HIV Sequence Compendium 2003

Editors

Thomas Leitner  
Los Alamos National Laboratory

Francine McCutchan  
Henry M. Jackson Foundation

Brian Foley  
Los Alamos National Laboratory

John W. Mellors  
University of Pittsburgh

Beatrice Hahn  
University of Alabama

Steven Wolinsky  
Northwestern University

Preston Marx  
ADARC

Bette Korber  
Los Alamos National Laboratory

Project Officer

James Bradac  
Division of AIDS  
National Institute of Allergy and Infectious Diseases

Los Alamos Database and Analysis Staff

Werner Abfalterer, Charles Calef, Brian Gaschen,  
Kristina Kommander, Dorothy Lang, 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 04-7420

http://hiv-web.lanl.gov
CONTENTS

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

PART I. REVIEWS .........................................................................................................................1

Vif and the Role of Antiviral Cytidine Deaminases in HIV-1 Replication ........................................2
Qin Yu, Nathaniel R. Landau, and Renate König

Forms and Function of Intracellular HIV DNA .................................................................14
Andreas Meyerhans, Tanja Breinig, Jean-Pierre Vartanian, and Simon Wain-Hobson

The Microbial Forensic Use of HIV Sequences ......................................................................22
Gerald H. Learn and James I. Mullins

Mutations in Retroviral Genes Associated with Drug Resistance ...........................................38
Urvi Parikh, Charles Calef, Shauna A. Clark, and John W. Mellors

PART II. HIV-1/SIVcpz COMPLETE GENOME ALIGNMENTS
Introduction ...........................................................................................................................123
Table of HIV-1/SIVcpz Sequences in the Nucleotide Alignment .................................................127
Notes on full-length HIV-1/SIVcpz Sequences in the Nucleotide Alignment ............................130
Nucleotide Alignment of HIV-1/SIVcpz Complete Genomes ...................................................146

PART III. HIV-2/SIVsmm COMPLETE GENOME ALIGNMENTS
Introduction ...........................................................................................................................319
Table of HIV-2/SIVsmm Sequences in the Nucleotide Alignment .................................................321
Nucleotide Alignment of HIV-2/SIVsmm Complete Genomes ...................................................322

PART IV. PRIMATE LENTIVIRUS COMPLETE GENOME ALIGNMENTS
Introduction ...........................................................................................................................405
Table of PLV Sequences in the Nucleotide Alignment ................................................................409
Nucleotide Alignment of PLV Complete Genomes ...................................................................411

PART V. HIV-1/SIVcpz AMINO ACID ALIGNMENTS
Introduction ...........................................................................................................................513
Explanation of annotation in the Amino Acid Alignments ..........................................................513
Amino Acid Alignments of HIV-1/SIVcpz ..............................................................................514

PART VI. HIV-2/SIVsmm AMINO ACID ALIGNMENTS
Introduction ...........................................................................................................................575
Table of HIV-2/SIVsmm Sequences in the Amino Acid Alignments ........................................576
Amino Acid Alignments of HIV-2/SIVsmm ............................................................................579

PART VII. PRIMATE LENTIVIRUS AMINO ACID ALIGNMENTS
Introduction ...........................................................................................................................609
Table of PLV Sequences in the Amino Acid Alignments ............................................................610
Amino Acid Alignments of PLV ..........................................................................................613
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. James Bradac, Project Officer) through an interagency agreement with the U.S. Department of Energy.

The Cover

This year’s cover is a photo of HIV infected T lymphocytes. The projections from the cell surface are microvilli. The HIV particles are seen among the microvilli as small, barely visible, rounded dots. Large clusters of virus can also be seen. The photo was taken by Dr. h.c. Lennart Nilsson, Karolinska University Hospital, Stockholm, Sweden. Lennart Nilsson has developed new photographic methods and technical improvements including electron microscopy, opening new dimensions to scientific photography. His work on medical subjects have become known worldwide through publication such as “Behold Man” and “A Child is Born.”

Citing this publication

This publication should be cited as HIV Sequence Compendium 2003, Leitner T, Foley B, Hahn B, Marx P, McCutchan F, Mellors J, Wolinsky S, and Korber B, editors. Published by Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, LA-UR number 04-7420.
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. In this years issue we have added a new section with HIV-2/SIVsmm complete genomes. Each of the alignments attempt to display the genetic variability within the different species, groups and subtypes of the virus.

At the time of publishing, there were 884 genomes longer than 7000 nucleotides available (and 907 sequences longer than 5000), and because of space limitations in the printed compendium we have omitted many sequences. These omissions were done considering redundant sequencing of certain isolates and patients as well as construction of phylogenetic trees with all available sequences. In section II, the HIV-1/SIVcpz complete alignment, we have included all subtypes and Circulating Recombinant Forms (CRFs), and thus the trend is to include at least the four reference sequences from each of the epidemiologically important virus variants. For some subtypes we have included more sequences to reflect their importance. In the future, if and when more subtypes and CRFs are described, there may be only space for the reference sequences. As always, tables with extensive background information gathered from the literature accompany the whole genome alignments. A more complete version of all alignments is 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. As always, we are open to complaints and suggestions for improvement. With the effort that goes into producing these volumes, we sincerely hope they will be widely used by the research community. Inquiries and comments regarding the Compendium should be addressed to:

Dr. Thomas Leitner
Theoretical Division, T-10, MS K710, LANL, 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
<table>
<thead>
<tr>
<th>NAME</th>
<th>SIZE</th>
<th>FUNCTION</th>
<th>LOCALIZATION</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gag</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MA</td>
<td>p17</td>
<td>membrane anchoring; Env interaction; nuclear transport of viral core.</td>
<td>virion</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(myristylated protein)</td>
<td></td>
</tr>
<tr>
<td>CA</td>
<td>p24</td>
<td>core capsid</td>
<td>virion</td>
</tr>
<tr>
<td>NC</td>
<td>p7</td>
<td>nucleocapsid, binds RNA</td>
<td>virion</td>
</tr>
<tr>
<td></td>
<td>p6</td>
<td>binds Vpr</td>
<td>virion</td>
</tr>
<tr>
<td><strong>Pol</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protease (PR)</td>
<td>p15</td>
<td>Gag/Pol cleavage and maturation</td>
<td>virion</td>
</tr>
<tr>
<td>Reverse transcriptase (RT), RNase H</td>
<td>p66</td>
<td>reverse transcription, RNase H activity</td>
<td>virion</td>
</tr>
<tr>
<td></td>
<td>p51</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>p15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Integrase (IN)</td>
<td>p31</td>
<td>DNA provirus integration</td>
<td>virion</td>
</tr>
<tr>
<td><strong>Env</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gp120/gp41</td>
<td>gp14</td>
<td>external viral glycoproteins bind to CD4 and secondary receptors</td>
<td>plasma membrane, virion envelope</td>
</tr>
<tr>
<td><strong>Tat</strong></td>
<td>p16/p14</td>
<td>viral transcription transactivator</td>
<td>primarily in nucleolus/nucleus</td>
</tr>
<tr>
<td><strong>Rev</strong></td>
<td>p19</td>
<td>RNA transport, stability and utilization factor(phosphoprotein)</td>
<td>primarily in nucleolus/nucleus shuttling between nucleolus and cytoplasm</td>
</tr>
<tr>
<td><strong>Vif</strong></td>
<td>p23</td>
<td>promotes virion maturation and infectivity</td>
<td>cytoplasm (cytosol, membranes) virion</td>
</tr>
<tr>
<td><strong>Vpr</strong></td>
<td>p10–15</td>
<td>promotes nuclear localization of preintegration complex, inhibits cell division, arrests infected cells at G2/M</td>
<td>virion, nucleus (nuclear membrane?)</td>
</tr>
<tr>
<td><strong>Vpu</strong></td>
<td>p16</td>
<td>promotes extracellular release of viral particles; degrades CD4 in the ER:(phosphoproteinonly in HIV-1 and SIVcpz)</td>
<td>integral membrane protein</td>
</tr>
<tr>
<td><strong>Nef</strong></td>
<td>p25–p27</td>
<td>CD4 and class I downregulation</td>
<td>plasma membrane, cytoplasm(virion?)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(myristylated protein)</td>
<td></td>
</tr>
<tr>
<td><strong>Vpx</strong></td>
<td>p12–16</td>
<td>Vpr homolog (not in HIV-1, only in HIV-2 and SIV)</td>
<td>virion (nucleus?)</td>
</tr>
</tbody>
</table>
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.

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 produced 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/or
Landmarks of the Genome

from the LTR promoter. 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. Some recent observations suggest that Vif functions late in replication to modulate assembly, budding, and/or maturation the N-terminal half of Vif (N'-Vif) specifically interacts with viral protease.

**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. It is found in HIV-1, HIV-2, SIVmac and SIVmnd. It is homologous to the Vpx protein.

**VPU** Vpu (viral protein U) is unique to HIV-1 and SIVcpz, a close relative of HIV-1. There is no similar gene in HIV-2 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 only in HIV-2/SIVmac/SIVsm and not in HIV-1 or SIVagm. This accessory gene is a homolog of HIV-1 vpr, and HIV-2/SIV 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 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.