JOURNAL OF VIROLOGY, Sept. 1995, p. 5320–5331 Vol. 69, No. 9
0022-538X/95/$04.00+0

Nucleotide Sequence and Protein Analysis of a Complex Piscine Retrovirus, Walleye Dermal Sarcoma Virus

DONALD L. HOLZSCHU,[1] DANIEL MARTINEAU,[2] SHARON K. FODOR,[3] VOLKER M. VOGT,[3] PAUL R. BOWSER,[4] AND JAMES W. CASEY[1]*

Section of Biochemistry, Molecular and Cell Biology,[3] Departments of Microbiology, Parasitology and Immunology,[1] and Avian and Aquatic Animal Medicine,[4] Cornell University, Ithaca, New York 14853, and Department de Microbiologie et Pathologie, Faculte de Medicine Veterinaire, University of Montreal, Saint-Hyacinthe, Montreal, Quebec J2S 7C5, Canada[2]

Received 14 February 1995/Accepted 12 May 1995

Walleye dermal sarcoma virus (WDSV) is a fish retrovirus associated with the development of tumors in walleyes. We have determined the complete nucleotide sequence of a DNA clone of WDSV, the N-terminal amino acid sequences of the major proteins, and the start site for transcription. The long terminal repeat is 590 bp in length, with the U3 region containing consensus sequences likely to be involved in viral gene expression. A predicted histidyl-tRNA binding site is located 3 nucleotides distal to the 3′ end of the long terminal repeat. Virus particles purified by isopycnic sedimentation followed by rate zonal sedimentation showed major polypeptides with molecular sizes of 90, 25, 20, 14, and 10 kDa. N-terminal sequencing of these allowed unambiguous assignment of the small polypeptides as products of the *gag* gene, including CA and NC, and the large polypeptide as the TM product of *env*. The 582-amino-acid (aa) Gag protein precursor is predicted to be myristylated as is found for most retroviruses. NC contains a single Cys-His motif like those found in all retroviruses except spumaviruses. The WDSV *pro* and *pol* genes are in the same translational reading frame as *gag* and thus apparently are translated after termination suppression. The *env* gene encodes a surface (SU) protein of 469 aa predicted to be highly glycosylated and a large transmembrane (TM) protein of 754 aa. The sequence of TM is unusual in that it ends in a very hydrophobic segment of 65 residues containing a single charged residue. Following the *env* gene are two nonoverlapping long open reading frames of 290 aa (*orf-A*) and 306 aa (*orf-B*), neither of which shows significant sequence similarity with known genes. A third open reading frame of 119 aa (*orf-C*) is located in the leader region preceding *gag*. The predicted amino acid sequence of reverse transcriptase would place WDSV phylogenetically closest to the murine leukemia virus-related genus of retroviruses. However, other members of this genus do not have accessory genes, suggesting that WDSV acquired *orf-A*, *orf-B*, and perhaps *orf-C* late in its evolution. We hypothesize by analogy with other complex retroviruses that the accessory genes of WDSV function in the regulation of transcription and in RNA processing and also in the induction of walleye dermal sarcoma.

Walleye dermal sarcoma virus (WDSV) is etiologically associated with the induction of a multifocal benign cutaneous tumor, walleye dermal sarcoma (WDS) (42). This disease is observed at a very high frequency in some localities; 5 to 10% of the walleyes in some Canadian waters and up to 27% of walleyes collected from Oneida Lake, New York, are affected (3, 46, 47). Although WDS often appears histologically malignant, this tumor has not been observed to be invasive or metastatic in feral fish (23). Tumor transmission has been accomplished by inoculating cell-free WDS filtrates into walleye fingerlings. Macroscopic tumors become visible 10 to 12 weeks after inoculation on greater than 80% of the fingerlings (22). The experimentally induced neoplasms are morphologically identical to the spontaneous tumors of adult fish, and type C retrovirus particles have been observed in some of the tumors (22).

Similar to several other neoplastic diseases of poikilotherms, WDS appears and regresses seasonally (30). Host physiological parameters contributing to tumor appearance and regression are unknown but may involve endocrine changes and variations in immune response at different water temperatures (1). The properties intrinsic to WDSV that affect tumor induction and regression are not well-studied, but it is possible that the replication cycle of some piscine retroviruses is restricted, thereby leading to the seasonal progression and regression of fish tumors (29).

Relative to avian and mammalian model systems that have been established to study retrovirus-induced oncogenesis, very little is known about the retroviruses that induce neoplasia in fish. We recently cloned the DNA genome of WDSV from unintegrated DNA in tumor tissue (21). The size of the DNA, ca. 13 kb, placed the WDSV genome among the largest known, similar to the genomes of spumaviruses (13, 34). The unintegrated DNA in tumors was shown to have a region susceptible to single-strand-specific nucleases (sometimes referred to as a gap structure), a feature common to spumaviruses and some lentiviruses (2, 4, 19). The presence of short transcripts in tumor tissues suggested that WDSV encodes accessory proteins, as do all the members of the three retrovirus genera typified by human spumaretrovirus (HSRV; spumaviruses), human immunodeficiency virus type 1 (HIV-1; lentiviruses), and human T-cell leukemia virus (HTLV and bovine leukemia virus [BLV]) (10, 14, 27, 35). As a first step in characterizing the properties of WDSV that may contribute to induction and subsequent regression of tumors, we have sequenced a full-length DNA clone of WDSV and analyzed the major virion proteins.

* Corresponding author. Phone: (607) 253-3579. Fax: (607) 253-3419. Electronic mail address: jwc3@cornell.edu.

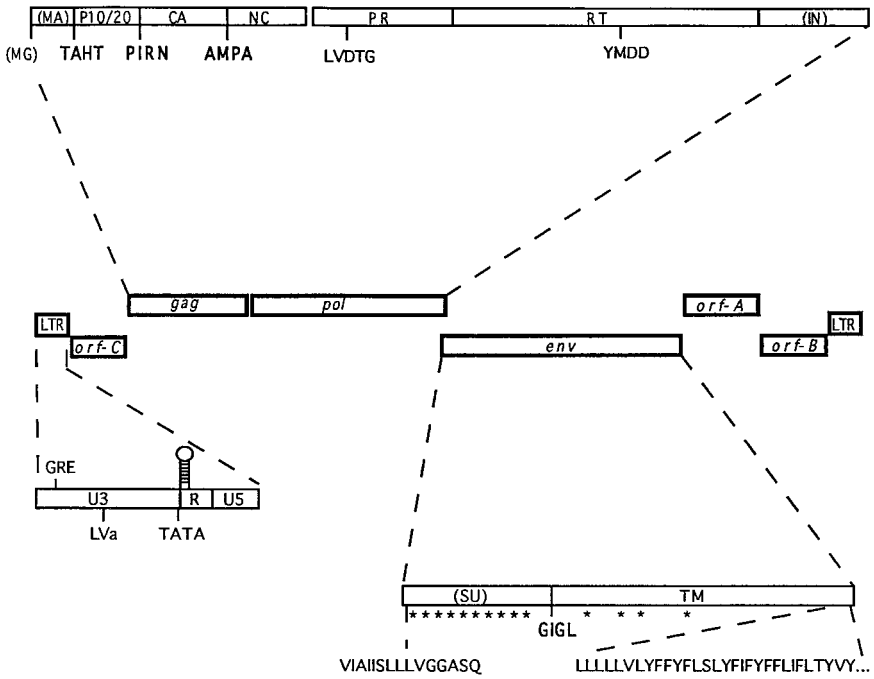

FIG. 1. Coding regions in WDSV. Experimentally determined N-terminal sequences of Gag and Env proteins are shown in bold type. Hallmark sequences for PR, RT, SU, and TM are described in the text and are shown in plain type. Asterisks below SU and TM denote N-linked glycosylation sites. *orf-C*, *orf-A*, and *orf-B* are predicted from the sequence data but have not been verified to be translated into proteins. Three consensus sequences within the U3 region, the predicted stem-loop structure at the start of transcription, and the U3-R junction are indicated above and below the WDSV LTR.

MATERIALS AND METHODS

DNA sequencing. All sequencing was performed on the previously described lambda clone WDSV-1 (21). A tissue culture system for WDSV is not yet available, thus, the clone WDSV-1 has not been tested for infectivity. Eleven *Bam*HI, *Hin*dIII, *Pst*I, *Eco*RI and *Cla*I fragments were subcloned into pBluescriptSK− and sequenced by using T3 and T7 primers and standard dideoxy sequencing techniques (38). These sequences were extended with WDSV-specific oligonucleotide primers. Subsequent rounds of primer synthesis and sequencing were used to processively sequence both strands of the WDSV genome. All WDSV-specific primers were synthesized on an ABI 392 automated DNA synthesizer. Plasmid DNA was prepared on Qiagen columns and sequenced by using either an ABI 393 at the University of Georgia core facility or a Pharmacia A.L.F. system at the Institut Armand-Frappier. Subsequently, the sequence was edited with data obtained by manual dideoxy sequencing of both strands of the WDSV genome (Sequenase; U.S. Biochemicals). Sequences were assembled using AssemblyLIGN and analyzed using MacVector, DNAStar, and the Wisconsin Genetics Computer Group package.

To verify that *gag*, *pro*, and *pol* are in the same reading frame, we have sequenced the PCR products encompassing this region from the DNAs of three different tumors. These data are not shown, but are consistent with the sequence of the lambda clone.

Primer extension. Virus isolation and RNA purification were carried out as previously described (24). Approximately 25 pmol of the ^{32}P-end-labeled oligonucleotides 5′ AGCCCAGGCCAGTCACATTTGTCTG 3′ and 5′ CTTATCTA AATTATATCCATGTTCG 3′ and approximately 100 ng of WDSV virion RNA were used in reverse transcription reactions to map the 5′ end of the WDSV genome (37). The products were separated on a 5% sequencing gel, and their lengths were determined by comparison with a pBluescriptSK− sequencing ladder.

Purification and analysis of WDSV proteins. Virus particles were collected by medium speed centrifugation from a tumor homogenate, banded on isopycnic sucrose gradients as previously described (24), and then collected by centrifugation. Further purification was achieved by rate zonal sedimentation. The redissolved pellet was sonicated for 2 min, layered over a 35-ml 5 to 25% (wt/vol) sucrose gradient in 10 mM Tris-HCl (pH 7.5)–1 mM EDTA–100 mM NaCl. The samples were centrifuged for 60 min at 43,000 × *g* in an SW-28 rotor. Sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) was carried out by standard procedures, and the gels were stained with 0.05% Coomassie blue–7% acetic acid–40% methanol. We found that destaining in 7% acetic acid–40% methanol led to disappearance of one of the Gag polypeptides, while destaining in 1% formic acid–40% methanol retained this polypeptide. N-terminal amino acid sequencing of polypeptides electrophoretically transferred to Immobilon membranes was performed by the Cornell Biotechnology Facility. Labeling of virions with NHS-biotin (*N*-hydroxysuccinimidobiotin; Pierce Biochemicals) was performed at 0°C for 1 h in 50 mM $NaHCO_3$–100 mM NaCl–5% sucrose. The reaction was quenched by the addition of ethanolamine to 25 mM. Labeled virions were digested with 6.2 U of *N*-glycosidase F (peptide *N*-glycosidase F; Boehringer) in 1.4% Nonidet P-40–20 mM NaH_2PO_4–50 mM EDTA at 37°C for 6 h. After SDS-PAGE, polypeptides were transferred to a polyvinylidine difluoride membrane (Immobilon; Millipore Corp.), which was then blocked with 10% nonfat dry milk–0.1% Tween 20. Subsequently, the membrane was probed with 0.2 μg of streptavidin-HRP (ImmunoPure Streptavidin, horseradish peroxidase conjugated; Pierce) per ml. The membranes were developed with enhanced chemiluminescence reagents (Amersham).

Searches of the GenBank database for proteins with sequence similarities to the novel WDSV open reading frames (ORFs) were performed using FASTA, one of the Genetics Computer Group package of programs. For the phylogenetic comparisons of reverse transcriptase amino acid sequences, the seven blocks of conserved sequences as defined by Xiong and Eickbush (45), consisting of 178 residues of reverse transcriptase, were located manually in the WDSV *pol* gene and lined up with the same sequences from other retroviruses as described previously (45). The 178 residues were treated as a single contiguous sequence without gaps, and trees were generated with the Megalign application in the DNAStar series of programs (DNAstar, Madison, Wis.), which uses a modification of the neighbor-joining method (36) to generate suggested phylogenetic trees.

Nucleotide sequence accession number. The accession number for the WDSV sequence is L41838.

RESULTS

The overall organization of WDSV DNA, as inferred from the nucleotide sequence coupled with the N-terminal amino acid sequences of structural genes, reveals the distinguishing features of this virus (Fig. 1). Long terminal repeats (LTRs) are found at each end, bracketing several long ORFs. Three ORFs are the *gag*, *pol*, and *env* genes found in all retroviruses. *orf-A*, *orf-B*, and *orf-C* encode putative proteins which have not been detected and whose functions remain to be determined.

LTR and leader region. Of the 12,708-bp sequence of the entire DNA, 590 bp at each end correspond to the LTRs (Fig. 2). These are composed of a 440-bp U3 region, a 77-bp R region, and a 73-bp U5 region. The boundary between U3 and R, which is the start site of transcription, was determined by

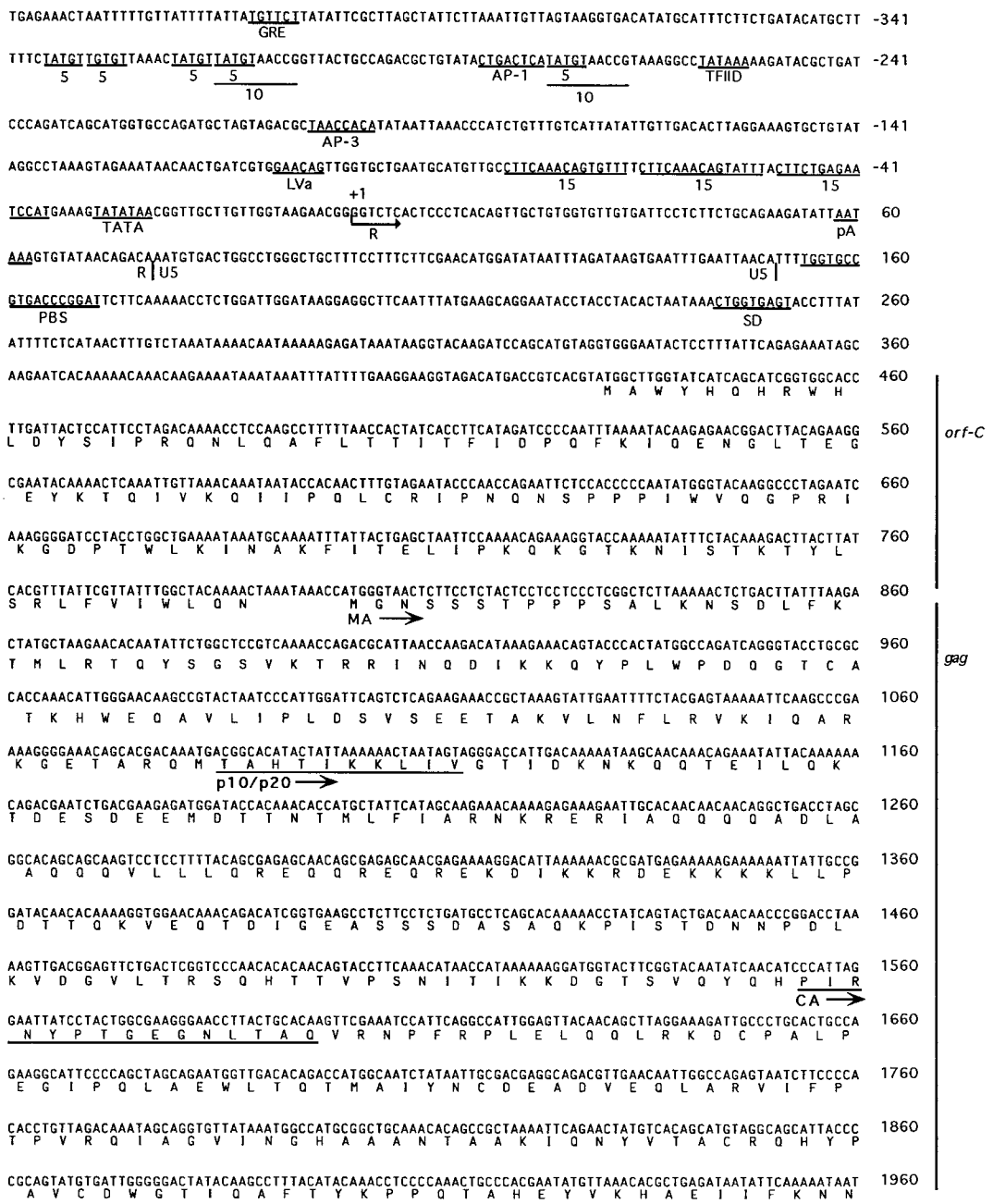

FIG. 2. The complete nucleotide sequence of WDSV-1 DNA. The deduced amino acid sequences of *gag*, *pol*, *env*, *orf-A*, *orf-B*, and *orf-C* are shown, with experimentally determined N-terminal sequences underlined and hallmark sequences boxed. The predicted boundaries of U3, R, and U5 are indicated along with the transcriptional start site (+1). Potential binding sites for transcription factors in the U3 region are underlined and abbreviated. Direct repeats of 5, 10, and 15 nucleotides are underlined and designated accordingly. The following annotations are used: pA, polyadenylation signal sequence; PBS, primer-binding site; SD, possible splice donor site; MA, matrix protein; CA, capsid protein; NC, nucleocapsid protein; SU, Env surface protein; TM, Env transmembrane protein; PPT, polypurine tract.

primer extension of virion RNA (Fig. 3). By definition, transcriptional initiation occurs at position 1, which is 23 bp downstream of the TATA box sequence TATATAA. A consensus polyadenylation sequence, AATAAA, occurs at position 58. In most retroviral RNAs, R ends with a CA dinucleotide approximately 20 nucleotides downstream of the polyadenylation signal. By analogy, in WDSV we assigned the dinucleotide CA at position 76, which occurs 13 b after the polyadenylation signal, as the 3′ end of R. The 3′ boundary of U5 was identified by comparison with the 3′ LTR sequence and its proximity to the $tRNA^{His}$ binding site. As with all retroviral LTRs, the WDSV LTR begins and ends with the dinucleotides TG. . . .CA, which are part of an imperfect inverted repeat.

We have identified a number of motifs within the WDSV LTR that may be involved in regulation of transcription (Fig. 2). In particular, U3 contains putative binding sites for the transcription factors AP-1 and AP-3 and the glucocorticoid receptor. An element identical to the LVa motif in murine leukemia virus (MLV) (40) is part of a larger segment that includes an MLV-like core enhancer sequence GTGGAAC for the C/EBP factor (Fig. 4A). Other sequences possibly involved in transcription are a pentanucleotide direct repeat, TRTGT,

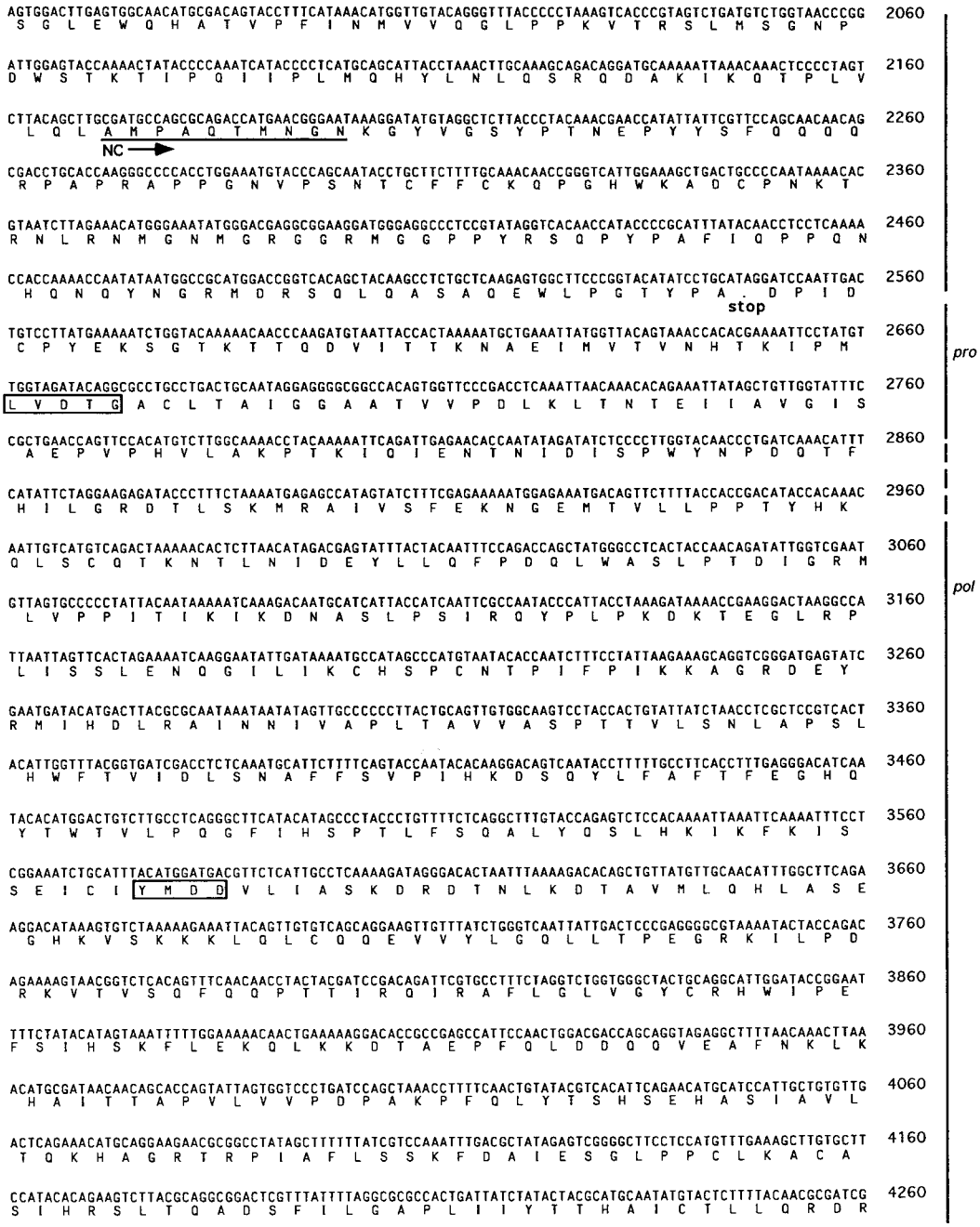

FIG. 2—*Continued.*

that occurs five times between nucleotides −337 and −272; two of these repeats are part of a 10-bp direct repeat starting at nucleotides −316 and −277 and part of a degenerate 15-bp direct repeat that occurs three times starting at nucleotide −82. Interestingly, the first 96 nucleotides at the 5′ end of the RNA can be drawn as a large stem-loop structure (Fig. 4B). The resemblance between this stem-loop and similar structures found at the same position in some lentiviruses is provocative.

The tRNA binding site used for initiation of minus-strand DNA synthesis begins three bases 3′ of the WDSV LTR. The 18-nucleotide WDSV tRNA binding site is 90% identical with $tRNA^{His}$ from *Drosophila melanogaster* and sheep. The next most closely related tRNA in the database is $tRNA^{Phe}$ from a plant, *Lupin minor*, which is 70% identical. On the basis of this information we conclude that $tRNA^{His}$ is used by WDSV as the primer for first-strand synthesis.

The leader sequence, by definition corresponding to the RNA segment proximal to the *gag* gene, is 799 nucleotides long. A potential splice donor sequence occurs at position 244. The leader RNA has seven methionine codons, six of which are closely followed by a translational termination codon. The methionine codon at position 430, which is in a reasonable translational context (18), is followed by an ORF for a protein of 119 amino acids (*orf-C*) with no obvious homology with proteins in the databases.

***gag* gene.** The internal structural proteins of all retroviruses are proteolytically cleaved from a polyprotein called Gag, the product of the *gag* gene. Among diverse retroviruses, Gag

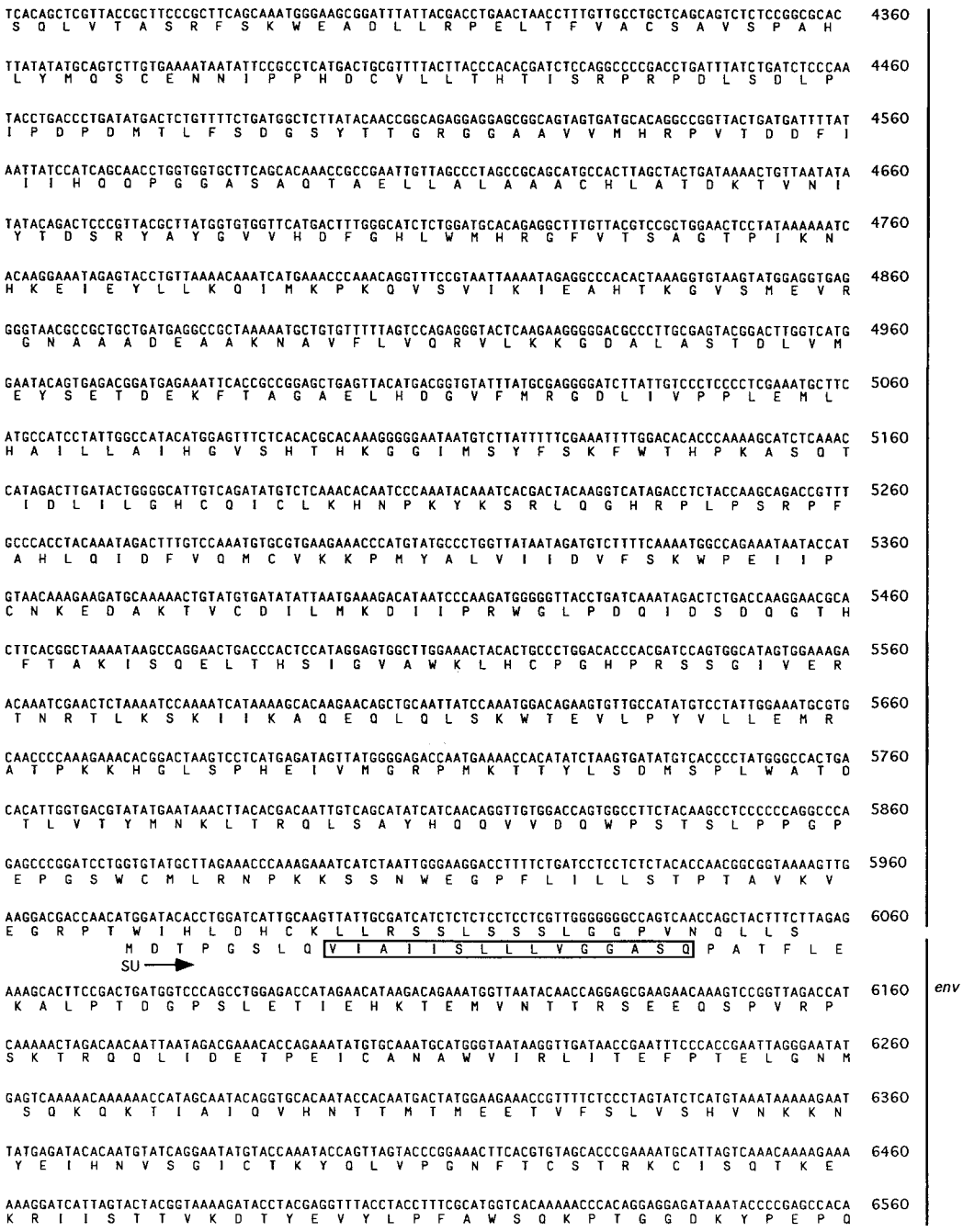

FIG. 2—*Continued.*

proteins show little amino acid sequence similarity, with the exception of a short stretch of ca. 20 residues (the major homology region) in the capsid protein (CA) (43) and the motif $CX_2CX_4HX_4C$ in the nucleocapsid protein (NC). The other Gag protein found in all retroviruses, matrix protein (MA), in general has no identifying consensus sequences except that it invariably forms the N-terminal domain of Gag and in most retroviruses is myristylated at the glycine occurring at residue 2. The *gag* gene of most retroviruses also encodes one or more other mature proteins of unknown function, between the MA and CA domains and sometimes also distal to the NC domain. In order to identify WDSV Gag proteins, we first extended the published purification scheme for virus particles (24) to include rate zonal sedimentation after the equilibrium banding step. The protein profile of virions sedimenting at about 600S showed several major low-molecular-weight polypeptides, presumed to be the products of *gag* (Fig. 5), plus one prominent polypeptide with an apparent size of 90 kDa. The latter was identified as a surface protein by virtue of its disappearance after digestion of virions with trypsin under conditions in which none of the Gag proteins was affected (data not shown).

To place coding sequences for these proteins on the DNA sequence, N-terminal amino acid sequence analysis was performed after electrophoretic transfer of the polypeptides to a membrane. All proteins yielded interpretable sequences that

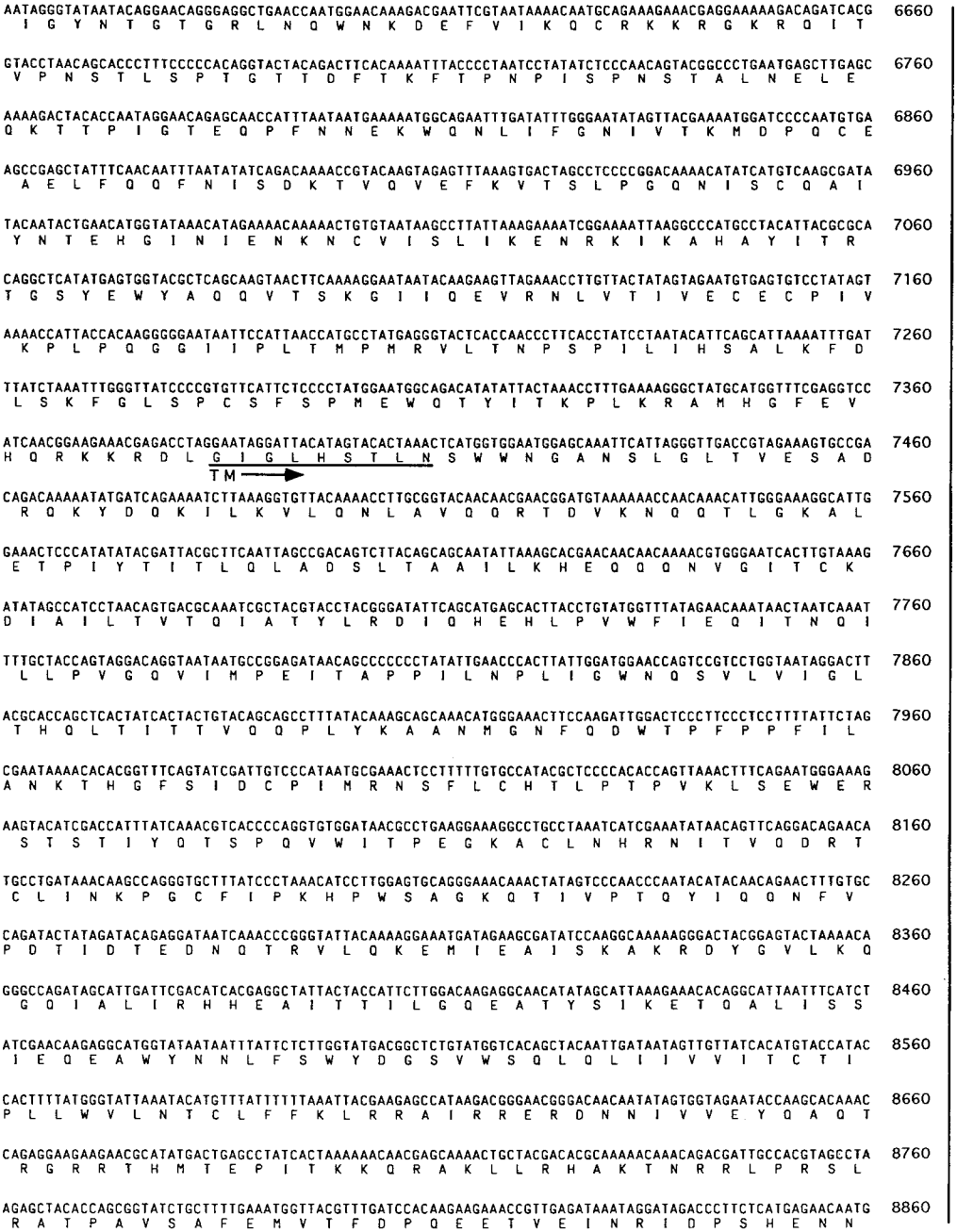

FIG. 2—*Continued.*

matched or nearly matched the sequence predicted from the DNA sequence (Fig. 2). We conclude that p25, p20, p14, and p10 are products of *gag*. On the basis of its size and relative position in Gag, p25 is CA [N terminus PIRNYPTG(D/E) GN(L/V)TAQ; predicted sequence at position 1553, PIRNYP TGEGNLTAQ]. Assuming there is no additional proteolytic processing at the C terminus (as there is for example in Rous sarcoma virus [RSV] as well as in HIV-1), the size of CA is 206 residues. The predicted major homology region for WDSV CA has the sequence VVQGLPPKVTRSLMSGNPDW, which is quite diverged in its C-terminal region from the consensus established for other retroviruses, ϕxQGxxExxxxFϕxRLxxxϕ, where ϕ is a hydrophobic residue and the other residues marked are universally or almost universally conserved (43). On the basis of its relative position and the presence of the characteristic Cys-His motif, p14 is NC (N terminus AMPA QNINGN; predicted sequence at position 2171, AMPAQTM NGN). The predicted size, 125 residues to the amber codon defining the end of Gag (position 2546), fits approximately with the size observed by SDS-PAGE. However, we cannot exclude the fact that a short sequence is processed from the C terminus of Gag to yield the final mature NC protein, as is observed in MLV (16). The two residues that do not match the predicted sequence (underlined) could be due to amino acid sequencing mistakes or alternatively to heterogeneity in the population of WDSV particles. The polypeptide p10 apparently is nested in p20, since both were found to have nearly the same start (N terminus of p10, GAHTIKKLIV, and that of

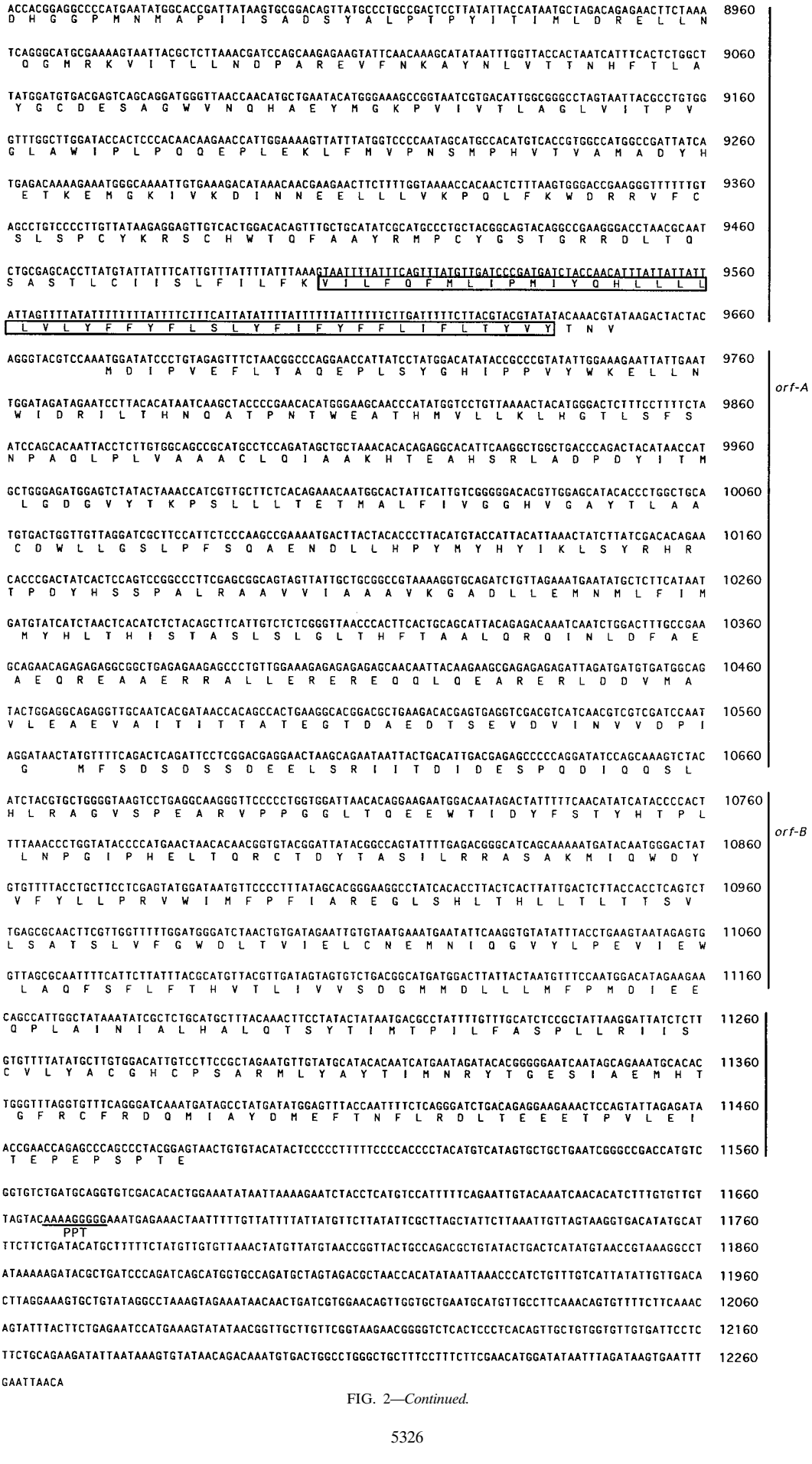

FIG. 2—*Continued.*

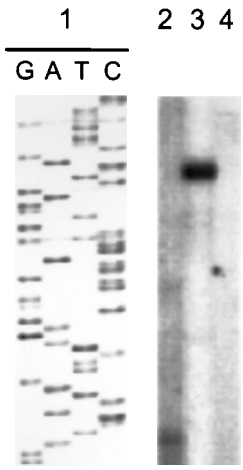

FIG. 3. Mapping of the transcriptional start site of WDSV. Primer extension analysis was carried out with virion RNA and the primers: AGCCCAGGCC AGTCACATTTGTCTG (lane 2) and CTTATCTAAATTATATCCATGTTCG (lane 3). The length of the extension products was determined from a T7-primed pBluescriptSK− sequencing reaction run in parallel (lane 1. A primer extension reaction run without the addition of a primer is shown in lane 4.

p20, TAHTIKKLI; predicted sequence at position 800, TAH TIKKLIV), implying incomplete proteolytic processing of this segment of Gag. (We presume that the first amino acid residue obtained for p10 was in error.) Most other retroviruses also have one or two mature Gag proteins that are derived from the region between MA and CA. The MA protein formed from the N-terminal domain of Gag (predicted N terminus, MGNS at position 1245) is expected to be 96 residues in length and thus presumably comigrates with p10. Because of the myristate-blocking group expected to be on the N-terminal Gly residue, MA should not have contributed to the N-terminal amino acid sequence determined for p10. The small predicted 90-residue protein immediately upstream of CA, which represents the difference between p20 and p10 and should be present in the same molar amounts as p10, may migrate with the buffer front under the conditions used here for gel electrophoresis, or perhaps this predicted protein is cleaved into smaller fragments by the viral protease.

***pro* and *pol* genes.** In all retroviruses the sequence following *gag* encodes the viral enzymes protease (PR), reverse transcriptase (RT), and integrase (IN), which are synthesized as a single large precursor polypeptide fused to Gag. In most viruses PR is encoded in the same frame as RT and IN. However, since PR in some viruses is in a different reading frame, it is by convention assigned its own gene. In WDSV *pro* and *pol* are contiguous in the same reading frame. For all but one genus of retroviruses, *pol* (and usually *pro*) is expressed by ribosomal frameshifting. Thus a small percentage of ribosomes stall at a sequence just before the termination codon, slip back to the −1 frame, and then proceed into the downstream genes. By contrast, in the MLV-related genus, termination suppression is used for *pro* and *pol* expression. In this case the downstream genes are in the same frame as *gag*, and a small fraction of ribosomes insert an amino acid in place of the stop codon (48). By sequence organization WDSV resembles MLV in this regard, with *gag*, *pro*, and *pol* all in the same frame. However, we have not been able to identify an obvious stem-loop structure similar to that reported to promote termination suppression in MLV (12, 44). The predicted PR protein of WDSV has the conserved features found for all retroviral proteases. The active-site aspartate residue is embedded in the conserved sequence LVDTGA (position 2660) (32). Two other conserved motifs, GISA and GRDxL, are found downstream as expected (positions 2753 and 2870, respectively). On the basis of the sizes of other retroviral proteases, ca. 100 to 120 amino acid residues, and the positions of these conserved sequences, the N terminus of WDSV PR should be very near or just distal to the termination codon for Gag and the C terminus should be about 9 to 25 residues after the third conserved repeat above, probably at the sequence IVSF.

The WDSV *pol* gene encodes the polymerase signature sequence YMDD, as expected (position 3575). Other conserved amino acid residues typical of retroviral reverse transcriptases are found as well: IKK (position 3233), IDL (position 3377), and LPQG (position 3476). The first block of easily recognizable amino acids found in all RTs starts at the sequence ENQG (position 3146). Since we have not purified the WDSV RT and analyzed its N terminus, it is possible to make only a rough estimate of where the RT polypeptide begins, which would be perhaps 40 residues upstream of this block. The 3′ end of *pol* presumably encodes IN. Known retroviral integrases have three conserved functional domains, an N-terminal zinc finger domain, a catalytic domain, and a C-terminal domain (17). We have not been able to unambigously identify these domains from the WDSV *pol* sequence. However, there is a histidine- and cysteine-rich segment within the putative IN coding region (position 5060 to about 5260). The density of these residues suggests that this segment may be analogous to

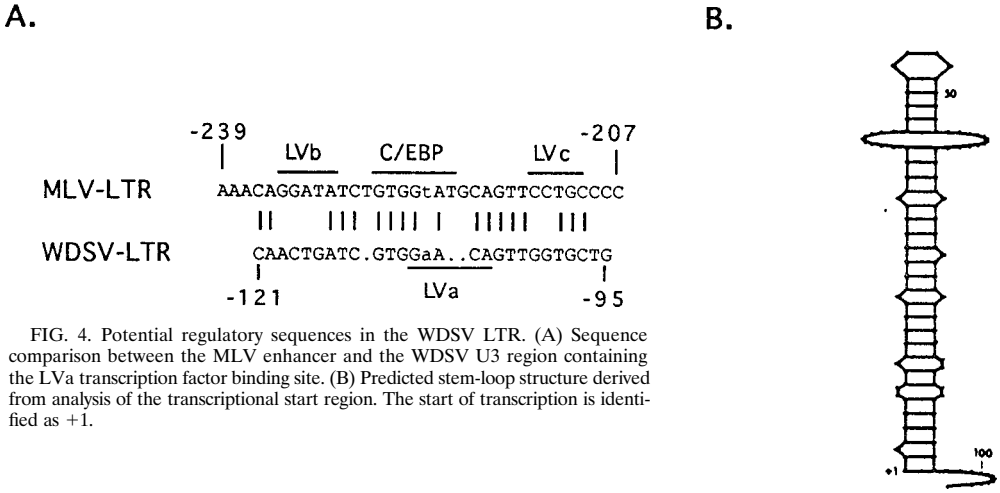

FIG. 4. Potential regulatory sequences in the WDSV LTR. (A) Sequence comparison between the MLV enhancer and the WDSV U3 region containing the LVa transcription factor binding site. (B) Predicted stem-loop structure derived from analysis of the transcriptional start region. The start of transcription is identified as +1.

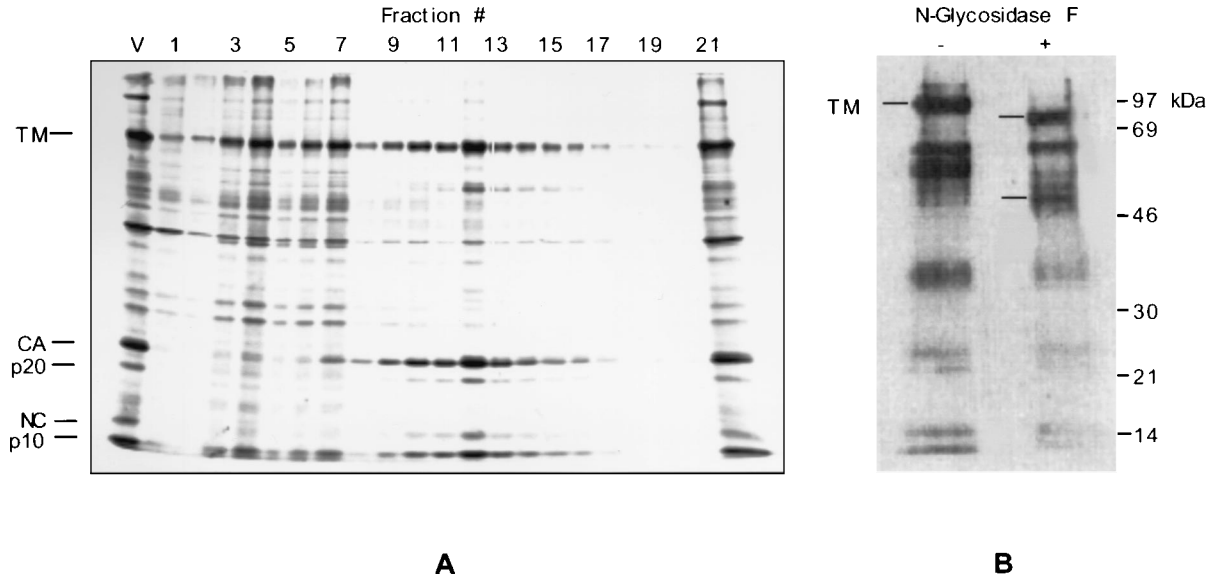

FIG. 5. WDSV polypeptides. (A) Partially purified virus after isopycnic banding in sucrose was further purified by rate zonal sedimentation at 18,000 rpm for 60 min in a 5 to 25% sucrose gradient. The collected fractions were analyzed by SDS-PAGE and Coomassie blue staining. Lanes: V, sample of virus loaded onto gradient; 1 to 19, fractions of gradient; 21, pellet fraction. The proteins identified by N-terminal sequence are marked. (B) Biotin-labeled WDSV from fractions 11 to 13 was digested with *N*-glycosidase F and then the proteins were separated by SDS-PAGE, blotted onto a membrane, and analyzed with streptavidin. Lanes: +, treated with glycosidase; −, parallel incubation without added glycosidase. TM, position of the TM component of Env.

the zinc finger domain of other retroviral integrase genes. Within this region and reminiscent of other IN proteins, two histidine residues occur that are separated by three amino acids (starting at position 5081), followed by two cysteine residues separated by two amino acids (starting at position 5183).

***env* gene.** The *env* gene of retroviruses is invariably downstream of *pol*. In many viruses these genes have a short overlapping region, but in some complex retroviruses accessory genes or portions of genes are sandwiched between *pol* and *env* and may overlap *env*. Being destined for the cell surface, retroviral Env proteins are synthesized with a signal sequence, extruded into the lumen of the endoplasmic reticulum, and transported through the Golgi apparatus to the plasma membrane. In all viruses Env becomes modified by N-linked glycosylation and becomes cleaved in the Golgi into the mature surface (SU) and transmembrane (TM) protein components, which remain associated in a complex both at the cell membrane and after incorporation into the virion. In all known cases, TM is anchored in the membrane by a single membrane-spanning segment near its C terminus.

We presumed originally that in WDSV virions the prominent p90 polypeptide was SU, on the basis of its sensitivity to trypsin. However, N-terminal sequencing showed p90 to start (N terminus, GIGLHSTN; predicted sequence at position 7384, GIGLHSTN) 469 amino acid residues from the beginning of the *env* ORF (position 5974), and thus we infer that it is TM (Fig. 2). Consistent with this inference is the observation that starting three residues upstream of the GIGL sequence is a series of contiguous basic residues, as found in other retroviral Env proteins near the cleavage site which separates SU from TM. The observation that p90 forms a relatively sharp band in SDS-polyacrylamide gels suggested that it carries few sugar modifications. Indeed, the sequence predicts only four N-linked glycosylation sites. Treatment of virions with *N*-glycosidase F, which cleaves N-linked oligosaccharides at the asparagine residue where they are attached to the polypeptide, reduced the apparent size of this polypeptide to approximately 80 kDa (Fig. 5B), consistent with the predicted size of 755 amino acid residues. At the very C terminus of the TM polypeptide is an extremely hydrophobic stretch of 65 hydrophobic residues containing only a single charged residue (Fig. 1 and 2). By its position and hydrophobicity this segment of TM is predicted to be or to harbor the membrane anchor domain.

We have not been able to unambiguously identify the predicted SU protein in WDSV virions. Assuming cleavage of a leader segment of about 23 amino acid residues (Fig. 2), SU should be about 446 residues in length. The sequence predicts 10 N-linked glycosylation sites, and hence this protein should migrate on SDS-PAGE as a diffuse band with an apparent mass larger than a protein of 50 kDa, perhaps 60 to 80 kDa. Coomassie-stained gels showed several minor polypeptides in this region, which were observed to be somewhat variable among different preparations of virus. Biotin labeling designed to be selective for surface proteins revealed two prominent bands in the size range expected for SU. The lower of these was sensitive to digestion with *N*-glycosidase F, with the digested polypeptide migrating at the position predicted for unglycosylated SU (Fig. 5B). If indeed this is SU, then the levels of this protein are at least 10-fold lower than those for TM. It is well-known that SU proteins, which in most retroviruses are bound noncovalently to TM, are easily lost by "shedding" upon purification or storage. This phenomenon could be responsible for the apparent dearth of SU in WDSV.

3′ ORFs. Distal to the *env* gene in WDSV are two nonoverlapping ORFs, *orf-A* (start at position 9674) and *orf-B* (start at position 10570) (Fig. 1 and 2). Assuming that translation is initiated at the first Met residue in each case, the sizes of these proteins would be 297 and 306 residues, respectively. However, given that splicing may lead to rearrangement of RNA segments, it is possible that the actual sequences of these proteins are not identical to those predicted by the nucleotide sequence. The putative initiating AUG codon for *orf-A* begins 23 nucleotides downstream of the *env* termination codon, and only 2 nucleotides separate *orf-A* and *orf-B* (Fig. 2). The products of these genes do not show significant sequence similarity with

entries in the database. We note that both proteins have stretches of acidic residues near their C termini. The last 40 amino acids of *orf-A* include 11 Asp or Glu residues and no positively charged amino acids. Similarly, the last 21 residues of *orf-B* include 8 Asp or Glu residues without positively charged amino acids. Acidic domains are found in numerous transcription factors, for example the viral transcription regulators VP16 of herpes simplex virus (31) and Taf of HSRV (26). By analogy with the function of accessory genes in other complex retroviruses, we surmise that the product of *orf-A* or *orf-B* may regulate viral gene expression.

DISCUSSION

The presence of ORFs other than *gag*, *pol*, and *env* identifies WDSV as a complex retrovirus (6). In all complex retroviruses, gene expression is regulated at the level of transcription and in the best-studied cases it is regulated at the level of RNA processing or transport. The interplay of *cis*-acting sequences in the viral DNA and RNA with products of viral accessory genes, as well as with host proteins, leads to the pattern of viral gene expression, which may vary according to growth conditions, cell type, and stage of infection. It is likely that in WDSV *cis*-acting sequences and *trans*-acting proteins also regulate gene expression and replication. The presence of a glucocorticoid response element in U3 is interesting, since tumor appearance and regression correlate with the reproductive cycle of the host, which is driven by steroids. In the HTLV-BLV group genus, transactivation by the viral Tax protein is mediated by a sequence of 21 bp present three times in U3 (7, 39). Conceivably, the 15-nucleotide repeat appearing three times upstream of the TATA box could play a similar role. In some lentiviruses, exemplified by HIV-1, transactivation by the viral Tat protein is mediated by a sequence in R, the TAR element, which is part of a stem-loop structure at the 5′ end of the RNA (11). The fact that the beginning of the WDSV RNA can also be drawn as a stable hairpin leads to the speculation that WDSV employs a similar mechanism of transactivation. These several similarities that WDSV may have with other complex retroviruses need to be examined experimentally.

In retrovirus replication a cellular tRNA acts as the primer for synthesis of the first or minus strand of DNA. In other retroviruses tRNAs for tryptophan, lysine, proline, or glutamine are used for this purpose and in general the tRNA primer is specific to the virus genus (20). Although the sequences of tRNAs from the walleye (*Stizostedion vitreum*) are not known, the similarity between the WDSV primer-binding site and the 3′ portion of $tRNA^{His}$ in insects and mammals leaves little doubt that this is the primer used by WDSV. No other retrovirus has yet been reported to have a $tRNA^{His}$ primer-binding site. Plus-strand priming in retroviruses is mediated by the polypurine sequence proximal to the 3′ LTR. In position and sequence the WDSV 3′ polypurine tract is similar to that in other retroviruses. However, in WDSV additional polypurine tracts are apparent, as they are in lentiviruses and spumaviruses (4, 19), in the 3′ portion of the *pol* gene. The sequences starting at nucleotides 4923 and 5101 differ from the 3′ polypurine tract by only one base. The position of these two sequences correlates approximately with the single-strand-specific nuclease site we reported previously for unintegrated WDSV DNA in spring tumors (21). By analogy with other complex retroviruses that utilize an additional priming site in *pol*, we hypothesize that the two extra polypurine tracts in WDSV also function in synthesis of the proviral DNA.

Of the three gene products common to all retroviruses—Gag, Pol, and Env—perhaps the WDSV TM component of Env is the most unusual. At more than 700 amino acid residues it is more than double the size of equivalent proteins in prototypic viruses like HIV (about 345 residues), RSV (about 200 residues), or MLV (about 170 residues). Only in primate spumaviruses are such large Env proteins found. TM proteins from well-characterized retroviruses have a single transmembrane segment of about 20 amino acid residues and a cytoplasmic domain of ca. 20 to 200 amino acid residues. In WDSV TM, the C-terminal sequence of this protein is extraordinarily hydrophobic. There are no charged residues in the last 48 amino acids of TM, and 40 of these 48 are either Leu, Ile, Phe, Met, Val, or Tyr. The amino acid 49 residues from the end is a Lys, and upstream of this is another stretch of 17 uncharged residues, 10 of these being hydrophobic. Hence we speculate that the last ca. 65 amino acids of TM all lie in the lipid bilayer surrounding the virion, with the possible exception of the single Lys residue. To our knowledge there is no precedent in retroviruses for a membrane anchor of this size, which could span the membrane two or perhaps three times. Also unprecedented in naturally occurring retroviruses is what would appear to be a complete absence of a cytoplasmic domain of TM. Another noteworthy aspect of TM is the amount of this protein in purified virions. Although we have not attempted to perform careful quantitations, it appears from the gels that TM is at least as heavily stained as the Gag protein CA. Given the predicted size differences and the assumption that Coomassie staining is not affected by sequence, the apparent molar ratio of Env to Gag in virions thus would be at least 1/4. Although in other model systems the amount of Env observed varies widely, few retroviruses have been described with this much virion-associated Env. The significance of these observations remains to be explored.

Because of the universality of *gag*, *pol*, and *env* genes in retroviruses, it is not difficult to predict the identity, size, and function of the protein products of these genes with reasonable precision. Predictions about the products of the other WDSV ORFs are much less certain. However, since members of the *Spumavirus* genus (27), *Lentivirus* genus (10), and the HTLV-BLV group genus (7, 15, 35) of retroviruses code for transactivators in the 3′ region of their genomes, it seems likely that the WDSV Orf-A and/or the Orf-B protein will be found to function similarly. In view of the complicated multiple splicing that has been observed in lentiviruses (10) and spumaviruses (27, 33), it may be that the products encoded mainly by *orf-A* and *orf-B* contain additional sequences contributed by other regions of the genome. Functional analysis of *orf-A* and *orf-B* by transfection and sequence analysis of WDSV subgenomic mRNAs will be needed to address these questions. Regardless of their exact structure, Orf-A and/or Orf-B can be hypothesized to be involved as causal agents in the induction of WDS. The observation that small subgenomic viral RNAs are present in walleye tumors taken in the spring supports this hypothesis, which is based largely on the models for tumor induction by the BLV and HTLV type 1 Tax proteins (28) and by the HIV Tat protein (41). However, it is important to emphasize that in these latter cases the tumors arising naturally as a consequence of infection rarely show evidence of expression of Tax or Tat. An alternative hypothesis that has not been examined is that the WDSV provirus must integrate into a specific site to induce tumors, as is the case for those avian leukemia viruses and MLVs that do not carry oncogenes.

We are aware of no precedent in retrovirology by which a biologically significant protein is encoded in the leader segment of the RNA analogous to Orf-C in WDSV. The avian sarcoma and leukemia viruses have three extremely short ORFs in the leader, and translation of at least one of these is

essential for packaging of viral RNA (8). However, in this case it appears not to be the translational products per se, but rather the event of translation that is important for the virus. One would presume by analogy with other retroviruses, all of which use the unspliced genomic RNA as a messenger for Gag, that the same ribosomes that translate *orf-C* continue on into *gag*. Alternatively, it is conceivable that the *orf-C* sequences are removed by splicing. It will be important to determine if the Orf-C protein actually is produced in infected cells and if splicing plays a role in the expression of Gag protein.

The several criteria used for phylogenetic comparisons of retroviruses do not allow a straightforward assignment of the closest relatives for WDSV, which is the first fish retrovirus completely characterized at the sequence level. Perhaps the most commonly used, as well as the most quantitative, parameter to suggest evolutionary relationships is the amino acid sequence of the conserved portions of RT. A stretch of about 200 amino acid residues can be recognized in the RTs of all retroviruses and retrotransposons (9). Though somewhat less well-conserved, major parts of this sequence are evident in more distantly related reverse transcriptases as well. We compared the predicted amino acid sequence of WDSV RT in the seven nearly contiguous blocks of sequence, comprising a total of 178 residues, described by Xiong and Eickbush (45), with RTs of a number of prototypic viruses in the seven recognized genera of retroviruses (Fig. 6) (5). This comparison places WDSV closest to viruses in the MLV-related virus genus. In quantitative terms (Table 1), the similarity index of WDSV RT and MLV RT over this stretch of 178 residues is 45%. If the same comparison is made with members of other retrovirus genera, this index ranges from 27% to 39%. This sequence similarity suggests that MLV-related viruses may share with WDSV a recent ancestor. Consistent with this idea, when the "tether," RNase H, and IN portions of *pol* (25) are used to search the database, the *pol* regions of murine and other mammalian type C viruses consistently score as the most similar. Nevertheless, more extensive analyses of these and other WDSV sequences will be needed to investigate this possible phylogenetic relationship in greater depth. In our judgment, the extent of similarity of WDSV with MLV should not be considered sufficient to warrant assignment of WDSV to the MLV-related virus genus. For viruses with this degree of similarity in RT sequence there are precedents both for inclusion in the same genus and for assignment to a different genus. For example, in the sample of viruses chosen here, visna virus RT is only 46% similar to the RT of some lentiviruses and nevertheless is considered a bona fide member of this genus. By contrast, RSV RT is 49% similar to the RTs of both mouse mammary tumor virus (MMTV) and Mason-Pfizer monkey virus (MPMV), all three of which are considered to represent

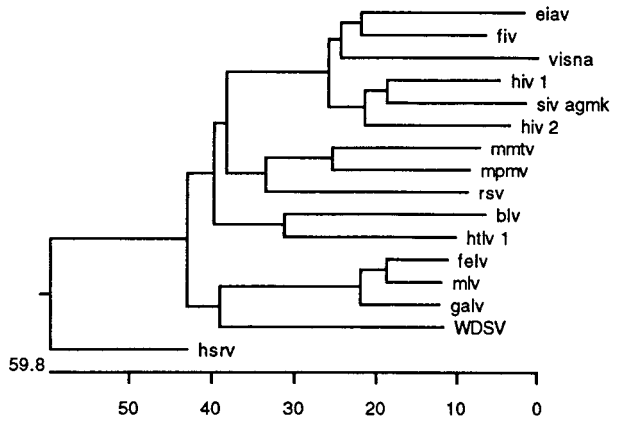

FIG. 6. Phylogenetic tree. Seven stretches of amino acids in reverse transcriptase, totally 178 residues, were analyzed and compared among the viruses shown, using software from DNAStar (36).

separate genera. Indeed MPMV and MMTV RTs themselves are 64% similar to each other. In all of these cases criteria other than reverse transcriptase sequence were considered in assignment of the viruses to the same or distinct genera.

While some additional criteria for retrovirus classification also suggest that WDSV is most closely related to MLV, others do not. On the one hand, MLV and its relatives are the only retroviruses known to use termination suppression as a mechanism of expression of *pro* and *pol* genes, as predicted for WDSV. Furthermore, some of the sequence motifs in the U3 portion of these viruses also are suggestive of close evolutionary relatedness. Finally, of the retroviruses that have Cys-His motifs in the NC domain of Gag, most have two motifs, whereas MLV and WDSV have only one. On the other hand, WDSV so far is unique in utilizing a $tRNA^{His}$ primer, quite different from the $tRNA^{Pro}$ found in MLV. In its C-terminal portion, the sequence of the 20-amino-acid major homology region in the CA domain of WDSV Gag is unlike all others previously reported. The abundant unintegrated linear DNA in spring tumors, as well as the single nuclease-sensitive site and additional polypurine tracts in this DNA (21), is a feature found in lentiviruses and spumaviruses but not MLV. Perhaps most important, WDSV is a complex virus as defined by the presence of accessory genes, unlike MLV and its relatives, all of which lack accessory genes. Taken together, these considerations lead us to propose WDSV as the prototype of a new genus. This genus and the MLV-related virus genus may share a recent ancestor, and if so it seems most plausible that the WDSV accessory genes were acquired after the time of divergence from this ancestor. The detailed molecular characteriza-

TABLE 1. Similarities in reverse transcriptase sequence among retroviruses

	% Similarity[a]						
	MLV	RSV	MMTV	MPMV	Visna virus	HSRV	HTLV
WDSV	**45 (45–46)**	34	33	37	27 (27–32)	33	39 (31–39)
Visna virus	25	32	31	33	**(46–60)**	28	31
RSV	35		**49**	**49**			
MMTV				**64**			
HTLV type 1							(55)

[a] The viruses listed at the top are representative of the seven currently recognized genera of retroviruses (5). The selected numbers give the sequence pair distances, using the Clustal method with the PAM250 residue weight table, between the virus listed at the left and the virus listed at the top. The numbers in parentheses represent the range of values obtained when the virus at the left is compared with the other members of the same genus that are given in Fig. 6. For visna virus (lentiviruses), the additional genus members compared are equine infectious anemia virus, HIV-1, HIV-2, and simian immunodeficiency virus SIV_{agmk}. For MLV, the additional genus members compared are gibbon ape leukemia virus and feline leukemia virus. For HTLV type 1, the additional member is BLV. The figures in boldtype are discussed in the text.

tion of other fish retroviruses will be needed to establish if WDSV has features that are diagnostic for fish retroviruses in general or if it is simply idiosyncratic.

ACKNOWLEDGMENTS

This work was supported by USPHS grant CA-53622 to J.W.C and NSERC grant 0GP0138236 to D.M. S.K.F. was supported by USPHS Training Grant GM-07273.

REFERENCES

1. **Anders, K., and M. Yoshimizu.** 1994. Role of viruses in the induction of skin tumours and tumour-like proliferations of fish. Dis. Aquat. Org. **19:**215–232.
2. **Blum, H. E., J. D. Haris, P. Ventura, K. Staskus, E. Retzel, and A. T. Haase.** 1985. Synthesis in cell culture of the gapped linear duplex DNA of the slow virus visna. Virology **142:**270–277.
3. **Bowser, P. R., M. J. Wolfe, J. L. Forney, and G. A. Wooster.** 1988. Seasonal prevalence of skin tumors from walleye (*Stizostedion vitreum*) from Oneida Lake, New York. J. Wildl. Dis. **24:**292–298.
4. **Charneau, P., and F. Clavel.** 1991. A single-stranded gap in human immunodeficiency virus unintegrated linear DNA defined by a central copy of the polypurine tract. J. Virol. **65:**2415–2421.
5. **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.
6. **Cullen, B. R.** 1991. Human immunodeficiency virus as a prototypic complex retrovirus. J. Virol. **65:**1053–1056.
7. **Derse, D.** 1987. Bovine leukemia virus transcription is controlled by virus-encoded *trans*-acting response elements. J. Virol. **61:**2462–2471.
8. **Donze, O., and P. F. Spahr.** 1992. Role of the open reading frames of Rous sarcoma virus leader RNA in translation and genome packaging. EMBO J. **11:**3747–3757.
9. **Doolittle, R. F., D. F. Feng, M. A. McClure, and M. S. Johnson.** 1990. Retrovirus phylogeny and evolution. *In* R. Swanstrom and P. K. Vogt (ed.), Retroviruses. Strategies of replication. Springer-Verlag, New York.
10. **Feinberg, M. B., R. F. Jarrett, A. Aldovini, R. C. Gallo, and F. Wong-Staal.** 1986. HTLV-III expression and production involve complex regulation at the levels of splicing and translation of viral RNA. Cell **46:**807–817.
11. **Feng, S., and E. C. Holland.** 1988. HIV-I tat *trans*-activation requires the loop sequences within tar. Nature (London) **334:**165–167.
12. **Feng, Y. X., H. Yuan, A. Rein, and J. G. Levin.** 1992. A bipartite signal for a readthrough 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.
13. **Flugel, R. M.** 1993. The molecular biology of the human spumavirus. *In* B. C. Cullen (ed.), Human retroviruses—frontiers in molecular biology. Oxford University Press, Great Britain.
14. **Franchini, G., F. Wong-Staal, and R. C. Gallo.** 1984. Human T-cell leukemia virus (HTLV-I) transcripts in fresh and cultured cells of patients with adult T-cell leukemia. Proc. Natl. Acad. Sci. USA **81:**6207–6211.
15. **Haseltine, W. A., J. Sodroski, R. Patarca, D. Briggs, D. Perkins, and F. Wong-Staal.** 1984. Structure of the 3′ terminal region of type II human T-lymphotropic virus: evidence for a new coding region. Science **225:**419–421.
16. **Henderson, L. E., R. Sowder, T. D. Copeland, G. Smythers, and S. Oroszlan.** 1984. Quantitative separation of murine leukemia proteins by reversed-phase high-pressure liquid chromatography reveals newly described *gag* and *env* cleavage products. J. Virol. **52:**492–500.
17. **Katz, R. A., and A. M. Skalka.** 1994. The retroviral enzymes. Annu. Rev. Biochem. **63:**133–173.
18. **Kozak, M.** 1987. An analysis of 5′-noncoding regions from 699 vertebrate messenger RNAs. Nucleic Acids Res. **15:**8125–8148.
19. **Kupiec, J., J. Tobaly-Tapiero, M. Canivet, M. Santillana-Hayat, R. M. Flugel, J. Peries, and R. Emanoil-Ravier.** 1988. Evidence for a gapped linear duplex DNA intermediate in the replicative cycle of human and simian spumaviruses. Nucleic Acids Res. **16:**9557–9565.
20. **Luciw, P. A., and N. J. Leung.** 1992. Mechanisms in retroviral replication, p. 159–298. *In* J. Levy (ed.), The Retroviridae, vol. 1. Plenum Press, New York.
21. **Martineau, D., P. R. Bowser, R. R. Renshaw, and J. W. Casey.** 1992. Molecular characterization of a unique retrovirus associated with a fish tumor. J. Virol. **66:**596–599.
22. **Martineau, D., P. R. Bowser, G. A. Wooster, and G. A. Armstrong.** 1990. Experimental transmission of a dermal sarcoma in fingerling walleyes (*Stizostedion vitreum vitreum*). Vet. Pathol. **27:**230–234.
23. **Martineau, D., P. R. Bowser, G. A. Wooster, and J. L. Forney.** 1990. Histologic and ultrastructural studies of dermal sarcoma of walleye (Pisces: *Stizostedion vitreum*). Vet. Pathol. **27:**340–346.
24. **Martineau, D., R. Renshaw, J. R. Williams, J. W. Casey, and P. R. Bowser.** 1991. A large unintegrated retrovirus DNA species present in a dermal tumor of walleye *Stizostedion vitreum*. Dis. Aquat. Org. **10:**153–158.
25. **McClure, M. A.** 1991. Evolution of retroposons by acquisition or deletion of retrovirus-like genes. Mol. Biol. Evol. **8:**835–856.
26. **Mergia, A., L. W. Rensaw-Gegg, M. W. Stout, R. Renne, and O. Herchenroeder.** 1993. Functional domains of the simian foamy virus type 1 transcriptional transactivator (Taf). J. Virol. **67:**4598–4604.
27. **Muranyi, W., and R. M. Flugel.** 1991. Analysis of splicing patterns of human spumaretrovirus by polymerase chain reaction reveals complex RNA structures. J. Virol. **65:**727–735.
28. **Nerenberg, M., S. H. Hinrichs, R. K. Reynolds, G. Khoury, and G. Jay.** 1987. The *tat* gene of human T-lymphotropic virus type 1 induces mesenchymal tumors in transgenic mice. Science **237:**1324–1329.
29. **Papas, T. S., J. E. Dahlberg, and R. A. Sonstegard.** 1976. Type C virus in lymphosarcoma in northern pike (*Esox lucius*). Nature (London) **261:**506–508.
30. **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.
31. **Ptashne, M., and A. A. F. Gann.** 1990. Activators and targets. Nature (London) **335:**329–331.
32. **Rao, J. K. M., J. W. Erickson, and A. Wlodawer.** 1991. Structural and evolutionary relationships between retroviral and eucaryotic aspartic proteinases. Biochemistry **30:**4663–4671.
33. **Renshaw, R. W., and J. W. Casey.** 1994. Transcriptional mapping of the 3′ end of the bovine synctial virus genome. J. Virol. **68:**1021–1028.
34. **Renshaw, R. W., M. A. Gonda, and J. W. Casey.** 1991. Structure and transcriptional status of bovine syncytial virus in cytopathic infections. Gene **105:**179–184.
35. **Rice, N. R., S. L. Simek, G. C. Dubois, S. D. Showalter, R. V. Gilden, and R. M. Stephens.** 1987. Expression of bovine leukemia virus X region in virus-infected cells. J. Virol. **61:**1577–1585.
36. **Saitou, N., and M. Nei.** 1987. The neighbor joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:**406–525.
37. **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.
38. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA **74:**5463–5467.
39. **Shimotohno, K., M. Takano, T. Teruuchi, and M. Miwa.** 1986. Requirement of multiple copies of a 21-nucleotide sequence in the U3 regions of human T-cell leukemia virus type I and type II long terminal repeats for *trans*-acting activation of transcription. Proc. Natl. Acad. Sci. USA **83:**8112–8116.
40. **Speck, N. A., and D. Baltimore.** 1987. Six distinct nuclear factors interact with the 75-base-pair repeat of the Moloney murine leukemia virus enhancer. Mol. Cell. Biol. **7:**1101–1110.
41. **Vogel, J., S. H. Hinrichs, R. K. Reynolds, P. A. Luciw, and G. Jay.** 1988. The HIV tat gene induces dermal lesions resembling Kaposi's sarcoma in transgenic mice. Nature (London) **335:**606–611.
42. **Walker, R.** 1969. Virus associated with epidermal hyperplasia in fish. Natl. Cancer Inst. Monogr. **31:**195–207.
43. **Wills, J. W., and R. C. Craven.** 1991. Form, function, and use of retroviral Gag proteins. AIDS **5:**639–654.
44. **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.
45. **Xiong, Y., and T. H. Eickbush.** 1990. Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J. **9:**3353–3362.
46. **Yamamoto, T., R. K. Kelley, and O. Nielsen.** 1985. Morphological differentiation of virus-associated skin tumors of walleye (*Stizostedion vitreum vitreum*). Fish Pathol. **20:**361–372.
47. **Yamamoto, T., R. D. MacDonald, D. C. Gillespie, and R. K. Kelly.** 1985. Viruses associated with lymphocystis and dermal sarcoma of walleye (*Stizostedion vitreum vitreum*). J. Fish Res. Board Can. **33:**2408–2419.
48. **Yoshhinaka, Y., I. Katoh, T. D. Copeland, and S. J. 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.