HIV Sequence Compendium 2013

Editors

Brian Foley
Los Alamos National Laboratory

Thomas Leitner
Los Alamos National Laboratory

Cristian Apetrei
University of Pittsburgh

Beatrice Hahn
University of Pennsylvania

Ilene Mizrachi
National Center for Biotechnology Information

James Mullins
University of Washington

Andrew Rambaut
University of Edinburgh

Steven Wolinsky
Northwestern University

Bette Korber
Los Alamos National Laboratory

Project Officer
Stuart Shapiro
Division of AIDS
National Institute of Allergy and Infectious Diseases

Los Alamos HIV Sequence Database and Analysis Staff
Werner Abfalterer, Mira Dimitrijevic, Shihai Feng, Will Fischer, Bob Funkhouser, Peter Hraber, Jennifer Macke, James J. Szinger, Hyejin Yoon

This publication is funded by the Division of AIDS, National Institute of Allergy and Infectious Diseases, through interagency agreement IAA Y1-AI-8309-1 "HIV/SIV Database and Analysis Unit" with the U.S. Department of Energy.

Published by
Theoretical Biology and Biophysics
Group T-6, Mail Stop K710
Los Alamos National Laboratory
Los Alamos, New Mexico 87545 U.S.A.

LA-UR-13-26007

http://www.hiv.lanl.gov/

HIV Sequence Compendium 2013

Published by
Theoretical Biology and Biophysics
Group T-6, Mail Stop K710
Los Alamos National Laboratory
Los Alamos, New Mexico 87545 U.S.A.

LA-UR-13-26007
Approved for public release; distribution is unlimited.

Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by Los Alamos National Security, LLC, for the National Nuclear Security Administration of the U.S. Department of Energy under contract DE-AC52-06NA25396.

This report was prepared as an account of work sponsored by an agency of the U.S. Government. Neither Los Alamos National Security, LLC, the U.S. Government nor any agency thereof, nor any of their employees make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by Los Alamos National Security, LLC, the U.S. Government, or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of Los Alamos National Security, LLC, the U.S. Government, or any agency thereof.

Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

This report was prepared as an account of work sponsored by NIH/NIAID/DAIDS under contract number IAA Y1-AI-8309-1 "HIV/SIV Database and Analysis Unit".

Contents

Contents **iii**

I Preface **1**
- I-1 Introduction 1
- I-2 Acknowledgements 1
- I-3 Citing the compendium or database 1
- I-4 About the PDF 1
- I-5 About the cover 2
- I-6 Genome maps 4
- I-7 HIV/SIV proteins 5
- I-8 Landmarks of the genome 6
- I-9 Amino acid codes 8
- I-10 Nucleic acid codes 8

II HIV-1/SIVcpz Complete Genomes **9**
- II-1 Introduction 9
- II-2 Annotated features 10
- II-3 Sequences 13
- II-4 Alignments 18

III HIV-2/SIV Complete Genomes **155**
- III-1 Introduction 155
- III-2 Annotated features 156
- III-3 Sequences 158
- III-4 Alignments 161

IV PLV Complete Genomes **227**
- IV-1 Introduction 227
- IV-2 Sequences 228
- IV-3 Alignments 231

V HIV-1/SIVcpz Proteins **311**
- V-1 Introduction 311
- V-2 Annotated features 312
- V-3 Sequences 314
- V-4 Alignments 320

VI HIV-2/SIV Proteins **373**
- VI-1 Introduction 373
- VI-2 Annotated features 374
- VI-3 Sequences 375
- VI-4 Alignments 378

VII PLV Proteins **405**
- VII-1 Introduction 405
- VII-2 Sequences 406
- VII-3 Alignments 416

I

Preface

I-1 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 before January 1, 2013. Hence, though it is published in 2013 and called the 2013 Compendium, its contents correspond to the 2012 curated alignments on our website.

The number of sequences in the HIV database is still increasing. In total, at the end of 2012, there were 518,163 sequences in the HIV Sequence Database, an increase of 12% since the previous year.

The number of near complete genomes (>7000 nucleotides) increased to 5013 by end of 2012. However, as in previous years, the compendium alignments contain only a small fraction of these. Included in the alignments are a small number of sequences representing each of the subtypes and the more prevalent circulating recombinant forms (CRFs) such as 01 and 02, as well as a few outgroup sequences (groups N, O, and P and SIV-CPZ). Of the rarer CRFs we included one representative each. A more complete version of all alignments is available on our website, `http://www.hiv.lanl.gov/content/sequence/NEWALIGN/align.html`

Reprints are available from our website in the form of PDF files. As always, we are open to complaints and suggestions for improvement. Inquiries and comments regarding the compendium should be addressed to seq-info@lanl.gov.

I-2 Acknowledgements

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 (Stuart Shapiro, Project Officer) through interagency agreement IAA Y1-AI-8309-1 “HIV/SIV Database and Analysis Unit” with the U.S. Department of Energy.

I-3 Citing the compendium or database

The LANL HIV Sequence Database may be cited in the same manner as this compendium:

HIV Sequence Compendium 2013. Brian Foley, Thomas Leitner, Cristian Apetrei, Beatrice Hahn, Ilene Mizrachi, James Mullins, Andrew Rambaut, Steven Wolinsky, and Bette Korber editors. 2013. Publisher: Los Alamos National Laboratory, Theoretical Biology and Biophysics, Los Alamos, New Mexico. LA-UR-13-26007.

I-4 About the PDF

The complete *HIV Sequence Compendium 2013* is available in Adobe Portable Document Format (PDF) from our website, `http://www.hiv.lanl.gov/`. The PDF version is hypertext enabled and features ‘clickable’ table-of-contents, indexes, references and links to external web sites.

This volume is typeset using LaTeX.

I-5 About the cover

Human Immunodeficiency Virus Illustration

On this year's cover is a cutaway view of an HIV-1 virion purchased from Visual Science. It won the Illustration category of Visualization Challenge in 2010 (Nesbit and Norman, 2011), and has appeared on the covers of Nature Medicine (October 2010) and the IAVI Report (May-June 2011). For additional details about the highlighted features, references, and all inquiries about the use of this image, please visit `http://visualscience.ru/en/projects/hiv/illustrations/`.

Virus-derived molecules (shades of orange)

Glycosylated envelope trimer The envelope protein is synthesized as a precursor protein, gp160, which is cleaved by the cellular enzyme furin into gp41 and gp120 subunits. Each trimer of the envelope "spike" complex is composed of 3 molecules each of gp41 and gp120. The gp41 transmembrane protein anchors the spike to the lipid membrane. Envelope trimers are central to viral fusion and cell entry. The envelope spike is glycosylated by the attachment of carbohydrate moieties, depicted in dark orange.

Matrix – Gag p17 The HIV-1 gag gene encodes the p55 polyprotein, which initiates formation of new virions. During maturation, p55 is cleaved into 4 structural proteins: p17, p24, p6, and p7. A layer of matrix protein p17 (MA) lies just below the lipid bilayer. In addition to its structural role, p17 has various functions throughout the replicative cycle of the virus (reviewed by Fiorentini et al. (2006)).

Capsid – Gag p24 The Gag p24 (CA) protein forms a cone-shaped shell, the capsid, which encloses the viral genomic RNA. While most p24 molecules are involved in forming the mature capsid (solid line), some exist as free dimers (dashed line). For recent information about capsid structure, see Zhao et al. (2013).

Nucleocapsid – Gag p7 The nucleocapsid protein binds to RNA to protect it from degradation by nucleases. It is closely associated with p6 and CA.

Gag p6 This small protein binds avidly to Vpr and membranes, and is important for virion budding.

Integrase Integrase is encoded by the pol gene. It forms a homotetrameric enzyme complex that integrates viral DNA into the host cell's genomic DNA.

Viral RNA The viral genome is encoded as a single strand of RNA, which is packed inside the capsid. Two unpaired strands are present per virion.

Reverse transcriptase After cell entry, this pol-encoded protein transcribes viral RNA to DNA. It is a heterodimer, composed of p66 (polymerase + RNase H) and p51 (polymerase only). After polymerase transcribes the genome to DNA, RNase H degrades the RNA part of RNA-DNA complexes, leaving a single strand of viral DNA to be integrated into the host genome.

Vpr Viral Protein R is a 96-amino acid protein that performs several crucial functions in the HIV-1 replicative cycle. In HIV-2, the function of Vpr is shared by another protein, Vpx.

Host-derived molecules (shades of gray)

Host proteins are incorporated into the virion during budding. A few of these host proteins that are that are directly relevant to HIV biology are illustrated in the figure.

CD55 This host protein helps protect HIV from the membrane-attack complex by attenuating the host's complement system (Marschang et al., 1995).

HLA-DR1 Human leukocyte antigens (HLA) molecules are best known for their role in presenting extracellular antigens to helper T cells. DR1 is just one example of a human protein that may be present on HIV virions; the specific HLA molecules present will depend on the alleles carried by the host. In HIV infection, these HLA molecules are significant for their binding to CD4 receptors.

Lipid bilayer The viral membrane is derived from the host cell.

ICAM-1 The human Intercellular Adhesion Molecule-1 is a member of the immunoglobulin superfamily of adhesion proteins. Its presence on the virion increases infectivity.

Lys-tRNA and lysyl-tRNA synthetase The cellular protein lysyl-tRNA synthetase and the corresponding transfer RNA, Lys-tRNA, are found inside the capsid; these molecules initiate reverse transcription.

Cyclophilin This host protein binds to the viral capsid. Cyclophilin is an isomerase involved in protein folding, and is required for HIV infectivity.

Actin Host cell actin is commonly captured during HIV budding.

References

S. Fiorentini, E. Marini, S. Caracciolo, and A. Caruso. Functions of the HIV-1 matrix protein p17. *New Microbiol.*, 29(1):1–10, Jan 2006.

P. Marschang, J. Sodroski, R. Würzner, and M. P. Dierich. Decay-accelerating factor (CD55) protects human immunodeficiency virus type 1 from inactivation by human complement. *Eur. J. Immunol.*, 25 (1):285–290, Jan 1995.

J. Nesbit and C. Norman. 2010 visualization challenge. *Science*, 331 (6019):847–856, 18 Feb 2011.

G. Zhao, J. R. Perilla, E. L. Yufenyuy, X. Meng, B. Chen, J. Ning, J. Ahn, A. M. Gronenborn, K. Schulten, C. Aiken, and P. Zhang. Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics. *Nature*, 497(7451):643–646, 30 May 2013.

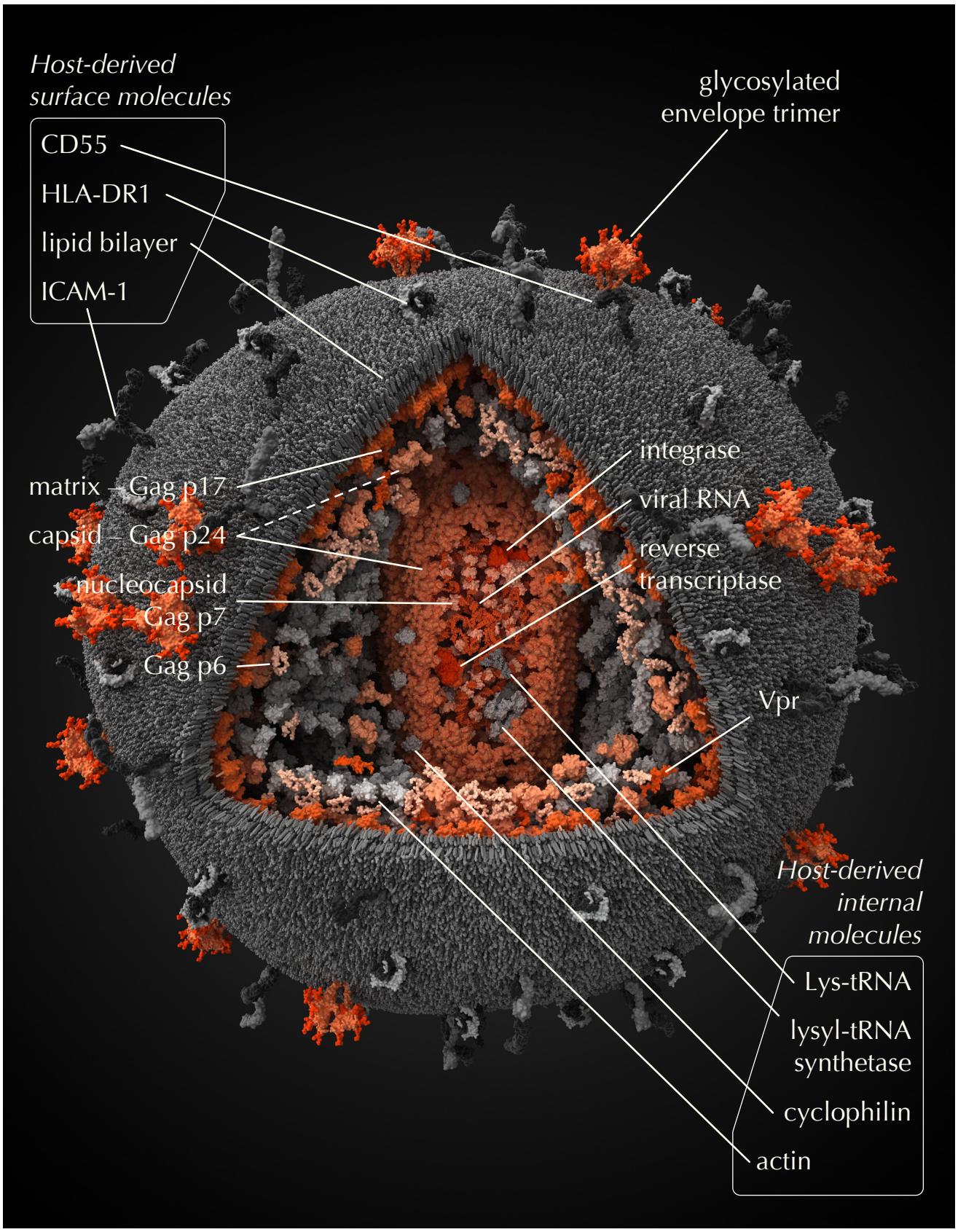
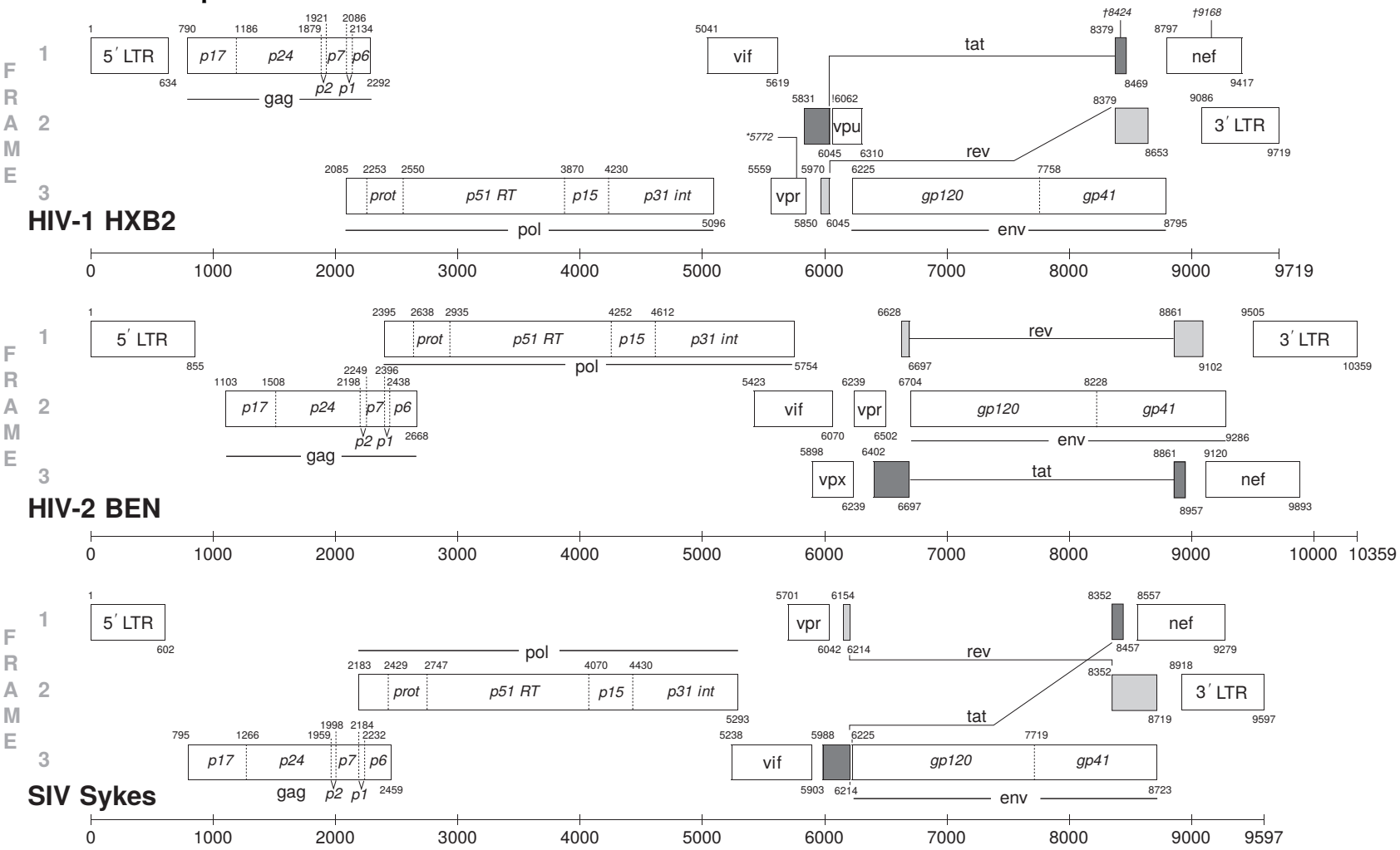


Figure I.1: A Human Immunodeficiency Virus.

I-6 Genome maps



Landmarks of the HIV-1, HIV-2, and SIV genomes. Open reading frames are shown as rectangles. 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 the position of a frameshift in the *vpr* gene caused by an "extra" `t` relative to most other subtype B viruses; !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/sequence/HIV/REVIEWS/HXB2.html`

I-7 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		virion
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 transcriptional transactivator	primarily in nucleolus/nucleus
Rev	p19	RNA transport, stability and utilization factor (phosphoprotein)	primarily in nuleolus/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; (phosphoprotein only in HIV-1 and SIVcpz)	integral membrane protein
Nef	p27-p25	CD4 and class I downregulation (myristylated protein)	plasma membrane, cytoplasm, (virion?)
Vpx	p12-16	Vpr homolog present in HIV-2 and some SIVs, absent in HIV-1	virion (nucleus?)
Tev	p28	tripartite tat-env-rev protein (also named Tnv)	primarily in nucleolus/nucleus

I-8 Landmarks of the genome

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, 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 NC 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, RNAse, 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 kDa phosphoprotein, localized primarily in the nucleolus/nucleus, Rev acts by binding to RRE and promoting the nuclear export, stabilization and utilization of the unspliced 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 kDa. 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 kDa) 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 kDa (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-kDa 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 kDa 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.

I-9 Amino acid codes

A	Alanine
B	Aspartic Acid or Asparagine
C	Cysteine
D	Aspartic Acid
E	Glutamic Acid
F	Phenylalanine
G	Glycine
H	Histidine
I	Isoleucine
K	Lysine
L	Leucine
M	Methionine
N	Asparagine
P	Proline
Q	Glutamine
R	Arginine
S	Serine
T	Threonine
V	Valine
W	Tryptophan
X	unknown or “other” amino acid
Y	Tyrosine
Z	Glutamic Acid or Glutamine
.	gap
-	identity
*	stop codon
#	incomplete codon

I-10 Nucleic acid codes

A	Adenine
C	Cytosine
G	Guanine
T	Thymine
U	Uracil
M	A or C
R	A or G
W	A or T
S	C or G
Y	C or T
K	G or T
V	A or C or G
H	A or C or T
D	A or G or T
B	C or G or T
N	unknown
.	gap
-	identity