JOURNAL OF VIROLOGY, June 1996, p. 3606–3616
0022-538X/96/$04.00+0

Vol. 70, No. 6

Complete Nucleotide Sequence and Transcriptional Analysis of the Snakehead Fish Retrovirus

DOUGLAS HART,[1]* G. NICOLAS FRERICHS,[2] ANDREW RAMBAUT,[3] AND DAVID E. ONIONS[1]

Department of Veterinary Pathology, University of Glasgow Veterinary School, Glasgow G61 1QH,[1] Institute of Aquaculture, University of Stirling, Stirling FK9 4LA,[2] and Department of Zoology, University of Oxford, Oxford OX1 3PS,[3] United Kingdom

Received 2 November 1995/Accepted 11 March 1996

The complete genome of the snakehead fish retrovirus has been cloned and sequenced, and its transcriptional profile in cell culture has been determined. The 11.2-kb provirus displays a complex expression pattern capable of encoding accessory proteins and is unique in the predicted location of the *env* initiation codon and signal peptide upstream of *gag* and the common splice donor site. The virus is distinguishable from all known retrovirus groups by the presence of an arginine tRNA primer binding site. The coding regions are highly divergent and show a number of unusual characteristics, including a large Gag coiled-coil region, a Pol domain of unknown function, and a long, lentiviral-like, Env cytoplasmic domain. Phylogenetic analysis of the Pol sequence emphasizes the divergent nature of the virus from the avian and mammalian retroviruses. The snakehead virus is also distinct from a previously characterized complex fish retrovirus, suggesting that discrete groups of these viruses have yet to be identified in the lower vertebrates.

The family *Retroviridae* comprises seven genera of viruses isolated and characterized from a variety of vertebrate species and which display a diverse range of host interactions and replication strategies (37, 57). The classification of retroviruses, formerly based on morphology and pathogenicity, is now more centered on a number of characteristics of the viral genome, including gene organization and sequence homology (9). Although retroviruses appear to be ubiquitous in vertebrates, there is an incomplete representation of the comparative relationship of these viruses outside of the mammalian and avian species.

The snakehead retrovirus (SnRV) was originally identified as a spontaneously productive infection of the SSN-1 cell line, derived from the Southeast Asian striped snakehead fish (*Ophicephalus striatus*) (18). Electron microscopy of morphologically normal SSN-1 cells showed the release of 85- to 90-nm-diameter type C retrovirus particles from the outer cell membranes. Cell culture supernatants displayed a high level of Mn^{2+}-dependent reverse transcriptase (RT) activity associated with a sucrose density gradient fraction of 1.16 g/ml. SnRV induced cytolytic changes in cultures of the BF-2 cell line, derived from the bluegill fry (*Lepomis machrochirus*), and productive infection was confirmed by electron microscopy of cells and RT assay of the cell culture supernatant. Southern blot and PCR results have suggested the retrovirus to be exogenous in nature to the striped snakehead species (unpublished results), and its potential pathogenicity is under investigation. We report here the cloning, complete nucleotide sequence, and transcriptional profile of SnRV from the SSN-1 cell line. In addition, we compare SnRV to the well-characterized mammalian and avian retroviruses and to a recently sequenced fish retrovirus, the walleye dermal sarcoma virus (WDSV) (25).

MATERIALS AND METHODS

Cell culture and virus purification. Cultures of retrovirus-infected SSN-1 cells and noninfected SSN-2 cells, both derived from striped snakehead fish, were grown and maintained in serum-supplemented Leibovitz L-15 medium. Retrovirus particles were purified by sedimentation velocity and equilibrium density gradient centrifugation. Pelleted virus from clarified SSN-1 medium was resuspended in TNE buffer (10 mM Tris-HCl [pH 7.5], 100 mM NaCl, 1 mM EDTA), layered onto a 5 to 20% step sucrose gradient, and centrifuged at 100,000 × *g* for 20 min. The visible viral band was collected, layered onto a 20 to 50% continuous sucrose gradient, and centrifuged at 150,000 × *g* for 3 h. Gradient fractions were collected and assayed for RT activity as described previously (18). Fractions with peak activity were pooled and stored at −80°C.

RT-PCR amplification of viral genomic RNA. Purified virus was diluted in TNE buffer, pelleted at 200,000 × *g* for 30 min and resuspended in disruption buffer (0.1% Nonidet P-40, 10 mM Tris-HCl [pH 8.3], 1 mM dithiothreitol, 1 U of Pharmacia Biotech RNAguard per μl) to release the viral genomic RNA. cDNA synthesis was performed with a first-strand cDNA synthesis kit (Pharmacia Biotech) with the nonspecific primer P3, 5′-AGTATCGATCTCGAGTT-3′. One third of the completed cDNA synthesis reaction mixture was added to a PCR reaction mix containing 1 × *Taq* buffer (10 mM Tris-HCl [pH 8.3], 50 mM KCl, 1.5 mM $MgCl_2$, 0.01% gelatin), 2.5 U of Ampli*Taq* (Perkin Elmer), and 30 to 40 picomol of primer PR3, 5′-CGTGCGGCCGCGAATTCNNNNGT-3′. Amplification was performed with an initial denaturation of 2 min at 95°C, followed by 3 cycles of 15 s at 95°C, 30 s at 37°C, and 90 s at 72°C and 27 cycles of 15 s at 95°C, 30 s at 50°C, and 90 s at 72°C, with a final extension of 6 min at 72°C.

DNA sequencing and analysis. PCR products were purified, cloned into pBluescript KS (Stratagene) or pGEM-T (Promega) vectors, and sequenced on both strands by the dideoxy chain termination method of Sanger et al. (53). A minimum of two clones were sequenced for all positions of the genome. Sequence data were compiled and analyzed using the University of Wisconsin Genetics Computer Group package. Nucleotide sequences and deduced amino acid sequences were compared with database entries by using the FASTA and BLAST search programs. Coiled-coil predictions were determined by using the COILS program (32, 43). Transcription factor sites were identified by using the SIGNAL SCAN program of Prestridge (47).

PCR amplification of proviral DNA. Genomic DNA was extracted from SSN-1 cells by the method of Miller et al. (36). Proviral DNA was PCR amplified with the following primers derived from the RT-PCR clone sequence data (Fig. 1): GPOL3, 5′-GAGTACCACACCTAGGTGGGTACCACCC-3′; DG11, 5′-TTTTGGTCACTCCTGTGGGTATGAACC-3′; DG10, 5′-CATACAGAAAGTGACTACACCTCGCC-3′; GU34, 5′-GAATAAGTCCTGTTCCCAGGGACTAGCG-3′; NPOL2, 5′-TCTACCTTTCCCTGTGATTGAGGATG-3′; NGAG2, 5′-CTAAGTTAAGACAAGGACCCACTGAG-3′; NU31, 5′-TCATAAGGCCAATCACGCTAGGAG-3′; NGAG3, 5′-GGGGTCATCTTTGTTACCGTCTCT-3′; LTR10, 5′-TGACTCATATCCTGCTTAGTAGAC-3′; LTR20, 5′-GAAAGTACGACTCAGGCTCAAGAC-3′; ML1, 5′-TGGTACCCATGGATACAGGTACCTCA-3′; and GPOL2, 5′-TGTCAGACATGGCCTGTACTTTAGCAGC-3′.

Inverse PCR. *Alu*I-digested SSN-1 genomic DNA (100 ng) was self-ligated in

* Corresponding author. Mailing address: Department of Veterinary Pathology, University of Glasgow Veterinary School, Bearsden Rd., Glasgow G61 1QH, United Kingdom. Phone: 141 330 6934. Fax: 141 330 5602. Electronic mail address: gvpv22@udcf.gla.ac.uk.

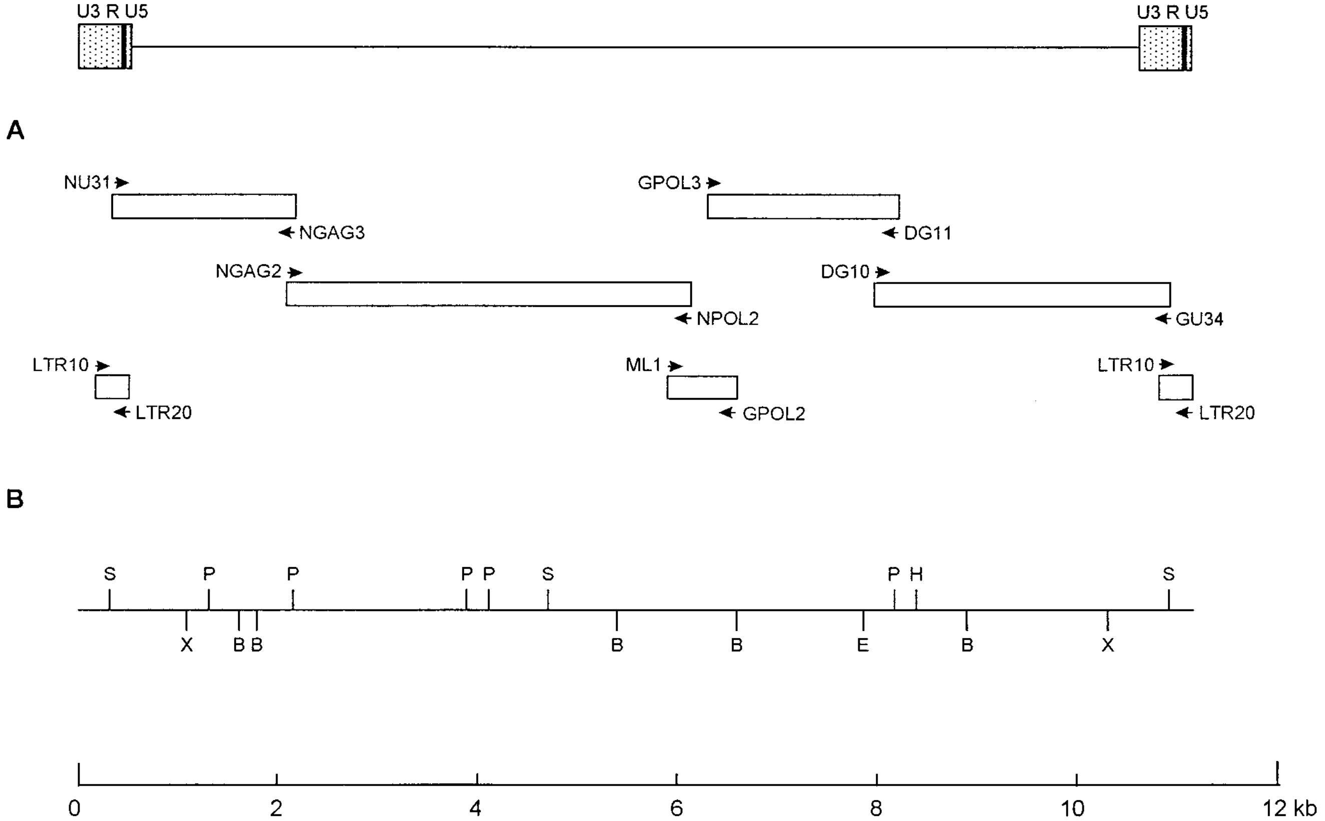

FIG. 1. Positions of PCR products and provirus genome restriction map. (A) The locations of the primers (arrows) used in PCR amplification of the proviral DNA and the resulting products (open boxes). (B) The positions of restriction sites used in Southern analysis of the provirus. In order to confirm that the genome structure as deduced from sequence analysis of the overlapping PCR products corresponded to that of the provirus, the restriction map of the provirus was compared with the deduced map by Southern hybridization analysis of digested SSN-1 DNA. The four major PCR products (in panel A) plus RT-PCR clones corresponding to positions 7076 to 8106 and 9948 to 10658 were used as probes in the analysis. B, *Bam*HI; E, *Eco*RI; H, *Hin*dIII; P, *Pst*I; S, *Sca*I; X, *Xho*I.

a volume of 50 μl containing 1 × T4 DNA ligase buffer and 1 U of T4 DNA ligase (Promega) at 14°C overnight. The ligase was heat inactivated at 70°C for 10 min, 10 U of *Sca*I was added, and the reaction mixture was incubated at 37°C for 2 h. The reaction mixture (5 μl) was amplified in a final volume of 50 μl containing 1× Ampli*Taq* buffer, 200 μM dNTPs, 2 mM $MgCl_2$, 2.5 U of Ampli*Taq*, and 25 pmol of primers NRP3 (5′-GGACTAGCGTATGTGCATGC TAAC-3′) and IPS-2 (5′-AGATTCGAGAAACGAAGCGTCAAC-3′). Thermocycling was carried out for 90 s at 95°C and for 35 cycles of 15 s at 95°C, 30 s at 65°C, and 1 min at 72°C, with a final extension of 7 min at 72°C.

Northern (RNA) hybridization. Polyadenylated mRNA was isolated from SSN-1 and SSN-2 cells with the Quickprep mRNA purification kit (Pharmacia Biotech). Each mRNA preparation (2.5 μg) was separated on a 2% formaldehyde gel and transferred to a nylon membrane. A 710-bp RT-PCR clone corresponding to nucleotides 9948 to 10658 of the SnRV long terminal repeat (LTR) region was randomly labelled with [α-^{32}P]dCTP by using an oligonucleotide-labelling kit (Pharmacia Biotech) and purified through a Sephadex G50 column (Nick column; Pharmacia Biotech). Hybridization in 50% formamide solution and high-stringency washes were carried out as described by Sambrook et al. (52).

Transcript analysis. Analyses of retroviral transcripts were performed using SSN-1 mRNA isolated as described above. The cap site of the retrovirus transcripts was determined by using the 5′-AmpliFINDER RACE kit (Clontech) with primer SLIC1, 5′-CTGGTGTTTCGTAAGCTGCATCTG-3′, for cDNA synthesis and primer SLIC2, 5′-ACGTGCTGCCTAGTCTACCAATAC-3′, for amplification. The polyadenylation site was determined by using the 3′-Ampli FINDER RACE kit (Clontech), with primer LTR10. Subgenomic transcripts were amplified with primer NRP3, 5′-GGACTAGCGTATGTGCATGCTAAC -3′, for cDNA synthesis and primers NRP2, 5′-GTCGTACTTTCACCGTTTA CAGTG-3′, and F4P01, 5′-CTCATGCTCAGACAGATCCGGACTGATC-3′, NRP2 and NRP3, or NRP2 and TENV1, 5′-CGTCCCTAAGGAGTATGTA TCCTG-3′, for amplification.

Phylogenetic analysis. Amino acid sequences were aligned with the CLUSTAL W package (58) by using the BLOSUM series of protein weight matrices (24) and adjusted manually to remove further gaps. The PHYLIP package (15) was used for the analysis with protein distances calculated by using the Dayhoff PAM 250 matrix (10), and the tree was constructed by the neighbor-joining method of Saitou and Nei (50). The SEQBOOT and CONSENSE programs were used for bootstrap analysis of the aligned sequences.

Nucleotide sequence accession number. The complete nucleotide sequence of the SnRV genome is available in the GenBank database under accession no. U26458.

RESULTS

Nonspecific RT-PCR of genomic RNA isolated from purified virions yielded clones that could be identified to putative *gag*, *pol*, *env*, and LTR regions by comparison to other retrovirus sequences. Primers were derived from the sequence data and used to amplify overlapping regions covering the predicted proviral genome (Fig. 1A). The size of the genome deduced from sequence analysis of the PCR products was confirmed as corresponding to that of the provirus by restriction mapping (Fig. 1B) and Southern blot analysis (data not shown). The complete nucleotide sequence is shown in Fig. 2, with indicated nucleotide positions relative to the cap site (+1) of the RNA genome.

The sizes of the RNA genome and provirus are 10,688 nucleotides and 11,157 bp, respectively. Sequence analysis of the provirus found that it displays a typical LTR-*gag-pol-env*-LTR retrovirus organization (Fig. 3). Separation of the *gag* and *pol* regions by an in-frame stop codon is analogous to the organization found in the mammalian type C retroviruses, where expression of *pol* occurs by translational suppression of the *gag* termination codon (27, 68). The genome sequence has a base

```
      →                     R                      ←→                 U5                 ←       PBS
   1  GTCTTTTGCTGCACCCAACTTGAAGAATAAAGATTATTGCATCTGACCCGTCTTGAGCCTGAGTCGTACTTTCACCGTTTACAGTGGCGAGCCAGCCAGG
                                                                                  LP  ↱
 101  AGCCTTTTCTTCATAAAGAGATTCGAGAAACGAAGCGTCAACCCCGGGGAGAAACGACAGGAGGAAGCGCGCCTGGAACATGAAGCTGGTTCTTCTGTTC
                                                                                     M  K  L  V  L  L  F
                         ↓SD
 201  AGCCTCAGCGTTCTACTTGGGTGAGTTCGTTCTAATCTATTTATCGGAAAGAATACAAGGCTACAAGAAACCTGGTAAAACAAAGAAAAGACATTACAAA
      S  L  S  V  L  L  G  *
                                     Gag   ↱   MA/?
 301  AGAAAGTATTGGTAGACTAGGCAGCACGTAAAAGAAATGGCCTCCAACAAATGGTTTGTGTACAGCGACGAGCCAACAAAAGTAATACTAAAACGAGACA
                                          M  A  S  N  K  W  F  V  Y  S  D  E  P  T  K  V  I  L  K  R  D  K

 401  AAAGCAAGGAAAAAGATGAAACTAAAAAGAAGAAAATTAAAACAGAACAAAATTCAGATGCAGCTTACGAAACACCAGGCACCGCGCCTGTACAGAAACC
        S  K  E  K  D  E  T  K  K  K  K  I  K  T  E  Q  N  S  D  A  A  Y  E  T  P  G  T  A  P  V  Q  K  P

 501  ACTGTTGGAAACCACGCCTGAAGCAGAATTAGAAAAGGTTCTTAAAGGGCTAGAAGAGTGGGGATACAAGGCTTTAGAGAAGAAGAGGGACCCAGAGCTT
       L  L  E  T  T  P  E  A  E  L  E  K  V  L  K  G  L  E  E  W  G  Y  K  A  L  E  K  K  R  D  P  E  L

 601  TGGAACCCCAATCAAGAAGGACTAGATGAGTACCTCTTTAGGGGCTGGGTTCAGGGAGGCCTAGCCGATTCTAAGAAAGCCCTCGAGAAAAACATGGAAA
      W  N  P  N  Q  E  G  L  D  E  Y  L  F  R  G  W  V  Q  G  G  L  A  D  S  K  K  A  L  E  K  N  M  E  K
                                                                                                      CC
 701  AATTTGTACCTTTGTTTGTGGTCACCATGTCCCAAGCAGTACCATATTGGCGGCAGACCATGCAGGCACGCAATCAAAATGGTAAAAAGCAAAAGAACCG
        F  V  P  L  F  V  V  T  M  S  Q  A  V  P  Y  W  R  Q  T  M  Q  A  R  N  Q  N  G  K  K  Q  K  N  R

 801  CATTGCTGAACTAGAAAAAGAGGTAGCAGACTTAACTAGCGCCGGTAGGGGAGCTGATCAGGTTATAGCAGGTATGGACAAAGAACTAAAGAAAACTGCA
       I  A  E  L  E  K  E  V  A  D  L  T  S  A  G  R  G  A  D  Q  V  I  A  G  M  D  K  E  L  K  K  T  A

 901  GAGAAATATCAAGCCAAATTAGAAGAACTAGAGGAGCAGTTAGCTGCGATGACGGTAGAAAAGGAGGAATTGGAAAGCCAGGTTGAAGGGTTAAAGGAAT
      E  K  Y  Q  A  K  L  E  E  L  E  E  Q  L  A  A  M  T  V  E  K  E  E  L  E  S  Q  V  E  G  L  K  E  S

1001  CTCTAGTAGAAGCAGAAACTAAGAAAGTCAGTCTCATGGAAGTATTGACTATGCCCACCAGATCAAAAGGACCGAAGAAACGGGGCCCTGATTTGAAACA
        L  V  E  A  E  T  K  K  V  S  L  M  E  V  L  T  M  P  T  R  S  K  G  P  K  K  R  G  P  D  L  K  Q

1101  GATTAGATCGTTGCACGTGATGGCTGATTCATTGGGCATGGACAGTGATGGCATAGACTGGGATTGGTTGGCTAGACAAGCCTGGGACTATGAAGGAGAT
       I  R  S  L  H  V  M  A  D  S  L  G  M  D  S  D  G  I  D  W  D  W  L  A  R  Q  A  W  D  Y  E  G  D
                                 MA/?  ↰ ↱  CA
1201  GAGGATCCACATGTTAAGGAAGCAGAGGAGGAGGCATGGCGTGAGAGACAAACATCTCAACCATCCCAACCCTCTCAGTTGAGACCTTTTGTAGCCGCAG
      E  D  P  H  V  K  E  A  E  E  E  A  W  R  E  R  Q  T  S  Q  P  S  Q  P  S  Q  L  R  P  F  V  A  A  G

1301  GGAACGGTCATAGGGAAGATCAATGGAGACCCCTGACAGTCACAGAATTACCCGCAGCCGTTACCGCAGTAGGAGGAGCATGGGATCCCACGAGGGAAAC
        N  G  H  R  E  D  Q  W  R  P  L  T  V  T  E  L  P  A  A  V  T  A  V  G  G  A  W  D  P  T  R  E  T

1401  AGGAAGTGCTAGGTGGAAAAAGATAGTGAAAGCAGCAGAGGCTATAGGATGGGGGACAGGCGATGTGTGTCAGGTTGTGACCGCTATGTCACCATCATGG
       G  S  A  R  W  K  K  I  V  K  A  A  E  A  I  G  W  G  T  G  D  V  C  Q  V  V  T  A  M  S  P  S  W

1501  GCAGATGTGCCTCCCGAAATAAGAAATAGGGTGGCTACTGAGAAAGAAATAAAAGCATGGTTGATGAAGCAGGGACCCGGAGGAGGACAGGGTTTGTTGG
      A  D  V  P  P  E  I  R  N  R  V  A  T  E  K  E  I  K  A  W  L  M  K  Q  G  P  G  G  G  Q  G  L  L  E
                                             MHR
1601  AATTTACTAAGTTAAGACAAGGACCCACTGAGAATCCTAGTAACTACTTGGAAAAGGCTCTAGAATTGTACCTAGACTCTCAACCCGGTGATAGAGACGG
        F  T  K  L  R  Q  G  P  T  E  N  P  S  N  Y  L  E  K  A  L  E  L  Y  L  D  S  Q  P  G  D  R  D  G

1701  TAACAAAGATGACCCCGCATTTCTGCAGCAGGCTACTCAAGGCCTGTTGCCTTGGTTAAAGAAAGCTGTGATATTAGGAGGAAAAAACACGTCATGGCAA
       N  K  D  D  P  A  F  L  Q  Q  A  T  Q  G  L  L  P  W  L  K  K  A  V  I  L  G  G  K  N  T  S  W  Q
                                   CA  ↰ ↱  NC
1801  GAGATGACTAGCTTCTGCCAGAGGTTGTGGTTGGTCAGAGATCAATTTGCCGACAAAACAGGTGTCTCAAAGGCCCGACCTATTGTCAGAAATGAAGGAC
      E  M  T  S  F  C  Q  R  L  W  L  V  R  D  Q  F  A  D  K  T  G  V  S  K  A  R  P  I  V  R  N  E  G  P

1901  CAAGACCACAACAAGGGCACAGCAAAATTGTTTTTGGGGGAAACTGCCGAAATTGTGGAAAAGCAGGACACATGGCTAGAGATTGTTGGGCCAAAGGTGG
        R  P  Q  Q  G  H  S  K  I  V  F  G  G  N  C  R  N  C  G  K  A  G  H  M  A  R  D  C  W  A  K  G  G

2001  TGGACAAGAAGGAAAAGGACCGCGTCAGAACACCACCTGGAAACCCAAATCTGGAGCAATAGCCAGCGCCCCACCTGCTGAAAGCCCTTATGCTGATTGT
       G  Q  E  G  K  G  P  R  Q  N  T  T  W  K  P  K  S  G  A  I  A  S  A  P  P  A  E  S  P  Y  A  D  C
                                                                                              Pol ↱
2101  GCTAAACAGTTAGCCGATATCGAAAAACGCCTTAAGGACCTCACCACTGCTGGTGGAGGACCCAAAGGACCCAACCCATTCCACAAACCATAGGGCGTTG
      A  K  Q  L  A  D  I  E  K  R  L  K  D  L  T  T  A  G  G  G  P  K  G  P  N  P  F  H  K  P  *  G  V  A
                   ↱  PR
2201  CGGCCCTCGCCGCCCCACTCTGCTCCTTGAAGGGACGGCCACACGTATCTGTAGAGATAGAGGGCCACAAGATCGAGTGTCTAGTGGACACAGGAGCAGA
        A  L  A  A  P  L  C  S  L  K  G  R  P  H  V  S  V  E  I  E  G  H  K  I  E  C  L  V  D  T  G  A  E

2301  GGTGTCATTAACATCCCTTCAATTACAGGCACAACGATTTGAACAGGTGGTTGGTTTGGGAGGCAAACCGGTGAGAGTAGGGATAGCGGATCATGTCGAT
       V  S  L  T  S  L  Q  L  Q  A  Q  R  F  E  Q  V  V  G  L  G  G  K  P  V  R  V  G  I  A  D  H  V  D

2401  ACCACAGTAGGACAAGTAAAAGGAAAAGGCTGCTGGCGAATATCACAGGAATTAGCGGAAAATATCCTAGGAAATGATCTGTTGAGGTCATTAGGATTAA
      T  T  V  G  Q  V  K  G  K  G  C  W  R  I  S  Q  E  L  A  E  N  I  L  G  N  D  L  L  R  S  L  G  L  I
          PR  ↰                                                 ↱  RT/RNH
2501  TAGTGGATCAATGCAACGGAGTTCTATGGCAGGCTAGTGAGGGCCTAGGACCCGATAATTGGATGGCAGCAGAAACACTAAGGATATACAGTATTAAATC
        V  D  Q  C  N  G  V  L  W  Q  A  S  E  G  L  G  P  D  N  W  M  A  A  E  T  L  R  I  Y  S  I  K  S

2601  ACCCGGTCATTATAACCTACCTGAACTATTAGCCACTAAAGATGAACAACTAGCTGATATCCTATGGAACAACGTTGAAGCCTTTGCTACTCACAGAAAT
       P  G  H  Y  N  L  P  E  L  L  A  T  K  D  E  Q  L  A  D  I  L  W  N  N  V  E  A  F  A  T  H  R  N

2701  GATTGCGGGAACTTACAGGGCATGACTGCCAGTTTTACCGCTGACCATCCAAAAATGATTAAACAATACCCGGTACCGGATGCATCACATGCTAGCATAA
      D  C  G  N  L  Q  G  M  T  A  S  F  T  A  D  H  P  K  M  I  K  Q  Y  P  V  P  D  A  S  H  A  S  I  K

2801  AGGAAACTGTGGAAGCATTATTGGAACAAGGTGTTCTTAGAAAGTGTAATAGCACAGTTAACAGTGCTATATGGCCAGTAGGCAAGCCGGATGGATCATG
        E  T  V  E  A  L  L  E  Q  G  V  L  R  K  C  N  S  T  V  N  S  A  I  W  P  V  G  K  P  D  G  S  W

2901  GAGGCTAACCATTGATTATAGGCCACTCAACTCGGCTGTCTCGTGTCCATATCCCACAGTAGCCTCAACCCCAGAGTTGTTTGCTAAACTAGAGAAGAAA
       R  L  T  I  D  Y  R  P  L  N  S  A  V  S  C  P  Y  P  T  V  A  S  T  P  E  L  F  A  K  L  E  K  K
```

FIG. 2. Complete nucleotide sequence and deduced amino acid sequences of the SnRV genome. The nucleotide sequence is shown from the cap site (position 1) to the polyadenylation site (position 10688). Features of the sequence are indicated by arrows or are underlined. The inverse repeats of the LTR are underlined twice. Vertical arrows denote splice junction sites and potential Env cleavage sites. Predicted methionine start residues are underlined, and termination codons are marked by an asterisk. The predicted positions of the coding domains are based on sequence homology. CA, capsid; CC, coiled-coil; LP, leader peptide; MA, matrix; MHR, major homology region; NC, nucleocapsid; PBS, primer binding site; PPT, polypurine tract; polyA, polyadenylation signal; PR, protease; RT/RNH, reverse transcriptase/RNase H; SA, splice acceptor; SD, splice donor; TMA, transmembrane anchor; ?, unknown coding domain.

```
3001  TACCAGGTGTATAGCTCACTGGACATTAGCAACGGCTTCTGGTCCATCAGACTGGAGGAAGAATGTCAATACTTGTTTGCATTTACTTTCGATACGCAAC
      Y  Q  V  Y  S  S  L  D  I  S  N  G  F  W  S  I  R  L  E  E  E  C  Q  Y  L  F  A  F  T  F  D  T  Q  Q

3101  AGTATACATGGACCAGGCTACCACAGGGATTTCATGCTTCACCAGGCATATTCCACCAAGCCCTGTACAATGGACTGGCCTCCTGTAAAACAGCGATTGA
        Y  T  W  T  R  L  P  Q  G  F  H  A  S  P  G  I  F  H  Q  A  L  Y  N  G  L  A  S  C  K  T  A  I  E

3201  AAGTCAGGGTTGTAAACTATTGCAATATGTAGATGATATCCTGTTGATGAGTGAAGATAGGGACCATCATTTAAGGTCGCTGGCAATATTACTGCAAGGC
       S  Q  G  C  K  L  L  Q  Y  V  D  D  I  L  L  M  S  E  D  R  D  H  H  L  R  S  L  A  I  L  L  Q  G

3301  CTCAAGGATCTAGGAGTAAAAATCAATCCTAAGAAGTCCCACTTTTGCAAAGATCAGGTGCAGTACCTGGGAGTCAATGTAGGAGCCGACACCAGATCAC
      L  K  D  L  G  V  K  I  N  P  K  K  S  H  F  C  K  D  Q  V  Q  Y  L  G  V  N  V  G  A  D  T  R  S  L

3401  TGATCGATGCCAGAAGCCAACTGATAAGAACGTTAGACATCCCATTGACGGTGCAAGGCTTAAGATCAGCGCTGGGGTTGTTTAATTTCTGCAGAGCATG
        I  D  A  R  S  Q  L  I  R  T  L  D  I  P  L  T  V  Q  G  L  R  S  A  L  G  L  F  N  F  C  R  A  W

3501  GATACCTGAGTTCAGTCGGAAAACACAGAGTTTGTATGATATGCTTAAAGGAGACTGTAAATCTACTGATAAGCTAAAGTGGACTGAGGATAATCTGAAT
       I  P  E  F  S  R  K  T  Q  S  L  Y  D  M  L  K  G  D  C  K  S  T  D  K  L  K  W  T  E  D  N  L  N

3601  AAATTTAAACTCTTAAAGGACGAAGTAGCCAGCGCATGTGTGTTGGGCCTACCTGATCCCACCTTACCTTTCAGACATCTGATAGGAATCAGACAGGGAC
      K  F  K  L  L  K  D  E  V  A  S  A  C  V  L  G  L  P  D  P  T  L  P  F  R  H  L  I  G  I  R  Q  G  H

3701  ATTTTCTCTGCAGCCTTGTCCAGAAAGATGATAATGGCATGTGGCACGTGCTAGGGTTCTATTCCAGGAAAATGACACCTGTAGAAAGTAACCTAGGAAT
        F  L  C  S  L  V  Q  K  D  D  N  G  M  W  H  V  L  G  F  Y  S  R  K  M  T  P  V  E  S  N  L  G  I

3801  ATGTGAACAATATGCTGAATGTGCTGCCTGGGCCATCAGTGCCTGCAATCTGGTCAGTGGGTTCGGCAGAAAGATCATAGTAACATCTCACTCACCTGTA
       C  E  Q  Y  A  E  C  A  A  W  A  I  S  A  C  N  L  V  S  G  F  G  R  K  I  I  V  T  S  H  S  P  V

3901  AAGTTTATCCTCCAGACGACGCCTAATGTGAGTAATCAACGTTTAGCCAGGTGGCATAGAATATTGACCCAAGAAGACATCACTATAGAAACTGATGCTA
      K  F  I  L  Q  T  T  P  N  V  S  N  Q  R  L  A  R  W  H  R  I  L  T  Q  E  D  I  T  I  E  T  D  A  S

4001  GTATCCAAGGATGGTTTGTCCCGGAGCCTTATGAGGGGGAACAGCACCAGTGCCGACCTCATGATGACACAATAACATGGAGAGTCAGCACTAGAGCTAT
        I  Q  G  W  F  V  P  E  P  Y  E  G  E  Q  H  Q  C  R  P  H  D  D  T  I  T  W  R  V  S  T  R  A  M

4101  GCCCACGGGTGAGAAGTGGTGGATAGACGGAAGTAGATATTGGGATCATGATAAGGGAGGCTACGTTACAGGATGGGCAGCATTGAGAGAGGATAAGAAA
       P  T  G  E  K  W  W  I  D  G  S  R  Y  W  D  H  D  K  G  G  Y  V  T  G  W  A  A  L  R  E  D  K  K

4201  AATCAGCTGGGGGGTGCCTTAGAGGGACATGTGAGTGCACAAGTGGCAGAGTTAGTGGCGCTAAGGGAAGCCCTAAGGCTACAAAGACCTCTGACCCTAT
      N  Q  L  G  G  A  L  E  G  H  V  S  A  Q  V  A  E  L  V  A  L  R  E  A  L  R  L  Q  R  P  L  T  L  Y

4301  ATACTGACAGTACTTACGTGCTCGGCATTTGTACCAAATACCTTGCTGTTTGGAAAAGGAGGGGTATGGTCAATGCAGATGGATCACAGATCAGCAACCA
        T  D  S  T  Y  V  L  G  I  C  T  K  Y  L  A  V  W  K  R  R  G  M  V  N  A  D  G  S  Q  I  S  N  Q

4401  GAATATCCTACAAGAAATTTGGCAATTGATTGAACATGATTCTACACAAACGCTTGGTATAGTCAAGGTAAAAGCACATACCCAGCGTAAATGTAGCACG
       N  I  L  Q  E  I  W  Q  L  I  E  H  D  S  T  Q  T  L  G  I  V  K  V  K  A  H  T  Q  R  K  C  S  T
                                                                                        RT/RNH  ↰  ↱   ?
4501  CATGAACAACAACTGAATAATGATGTAGATCAGCCGGCAAAACAATATGCTAAGGAAGAACCTAACATGTCAGTCATAGCACCTCTGCAAGTCTATCCAT
      H  E  Q  Q  L  N  N  D  V  D  Q  P  A  K  Q  Y  A  K  E  E  P  N  M  S  V  I  A  P  L  Q  V  Y  P  L

4601  TATGGATAGGTTTGGTACCATGTAAAGAACCAAAATTGTGGGAAAATATACAACATCACATCACAAAGGTTGACCTGCCCGACTTCCAGAAACAACTTGC
        W  I  G  L  V  P  C  K  E  P  K  L  W  E  N  I  Q  H  H  I  T  K  V  D  L  P  D  F  Q  K  Q  L  A

4701  AGAAATAATGCCACAACAGGATATTTCACACTGCACACTAGCATATTTTGATAAACCATCCCCAGAAGCAACCAAATATCATGATAAAATAAAACCCTAC
       E  I  M  P  Q  Q  D  I  S  H  C  T  L  A  Y  F  D  K  P  S  P  E  A  T  K  Y  H  D  K  I  K  P  Y

4801  CTGGGAAAAGGACAGCAGCTCACACTGTGTGACACCTACATAGGCAAAGAAGGAGCTGCGATACTGGGACAATTGAGACCAGACATGCAAGCTCTTCATC
      L  G  K  G  Q  Q  L  T  L  C  D  T  Y  I  G  K  E  G  A  A  I  L  G  Q  L  R  P  D  M  Q  A  L  H  Q

4901  AGGCAGAAGGGGAAGTGCACGTTAGCCTAGGTACTCGTGCTGGGCACTGTCCACAGGAATTAGGTACAATGTTAACTAACCTGTTGAAAAGCACCCAAGA
        A  E  G  E  V  H  V  S  L  G  T  R  A  G  H  C  P  Q  E  L  G  T  M  L  T  N  L  L  K  S  T  Q  E

5001  ACGAATATGGCAGGATCCACCTGTTTTTAAACTGCATGATGAGACAGGAAAAGTACAAGGCTACGTCATTAAGACAACCTTAGGCATGAAAACATGGATA
       R  I  W  Q  D  P  P  V  F  K  L  H  D  E  T  G  K  V  Q  G  Y  V  I  K  T  T  L  G  M  K  T  W  I
                                 ?  ↰  ↱   IN
5101  ATGAATGATCACCTGGTGACTCAGAGTGAACAGACTGAGGGTAGGGCTAAGCTAAGCAGCACGGAAGGCTACGCCTTAGCCCAACAATACCATCATCTGT
      M  N  D  H  L  V  T  Q  S  E  Q  T  E  G  R  A  K  L  S  S  T  E  G  Y  A  L  A  Q  Q  Y  H  H  L  Y

5201  ACGGTCACCCCTCAGAGGAAAGCCTTAGAAAGGTGTTGACCAAACGATTTGTGTGGGAAGATATGGGACAACATTGTAAGGAAATAACTAACACCTGTCT
        G  H  P  S  E  E  S  L  R  K  V  L  T  K  R  F  V  W  E  D  M  G  Q  H  C  K  E  I  T  N  T  C  L
                         ↓SA       ORF1 exon1  ↱                                      ORF2 exon1   ↱
5301  AACCTGTGCAAAATATAAAGTTCTCAGAGCAGGACCACCAATGGGCGTAGGACGATCGGCTGAAGGGCCCTGCCAAAAACTACAGGTAGACCATGTGGGA
       T  C  A  K  Y  K  V  L  R  A  G  P  P  M  G  V  G  R  S  A  E  G  P  C  Q  K  L  Q  V  D  H  V  G
                                                                                                  M  W  D
                                  ↓SD
5401  CCTTTGGGACCTGGTACCCATGGATACAGGTACCTCACAACCATGGTTGATGTGTACACAGGTTGGTTCTGGGCCAAACCCTGTAGAGGTCCAACTACGG
      P  L  G  P  G  T  H  G  Y  R  Y  L  T  T  M  V  D  V  Y  T  G  W  F  W  A  K  P  C  R  G  P  T  T  G
       L  W  D  L  V  P  M  D  T  E

5501  GTGCCACTATAGCGGCCTTAGAGGAACACATCAGTATTTGGGGGGTACCCTATTCTATACAATCTGATAATGGCACCGCGTTCACTAGTAAAGCCATGCA
        A  T  I  A  A  L  E  E  H  I  S  I  W  G  V  P  Y  S  I  Q  S  D  N  G  T  A  F  T  S  K  A  M  Q

5601  AGAATGGGCCAATACGTATGGCATAGAATGGAAGGTGGGGGCTATATACCATCCTCAATCACAGGGAAAGGTAGAAAGAAAACATCGACTGCTCAAAGAC
       E  W  A  N  T  Y  G  I  E  W  K  V  G  A  I  Y  H  P  Q  S  Q  G  K  V  E  R  K  H  R  L  L  K  D

5701  AGACTTAAAAGGGCCACACATGAAGGAAAGAACTGGGTTCAAGCACTACCCTCCATACTACTATTTATTAACTCTATGCATCCACGAGATCAGTTTTCTG
      R  L  K  R  A  T  H  E  G  K  N  W  V  Q  A  L  P  S  I  L  L  F  I  N  S  M  H  P  R  D  Q  F  S  A

5801  CCTATGAACTAATGACAGGGAGAGTACCACACCTAGGTGGGTACCACCCACACCCCCTAGAGACAGCCAAAGAAGAAGAAGTACGGATATTCTTAAGAGC
        Y  E  L  M  T  G  R  V  P  H  L  G  G  Y  H  P  H  P  L  E  T  A  K  E  E  E  V  R  I  F  L  R  A

5901  AACACATGATTGCATGCAAAATAAGAAGTGGGAAGAACAGTTAGAAAAGGTCACTAAAGCAGAAGCCCACTCACACTGGACTCAGAGGAGCCCTAATTTA
       T  H  D  C  M  Q  N  K  K  W  E  E  Q  L  E  K  V  T  K  A  E  A  H  S  H  W  T  Q  R  S  P  N  L
```

FIG. 2—*Continued.*

```
6001  GAACCTGGCTGCATTGTGTTAGTAAGAAAGTTCACCGGGGATGCCTTCTCACCTAAATGGGAAGGACCATATGTGATAACAGAAACTACCAAATATGCTG
      E  P  G  C  I  V  L  V  R  K  F  T  G  D  A  F  S  P  K  W  E  G  P  Y  V  I  T  E  T  T  K  Y  A  A
                                                                                        ↓SA
6101  CTAAAGTACAGGCCATGTCTGACAAAGTAACACAACATTCTGGGTGGATACATAGGACTCACCTTGTGTTGTTTCCTTCACAGAACAAGCGTTGGGCGGA
        K  V  Q  A  M  S  D  K  V  T  Q  H  S  G  W  I  H  R  T  H  L  V  L  F  P  S  Q  N  K  R  W  A  D
                                                                         Env, ORF1 exon2→ T  S  V  G  R  I
                                                                               ORF2 exon2→ Q  A  L  G  G
                                              ↓SD
6201  TCCTGGAAATCCCGGAAACCAACCAGACGAGGACTGTACAGGTAAGAAAGGGACAGTTAGTACAGCTGACATGTCCCCAACTACCTCCACCACAAGGGAC
       P  G  N  P  G  N  Q  P  D  E  D  C  T  G  K  K  G  T  V  S  T  A  D  M  S  P  T  T  S  T  T  R  D
        L  E  I  P  E  T  N  Q  T  R  T  V  Q  V  R  K  G  Q  L  V  Q  L  T  C  P  Q  L  P  P  P  Q  G  T
      S  W  K  S  R  K  P  T  R  R  G  L  Y  R

6301  CGGGGTATTAATATGGGGACGGAACAAACGCACAGGAGGAGGAGCCCTAGATTTCAACGGGGTTCTGACAGTCCCAGTGGGGGACAATGAGAACACCTAT
      R  G  I  N  M  G  T  E  Q  T  H  R  R  R  S  P  R  F  Q  R  G  S  D  S  P  S  G  G  Q  *
       G  V  L  I  W  G  R  N  K  R  T  G  G  G  A  L  D  F  N  G  V  L  T  V  P  V  G  D  N  E  N  T  Y

6401  CAGTGCATGTGGTGCCAAAACACTACCAGCAAAAACGCACCAAGGCAAAAACGTAGCCTAAGGAACCAACCCACAGAATGGCATCTCCATATGTGTGGAC
      Q  C  M  W  C  Q  N  T  T  S  K  N  A  P  R  Q  K  R  S  L  R  N  Q  P  T  E  W  H  L  H  M  C  G  P

6501  CACCAGGTGATTACATATGCATGTGGACCAATAAGAAACCAGTGTGTACTACCTATCATGAAGGACAGGATACATACTCCTTAGGGACGCATAGGAAGGT
        P  G  D  Y  I  C  M  W  T  N  K  K  P  V  C  T  T  Y  H  E  G  Q  D  T  Y  S  L  G  T  H  R  K  V

6601  GCTCCCCAAAGTAACTGAAGCCTGTGCGGTTGGACAACCTCCTCAGATACCCGGAACCTATGTAGCCAGTAGTAAAGGATGGACTATGTTTAATAAATTT
       L  P  K  V  T  E  A  C  A  V  G  Q  P  P  Q  I  P  G  T  Y  V  A  S  S  K  G  W  T  M  F  N  K  F

6701  GAAGTCCATTCCTACCCCGCTAATGTCACCCAGATTAAAACAAACAGAACACTGCATGACGTAACATTATGGTGGTGTCATGACAACTCCATATGGAGAT
      E  V  H  S  Y  P  A  N  V  T  Q  I  K  T  N  R  T  L  H  D  V  T  L  W  W  C  H  D  N  S  I  W  R  C

6801  GTACACAGATGGGGTTTATACACCCGCATCAAGGGAGAAGAATACAGCTAGGCGATGGGACCAGATTTAGGGATGGGTTGTATGTCATAGTATCCAATCA
        T  Q  M  G  F  I  H  P  H  Q  G  R  R  I  Q  L  G  D  G  T  R  F  R  D  G  L  Y  V  I  V  S  N  H

6901  TGGGGACCATCACACAGTACAGCATTATATGTTAGGCTCAGGATACACTGTGCCAGTCTCCACCGCCACCCGTGTCCAAATGCAGAAAATAGGACCAGGG
       G  D  H  H  T  V  Q  H  Y  M  L  G  S  G  Y  T  V  P  V  S  T  A  T  R  V  Q  M  Q  K  I  G  P  G

7001  GAATGGAAAATAGCAACTAGCATGGTAGGGCTATGCCTGGACGAATGGGAAATAGAGTGCACCGGCTTCTGTAGTGGTCCTCCTCCTTGCAGTTTAAGCA
      E  W  K  I  A  T  S  M  V  G  L  C  L  D  E  W  E  I  E  C  T  G  F  C  S  G  P  P  P  C  S  L  S  I

7101  TAACTCAACAGCAGGATACAGTGGGAGGATCTTATGACTCATGGAACGGGTGTTTTGTCAAATCAATACACACACCAGTTATGGCTCTGAACCTTTGGTG
        T  Q  Q  Q  D  T  V  G  G  S  Y  D  S  W  N  G  C  F  V  K  S  I  H  T  P  V  M  A  L  N  L  W  W

7201  GAGAAGGAGCTGTAAAGGGCTACCTGAAGCGACTGGCATGGTAAAAATATACTATCCTGACCAATTTGAAATAGCACCATGGATGAGACCTCAGCCCAGA
       R  R  S  C  K  G  L  P  E  A  T  G  M  V  K  I  Y  Y  P  D  Q  F  E  I  A  P  W  M  R  P  Q  P  R
                                                              ↓
7301  CAACCTAAACTGATATTACCTTTTACTGTAGCACCAAAATATAGACGACAACGAAGGGGACTTAACCCCTCAACCACACCTGATTATTACACTAATGAAG
      Q  P  K  L  I  L  P  F  T  V  A  P  K  Y  R  R  Q  R  R  G  L  N  P  S  T  T  P  D  Y  Y  T  N  E  D

7401  ATTATAGTGGATCAGGGGGGTGGGAAATAAATGACGAGTGGGAATATATACCACCAACAGTGAAACCAACCACGCCTAGTGTTGAATTCATACAGAAAGT
        Y  S  G  S  G  G  W  E  I  N  D  E  W  E  Y  I  P  P  T  V  K  P  T  T  P  S  V  E  F  I  Q  K  V
                                                         ↓
7501  GACTACACCTCGCCAGGATAAATTGACCACTGTCCTGAGCCGGAATAAAAGGGGAGTGAACATTGCCTCTAGTGGCAACAGCTGGAAGGCAGAAATAGAT
       T  T  P  R  Q  D  K  L  T  T  V  L  S  R  N  K  R  G  V  N  I  A  S  S  G  N  S  W  K  A  E  I  D

7601  AAGATAAGGAAACAAAAATGGCAGAAATGCTATTTCTCAGGAAAACTAAGAATAAAAGGAACAGACTATGAGGAAATAGATACATGCCCAAAACCATTAA
      K  I  R  K  Q  K  W  Q  K  C  Y  F  S  G  K  L  R  I  K  G  T  D  Y  E  E  I  D  T  C  P  K  P  L  I

7701  TAGGACCACTGTCAGGGTTCATACCCACAGGAGTGACCAAAACCCTAAAAACAGGGGTAACATGGACCACCGCTGTTGTAAAAATAGATCTGCAGCAGTG
        G  P  L  S  G  F  I  P  T  G  V  T  K  T  L  K  T  G  V  T  W  T  T  A  V  V  K  I  D  L  Q  Q  W

7801  GGTGGATATCCTTAATAGCACCTGTAAAGATACACTTATAGGGAAACACTGGATTAAGGTAATTCAGCGGCTTTTACGGGAATATCAGAAGACAGGGGTT
       V  D  I  L  N  S  T  C  K  D  T  L  I  G  K  H  W  I  K  V  I  Q  R  L  L  R  E  Y  Q  K  T  G  V
                                                                                                    ↓
7901  ACCTTTAATCTCCCACAGGTTCAATCATTACCCAATTGGGAAACCAAAAACAAAGATAATCCTGGACATCACATTCCCAAAAGCCGGCGGAAGAGAATTA
      T  F  N  L  P  Q  V  Q  S  L  P  N  W  E  T  K  N  K  D  N  P  G  H  H  I  P  K  S  R  R  K  R  I  R
          ↓
8001  GGCGAGGATTAGGTGAAGCTTTAGGATTAGGTAACTTTGCTGATAACAGATGGAAGGATTTACAGATAGCAGGATTAGGGGTAGAACAGCAGAAATTAAT
        R  G  L  G  E  A  L  G  L  G  N  F  A  D  N  R  W  K  D  L  Q  I  A  G  L  G  V  E  Q  Q  K  L  M

8101  GGGATTAACTAGAGAAGCCACGTTTGAGGCATGGAATGCCCTGAAGGGTATTTCTAACGAATTAATTAAATGGGAAGAAGATATGGTAGCCACTCTCAGA
       G  L  T  R  E  A  T  F  E  A  W  N  A  L  K  G  I  S  N  E  L  I  K  W  E  E  D  M  V  A  T  L  R

8201  CAGCTTCTATTACAAATTAAAGGCACTAATACTACCCTCTGTAGTGCGATGGGACCGTTAATGGCTACCAATATACAACAAATAATGTTCGCATTACAAC
      Q  L  L  L  Q  I  K  G  T  N  T  T  L  C  S  A  M  G  P  L  M  A  T  N  I  Q  Q  I  M  F  A  L  Q  H

8301  ATGGTAATCTACCTGAAATGTCTTACTCTAACCCTGTGTTGAAGGAAATAGCTAAACAATATAATGGACAAATGTTAGGTGTACCAGTAGAAACTACAGG
        G  N  L  P  E  M  S  Y  S  N  P  V  L  K  E  I  A  K  Q  Y  N  G  Q  M  L  G  V  P  V  E  T  T  G

8401  AAATAACTTGGGAATAATGTTATCATTACCTACCGGGGGAGAGAACATAGGAAGAGCAGTAGCTGTATACGATATGGGAGTAAGACATAACCGTACGCTA
       N  N  L  G  I  M  L  S  L  P  T  G  G  E  N  I  G  R  A  V  A  V  Y  D  M  G  V  R  H  N  R  T  L

8501  TACCTCGATCCCAACGCTAGATGGATCCACAACCACACGGAAAAGAGTAATCCTAAAGGCTGGGTAACTATAGTAGACTTATCTAAGTGTGTCGAAACCA
      Y  L  D  P  N  A  R  W  I  H  N  H  T  E  K  S  N  P  K  G  W  V  T  I  V  D  L  S  K  C  V  E  T  T

8601  CAGGAACTATTTATTGTAATGAACACGGATTTAGAGACAGGAAATTCACTAAAGGACCTTCAGAATTAGTACGGCATTTAGCTGGTAATACATGGTGTTT
        G  T  I  Y  C  N  E  H  G  F  R  D  R  K  F  T  K  G  P  S  E  L  V  R  H  L  A  G  N  T  W  C  L

8701  AAATTCAGGAACATGGTCATCACTGAAAAATGAGACACTGTATGTAAGTGGACGAAATTGCTCCTTCTCTCTCACCAGTAGGCGGCGACCTGTGTGTTTT
       N  S  G  T  W  S  S  L  K  N  E  T  L  Y  V  S  G  R  N  C  S  F  S  L  T  S  R  R  R  P  V  C  F

8801  CACCTAAATAGCACCGCCCAATGGCGAGGACACGTTTTGCCTTTTGTAGGTAACTCCCAGGAAGCACCCAACACTGAGATATGGGAGGACTCATAGAAG
      H  L  N  S  T  A  Q  W  R  G  H  V  L  P  F  V  G  N  S  Q  E  A  P  N  T  E  I  W  E  G  L  I  E  E
```

FIG. 2—*Continued.*

```
 8901  AAGCAATTCGAGAACATAATAAGGTACAAGATATTCTGACCAAATTGGAACAGCAACACCAAAATTGGAAACAAAACACAGACAACGCCCTACAAAATAT
         A  I  R  E  H  N  K  V  Q  D  I  L  T  K  L  E  Q  Q  H  Q  N  W  K  Q  N  T  D  N  A  L  Q  N  M
                                                                                      TMA
 9001  GAAGGATGCAATTGACTCCATGGATAATAACATGTTAACATTTAGATACGAGTACACCCAGTACGGGTTGTTTATTGTATGTCTGCTTGCTTTTCTGTTT
        K  D  A  I  D  S  M  D  N  N  M  L  T  F  R  Y  E  Y  T  Q  Y  G  L  F  I  V  C  L  L  A  F  L  F

 9101  GCTGTAATCTTCGGATGGTTATGTGGGGTTACCGTGAGACTAAGGGAGGTCTTTACAATCTTATCTGTAAAAATACATGCTTTAAAGTCACAAGCACATC
       A  V  I  F  G  W  L  C  G  V  T  V  R  L  R  E  V  F  T  I  L  S  V  K  I  H  A  L  K  S  Q  A  H  Q

 9201  AGTTGGCTATGTTGCGAGGTCTACGTGACCCAGAGACAGGGGAACAAGATCGACAGGCACCCGCCTACAGGGAGCCACCCACATGTCAGGAATGGGCGAG
         L  A  M  L  R  G  L  R  D  P  E  T  G  E  Q  D  R  Q  A  P  A  Y  R  E  P  P  T  C  Q  E  W  A  R
                                      ORF2 exon3 →↓SA
 9301  ACGCAGGGGGGGAAGACCACCTATTGTCACTTTTCTGATTGACAGAGAGACTGGTGAACGACACGACGGACAGATCTTCCAACCTATAAGAAACAGATCT
        R  R  G  G  R  P  P  I  V  T  F  L  I  D  R  E  T  G  E  R  H  D  G  Q  I  F  Q  P  I  R  N  R  S
                                      ORF1 exon3 → R  D  W  *
                                                                                               3' ORF ↱
 9401  AACCAGGTACATCGACCACAACCGCCCAGACCTACGGCCCCTAACCCAGATAACCAACGACCTATTAGGGAACCACGTCCCGGGGAGCCAGAACATGGCG
       N  Q  V  H  R  P  Q  P  P  R  P  T  A  P  N  P  D  N  Q  R  P  I  R  E  P  R  P  G  E  P  E  H  G  D
                                                                                                     M  A

 9501  ACTTCCTTCAAGGAGCCTCCTGGATGTGGCAGTGATGAAGATAAGGACAATGACACACCTTCACGAGTACCCAAACCAGAAGAAGTGATAAGTTTATATA
         F  L  Q  G  A  S  W  M  W  Q  *
       T  S  F  K  E  P  P  G  C  G  S  D  E  D  K  D  N  D  T  P  S  R  V  P  K  P  E  E  V  I  S  L  Y  N

 9601  ATCCTGAATTATTTGCTTGGGGGGGAACACTGGGAGATTATGTATGGTGTAAAACTACTACTATTCGATGGGGACCTATTGTAATGAAGTATCGTGTGTG
         P  E  L  F  A  W  G  G  T  L  G  D  Y  V  W  C  K  T  T  T  I  R  W  G  P  I  V  M  K  Y  R  V  C

 9701  TGGGCACCTCGAATCAGATTATACCACGAGTCAGTATTACAGAGAATCCGTTAATCAAACGTGGATAGTAGGTCATAGTGGAAGAGAAGTGACACATACC
        G  H  L  E  S  D  Y  T  T  S  Q  Y  Y  R  E  S  V  N  Q  T  W  I  V  G  H  S  G  R  E  V  T  H  T

 9801  TGTGATCATTACTGTCGGTGTTTGAATCCACCCAGGTTCTGCTGGTGTAGAAAACACAAAAGGTGGCATCCCTTCATATCGCCCCTAGGTCTAGTTCACT
       C  D  H  Y  C  R  C  L  N  P  P  R  F  C  W  C  R  K  H  K  R  W  H  P  F  I  S  P  L  G  L  V  H  W

 9901  GGAGTTGGATGGAGAAGTTCGTGCAACCGGTGATAAGGCTCGAGGGTTGGAAAGAACTGAAGGAGGAATTGCTGACACACCCTGACAAAATAAAAATCTG
         S  W  M  E  K  F  V  Q  P  V  I  R  L  E  G  W  K  E  L  K  E  E  L  L  T  H  P  D  K  I  K  I  W

10001  GGTGGGACGTGTACACGGCCGGATCAGTCCGGATCTGTCTGAGCATGAGTTAATGGTTAACATGTGTGCTCAGGGTACACTTAAACCGACACTGTTACCA
        V  G  R  V  H  G  R  I  S  P  D  L  S  E  H  E  L  M  V  N  M  C  A  Q  G  T  L  K  P  T  L  L  P

10101  GCCCTCTGGTGAAACTGAAAATGTGTCATTTCTGGTGAAACGAAACCTATGACTAATTCACAAATAGAATCCAGCACCGCCTTTGTCTAGGTGTGTATTC
       A  L  W  *
          PPT   →              U3                    E BOX OCT-1/2       ISGF-1/2
10201  TAGGGGGGCTTGTGACACCACTATACTTCAACTCTGGTCTCCATCTCATATGCAAAGCATTATATGAAACTGAAACTAAGCATGACTTGCTATTCTCTAA

                                                      CORE             AP-1  LVb/c                    TEF-1
10301  TGTAACGAGAGATGCCTGTTTGTGCACCATTGTAAAGGTACAAACTGTGGTCATAAACATGATGACTCATATCCTGCTTAGTAGACTATTGTTAGCATGC

                           LVa                  E BOX                                       Myogenin/E BOX
10401  ACATACGCTAGTCCCTGGGAACAGGACTTATTCCTATGATACATATGCTCTTAATGTACTTGTTTACTGTTTATACGCACCCAAGTAACAACTGTATTTA

                      AP-1    CCAAT                                                                    GC
10501  ATTATAACCTATGTGAATCATAAGGCCAATCACGCTAGGAGGCGGCCTCTAGTACTTTTCCATTGTATGAACTATTGTTATTAGCCCCTCCCTGAAGGAG

           E BOX       TATA                    ←→                   R     polyA              ←
10601  GAGCCACTTGCCTATATATTGTAGAGCACGGCTGTGCTCATTGTCTTTTGCTGCACCCAACTTGAAGAATAAAGATTATTGCATCTGA
```

FIG. 2—*Continued.*

composition of high A content (32.7%) and low C content (20.8%). Analyses of the codon usage in the *gag*, *pol*, and *env* coding regions revealed a bias towards A or T in the codon third position, with only 41.2% of codons ending in G or C.

LTR. The LTR has a size of 518 bp and comprises U3, R, and U5 regions with sizes of 434, 46, and 38 bp, respectively. The U3 region of retroviruses is preceded in the genome by a polypurine tract which serves as the priming site for plus-strand DNA synthesis during replication. The presence of several polypurine stretches upstream of the putative SnRV U3 region made the prediction of the start of U3 difficult by sequence analysis alone. The location of this site was therefore determined by amplification of the U3 flanking cellular region by inverse PCR, which defined the proviral start of U3 at position 10211. The polypurine tract is therefore located at positions 10202 to 10208 (AGGGGGG). Within U3, a putative basal promoter region comprises CCAAT, GC, and TATA boxes located at positions 10521 to 10532, 10593 to 10606, and 10612 to 10626, respectively (7). In addition, several possible sites for transcription factors were identified upstream of this region (Fig. 2).

The cap site, corresponding to the U3/R junction, and the polyadenylation site, corresponding to the R/U5 junction, were determined by 5′ and 3′ rapid amplification of cDNA ends (RACE) methods, respectively. The U3/R junction is located at position 10643, and a consensus polyadenylation signal (AATAAA) is located within R at positions 26 to 31 (repeated at positions 10668 to 10673), 15 nucleotides upstream of the R/U5 junction at position 46. Comparable to other retroviruses, the 3′ end of U5 is defined by the presence of a tRNA primer binding site at positions 85 to 102. The primer binding site of SnRV is a perfect complement to the 3′-terminal 18 nucleotides of murine Arg^{1,2} tRNA (23). The LTR is delineated by the short inverse repeats 5′-TGT and ACA-3′ at positions 10211 to 10213 and 81 to 83, respectively. The inverse repeats, in common with all retroviruses, contain the conserved 5′-TG and CA-3′ dinucleotides found at the termini of the provirus. The spacing of the 5′ TG 2 nucleotides from the polypurine tract and the 3′ CA 1 nucleotide from the primer binding site is an organization similar to that found in the human immunodeficiency virus type 1 genome. SnRV replication may therefore also be similar to that of human immuno-

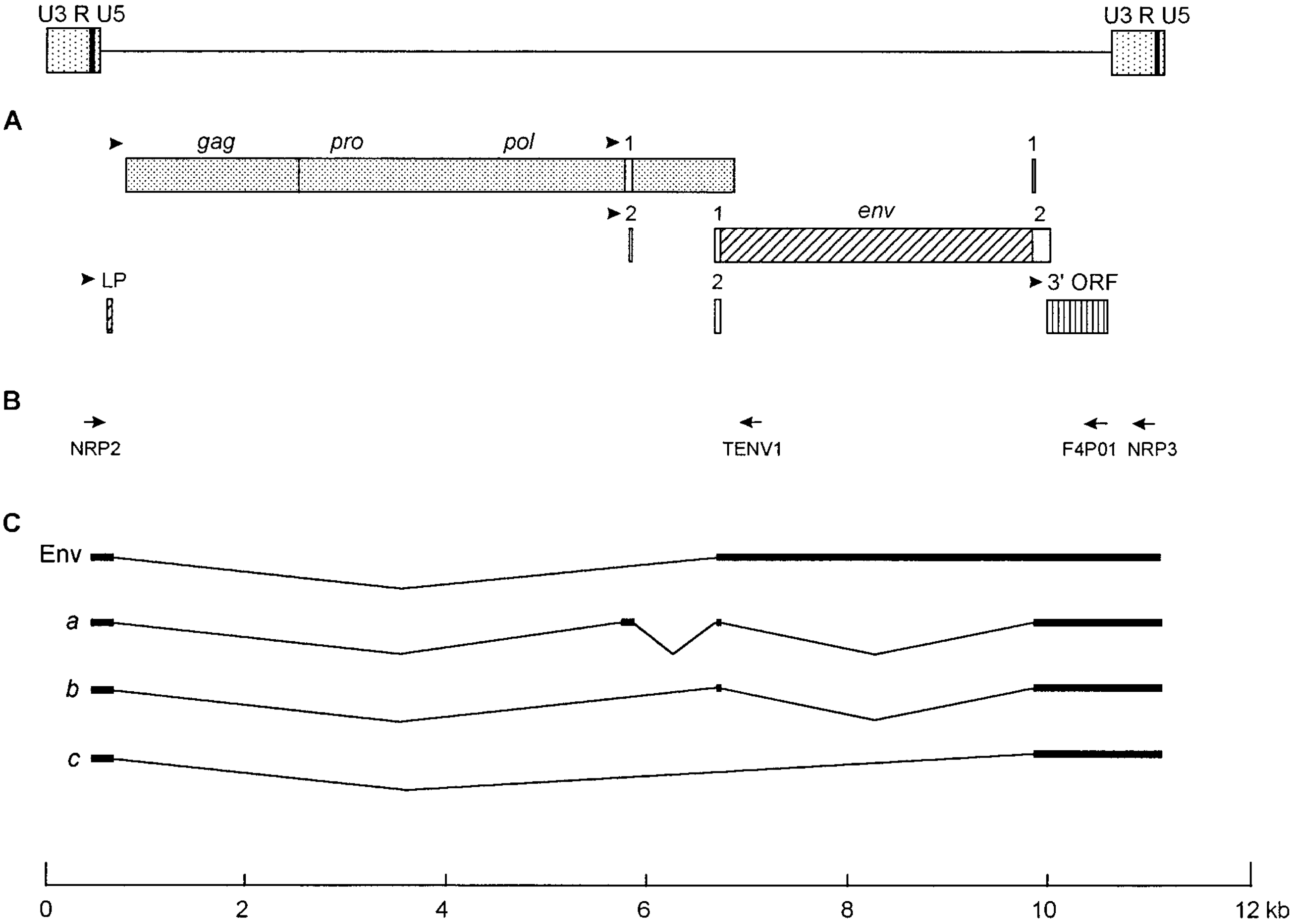

FIG. 3. SnRV genome organization and splicing pattern. (A) Predicted coding regions are shown as open or shaded boxes, with potential start codons indicated by arrowheads. Numbered boxes represent the positions of coding regions for the proteins ORF 1 and ORF 2. (B) Location of primers used for RT-PCR amplification of subgenomic transcripts. (C) Structure of spliced transcripts.

deficiency virus type 1, where the RNase H specifically cleaves 1 nucleotide from the tRNA primer end, leaving the unintegrated linear DNA with a 5′-terminal ribonucleotide (19, 48).

Leader and *gag*. In all retroviruses there is a leader region between the 5′ LTR and the start of the *gag* open reading frame (ORF) within which reside signals important for the replication, processing, and packaging of the viral RNA. For most retroviruses the *gag* initiation codon is the first codon downstream from the cap site in suitable context for the initiation of translation (31). In the case of SnRV, the first such start codon is located at position 180, potentially encoding a hydrophobic leader peptide (LP) of 14 amino acids (MKLV LLFSLSVLLG). The next start codon is located 105 nucleotides downstream of the LP at position 337. This initiation codon is in a strong sequence context for translation initiation (29) and is the putative start codon of *gag*, potentially coding for a 618-amino-acid (69-kDa) polyprotein.

The N terminus of the Gag protein is not predicted to be myristylated, but a polybasic region is located from amino acids 15 to 35 that may form a region of interaction between the Gag matrix (MA) protein and the plasma membrane (33, 69). Interestingly, the polybasic region contains a consensus bipartite nuclear localization sequence (11, 49), consisting of two basic residues spaced 10 amino acids from a further cluster of four basic residues (KRDKSKEKDETKKKKIK). Database searches with the N-terminal 300 amino acids of Gag did not reveal any similarities to the matrix (MA) proteins of other retroviruses but predominantly indicated similarities to the rod-like domain of myosins, tropomyosins, lamins, and other intermediate filament-type proteins. Analysis of the amino acid sequence within this region revealed a high content (58%) of predicted α-helical secondary structures. Sequence analysis also located a hydrophobic heptad repeat pattern from residues 154 to 238, predicted to form a two-stranded coiled-coil structure typical of the rod domain of myosin-like proteins (32, 43).

A low level of similarity (15 to 20% amino acid identity) to the capsid (CA) region of the mammalian type C retroviruses was found approximately from amino acids 300 to 500, and this region contains the major homology region (amino acids 426 to 445) found in all retroviruses except the spumaviruses (44, 64).

The C-terminal region of Gag contains a single Cys-His box (amino acids 537 to 550) characteristic of the mammalian type C retrovirus nucleocapsid (NC) domains. Database searches revealed this region of Gag to have sequence similarity to that of the lentiviruses, however, producing the highest amino acid identity scores to simian immunodeficiency virus (34.7%) and human immunodeficiency virus type 2 (27.8%). The region of similarity extended from the second Cys-His box of the lentivirus NC domain into the C-terminal protein region. This included an ASAPP sequence (amino acids 576 to 580) that is a close approximation of the conserved PS/TAPP motif found in the lentiviral C-terminal proteins (39).

pol. The *pol* ORF is predicted to start at position 2194, encoding a large polyprotein of 1,398 amino acids (157 kDa), and to be expressed by translational suppression of the *gag* TGA termination codon. A common feature of retroviruses expressing *pol* this way, and for other retroviruses that express *pol* via ribosomal frameshifting of the *gag* ORF, is the presence of a potential RNA secondary structure motif downstream of the site of suppression or frameshifting (5, 16, 65, 66). The SnRV sequence shows a high G and C content in this region, suggesting that a potential secondary structure is likely. However, analysis of this region did not reveal any significant similarities of sequence or potential secondary structure typical of the models proposed for other retroviruses.

Comparisons of the deduced Pol protein sequence to those of other retroviruses identified conserved regions typical of retroviral protease, RT, RNase H, and integrase (IN) enzyme domains (28, 61). The protease domain is located approximately from amino acids 1 to 110, analogous to the location in the mammalian type C viruses, the lentiviruses, and simian foamy virus. The protease sequence is most similar to those of the primate lentiviruses, with simian immunodeficiency virus and human immunodeficiency virus sequences giving the highest amino acid identity scores of 32.3 and 28.1%, respectively. The RT/RNase H domain is located approximately from amino acids 110 to 800 and shows closest identity to the mammalian type C viruses (34.4%), WDSV (33.5%), and the spumaviruses (28.2%).

The IN domain is located approximately from residues 980 to 1330 and, as with the RT/RNase H domain, is closest in identity to that of the mammalian type C retroviruses (30.1%), WDSV (29.4%), and the spumaviruses (25.9%). Thus, there exists in Pol a region of approximately 180 amino acids separating the RT/RNase H and IN domains. The nonprimate lentiviruses exhibit a similarly located domain which studies have shown to encode a dUTPase (14, 35). The SnRV protein sequence did not, however, show any significant sequence similarity to a dUTPase or to any other viral or nonviral protein in the database.

env. Unique to SnRV, *env* is predicted to be expressed by fusion of the LP to the downstream *env* region (Fig. 3). All retroviruses express *env* as a singly spliced transcript with the splice donor site typically located upstream of *gag* and the *env* initiation codon located downstream of the splice acceptor site. An exception to this is found in the avian leukosis and sarcoma retroviruses, where the splice donor site is located six codons into *gag* and expression of *env* fuses these residues to the downstream region (54). Analysis of SnRV transcripts was performed by RT-PCR on SSN-1 mRNA with primers based on R/U5 (primer NRP2) and U3 (primers F4P01, NRP3) sequences (Fig. 3B). Sequence analysis of the cDNA clones revealed a common splice donor site at position 221, 1 nucleotide upstream of the LP stop codon. A potential *env* splice acceptor site was located at position 6184, and RT-PCR with an *env*-specific downstream primer (TENV1) produced a single product which contained this splice junction site. The first ATG codon in suitable context for translation initiation 3′ of the acceptor site is located more than 500 nucleotides downstream, however, and the protein sequence in this region does not contain a potential N-terminal signal peptide, necessary for transfer of the Env precursor into the endoplasmic reticulum (ER). The initiation codon for *env* is therefore predicted to be the LP initiation codon, the singly spliced *env* mRNA fusing the LP into frame with the downstream reading frame and encoding a protein of 1,130 amino acids (128 kDa). Moreover, the LP provides the hydrophobic core of the signal peptide for Env, with the peptidase cleavage site predicted to be located four amino acids downstream of the fusion site (60). The size of the predicted *env* transcript is in accordance with the 4.7-kb band observed in Northern blots (Fig. 4).

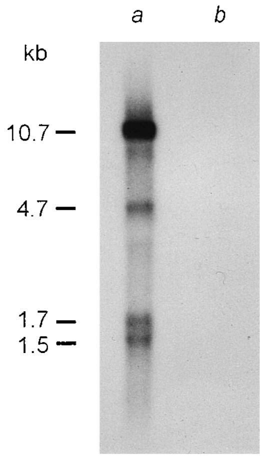

FIG. 4. Northern blot analysis of poly(A) RNA isolated from SSN-1 (lane *a*) and SSN-2 (lane *b*) cell cultures (2.5 μg each) with an SnRV LTR probe. SSN-2 is a striped snakehead cell line which shows no evidence of a retrovirus infection (reference 18 and unpublished results). Analysis of SSN-1 mRNA shows putative retroviral genomic and *env* transcripts with sizes of 10.7 and 4.7 kb, respectively, plus two smaller bands in the 1.5- to 1.7-kb range.

There are four potential sites for cleavage of the Env precursor into the surface (SU) and transmembrane (TM) proteins at amino acid positions 405, 470, 618, and 621 (consensus sequence R-X-R/K-R). The Env amino acid sequence did not show any significant similarities to those of other retroviruses in database searches, but the C-terminal half of this protein could be compared with models of TM based on predicted and known protein structures (20, 41). The hydrophobic fusion peptide region, found in other retroviral TM proteins downstream of the SU/TM cleavage site, could not be easily identified in the SnRV protein. A region containing potential α-helical secondary structures is located from amino acids 646 to 688, corresponding to the predicted fibrous core of retroviral TM proteins. A second α-helical region with a high probability of forming a coiled coil is located from amino acids 910 to 964. The hydrophobic membrane-spanning region of the TM protein is likely to span amino acids 974 to 994, leaving a 136-amino-acid cytoplasmic domain (CD) which is highly hydrophilic and proline rich. Adjacent and internal to the putative membrane anchor is a leucine zipper motif from amino acids 993 to 1024. As with all retroviral Env proteins there are multiple predicted N-glycosylation sites, 11 external and 1 internal.

ORFs. Northern blot analyses of SSN-1 mRNA using an SnRV LTR probe produced bands corresponding to genomic RNA, singly spliced *env* plus two smaller bands in the 1.5- to 1.7-kb range (Fig. 4). Analysis of RT-PCR clones from SSN-1 mRNA subsequently identified three cDNA species which were predicted to be formed by a complex splicing process (Fig. 3C) (55). Transcript *a* contains four exons with two potential initiator codons located in the second exon at positions 5341 and 5393. Translation from these positions would produce two proteins (ORF 1 and ORF 2) with sizes of 52 (5.7 kDa) and 94 (11 kDa) amino acids, respectively. Transcript *b* contains three exons and is a potential mRNA for the 3′ ORF, which has an ATG codon in a strong initiation context at position 9495, downstream of the splice acceptor site at position 9346. Translation of the 3′ ORF would produce a protein with a size of 205 amino acids (24 kDa). Transcript *c* may additionally encode the 3′ ORF but also brings the LP coding region into frame with the Env CD to potentially code for a 76-amino-acid (8.8-kDa) LP/CD fusion protein.

Database searches of the amino acid sequences of the ORFs showed no significant homology to any viral or nonviral proteins. The ORF 1 protein has a potential N-terminal myristylation site, and ORF 2 contains a basic region from amino acids 21 to 32. ORF 2 shares with Env the C-terminal 62 residues of the CD. Comparable to Env, fusion of the LP to the CD creates a potential signal peptide, with peptidase cleavage predicted at the fusion site (60). The 3′ ORF protein contains an N-terminal acidic region, and a cluster of cysteine residues is located at amino acids 103 to 118, followed immediately by a small basic region (RKHKR). An α-helix is predicted from amino acids 148 to 162 and displays an acidic helical face, comprising four glutamate residues occupying one side of the structure.

Phylogenetic analysis. In order to attempt to determine the relationship of SnRV to the known retrovirus groups and to WDSV, a phylogenetic tree was reconstructed on the basis of multiple alignments of Pol amino acid sequences. The relatively conserved RT/RNH domain of SnRV (amino acids 186 to 743) was aligned with that of WDSV plus representative retroviruses from each genus. The neighbor-joining method of Saitou and Nei (50) was used to construct a tree using protein distances calculated from the PAM 250 matrix (10). Bootstrap analysis was performed to test the variation around the tree, with the initial data set resampled 2,000 times. The resulting tree (Fig. 5) is in accordance with previously published analyses (12, 34, 67) and illustrates the phylogenetic relationship of the seven characterized retrovirus genera. The two fish retroviruses appear grouped with the mammalian type C clade but are clearly divergent both from these viruses and from each other. While the tree illustrates SnRV diverging from the WDSV branch, the low bootstrap value for this fork indicates that the alignment cannot unambiguously resolve the branching order of SnRV, WDSV, and the mammalian type C retroviruses. The same alignment was also analyzed using the maximum likelihood and parsimony tree reconstruction methods. Both of these methods produced topologies comparable to that of the neighbor-joining tree (data not shown), and the differences in branching order that were observed corresponded to the low bootstrap values of the nodes concerned.

DISCUSSION

The genomic organization, tRNA primer, sequence homology, and transcriptional profile restrict the placement of SnRV in any of the known retrovirus groups. The size of the genome is approximately between those of the lentiviruses and spumaviruses, and analyses of the base composition and codon usage also reveal similarities to these groups (6, 38, 59). The codon usage is distinct from that found in teleosts, where 65% of codons in analyzed gene sequences end in G or C (17). The organization of the *gag*, *pol*, and *env* coding regions of SnRV is comparable to that of the mammalian type C retroviruses but differs in the presence of a 3′ ORF between *env* and the LTR. The Arg tRNA primer binding site distinguishes SnRV from all known retrovirus groups, which tend to share common types of tRNA primer between related viruses (34).

A unique feature of SnRV is the unusual location of the *env* start codon and signal peptide in the leader region. The location of the LP upstream of *gag* and the common splice donor site suggests that a level of translational control exists for the expression of downstream ORFs. A similar situation is known to occur in the equine infectious anemia virus genome, where the first coding exon of *tat* is also located upstream of *gag* and the donor site (13, 40, 56). In this case the expression of downstream equine infectious anemia virus ORFs occurs by leaky scanning of ribosomes past the inefficient CTG *tat* start codon (8). The sequence context (AAC<u>ATG</u>A) of the SnRV LP start codon suggests that it is likely to be used efficiently for translation initiation (29), however, and we suspect that the size and location of the LP ORF would be more consistent with a translation reinitiation mechanism (30). An alternative possibility for the expression of downstream ORFs is a cap-independent internal ribosome entry mechanism, which has recently been found to occur in the expression of murine leukemia virus *gag* (3). It is of interest to note that in addition to Env, fusion of the LP to the CD also creates a potential signal peptide for this protein. The existence and location of the LP may therefore enable both Env and the CD to be independently expressed and directed to the ER by the same N-terminal signal peptide.

The Gag protein is distinguishable by the presence of a large region that is similar in sequence to the coiled-coil structures found in myosin-like proteins. Comparable to these proteins, this region of Gag is likely to function as a multimerization domain, allowing the formation of homodimeric or higher-order complexes. The coiled-coil structure may play a role in the transport and assembly of the Gag/Pol polyproteins during virus release from cells. It is not known whether this region is part of the MA or CA proteins or is an extra Gag protein located between these domains.

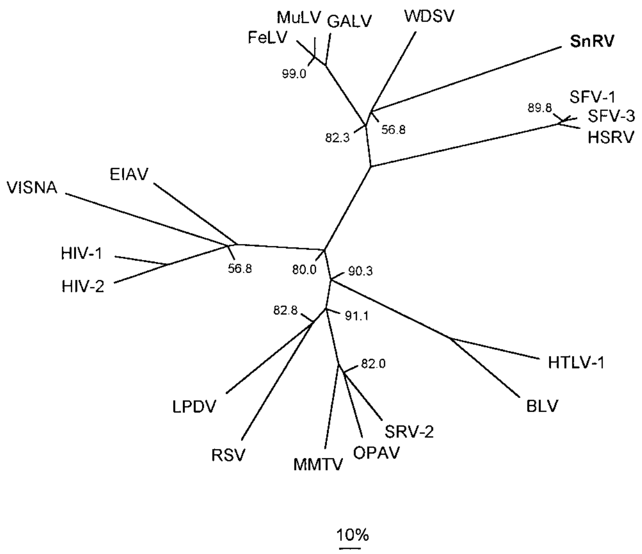

FIG. 5. Unrooted phylogenetic tree of representative retroviruses based on alignment of Pol RT/RNH amino acid sequences. The values given at the branch forks are the bootstrap percentages (for 2,000 samplings) with missing values all being 100%. The solid line below the diagram represents 10% divergence. The tree was constructed from an alignment of 19 sequences, each 615 amino acids in length, corresponding to amino acids 186 to 743 of SnRV Pol. SFV-1 and SFV-3, simian foamy virus types 1 and 3, respectively; HSRV, human spumaretrovirus; HTLV-1, human T-cell leukemia virus type 1; BLV, bovine leukemia virus; SRV-2, simian retrovirus type 2; OPAV, ovine pulmonary adenocarcinoma virus (Jaagsiekte sheep retrovirus); MMTV, mouse mammary tumor virus; RSV, Rous sarcoma virus; LPDV, lymphoproliferative disease virus; HIV-1 and HIV-2, human immunodeficiency virus types 1 and 2, respectively. VISNA, visna sheep lentivirus; EIAV, equine infectious anemia virus; FeLV, feline leukemia virus; MuLV, murine leukemia virus; GALV, gibbon ape leukemia virus.

The significance of the similarities of the Gag NC and Pol PR domains to those of the lentiviruses is not clear. The observed similarities within the relatively small NC and PR domains may be due to the comparable nucleotide compositions of these viruses, resulting in similar amino acid compositions of the viral proteins (2, 6). The lentivirus PS/TAPP motif region to which SnRV shows similarity to is an assembly domain required late in the virus particle budding process (22, 26, 42). This domain appears to be located in other retroviruses between MA and CA and is characterized by a PPPY/W motif (63), which SnRV lacks. SnRV may therefore have this domain at a location similar to that in the lentiviruses.

The Pol polyprotein is distinctive from that of other retroviruses in containing a region of approximately 180 amino acids between the RT/RNaseH and IN domains. This region, which bears no significant similarity to the similarly located dUTPases of the nonprimate lentiviruses, may represent a unique domain of unknown function.

The Env protein is unusual in the multiplicity of potential SU/TM cleavage sites, the lack of a distinct hydrophobic fusion peptide sequence, and the existence of a long CD which has previously been known to occur only in lentiviral genomes. The CD is proline rich, suggesting an extended secondary structure that is likely to be involved in protein binding (62). The role, if any, of the CD of lentiviral TM proteins in the virus life cycle is not fully understood (41). The identification of proteins expressed by lentiviruses that contain the CD separate from the TM protein has raised the possibility of an independent function for this region (1, 21, 51). Interestingly, the SnRV CD may also be potentially expressed independently of Env, both as the C-terminal region of ORF 2 and as an LP/CD fusion protein. In a remarkable comparison, in vivo studies of equine infectious leukemia virus expression have identified a transcript consisting of the leader *tat* exon spliced to the *env* CD (1). The resulting protein was found to be located within the ER and Golgi complex of infected cells. As noted previously, the SnRV CD may also be directed to the ER by the LP signal peptide, suggesting that a common role for these regions is possible.

The transcriptional profile of SnRV is complex and is similar to the splicing patterns observed in the lentiviruses, bovine leukemia virus–human T-cell leukemia virus group, and spumaviruses. Comparable to expression in these viruses, the expression of additional subgenomic mRNAs potentially allows production of accessory proteins. The expression of ORFs 1 and 2 and the 3′ ORF would depend on the translational regulation of the LP, as this exon is present in all viral transcripts. All complex retroviruses characterized to date encode a transcriptional transactivator which plays an important role in the regulation of virus expression. The N-terminal acidic region, cysteine cluster, and basic region of the 3′ ORF are features also found in the Tat proteins of lentiviruses (45). This protein may likewise function as a transcriptional transactivator, although it lacks any significant sequence similarities to the lentiviral Tat proteins.

Phylogenetic analyses using alignments of RT amino acid sequences indicated a distant relationship of SnRV to the mammalian type C retroviruses. However, the virus is clearly divergent from this group, and the bootstrap results imply that with the present data it is not possible to accurately place the SnRV branch in relationship to these viruses. SnRV also appears divergent from WDSV, the only other fish retrovirus for which sequence data are available. These viruses share a similar genetic organization and contain additional potential ORFs in accord with their classification as complex retroviruses. The viruses differ, however, in the class of tRNA primer and do not exhibit high levels of sequence similarity. Indeed, the *pol* region of SnRV is no closer in similarity to that of WDSV than to that of the mammalian type C viruses, and comparisons of the *gag* and *env* regions have revealed even less similarity.

A growing number of retroviruses, many of which are associated with neoplastic diseases, have been observed in fish species (4, 46). Two of these retroviruses, WDSV and the SnRV reported here, have now been characterized at the molecular level. The divergent nature of these complex viruses, not only from the known retrovirus groups but also from each other, suggests that there exist groups of such viruses that have yet to be identified among the lower vertebrates. The elucidation of the replication strategies employed by these viruses is likely to extend the knowledge of retrovirus-host interactions and the basis of disease. Further characterization of fish retroviruses will also determine if there exists a diversification of viral types parallel to that found in the higher vertebrates and may help in the understanding of the evolutionary history of retroviruses as a whole.

ACKNOWLEDGMENT

This work was supported by the UK Overseas Development Administration.

REFERENCES

1. **Beisel, C. E., J. F. Edwards, L. L. Dunn, and N. R. Rice.** 1993. Analysis of multiple mRNAs from pathogenic equine infectious anemia virus (EIAV) in an acutely infected horse reveals a novel protein, Ttm, derived from the carboxy terminus of the EIAV transmembrane protein. J. Virol. **67:**832–842.
2. **Berkhout, B., and F. J. Van Hemert.** 1994. The unusual nucleotide content of the HIV RNA genome results in a biased amino acid composition of HIV proteins. Nucleic Acids Res. **22:**1705–1711.
3. **Berlioz, C., and J.-L. Darlix.** 1995. An internal ribosomal entry mechanism promotes translation of murine leukemia virus *gag* polyprotein precursors. J. Virol. **69:**2214–2222.
4. **Bowser, P. R., and J. W. Casey.** 1993. Retroviruses of fish. Annu. Rev. Fish Dis. **3:**209–224.
5. **Brierly, I.** 1995. Ribosomal frameshifting on viral RNAs. J. Gen. Virol. **76:**1885–1892.
6. **Bronson, E. C., and J. N. Anderson.** 1994. Nucleotide composition as a driving force in the evolution of retroviruses. J. Mol. Evol. **38:**506–532.
7. **Bucher, P.** 1990. Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. J. Mol. Biol. **212:**563–578.
8. **Carroll, R., and D. Derse.** 1993. Translation of equine infectious anemia virus bicistronic *tat-rev* mRNA requires leaky ribosome scanning of the *tat* CTG initiation codon. J. Virol. **67:**1433–1440.
9. **Coffin, J. M.** 1992. Structure and classification of retroviruses, p. 19–49. *In* J. A. Levy (ed.), The Retroviridae, vol. 1. Plenum Press, New York.
10. **Dayhoff, M. O.** 1978. Atlas of protein sequence and structure, vol. 5, suppl. 3. National Biomedical Research Foundation, Washington, D.C.
11. **Dingwall, C., and R. A. Laskey.** 1991. Nuclear targeting sequences—a consensus? Trends Biochem. Sci. **16:**478–481.
12. **Doolittle, R. F., D.-F. Feng, M. S. Johnson, and M. A. McClure.** 1989. Origins and evolutionary relationships of retroviruses. Q. Rev. Biol. **64:**1–30.
13. **Dorn, P., L. DaSilva, L. Martarano, and D. Derse.** 1990. Equine infectious anemia virus *tat*: insights into the structure, function, and evolution of lentivirus *trans*-activator proteins. J. Virol. **64:**1616–1624.
14. **Elder, J. H., D. L. Lerner, C. S. Hasselkus-Light, D. J. Fontenot, E. Hunter, P. A. Luciw, R. C. Montelaro, and T. R. Phillips.** 1992. Distinct subsets of retroviruses encode dUTPase. J. Virol. **66:**1791–1794.
15. **Felsenstein, J.** 1993. PHYLIP manual, version 3.52c. University of Washington, Seattle.
16. **Feng, Y.-X., H. Yuan, A. Rein, and J. G. Levin.** 1992. Bipartite signal for read-through suppression in murine leukemia virus mRNA: an eight-nucleotide purine-rich sequence immediately downstream of the *gag* termination codon followed by an RNA pseudoknot. J. Virol. **66:**5127–5132.
17. **Fitzgerald, L. M., A. Rodríguez, and G. Smutzer.** 1993. Codon usage in bony fish. Mol. Mar. Biol. Biotechnol. **2:**112–119.
18. **Frerichs, G. N., D. Morgan, D. Hart, C. Skerrow, R. J. Roberts, and D. E. Onions.** 1991. Spontaneously productive C-type retrovirus infection of fish cell lines. J. Gen. Virol. **72:**2537–2539.
19. **Furfine, E. S., and J. E. Reardon.** 1991. Human immunodeficiency virus

reverse transcriptase ribonuclease H: specificity of tRNA^{Lys3}-primer excision. Biochemistry **30:**7041–7046.

20. **Gallaher, W. R., J. M. Ball, R. F. Garry, M. C. Griffin, and R. C. Montelaro.** 1989. A general model for the transmembrane proteins of HIV and other retroviruses. AIDS Res. Hum. Retroviruses **5:**431–440.
21. **Gonda, M. A.** 1994. The lentiviruses of cattle, p. 83–109. *In* J. A. Levy (ed.), The Retroviridae, vol. 3. Plenum Press, New York.
22. **Göttlinger, H. G., T. Dorfman, J. G. Sodroski, and W. A. Haseltine.** 1991. Effect of mutations affecting the p6 *gag* protein on human immunodeficiency virus particle release. Proc. Natl. Acad. Sci. USA **88:**3195–3199.
23. **Harada, F., and S. Nishimura.** 1980. tRNAs containing the g-psi-psi-c sequence: two arginine tRNAs of mouse leukemia cells. Biochem. Int. **1:**539–546.
24. **Henikoff, S., and J. G. Henikoff.** 1992. Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. USA **89:**10915–10919.
25. **Holzschu, D. L., D. Martineau, S. K. Fodor, V. M. Vogt, P. R. Bowser, and J. W. Casey.** 1995. Nucleotide sequence and protein analysis of a complex retrovirus, walleye dermal sarcoma virus. J. Virol. **69:**5320–5331.
26. **Huang, M., J. M. Orenstein, M. A. Martin, and E. O. Freed.** 1995. p6^{Gag} is required for particle production from full-length human immunodeficiency virus type 1 molecular clones expressing protease. J. Virol. **69:**6810–6818.
27. **Jacks, T.** 1990. Translational suppression in gene expression in retroviruses and retrotransposons. Curr. Top. Microbiol. Immunol. **157:**93–124.
28. **Johnson, M. S., M. A. McClure, D.-F. Feng, J. Gray, and R. F. Doolittle.** 1986. Computer analysis of retroviral *pol* genes: assignment of enzymatic functions to specific sequences and homologies with nonviral enzymes. Proc. Natl. Acad. Sci. USA **83:**7648–7652.
29. **Kozak, M.** 1986. Point mutations define a sequence flanking the ATG initiator codon that modulates translation by eukaryotic ribosomes. Cell **44:**283–292.
30. **Kozak, M.** 1987. Effects of intercistronic length on the efficiency of reinitiation by eukaryotic ribosomes. Mol. Cell. Biol. **7:**3438–3445.
31. **Kozak, M.** 1989. The scanning model for translation: an update. J. Cell Biol. **108:**229–241.
32. **Lupas, A., M. Van Dyke, and J. Stock.** 1991. Predicting coiled coils from protein sequences. Science **252:**1162–1164.
33. **Matthews, S., P. Barlow, J. Boyd, G. Barton, R. Russell, H. Mills, M. Cunningham, N. Meyers, N. Burns, N. Clark, S. Kingsman, A. Kingsman, and I. Campbell.** 1994. Structural similarity between the p17 matrix protein of HIV-1 and interferon-gamma. Nature (London) **370:**666–668.
34. **McClure, M. A., M. S. Johnson, D.-F. Feng, and R. F. Doolittle.** 1988. Sequence comparisons of retroviral proteins: relative rates of change and general phylogeny. Proc. Natl. Acad. Sci. USA **85:**2469–2473.
35. **McGeoch, D. J.** 1990. Protein sequence comparisons show that the "pseudoproteases" encoded by poxviruses and certain retroviruses belong to the deoxyuridine triphosphatase family. Nucleic Acids Res. **18:**4105–4110.
36. **Miller, S. A., D. D. Dykes, and H. F. Polesky.** 1988. A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res. **16:**1215.
37. **Murphy, F. A., C. M. Fauquet, M. A. Mayo, A. W. Jarvis, S. A. Ghabrial, M. D. Summers, G. P. Martelli, and D. H. L. Bishop (ed.).** 1995. Virus taxonomy—the classification and nomenclature of viruses: sixth report of the International Committee on Taxonomy of Viruses. Springer-Verlag, New York.
38. **Myers, G., and G. N. Pavlakis.** 1992. Evolutionary potential of complex retroviruses, p. 51–105. *In* J. A. Levy (ed.), The Retroviridae, vol. 1. Plenum Press, New York.
39. **Myers, G., S. Wain-Hobson, B. Korber, R. F. Smith, and G. N. Pavlakis (ed.).** 1993. Human retroviruses and AIDS: a compilation and analysis of nucleic acid and amino acid sequences. Los Alamos National Laboratory, Los Alamos, N.Mex.
40. **Noiman, S., A. Gazit, O. Tori, L. Sherman, T. Miki, S. R. Tronick, and A. Yaniv.** 1990. Identification of sequences encoding the equine infectious anemia virus *tat* gene. Virology **176:**280–288.
41. **Pancino, G., H. Ellerbrok, M. Sitbon, and P. Sonigo.** 1994. Conserved framework of envelope glycoproteins among lentiviruses. Curr. Top. Microbiol. Immunol. **188:**77–105.
42. **Parent, L. J., R. P. Bennett, R. C. Craven, T. D. Nelle, N. K. Krishna, J. B. Bowzard, C. B. Wilson, B. A. Puffer, R. C. Montelaro, and J. W. Wills.** 1995. Positionally independent and exchangeable late budding functions of the Rous sarcoma virus and human immunodeficiency virus Gag proteins. J. Virol. **69:**5455–5460.
43. **Parry, D. A. D.** 1982. Coiled-coils in α-helix-containing proteins: analysis of the residue types within the heptad repeat and the use of these data in the prediction of coiled-coils in other proteins. Biosci. Rep. **2:**1017–1024.
44. **Patarca, R., and W. A. Haseltine.** 1985. A major retroviral core protein related to EPA and TIMP. Nature (London) **318:**390.
45. **Peterlin, B. M., M. Adams, A. Alonso, A. Baur, S. Ghosh, X. Lu, and Y. Luo.** 1993. Tat *trans*-activator, p. 75–100. *In* B. R. Cullen (ed.), Human retroviruses. IRL Press, Oxford.
46. **Poulet, F. M., P. R. Bowser, and J. W. Casey.** 1994. Retroviruses of fish, reptiles, and molluscs, p. 1–38. *In* J. A. Levy (ed.), The Retroviridae, vol. 3. Plenum Press, New York.
47. **Prestridge, D. S.** 1991. SIGNAL SCAN: a computer program that scans DNA sequences for eukaryotic transcriptional elements. Comput. Appl. Biol. Sci. **7:**203–206.
48. **Pullen, K. A., L. K. Ishimoto, and J. J. Champoux.** 1992. Incomplete removal of the RNA primer for minus-strand DNA synthesis by human immunodeficiency virus type 1 reverse transcriptase. J. Virol. **66:**367–373.
49. **Robbins, J., S. M. Dilworth, R. A. Laskey, and C. Dingwall.** 1991. Two interdependent basic domains in nucleoplasmin nuclear targeting sequence: identification of a class of bipartite nuclear targeting sequence. Cell **64:**615–623.
50. **Saitou, N., and M. Nei.** 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:**406–425.
51. **Saltarelli, M., G. Querat, D. A. M. Konings, R. Vigne, and J. E. Clements.** 1990. Nucleotide sequence and transcriptional analysis of molecular clones of CAEV which generate infectious virus. Virology **179:**347–364.
52. **Sambrook, J., E. F. Fritsch, and T. Maniatis.** 1989. Molecular cloning: a laboratory manual, 2nd ed. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
53. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA **74:**5463–5467.
54. **Schwartz, D. E., R. Tizard, and W. Gilbert.** 1983. Nucleotide sequence of Rous sarcoma virus. Cell **32:**853–869.
55. **Shapiro, M. B., and P. Senapathy.** 1987. RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. Nucleic Acids Res. **15:**7155–7174.
56. **Stephens, R. M., D. Derse, and N. R. Rice.** 1990. Cloning and characterization of the cDNAs encoding equine infectious anemia virus Tat and putative Rev proteins. J. Virol. **64:**3716–3725.
57. **Temin, H. M.** 1992. Origin and general nature of retroviruses, p. 1–18. *In* J. A. Levy (ed.), The Retroviridae, vol. 1. Plenum Press, New York.
58. **Thompson, J. D., D. G. Higgins, and T. J. Gibson.** 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22:**4673–4680.
59. **Van Hemert, F. J., and B. Berkhout.** 1995. The tendency of lentiviral open reading frames to become A-rich: constraints imposed by viral genome organization and cellular tRNA availability. J. Mol. Evol. **41:**132–140.
60. **Von Heijne, G.** 1986. A new method for predicting signal sequence cleavage sites. Nucleic Acids Res. **14:**4683–4690.
61. **Weber, I. T.** 1989. Structural alignment of retroviral protease sequences. Gene **85:**565–566.
62. **Williamson, M. P.** 1994. The structure and function of proline-rich regions in proteins. Biochem. J. **297:**249–260.
63. **Wills, J. W., C. E. Cameron, C. B. Wilson, Y. Xiang, R. P. Bennett, and J. Leis.** 1994. An assembly domain of the Rous sarcoma virus Gag protein required late in budding. J. Virol. **68:**6605–6618.
64. **Wills, J. W., and R. C. Craven.** 1991. Form, function and use of retroviral Gag proteins. AIDS **5:**639–654.
65. **Wills, N. M., R. F. Gesteland, and J. F. Atkins.** 1991. Evidence that a downstream pseudoknot is required for translational read-through of the Moloney murine leukemia virus *gag* stop codon. Proc. Natl. Acad. Sci. USA **88:**6991–6995.
66. **Wills, N. M., R. F. Gesteland, and J. F. Atkins.** 1994. Pseudoknot-dependant read-through of retroviral *gag* termination codons: importance of sequences in the spacer and loop 2. EMBO J. **13:**4137–4144.
67. **Xiong, Y., and T. H. Eickbush.** 1990. Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J. **9:**3353–3362.
68. **Yoshinaka, Y., I. Katoh, T. D. Copeland, and S. Oroszlan.** 1985. Murine leukemia virus protease is encoded by the *gag-pol* gene and is synthesized through suppression of an amber termination codon. Proc. Natl. Acad. Sci. USA **82:**1618–1622.
69. **Zhou, W., L. J. Parent, J. W. Wills, and M. D. Resh.** 1994. Identification of a membrane-binding domain within the amino-terminal region of human immunodeficiency virus type 1 Gag protein which interacts with acidic phospholipids. J. Virol. **68:**2556–2569.