Acids Res.* **12**, 857-872.
- KYTE, J., and DOOLITTLE, R. F. (1982). A simple method for displaying the hydrophobic character of a protein. *J. Mol. Biol.* **157**, 105-132.
- LENZ, J., CROWTHER, R., STRACESKI, A., and HASELTINE, W. (1982). Nucleotide sequence of the Akv *env* gene. *J. Virol.* **42**, 519-529.
- LENER, R. A., WILSON, C. B., DEL VILLANO, B. C., MCCONAHEY, P. J., and DIXON, F. J. (1976). Endogenous oncornaviral gene expression in adult and fetal mice: Quantitative, histologic, and physiologic studies of the major viral glycoprotein, gp70. *J. Exp. Med.* **143**, 151-166.
- MARSHALL, R. D. (1974). The nature of metabolism of carbohydrate-peptide linkage of glycoproteins. *Biochem. Soc. Symp.* **40**, 17-26.
- MARTIN, M., BRYAN, T., RASHEED, S., and KHAN, A. S. (1981). Identification and cloning of endogenous retroviral sequences present in human DNA. *Proc. Natl. Acad. Sci. USA* **78**, 4892-4896.
- MAXAM, A., and GILBERT, W. (1980). Sequencing end-labeled DNA with base-specific chemical cleavages. In "Methods in Enzymology" (L. Grossman and K. Moldave, eds.), Vol. 65, pp. 499-560. Academic Press, New York.
- MOUNT, S. M. (1982). A catalogue of splice junction sequences. *Nucleic Acids Res.* **10**, 459-472.
- MUESING, M. A., SMITH, D. H., CARRADILLA, C. D., BENTON, C. V., LASKY, L. A., and CAPON, D. J. (1985). Nucleic acid structure and expression of the human AIDS/lymphadenopathy retrovirus. *Nature (London)* **313**, 450-458.
- NODA, M., KURIHARA, M., and TAKANO, T. (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.
- O'BRIEN, S. J., BONNER, T. I., COHEN, M., O'CONNELL, C., and NASH, W. G. (1983). Mapping of an endogenous retroviral sequence to human chromosome 18. *Nature (London)* **303**, 74-77.
- O'CONNELL, C. D., and COHEN, M. (1984). The LTR sequences of a novel human endogenous retrovirus. *Science (Washington, D. C.)* **226**, 1204-1206.
- O'CONNELL, C. D., O'BRIEN, S. J., NASH, W. G., and COHEN, M. (1984). ERV3, a full-length human endogenous provirus: Chromosomal localization and evolutionary relationships. *Virology* **138**, 225-235.
- O'NEILL, R. R., BUCKLER, C. E., THEODORE, T. S., MARTIN, M. A., and REPASKE, R. (1985). Envelope and long terminal repeat sequences of a cloned infectious NZB xenotropic murine leukemia virus. *J. Virol.* **53**, 100-106.
- RATNER, L., HASELTINE, W., PATARCA, R., LIVAK, K. J., STARCICH, B., JOSEPHS, S. F., DORAN, E. R., RAFALSKI, J. A., WHITEHORN, E. A., BAUMEISTER, K., IVANOFF, L., PETTEWAY, JR., S. R., PEARSON, M. L., LAUTENBERGER, L. A., PAPAS, T. S., GHRAIEB, J., CHANG, N. T., GALLO, R. C., and WONG-STAAAL, F. (1985). Complete nucleotide sequence of the AIDS virus, HTLV-III. *Nature (London)* **313**, 277-284.
- REDMOND, S. M. S., and DICKSON, C. (1983). Sequence and expression of the mouse mammary tumor virus *env* gene. *EMBO J.* **2**, 125-131.
- REPASKE, R., O'NEILL, R. R., KHAN, A. S., and MARTIN, M. A. (1983a). Nucleotide sequence of the *env*-specific segment of NFS-Th-1 xenotropic murine leukemia virus. *J. Virol.* **46**, 204-211.
- REPASKE, R., O'NEILL, R. R., STEELE, P. E., and MARTIN, M. A. (1983b). Characterization and partial nucleotide sequence of endogenous type C retrovirus segments in human chromosomal DNA. *Proc. Natl. Acad. Sci. USA* **80**, 678-682.
- REPASKE, R., STEELE, P. E., O'NEILL, R. R., RABSON, A. B., and MARTIN, M. A. (1985). Nucleotide sequence of a full-length human endogenous retroviral segment. *J. Virol.* **54**, 764-772.
- RICE, N. R., STEPHENS, R. M., COUEZ, D., DESCHAMP, J., KETTMANN, R., BURNY, A., and GILDEN, R. V. (1984). The nucleotide sequence of the *env* gene and post-*env* region of bovine leukemia virus. *Virology* **138**, 82-93.
- SANCHEZ-PESCADOR, R., POWER, M. D., BARR, P. J., SEIMER, K. S., STEMPIEN, M. M., BROWN-SHIMER, S. L., GEE, W. W., RENARD, A., RANDOLPH, A., LEVY, J. A., DINA, D., and LUCIW, P. A. (1985). Nucleotide sequence and expression of an AIDS-associated retrovirus (ARV-2). *Science (Washington, D. C.)* **227**, 484-492.
- SCHWARTZ, D. E., TIZARD, R., and GILBERT, W. (1983). Nucleotide sequence of Rous Sarcoma virus. *Cell* **32**, 853-869.
- SEGREST, J. P., and FELDMANN, R. J. (1974). Membrane proteins: Amino acid sequence and membrane penetration. *J. Mol. Biol.* **87**, 853-858.

- SHINNICK, T. M., LERNER, R. A., and SUTCLIFF, J. G. (1981). Nucleotide sequence of Moloney murine leukemia virus. *Nature (London)* **293**, 543-548.
- SODROSKI, J., PATARCA, R., PERKINS, D., BRIGGS, D., LEE, T.-H., ESSEX, M., COLIGAN, J., WONG-STAAAL, F., GALLO, R. C., and HASELTINE, W. A. (1984). Sequence of the envelope glycoprotein gene of type II human T lymphotropic virus. *Science (Washington, D. C.)* **225**, 421-424.
- STEELE, P. E., RABSON, A. B., BRYAN, T., and MARTIN, M. A. (1984). Distinctive termini characterize two families of human endogenous retroviral sequences. *Science (Washington, D. C.)* **225**, 943-947.
- SUNI, J., NARVANEN, A., WAHLSTROM, T., AHO, M., PAKKANEN, R., VAHERI, A., COPELAND, T., COHEN, M., and OROSZLAN, S. (1984). Human placental syncytiotrophoblastic Mr 75000 polypeptide defined by antibodies to a synthetic peptide based on a cloned human endogenous retroviral DNA sequence. *Proc. Natl. Acad. Sci. USA* **81**, 6197-6201.
- VIEIRA, J., and MESSING, J. (1982). The pUC plasmids, an M13mp7-derived system for insertion mutagenesis and sequencing with synthetic universal primers. *Gene* **19**, 259-268.
- WAIN-HOBSON, S., SONIGO, P., DANOS, O., COLE, S., and ALIZON, M. (1985). Nucleotide sequence of the AIDS virus, LAV. *Cell* **40**, 9-17.
- WESTLEY, B., and MAY, F. E. B. (1984). The human genome contains multiple sequences of varying homology to mouse mammary tumour virus DNA. *Gene* **28**, 221-227.
- WILHELMSEN, K. C., EGGLETON, K., and TEMIN, H. (1984). Nucleic acid sequences of the oncogene *v-rel* in reticuloendotheliosis virus strain T and its cellular homolog, the proto-oncogene *c-rel*. *J. Virol.* **52**, 172-182.