Human Retroviruses, Cancer, and AIDS:
Approaches to Prevention and Therapy, pages 75–88
© 1988 Alan R. Liss, Inc.

CHARACTERIZATION OF THE ENV GENE AND OF TWO NOVEL CODING REGIONS OF THE HUMAN SPUMARETROVIRUS

B. Maurer, H. Bannert, A. Rethwilm,
G. Darai*, and R.M. Flügel

Institute of Virus Research, German Cancer Research Center,
P.O. Box 101949, 6900 Heidelberg, FRG, and
*Institute of Medical Virology, University of Heidelberg,
Im Neuenheimer Feld 324, 6900 Heidelberg, FRG

ABSTRACT Recombinant clones harboring retroviral DNA were established. The nucleotide sequence of the central and 3' region of the genome of the human spumaretrovirus was determined. The 5' end of the deduced protein sequence was homologous to the endonuclease domain of retroviral reverse transcriptases. A small intergenic region is followed by a long open reading frame of 985 amino acid residues that according to its genomic location and structural features is a typical retroviral env gene. Surprisingly, the post-env region contains two open reading frames that encodes two novel retroviral genes, termed bel-1 and bel-2. The 3' LTR is 963 nucleotides long and contains the signal sequences characteristic for transcriptional regulation of retrovirus genomes.

INTRODUCTION

Spumaretroviruses, also called foamy viruses which constitute the third subfamily of the retrovirus family have been isolated from a number of mammalian species, including primates (1).

[1] This work was supported by the Deutsche Forschunggemeinschaft (Re 627/7).

They seem to induce persistent infections in their natural hosts without a recognizable pathological disease. They induce, however, charcteristic cytopathic effects that lead to the degeneration of the infected cells. In contrast to the onco-, and lentiviruses, which have been characterized in detail, spumaretroviruses have been neglected. There are a number of different reasons why research on this subfamily has been neglected. The reasons are the poor growth properties of the virus in cell culture and in vivo and the apparent absence of pathological disease. In addition, there have been attempts to molecularly clone the virus by different groups without much success. It turns out that the genome of the foamy virus has structural features that makes the molecular cloning difficult when standard cloning procedures are used, e.g. parts of the long terminal repeat sequences (U3) have been found to be relatively unstable during this establishment in bacterial plasmid vectors.

Furthermore, while molecular cloning the foamy virus genome, we happened to isolate a clone that contained part of a human retrotransposon (2). Another unusual feature of the foamy viral genome is its large size. To date, the genome of the human spumaretrovirus (HSRV) that was isolated in 1971 by Achong and Epstein (3) from a nasopharynx carcinoma patient and that we work with exceeds even the lentiviral genomes in size. In order to study the biological role of foamy viruses and to ascertain whether or not it really has no pathological role in vivo, we decided to characterize the viral genome of HSRV in more detail.

RESULTS

The stragey used for the molecular cloning of HSRV DNA relies on two independent sources of viral DNA. As a source for proviral and viral DNAs, HSRV-infected human embryonic lung (HEL) cells were used to prepare DNA fragments that hybridized to ^{32}P-labeled DNA complimentary to HSRV DNA (cDNA). As a second source, cDNA was separately synthesized from purified HSRV virions in larger amounts by using reverse transcriptase. Subsequently, and separately, recombinant lambda clones were constructed that harbored inserts of either viral or cDNA (4). Those phage clones that hybridized to separately synthesized cDNA were subcloned into bacterial plasmid vectors, e.g. pAT153 and pSP65. In this way, a number of recombinant plasmid clones were established two of which (C55 and B52) were characterized in more detail and will be described here. The molecular cloning of the HSRV genome is the subject of a manuscript in preparation.

1. Nucleotide Sequence Analysis

Recombinant clone C55 was characterized by restriction mapping. It contained a 5.4 kbp insert of HSRV DNA that had originally been cloned from cDNA, and subcloned into the Hind III sites of the bacterial plasmid pAT153. The detailed analysis revealed that C55 DNA overlaps part of B52 DNA, a recombinant clone that was established from viral DNA. Fig. 1 shows the restriction mapping of both recombinant clones that resulted in a unique and unambiguous physical map. The determination of the nucleotide sequence of C55 DNA allowed us to correlate its physical map to its genetic map by analogy to other retroviral genomes (Fig. 1).

In the following, the DNA sequence shown in fig. 2 will be described starting from the 5' end.

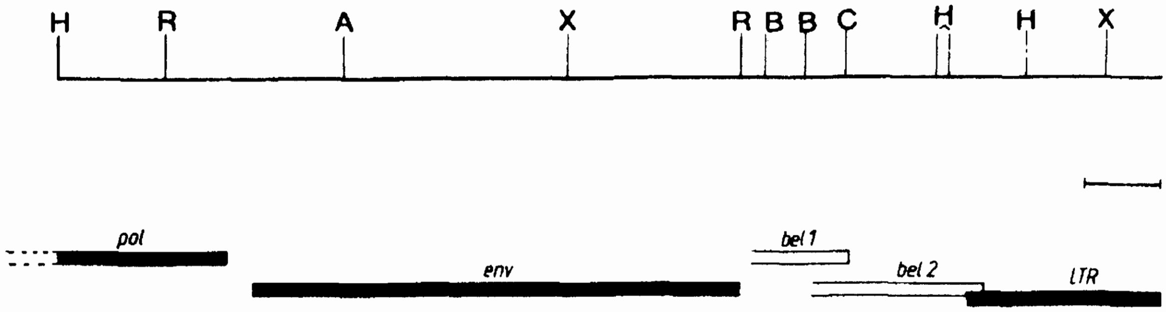

FIGURE 1. Restriction mapping of some enzymes (upper part) and proposed genetic map of the central and 3' region of the HSRV genome (lower part). H = Hind III, R = Eco RI, A = Hpa I, X = Xba I, B = Bam HI, and C = Cla I. The scale marks 500 bp.

a) The 3' pol and pol/env intergenic region. The nucleotide sequence shown in fig. 2 starts with the Hind III site of C55 DNA. The first 1032 nucleotides are homologous to the endonuclease (integrase) domain of the pol genes of most retroviruses (4). The deduced amino acid sequences show a relatively high degree of homology of 21-31% to the corresponding protein sequences of the retroviral reverse transcriptases (data not shown) (4).

```
              10         20         30         40         50         60         70         80         90        100
   1 AAGCTTGCCA CCCAAGGAAG TTATGTGGTT AATTGTAATA CCAAAAAACC AAACCTGGAT GCAGAGTTGG ATCAATTATT ACAGGGTCAT TATATAAAAG
 101 GATATCCCAA ACAATATACA TATTTTTTAG AAGATGGCAA AGTAAAAGTT TCCAGACCTG AAGGGGTTAA AATTATTCCC CCTCAGTCAG ACAGACAAAA
 201 AATTGTGCTT CAAGCCCACA ATTTGGCTCA CACCGGACGT GAAGCCACTC TTTTAAAAAT TGCCAACCTT TATTGGTGGC CAAATATGAG AAAGGATGTG
 301 GTTAAACAAC TAGGACGCTG TCAACAGTGT TTAATCACAA ATGCTTCCAA CAAAGCCTCT GGTCCTATTC TAAGACCAGA TAGGCCTCAA AAACCTTTTG
 401 ATAAATTCTT TATTGACTAT ATTGGACCTT TGCCACCTTC ACAGGGATAC CTATATGTAT TAGTAGTTGT TGATGGAATG ACAGGATTCA CTTGGTTATA pol
 501 CCCCACTAAG GCTCCTTCTA CTAGCGCAAC TGTTAAATCT CTCAATGTAC TCACTAGTAT TGCAATTCCA AAGGTGATTC ACTCTGATCA AGGTGCAGCA
 601 TTCACTTCTT CAACCTTTGC TGAATGGGCA AAGGAAAGAG GTATACATTT GGAATTCAGT ACTCCTTATC ACCCCCAAAG TGGTAGTAAG GTGGAAAGGA
 701 AAAATAGTGA TATAAAACGA CTTTTAACTA AACTGCTAGT AGGAAGACCC ACAAAGTGGT ATGACCTATT GCCTGTTGTA CAACTTGCTT TAAACAACAC
 801 CTATAGCCCT GTATTAAAAT ATACTCCACA TCAACTCTTA TTTGGTATAG ATTCAAATAC TCCATTTGCA AATCAAGATA CACTTGACTT GACCAGAGAA
 901 GAAGAACTTT CTCTTTTACA GGAAATTCGT ACTTCTTTAT ACCATCCATC CACCCCTCCA GCCTCCTCTC GTTCCTGGTC TCCTGTTGTT GGCCAATTGG
1001 TCAGGAGAGG GTGGCTAGGC CTGCTTCTTT GAGACCTCGT TGGCATAAAC CGTCTACTGT ACTTAAGGTG TTGAATCCAA GGACTGTTGT TATTTTGGAC
1101 CATCTTGGCA ACAACAGAAC TGTAAGTATA GATAATTTAA AACCTACTTC TCATCAGAAT GGCACCACCA ATGACACTGC AACAATGGAT CATTTGGAAA
1201 AAAATGAATA AAGCGCATGA GGCACTTCAA AATACAACAA CTGTGACTGA ACAGCAGAAG GAACAAATTA TACTGGACAT TCAAAATGAA GAAGTACAAC
1301 CAACTAGGAG AGATAAATTT AGATATCTGC TTTATACTTG TTGTGCTACT AGCTCAAGAG TATTGGCCTG GATGTTTTTA GTTTGTATAT TGTTAATCAT
1401 TGTTTTGGTT TCATGCTTTG TGACTATATC CAGAATACAA TGGAATAAGG ATATTCAGGT ATTAGGACCT GTAATAGACT GGAATGTTAC TCAAAGAGCT
1501 GTTTATCAAC CCTTACAGAC TAGAAGGATT GCACGTTCCC TTAGAATGCA GCATCCTGTT CCAAAATATG TGGAGGTAAA TATGACTAGT ATTCCACAAG
1601 GTGTATACTA TGAACCCCAT CCGGAACCCA TAGTGGTGAA GGAGAGGGTC CTAGGTCTTT CTCAAATTCT GATGATTAAT TCAGAAAACA TTGCTAATAA
1701 TGCTAATTTG ACACAAGAAG TAAAGAAGTT GTTAACTGAA ATGGTTAATG AAGAAATGCA AAGTTTGTCA GATGTAATGA TTGACTTTGA AATTCCTTTA
1801 GGAGACCCTC GTGATCAAGA ACAATATATA CATAGAAAAT GCTATCAAGA ATTTGCAAAT TGTTATTTAG TAAAATATAA AGAACCCAAA CCGTGGCCTA
1901 AGGAGGGCCT TATAGCTGAT CAATGCCCAT TACCAGGTTA CCATGCTGGA TTAACCTATA ATAGACAGTC TATTTGGGAT TACTATATTA AAGTGGAGAG
2001 TATTAGACCT GCAAATTGGA CAACAAAGAG TAAATATGGA CAAGCTAGAC TAGGAAGTTT TTATATTCCT AGCAGCCTGA GACAAATCAA TGTTAGTCAT
2101 GTACTATTCT GTAGTGATCA ATTATATTCT AAATGGTATA ATATAGAAAA TACCATAGAA CAAAACGAGC GGTTTCTGCT TAATAAACTA AATAACCTTA
2201 CATCTGGAAC CTCAGTATTG AAGAAAAGAG CTCTTCCGAA GGATTGGAGT TCTCAAGGTA AAAATGCTCT GTTTAGAGAA ATCAATGTGT TAGATATCTG
2301 CAGTAAACCT GAATCTGTAA TACTATTGAA TACTTCATAC TATTCCTTCT CTTTATGGGA AGGAGATTGT AATTTTACTA AAGATATGAT TTCTCAGTTG
2401 GTTCCAGAAT GTGATGGATT TTATAACAAT TCTAAGTGGA TGCATATGCA TCCATATGCT TGTAGATTCT GGAGAAGTAA GAAGAATGAA AAAGAAGAAA
2501 CTAAATGTAG AGATGGGGAA ACTAAGAGAT GTCTGTATTA TCCTTTATGG GACAGTCCCG AATCTACATA TGATTTTGGT TATTTAGCAT ACCAAAAGAA
2601 TTTTCCTTCC CCTATCTGTA TAGAACAACA GAAAATTAGA GATCAAGATT ATGAAGTCTA TTCTTTGTAT CAAGAACGCA AAATAGCTTC TAAAGCATAT env
2701 GGAATTGATA CAGTTTTATT CTCTCTAAAG AATTTTCTTA ATTATACAGG AACTCCTGTA AATGAAATGC CTAATGCAAG AGCTTTTGTA GGCCTAATAG
2801 ATCCCAAGTT TCCTCCTTCC TATCCCAATG TTACTAGGGA ACATTATACT TCCTGTAATA ATAGGAAAAG AAGAAGTGTT GATAATAACT ATGCTAAGTT
2901 AAGGTCTATG GGGTATGCAC TTACAGGAGC AGTGCAAACC TTATCTCAAA TATCAGATAT TAATGATGAA AACTTACAGC AAGGAATATA TTTATTAAGG
3001 GATCATGTAA TAACCTTAAT GGAAGCTACA TTGCATGATA TATCTGTTAT GGAAGGAATG TTTGCTGTAC AACATTTGCA TACACATTTG AATCATTTGA
3101 AGACAATGCT TCTAGAAAGA AGAATAGACT GGACCTATAT GTCTAGTACT TGGCTACAAC AACAATTACA GAAATCTGAT GATGAGATGA AAGTAATAAA
3201 GAGAATTGCT AGAAGTTTGG TATATTATGT TAAACAAACC CATAGTTCTC CCACAGCTAC AGCCTGGGAG ATTGGATTAT ATTATGAATT GGTTATACCT
3301 AAACATATTT ACTTGAATAA TTGGAATGTT GTCAATATAG GTCACTTAGT TAAATCAGCT GGACAATTGA CTCATGTAAC TATAGCTCAT CCTTATGAAA
3401 TAATCAATAA GGAATGTGTA GAGACTATAT ATCTGCATCT TGAGGACTGC ACAAGACAAG ATTATGTCAT ATGTGATGTG GTAAAGATAG TGCAGCCTTG
3501 TGGCAATAGC TCAGACACGA GTGATTGTCC TGTCTGGGCT GAAGCTGTAA AAGAACCATT TGTGCAAGTC AATCCTCTGA AAAACGGAAG TTATCTGGTT
3601 TTGGCAAGTT CCACAGACTG TCAGATCCCA CCATATGTTC CTAGCATCGT GACTGTTAAT GAAACAACGT CATGCTTTGG ACTGGACTTT AAAAGGCCAC
3701 TGGTTGCGGA AGAAAGATTG AGCTTTGAGC CACGACTGCC AAATCTACAA CTAAGATTAC CACATTTGGT TGGAATTATT GCAAAAATCA AAGGGATAAA
3801 AATAGAAGTC ACATCCTCTG GAGAAAGTAT AAAAGAGCAG ATTGAAAGAG CAAAAGCTGA GCTCCTTCGA CTGGACATTC ACGAGGGAGA TACTCCTGCC
3901 TGGATACAAC AGCTAGCTGC AGCAACAAAG GACGTCTGGC CAGCAGCAGC TTCTGCTCTA CAAGGAATTG GTAACTTTTT ATCTGGGACT GCCCAAGGAA
4001 TATTTGGAAC TGCCTTTAGT CTCTTGGGAT ACTTAAAGCC TATCCTAATA GGAGTAGGGG TCATTCTCTT GGTTATTCTT ATATTTAAGA TTGTATCATG
4101 GATTCCTACG AAAAAGAAGA ATCAGTAGCC TCCACCTCTG GAATTCAGGA CCTGCAGACT CTGAGTGAGC TTGTTGGCCC TGAAAATGCC GGAGAGGGAG
4201 AGCTGACTAT TGCTGAGGAA CCTGAAAGAA AATCCTCGAC GCCCCAGACG ATATACCAAA AGAGAAGTCA AATGTGTGTC TTATCATGCA TATAAAGAAA
4301 TTGAGGACAA ACATCCTCAA CATATTAAAC TGCAGGATTG GATCCCCACA CCAGAGGAAA TGAGTAAGTC ACTCTGTAAA AGACTTATTT TATGTGGATT
4401 GTATAGTGCA GAAAAGGCCT CAGAGATTTT AAGGATGCCT TTTACAGTTT CTTGGGAACA ATCAGATACT GACCCTGACT GTTTTATTGT AAGCTATACA
4501 TGTATATTTT GTGATGCTGT AATACATGAT CCCATGCCCA TAAGATGGGA TCCTGAAGTT GGAATTTGGG TAAAATATAA ACCCCTCAGA GGAATTGTTG bel-1
4601 GATCTGCTGT GTTTATTATG CATAAACATC AAAGAAACTG TTCTCTTGTT AAACCTTCTA CCAGTTGCTC AGAAGGTCCA AAACCAAGAC CTAGGCACGA
4701 TCCTGTCCTT CGCTGTGACA TGTTTGAAAA GCATCACAAG CCTCGGCAGA AACGACCCAG GAGACGATCC ATCGATAATG AGTCATGTGC TTCCAGTAGT
4801 GACACCATGG TCAATGAGCC ATGATCACTA TGCACCAACC CTCTTTGGAA TCCTTGACCG CTACTATCAG GGCTACTTGA AGAGTCCAGC AACCTACCAA
4901 ACTTGGAAGT TCACATGTCA GGTGGACCCT TCTGGGAAGA GGTTTATGGA GACTCAATTT TGGGTCCCCC CTCTGGGTCA GGTGAACATT CAGTTTTATA
5001 AGAATTATCA GATTCTAACT TGCTGTCAGG CTGTAGACCC ATTTGCTAAT ATTTTTCATG GTACTGATGA AGAAATGTTT GACATTGATT CAGGTCCTGA
5101 TGTTTGGTGT TCTCCCTCTT TGTGTTTCAA GGTAATTTAT GAAGGGGCAA TGGGCCAAAA GCAAGAACAA AAAACATGGC TGTGCAGACT AGGACATGGT bel-2
5201 CATCGTATGG GAGCATGCGA TTACCGTAAA GTAGATCTGT ATGCAATGAG ACAAGGAAAA GAAAACCCTT ATGGAGATAG GGGTGATGCA GCTTTGCAAT
5301 ATGCTTATCA GGTTAAAAGG GGCTGTAAAG CAGGGTGCTT GGCATCACCT GTACTTAACT ACAAAGCTTT GCAGTTTCAT AGAACCATTA TGGCAGACTT
```

Fig. 2

Fig. 2 (continued)

```
5401 CACCAATCCT AGGATTGGAG AAGGACATCT TGCTCATGGT TACCAAGCAG CTATGGAAGC TTATGGACCT CAGAGAGGAA GTAACGAGGA GAGGGTGTGG
5501 TGGAATGTCA CTAGAAACCA GGGAAAACAA GGAGGAGAGT ATTACAGGGA AGGAGGTGAA GAACCTCATT ACCCAAATAC TCCTGCTCCT CATAGACGTA
5601 CCTGGGATGA GAGACACAAG GTTCTTAAAT TGTCCTCATT CGCTACTCCC TCTGACATCC AACGCTGAAC TACTAAAGCA TTGCCTTATG GCTGGAAAGT
5701 GGTCACCGAA AGCGGAAATG ATTATACTAG CCGCAGAAAG ATCAGAACAT TGACAGAGAT GACTCAGGAT GAAATTAGAA AAAGGTGGGA AAGTGGATAT
5801 TGTGACCCCT TCATTGACTC AGGAAGTGAC TCAGATGGAC CCTTCTAAAA GCCACAGACA GTAAAAATGT GTTAGCACTT TATACAATAT TATATCTGCT
5901 TAAGCTATAG AAGCTTTCAC ATACTCAGTA\GCTGTTTCAC AATCAACAAA ACAATGATGA TGTAATCATA AGGAAGTAGT TTAAAATAGG TTAAGTAAGT
6001 TTACTGCAGT AGATAATCCC TGGGGAGGAT CTGGCTCTGT AAGCTGGAAC AGCAATGTTT TCAGTTCCAA TCCTCTCAAA GGAGAACCAG AGGGATGATG
6101 TGTTAGTTCA AATCCCATTA TCCTCATGGT TCCCTTTTCC ATAGTTTACT ATATTAATTT AAGGATAAGG TATAAGGATT AAGGTATGAG GTGTGTGGCT
6201 CAACACGTAG GGTGACAAGA AAATCTACTG TAATAGGACA CAACACCTCT AAAGTTGCCC GTGGGAAGGT GAAGTGAGAT CGAATCTTTC CTTAACGCAG  LTR
6301 ACAGCTTTTT ATCCACTAGG GATAATGTTT TAAGGAATAC TATAGTAATA GATTGATAGT TTTAACAATG ATAGAAATAG TATATAAGGA TAGTTTCTAG
6401 ATTGTACGGG AGGCTCTTCA CTACTCGCTG CGTCGAGAGT GTACGAGACT CTCCAGGTTT GGTAAGAAAT ATTTTATATT GTTATAATGT TACTATGATC
6501 CATTAACACT CTGCTTATAG ATTGTAAGGG TGATTGCAAT GCTTTCTGCA TAAAACTTTG GTTTTCTTGT TAATCAATAA ACCGACTTGA TTCGAGAACC
6601 AACTCCTATA TTATTGTCTC TTTTATACTT TATTAAGTAA AAGGATTTGT ATATTAGCCT TGCTAAGGGA GACATCTAGT GATATAAGTG TGAACTACAC
6701 TTATCTTAAA TGATGTAACT CCTTAGGATA ATCAATATAC AAAATTCCAT GACAA
```

FIGURE 2. Nucleotide sequence of the HSRV DNA. The sequence starts with the Hind III site at position 1 as the 5' terminus and extends to the end of the 3' LTR. Gene locations are given at the right margin. Stop codons are boxed, the putative env initiator is marked by horizontal arrow; potential splice acceptor sites for the bel genes are indicated by vertical arrowheads. The polypurine tract upstream of the 3' LTR is doubly underlined. The TATA-box and the polyA-addition signal is marked by a broken line. Cleavage sites for the restriction enzymes shown in fig. 1 are underlined.

The 3' pol region is followed by a relatively short intergenic sequence of 127 nucleotides. It is at this location that a long open reading frame (orf) coding for 995 amino acid residues is identified. This long orf lies in a frame different from that of the pol gene product and will be described below in detail. Overlapping the pol gene and the long orf in the third reading frame a short orf (S1) of only 87 amino acids was found. Since it is so short and not homologous to any of the other lentiviral genes that have been described to occur at a comparable genetic location including the S1 gene of equine infectious anemia virus (EIV) (5) we consider it unlikely that S1 of HSRV is expressed.

The env gene. There are three initiation codons in the first 70 nucleotides of the major long open reading frame that opens at position 1141 (Fig. 2). Only the second ATG and its neighboring sequences conform to those of a typical eukaryotic initiator (6). This initiator is located at position 1171, eleven tripletts inside the major orf that according to its size and to certain peculiar but typical structural features is the HSRV env precursor protein.

It extends for 985 codons to position 4126 where it stops (fig. 3). Its hydrophobicity profile was calculated according to Kyte and Doolittle (7) and is shown in fig. 4. There are two long hydrophobic domains and at least one that is less hydrophobic that deserve to be described. The first hydrophobic region corresponds to the signal peptide, since it is located close to the NH_2-terminus of the env precursor. It is of interest that the first 62 amino acids residues after the initiator at 1171 are clearly hydrophilic in character as is the case for the env precursors of visna virus (VIV) and mouse mammary tumour virus (MMTV) which are of similar lengths and hydrophilicity (8-11). The proper signal peptide comprises residues 74 to 87 (see fig. 3). There is a strongly basic

```
               10         20         30         40         50         60
  1 NLLLIRMAPP MTLQQWIIWK KMNKAHEALQ NTTTVTEQQK EQIILDIQNE EVQPTRRDKF
 61 RYLLYTCCAT SSRVLAWMFL VCILLIIVLV SCFVTISRIQ WNKDIQVLGP VIDWNVTQRA
121 VYQPLQTRRI ARSLRMQHPV PKYVEVNMTS IPQGVYYEPH PEPIVVKERV LGLSQILMIN
181 SENIANNANL TQEVKKLLTE MVNEEMQSLS DVMIDFEIPL GDPRDQEQYI HRKCYQEFAN
241 CYLVKYKEPK PWPKEGLIAD QCPLPGYHAG LTYNRQSIWD YYIKVESIRP ANWTTKSKYG
301 QARLGSFYIP SSLRQINVSH VLFCSDQLYS KWYNIENTIE QNERFLLNKL NNLTSGTSVL  OMP
361 KKRALPKDWS SQGKNALFRE INVLDICSKP ESVILLNTSY YSFSLWEGDC NFTKDMISQL
421 VPECDGFYNN SKWMHMHPYA CRFWRSKKNE KEETKCRDGE TKRCLYYPLW DSPESTYDFG
481 YLAYQKNFPS PICIEQQKIR DQDYEVYSLY QERKIASKAY GIDTVLFSLK NFLNYTGTPV
541 NEMPNARAFV GLIDPKFPPS YPNVTREHYT SCNNRKRRSV DNNYAKLRSM GYALTGAVQT
601 LSQISDINDE NLQQGIYLLR DHVITLMEAT LHDISVMEGM FAVQHLHTHL NHLKTMLLER
661 RIDWTYMSST WLQQQLQKSD DEMKVIKRIA RSLVYYVKQT HSSPTATAWE IGLYYELVIP
721 KHIYLNNWNV VNIGHLVKSA GQLTHVTIAH PYEIINKECV ETIYLHLEDC TRQDYVICDV  TMP
781 VKIVQPCGNS SDTSDCPVWA EAVKEPFVQV NPLKNGSYLV LASSTDCQIP PYVPSIVTVN
841 ETTSCFGLDF KRPLVAEERL SFEPRLPNLQ LRLPHLVGII AKIKGIKIEV TSSGESIKEQ
901 IERAKAELLR LDIHEGDTPA WIQQLAAATK DVWPAAASAL QGIGNFLSGT AQGIFGTAFS
961 LLGYLKPILI GVGVILLVIL IFKIVSWIPT KKKNQ*
```

FIGURE 3. Deduced amino acid sequence of the HSRV env gene. The methionine residue that initiates the env precursor is marked by a flag. Hydrophobic regions are underlined, sites for N-linked glycosylations are boxed. The proteolytic cleavage sites for generating the OMP and TMP is marked by a vertical arrow.

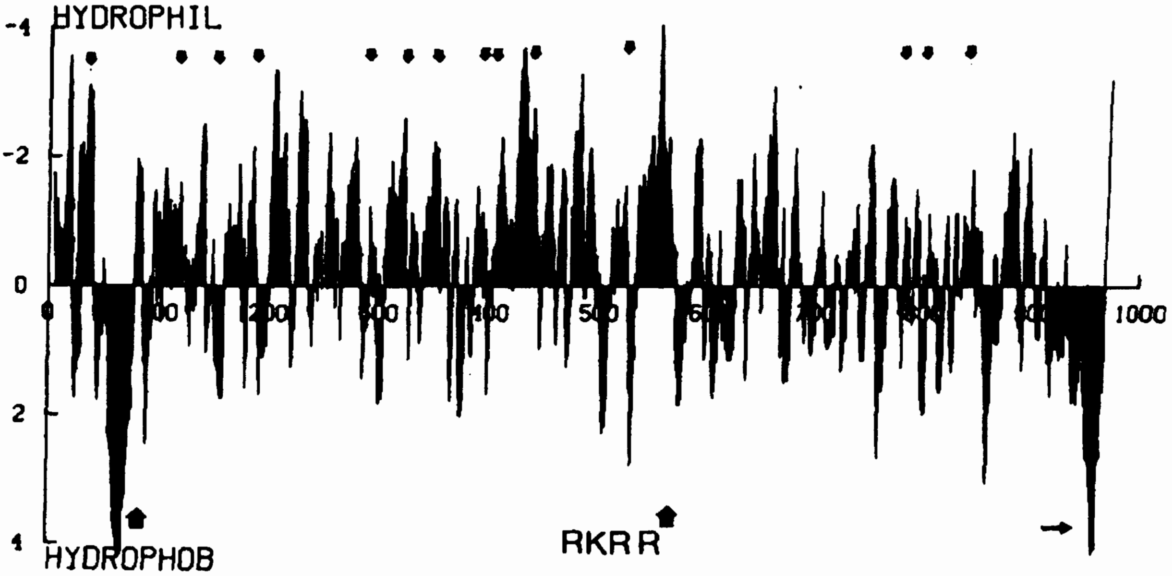

FIGURE 4. Hydropathic profile of the HSRV env precursor polypeptide. The hydrophilic (upper) and hydrophobic (lower) profile was calculated using a 5-amino acid-window size (7). Potential N-linked glycosylation sites are marked with arrows in the upper part. The two thick arrows in the lower part mark the potential processing sites of the env precursor into the mature OMP and TMP. The horizontal arrow indicates the transmembrane peptide.

sequence Arg-Lys-Arg-Arg (fig. 3 and 4) that represents the site for proteolytic processing of the HSRV env precursor into the outer membrane protein (OMP) and into the transmembrane protein (TMP). TMP is located at the carboxy terminus of the env polypeptide (12).

This proteolytic cleavage site is followed by a second less hydrophobic region that corresponds to the so-called fusion peptide region of other retroviral TMPs (13). The third, carboxyterminal, hydrophobic domain of the HSRV env protein is 37 residues long and shows a remarkable resemblance to the corresponding transmembrane domains of HIV-1 and VIV (14,8).

The molecular weight of the OMP of HSRV is predicted to be 53.7 kd, whereas that of a related simian foamy virus type 1 was found to be 70 kd (15). However, this comparison does not take into account the varying degree of glycosylation of these env proteins.

There are 14 potential glycoslyation signals of the type Asn-X-Ser/Thr in the env precursor protein sequence of HSRV (fig. 4). All but three are found in the OMP part of the env protein. The number and distribution of these signal sites resemble the pattern found in other retroviral env protein sequences, and particularly the lentiviruses, HIV, VIV, and EIV. As anticipated, there is little, if any homology at the level of amino acids, but it is worth mentioning that on the nucleotide level the degree of homology of the HSRV env gene to the VIV env gene is 52.1%.

The bel-1 and bel-2 sequences. The post-env region of HSRV is unusually long. It encompasses 1663 nucleotides (fig. 2), ranging from the end of the env gene to the start of the 3' LTR. Two open reading frames are identified in this region that are 205 (bel-1) and 364 (bel-2) amino acid residues long (fig. 5 and 6). It is interesting that both putative bel genes could be generated from spliced viral mRNAs or alternatively could start from initiators, resulting in slightly smaller versions of the predicted polypeptides. The presumed splice acceptor sites conform to those that are found in the HTLV-1 genome for mRNAs from which the px or tat proteins are derived and in which the corresponding initiator codon is located on a RNA leader sequence that is derived from the 5' end of the genome (16,17).

The deduced protein sequences are interesting for a number of different reasons. With respect to their genomic location they could correspond to either the x genes of HTLV-I and II or the 3'-orf of HIV. Alternatively, they can be compared to v-oncogenes of oncoviruses. It is noteworthy that bel-1 and bel-2 do not show any significant homology to any protein sequences of the NBFR protein data bank. On the other hand, they display some structural peculiarities (Fig. 5 and 6). Both are relatively rich in cysteine residues (there are 10 cysteine residues in bel-1 and 12 cys residues in bel-2), and have either one (bel-2) or two (bel-1) glycosylation signals, properties that are reminiscent of certain hormone sequences. The bel-2 has one Arg-Gly-Asp sequence, a versatile cell recognition signal that is crucial for interacting with its cell surface receptor (18). Furthermore, both bel sequences possess potential phosphorylation sites of the type Arg-Arg-Ser or -Thr. Finally, on the other hand, both putative genes show remarkably long stretches of strong basic residues that are reminiscent of similar regions in the tat proteins of HIV (16). It remains to be studied if the bel genes are expressed and what functions they possess.

```
                            bel-1
                  10          20          30          40

   1  LLLRNLKENP  RRPRRYTKRE  VKCVSYHAYK  EIEDKHPQHI

  41  KLQDWIPTPE  EMSKSLCKRL  ILCGLYSAEK  ASEILRMPFT

  81  VSWEQSDTDP  DCFIVSYTCI  FCDAVIHDPM  PIRWDPEVGI

 121  WVKYKPLRGI  VGSAVFIMHK  HQRNCSLVKP  STSCSEGPKP

 161  RPRHDPVLRC  DMFEKHHKPR  QKRPRRRSID  NESCASSSDT

 201  MVNEP*
```

FIGURE 5. Amino acid sequence of the bel-1 coding region. The thick vertical arrow marks the position of the splice acceptor site corresponding to that shown in fig. 2. N-linked glycosylation sites are boxed; cysteine residues are underlined. The strongly basic amino acid stretches is doubly underlined.

The LTR Region

There are 107 nucleotides between the termination codon of the bel-2 open reading frame and the start of the 3' LTR. The LTR sequence is 963 nucleotides long and is located at the 3' end of the B52 DNA. It contains a number of key features of regulatory signals shown to be required for retroviral replication and transcription. There is a characteristic sequence, AGAA AAAGGTGGGAAAG, the polypurine tract (*figure 3*), that defines the 5' boundary of the HSRV LTR and is found in all retroviral genomes at this location and is the putative primer for plus strand DNA synthesis of retroviruses. Within the 3' presumed LTR there is one poly(A)-addition signal AATAAA (position 6576 to 6581) which perfectly matches the canonical consensus sequence. If the location of the poly(A)-signal is assumed to be nonvariant with respect to the polyadenylation-site of HSRV, one can draw an analogy to the LTRs of HIV and VIV (14,8), and place the polyadenylation site 24 bp downstream of the poly(A)-addition signal at CA (position 6600). Similar arguments indicate that the cap site starts with a G close to position 6412. A perfect and multiple TATATA-box precedes the presumptive RNA initiation

```
             10          20          30          40
              ↓
  1  NINPSEELLD  LLCLLCINIK  ETVLLLNLLP  VAQKVQNQDL
 41  GTILSFAVTC  LKSITSLGRN  DPRDDSSIMS  HVLSVVTPWS
 81  MSHDHYAPTL  FGILDRYYQG  YLKSSATYQT  WKFTCTVDPS
121  GKRFMGTQFW  VPPLGQVNIQ  FYKNYQILTC  CQAVDPFANI
161  FHGTDEEMFD  IDSGPDVWCS  PSLCFKVIYE  GAMGQKQEQK
201  TWLCRLGHGH  RMGACDYRKV  DLYAMRQGKE  NPYGDRGDAA
241  LQYAYQVKRG  CKAGCLASPV  LNYKALQFHR  TIMADFTNPR
281  IGEGHLAHGY  QAAMEAYGPQ  RGSNEERVWW  NVTRNQGKQG
321  GEYYREGGEE  PHYPNTPAPH  RRTWDERHKV  LKLSSFATPS
381  DIQR*
```

bel-2

FIGURE 6. Amino acid sequence of the bel-2 gene. The thick arrow marks the position of the splice acceptor site as shown in fig. 2. The cell recognition Arg-Gly-Asp (RGD) (18) is doubly underlined. A putative N-linked glycosylation signal is boxed and cysteine residues are underlined.

site by 31 bp. Thus, the HSRV LTR can be subdivided into three regions. The R and U5 regions were calculated to be 193 and 154 bp, respectively. The U3 region was estimated to be 616 bp long. Restriction fine mapping of another recombinant clone that was derived from viral DNA and that contained sequences from the 5'-region of the HSRV genome resulted in DNA fragment sizes that were consistent with those obtained with the 3' LTR sequences from B52 DNA (R.M. Flügel, unpublished observation).

Although the HSRV LTR is unusually long, open reading frames for encoding proteins of more than 30 amino acid residues were not found. Furthermore, the HSRV LTR has little if any sequence homology to the LTRs of other retrovirus, in particular to those of HIV-I, HTLV-II, VIV, MMTV, and HIV-2 (14,17,18,10,19). In addition, no significant sequence homology was found to human endogenous retrovirus-like sequences including the 968 bp long LTR sequence of human endogenous retroviral HERV genes (20).

There are ten octamer sequences and six nonamers that form direct repeats within the LTR. The indirect repeats include a decamer at position 6646 and 6743 with a loop size of 78 bp and two octamers with a relatively small loop size of 34 and 5 bp, respectively. The significance of these repeats with respect to the regulation of transcription remains to be elucidated.

DISCUSSION

There are a number of common as well as distinguishing features when the HSRV genome is compared to the genomes of other retroviruses. In general, the characteristics typical for a retrovirus have been readily identified in foamy viruses in the past. It has an RNA genome composed of two subunits and a virion reverse transcriptase, concomitant with a RNAse H activity (1, and references cited therein). The fact that we succeeded in molecular cloning proviral DNA from HSRV-infected human cells and that we can sequence across the 5' LTR into the cellular flanking regions unambiguously proves that HSRV integrates into the host DNA. Furthermore, the protein sequences deduced from the DNA sequence of the HSRV pol gene and their comparison with the corresponding reverse transcriptase domains of other retroviruses clearly show that the HSRV pol gene product can be subdivided into a RNAse H and an endonuclease (integrase) domain.

Furthermore, the HSRV genome contains 3' to the pol gene an env gene, that is about as long as the env genes of lentiviruses. Although the primary structure of the HSRV env gene is unique, its overall structural organization and hydropathic profile is that of a typical retroviral env gene. It is rich in cysteine residues, contains many glycosylation sites, a signal peptide at its NH_2-terminus, and a proteolytic processing site of the same type as HIV-1 and VIV env gene products, that generates the mature env proteins that are essential for the budding process and necessary for the attachment of infectious virus to the surface receptors of the host cell.

On the other hand, spumaretroviruses seem to have properties that set them apart from the other subfamilies of retroviruses, namely from the onco- and lentiviruses, and from other members of the retrovirus family, e.g. the HTLV/BLV and the D- and B-type viruses. The large size of the HSRV genome can be accounted for in broad terms, since the different genes (gag, pol, env) and genomic regions (LTR and post-env) were found to be of a correspondingly greater size. The LTR is quite large, with a size of 963 bp as are the gag, pol, and env genes. Moreover, the bel coding regions that are located in the post-env part of the HSRV genome are unique with respect to size when compared to other retroviral genomes. Apparently, HSRV does not have a sor gene like HIV-1 and VIV do.

In conclusion, HSRV has a unique genomic structure that differs from those of other retroviruses in certain respects. A final and definitive phylogenetic placement will be possible when the 5' part of the HSRV DNA sequence will be determined which is currently underway. The fascinating aspect of the elucidating functions for the bel coding regions will allow to determine whether or not they can be correlated with the regulation of foamy virus expression. It will be of interest to study what regions of the HSRV genome are responsible for the cytopathic effects and to find out if foamy viruses cause a human disease.

ACKNOWLEDGEMENT

We thank Dr. Ronald C. Montelaro, Louisiana State University, for communicating his results to us. The expert editorial assistance of Ms. M. Kasamasch is gratefully acknowledged.

REFERENCES

1. Weiss R, Teich N, Coffin J (eds) (1985) Taxonomy of retroviruses. In "RNA Tumor Viruses", New York: Cold Spring Harbor Laboratory, 2nd edition, 2, p 25.
2. Flügel RM, Maurer B, Bannert H, Rethwilm A, Schnitzler P, Darai G (1987). Nucleotide sequence analysis of a cloned DNA fragment from human cells reveals homology to retrotransposons. Mol Cell Biol 7:231.
3. Achong BG, Mansell PW, Epstein MA, Clifford P (1971). An unusual virus in cultures from a human nasopharyngeal carcinoma. J Nat Cancer Inst 46:299.

4. Flügel RM, Rethwilm A, Maurer B, Darai G (1987). Nucleotide sequence analysis of the env gene and its flanking regions of the human spumaretrovirus reveals two novel genes. EMBO J

5. Rushlow K, Olsen K, Steigler G, Payne SL, Montelaro RC, Issel CJ (1986). Lentivirus genomic organization: the complete nucleotide sequence of the env gene region of equine infectious anemia virus. Virology 155:309.

6. Kozak M (1984). Compilation and analysis of the sequences upstream from the transcriptional start site in eukaryotic mRNAs. Nucleic Acids Res 12:857.

7. Kyte J, Doolittle RF (1982). A simple method for displaying the hydropathic character of a protein. J Mol Biol 157:105.

8. Sonigo P, Alizon M, Staskus K, Klatzmann C, Cole S, Danos O, Retzel E, Tiollais P, Haase A, Wain-Hobson S (1985). Nucleotide sequence of the visna lentivirus: relationship to the AIDS virus. Cell 42:369.

9. Fasel N, Pearson K, Buetti E, Diggelmann H (1982). The region of mouse mammary tumor virus DNA containing the long terminal repeat includes a long coding sequence and signals for hormonally regulated transcription. EMBO J 1:3.

10. Majors JE, Varmus HE (1983). Nucleotide sequencing of an apparent proviral copy of env mRNA defines determinants of expression of the mouse mammary tumor virus env gene. J Virol 47:495.

11. Redmond SMS, Dickson C (1983). Sequence and expression of the mouse mammary tumor virus env gene. EMBO J 2:125.

12. Coffin JM (1986). Genetic variation in AIDS viruses. Cell 46:1.

13. Bolognesi DP, Montelaro RC, Frank H, Schäfer W (1978). Assembly of type C oncornaviruses: a model. Science 199:183.

14. Ratner L, Haseltine W, Patarca R, Livak KJ, Starcich B, Josephs SF, Doran ER, Rafalski JA, Whitehorn EA, Baumeister K, Ivanoff L, Petteway SR jr, Pearson ML, Lautenberger JA, Papas TS, Ghrayeb J, Chang NT, Gallo RC, Wong-Staal F (1985). Complete nucleotide sequence of the AIDS virus, HTLV-III. Nature 313:277.

15. Benzair A-B, Rhodes-Feuillette A, Lasneret J, Emanoil-Ravier R, Périès J (1985). Purification and characterization of the major envelope glycoprotein of simian foamy virus type I. J Gen Virol 66:1449.

16. Sodroski J, Rosen C, Wong-Staal F, Salahuddin SZ, Popovic M, Arya SK, Gallo RC, Haseltine WA (1985). Trans-acting

transcriptional regulation of human T-cell leukemia virus type III long terminal repeat. Science 227:171.

17. Shimotohno K, Wachsman W, Takahashi Y, Golde DW, Miwa M, Sugimura T, Chen ISY (1984). Nucleotide sequence of the 3' region of an infectious human T-cell leukemia virus type II genome.

18. Ruoshlat E, Pierschbacher MD (1986). Arg-Gly-Asp: a versatile cell recognition signal. Cell 44:517.

19. Clavel F, Guyader, M, Guétard D, Sallé M, Montagnier L, Alizon M (1986). Molecular cloning and polymorphism of the human immune deficiency virus type 2. Nature 324:691.

20. Ono M. (1986). Molecular cloning and long terminal repeat sequences of human endogenous retrovirus genes related to types A and B retrovirus genes. J Virol 58:937.

CORRECTION

The nucleotide sequence analysis of another recombinant clone of proviral DNA of HSRV revealed that the LTR region is longer by 298 bp. This results in a total length of 1216 bp for the LTR, with a U3 region of 814 bp. Correspondingly, the 3' LTR overlaps the bel-2 gene by 53 nucleotides, so that the polypurine tract at position 5472 to 5496 has to be postulated to be primer for plus strand DNA synthesis in analogy to all other retroviruses.

Accordingly, the 3'-LTR regions shown in fig. 1 and 2 should be redrawn to represent the overlap into the bel-2 part of the HSRV genome.