HIV Sequence Compendium 2006/2007

Editors

Thomas Leitner
Los Alamos National Laboratory

Brian Foley
Los Alamos National Laboratory

Beatrice Hahn
University of Alabama

Preston Marx
Tulane National Primate
Research Center

Francine McCutchan
Henry M. Jackson Foundation

John W. Mellors
University of Pittsburgh

Steven Wolinsky
Northwestern University

Bette Korber
Los Alamos National Laboratory

Project Officers
Geetha Bansal, James Bradac
Division of AIDS
National Institute of Allergy and Infectious Diseases

Los Alamos Database and Analysis Staff
Werner Abfalterer, Gayathri Athreya, Irina Maljkovic Berry, Charles Calef,
Brian Gaschen, Carla Kuiken, Jennifer Macke, Ming Zhang

This publication is being funded by the Division of AIDS, National Institute of Allergy and Infectious Diseases, through an interagency agreement with the U.S. Department of Energy.

Published by Theoretical Biology and Biophysics
Group T-10, Mail Stop K710
Los Alamos National Laboratory, Los Alamos, New Mexico 87545 U.S.A.

LA-UR-07-4826

http://hiv.lanl.gov

CONTENTS

Acknowledgments ii
Introduction iii
Maps of HIV and SIV Genomes iv
Landmarks of the Genome v

PART I. REVIEWS 1
Sequence Alignment in HIV Computational Analysis 2
Ana Abecasis, Anne-Mieke Vandamme, and Philippe Lemey
The Epidemiology of Transmission of Drug Resistant HIV-1 17
David A.M.C. van de Vijver, Annemarie M.J. Wensing, Charles A.B. Boucher
Search Tools in the HIV Databases 37
Jennifer Macke, Charles Calef, Karina Yusim, Robert Funkhouser, Thomas Leitner, James Szinger, Brian Gaschen, Werner Abfalterer, John Mokili, Brian Foley, Bette Korber, Carla Kuiken
Mutations in Retroviral Genes Associated with Drug Resistance 58
Shauna A. Clark, Charles Calef, John W. Mellors

PART II. HIV-1/SIVcpz COMPLETE GENOME ALIGNMENT
Introduction 159
Table of HIV-1/SIVcpz Sequences in the Nucleotide Alignment 161
Notes on full-length HIV-1/SIVcpz Sequences in the Nucleotide Alignment 164
Nucleotide Alignment of HIV-1/SIVcpz Complete Genomes 180

PART III. HIV-2SIVsmm COMPLETE GENOME ALIGNMENT
Introduction 355
Table of HIV-2/SIVsmm Sequences in the Nucleotide Alignment 356
Nucleotide Alignment of HIV-2/SIVsmm Complete Genomes 358

PART IV. PRIMATE LENTIVIRUS COMPLETE GENOME ALIGNMENT
Introduction 443
Table of PLV Sequences in the Nucleotide Alignment 447
Nucleotide Alignment of PLV Complete Genomes 449

PART V. HIV-1/SIVcpz AMINO ACID ALIGNMENT
Introduction 555
Amino Acid Alignments of HIV-1/SIVcpz 556

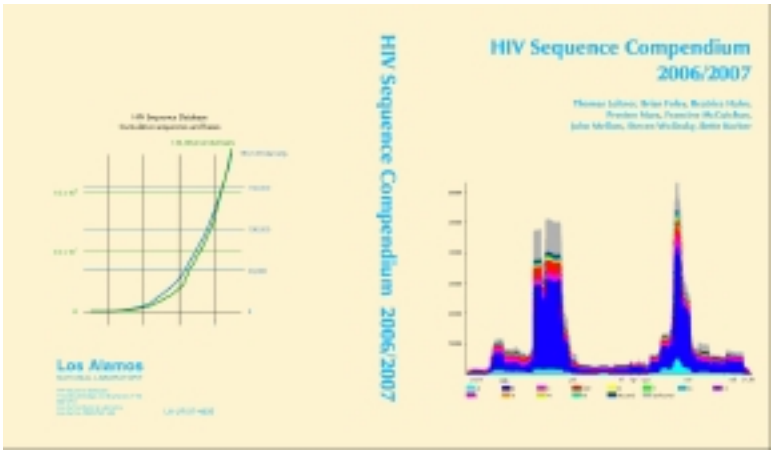
PART VI. HIV-2/SIVsmm AMINO ACID ALIGNMENT
Introduction 615
Table of HIV-2/SIVsmm Sequences in the Amino Acid Alignments 616
Amino Acid Alignments of HIV-2/SIVsmm 620

PART VII. PRIMATE LENTIVIRUS AMINO ACID ALIGNMENT
Introduction 647
Table of PLV Sequences in the Amino Acid Alignments 648
Amino Acid Alignments of PLV 651

Acknowledgments

The HIV Sequence Database and Analysis Project is funded by the Vaccine and Prevention Research Program of the AIDS Division of the National Institute of Allergy and Infectious Diseases (Dr. Geetha Bansal and Dr. James Bradac Project Officers) through an interagency agreement with the U.S. Department of Energy.

The Cover



The front cover displays a histogram of the number of sequences of each subtype versus the genomic position in HIV-1. Each pixel position along the genomic axis represents 15 base pairs. For each window of 15 base pairs the number of sequences of each subtype that occured in that window was tallied and plotted in different colors. The large peak on the right corresponds to the often-sequenced V3 region of Env. The tall double peak on the left corresponds to the protease and reverse transcriptase regions of Pol, which are of interest to sequencers because of mutations there that confer resistance to commonly used antiretroviral drugs. Each subtype (color) is plotted above the preceeding subtype which means the visible height of each color streak represents the amount of that subtype in that genomic region. Sequences of subtype B, shown as dark blue, continue to be the most commonly sequenced throughout the genome. Note that all circulating recombinant forms (CRFs) have been lumped into a single group shown as a bright red. Other non-CRF recombinants form a separate group colored navy blue. Sequences which have had no subtype assigned to them or do not fall into any of the subtypes defined in the illustration are plotted as gray and labeled as unknown. Because of the scale of the figure, rare subtypes like J and K cannot actually be discerned on the histogram. The graph on the back cover shows the growth in the number of sequences (blue) and bases (green) in the database over the last two decades. As of spring 2007 when this figure was drawn there were 190,749 sequences representing 136,606,452 total bases.

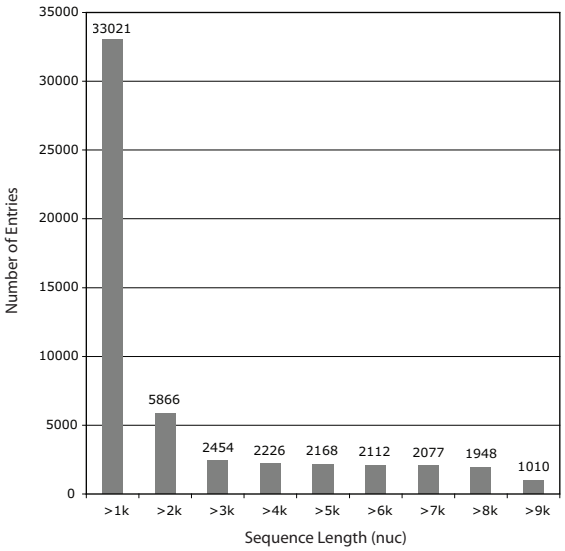
Citing this publication

This publication should be cited as *HIV Sequence Compendium 2006/2007*, Leitner T, Foley B, Hahn B, Marx P, McCutchan F, Mellors J, Wolinsky S, and Korber B, editors. 2007. Published by Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, LA-UR number 07-4826.

Introduction

This compendium is an annual printed summary of the data contained in the HIV sequence database. In these compendia we try to present a judicious selection of the data in such a way that it is of maximum utility to HIV researchers. Each of the alignments attempts to display the genetic variability within the different species, groups, and subtypes of the virus.

This compendium contains sequences published until the end of 2006.



The number of sequences in the HIV database is still increasing exponentially (see back cover). In total, at the time of printing, there were 195,633 sequences in the HIV Sequence Database.

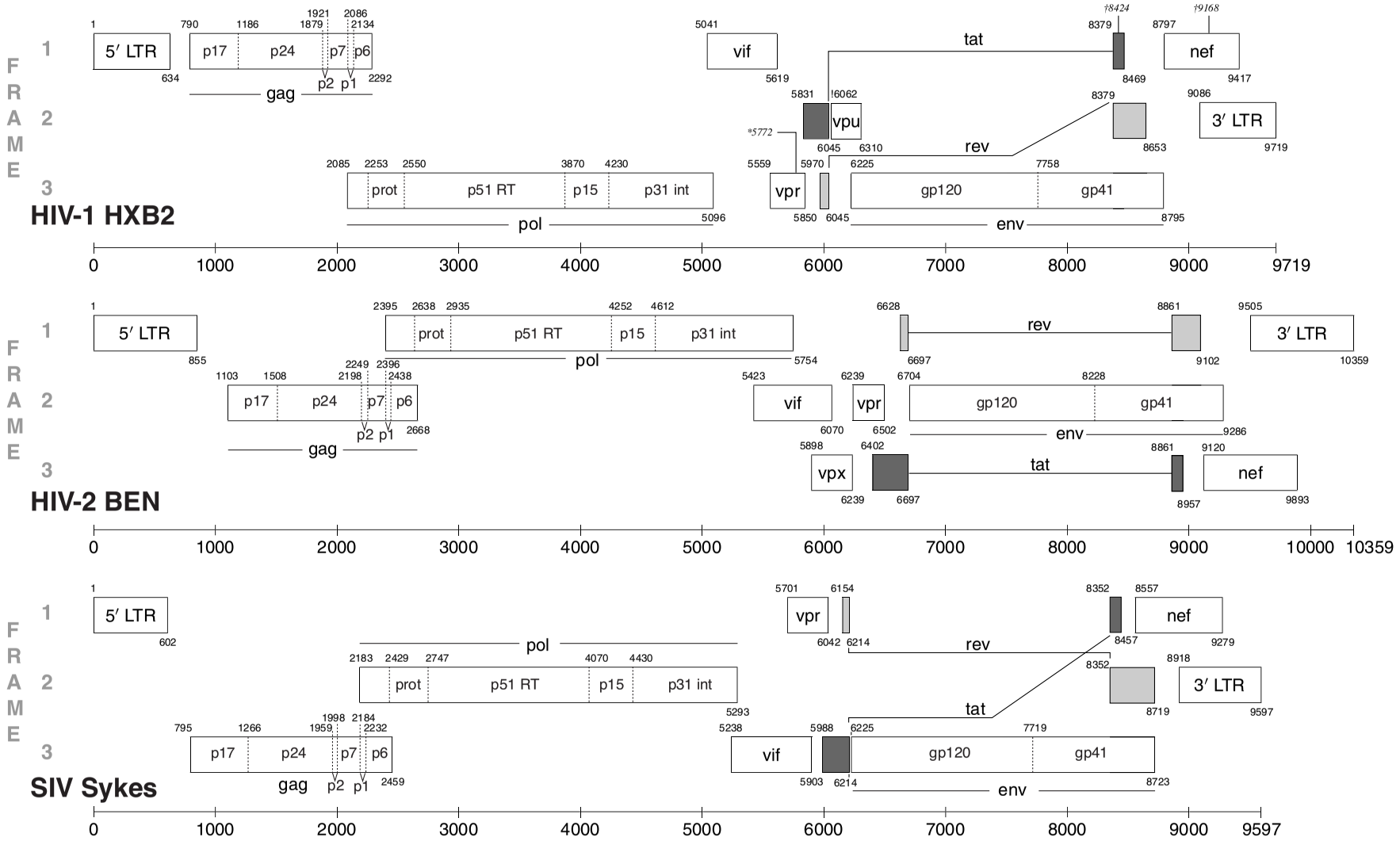
Since the previous compendium, the number of near-complete genomes increased again by approximately 50%. Hence, as last year, we have omitted many sequences in the compendium alignments. Sequences to be omitted are chosen so as to eliminate redundant sequences and maximize phylogenetic diversity. As always, tables with extensive background information gathered from the literature accompany the whole genome alignments. More complete versions of all alignments are available on our website:

http://www.hiv.lanl.gov/content/hiv-db/ALIGN_CURRENT/ALIGN-INDEX.html

Reprints of all reviews are available from our website in the form of both HTML and PDF files.
http://www.hiv.lanl.gov/content/hiv-db/REVIEWS/reviews.html

As always, we are open to complaints and suggestions for improvement. Inquiries and comments regarding the Compendium should be addressed to:

Dr. Thomas Leitner
Theoretical Biology and Biophysics, T-10
Mail Stop K710
Los Alamos National Laboratory
Los Alamos, NM 87545, USA
Ph: (505)-667-3898; fax: (505)-665-3493; e-mail: tkl@lanl.gov



Landmarks of the HIV-1, HIV-2, and SIV genomes. The gene start, indicated by the small number in the upper left corner of each rectangle normally records the position of the `a` in the `atg` start codon for that gene while the number in the lower right records the last position of the stop codon. For *pol*, the start is taken to be the first `t` in the sequence `ttttttag` which forms part of the stem loop that potentiates ribosomal slippage on the RNA and a resulting –1 frameshift and the translation of the gag-pol polyprotein. The *tat* and *rev* spliced exons are shown as shaded rectangles. In HXB2, **5772* marks position of frameshift in the *vpr* gene; !6062 indicates a defective `acg` start codon in *vpu*; *†8424*, and *†9168* mark premature stop codons in *tat* and *nef*. See Korber *et al.*, Numbering Positions in HIV Relative to HXB2CG, in *Human Retroviruses and AIDS, 1998* p. 102. Available from `http://www.hiv.lanl.gov/content/hiv-db/HTML/reviews/HXB2.html`

HIV/SIV PROTEINS

NAME	SIZE	FUNCTION	LOCALIZATION
Gag			
MA	p17	membrane anchoring; Env interaction; nuclear transport of viral core. (myristylated protein)	virion
CA	p24	core capsid	virion
NC	p7	nucleocapsid, binds RNA	virion
	p6	binds Vpr	virion
Pol			
Protease (PR)	p15	Gag/Pol cleavage and maturation	virion
Reverse transcriptase (RT),	p66 p51	reverse transcription, RNase H activity	virion
RNase H	p15		
Integrase (IN)	p31	DNA provirus integration	virion
Env	gp120/ gp41	external viral glycoproteins bind to CD4 and secondary receptors	plasma membrane, virion envelope
Tat	p16/p14	viral trancriptional transactivator	primarily in nucleolus/nucleus
Rev	p19	RNA transport, stability and utilization factor(phosphoprotein)	primarily in nucleolus/nucleus shuttling between nucleolus and cytoplasm
Vif	p23	promotes virion maturation and infectivity	cytoplasm (cytosol, membranes) virion
Vpr	p10–15	promotes nuclear localization of preintegration complex, inhibits cell division, arrests infected cells at G2/M	virion, nucleus (nuclear membrane?)
Vpu	p16	promotes extracellular release of viral particles;degrades CD4 in the ER;(phosphoproteinonly in HIV-1 and SIVcpz)	integral membrane protein
Nef	p25–p27	CD4 and class I downregulation (myristylated protein)	plasma membrane, cytoplasm(virion?)
Vpx	p12–16	Vpr homolog (not in HIV-1, only in HIV-2 and SIV)	virion (nucleus?)

LANDMARKS:

HIV GENOMIC STRUCTURAL ELEMENTS

LTR Long terminal repeat, the DNA sequence flanking the genome of integrated proviruses. It contains important regulatory regions, especially those for transcription initiation and polyadenylation.

TAR Target sequence for viral transactivation, the binding site for Tat protein and for cellular proteins; consists of approximately the first 45 nucleotides of the viral mRNAs in HIV-1 (or the first 100 nucleotides in HIV-2 and SIV.) TAR RNA forms a hairpin stem-loop structure with a side bulge; the bulge is necessary for Tat binding and function.

RRE Rev responsive element, an RNA element encoded within the env region of HIV-1. It consists of approximately 200 nucleotides (positions 7327 to 7530 from the start of transcription in HIV-1, spanning the border of gp120 and gp41). The RRE is necessary for Rev function; it contains a high affinity site for Rev; in all, approximately seven binding sites for Rev exist within the RRE RNA. Other lentiviruses (HIV-2, SIV, visna, CAEV) have similar RRE elements in similar locations within env, while HTLVs have an analogous RNA element (RXRE) serving the same purpose within their LTR; RRE is the binding site for Rev protein, while RXRE is the binding site for Rex protein. RRE (and RXRE) form complex secondary structures, necessary for specific protein binding.

PE Psi elements, are a set of 4 stem-loop structures preceding and overlapping the Gag start codon which are the sites recognized by the cysteine histidine box, a conserved motif with the canonical sequence CysX2CysX4HisX4Cys, present in the Gag p7 MC protein. The Psi Elements are present in unspliced genomic transcripts but absent from spliced viral mRNAs.

SLIP A TTTTTT slippery site, followed by a stem-loop structure, is responsible for regulating the -1 ribosomal frameshift out of the Gag reading frame into the Pol reading frame.

CRS Cis-acting repressive sequences postulated to inhibit structural protein expression in the absence of Rev. One such site was mapped within the pol region of HIV-1. The exact function has not been defined; splice sites have been postulated to act as CRS sequences.

INS Inhibitory/Instability RNA sequences found within the structural genes of HIV-1 and of other complex retroviruses. Multiple INS elements exist within the genome and can act independently; one of the best characterized elements spans nucleotides 414 to 631 in the *gag* region of HIV-1. The INS elements have been defined by functional assays as elements that inhibit expression posttranscriptionally. Mutation of the RNA elements was shown to lead to INS inactivation and up regulation of gene expression.

GENES AND GENE PRODUCTS

GAG The genomic region encoding the capsid proteins (group specific antigens). The precursor is the p55 myristylated protein, which is processed to p17 (MAtrix), p24 (CApsid), p7 (NucleoCapsid), and p6 proteins, by the viral protease. Gag associates with the plasma membrane where the virus assembly takes place. The 55 kDa Gag precursor is called assemblin to indicate its role in viral assembly.

POL The genomic region encoding the viral enzymes protease, reverse transcriptase and integrase. These enzymes are produced as a Gag-Pol precursor polyprotein, which is processed by the viral protease; the Gag-Pol precursor is produced by ribosome frameshifting near the 3' end of *gag*.

ENV Viral glycoproteins produced as a precursor (gp160) which is processed to give a noncovalent complex of the external glycoprotein gp120 and the transmembrane glycoprotein gp41. The mature gp120-gp41 proteins are bound by non-covalent interactions and are associated as a trimer on the cell surface. A substantial amount of gp120 can be found released in the medium. gp120

contains the binding site for the CD4 receptor, and the seven transmembrane domain chemokine receptors that serve as co-receptors for HIV-1.

TAT Transactivator of HIV gene expression. One of two essential viral regulatory factors (Tat and Rev) for HIV gene expression. Two forms are known, Tat-1 exon (minor form) of 72 amino acids and Tat-2 exon (major form) of 86 amino acids. Low levels of both proteins are found in persistently infected cells. Tat has been localized primarily in the nucleolus/nucleus by immunofluorescence. It acts by binding to the TAR RNA element and activating transcription initiation and elongation from the LTR promoter, preventing the 5' LTR AATAAA polyadenylation signal from causing premature termination of transcription and polyadenylation. It is the first eukaryotic transcription factor known to interact with RNA rather than DNA and may have similarities with prokaryotic anti-termination factors. Extracellular Tat can be found and can be taken up by cells in culture.

REV The second necessary regulatory factor for HIV expression. A 19 kD phosphoprotein, localized primarily in the nucleolus/nucleus, Rev acts by binding to RRE and promoting the nuclear export, stabilization and utilization of the viral mRNAs containing RRE. Rev is considered the most functionally conserved regulatory protein of lentiviruses. Rev cycles rapidly between the nucleus and the cytoplasm.

VIF Viral infectivity factor, a basic protein of typically 23 kD. Promotes the infectivity but not the production of viral particles. In the absence of Vif the produced viral particles are defective, while the cell-to-cell transmission of virus is not affected significantly. Found in almost all lentiviruses, Vif is a cytoplasmic protein, existing in both a soluble cytosolic form and a membrane-associated form. The latter form of Vif is a peripheral membrane protein that is tightly associated with the cytoplasmic side of cellular membranes. In 2003, it was discovered that Vif prevents the action of the cellular APOBEC-3G protein which deaminates DNA:RNA heteroduplexes in the cytoplasm.

VPR Vpr (viral protein R) is a 96-amino acid (14 kD) protein, which is incorporated into the virion. It interacts with the p6 Gag part of the Pr55 Gag precursor. Vpr detected in the cell is localized to the nucleus. Proposed functions for Vpr include the targeting the nuclear import of preintegration complexes, cell growth arrest, transactivation of cellular genes, and induction of cellular differentiation. In HIV-2, SIV-SMM, SIV-RCM, SIV-MND-2 and SIV-DRL the Vpx gene is apparently the result of a Vpr gene duplication event, possibly by recombination.

VPU Vpu (viral protein U) is unique to HIV-1, SIVcpz (the closest SIV relative of HIV-1), SIV-GSN, SIV-MUS, SIV-MON and SIV-DEN. There is no similar gene in HIV-2, SIV-SMM or other SIVs. Vpu is a 16-kd (81-amino acid) type I integral membrane protein with at least two different biological functions: (a) degradation of CD4 in the endoplasmic reticulum, and (b) enhancement of virion release from the plasma membrane of HIV-1-infected cells. Env and Vpu are expressed from a bicistronic mRNA. Vpu probably possesses an N-terminal hydrophobic membrane anchor and a hydrophilic moiety. It is phosphorylated by casein kinase II at positions Ser52 and Ser56. Vpu is involved in Env maturation and is not found in the virion. Vpu has been found to increase susceptibility of HIV-1 infected cells to Fas killing.

NEF A multifunctional 27-kd myristylated protein produced by an ORF located at the 3' end of the primate lentiviruses. Other forms of Nef are known, including nonmyristylated variants. Nef is predominantly cytoplasmic and associated with the plasma membrane via the myristyl residue linked to the conserved second amino acid (Gly). Nef has also been identified in the nucleus and found associated with the cytoskeleton in some experiments. One of the first HIV proteins to be produced in infected cells, it is the most immunogenic of the accessory proteins. The *nef* genes of HIV and SIV are dispensable *in vitro*, but are essential for efficient viral spread and disease progression *in vivo*. Nef is necessary for the maintenance of high virus loads and for the development of AIDS in macaques, and viruses with defective Nef have been detected in some HIV-1 infected long term survivors. Nef downregulates CD4, the primary viral receptor, and

MHC class I molecules, and these functions map to different parts of the protein. Nef interacts with components of host cell signal transduction and clathrin-dependent protein sorting pathways. It increases viral infectivity. Nef contains PxxP motifs that bind to SH3 domains of a subset of Src kinases and are required for the enhanced growth of HIV but not for the downregulation of CD4.

VPX A virion protein of 12 kD found in HIV-2, SIV-SMM, SIV-RCM, SIV-MND-2 and SIV-DRL and not in HIV-1 or other SIVs. This accessory gene is a homolog of HIV-1 vpr, and viruses with Vpx carry both vpr and vpx. Vpx function in relation to Vpr is not fully elucidated; both are incorporated into virions at levels comparable to Gag proteins through interactions with Gag p6. Vpx is necessary for efficient replication of SIV-SMM in PBMCs. Progression to AIDS and death in SIV-infected animals can occur in the absence of Vpr or Vpx. Double mutant virus lacking both vpr and vpx was attenuated, whereas the single mutants were not, suggesting a redundancy in the function of Vpr and Vpx related to virus pathogenicity.

STRUCTURAL PROTEINS/VIRAL ENZYMES The products of *gag*, *pol,* and *env* genes, which are essential components of the retroviral particle.

REGULATORY PROTEINS Tat and Rev proteins of HIV/SIV and Tax and Rex proteins of HTLVs. They modulate transcriptional and posttranscriptional steps of virus gene expression and are essential for virus propagation.

ACCESSORY OR AUXILIARY PROTEINS Additional virion and non-virion- associated proteins produced by HIV/SIV retroviruses: Vif, Vpr, Vpu, Vpx, Nef. Although the accessory proteins are in general not necessary for viral propagation in tissue culture, they have been conserved in the different isolates; this conservation and experimental observations suggest that their role *in vivo* is very important. Their functional importance continues to be elucidated.

COMPLEX RETROVIRUSES Retroviruses regulating their expression via viral factors and expressing additional proteins (regulatory and accessory) essential for their life cycle.