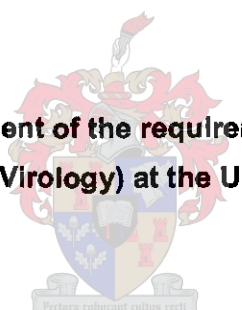
Characterisation of new full-length HIV-1 subtype D viruses from South Africa

André Gareth Loxton



Thesis presented in fulfillment of the requirements for the degree of Masters of Science (Medical Virology) at the University of Stellenbosch

Promoter:

Professor E. Janse van Rensburg

Co-promoter:

Professor S. Engelbrecht

December 2004

Declaration

I, the undersigned, hereby declare that the work contained in this thesis is my own original work and that I have not previously in its entirety or in part submitted it at any university for a degree.

Signature .....

Date 1/9/2004

Summary

The first episode of HIV-1 in South Africa was documented in 1982. Homosexual transmission of the virus was the predominate mode of transmission in an epidemic of mainly HIV-1 subtype B and D infections. To date, no full-length sequences of subtype D strains from South Africa has been reported. Here we describe the characterization and some of the unique features of the Tygerberg HIV-1 subtype D strains.

A near full-length 9 kb fragment was obtained through a one step PCR using high molecular weight DNA. Cloning was done successfully with the pCR-XL-TOPO cloning kit. Large quantities of plasmid DNA was grown and sequenced on both strands of the DNA. ORF determination and subtyping was followed by standard phylogenetic methods to construct evolutionary phylogenetic trees.

Subtyping and similarity plots revealed that the sequences from Tygerberg are pure subtype D. All the Tygerberg strains had intact genes with no premature stop codons. At the tip of the V3 loop, the Tygerberg strains have the GQGQ motif. R214 has a more variable *vpu* gene than the rest of the Tygerberg strains, but is still subtype D in this region. No premature stop codons have been observed in the *tat* gene and the glycosilation of the strains are less than the subtype D consensus.

We are the first to report full-length sequences of HIV-1 subtype D strains from South Africa. The sequences represent non-mosaic genomes of subtype D. Our results confirm that the subtype D sequences from the beginning of the HIV-1 epidemic differ from the subtype D sequences from recent isolates.

Opsomming

Die eerste episode van HIV-1 infeksie in Suid Afrika is in 1982 gedokumenteer. Die epidemie het hoofsaaklik uit subtipe B en D bestaan en was deur homoseksuele kontak oorgedra. Geen vollengte subtipe D DNS volgordes van Suid Afrika is tans beskryf nie. Hier beskryf ons die karakterisering van vollengte subtipe D stamme asook sommige van die unieke eienskappe van dié virusse.

Die vollengte 9 kb genoom volgorde was verkry deur 'n eenstap PKR reaksie met hoë molekulêre gewig DNS uit te voer. Die 9 kb fragment was suksesvol gekloneer met behulp van die pCR-XL-TOPO klonerings toetsstel. Groot hoeveelhede plasmied DNS was opgegroei en die nukleotied volgorde bepaal op beide stringe van die genoom. Die stamme was gesubtipeer en filogenetiese analise was uitgevoer met standaard metodes.

Die volledige DNS volgordes was bepaal en subtipering het daarop gedui dat die stamme van Tygerberg suiwer subtipe D is. Geen premature stop kodons is in die nukleotied volgordes van die Tygerberg stamme gevind nie. By die draai van die varieërbare deel (V3) het al die Tygerberg stamme die GQGQ motief gehad. R214 het 'n meer varieërbare *vpu* geen, maar behoort steeds tot die subtipe D groep in dié gedeelte. Daar was geen premature stop kodons in die *tat* geen gevind nie en die glikosilasie van die stamme is minder as dié van die konsensus subtipe D stam.

Ons is die eerste groep om vollengte subtipe D stamme van Suid Afrika te karakteriseer. Die DNS volgordes verteenwoordig suiwer subtipe D genome. Ons resultate bevestig die van ander dat die nukleotied volgordes van die ouer subtipe D stamme verskil van die nuwer stamme.

Acknowledgements

I wish to extent my sincere thanks and appreciation to the following people and institutions without whom this thesis would not have been possible:

Prof. S. Engelbrecht, my co-promoter, for her continued advice, assistance and guidance throughout the course of the project.

Prof. E. Janse van Rensburg, my promoter, for the opportunity to do a project in the department.

Florette Treurnicht and my colleagues at the Department of Medical Virology for their support and encouragement throughout my project.

My parents, Gert and Johanna, for supporting me and giving me the opportunity to attend university.

To my girlfriend, Natalie, for all the words of encouragements and the beautiful cards she made.

The South African AIDS Vaccine Initiative (SAAVI) and the Poliomyelitis Research Foundation (PRF) for funding the study.

"Humanities ancient enemies are, after all, microbes. They didn't go away just because science invented drugs, antibiotics, and vaccines (with the notable exception of smallpox). They didn't disappear from the planet when Americans and Europeans cleaned up their towns and cities in the post-industrial era. And they certainly won't become extinct simply because human beings choose to ignore their existence. "

***Laurie Garrett**, "The Coming Plague", Farrar, Strauss and Giroux, New York, 1994.*

"AIDS cannot be explained by a single virus causing a single and continuous epidemic. Instead, worldwide spread is the work of a virus family of types, subtypes, and strains that cause more or less related epidemics. Each member of the family has its own distinctive behaviour, and each epidemic runs its own distinctive course."

***J. Goudsmit**. 'Viral Sex: The Nature of AIDS.' 1997*

CONTENTS

Chapter 1

INTRODUCTION AND LITERATURE REVIEW

Chapter 1

INTRODUCTION AND LITERATURE REVIEW

INTRODUCTION

The Acquired immune deficiency syndrome (AIDS) is caused by two related viruses, human immunodeficiency virus (HIV) type 1 and type 2. Epidemiological analyses indicate that HIV-1 has spread all over the world, while HIV-2 is largely restricted to West Africa (Essex and Mboup, 2002). The majority of HIV-infections worldwide are caused by HIV-1 group M (major) viruses. Group M can be further divided into nine genetic subtypes and 15 circulating recombinant forms (CRF) (HIV sequence compendium, 2002).

The World Health Organization estimates that a total of 40 million people are currently infected with HIV-1, of which 26.6 million infected individuals live in sub-Saharan Africa (UNAIDS, 2003). In Africa, the most prevalent subtype is HIV-1 subtype C (Esparza and Bhamarapravati, 2000; McCutchan, 2000; Novitsky *et al*, 1999). The second most common subtype is the CRF02_AG (Essex and Mboup, 2002; Moore *et al*, 2001; Cornelissen *et al*, 2000). Other common subtypes are HIV-1 subtypes A and D (Hu *et al*, 2000; Rayfield *et al*, 1998).

Biological markers in epidemiological research and the tools to study the etiology, prevention and surveillance of infectious diseases comprise a major part of molecular epidemiology. HIV-1 subtypes and the genetic similarity between these strains can be used as such markers. Molecular epidemiology of HIV can provide detailed information on the spread and variation of HIV-1 and the data may help in designing vaccine trials and may even have relevance for understanding the biology of the virus (Ho and Huang, 2002).

In South Africa, the HIV-1 epidemic was initially associated with the homosexual population (Sher, 1989). Subtypes B and D were sequenced between 1984-1989 from homosexual men, who are thought to have introduced HIV-1 into South Africa from other countries (Becker *et al*, 1985).

The aim of this study was to characterise the HIV-1 subtype D strains sequenced from the beginning (1984-1986) of the epidemic in South Africa. The literature review of chapter 1 attempts to give an overview of the history of the HIV/AIDS epidemic and the origin of the virus. The diversity of HIV-1 subtype D are highlighted as part of the molecular epidemiology of the virus. The section on phylogenetic analysis gives an overview of some of the techniques that are used to model the HIV epidemic.

LITERATURE REVIEW

1.1 History

1.1.1 The beginning of the AIDS epidemic

AIDS was first recognised as a new and distinct clinical entity in 1981, when a clustering of an unusual opportunistic infection (*Pneumocystis carinii* pneumonia) and a rare neoplasm (Kaposi's sarcoma) was observed in young homosexual men in the United States of America (USA) (Gottlieb *et al.* 1981). Because this new clinical manifestation involved gay men, it was thought that the cause of this syndrome might be related to a life-style habit unique to this cohort of people. AIDS cases were soon reported in other groups as well, including intravenous (IV) drug users (CDC, 1982), haemophiliacs, blood transfusion recipients (Curran *et al*, 1984) and infants (Oleske *et al*, 1983). In Africa, the same clinical manifestations were observed, not only in homosexual men, but also in the heterosexual population (Piot *et al*, 1984). AIDS was subsequently defined as the appearance of certain dramatic and often life-threatening infections and cancers accompanied by a measurable depletion of immune competence (Ammamm *et al*, 1983). These observations made it clear that an infectious aetiology for AIDS should be considered.

In 1983, the first indication that AIDS could be caused by a retrovirus came, when Barre-Sinoussi *et al* (1983) recovered a reverse transcriptase containing virus from the lymph node of a man with persistent lymphadenopathy syndrome (LAS) which they later called Lymphadenopathy virus (LAV). A year later, Robert Gallo and his colleagues independently postulated that a variant T-lymphotropic retrovirus might be the causative agent of AIDS (Gallo *et al*, 1984).

Levy and co-workers also reported the identification of retroviruses, which they called AIDS-associated retroviruses (ARV) (Levy *et al*, 1984). By this time there were three prototype viruses: (a) LAV, (b) HTLV- III and (c) ARV. Because of the widespread interest in AIDS and its origins, the International Committee on the Taxonomy of viruses proposed that the AIDS retroviruses be officially designated as the human immunodeficiency viruses (HIV) (Coffin *et al*, 1986).

1.1.2 The origin of HIV

HIV is a member of the lentivirus subfamily of retroviruses (*Retroviridae*), with a diploid genome comprising two single-stranded RNA molecules. Since the discovery and subsequent recognition of HIV as the etiological agent of AIDS, more than 40 million people have been infected with HIV-1 (UNAIDS, 2003). A plasma sample from 1959 obtained from central Africa (Democratic Republic of the Congo, DRC) highlighted the fact that the epidemic might have originated in Africa (Nahmias *et al*, 1986). Even though AIDS was first described in America, the European epidemic may have started in the 1960's as the first reported case came from a Norwegian family who were missionaries in Africa, before 1970 (Froland *et al*, 1988).

There is now considerable evidence for a simian origin of HIV (Hahn *et al*, 2000; Myers *et al*, 1992). Viruses related to HIV, the simian immunodeficiency viruses (SIV), are found in many species of non-human primates. (**Fig.1.1**). It seems that HIV originated through cross-species transmission from naturally infected African primates to human (Hahn *et al*, 2000), a process referred to as zoonotic infections. Phylogenetic analysis indicates that multiple interspecies transmissions from simian species have introduced two genetically distinct types of HIV into the human population: HIV-1 and HIV-2, which are closely related to primate lentiviruses infecting chimpanzees (SIV_{cpz}) and sooty mangabeys (SIV_{sm}) respectively (Korber *et al*, 2000; Gao *et al*, 1992). Chimpanzees are commonly hunted for food, especially in west equatorial Africa (Hahn *et al*, 2000) and as a consequence represent a ready source for zoonotic transmission of SIVcpz to man. HIV-1 is most similar to SIV sequenced from chimpanzees, particularly to the strains sequenced from the subspecies,

Pan troglodytes troglodytes (Gao *et al*, 1999). The phylogenetic positions of

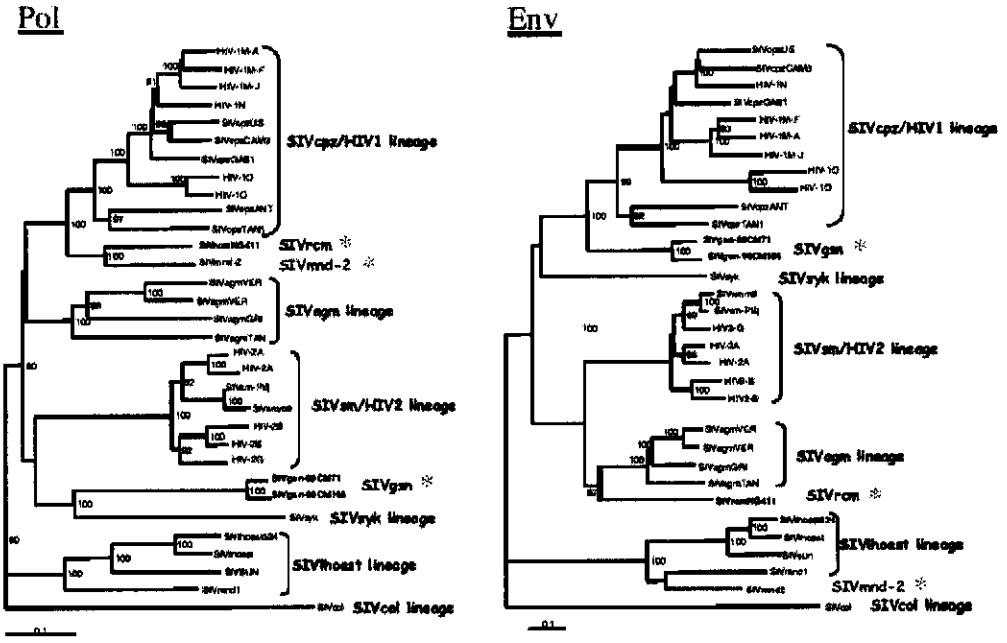


Figure 1.1. Evolutionary relationship of primate lentiviruses for which full-length sequences are available. Relationship based on the neighbour-joining phylogenetic analysis of full-length Pol and Env amino acid sequences. The six major lineages are indicated in black and the recently described SIVs with discordant phylogenies are in grey using an asterisks (*). Branch lengths are drawn to scale and only bootstrap values above 80% are shown. (Peeters and Courgnaud, 2002).

HIV-1 groups M, N and O within the HIV-1/SIVcpz radiation indicate that the three HIV-1 groups have each arisen as a consequence of independent zoonotic transmissions (Gao *et al*, 1999). More support for the zoonotic infections of humans is the fact that natural SIV infections fail to cause disease in infected animals (Rey-Cuille *et al*, 1998; Cichutek and Norley, 1993), which indicates that the virus has learned to adapt to the host or that they co-exist to mutual benefit.

The timing of SIVcpz transmission to humans, leading ultimately to the HIV-1 pandemic, has been a challenging question. Phylogenetic methodology has estimated 1930 +/- 20 years as the timing of the last common ancestor of the HIV-1 group viruses (Hahn *et al*, 2000; Korber *et al*, 2000). This estimation relies on the assumption of a molecular clock, which postulates that molecular

change is a linear function of time and that substitution accumulates according to a Poisson distribution (Korber *et al*, 2000). The date of the most recent ancestor of HIV-2 subtype A strains was estimated to be 1940 +/- 16 years and that of the B strains was estimated to be 1945 +/- 14 years (Lemey *et al*, 2003).

A recent article by Salemi *et al* (2003) clarified the origin and evolution of the primate lentiviruses (PLV), the group which include the human immunodeficiency virus type 1 and 2 as well as their simian relatives. The PLV strains are currently assigned to six approximately equidistant phylogenetic lineages (Hahn *et al*, 2000): the SIVcpz, (ii) the SIVsm clade, (iii) the SIVagm clade, the SIVlhoest clade, (v) a SIVsyk clade and (vi) the divergent SIVcol strain. Salemi and colleagues' (2003) analysis confirmed the existence of at least five putative recombinant fragments in the PLV genome with different clustering patterns. The findings not only imply that the six so-called pure PLV lineages have in fact mosaic genomes, but also make more unlikely the hypothesis of co-speciation of SIVs and their simian hosts. This is in correlation with Bailes *et al* (2003) who, through phylogenetic analyses found a hybrid origin of SIV in chimpanzees. The findings of these two groups has important implications: first, it provides evidence that, in addition to humans, other ape species acquired SIV by cross-species transmission which caused the formation of recombinant viruses and secondly, it showed that recombinant chimpanzee viruses is capable of spreading to humans.

HIV-2, the other causative agent of AIDS, has been found predominantly in the heterosexual populations in West Africa (countries such as: Guinea Bissau, Ivory Coast, and Senegal), but has spread very little to other areas (De Cock *et al*, 1993). Clavel and co-workers (1986) also reported on the isolation of this new human retrovirus from West African patients. HIV-2 infection of humans in western Africa may have arisen, and may still be occurring, by cross-species transmission from sooty mangabey (sm) monkeys. The natural habitat of mangabey monkeys, the forested regions of western Africa, is nearly coincident with the region where human infection with HIV-2 is endemic, and the sequences of HIV-2 sequences are within the range of variation of known SIVsm sequences. The immunologic abnormalities associated with HIV-2 are

similar but milder than those in persons with HIV-1 infections (Egboga *et al*, 1992).

1.2 The HIV-1 virus

1.2.1 The Virion Structure

Studies have shown that HIV exhibits a characteristic cone-shaped core that is surrounded by a bilayer lipid envelope derived from the host cell membrane (**Fig. 1.2**). The inner core is comprised of the major capsid (CA) protein p24 (Gag protein), which surrounds two copies of the viral RNA (Briggs *et al*, 2003). Closely associated with the RNA strands is the viral RNA-dependant DNA polymerase (Pol) including the protease, reverse transcriptase (RT) and integrase and the nucleocapsid (NC) proteins (Briggs *et al*, 2003; Levy, 1994, Hahn, 1994). The inner portion of the viral membrane is surrounded by a myristolated p17 core (Gag) protein that provides the matrix (MA) for the viral structure and is vital for the integrity of the virion. MA is required for the incorporation of the Env proteins into the mature virions. The surface of the virus is characteristically made up of 72 knobs containing trimers or tetramers of the envelope glycoproteins. They are derived from a gp160 precursor, which is cleaved inside the cell into a gp120 external surface (SU) envelope protein and a gp41 transmembrane (TM) protein (Goettlinger, 2001; Levy, 1994). These proteins are transported to the cell surface, where part of the central and N-terminal portion of gp41 is also expressed on the outside of the virion. The central region of the TM protein binds to the external viral gp120 in a noncovalent manner.

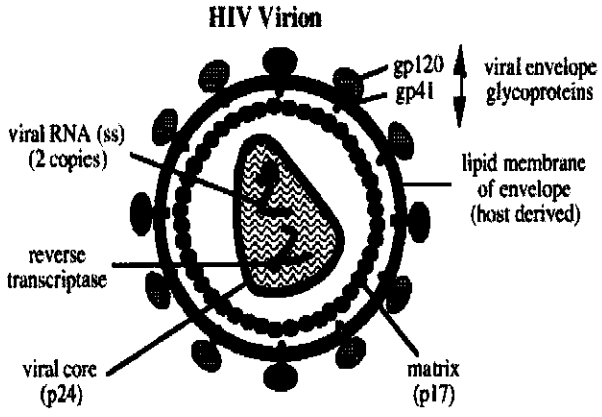


Figure 1.2. A cartoon illustration of the HIV-1 virion displaying the viral *envelope*, *gag*, and *pol* proteins. Also visible in the cartoon is the single stranded RNA molecules.
(http://www.chemsoc.org/exemplarchem/entries/2002/levasseur/images/hiv. GIF)

It is estimated that a single HIV-1 virion contains about 1200 molecules of p24, roughly 80 molecules of the reverse transcriptase and up to 280 molecules of gp120 (Hahn, 1994).

1.2.3 HIV-1 genome organisation

The genomic size of the HIV virion is about 9.2 kb, with open reading frames coding for several proteins (**Fig.1.3**). HIV contains long terminal repeats (LTR) that do not encode proteins but are essential for the regulation of viral gene expression (Briggs *et al*, 2003). The LTRs are on both sides of the HIV genes: structural genes (*gag*, *pol* and *env*), the regulatory genes (*tat*, *rev* and *nef*) and the accessory genes (*vif*, *vpr* and *vpu*). An overview of the HIV-1 genes and their products, as well as their function in the life cycle of the virus is given in **Table 1.**

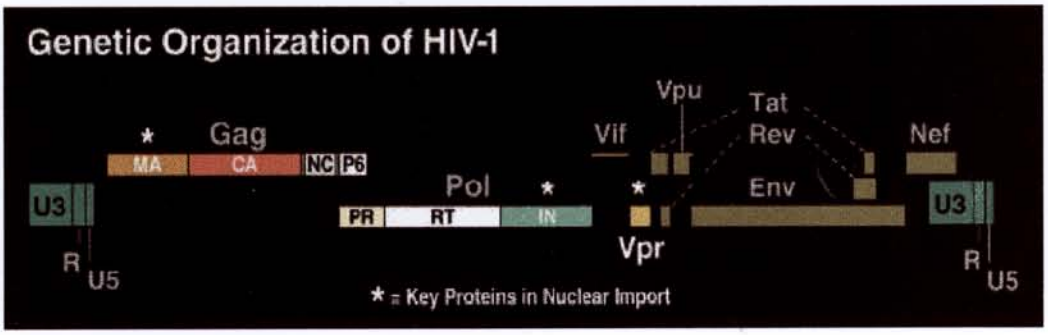


Figure 1.3. The HIV-1 genome. The different genes of the virus as well as the U3, U5 and R regions of the LTR's are showed on the figure (Briggs *et al*, 2003).

Table 1. The HIV-1 genes and their products (HIV Sequence Compendium, 2002)

Gene	Protein	Function
gag MA	p17	Membrane anchoring; env interaction; nuclear transport of viral core, (myristylated protein)-
CA	p24	Core capsid
NC	p7	Nucleocapsid, binds RNA
	p6	Binds Vpr
protease (PR)	p15	*gag/pol* cleavage and maturation
reverse transcriptase (RT)	p66 p51	Reverse transcription, RNase H activity
RNase H *integrase* (IN)		DNA provirus integration
env	gp120 gp41	External viral glycoproteins bind to CD4 and secondary receptors
tat	p16/p14	Viral transcriptional transactivator
Rev	p19	RNA transport, stability and utilisation factor (phosphoprotein)
Vif	p23	Promotes virion maturation and infectivity
Vpr	p10-15	Promotes nuclear localization of preintegration complex, inhibits cell division, arrests infected cells at G2/M
Vpu	p16	Promotes extra cellular release of viral particles; degrades CD4 in the ER; (phosphoprotein only in HIV-1 and SIV cpz
Nef	p27-p25	CD4 and class I down regulation (myristylated protein)
Vpx	p12-16	Vpr homologue (not in HIV-1, only in HIV-2 and SIV)

1.2.4 HIV-1 replication (The life cycle)

The life cycle of HIV-1 can be divided into two phases, establishment of infection and productive infection (Goettlinger, 2001; Haseltine, 1992). Infection of the target cell is established through a set of virus-cell interactions that include binding of the virus to the cell surface, fusion of the virus and cell membranes, entry of the virus capsid into the cytoplasm, conversion of viral RNA into DNA and the entry of the viral DNA into the nucleus (**Fig.1.4**).

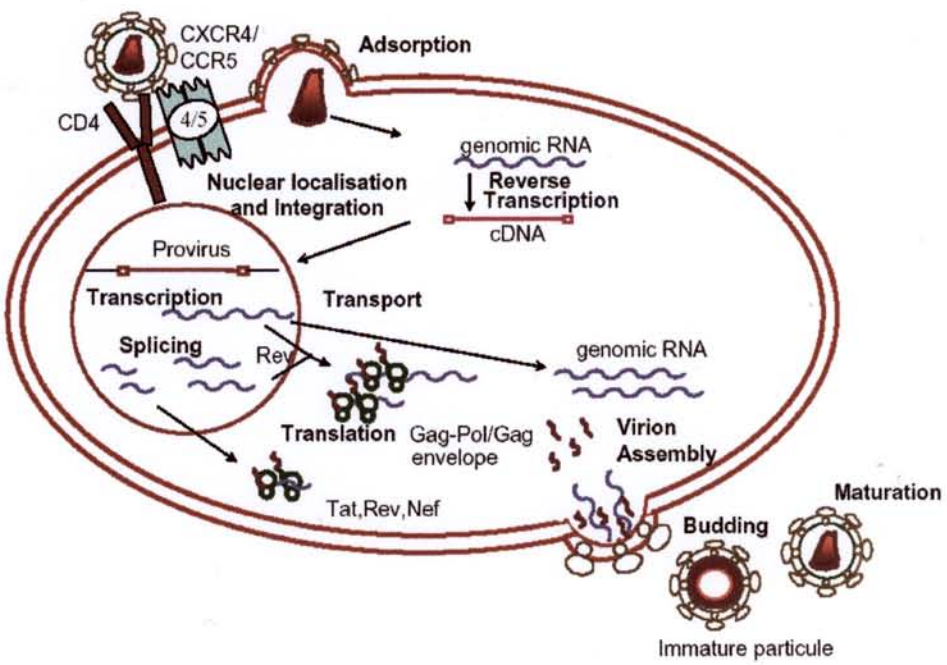


Figure 1.4. The HIV-1 replication cycle. The stages of viral replication are depicted in the cartoon. Viral attachment and integration, transcription, translation, transport and budding of the virus are illustrated on the cartoon (Gatignol and Jeang, 2000).

Once the viral DNA enters the nucleus, infection is established. The viral DNA may be integrated into the host DNA or may form stable circles. Once integration has occurred, the progeny of the infected cell will also be infected (Goettlinger, 2001). Viral expression begins when viral DNA is transcribed into RNA by the host DNA polymerase II. The viral RNA is processed by splicing and exported to the cytoplasm, where it is translated into viral protein. The virus capsid, which assembles on the inner surface of the membrane, incorporates

full-length viral RNA into newly formed particles (Briggs *et al*, 2003; Luciw, 1996).

New virions are produced as the virus buds through a region of the cell membrane. The outer surface protein of the virus, located on the surface of the cell membrane, becomes associated with the progeny virus particles during the budding process (Goettlinger, 2001; Haseltine, 1992). Replication of the virus is controlled by host cell as well as by viral genes. The state of differentiation and activation of the infected cell may determine the rate of each step in the virus life cycle. Additionally, some of the proteins specified by the virus affect the rate of accumulation of the primary RNA transcript in the nucleus, the processing and export of the viral transcripts and the rate of assembly and budding of the virus particles (Briggs *et al*, 2003).

1.3 The diversity of HIV-1

The HIV-1 genome can accommodate a high degree of sequence variation while maintaining replication competence and structural integrity (Balfe *et al*, 1990). The development of variation is facilitated by the "infidelity" of HIV-1 reverse transcriptase, which lacks an editing function (Preston *et al*, 1988). Therefore, no two HIV strains are alike and even within a single individual, HIV is present as a 'quasispecies' - a swarm of micro variants that are highly related, yet genetically distinct from each other (Goodenow *et al*, 1989; Vartanian *et al*, 1992).

In Africa, where the effects of HIV-1 have been most devastating, there are multiple subtypes of the virus. The distribution of different subtypes within African populations is generally not linked to particular risk behaviours (Neilson *et al*, 1999) unlike in some Asian countries where the spread of HIV-1 subtype E (CRF01_AE) is linked to the intravenous drug users (Weniger *et al*, 1994). Africa is therefore an ideal setting in which to examine in more detail the diversity and mixing of viruses of different subtypes on a population basis (Neilson *et al*, 1999).

Phylogenetic analyses of numerous strains of HIV-1, sequenced from diverse geographical origins, have revealed that they can be subdivided into groups, subtypes and sub-subtypes (Robertson *et al*, 1999). HIV-1, the variant responsible for the majority of HIV/AIDS infections (99% worldwide) (Moore *et al*, 2001), has been further divided into 3 groups: M (major), N (New, or non-M, non-O) and O (outlier) (Peeters, 2001). Within group M, at least nine subtypes have been identified: A-D, F-H, J and K) (Peeters, 2000). Subsequent to the designation of group M subtypes, it was realized that certain sequences do not display a single subtype cluster pattern when different regions of their genomes where phylogenetically analysed. These mosaic HIV-1 genomes have been identified in several, apparently unlinked, individuals and some (A/E; B/C etc) play a major role in the global AIDS epidemic and are now designated circulating recombinant forms (CRFs) (Robertson *et al*, 1999; Moore *et al*, 2001). Separate sub-clusters are distinguished within subtypes A and F (A1 and A2, F1 and F2), each pair of sub-subtypes being more related to each other than with other subtypes. Subtypes B and D should be the same subtype, but their original designation as different subtypes has been retained for consistency with earlier published work. The identification of new clades of HIV-1 and the realisation of the existence of CRFs, characterised by full-length genome sequence analysis, have led to several re-adjustments in the taxonomy of HIV-1 (Thomson *et al*, 2002).

1.3.1 Distribution of HIV-1 subtypes

The distribution of subtypes varies from country to country and sometimes also between different risk groups in a specific area (van Harmelen *et al*, 1997; Williamson *et al*, 1995). The occurrence of a certain subtype in a population can be a consequence of a founder effect: introduction and rapid spread of a pathogen in a virgin population, which leads into a genetically highly homogenous epidemic (Daniels *et al*, 2003).

Central Africa is thought to be the origin of all subtypes of HIV-1. The initial diversification of group M may have occurred within or near the territory of the

DRC, where the highest diversity of group M has been reported (Vidal *et al*, 2000), and the earliest case of HIV-1 infection been documented (Nahmias *et al*, 1986).

Since 1992, the *env* coding sequence of HIV has been used to classify globally prevalent viruses (Janssens *et al*, 1997). Subtypes form clusters roughly equidistant with each other in phylogenetic trees, being separated by 25-35% amino-acid distance between *env* sequences (Gaschen *et al*, 2002; Thomson *et al*, 2002). Of the 9 subtypes identified for HIV-1, the subtype C viruses have been implicated for causing 47.2% of infections in 2002 (Osmanov *et al*, 2002). The highest incidence of subtype C was observed in the southern part of Africa, Ethiopia, India and China (Thomson *et al*, 2002). In the regions with the highest incidence of subtype C, the prevalence of subtype C infections can exceed 30% of the adult populations (UNAIDS, 2002). The subtype C virus is also circulating as a minor form in Brazil and Russia (**Fig.1.5**). The second most prevalent genetic variant of HIV-1 is represented by *env* subtype A, which is in a large proportion of cases, is represented by CRF02_AG strains (Osmanov *et al*, 2002). Subtype B is the main genetic form in western and central Europe, the Americas and Australia, and is common in several countries of Southeast Asia, North Africa, and the Middle East. In South Africa and Russia, subtype B infections are almost exclusively seen in homosexual men (Thomson *et al*, 2002). This subtype has accounted for a significant number of HIV-1 infections in 2000, estimated at around 12.3% of global cases. Other globally prevalent HIV-1 genetic forms, common on a localized scale, are subtype D (Uganda, Tanzania and Kenya; 34% to 53% of infections in east Africa), Subtype F (Romania), subtype G (west and central Africa), and the circulating recombinant form, CRF12_BF (Thomson *et al*, 2002).

Figure 1.5. Regional spread of HIV-1 genetic subtypes. The different subtypes are indicated in black and the circulating recombinant forms of South America and Western Europe indicated in red. (Thomson *et al*, 2002)

1.4 HIV-1 subtype D

Subtype D viruses was first recognized in Zairian patients in 1983 when Alizon *et al* (1986) described the LAV_{ELI} and the recombinant LAV_{MAL} sequences. The LAV_{ELI} strain (DRC) became the first HIV-1 subtype D strain to be sequenced fully. This allowed the comparison of full-length clones from Africa and the United States. Partial sequencing of the *gag* and *env* genes of an HIV-1 subtype D sequence, obtained from a Zairian male student in Alabama, by Gao *et al* (1994) underlined the fact that the early HIV-1 epidemic in America could have been due to the introduction of HIV-1 from Africa. Currently, subtype D sequences accounts for 7.43% of the full-length viruses characterised thus far, all of which are from Africa. Even though subtype B and subtype D was associated with the initial HIV-1 epidemic in South Africa (Puren, 2002), to date, not a single full-length genome has been sequenced for the early strains from South Africa.

It has been suggested that HIV-1 subtypes could influence viral transmissibility and pathogenesis, but the existence of many other factors that influence these

features makes it difficult to establish the true effect of viral subtypes. Factors such as the V3 loop sequence and chemokine receptor usage have been shown to play a role in syncytium inducing phenotype and viral tropism (O'Hagen *et al*, 2003; Dragic *et al*, 1996). The two principal co-receptors used by HIV-1 are CXCR4 and CCR5, members of the CXC and CC chemokine receptor family, respectively (Fenyo *et al*, 1997). Tscherning *et al* (1998) showed that subtype D sequences do not show dual tropism for CXCR4 and CCR5. The particular co-receptor used by a strain of HIV-1 to enter a host cell is primarily determined by the amino acid sequence of the V3-loop region (35 amino acids) of the viral envelope (Pillia *et al*, 2003). Compared to other group M subtypes, subtype D strains demonstrate a highly variable pattern of V3-loop amino acids (Spira *et al,* 2003). There is an elevated rate of nonsynonomous (amino acid altering) substitutions in the third variable region of subtype D viruses (Korber *et al*, 1994). The number of amino acid changes within the V3-loop regions compared to changes outside the V3-loop region in subtype D genomes is larger than in other subtypes of HIV-1 (Korber *et al*, 1994).

A recent study in Tanzania suggested that the maternal subtype could play a role in the incidence of vertical transmission, with subtypes A, C and recombinant viruses being more likely to be perinatally transmitted than subtype D (Renjifo *et al*, 2001). Viruses containing subtype C LTR's are 6.1 times more likely to be transmitted than those with subtype D LTR's (Blackard *et al*, 2001; Gordon *et al*, 2003). A prospective study of female sex workers in Senegal showed that women infected with C, D or G subtypes were eight-fold more likely to develop AIDS than were those infected with subtypes A, suggesting that HIV-1 subtypes differ in rates of progression to AIDS (Kanki *et al*, 1999). An Ugandan study, looking at 1045 adults infected with subtypes A or D showed that subtype D was associated with faster progression to death and with a lower CD4 cell count than subtype A. In contrast to Kanki *et al* (1999), a study by Kaleebu and co-workers (2001) found no significant difference in disease progression between individuals infected with subtype A and D. Subtypes A and D are also the predominant HIV-1 subtypes in Uganda (Hu *et al*, 2000; Kaleebu *et al*, 2002).

The neutralization profile of a specific subtype of HIV plays an important role in the diversity of HIV. Even though Kitabwalla *et al* (2003) showed that a quadruple combination of human monoclonal antibodies (MAb) raised against subtype B were able to neutralize subtypes A – D, Zwick *et al* (2001) showed that a neutralizing MAb, Fab Z13, wasn't able to neutralize any of the primary subtype D sequences. A study by Palmer *et al*, (1998) found subtype D viruses to function with diminished drug sensitivity owing to rapid growth kinetics, whereas subtypes A, B, C and E demonstrated comparable results.

1.5 Phylogenetic analysis of HIV

1.5.1 Concepts of molecular evolution

The idea of evolution originated early in the 1800s when naturalists realised that species have changed over time but was uncertain as to what have changed. Since the time of Charles Darwin, it has been a dream for many biologists to reconstruct the evolutionary history of all organisms on earth and express it in the form of a phylogenetic tree (Ayala and Fitch, 1997). The primary cause of evolution is the mutational change of genes. A mutant gene or DNA sequence caused by nucleotide substitution, insertions/deletions (indels), recombination or gene conversion may spread through the population by genetic drift and/or natural selection and eventually be fixed in a species (Hartl and Clark, 1997).

A phylogenetic tree is a mathematical structure, which is used to model the actual evolutionary history of a set of relationships among groups or organisms (Posada *et al*, 2001; Page and Holmes, 1998). The tree consists of nodes connected by branches (or edges). Terminal nodes (operational taxonomic unit, OTU) represent sequences or organisms for which we have data; they may either be extant or extinct (Nei and Kumar, 2000; Page and Holmes, 1998; Vandamme, 2003). Internal nodes represent hypothetical ancestors; the ancestors of all the sequences that comprise the tree are the roots of the tree. An unrooted tree only positions the individual taxa relative to each other without indicating the direction of the evolutionary process. In an unrooted tree, there is no indication of which node represents the ancestor of all OTUs (Vandamme, 2003). The easiest way to calculate divergence times is to assume that

sequences divergence accumulates linearly over time; this is called a molecular clock. When the molecular clock holds, all lineages in the tree have accumulated substitutions at the same rate, so that the evolutionary rate is constant (Vandamme, 2003; Page and Holmes, 1998). The molecular clock theory is an assumption of evolution, therefore for each set of data to be analysed, the molecular-clock hypothesis should be tested with the statistical methods available (Nei and Kumar, 2000).

1.5.2 The multiple alignment

Once sequences are obtained, the sequences need to be error-checked and assembled into contiguous fragments (contigs). With HIV sequences it is important to check if any of the sequences are potential contaminants (Korber *et al*, 1995). In addition, if multiple HIV sequences have been obtained, these need to be aligned so that homologous sites appear in the same column. Sequences normally have different lengths, which mean that gaps must be used in some positions to achieve the alignment. The generation of alignments is one of the most common tasks in computational sequence analysis because alignments are required for many other analyses, such as structure predictions or to demonstrate sequence similarity within a family of sequences (Higgins, 2003). The most commonly used software to do alignments with is the Clustal W (Thompson *et al*, 1994) and Clustal X (Thompson *et al*, 1997) programmes.

1.5.3 Nucleotide substitution models

DNA sequences are not very informative about their evolutionary history. When comparing homologous sites in DNA sequences, we simply observe that the sequences are the same or not (Page and Holmes, 1998). A basic process in the evolution of DNA sequences is the substitution of one nucleotide for another (transitions and transversions) during the evolutionary time (Graur and Li, 1999). To study the dynamics of nucleotide substitutions, it is necessary to use a mathematical model of nucleotide substitution. For this reason, many scientists have developed different substitution models (Li, 1997). The models range from the simple Jukes and Cantor method to the more sophisticated

general time-reversible (GTR) model. The most frequently used model for HIV datasets, Kimura's two-parameter model will be discussed further.

1.5.3.1 Kimura 2-parameter model (K2P)

The rate of transitional nucleotide substitution is often higher than that of transversional substitution in real data (Nei and Kumar, 2000). Kimura (1980) proposed a method for estimating the number of nucleotide substitutions per site, taking into account this observation. Kimura's 2-parameter model assumes that the rate of transitions per site (α) may differ from the rate of transversions (β), giving a total rate of substitution per site of $\alpha + 2\beta$ (Page and Holmes, 1998). One should keep in mind that for any nucleotide there are three possible changes, one of which is a transition, the remaining two being transversions. The K2P model is the most widely used model to study HIV phylogenies.

1.5.4 Phylogeny inference based on distance methods

1.5.4.1 Tree-inferring methods based on genetic distance

Neighbour-joining (NJ)

Saitou and Nei (1987) developed an efficient tree-building method that is based on the minimum evolution principal. This method does not examine all possible topologies, but at each stage of taxon clustering a minimum evolution principal is used. One of the important concepts in the NJ method is neighbours, which are defined as two taxa that are connected by a single node in an unrooted tree. The algorithm to construct a NJ tree begins with a star tree, which is produced under the assumption that there is no clustering of taxa (Nei and Kumar, 2000).

1.5.4.2 Evaluation of inferred trees using Bootstrap analysis

One way to measure sampling error is to take multiple samples from the population being studied and compares the estimates obtained from the different samples. The spread of those estimates gives an indication of the extent of sampling error, that is, how much our conclusions would vary depending on what sample we took (Page and Holmes, 1998). The bootstrap is a computational technique for estimating a statistic for which the underlying

distribution is unknown or difficult to derive analytically (Felsenstein, 1985; Efron *et al*, 1996). The bootstrap belongs to a class of methods called resampling techniques because it estimates the sampling distribution by repeatedly resampling data from the original sample data set (Graur and Li, 1999). The value obtained for this repeated process is called the bootstrap confidence value (P_B) or simply the bootstrap value (Nei and Kumar, 2000) and is expressed as percentages.

AIM OF THE STUDY

In 2002, when this study was initiated, no full-length sequences for subtype D from South Africa were available. At that stage, only five subtype D sequences were described in the Los Alamos database. Three of the sequences are from the DRC (ELI, NDK and Z2) and the other sequence, 94UG114, from Uganda. The fifth sequence, MB2059 is from Kenya. Other full-length subtype D sequences available were published later in 2002 and 2003 (Kijak and McCutchan, 2003; Koulinska *et al*, 2003; Vidal, 2003; Dowling *et al*, 2002; Harris *et al*, 2002; Novelli *et al*, 2002). Therefore, determining the full-length sequences of HIV-1 viruses from the beginning of the epidemic may shed light on the origin of HIV-1 in South Africa.

The objective of the study was to characterise HIV-1 subtype D sequences, by means of cloning, sequencing and phylogenetic analysis of the clones of HIV-1 subtype D, sequenced at the start (1984-1986) of the HIV-1 epidemic in South Africa.

Chapter 2

MATERIALS AND METHODS

Chapter 2

MATERIALS AND METHODS

2.1 Viral Isolates and patient data

Since 1984, blood samples from HIV-1 infected patients were obtained at the Tygerberg Academic Hospital in the Western Cape. From 1984 to 1986, four HIV-1 subtype D viruses were sequenced. Viruses R2, R214 and R286 were isolated by Brenda Robson and virus R482 was isolated by Susan Engelbrecht (Engelbrecht, 1992). Viruses were co-cultured with donor peripheral blood mononuclear cells (PBMC) obtained from healthy HIV negative individuals. High molecular weight (hmw) genomic DNA was sequenced from virus-infected cell cultures and stored at 4 °C.

The current study makes use of the same DNA used in an earlier study (Engelbrecht *et al*, 1995). The project was approved by the Ethical Committee of the University of Stellenbosch (Research Committee C) on 23/05/1995, with project number, 95/127 and Susan Engelbrecht as responsible person. The title of the project was: Molecular epidemiology and analysis of the HIV-1 *env* gene. The demographic and clinical data of the patients used in this study as well as the viral phenotypes, are summarised in **Table 2.1**.

2.2 Plasmid vectors and Bacterial strains

The pCR-XL-TOPO plasmid (Invitrogen Corporation, Carlsbad, CA, USA) is a 3519 bp expression plasmid, which is linearised for TA-cloning in the TOPO® XL PCR Cloning Kit. The plasmid has a T7 promoter site for *in vitro* RNA transcription and sequencing as well as the M13 forward and reverse sites for sequencing. Kanamycin and Zeocin® resistance genes for flexible antibiotic selection are included in the vector.

The competent bacterial strain *E.coli* Top 10 was used in the transformation reactions. The bacterial strain is provided at a transformation efficiency of 1 x 10^9 cfu/µg super coiled DNA and is used for high-efficiency cloning and plasmid

propagation. The genotype of the *E.coli* Top 10 competent cells stored at -80°C is:

F⁻*mcr*A Δ(*mrr-hsd*RMS-*mcr*BC) Φ80*lac*ZΔM15 Δ*lac*X74 *deo*R *rec*A1 *ara*D139 Δ(*ara-leu*)7697 *gal*U *gal*K *rps*L (Str^R) *end*A1 *nup*G

Table 2.1. Demographic and Clinical data of patients and viral phenotype of HIV-1 isolates (Adapted from Engelbrecht *et al*, 1995)

Patient number	Sample date	Demographic data[a] at isolation	Clinical stage[b]	Viral phenotype[c]	*env* subtype
R2	15-11-1984	24 W M Bi	AIDS	SI	D
R214	20-6-1985	36 W M Ho	AIDS	SI	D
R286	17-6-1985	33 W M Ho	AIDS	SI	D
R482	30-1-1986	37 W M Ho	AIDS	SI	D

a) Demographics (number indicates age in years): W, white; M, male; Ho, homosexual; Bi, bisexual

b) Clinical data: AIDS, acquired immune deficiency syndrome

c) Phenotype: SI, syncytium inducing

2.3 PCR amplification and purification of the PCR fragment

To amplify virtually full-length HIV-1 genomes in one continuous segment, three primer pair combinations were tested initially on DNA from sequence R286. These included: MSF12/MSR5 (Salminen *et al*, 1995b), UP1A/LOW2 (Gao *et al*, 1998) and UP1A/S2Full (zur Megede *et al*, 2002). The primer pair that gave the best amplification of the DNA fragments was used further to amplify the other sequences (R2, R214 and R482). MSF 12 (5'- AAA TCT CTA GCA GTG GCG CCC CGA ACA – 3') primer was used as the forward primer with the MSR 5 (GCA CTC AAG GCA AGC TTT ATT GAG GCT –3') as the reverse primer. For the PCR reaction, 3 µl (0.3 µg/µl) of hmw DNA with the Expand long

template PCR system (Roche Molecular Biochemicals, Mannheim, Germany) with buffer 2 was used. The reaction was performed on a GeneAmp PCR System 9600 thermal cycler (Perkin Elmer, Boston, MA, USA) using a method adapted from Salminen *et al* (1995b): template DNA was denatured at 94°C for 2 minutes, followed by ten cycles of denaturing at 94°C for 2 minutes, annealing at 60°C for 30 seconds and elongation at 68°C for 8 minutes. This was followed by 20 cycles of denaturing at 94°C for 10 seconds, annealed 60°C for 30 seconds and elongated at 68°C for 8 minutes with 15-second increments per cycle. A final elongation step at 68°C for 30 minutes was added. After amplification, the DNA was stored at 4°C.

The amplified DNA was visualised by electrophoresis through a 0,6% agarose gel containing 5 µg/ml ethidium bromide in TAE buffer (0.04M Tris-acetate, 0.001M EDTA). A 1 kb DNA ladder (Promega, Madison, WI, USA) was included in the electrophoresis to compare DNA fragment sizes. After electrophoresis, the DNA fragments were purified from the gel, using the QIAEX II Gel Extraction kit (Qiagen, GmbH, Germany). The manufacturer's protocol was used without any modifications. The pellet was air-dried, the DNA eluted in TE buffer, and the DNA concentration determined using the following equation (Sambrook *et al*, 1989):

$$\text{DNA concentration} = \frac{\text{OD 260}}{20} \times \text{dilution factor}$$

$$\text{DNA purity} = \frac{\text{OD260}}{\text{OD280}}$$

OD= optical density, measured in a Spectronic® Genesys 5 spectrophotometer (Spectronic Instruments, Rochester, NY, USA). Optical density readings of the DNA at 260nm were used to calculate the concentration of the DNA.

2.4 Cloning of the PCR fragments

2.4.1 Cloning

For efficient cloning of the near full-length genome of HIV-1, the TOPO® XL PCR Cloning kit (Invitrogen Corporation, Carlsbad, CA, USA) designed for cloning large fragments was used. The manufacturer's protocol was followed. Briefly, cloning reactions were prepared for sequences R2, R214, R286 and R482. The purified DNA products was cloned into the pCR-XL-TOPO cloning vector at a 1:4 (vector: insert) ratio and transformed into the Top10 chemically competent cells. These reactions were then plated onto Luria-Bertani (LB) agar (10g/L bacto-tryptone, 5g/L bacto-yeast-extract, 10g/L NaCl, 15g/L bacto-agar) (Hispanlab, SA) plates for growth overnight at 33°C.

2.4.2 Plasmid DNA Isolations

Following an overnight incubation, single colonies from the LB agar plates were inoculated into 3 ml LB media (10g/L bacto-tryptone, 5 g/L bacto-yeast extract, 10 g/L NaCl) (Hispanlab, SA) containing Kanamycin (50ug/ml) and incubated in a Labcon shaking incubator (Labmark, Roodepoort, RSA) at 33°C for 16 hours. DNA extractions were done using the small-scale plasmid DNA protocol (Sambrook *et al*, 1989). Plasmid DNA was separated by gel electrophoresis on a 0.6% agarose gel.

2.4.3 Preparation of glycerol stocks

Glycerol stocks were prepared from all positive clones by adding the bacterial culture in a 1:3 ratio to the glycerol. (Adapted from Sambrook *et al,* 1989). The stocks were stored in cryogenic vials at –80°C.

2.4.4 Preparation of plasmid DNA for sequencing

To prepare large enough volumes of plasmid DNA of high purity for sequencing, we used the QIAfilter plasmid Midi kit and protocol (Qiagen, Heidelberg, Germany). The concentration, as well as the purity of the plasmid DNA was determined as described before.

2.5 DNA Sequencing and analysis

2.5.1 DNA sequencing

The near complete genome was fully sequenced using the ABI Prism 310 Genetic Analyzer (Applied Biosystems, USA) and the BigDye® terminator cycle sequencing kit. Sanders-Buell *et al* (1995) described the HIV-1 sequencing primers that were used. Additional primers designed for sequencing the gaps in the near full-length genome are shown in **Table 2.2**. The primers for sequencing

Table 2.2 Additional primers designed to sequence the HIV-1 subtype D genomes

Primer	Primer sequence	Strand	Tm- ^{o}C
G05	5'- ATG CAG AGA GGC AAT TTT AAG G- 3'	+	54.9
Pol1D	5'- TCC CTC AAA TCA CTC TTT GGC - 3'	+	56.3
Pol2D	5'- CTA TTG AAA CTG TAC C - 3'	+	40.1
Pol2Drev	5'- CCA TCC ATT CCT GGC - 3'	-	49.0
Pol3D	5'- CAG TAC TGG ATG TGG G- 3'	+	48.5
Pol3Drev	5'- CCC ACA TCC AGT ACT G - 3'	-	48.5
Pol-DF	5'- TTG TAC AGA TAT GGA AAA GGA AGG- 3'	+	54.1
Pol-DR	5'-AAT TTA GGA GTC TTT CCC - 3'	-	46.6
Env-DF	5'- GGT CAC AGT TTA TTA TGG G- 3'	+	48.5
Env-DR	5'- GAA TTG CAA AAC CAG CTG G - 3'	-	53.6

Pol = Polymerase; G=Gag; Env = Envelope

the pCR-XL-TOPO vector were obtained with the kit. The primers for sequencing the accessory genes (TatX1F, Nef F and Nef R) were described by Scriba *et al* (2001).

2.5.2 Sequence analysis

2.5.2.1 Full-length sequence assembly

The DNA sequences obtained was edited using the Chromas program (Griffith University, Brisbane, Queensland, Australia). The short sequence fragments were then incorporated into the Auto Assembler program (Applied Biosystems, Foster City, CA, USA), to put together longer fragments of DNA that overlap one another. These sequences were then put together as contigs. The different contigs were adjusted manually and full-length DNA sequences were verified using the DNA strider program (Marck, 1998).

2.5.2.2 Annotation of the genes

After the full-length sequences had been constructed, it was necessary to determine the open-reading frames of the full-length sequences. Viral sequences were imported to the DNAman program (Lynnon Biosoft, Vaudreuil-Dorion, Quebec, Canada), and converted to amino acid sequences for the three different reading frames. The HIV/SIV sequence locator tool at the HIV sequence database website (http://hiv-web.lanl.gov) was used to give an indication of the starting points of the different genes. Viral genes were then annotated from the first 'atg' codon observed, after the long terminal repeat region.

2.5.2.3 The NCBI subtyping of full-length sequences

The full-length sequences obtained were compared to other full-length sequences in Los Alamos, using the NCBI Subtyping Tool available at the Los Alamos website (http://hiv-web.lanl.gov). This subtyping program gives a fast indication of the composition (whether the sequence is that of HIV) of the DNA and to what subtype the sequence belongs to.

2.5.2.4 Simplot

To identify any recombination breakpoints, we used the similarity plot method as implemented in the SIMPLOT program for Microsoft Windows (Salminen *et al*, 1995a). In this program, a panel of reference sequences is moved across the query sequence. Analysis was done with a window of 400 bp moving along the alignment in increments of 20 bp. A total of 100 replicates were generated for each query sequence, plotting the percent similarity values of the query sequence with the sequence from the reference panel. The program uses the Kimura 2-parameter nucleotide substitution model (Kimura, 1980), with a transition/transversion value of 2.

2.6 Phylogenetic analysis

2.6.1 Datasets used for phylogenetic analysis

Phylogenetic analysis of the full-length sequences was carried out with the current 2001 HIV-1 subtype reference alignments obtained from the HIV sequence database (http://hiv-web.lanl.gov). Full-length subtype D sequences (**Table 2.3**) were downloaded to perform subtype specific phylogenetic analysis. From the full-length subtype D, the individual gene sequence of each strain was excised, for the comparison of subtype D specific genes.

2.6.2 Multiple alignment

Multiple alignments of the sequences were done using the Clustal X program (Thompson *et al*, 1997). All the gaps in the sequences were removed and full alignments were performed. Alignments were checked manually for any inconsistencies. The subtype D sequences from Tygerberg were compared to the current 2001 HIV-1 subtype reference sequences (http://hiv-web.lanl.gov) in a multiple alignment.

2.6.3 Phylogenetic tree analysis

Eleven phylogenetic trees were constructed for the following comparisons:

1. Tygerberg HIV-1 subtype D with 2001 HIV-1 subtype reference set

Table 2.3 Full-length Subtype D isolates (http://www.hiv.lanl.gov) (Accessed: 13-2-2004)

Subtype	Strain	Accession number	Country	Author
HIV-1 D	MB 2059	AFD 133821	KE	Neilson *et al* (1999)
	01KE_NKU3006	AF 457090	KE	Dowling *et al* (2002)
	99UGA07412	AF 484477	UG	Harris *et al* (2002)
	99UGB21875	AF 484480	UG	Harris *et al* (2002)
	99UGB25647	AF 484481	UG	Harris *et al* (2002)
	99UGB32394	AF 484483	UG	Harris *et al* (2002)
	99UGD23550	AF 484485	UG	Harris *et al* (2002)
	99UGD26830	AF 484486	UG	Harris *et al* (2002)
	99UGE08364	AF 484487	UG	Harris *et al* (2002)
	99UGE23438	AF 484489	UG	Harris *et al* (2002)
	99UGF05734	AF 484490	UG	Harris *et al* (2002)
	99UGF10555	AF 484494	UG	Harris *et al* (2002)
	99UGG35093	AF 484495	UG	Harris *et al* (2002)
	99UGJ27597	AF 484497	UG	Harris *et al* (2002)
	99UGK09259	AF 484498	UG	Harris *et al* (2002)
	99UGK09958	AF 484499	UG	Harris *et al* (2002)
	98UG57128	AF 484502	UG	Harris *et al* (2002)
	98UG57130	AF 484504	UG	Harris *et al* (2002)
	98UG57131	AF 484505	UG	Harris *et al* (2002)
	98UG57132	AF 484506	UG	Harris *et al* (2002)
	98UG57140	AF 484511	UG	Harris *et al* (2002)
	98UG57146	AF 484513	UG	Harris *et al* (2002)
	98UG57143	AF484514	UG	Harris *et al* (2002)
	99UGE 13613	AF 484515	UG	Harris *et al* (2002)
	99UGJ32228	AF 484516	UG	Harris *et al* (2002)
	99UGA03349	AF 484518	UG	Harris *et al* (2002)
	99UGF03726	AF 484519	UG	Harris *et al* (2002)
	92UG001 1-2	AJ 320484	UG	Novelli *et al* (2002)
	99TCD.MN011	AJ 488926	TD	Vidal (2003)
	99TCD.MN012	AJ 488927	TD	Vidal (2003)
	TZBFL0170-3-2	AY 237166	TZ	Koulinska *et al* (2003)
	99UGA08483	AY 304496	UG	Unpublished
	ELI	K 03454	CG	Alizon *et al* (1986)
	Z2	M 22639	CG	Srinivasan *et al* (1987)
	NDK	M 27323	CG	Spire *et al* (1989)
	84zr085	U 88822	ZR	Gao *et al* (1998)
	94UG114	U88824	UG	Gao *et al* (1998)

KE= Kenya; UG= Uganda; TD= Chad; TZ=Tanzania; CG=Congo; ZR=Zaire

2. Full length Tygerberg HIV-1 subtype D with Full length subtype D sequences from Los Alamos
3. Tygerberg HIV-1 D gag with Los Alamos HIV-1 D gag
4. Tygerberg HIV-1 D *pol* with Los Alamos HIV-1 D *pol*
5. Tygerberg HIV-1 D *env* with Los Alamos HIV-1 D *env*
6. Tygerberg HIV-1 D *vif* with Los Alamos HIV-1 D *vif*
7. Tygerberg HIV-1 D *vpr* with Los Alamos HIV-1 D *vpr*
8. Tygerberg HIV-1 D *vpu* with Los Alamos HIV-1 D *vpu*
9. Tygerberg HIV-1 D *tat* with Los Alamos HIV-1 D *tat*
10. Tygerberg HIV-1 D *rev* with Los Alamos HIV-1 D *rev*
11. Tygerberg HIV-1 D *nef* with Los Alamos HIV-1 D *nef*

For the comparison of the complete genomes with the reference set and the full-length sequences with the full-length subtype D sequences, neighbour-joining phylogenetic trees (Saitou and Nei, 1987) were constructed using the Treecon W program (Van de Peer and De Wachter, 1994). In the construction of the tree, all alignment positions were used to calculate the best tree. The Kimura 2-parameter nucleotide substitution model was used and 100 bootstrap replicates were performed. The subtype O sequence, O.CM.91.MVP.5180, was used as out-group to root all the trees with. This sequence is represented as follows: O indicates the subtype; CM represents Cameroon, the country of origin; 91 indicate the year of the sample followed by the sequence name MVP.5180.

The sub genomic regions of the virus was analysed with the same method as described above. These regions included were compared to the sub genomic regions of the LANL HIV-1 D sequences. The regions are: structural genes (*gag*, *pol* and *env*), regulatory genes (*vif*, *vpr* and *vpu*) and the accessory genes (*tat*, *rev* and *nef*).

2.6.4 Similarity between HIV subtypes

The similarity between the full-length sequences and the reference set as well as the similarity between the different genes was determined with the BioEdit

program (Hall, 1999). The PAM 250 matrix was used as the model to determine the similarity between the sequences.

2.7 Amino acid alignment and analysis of the subtype D Env protein

2.7.1 V3 alignment

The nucleotide sequences for the *env* gene of the subtype D sequences were converted to amino acid sequences in the DNAMAN program (Lynnon Biosoft, Vaudreuil-Dorion, Quebec, Canada). An alignment of the Env protein was then constructed in Clustal X (Thompson *et al*, 1997). The V3 region was excised from the alignment for comparison with the other subtype D sequences.

2.7.2 Glycosylation

The total number of N-linked glycosylation sites (Marshall, 1974) was determined with the N-GLYCOSITE program implemented in the HIV sequence database. The glycosylation was determined for each of the Tygerberg sequences as well as for the subtype A-D, F-H and K consensus sequences.

Chapter 3

RESULTS

Chapter 3

RESULTS

3.1 PCR Amplification and purification of the HIV-1 genome

PCR amplification was the first step in the project to characterise the HIV-1 subtype D sequences from South Africa. High molecular weight DNA was available and through a modification of the protocol by Salminen *et al* (1995a), it was possible to amplify the near full-length 9 kb genome of sequences: R2, R214, R286 and R482 with a single amplification reaction using the Expand long template PCR system. The PCR samples were loaded on an agarose gel (**Fig 3.1**) and a DNA fragment of the correct size was observed.

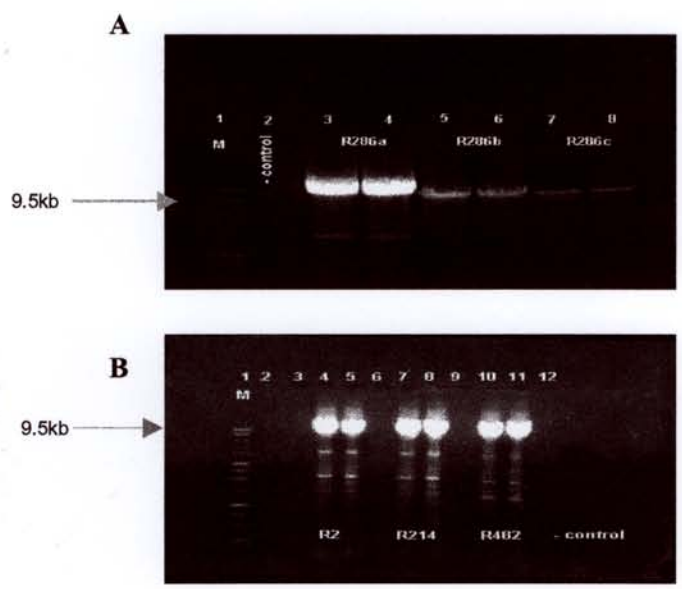


Figure 3.1. Near full-length PCR amplification of the HIV-1 subtype D DNA. Gel A represents R286 DNA amplified with three primer pairs: a) MSF12/MSR5 (lanes 3-4), b) UP1A/LOW2 (lanes 5-6) and c) UP1A/S2Full (lanes 7-8). Two reactions were loaded for each of the primer combinations. Gel B represents 2 reactions of each sample R2 (lanes4-5), R214 (lanes7-8) and R482 (lane10-11) amplified with the MSF12/MSR5 primer pair. On both gels, M is the 1kb molecular weight marker. The negative control is marked (-). The arrow points to the band size between the top (10kb) and second band (8 kb) of the DNA ladder.

In fig 3.1(A), it is clear that the primer pair MSF12/MSR5 gave the best DNA amplification of R286. The UP1A/LOW2 combination also gave a better amplification result than the UP1A/S2Full pair that was designed by J. zur Megede for subtype C (zur Megede *et al*, 2002). From this result, it was decided that the remaining sequences would all be amplified with the MSF12/MSR5 primer pair. Visible on the top gel is the non-specific amplification observed for each of the primer pair sets. In fig 3.1 B, sequences R2, R214 and R482 gave clear amplification results with the MS-primer set, even though some non-specific amplification was observed here as well. In both gels, no bands are visible in the negative control lane, indicating that no contamination was present. The PCR products that displayed the expected 9 kb banding size were excised from the agarose gel with a sterile razor. After purification with the QIAEX II Agarose Gel Extraction Kit (Qiagen, GmbH, Germany), the concentrations of the purified plasmid are indicated in **Table 3.1.**

Table 3.1. DNA concentration of the 9kb gel purified PCR fragments

DNA	OD260	Concentration (µg/µl)
R2	0.019	0.066
R214	0.048	0.168
R286	0.033	0.115
R482	0.015	0.052

3.2 Cloning the PCR fragments and small-scale DNA preparations

The TOPO cloning vector and cloning kit, designed for cloning large fragments, made it easy to clone all four purified DNA products (R2, R214, R286 and R482) into the vector. The recombinant vector grew stable at 33°C making it possible to grow and extract large quantities of plasmid DNA for sequencing. Because of the large size of the insert (9kb), only a few colonies were observed on the agar plates. It is suggested that less than 20 colonies per plate should be

observed if the correct insert had been cloned (Salminen *et al,* 1995a). The total number of minipreps performed for each sequence is indicated in **Table 3.2**.

Table 3.2. The number of recombinant clones obtained for each strain

Strain	Number of minipreps	Number of positive clones	Names of clones	% Efficiency
R2	18	6	pR2.7 pR2.8 pR2.9 pR2.10 pR2.11 pR2.12	33
R214	6	1	pR214.5	16
R286	21	1	pR286.2	4
R482	15	3	pR482.3 pR482.7 pR482.9	20

In total, 60 DNA miniprep isolations had been performed for the 4 samples. R2 yielded the most clones. One clone was obtained for each of the recombinant plasmids, pR214 and pR286.

One clone of each of the 4 samples was randomly selected for sequencing. The optical density and concentration of the selected clones are indicated in **Table 3.3.**

Table 3.3. Concentration of HIV-1 plasmid DNA for sequencing

Plasmid	OD260	OD280	Concentration (µg/µl)	Purity
R2.7	0,637	ND	1.1	ND
R214.5	0.454	ND	0.7	ND
R286.2	0.281	0.149	0.491	1.88
R482.9	0.620	0.721	1.086	1.16

ND=Not determined.

The plasmids generally gave OD_{260} values above 0.450, except for plasmid pR286 whose reading was 0.281. Even though plasmid pR482 gave a higher OD_{260} reading than plasmid pR286, it was less pure. For sequencing only 1µg DNA per reaction is needed, which means that plasmids pR2 and pR482 had to be diluted to obtain the correct input concentration. The input volume of plasmids pR214 and pR286 had to be increased in the reaction to obtain the correct concentration for sequencing.

3.3 DNA sequencing

The primers used for sequencing the complete genome worked well. In total, 81 different primers were used to sequence the 4 plasmids. All the primers designed to fill the gaps in the genome gave readable sequence electropherogram results. The primers used to sequence the DNA are listed in **Appendix A.**

3.4 Analysis

3.4.1 Annotation of genes

The Auto Assembler program enabled us to assemble the sequenced DNA fragments to construct contiguous fragments. Once the near full-length sequence was obtained, the HIV/SIV sequence locator tool at the HIV sequence database was used to give an indication of the starting points of the different viral genes. The viral genes were then annotated from the first 'atg' codon observed. The full-length sequence for each plasmid pR2, pR214, pR286 and pR482 has been determined as well as the open reading frames of the sequences. The full-length sequences with the coding amino acids are given in **Appendix B**. In **Tables 3.4 – 3.7** the gene positions of the sequences from the *gag* gene are indicated. All the sequences had genes of similar length as the full-length HIV-1 subtype D sequences from Los Alamos. No premature stop codons had been observed in any of the genes for the plasmids pR2, pR214, pR286 and pR482.

3.4.2 NCBI Subtyping of the Tygerberg plasmid sequences

The subtyping results for plasmids pR2, pR214, pR286 and pR482 are shown in **Appendix C**. The results show that the 4 plasmids are complete HIV-1 subtype D sequences. It should be noted that the subtyping tool might give very misleading results in cases where the query sequence has large inserts or deletions and should only be used for exploratory work and should be followed up by analyses based on aligned sequences (Kuiken and Leitner, 2001).

Table 3.4 Nucleotide position on the HIV genome relative to plasmid pR2

Region of the genome	Start nucleotide	End nucleotide
***Gag* gene**		
gag Pr55 precursor	234	1739
gag p17 Matrix	234	632
gag p24 Capsid	633	1325
gag p2	1326	1370
gag p7	1371	1535
gag p1	1536	1583
gag p6	1584	1739
***Pol* gene**		
Pol polyprotein	1535	4546
Pol p10 Protease	1700	1996
Pol p51 RT	1997	3319
Pol p15 Rnase	3320	3679
Pol p31 integrase	3680	4546
***Vif* gene**	4491	5069
***Vpr* gene**	5009	5299
***Tat* gene**		
Exon 1	5280	5494
Exon 2	7798	7888
***Rev* gene**		
Exon 1	5419	5494
Exon 2	7798	8072
***Vpu* gene**	5511	5756
***Env* gene**		
gp 160	5674	8214
gp 41	7201	8214
***Nef* gene**	8216	8839

Table 3.5 Nucleotide position on the HIV genome relative to plasmid pR214

Region	Start nucleotide	End nucleotide
***Gag* gene**		
gag Pr55 precursor	214	1710
gag p17 Matrix	214	612
gag p24 Capsid	613	1302
gag p2	1303	1344
gag p7	1345	1506
gag p1	1507	1554
gag p6	1555	1710
***Pol* gene**		
Pol polyprotein	1506	4502
Pol p10 Protease	1671	1967
Pol p51 RT	1968	3278
Pol p15 Rnase	3279	3638
Pol p31 integrase	3639	4502
***Vif* gene**	4447	5022
***Vpr* gene**	4962	5252
***Tat* gene**		
Exon 1	5233	5444
Exon 2	7748	7838
***Rev* gene**		
Exon 1	5369	5444
Exon 2	7748	8018
***Vpu* gene**	5461	5709
***Env* gene**		
gp 160	5624	8155
gp 41	7139	8155
***Nef* gene**	8157	8780

Table 3.6 Nucleotide position on the HIV genome relative to plasmid pR286

Region	Start nucleotide	End nucleotide
***Gag* gene**		
gag Pr55 precursor	235	1743
gag p17 Matrix	235	633
gag p24 Capsid	634	1326
gag p2	1327	1368
gag p7	1369	1539
gag p1	1540	1587
gag p6	1588	1743
***Pol* gene**		
Pol polyprotein	1539	4547
Pol p10 Protease	1704	2003
Pol p51 RT	2004	3320
Pol p15 Rnase	3321	3680
Pol p31 integrase	3681	4547
***Vif* gene**	4492	5070
***Vpr* gene**	5010	5300
***Tat* gene**		
Exon 1	5281	5495
Exon 2	7805	7892
***Rev* gene**		
Exon 1	5420	5495
Exon 2	7805	8076
***Vpu* gene**	5513	5758
***Env* gene**		
gp 160	5676	8225
gp 41	7191	8225
***Nef* gene**	8227	8850

Table 3.7 Nucleotide position on the HIV genome relative to plasmid pR482

Region	Start nucleotide	End nucleotide
***Gag* gene**		
gag Pr55 precursor	231	1736
gag p17 Matrix	231	629
gag p24 Capsid	630	1322
gag p2	1323	1367
gag p7	1368	1532
gag p1	1533	1580
gag p6	1581	1736
***Pol* gene**		
Pol polyprotein	1532	4540
Pol p10 Protease	1697	1993
Pol p51 RT	1994	3313
Pol p15 Rnase	3314	3673
Pol p31 integrase	3674	4540
***Vif* gene**	4485	5069
***Vpr* gene**	5009	5293
***Tat* gene**		
Exon 1	5274	5488
Exon 2	7807	7897
***Rev* gene**		
Exon 1	5413	5488
Exon 2	7807	8081
***Vpu* gene**	5505	5750
***Env* gene**		
gp 160	5668	8223
gp 41	7186	8223
***Nef* gene**	8225	8848

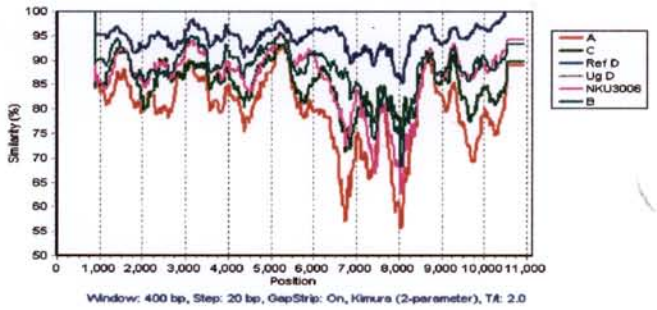
3.4.3 Simplot

The similarity plots of the Tygerberg full-length sequences with the selected reference full-length sequences are depicted in Figure 3.4 (A-D). The grouping system implemented in Simplot was used to screen our sequences against full-length subtypes A, B, C and D strains from Los Alamos. The subtype D strains consisted of reference subtype D (Eli, NDK, Z2Z6 and 84ZR085), Ugandan subtype D (subtype D sequences from Uganda between 1998-1999) and NKU (a subtype D strain from Kenya 2001). The Simplot graphs indicate that there is more than 95% similarity between the Tygerberg sequences and the reference subtype D sequences. An average of 90% similarity is seen between the Tygerberg and the subtype B sequences in the graphs. The Simplot graphs also indicate that the Tygerberg sequences are least similar to the subtype A sequence. A window size of 400 bp and 20 bp increments for the similarity plots was enough to show that the Tygerberg plasmid sequences displayed nonmosaic sequences.

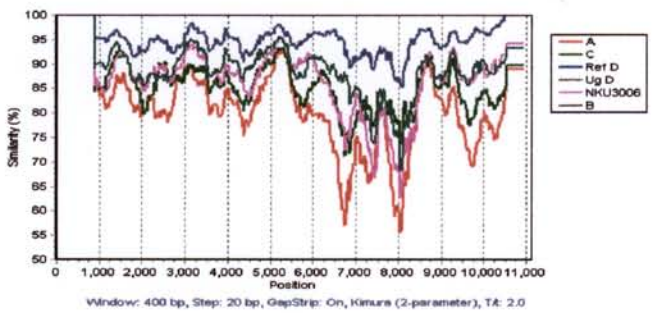
3.5 Phylogenetic analysis

We constructed evolutionary phylogenetic trees to determine the relationship between the near full-length genomes of sequences R2, R214, R286 and R482 and non-recombinant reference and Ugandan subtype D strains from the database. A total of 11 multiple alignments had been performed. Two full-length alignments were also performed: one alignment to compare the Tygerberg subtype D sequences to the full-length reference alignment and the other to compare the Tygerberg full-length strains to the full-length subtype D sequences in the database.

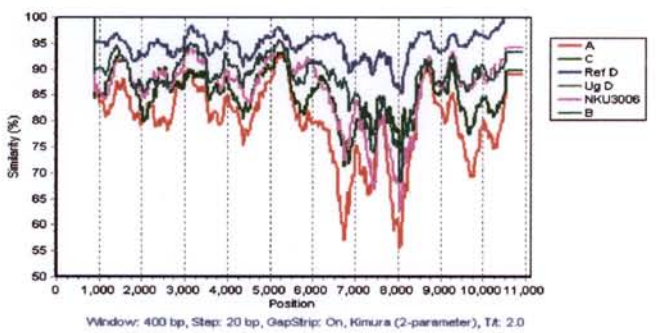
A: pR2



B: pR214



C: pR286



D: pR482

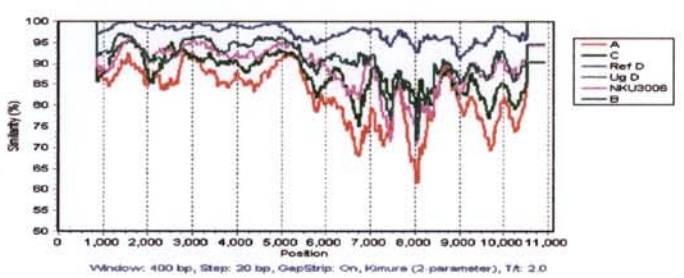


Figure 3.4 (A-D). Similarity Plot (Simplot) of the HIV-1 plasmids (R-strains). The legend on the right hand side indicates the different subtypes. The Y-axis represents the similarity (%) of the sequences. On the x-axis is the position relative to the HIV-1 sequence in question. A window size of 400 bp with 20 bp step increments was used. The Kimura (2-parameter) model with a transition: transversion ration of 2 was used.

3.5.1 Complete genomes

3.5.1.1 Alignment of Tygerberg complete genomes with the reference set

The phylogenetic tree depicting the sequences of the Tygerberg sequences compared to the 2001 reference strains is shown in **figure 3.6**. From figure 3.6, the genetic subtypes of HIV-1 can clearly be distinguished, as the different subtypes cluster together. Comparing the full-length reference dataset from Los Alamos with our sequences, a clear cluster with 100% bootstrap values of our sequences can be observed. Closely related to the Tygerberg sequences are the reference subtype D strains, which forms a cluster with 100% bootstrap value. Sequence R2 forms a branch with sequence R214 and sequences R286 and R482 cluster together. From the figure, it is also evident that the NDK and Eli strains are closer related to the Tygerberg strains than to the other subtype D reference strains (94UG114 and 84ZR085). The subtype B strains cluster with high bootstrap support (100%) close to the subtype D sequences. A 10% sequence divergence is indicated on the scale in figure 3.6. The phylogenetic tree is rooted with the subtype O strain from Cameroon.

3.5.1.2 Full-length subtype D strains

The phylogenetic comparison between the full-length Tygerberg strains and the other full-length subtype D sequences in the LANL database are shown in **figure 3.7**. Compared only to full-length subtype D sequences, the Tygerberg strains are more than 70% related to the other full-length sequences and up to 92% similar to each other. The Tygerberg strains again forms a separate cluster with 100% bootstrap values. From the tree, three groups can be seen. The top group (indicated with a blue bracket) with a bootstrap value of 94% consist mainly of the 1998-1999 full-length subtype D sequences from Uganda. Also present in this group is the Kenya strains (MB2059 and NKU3006). This group represent strains that seem to have evolved at the same rate or was sampled at the same time, as indicated by the almost similar branch lengths. The bottom group (indicated by the red and green brackets) of the tree is divided into two sections. The group indicated by the red bracket contains the strains from Chad

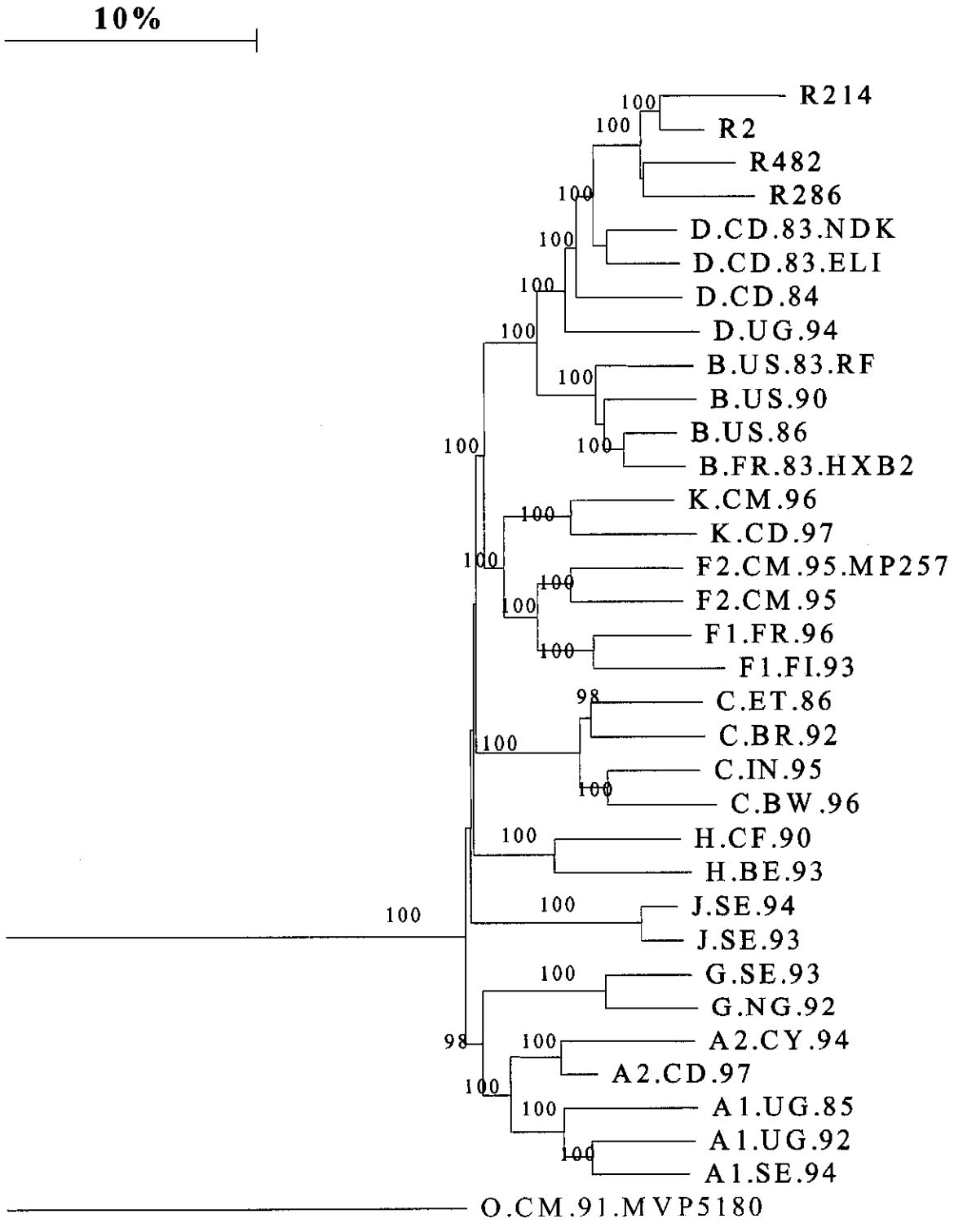


Figure 3.6. A neighbour-joining phylogenetic tree of the full-length dataset and the Tygerberg sequences (Red). In light blue are the HIV-1 subtype D full-length strains. Bootstrap values greater than 70% are shown at the major nodes. Branch lengths are drawn to scale.

and the group indicated by the green bracket contains the Tygerberg and the reference D strains. From figure 3.7 there seems to be a separation between the older (1983-1985) subtype D sequences and the newer (1999-2001) sequences. An interesting observation is the fact that the strains from Chad (99TCD.MN011 and 99TCD.MN012) are more closely related to the reference subtype D strains then to the strains from Rakai (Harris *et al*, 2002), Uganda from the same year. This is in conjunction with Vidal *et al* (2003) who found the full-length sequences from Chad to be different from the sequences from East Africa. The strains from Chad share a similarity with the Tygerberg strains ranging between 72 and 77%. Bootstrap values greater than 70% are indicated on the tree.

3.5.2 Subgenomic fragments

A) *Gag* gene

The *gag* gene is 1.5 kb in length and is translated into a 55 kDa polyprotein precursor ($Pr55^{Gag}$), which can produce non-infectious, virus-like particles in the absence of other viral proteins or packageable viral RNA (Freed, 1998). Comparing to the other genes in the HIV-1 genome, intersubtype diversity within the *gag* gene ranges from 15% (Caumont *et al*, 2001; Harris *et al*, 2002), making *gag* one of the most conserved genes of the virus. Phylogenetic analysis of the *gag* gene shows that the Tygerberg strains and the reference subtype D strains cluster with a 100% bootstrap value. The similarity between the Tygerberg sequences range from 86.4% (R214 with R286) to 92.4% (R2 with R214). The structure of the *gag* phylogenetic tree resembles that of the full-length subtype D tree. The *gag* phylogenetic tree is shown in **figure 3.8**. In the gag p7 region, a duplication of 12 nucleotides corresponding to a 4 amino acid duplication, NFKG, is seen in the Tygerberg sequences, except for R286 who has the sequence NFYG (data not shown). The sequence is situated close to the first zinc finger motive in HIV. Conservation of the sequence suggests that the region has functional significance, perhaps the NFKG sequence has a role in binding to the viral RNA (Laukkanen *et al*, 1996).

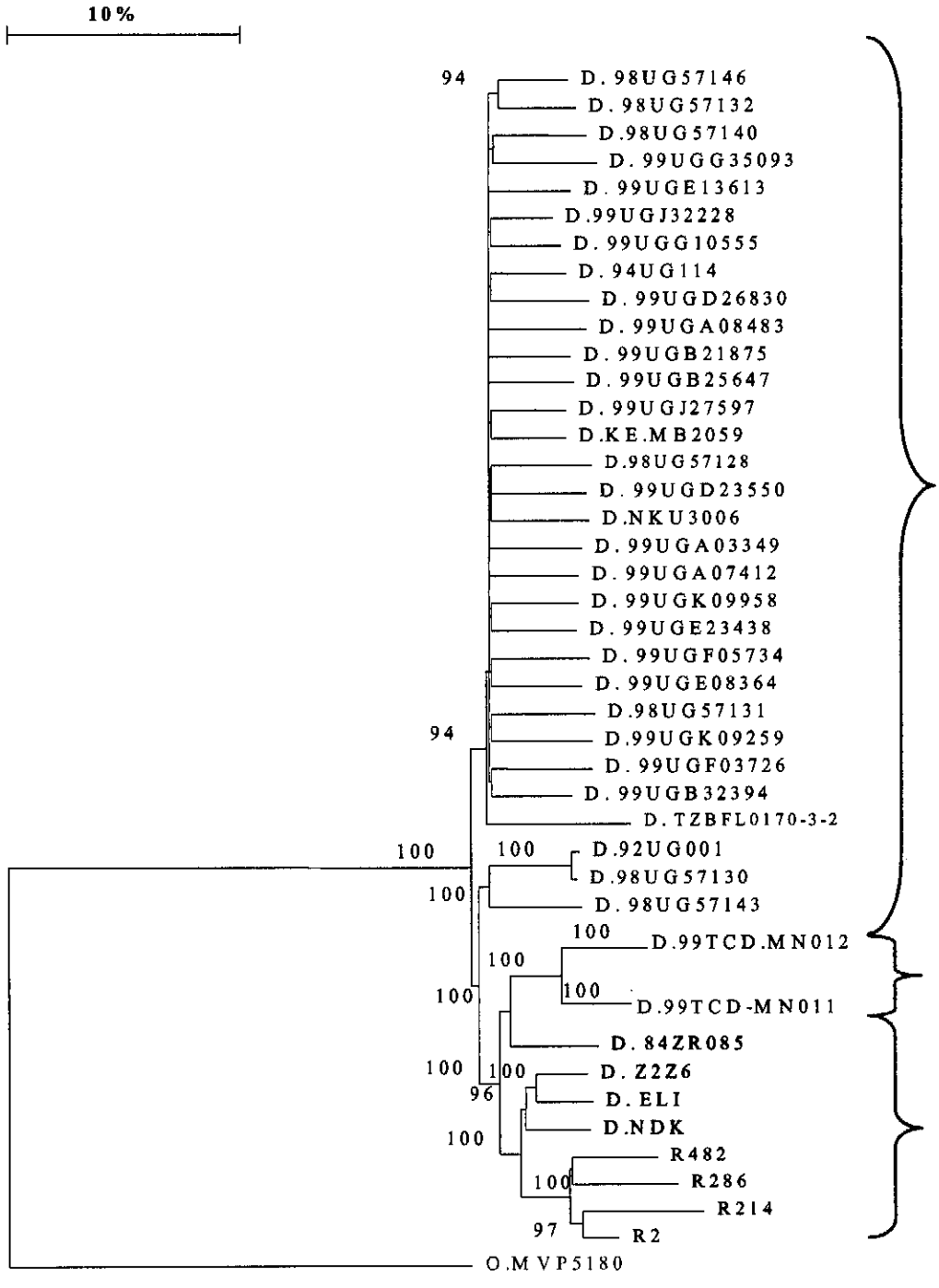


Figure 3.7. A neighbour joining phylogenetic tree comparing the complete genome DNA sequences of the Tygerberg sequences, indicated in red colour and the available full-length HIV-1 subtype D sequences. The reference subtype D sequences are indicated in dark blue and the subtype O sequence in light blue. Bootstrap values greater than 70% are indicated. The horizontal scale indicates the percentage variation between sequences.

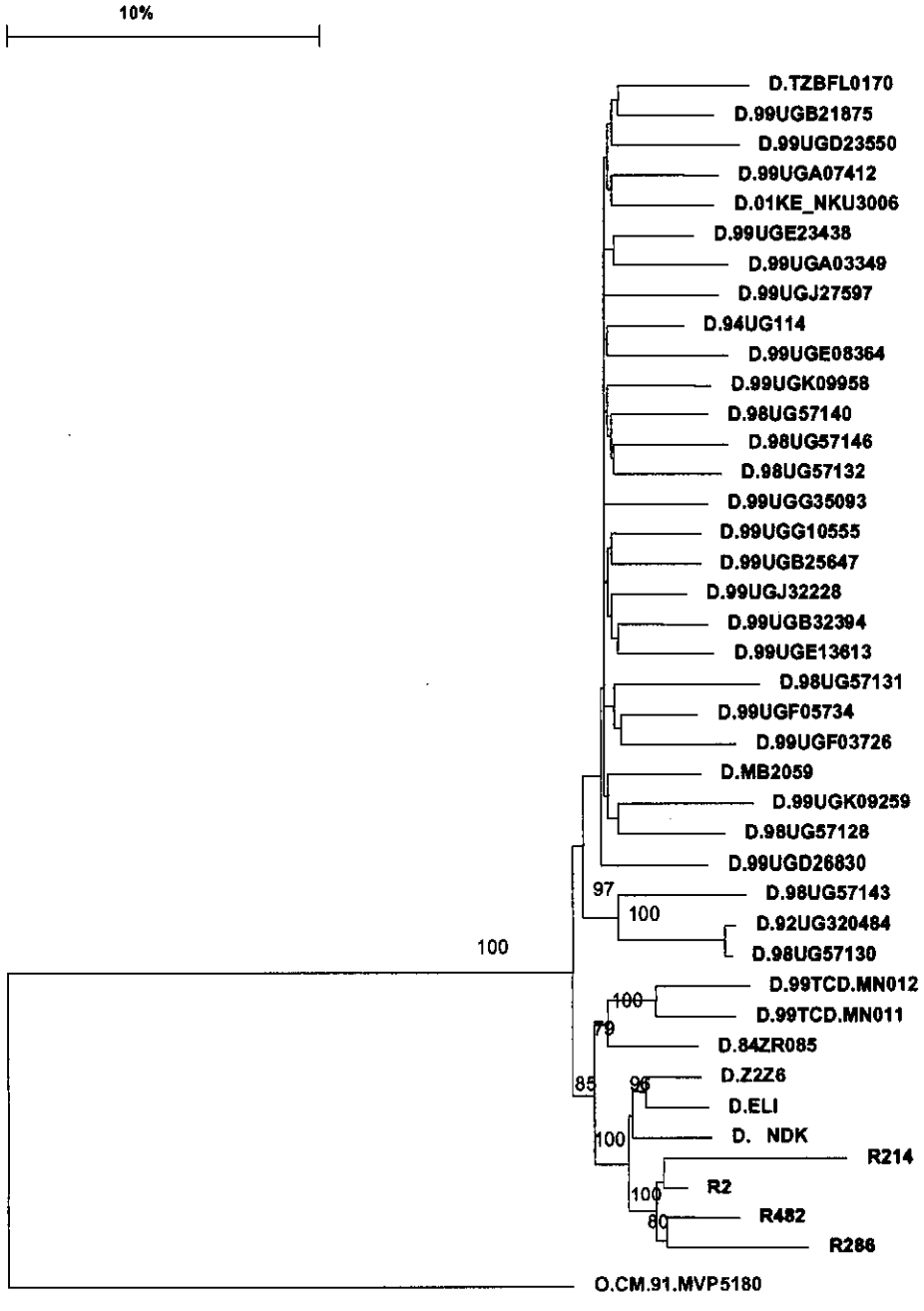


Figure 3.8. A neighbour joining phylogenetic tree comparing the complete *gag* DNA sequences of the Tygerberg sequences indicated with the red colour, with *gag* HIV-1 subtype D sequences. The reference subtype D sequences are indicated in dark blue and the subtype O sequence in light blue. Bootstrap values greater than 70% are indicated. The horizontal scale indicates the percentage variation between sequences.

B) *Pol* gene

The *pol* phylogenetic tree is depicted in **figure 3.9**. The *pol* sequences of the Tygerberg strains had greater similarity than the *gag* gene. The similarity between the Tygerberg strains ranged from 89.6% (R214 with R286) to 95.7% (R2 with R482). The phylogenetic analysis resulted in a similar tree as *gag*. In comparison with the other strains, the Tygerberg strains are more than 90% similar in gene sequence, except for strain R214 who had similarities of 80% with the other subtype D *pol* sequences.

C) *Env* gene

Sequence heterogeneity is a characteristic of the *env* gene and five variable regions (V1-V5) interspersed with more conserved regions (C1-C5) have been identified (Starcich *et al*, 1986). Great similarities (86%-90%) had been achieved between the Tygerberg strains. Similarities between the Tygerberg strains and the other subtype D *env* sequences *ranged* from 78% to 86%. The Tygerberg strains forms a separate cluster with 100% bootstrap value.

The V3 loop plays an important role in syncytium inducing phenotype and viral tropism (O'Hagen *et al*, 2003; Dragic *et al*, 1996). Compared to other group M subtypes, subtype D strains demonstrate a highly variable pattern of V3 loop amino acids (Spira *et al*, 2003). This is also evident from the Env alignment in **figure 3.11**. The alignment shows the V3 region and the flanking amino acids. The length of the V3 loop in the alignment below varied from 34 - 37 amino acids. All the sequences have a cysteine residue on both sides of the loop. At the crown of the V3 loop, seven different tetrameric sequences are visible. Most of the sequences have the GPGQ motif. All four the Tygerberg sequences have the GQGQ motif. The other motifs seen are the GPGA, GIGQ GPGL, GPGR and GLGQ. At position 11 and 25 of the V3 loop, all the Tygerberg sequences have positively charged amino acids (arginine (R) and lysine (K)), and their viral phenotype are therefore of the syncytium inducing (SI) type. This is in correlation with the results obtained by Engelbrecht *et al*, 1995 who grew the isolates in culture.

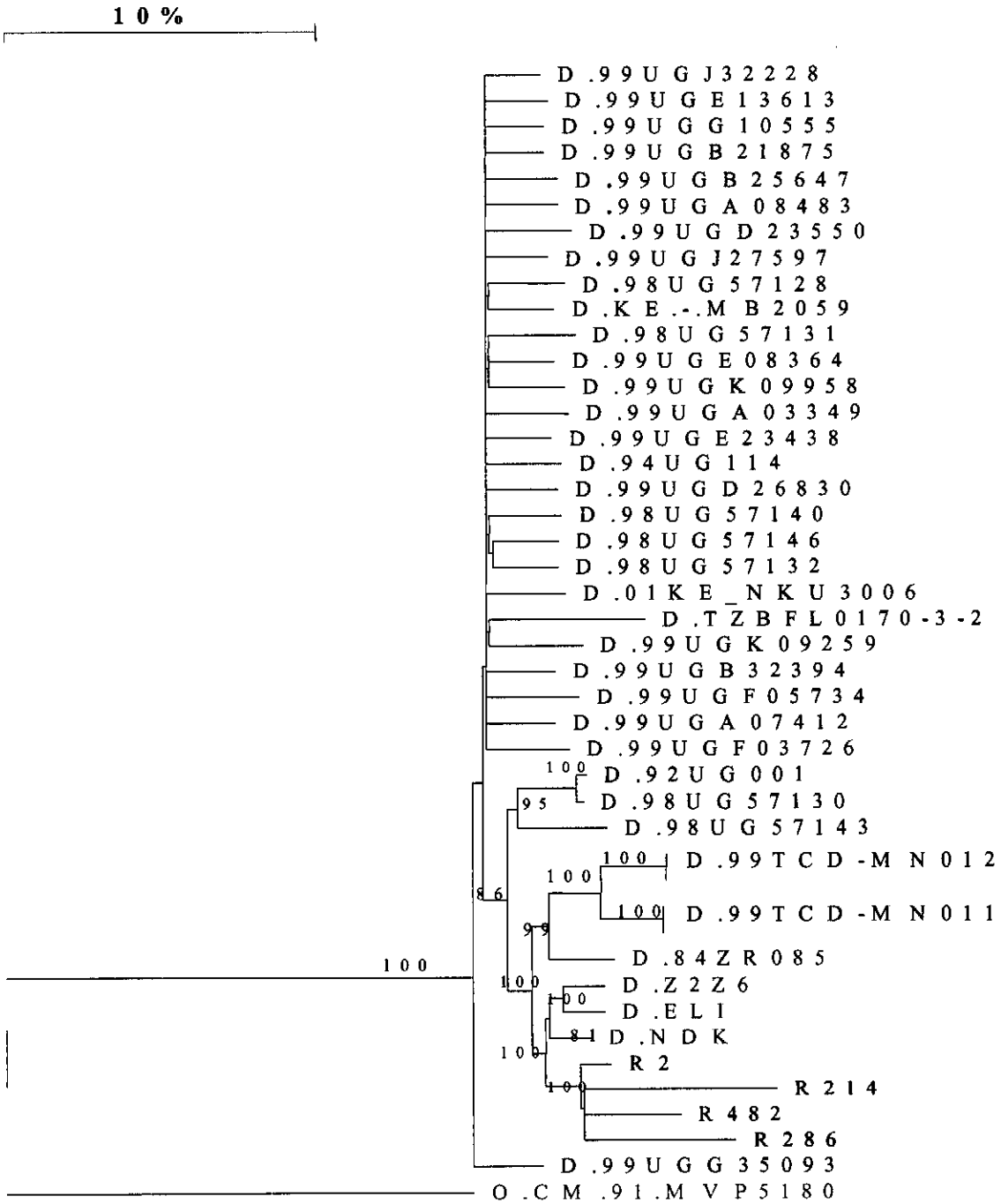


Figure 3.9. A neighbour joining phylogenetic tree comparing the complete *pol* DNA sequences of the Tygerberg sequences (Red) with the available complete *pol* HIV-1 subtype D sequences. The reference subtype D sequences are indicated in dark blue and the subtype O sequence in light blue. Bootstrap values greater than 70% are indicated. The horizontal scale indicates the percentage variation between sequences.

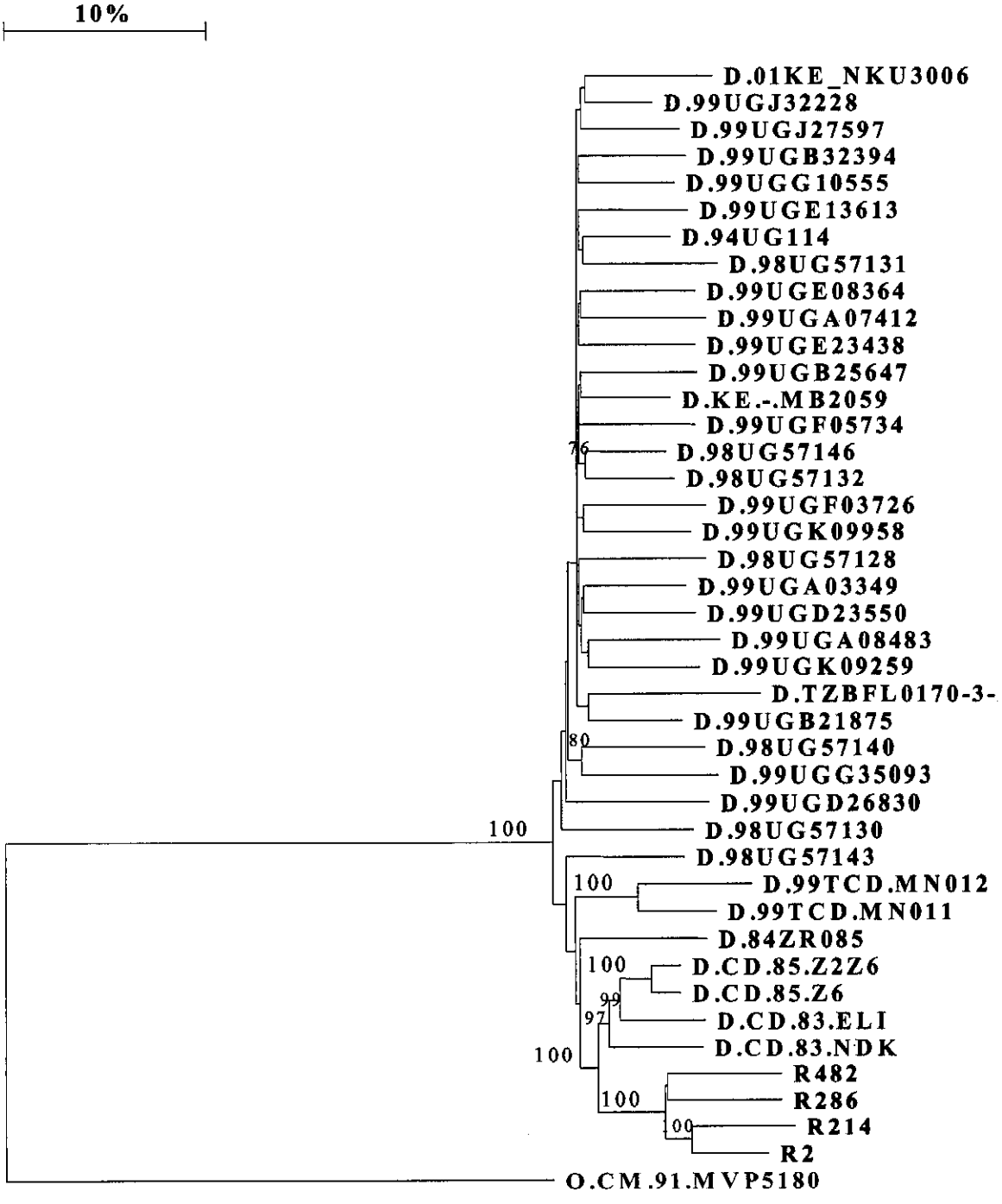


Figure 3.10. A neighbour joining phylogenetic tree comparing the complete *env* DNA sequences of the Tygerberg sequences indicated in red colour with the HIV-1 subtype D sequences. The reference subtype D sequences are indicated in dark blue and the subtype O sequence in light blue. Bootstrap values greater than 70% are indicated. The horizontal scale indicates the percentage variation between sequences.

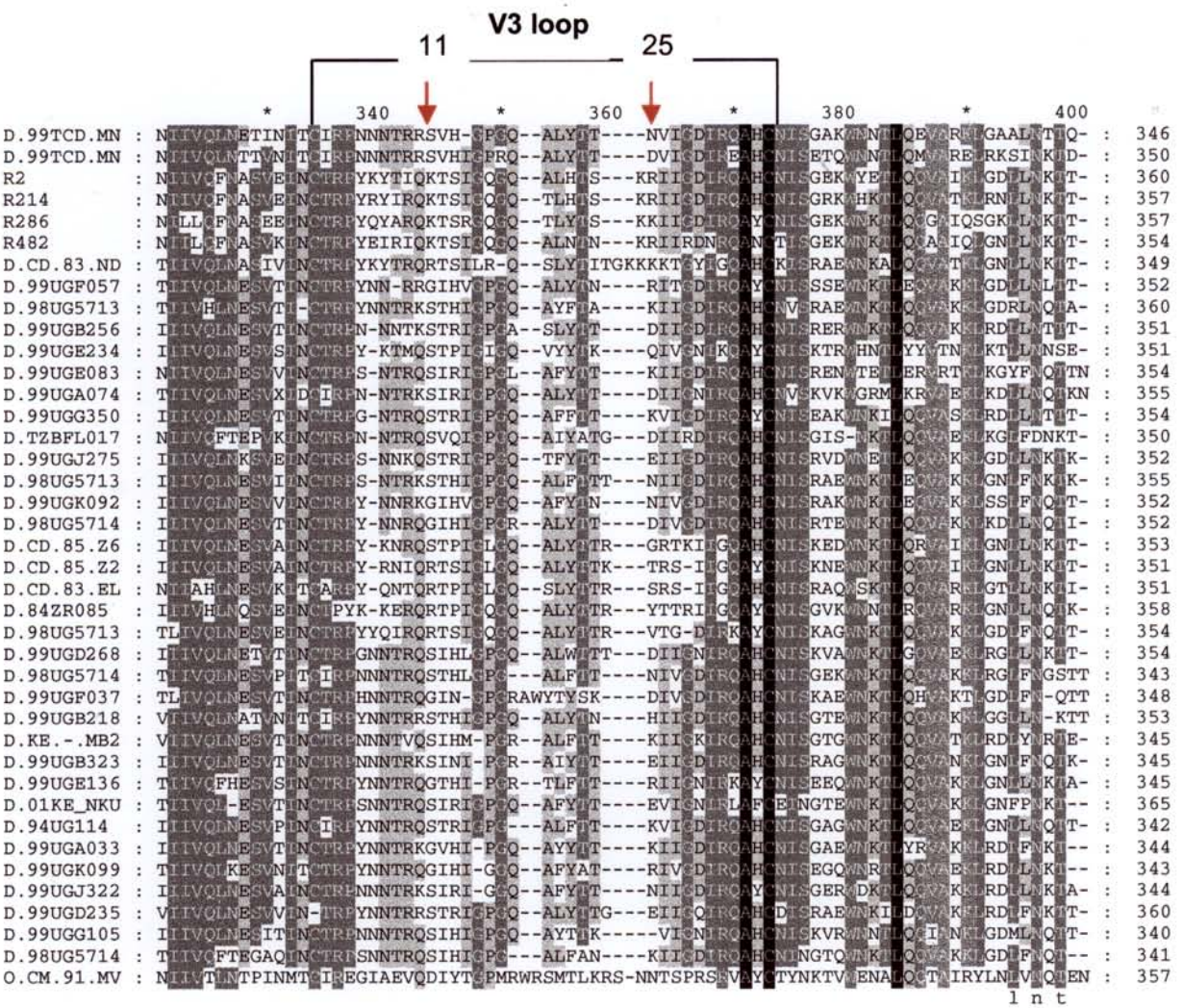


Figure 3.11. The partial *env* alignment of the different subtype D strains showing the V3 loop. Indicated in the figure is the 33-37 amino acid V3 loop, with the cysteine residues on both ends of the loop. Indicated with the red arrows are positions 11 and 25 relative to the Tygerberg strains.

D) *Vif* gene

Vif (virion infectivity factor) protein is essential for productive HIV-1 infection of peripheral blood lymphocytes and macrophages, the two major HIV-1 target cells *in vivo*. However, Vif is not required for production of infectious particles in several human cell lines *in vitro*. In spite of the dominant genotype of Vif mutations, the mechanism of its action remains unknown (Baraz and Kotler, 2004). In the phylogenetic tree of the *vif* gene, the Tygerberg strains again forms a separate cluster, this time with lower bootstrap values (85%). Similarities between the Tygerberg sequences ranged from 90.4% (R214 with R482) to 94.3% (RR2 with R286). Compared to the other strains, the *vif* similarity ranged from 88.7% to 93.2%.

E) *Vpr* gene

HIV-1 viral protein R (Vpr) is a small, highly conserved accessory protein encoded by the HIV genome that serves many functions in the viral life cycle. Vpr induces G2 cell cycle arrest, which is thought to indirectly enhance viral replication by increasing transcription from the LTR. Vpr has also been implicated in facilitating infection of non-dividing cells, most notably macrophages (Heinzinger *et al*, 1994). Because *Vpr* is a nucleo-cytoplasmic shuttling protein, its role in enhancing viral replication in macrophages may be mediated through enhanced entry of the HIV preintegration complex through the limiting nuclear pore (Sherman *et al*, 2002). In the *vpr* phylogenetic tree, the gene is conserved amongst the Tygerberg strains as can be seen from the high sequence similarity between the strains. Strains R2 and R214 share a 97% similarity. In the tree, R2 and R214 group together, while strains R286 and R482 group together. Similarities between the Tygerberg strains and the other subtype D *vpr* sequences are as high as 94%.

F) *Vpu* gene

Vpu, a membrane protein from HIV-1, folds into two distinct structural domains with different biological activities: a transmembrane (TM) helical domain involved in the budding of new virions from infected cells, and a cytoplasmic

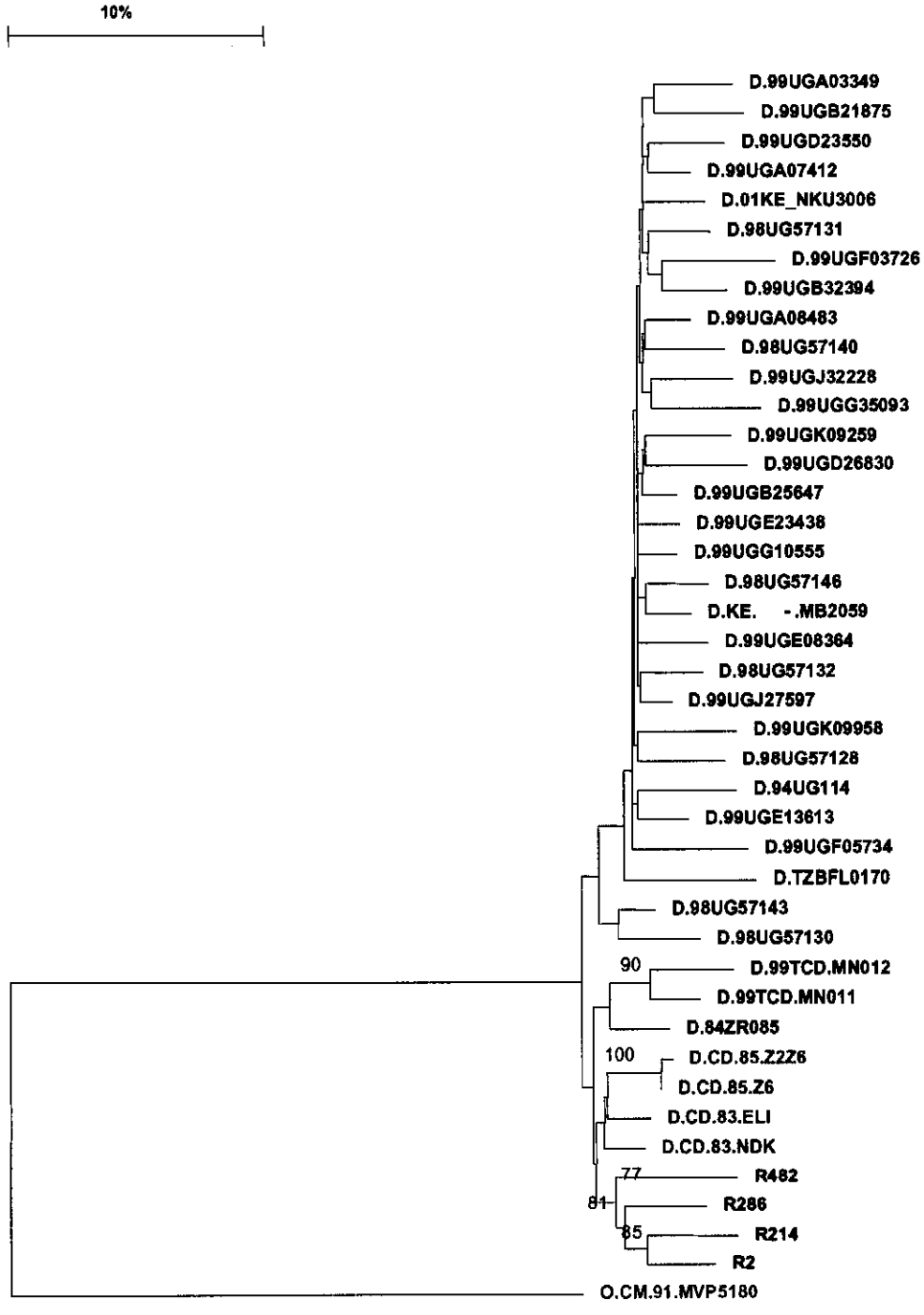


Figure 3.12. A neighbour joining phylogenetic tree comparing the complete *vif* DNA sequences of the Tygerberg sequences (R-strains; Red) with the HIV-1 subtype D sequences. Bootstrap values greater than 70% are indicated. The horizontal scale indicates the percentage variation between sequences.

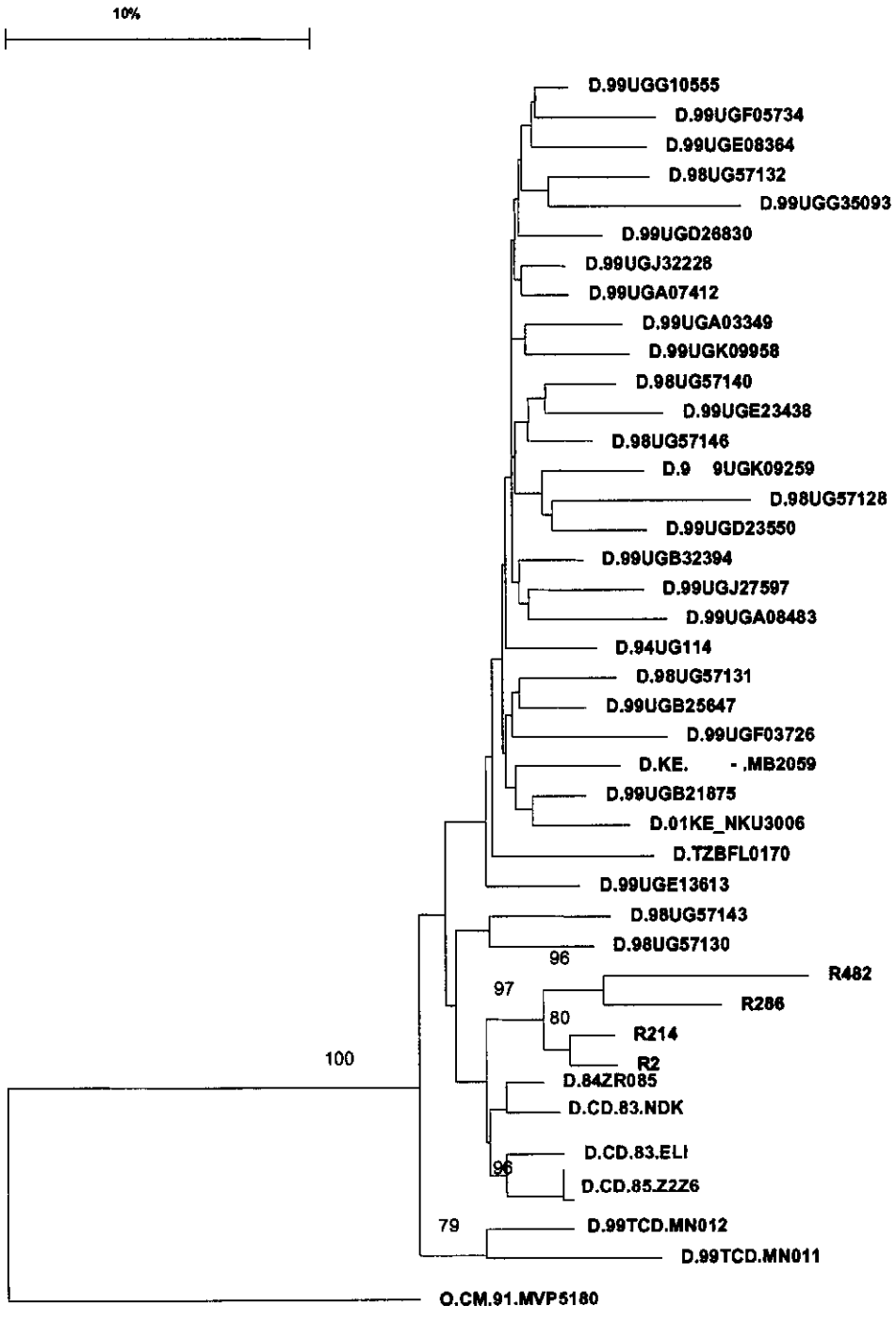


Figure 3.13. A neighbour joining phylogenetic tree comparing the complete *vpr* DNA sequences of the Tygerberg sequences indicated in red colour with the HIV-1 subtype D sequences. Bootstrap values greater than 70% are indicated. The horizontal scale indicates the percentage variation between sequences.

domain encompassing two amphipathic helices, which is implicated in CD4 degradation. The molecular mechanism by which Vpu facilitates virion budding is not clear. This activity of Vpu requires an intact TM helical domain. In addition, it is known that oligomerisation of the Vpu TM domain results in the formation of sequence-specific, cation-selective channels. It has been shown that the channel activity of Vpu is confined to the TM domain, and that the cytoplasmic helices regulate the lifetime of the Vpu channel in the conductive state (Montal, 2003). In the *vpu* phylogenetic tree the Tygerberg strains share a similarity above 83%. In the tree, strain R214 groups with NDK, while R2 and R286 group. The Tygerberg sequences share similarities of greater than 75% with the other subtype D *vpu* sequences. R214, however, is very different from strain 99UGD23550 whit whom it shares a similarity of only 68.4%. There is also a low similarity between 99UGD23550 and the other sequences if taken into account that *vpu* are conserved.

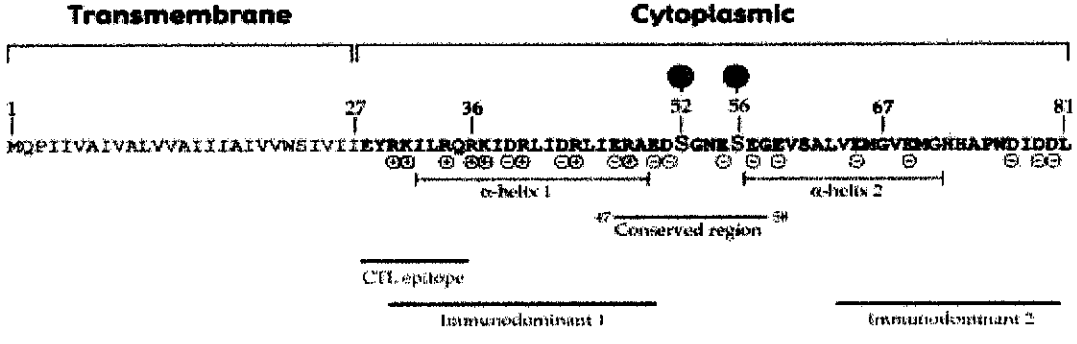


Figure 3.14. Annotated sequence of the HIV-1 (NL4-3) Vpu protein. The + and – symbols represent the global charge of the amino acid residues depicted. The two highly conserved and phosphorylated (P) serine residues are indicated at positions 52 and 56. The location of the two alpha-helical structures and the three immunodominant epitopes is also indicated (Bour and Strebel, 2003).

The amino acid alignment of the *vpu* gene of the subtype D strains is given in **figure 3.16**. In the alignment, it is clear that the amino acid sequence of R214 is different to the other subtype D *vpu* sequences. The sequence of R214 differs from the other D sequences in an area of conserved amino acids. Strain UGD23550 has an 8 amino acid insertion in the Vpu sequence. It is the only sequence with the insertion and may account why it differs so much from R214.

G) *Tat* gene

Tat is an 86-101 amino acid protein (Bayer *et al*, 1995). The amino terminus of tat, which extends from residues 1 to 21, contains three repeats of Proline-XXX-Proline in addition to acidic amino acids at positions 2, 5 and 9 (Gaynor, 1995). This region is followed by a domain extending from residues 22 to 37, which contains seven cysteine residues potentially capable of binding divalent ions such as cadmium and zinc (Gaynor, 1995). The *tat* phylogenetic tree has the same picture as the full-length tree, indicating that the Tygerberg strains are non-mosaic in the *tat* region. Similarities between the Tygerberg sequences are as high as 95.7%. When compared to the other sequences, similarities between the Tygerberg sequences and the other subtype D *tat* sequences reach 94%.

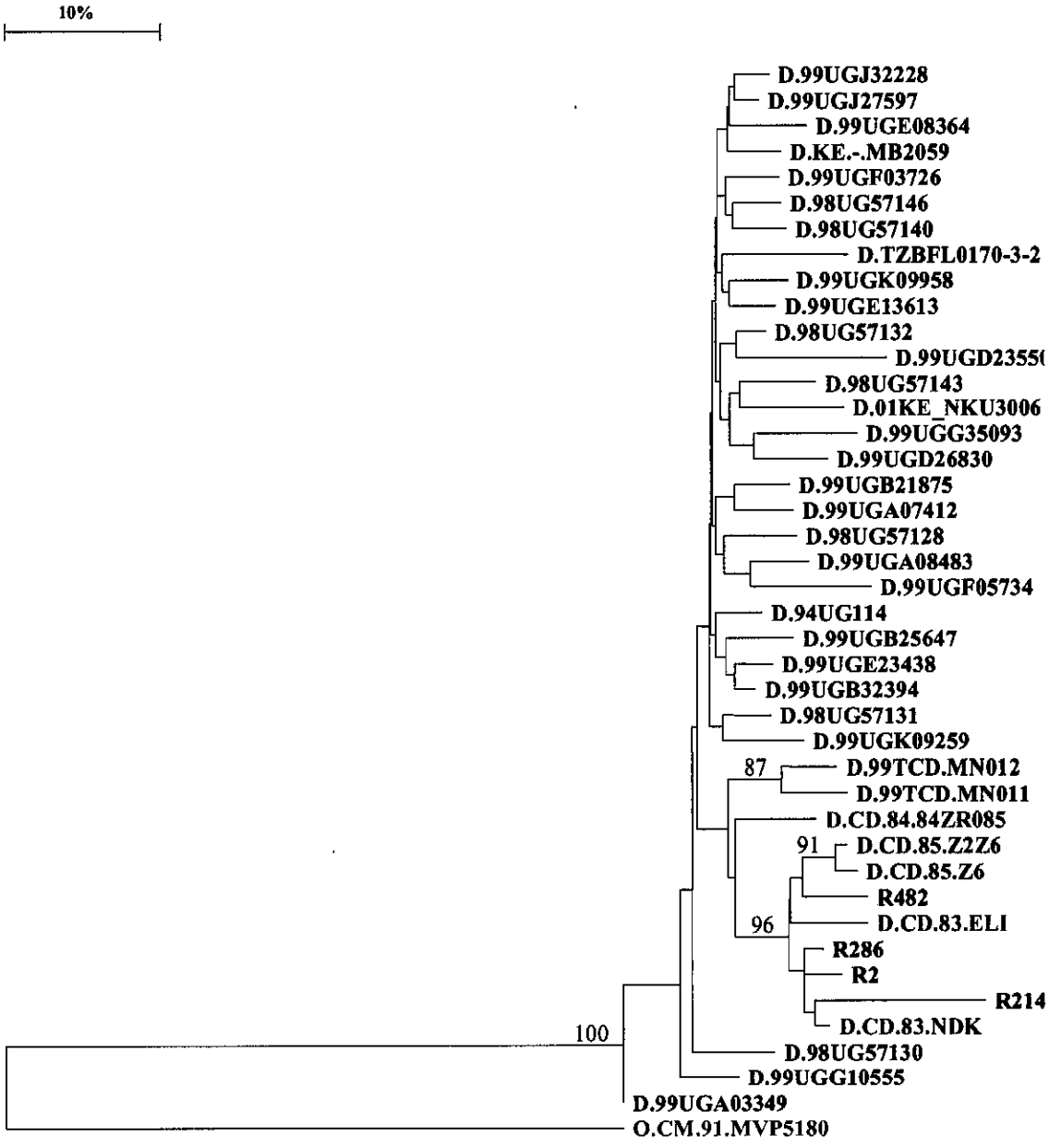


Figure 3.15. A neighbour joining phylogenetic tree comparing the complete *vpu* DNA sequences of the Tygerberg sequences indicated in red colour with the HIV-1 subtype D sequences. Bootstrap values greater than 70% are indicated. The horizontal scale indicates the percentage variation between sequences.

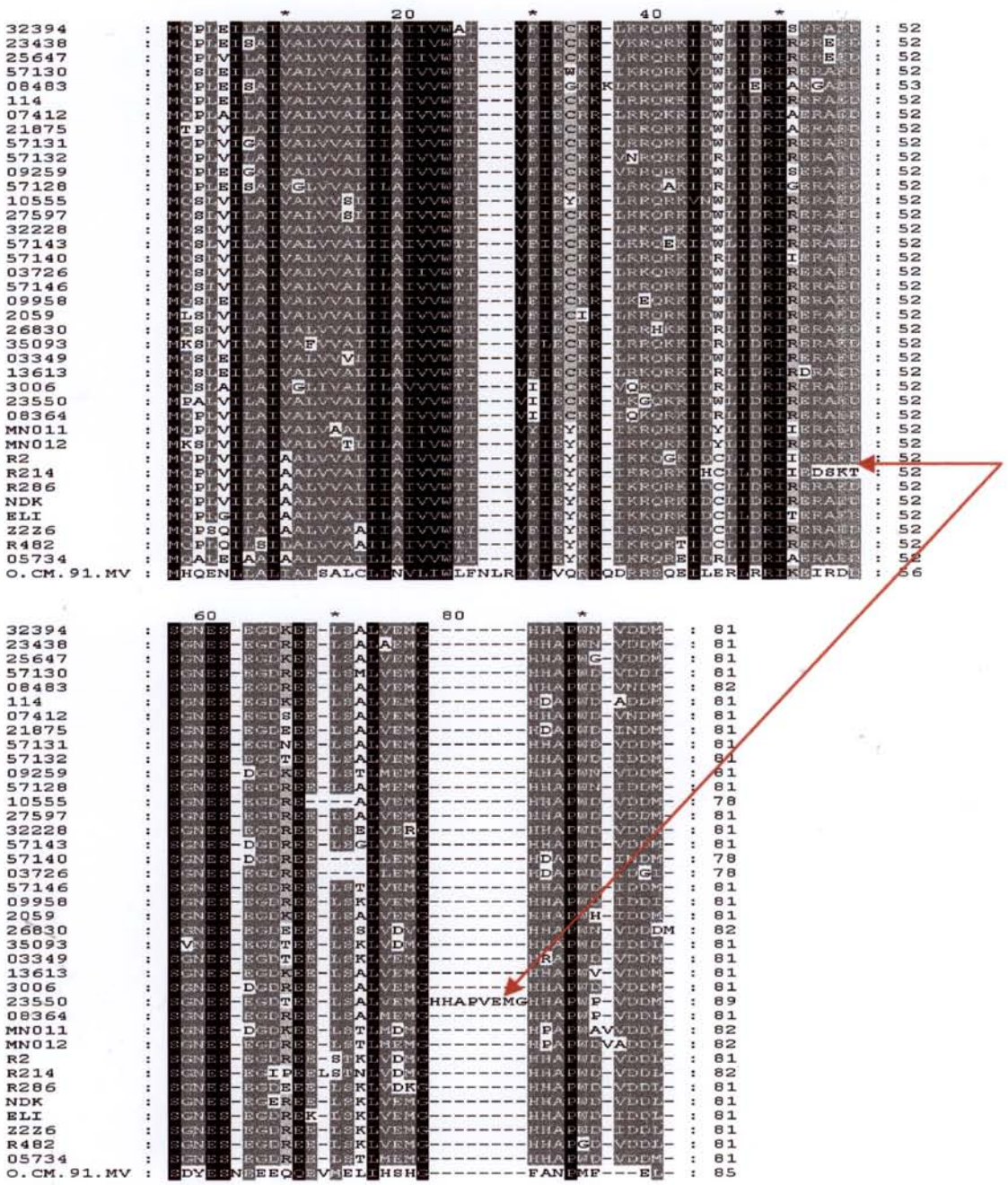


Figure 3.16. The Vpu amino acid alignment of the Tygerberg strains and the other subtype D strains. Indicated by the red arrows are the major differences in the sequences of strains R214 and UGD23550.

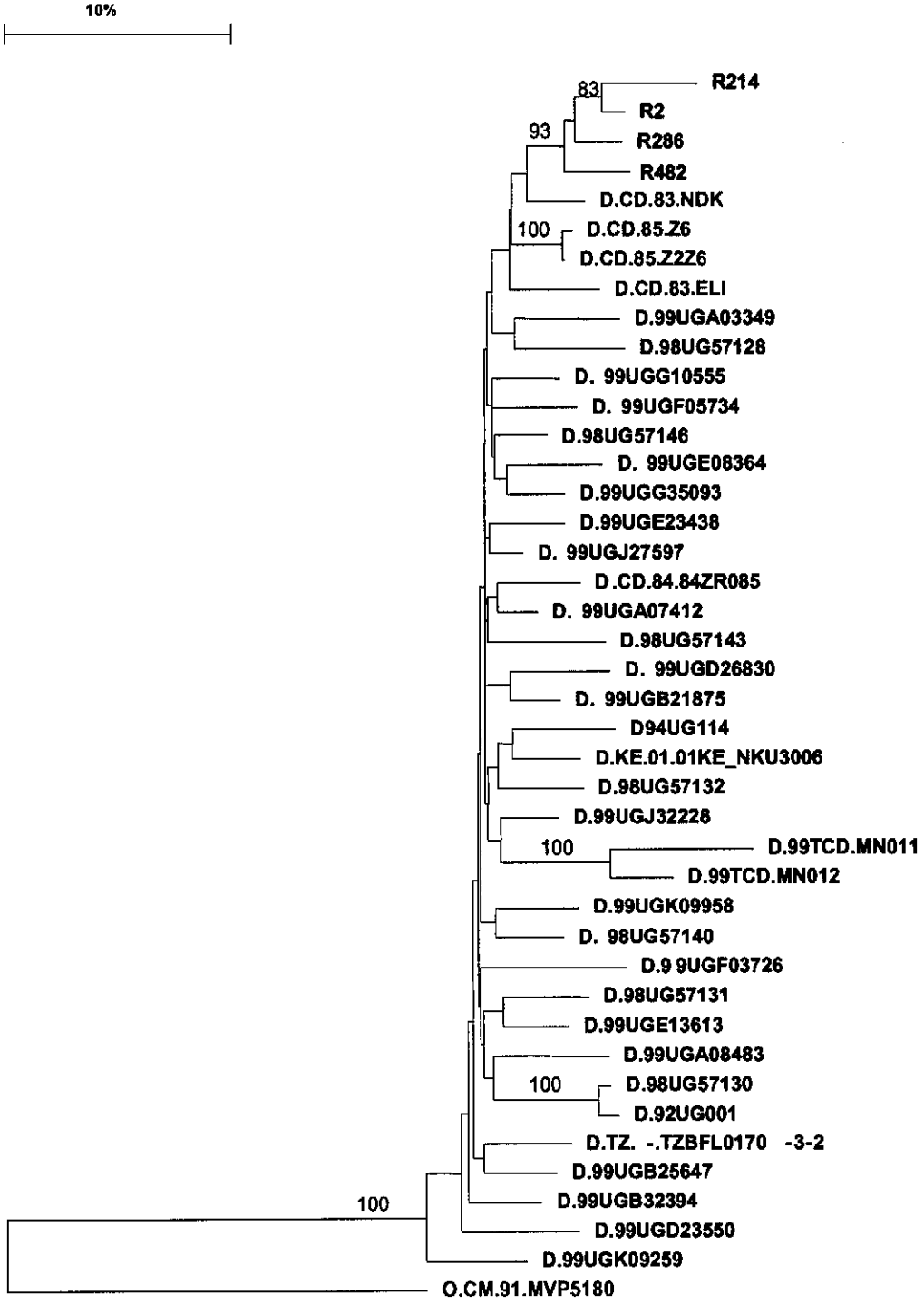


Figure 3.17. A neighbour joining phylogenetic tree comparing the complete *tat* DNA sequences of the Tygerberg sequences indicated in red colour with the HIV-1 subtype D sequences. Bootstrap values greater than 70% are indicated. The horizontal scale indicates the percentage variation between sequences.

Most subtype D viruses contain an in-frame stop codon in the second exon of *tat*, which removes 13 to 16 amino acids from the carboxy terminus. The Tygerberg strains do not possess a stop codon, but a glutamine, which makes the Tygerberg sequences to have the complete *tat* gene (**Figure 3.18**).

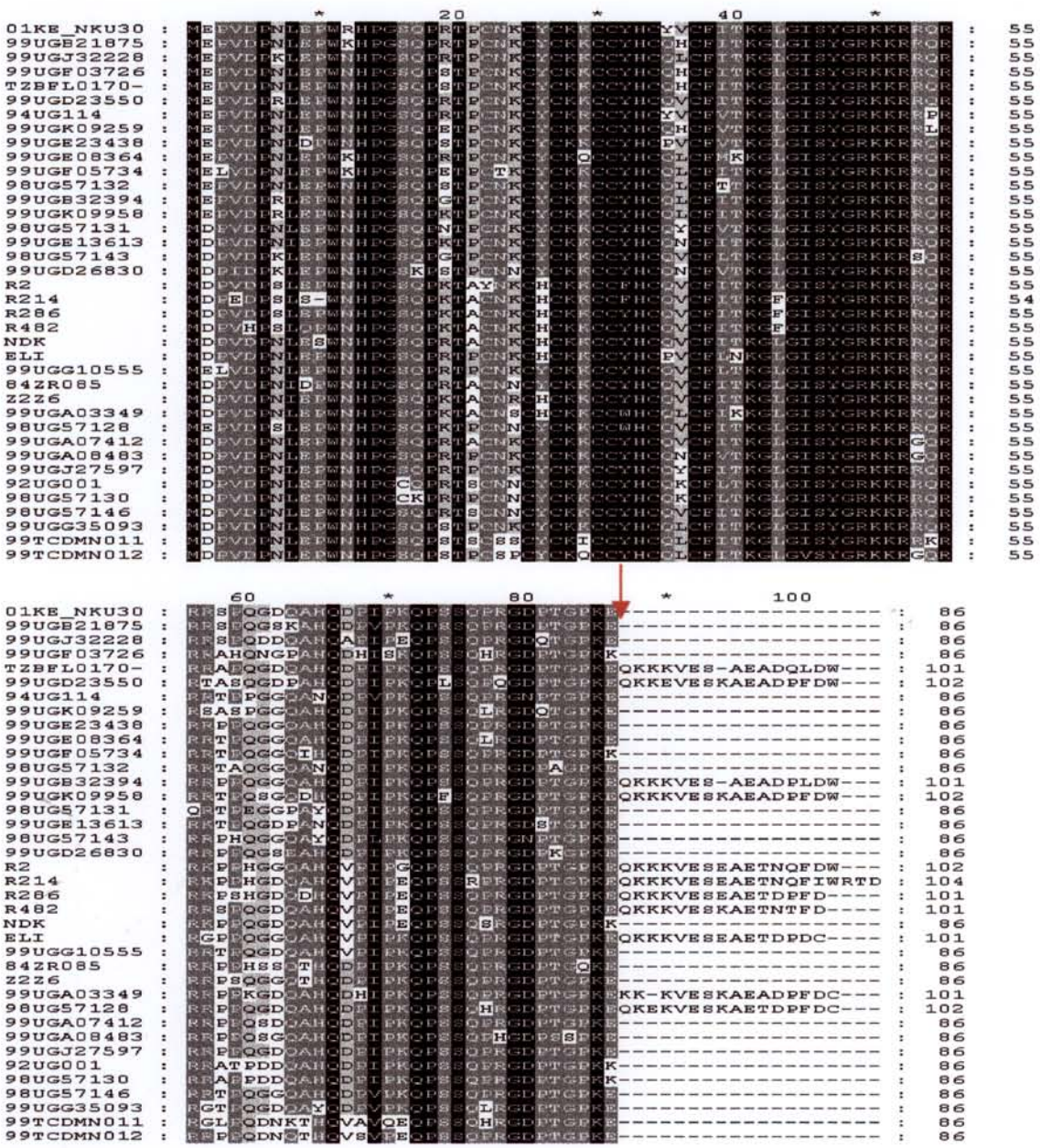


Figure 3.18 The complete Tat protein alignment of the HIV-1 subtype D amino acid sequences. Indicated by the red arrow is the position in exon 2 of Tat where most strains have a stop codon. Also visible is the fact that the tat protein is 101 or more amino acids if the stop codon is not present.

H) *Rev* and *Nef* genes

Rev is the second necessary regulatory factor for HIV expression and is a 19 kD phosphoprotein, localized primarily in the nucleolus. Rev acts by binding to the Rev Responsive Element (RRE) and promoting the nuclear export, stabilization and utilisation of the viral mRNAs containing RRE. Rev is considered the most functionally conserved regulatory protein of the lentiviruses (HIV Sequence compendium, 2002). The *rev* phylogenetic tree shows a cluster of the Tygerberg strains, with a bootstrap value of 84%. The *rev* phylogenetic tree is depicted in **figure 3.19**.

Nef is a multifunctional 27 kD myristylated protein produced by an ORF located at the 3' end of the primate lentiviruses. Nef is predominantly cytoplasmic and associated with the plasma membrane via the myristyl residue linked to the conserved second amino acid (glycine). Again, in the *nef* phylogenetic tree, the Tygerberg strains forms a separate cluster with bootstrap values of 95%. Closely related to the Tygerberg strains are the subtype D reference strains (**figure 3.20**).

3.5.3 N-linked glycosylation of the Tygerberg amino acid sequences

The glycosylation patterns of the Tygerberg sequences are depicted in **figure 3.21**. The number of glycosylation sites over the *env* gene is indicated in section of 100 base pairs. Sequence R2 has 29 glycosylation sites over the *env* gene. Sequence R214 has 28 sites, sequence R286 has 25 sites and sequence R482 has 27 sites over the *env* gene. The subtype D sequences are generally highly glycosylated as can be seen in **Appendix E**. The first 500-600 bases of the *env* gene are the most glycosylated areas for the subtype D sequences. The glycosylation patterns of consensus sequences of subtypes A – K are also indicated in Appendix E. When compared to the other subtypes, subtype D consensus does not have the most glycosylation sites, even though it is highly glycosylated. The only subtype to have more sites is subtype B.

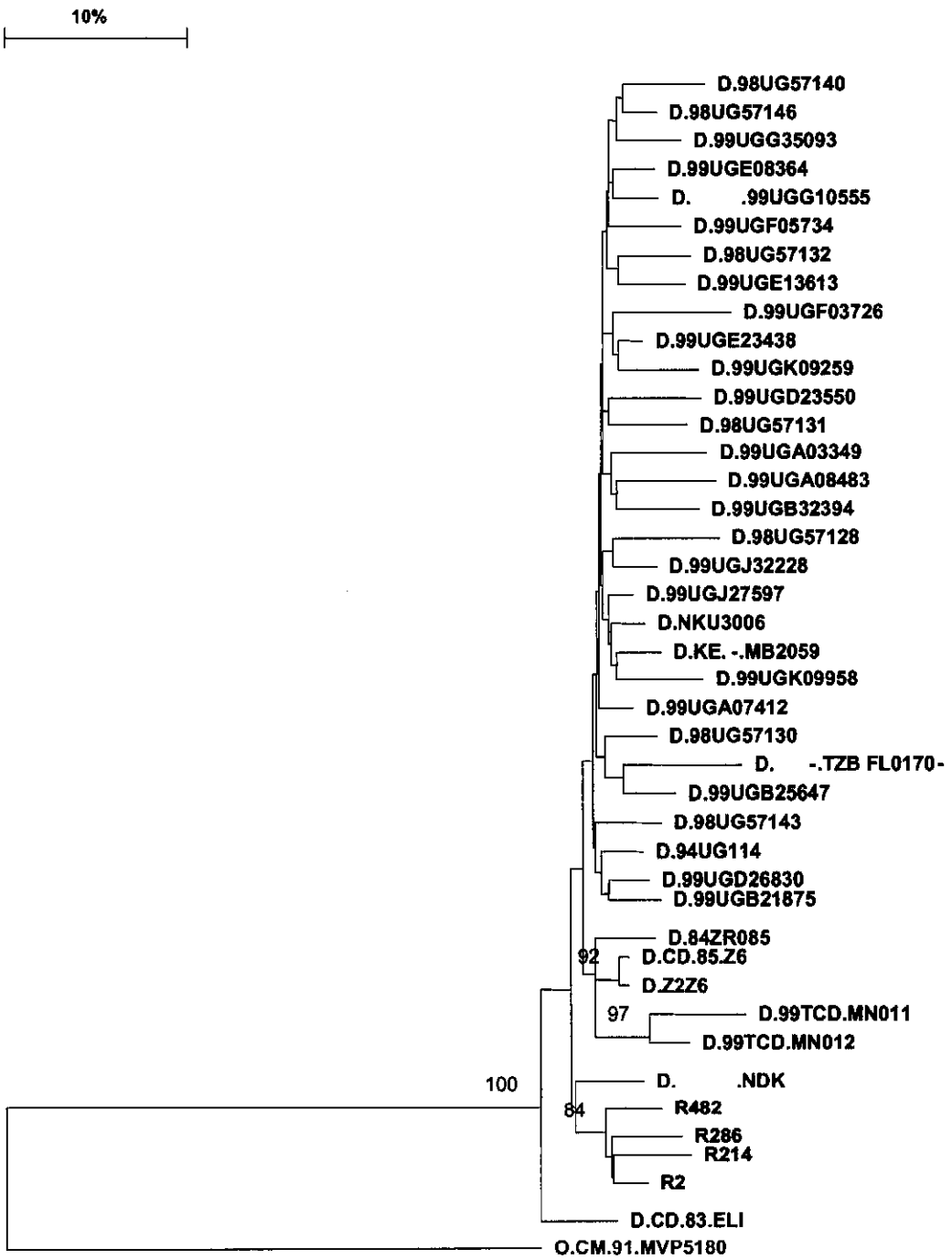


Figure 3.19. A neighbour joining phylogenetic tree comparing the complete *rev* DNA sequences of the Tygerberg sequences indicated in the red colour with the HIV-1 subtype D sequences. Bootstrap values greater than 70% are indicated. The horizontal scale indicates the percentage variation between sequences.

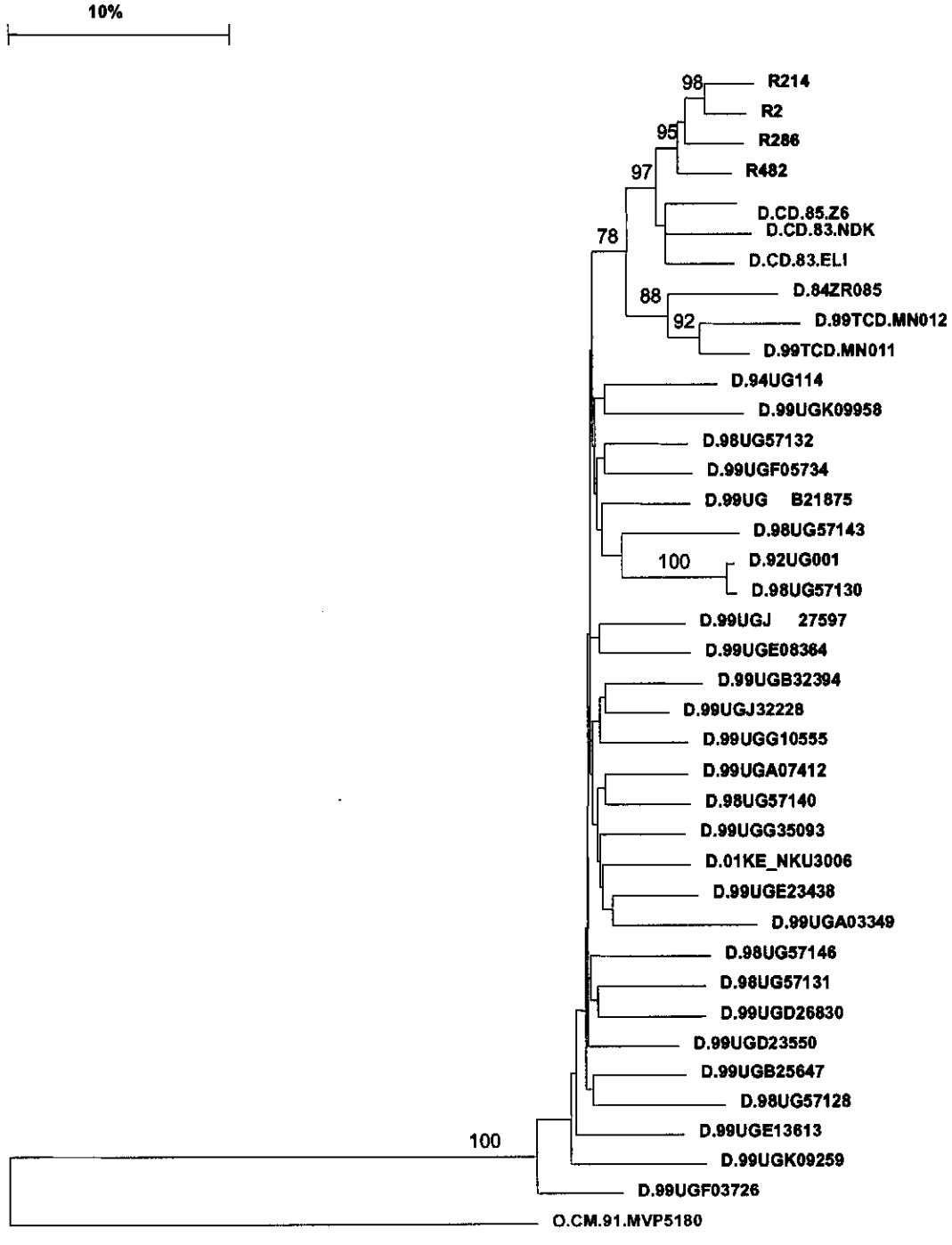
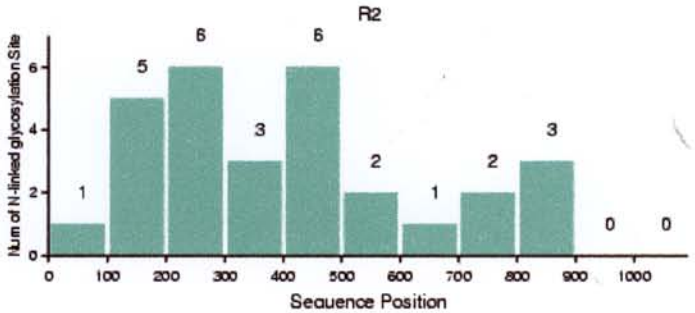
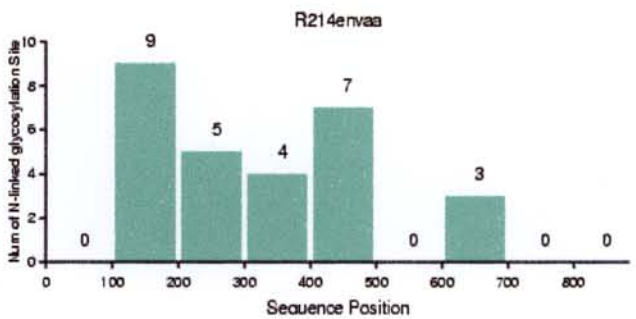


Figure 3.20. A neighbour joining phylogenetic tree comparing the complete *nef* DNA sequences of the Tygerberg sequences indicated with the red colour with the HIV-1 subtype D sequences. Bootstrap values greater than 70% are indicated. The horizontal scale indicates the percentage variation between sequences.

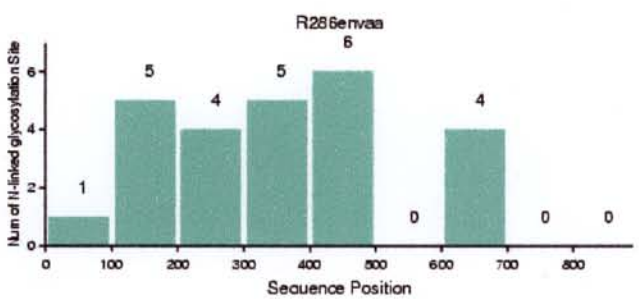
A) R2



B) R214



C) R286



D) R482

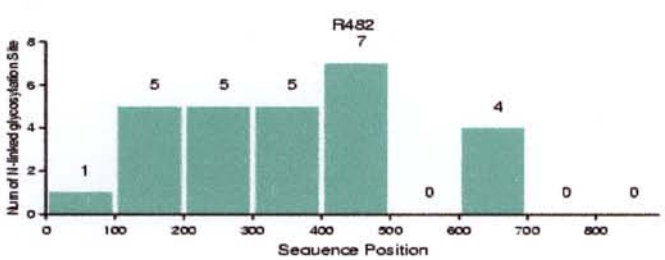


Figure 3.21. Graphs depicting the number of N-linked glycosylation sites in the complete *env* DNA sequences of the Tygerberg strains. On the y-axis is the number of glycosylation sites and on the x-axis the sequence position in the *env* gene. A) R2, B) R214, C) R286 and D) R482.

Chapter 4

DISCUSSION AND CONCLUSION

Page

Chapter 4

DISCUSSION AND CONCLUSION

In the present study, four HIV-1 subtype D strains obtained at the Tygerberg Academic Hospital between 1984 and 1986 were sequenced and characterised. The four full-length strains characterised indicated no intersubtype recombination. Evolutionary phylogenetic trees and sequence identity matrices proved useful to determine the similarity of the Tygerberg full-length sequences and the reference sequences from the Los Alamos database and highlighted the differences.

4.1 The HIV-1 epidemic in South Africa

The first reported cases of HIV-1 infection in South Africa occurred in 1982 (Ras *et al,* 1983). In South Africa, unlike the rest of sub-Saharan Africa, HIV-1 was initially spread by homosexual contact (Kustner, 1994). HIV-1 subtypes B and D were sequenced from these patients between 1984 and 1990 (Becker *et al*, 1985; Becker *et al*, 1995; Engelbrecht *et al*, 1995). Subtype D viruses were reported in five out of 11 South African male homosexual patients diagnosed in the early to mid 1980s (Engelbrecht *et al*, 1995). The first epidemic in the early eighties was almost exclusively confined to HIV-1 infections in men (Kustner, 1994).

By 1989, the second HIV-1 epidemic in South Africa was recognised primarily in the black population (Williamson *et al*, 1995). Infections of the second epidemic were predominantly heterosexual in origin and involved mainly HIV-1 subtype C (van Harmelen *et al,* 1997). This epidemic had attained a rapid global distribution and, whereas the transmission of the initial subtypes B and D seemed to be on the decline, HIV-1 subtype C spread at alarming rates (McCutchan *et al*, 1996). On the basis of the age of the epidemic and the genetic distance between *gag* sequences, an early study suggested multiple introductions of subtype C strains into South Africa (van Harmelen *et al*, 1999). HIV-1 subtype C has established itself as the most prevalent subtype in Africa (Esparza and Bhamarapravati, 2000; McCutchan, 2000).

4.2 HIV-1 subtype D and complete genomes in South Africa

Apart from the five subtype D viruses described by Engelbrecht *et al* (1995), not a lot of focus has been placed on the subtype D viruses from this country. In 1997, van Harmelen *et al* (1997), found subtype D in one male homosexual patient and one heterosexual patient through analysis of the partial *gag* sequences and heteroduplex mobility assays (HMA) of the V3-V5 region. Bredell *et al* (2002) identified subtype D viruses as recombinant viruses in *gag* and *env*. These recombinants included a C/D and D/U strain and one subtype D not amplifiable in the *env* region (D/-).

In earlier classifications, HIV-1 sequences were grouped in different subtypes based on partial *gag* and *env* sequences, representing clusters branching from a common node in phylogenetic trees, which suggest common ancestry. Subsequently, the characterisation by full-length genome sequencing has led to the identification of new HIV-1 clades and the realisation of the existence of CRFs (Thomson *et al*, 2002). Full-length HIV-1 genomes have been used to study the genomic organization of the virus, the structure and functions of viral genes and pathogenesis. The recognition of dual infections (Zhu *et al*, 1995) and the occurrence of recombination between subtypes (Robertson *et al*, 1995) suggest that cloning an intact plasma virus genome as a single full-length and determining the sequence thereof is desirable. Full-length genomes have now been obtained for 9 subtypes and about 15 recombinant forms of HIV-1. The HIV sequence database contains nineteen full-length sequences from South Africa mostly of subtype C, the strain responsible for the current epidemic (**Table 4.1**).

Alizon *et al* (1986) described the first full-length subtype D sequence, Eli, which was recovered in 1983 from a 24-year-old woman with AIDS. Today, 42 full-length subtype D sequences have been described, 27 of which are from Uganda (Harris *et al*, 2002). The other full-length sequences are from: Kenya (Dowling *et al*, 2002; Neilson *et al*, 1999), Chad (Vidal *et al*, 2003), Democratic republic of the Congo (Gao *et al*, 1998; Spire *et al*, 1989; Srinivasan *et al*, 1987; Alizon *et al*, 1986) and Tanzania (Koulinska *et al*, 2003). These 42 sequences are also pure subtype D sequences.

Table 4.1 Full-length HIV-1sequences from South Africa (http://www.lanl.gov)

Sequence name	Accession number	Subtype	Year	Reference
97ZA012	AF286227	C	1997	Rodenburg 2001
CM4	AF411964	A1CDGKU	1999	Papathanasopoulus 2002
DU178	AF411965	A2C	1998	Papathanasopoulus 2002
SW7	AF411966	C	1999	Papathanasopoulus 2002
99ZACM9	AF411967	C	1999	Papathanasopoulus 2002
TV001	AY162223	C	1998	zur Megede 2002
TV002	AX455929	C	1998	zur Megede 2002
DU151	AY043173	C	1999	van Harmelen 2001
DU179	AY043174	C	1999	van Harmelen 2001
DU422	AY043175	C	1999	van Harmelen 2001
CTSC2	AY043176	C	1999	van Harmelen 2001
97ZA003	AY118165	C	1997	Unpublished
97ZA009	AY118166	C	1997	Unpublished
98ZA445	AY158533	C	1998	Hunt 2003
98ZA502	AY158534	C	1998	Hunt 2003
98ZA528	AY158535	C	1998	Hunt 2003
TV012	AY162225	C	1998	zur Megede 2002
99ZATM10	AY228556	C	1999	Papathanasopoulos 2003
01ZATM45	AY228557	C	2001	Papathanasopoulos 2003

In Sudan, the largest country in Africa little is known about the HIV epidemic. In the capital Khartoum, the prevalence among antenatal clinics was between 1 and 5% in the 1996 to 1998 time frame, with no data about HIV-1 subtypes. This would be interesting because Hierholzer *et al* (2002) found that 50% of the samples from Sudan were subtype D in partial analysis of the *pol* and *env* genes. Globally subtype D consists of two different lineages, one circulating in East and another in West Central Africa (Vidal *et al*, 2003; Hierholzer *et al*, 2002). Genetically they are distinguishable as two significant subclusters within subtype D (Hierholzer *et al*, 2002), illustrating different founder effects of subtype D in East and West Central Africa (Vidal *et al*, 2003).

Subtype D sequences has also been described in CRFs. Eight viruses of the CRF05_DF type have been described by Laukkanen *et al* (2000) and Casado *et al* (2003). These viruses are restricted to Europe, even though virus X492, from a 49-year-old woman is suspected to be infected by a sailor who had travelled to Africa (Casado *et al*, 2003). Another suspected case is of virus, R890820, which was sequenced from a Dutch man with a female partner from the DRC (Bikandou *et al*, 2000; Laukkanen *et al*, 2000). The DRC has been reported to have a relatively high prevalence of subtype D compared to many other African countries (Vidal *et al*, 2000). The genetic distances in the phylogenetic trees drawn by Laukkanen *et al* (2000) suggest that the recombination event leading to the putative D/F CRF occurred relatively long ago, close to the divergence of the F1 and F2 subclusters. The fact that these recombinants are linked to the DRC suggests that the original recombination event took place in central Africa.

The second form of subtype D recombinants in the database, CRF10_CD has been mostly described by the group of Essex (Koulinska *et al*, 2001; Renjifo *et al*, 1999). Eleven CD recombinants have been described, 10 of which are from Tanzania. Bums *et al* (2002) described the other recombinant, from Kenya,, when they looked at sequence variability of the integrase protein from a diverse collection of HIV-1 sequences that represent several subtypes. In 1982 to 1984, subtypes A and D were present in Malawi. In 1987 to 1989 a survey found only eight more individuals who had been infected with subtypes A and D and by that time there were also recombinant viruses of the AD and DC

(*gag/env*) type (McCormack *et al*, 2002). Although subtype D was present early in the epidemic in Malawi, it did not spread in a comparable fashion as did subtype C. In Tanzania, Koulinska *et al* (2002) found that five out of six full-length recombinants were mostly subtype D in the *gag, pol, tat, rev* and the intracytoplasmic domain of gp41. The most common recombination patterns observed were D (*gag*) – A (*env*) and D (*gag*) – D/C/D (*env*).

4.3 Unique features of the HIV-1 subtype D genome

4.3.1 Tat exon 2

Tat is a small protein of 80 to 101 amino acids, which is encoded from two separate exons. Studies have shown that the Tat protein separately is largely unfolded (Metzger *et al*, 1997; Bayer *et al*, 1995). The Tat sequence has been subdivided into several distinct sequences on the basis of its amino acid composition: a N-terminal activation region (aa 1 - 19), a cysteine-rich role domain (aa 20 – 31), a core region (aa 32 – 47), a basic region (aa 48 – 57) and a glutamine-rich region (aa 60 – 76) (Metzger *et al*, 1997; Klostermeier *et al* 1997), each of these regions being essential for Tat function. In comparison with exon 1, the second coding exon of Tat has been less studied, since it is assumed that the second exon of Tat does not greatly alter measurements of Tat activity. Findings from HIV-2 Tat, however, are quite clear in demonstrating that this exon contributes towards optimal trans-activation (Tong-Starksen *et al* 1993). In other assays, the second exon of HIV-1 Tat has been shown to be important for trans-activation (Jeang *et al* 1993) and virus replication (Neuveut and Jeang, 1996). Two short motifs in the second exon of HIV-1 Tat could have been identified. The first is an RGD sequence (Fig. 3.18; pos 78-80) that is used as a cell adhesion signal for binding to cellular integrins (Brake *et al*, 1990). This RGD motif, however, is not found in HIV-2 or SIV Tat proteins. The second exon also had an E (Q/S) KKKVE motif, which is conserved in most HIV-1 Tat proteins. The functional significance of this motif has not been examined in detail. Most HIV-1 subtype D viruses contain an in-frame stop codon in the second exon of Tat, which removes 13 to 16 amino acids from the carboxyl terminus of the Tat protein (Fig. 3.17; Gao *et al*, 1998; Spira *et al*, 2003). The Tygerberg sequences (R2, R214, R256 and R482) all contain a Q

(glutamine) instead of the stop codon and have the complete *tat* gene. Although this change is unlike to alter the function of the respective gene products in a major way, it is possible that they could influence their mechanism of action in a subtle (but nevertheless biological important) manner (Gao *et al*, 1998).

4.3.2 HIV-1 V3 Loop

The third hyper variable (V3) domain of HIV-1 gp120 is a disulfide-linked loop of approximately 40 amino acids with a high degree of sequence diversity among different viral sequences (Stanfield *et al*, 1999). The V3 loop of all four the Tygerberg sequences has 35 amino acids (**Fig**. 3.11). The V3 loop is one of the major immunogenic sites on the virus and is sometimes called the principal-neutralizing determinant (PND) (Jahaverian *et al*, 1989). The accessibility or exposure of the V3 loop on gp120 appears to vary depending on the viral sequence type and increases significantly when the virus interacts with CD4 through a conformational change that is triggered in gp120 (Sattentau and Moore, 1991). The variation in the V3 loop has been the focus of extensive research efforts because sequence changes in the V3 can alter viral cell tropism, antibody neutralization, syncytium formation and chemokine receptor usage (Hoffman *et al*, 2002; Janse van Rensburg *et al*, 2002; Treurnicht *et al*, 2002; Fouchier *et al*, 1995; Zhong *et al*, 2003; Milich *et al*, 1993). The turn at the apex of the loop is characterised by a range of tetrameric sequences including: GPGQ, GPGR, GLGQ and GPGL. All four of the Tygerberg strains share the GQGQ motif with positive amino acids at positions 11 and 25. The amino acid at position 25 in HIV sequences is usually different for macrophage tropic and T-cell-line tropic viruses (Stanfield *et al*, 1999). Most of the macrophage tropic viruses have either an acidic amino acid or alanine at position 25, in contrast to the T cell-line tropic viruses, which usually have a non acidic amino acid at this position (Milich *et al*, 1993). Positively charged amino acids in these positions in the V3 loop are therefore correlated with syncytium-inducing (SI) viruses and negatively charged amino acids with the non syncytium-inducing viruses (NSI). Compared to other group M subtypes,

subtype D strains demonstrate a highly variable pattern of V3 loop amino acids (**Fig**. 3.11; Spira *et al*, 2003).

4.3.3 HIV-1 glycosylation

In the course of co-translational transfer into the lumen of the rough endoplasmic reticulum (RER), retroviral *env* gene products are modified by the addition of oligosaccharide side chains through N-linked glycosylation of asparagine residues in the nascent polypeptide (Hunter and Swanstrom, 1990). The number and distribution of N-linked glycosylation sites varies widely between different retroviruses, with the HIV-1 gp120 being one of the most extensively glycosylated proteins known (Lee *et al*, 1992; Myers and Lenroot, 1992). HIV-1 has as many as 30 of the canonical Asn-X-Ser/Thr oligosaccharide addition sites, with the majority (25) located in gp120. The glycans attached to these sites account for approximately 50% of the protein's total mass (Ogert *et al*, 2001). Numerous studies using glycosylation and glycosidase inhibitors have revealed the importance of the carbohydrate moieties in determining the conformation of the HIV-1 envelope glycoprotein, a property that undoubtedly affects its processing, intracellular transport and ability to interact with CD4 (Pal *et al*, 1989; Montefiori *et al*, 1988). The N-glycosylation site at N306 protects HIV-1 from neutralizing antibodies and the elimination of this particular glycan may influence HIV-1 infectivity (Schonning *et al*, 1996). In the present study, we determined the glycosylation of the Tygerberg strains and compared it with the consensus sequences of subtypes A-K. The glycosylation patterns of the Tygerberg sequences compared well: R2 has 29 sites, R214 has 28 sites, R286 has 25 sites and R482 has 27. The Tygerberg sequences had generally less glycosylation sites than the consensus subtype D, which had 32 sites. Most of the glycosilation sites of the other subtypes vary between 30 and 32, except for the subtype B consensus that has 33 and the subtype C consensus that has 29 sites.

4.3.4 R214 vpu gene

The Vpu protein of HIV-1 is a small integral membrane protein of 81 residues that is synthesized and localised in the RER of infected cells (Strebel *et al*,

1988). Vpu is unique to HIV-1. The vpu protein has one transmembrane hydrophobic helix and two amphipathic helices in its cytoplasmic domain (Ma *et al*, 2002). Residues 1-27 constitute the N-terminal hydrophobic membrane anchor, followed by 54 residues that protrude into the cytoplasm. Experiments with canine microsomal membranes have revealed that the 27 amino acid region of Vpu is responsible for membrane association (Strebel *et al*, 1989). The Vpu cytoplasmic domain contains a high proportion of charged residues followed by a series of acidic residues in the C-terminal part of the protein that confer an overall negative electrostatic charge to the molecule (**Fig. 4.1**). A highly conserved region spanning residues 47-58 contains a pair of serine residues that are constitutively phosphorylated by casein kinase II (Schubert and Strebel, 1994). *Vpu* and *env* are expressed from the same bicistronic mRNA in a Rev-dependant manner (Schwartz *et al*, 1990) and it is possible that this unusual utilisation of viral transcripts might reflect a requirement for the coordinate action of the two viral gene products (Strebel, 1996). Several HIV-1 sequences were found to carry point mutations in the Vpu translation initiation codon but have otherwise intact *vpu* genes (Strebel, 1996). The Tygerberg sequence, R214, has an intact vpu gene, but differs considerably (up to 30% from strain 99UGD23550) from the other subtype D sequences (**Appendix D7**). The serine residues of strain R214 is in place at positions 52 and 56, indicating that the phosphorylation by the ubiquitous casein kinase II can continue, but the RAED sequence prior to the first serine had changed to DSKT. The change in amino acid sequence might play a role in the assembly of the Vpu protein of R214, yielding it more efficient for particle release from the plasma membrane of infected cells (Bour and Strebel, 2003; Paul *et al*, 1998).

CONCLUSION

This work represents the first full-length characterisation of HIV-1 subtype D from South Africa. The study points out that the Tygerberg sequences (R2, R214, R286 and R482) are more closely related to the subtype D strains from West Central Africa than to the strains from East Africa, indicating two different founder effects for the viruses from east (more recent subtype D) and west (older subtype D) Africa. Given the potential impact of nonsubtype C viruses on ongoing vaccine and natural history studies, the extent of HIV-1 diversity in South African populations should be closely monitored. It would therefore be necessary to characterise in full, the subtype B strains sequenced at the beginning of the epidemic in South Africa in our attempt to reconstruct the epidemiology and evolutionary history of HIV in South Africa and the rest of the world. This will allow us to track the diversity and early evolution of the HIV-1 epidemic in South Africa so that: 1) The ancestoral subtype B/D strains can be used for vaccine design, 2) various issues regarding public health policy and planning can be addressed and 3) a more accurate estimation of the origin of the epidemic can be made.

REFERENCES

Alizon M, Wain-Hobson S, Montagnier L and Sonigo P. 1986. Genetic variability of the AIDS virus: nucleotide sequence analysis of two sequences from African patients. Cell 46 (July 4): 63-74

Ammamm AJ, Abrams DI, Conant M, Chudwin D, Cowan M, Volberding P, Lewis B and Casavant C. 1983. Acquired immune dysfunction in homosexual men: Immunologic profiles. Clin Immunol Immunopathol 27: 315-320

Ayala FJ and Fitch WM. 1997. Genetics and the origin of species: From Darwin to molecular biology 60 years after Dobzhansky. PNAS 94: 7691-7806

Bailes E, Gao F, Bibollet-Ruche F, Courgnaud V, Peeters M, Marx PA, Hahn BH and Sharp PM. 2003. Hybrid origin of SIV in chimpanzees. Science Jun 13; 300(5626): 1713

Balfe P, Simmonds P, Ludlam CA, Bishop JO and Brown AJ. 1990. Concurrent evolution of human immunodeficiency virus type 1 in patients infected from the same source: rate of sequence change and low frequency of inactivating mutations. J.Virol 64:6221-6233

Baraz L and Kotler M. 2004. The Vif protein of human immunodeficiency virus type 1 (HIV-1): enigmas and solutions. Curr Med Chem 11(2): 221-31

Barre-Sinoussi F, Chermann JC, Rey F, Nugeyve MT, Chamaret S, Gruist J, Dauguet C, Axler-Blin C, Vezinet-Brun F, Rouzloux C, Rozanbaum W and Montagnier L. 1983. Isolation of a T- lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). Science 220:868-871

Bayer P, Kraft M, Ejchart A, Westendorp M, Frank R and Rosch P. 1995. Structural studies of HIV-1 Tat protein. J Mol Biol 247(4):529-35

Becker ML, de Jager G and Becker WB. 1995. Analysis of partial *gag* and *env* gene sequences of HIV type 1 strains from Southern Africa. AIDS Res Hum Retroviruses 11(10): 1265-1267

Becker MLB, Spracklen FHN and Becker WB. 1985. Isolation of a lymphadenopathy associated virus from a patient with acquired immune deficiency syndrome. S Afr Med J 68: 144-147

Bikandou B, Takehisa J, Mboudjeka I, Ido E, Kuwata T, Miyazaki Y, Moriyama H, Harada Y, Taniguchi Y, Ichimura H, Ikeda M, Ndolo PJ, Nzoukoudi MY, M'Vouenze R, M'Pandi M, Parra HJ, M'Pele P, Hayami M. 2000. Genetic subtypes of HIV type 1 in Republic of Congo. AIDS Res Hum Retroviruses 16 (7):613-9

Blackard JT, Renjifo B, Fawzi W, Hertmerk E, Msamanga G, Mwakagile D, Hunter D, Spiegelman D, Sharghi N, Kagoma C and Essex M. 2001. HIV-1 LTR subtype and perinatal transmission. Virology 287: 261-265

Bour S and Strebel K. 2003. The HIV-1 Vpu protein: a multifunctional enhancer of viral particle release. Microbes Infect 5 (11): 1029-39

Brake DA, Debouck C and Biesecker G. 1990. Identification of an Arg-Gly-Asp (RGD) cell adhesion site in human immunodeficiency virus type 1 transactivation protein, tat. J Cell Biol 111(3): 1275-81

Bredell H, Hunt G, Casteling A, Cilliers T, Rademeyer C, Coetzer M, Miller S, Johnson D, Tiemessen CT, Martin DJ, Williamson C and Morris L. 2002. HIV-1 Subtype A, D, G, AG and unclassified sequences identified in South Africa. AIDS Res Hum Retroviruses 18(9): 681-3

Briggs JAG, Wilk T, Welker R, Krausslich H and Fuller SD. 2003. Structural organization of authentic, mature HIV-1 virions and cores. The EMBO Journal 22 (7): 1707-1715

Burns CC, Gleason LM, Mozaffarian A, Giachetti C, Carr JK and Overbaugh J. 2002. Sequence variability of the integrase protein from a diverse collection of HIV type 1 sequences representing several subtypes. AIDS Res. Hum. Retroviruses 18 (14): 1031-41

Casado G, Thomson MM, Delgado E, Sierra M, Vazquez-De Parga E, Perez-Alvarez L, Ocampo A and Najera R. 2003. Near full-length genome characterisation of an HIV type 1 CRF05_DF virus from Spain. AIDS Res Hum Retroviruses Aug 19 (8):719-25

Caumont A, Lan NT, Uyen NT, Hung PV, Schvoerer E, Urriza MS, Roques P, Schrive MH, Lien TT, Lafon ME, Dormont D, Barre-Sinoussi F and Fleury HJ. 2001. Sequence analysis of env C2/V3, gag p17/p24, and pol protease regions of 25 HIV type 1 sequences from Ho Chi Minh City, Vietnam. AIDS Res Hum Retroviruses 17(13):1285-91

Centers for Disease Control. 1982. Centers for Disease Control Task Force on Kaposi's sarcoma and opportunistic infections. N. Eng J Med 306: 248- 252

Cichutek K and Norley S. 1993. Lack of immune suppression in SIV infected natural hosts. AIDS 7 (suppl1): S25-S35

Clavel F, Guetard D, Brun-Vezinet F, Chamaret S, Rey M, Santos-Ferreira MO, Laurent AG, Dauget C, Katlama C, Rouzioux C, Klatzmann D, Champalimand JL and Montagnier L. 1986. Isolation of a new human retrovirus from West African patients with AIDS. Science 228: 343-346

Coffin J, Haase A, Levy JA, Montagnier L, Oroszlan S, Teich N, Temin H, Toyoshima K, Varmus H, Vogt P and Weiss R. 1986.What to call the AIDS virus? Nature 321 (1 May):10

Cornelissen M, van den Burg R, Zorgdrager F and Goudsmit J. 2000. Spread of distinct human immunodeficience virus type 1 AG recombinant lineages in Africa. Journal of Gen Virol 81: 515-523

Curran JW, Lawrence DN, Jaffe H, Kaplan JE, Zyla LD, Chamberland M, Weinstein R, Lui KJ, Schonberger LB and Spira TJ. 1984. Acquired immunodeficiency syndrome (AIDS) associated with transfusions. N. Eng J Med 310: 69-74

Daniels RS, Kang C, Patel D, Xiang Z, Douglas NW, Zheng NN, Cho HW and Lee JS. 2003. An HIV type 1 subtype B founder effect in Korea: gp160 signature patterns infer circulation of CTL-escape strains at the population level. AIDS Res Hum Retroviruses. Aug 19(8): 631-41

De Cock KM, Adjorlolo G, Ekpini E, Sibailly T, Konadio J, Maran M, Brattegaard K, Vetter KM, Doorby R and Gayle HD. 1993. Epidemiology and Transmission of HIV-2: Why there is no HIV-2 pandemic. JAMA 270 (17): 2083-2086

Dowling WE, Kim B, Mason CJ, Wasunna KM, Alam U, Elson L, Birx DL, Robb ML, McCutchan FE and Carr JK. 2002. Forty-one near full-length HIV-1 sequences from Kenya reveal an epidemic of subtype A and A-containing recombinants. AIDS 16 (13): 1809-1820

Dragic T, Litwin V, Allaway GP, Martin SR, Huang Y, Nagashima KA, Cayanan C, Maddon PJ, Koup RA, Moore JP and Paxton WA. 1996. HIV-1 entry into $CD4^+$ cells is mediated by chemokine receptor CC-CKR-5. Nature 381: 583-590

Efron B. 1996. Bootstrap confidence levels for phylogenetic trees. PNAS 93(14):7085-7090

Egboga A, Corrah T and Todd A. 1992. Immunological findings in African patients with pulmonary tuberculosis and HIV-2 infection. AIDS 6: 1045-1046

Engelbrecht S. 1992. Karakterisering van sekere aspekte van menslike immuungebrek virus tipe 1 (HIV-1) sequence. PhD Thesis, University of Stellenbosch

Engelbrecht S, Laten JD, Smith T-L and van Rensburg EJ. 1995. Identification of *env* subtypes in fourteen HIV type 1 sequences from South Africa. AIDS Res Hum Retroviruses 11(10): 1269-127

Essex M and Mboup S. 2002. Regional variation in the Africa epidemics. P631-640. In: AIDS in Africa (2^{nd} edition). Eds: Essex M, Mboup S, Kanki PJ, Marlink RG and Tlou SD. Kluwer Academic/ Plenum Publishers, New York (NY), USA

Esparza J and Bhamarapravati N. 2000. Accelerating the development and future of HIV-1 vaccines: why, when, where and how? Lancet 355: 2061-2066

Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39: 783-791

Fenyo EM, Schuitemaker H, Asjo B, Mckeating J, Sattentau Q, and the EC Concerted Action HIV Variability. 1997. The History of HIV-1 Biological Phenotypes: Past, present and future. pp III-I. In:Human retroviruses and AIDS 1997: a compilation and analysis of nucleic acid and amino acid sequences. Korber B, Hahn B, Foley B, Mellors JW, Leitner T, Myers G, McCuthchan F, Kuiken C. Eds. Theoretical biology and biophysics group. Los Alamos National laboratory, Los Alamos, NM

Fouchier RA, Brouwer M, Broersen SM and Schuitemaker H. 1995. Simple determination of human immunodeficiency virus type 1 syncytium-inducing V3 genotype by PCR. J Clin Microbiol 33 (4): 906-11

Freed EO. 1998. HIV-1 Gag proteins: Diverse functions in the virus life cycle. Virology 251: 1-15

Froland SS, Jenum P, Lindboe CF, Wefring KW, Linnestad PJ and Bohmer T. 1988. HIV-1 infection in Norwegian family before1970. Lancet Jun 11; 1 (8598): 1344-5.

Gallo RC, Salahuddin SZ, Popovic M, Shearer GM, Kaplan M, Haynes BF, Palker TJ, Redfield R, Oleske J and Safai B. 1984. Frequent detection and isolation of cytopathic retroviruses (HTLV-III) from patients with AIDS and at risk for AIDS. Science 224 (4648):500-3

Gao F, Robertson DL, Carruthers CD, Morrison SG, Jian B, Chen Y, Barré-Sinoussi F, Girard M, Srinivasan A, Abimiku AG, Shaw GM, Sharp PM and Hahn BH. 1998. A comprehensive panel of near-full-length clones and reference sequences for non-subtype B sequences of Human Immunodeficiency Virus type J. Virol 72(7): 5680-5698

Gao F, Bailes E, Robertson DL, Chen Y, Rodenburg CM, Michael SF, Cummins LB, Arthur LO, Peeters M, Shaw GM, Sharp PM and Hahn BH. 1999. Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. Nature 397: 4360-441

Gao F, Yue L, Sherri CH, Robertson DL, Graves AH, Saag MS, Shaw GM, Sharp PM and Hahn BH.1994. HIV-1 sequence subtype D in the United States. AIDS Res Hum Retroviruses 10(5): 625-627

Gao F, Yue L and White AT. 1992. Human infection by genetically diverse SIWith$_m$ related HIV- 2 in West Africa. Nature 358:495-499

Gashen B, Taylor J, Yusim K, Foley B, Gao F, Lang D, Novitsky V, Haynes B, Hahn BH, Bhattacharya T and Korber B. 2002. Diversity considerations in HIV-1 vaccine selection. Science (28 June) 296:2354-2360

Gatignol A and Jeang KT. 2000. Tat as a transcriptional activator and a potential therapeutic target for HIV-1. Adv Pharmacol 48:209-27. Review.

Gaynor RB. 1995. Regulation of HIV-1 gene expression by the transactivator protein tat. pp 51-77. In Transacting functions of human retroviruses. Eds Chen ISY, Koprowski H, Srinivasan A and Vogt PK. Current Topics in Microbiology and Immunology 193

Goettlinger H. 2001. The HIV-1 assembly machine. AIDS 15:S13-20

Goodenow M, Huet T, Saurin W, Kwok S, Sninsky J and Wain-Hobson S. 1989. HIV-1 sequences are rapidly evolving quasispecies: evidence for viral mixtures and preferred nucleotide substitutions. J Acquir Immune Defic Syndr. 2(4):344-52

Gordon M, de Oliveira T, Bishop K, Coovadia HM, Madurai L, Engelbrecht S, Janse van Rensburg E, Mosam A, Smith A and Cassol S. 2003. Molecular characteristics of Human immunodeficiency Viruses type 1 subtype C viruses from Kwazulu-Natal, South Africa: Implications for vaccine and antiretroviral control strategies. J. Virol 77(4): 2587-2599

Gottlieb MS, Schroff R, Schanker HM, Weisman JD, Fan PT, Wolf RA and Saxon A. 1981. Pneumocystis carinii pneumonia and mucosal cadidiasis in previously healthy homosexual men: Evidence of a new acquired cellular immune deficiency. N. Eng J. Med 305: 1425-1428

Graur D and Li WH. 1999. Fundamentals of molecular evolution. Sinauer Associates, Sunderland, MA, USA

Hahn BH. 1994. Viral genes and their products. In: Eds Broder S, Merigan TC and Bolognesi D. Textbook of AIDS medicine.pp21-35. Williams and Wilkins, Baltimore, Maryland, USA.1994

Hahn BH, Shaw GM, De Cock KM and Sharp PM. 2000. AIDS as a Zoonosis: Scientific and public health implications. Science 287 (January 28): 607-614

Hall TA. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl. Acids. Symp. Ser. 41:95-98

Harris ME, Serwadda D, Sewankambo N, Kim B, Kigozi G, Kiwanuka N, Phillips JB, Wabwire F, Meehan M, Lutalo T, Lane JR, Merling R, Gray R, Wawer M, Birx DL, Robb ML and McCutchan FE. 2002. Among 46 near full-length HIV type 1 genome sequences from Rakai district, Uganda, subtype D and AD recombinants predominate. AIDS Res Hum Retroviruses 18(17): 1281-1290

Harti DL and Clark AG. 1997. Principals of population genetics. Sinauer Associates, Sunderland, MA

Haseltine WA. 1992. The molecular biology of HIV-pp39-59 In: AIDS (Etiology, Diagnosis, Treatment and Prevention), third edition. Edited by DeVita VT, Hellman S, Rosenberg SA. JB Lippincott Company, Philadelphia

Heinzinger NK, Bukinsky MI, Haggerty SA, Ragland AM, Kewalramani V, Lee MA, Gendelman HE, Ratner L, Stevenson M and Emerman M. 1994. The vpr protein of human immunodeficiency virus type 1 influences nuclear localization of viral nucleic acids in non-dividing cells. Proc Nat Acad Sci USA 91: 7311-7315

Hierholzer M, Graham RR, El Khidir I, Tasker S, Darwish M, Chapman GD, Fagbami AH, Soliman A, Birx DL, McCutchan F and Carr JK. 2002. HIV type 1 strains from East and West Africa are intermixed in Sudan. AIDS Res Hum Retroviruses 18 (15):1163-6.

Higgins D. 2003. Basic concepts of molecular evolution. p1-23. In: The phylogenetic handbook: A practical approach to DNA and protein phylogeny. Eds: Salemi M and Vandamme A. Cambridge University Press, Cambridge, UK

HIV Sequence Compendium. 2002. Kuiken CL, Foley B, Freed E, Hahn B, Marx PA, McCutchan F, Mellors JW, Wolinksy S and Korber B, Eds. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM, LA-UR 03-3564.

Ho DD and Huang Y. 2002. The HIV-1 vaccine race. Cell 110 (2): 135-138

Hoffman NG, Seillier-Moiseiwitsch F, Ahn J, Walker JM and Swanstrom R. 2002. Variability in the human immunodeficiency virus type 1 gp120 Env protein linked to phenotype-associated changes in the V3 loop. J Virol. 76 (8) :3852-64

http://www.hiv.lanl.gov. Los Alamos National Laboratory HIV Database.

http://www.chemsoc.org/exemplarchem/entries/2002/levasseur/images/hiv. GIF. A cartoon illustration of the HIV-1 virion displaying the viral envelope, gag and pol proteins. Accessed: 10-3-2004

Hu DJ, Baggs J, Downing RG, Pieniazek D, Dorn J, Fridlund C, Birayakwaho B, Sempula SDK, Rayfield MA, Dondero TJ and Lal R. 2000. Predominance of HIV-1 subtype A and D infections in Uganda. Emerging Infect Disease (November-December): 609-615

Hunt GM, Papathanasopoulos MA, Gray GE and Tiemessen CT. 2003. Characterisation of near-full length genome sequences of three South African human immunodeficiency virus type 1 subtype C sequences. Virus Genes. 26(1): 49-56

Hunter E and Swanstrom R. 1990. Retrovirus envelope glycoproteins. Curr Top Microbiol Immunol 157:187-253

Janse van Rensburg E, Smith TL, Zeier M, Robson B, Sampson C, Treurnicht F and Engelbrecht S. 2002. Change in co-receptor usage of current South African HIV-1 subtype C primary sequences. AIDS 16(18): 2479-80

Janssens W, Buvé A and Nkengasong JN. 1997. The puzzle of HIV-1 subtypes in Africa. AIDS 11:705-712

Javaherian K, Langlois A J, McDanal C, Ross KL, Eckler LI, Jellis CL, Profy AT, Rusche JR, Bolognesi DP, Putney SD and Matthews TJ. 1989. Principal neutralizing domain of the human immunodeficiency virus type 1 envelope protein. PNAS 86 (17): 6768-72

Jeang KT, Chun R, Lin NH, Gatignol A, Glabe CG and Fan H. 1993. In vitro and in vivo binding of human immunodeficiency virus type 1 Tat protein and Sp1 transcription factor. J Virol 67(10): 6224-33.

Kaleebu P, French N, Make C, Yirrell D, Watera C, Lyagoba F, Nakiyingi J, Rutebemberwa A, Morgan D, Weber J, Gilks C and Whitworth J. 2002. Effect of HIV type 1 envelope subtype A and D on disease progression in a large cohort of HIV-1-positive persons in Uganda. J Infect Dis 185: 1244-1250

Kaleebu P, Ross A, Morgan D, Yirrell D, Oram J, Rytebemberwa A, Lyagoba F, Hamilton L, Biryahwaho B and Whitwoth J. 2001. Relationship between HIV-1 Env subtypes A and D and disease progression in a rural Ugandan cohort. AIDS 15: 293-299

Kanki PJ, Hamel DJ, Sankale JL, Hsieh C, Thior I, Barin F, Woodcock SA, Gueye-Ndiaye A, Zhang E, Montano M, Siby T, Marlink R, NDoye I, Essex ME and MBoup S. 1999. Human immunodeficiency virus type 1 subtypes differ in disease progression. J Infect Dis 179: 68-73

Kimura M. 1980. A simple method for estimating volutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of Mol Evol 15: 111-120

Kitabwalla M, Ferrantelli F, Wang T, Chalmers A, Katinger H, Stiegler G, Cavacini CA, Chen T and Ruperth RM. 2003. Primary African HIV clade A and D Sequences: Effective cross-clade neutralization with a quadruple combination of human monoclonal antibodies raised against clade B. AIDS Res Hum Retroviruses 19 (2): 125-131

Klostermeier D, Bayer P, Kraft M, Frank RW and Rosch P. 1997. Spectroscopic investigations of HIV-1 trans-activator and related peptides in aqueous solutions. Biophys Chem 63(2-3): 87-96

Korber BTM, Learn G, Mullins JI, Hahn BH and Wolinsky S. 1995. Protecting HIV sequence databases. Nature 378:242-243

Korber B, MacInnes K, Smith RF and Myers G. 1994. Mutational trends in V3 loop protein sequences observed in different lineages oh human immunodeficiency virus type 1. J Virol 68:6730-6744

Korber B, Muldoon M, Theiler J, Gao F, Gupta R, Lapedes A, Hahn BH, Wolinsky S and Bhuttacharga T. 2000.Timing the ancestor of the HIV-1 pandemic strains. Science 288 (June 9): 1789- 1795

Koulinska IN, Chaplin B, Essex M and Renjifo B. 2003. Hypermutation of HIV-1 genomes sequenced from infants soon after vertical transmission. AIDS Res Hum Retroviruses 19(12): 1115-23

Koulinska IN, Msamanga G, Mwakagile D, Essex M and Renjifo B. 2002. Common genetic arrangements among human immunodeficiency virus type 1 subtype A and D recombinant genomes vertically transmitted in Tanzania. AIDS Res Hum Retroviruses 18(13): 947-56

Koulinska IN, Ndung'u T, Mwakagile D, Msamanga G, Kagoma C, Fawzi W, Essex M and Renjifo B. 2001. A new human immunodeficiency virus type 1 circulating recombinant form from Tanzania. AIDS Res. Hum. Retroviruses 17(5): 423-31

Kuiken CL and Leitner T. 2001. HIV-1 subtyping. In: Computational and evolutionary analysis of HIV molecular sequences. Eds Rodrigo AG and Learn GH Jnr. Kluwer Academic Publishers, Massachusetts (MA).

Kustner H. 1994. Epidemiological Comments. Pretoria: Department of Health and Welfare. 21 (11) : 223-246

Laukkanen T, Carr JK, Janssens W, Liitsola K, Gotte D, McCutchan FE, Op de Coul E, Cornelissen M, Heyndrickx L, van der Groen G and Salminen MO. 2000. Virtually full-length subtype F and F/D recombinant HIV-1 from Africa and South America. Virology Mar 30;269(1):95-104

Laukkanen T, Liitsola K, Salminen MO and Leinikki P. 1996. HIV-1 D subtype in Finland. Clinical and Diagnostic Virology 5: 206-210

Lee WR, Syu WJ, Du B, Matsuda M, Tan S, Wolf A, Essex M and Lee TH. 1992. Nonrandom distribution of gp120 N-linked glycosylation sites important for viral infectivity of human immunodeficiency virus type 1. Pro Nat Acad Sci USA 89: 2213-2217

Lemey P, Pybus OG, Wang B, Saksena NK, Salemi M and Vandamme A. 2003. Tracing the origin and history of the HIV-2 epidemic. PNAS 100 (11): 6588-6952

Levy JA. 1994. HIV and the pathogenesis of AIDS. pp 5- 12. ASM Press. Washington DC.

Levy JA, Hoffman AD, Kramer SM, Landis JA, Shimabukuro JM and Oshiro LS. 1984.Isolation of lymphocytopathic retroviruses from San Francisco patients with AIDS. Science 225:840-842

Li WH. 1997. Molecular evolution. Sinauer Associates, Sunderland, Massachusetts, (MA), USA

Luciw P.1996. Human Immunodeficiency Viruses and their Replication. pp 1881-1952. In: Fields Virology, third edition. Edited by Fields BN, Knipe DM, Howley PM. Lippincott-Raven publishers, Philadelphia

Ma C, Marassi FM, Jones DH, Straus SK, Bour S, Strebel K, Schubert U, Oblatt-Montal M, Montal M and Opella SJ. 2002. Expression, purification, and activities of full-length and truncated versions of the integral membrane protein Vpu from HIV-1. Protein Sci 11(3): 546-57

Marck C. 1998. DNA Strider: a 'C' program for the fast analysis of DNA and protein sequences on the Apple Macintosh family of computers. Nucleic Acids Res 16: 1829-1836

Marshall RD. 1974. The nature and metabolism of the carbohydrate-peptide linkages of glycoproteins. Biochem Soc Symp 40: 17-26

McCormack GP, Glynn JR, Crampin AC, Sibande F, Mulawa D, Bliss L, Broadbent P, Abarca K, Ponnighaus JM, Fine PE and Clewley JP. 2002. Early evolution of the human immunodeficiency virus type 1 subtype C epidemic in rural Malawi. J Virol 76(24):12890-9.

McCutchan FE. 2000. Understanding the genetic diversity of HIV-1. AIDS 14 (Suppl. 3): S31-S44

McCutchan FE, Salminen MO, Carr JK, and Burke DS. 1996. HIV-1 genetic diversity. AIDS 10 (Suppl. 3):S13–S20.

Metzger AU, Bayer P, Willbold D, Hoffmann S, Frank RW, Goody RS and Rosch P. 1997. The interaction of HIV-1 Tat(32-72) with its target RNA: a fluorescence and nuclear magnetic resonance study. Biochem Biophys Res Commun 241(1): 31-6

Milich L, Margolin B and Swanstrom R. 1993. V3 loop of the human immunodeficiency virus type 1 Env protein: interpreting sequence variability. J Virol 67(9): 5623-34

Montal M. 2003. Structure-function correlates of Vpu, a membrane protein of HIV-1. FEBS Lett 552(1):47-53

Montefiori DC, Robinson WE Jr and Mitchell WM. 1988. Role of protein N-glycosylation in pathogenesis of human immunodeficiency virus type 1. PNAS 85(23):9248-52

Moore JP, Parren PWHI and Burton DR. 2001. Genetic subtypes, Humoral immunity and human immunodeficiency virus type 1 vaccine development. J Virol 75 (13): 5721-5729

Myers G and Lenroot R. 1992. HIV glycosylation: what does it portend? AIDS Res Hum Retroviruses 8: 1459-1460

Myers G, MacInnes K and Korber B. 1992. The emrgence of simian/human immunodeficiency viruses. AIDS Res Hum Retroviruses 8:373-386

Nahmias AJ, Weiss J Yao X, Lee F, Kodsi R, Schanfield M, Matthews T, Bolognesi D, Durack D and Motulsky A. 1986. Evidence for human infection with an HTLV III/ LAV like virus in central Africa, 1952. Lancet I: 1279-1280

Palmer S, Alaeus A, Albert J and Cox S. 1998. Drug susceptibility of subtypes A, B, C, D and E human immunodeficiency virus type 1 primary sequences. AIDS Res Hum Retroviruses 14 (2): 157-162

Papathanasopoulos MA, Patience T, Meyers TM, Morris L and McCutchan F. 2003. Full-length genome characterisation of HIV type 1 subtype C sequences from two slow-progressing perinatally infected siblings in South Africa. AIDS Res. Hum. Retroviruses 19 (11): 1033-7

Papathanasopoulus MA, Cilliers T, Morris L, Mokili JL, Dowling W, Birx DL and McCutchan FE. 2002. Full-length analysis of HIV-1 subtype C utilizing CXCR4 and intersubtpye recombinants sequenced in South Africa. AIDS Res Hum Retroviruses 18(12): 879-886

Paul M, Mazumder S, Raja N and Jabbar MA. 1998. Mutational analysis of the humn immunodeficiency virus type 1 vpu transmembrane domain that promotes the enhanced release of virus-like particles from the plasma membrane of mammalian cell. J Virol 72: 1270-1279

Peeters M. 2000. Genetic diversity of HIV-1: the moving target. AIDS 14(Suppl3): S129-S140

Peeters M. 2001. The genetic variability of HIV-1 and its implications. Transfus Clin Biol 8:222-225

Peeters M and Courgnaud V. 2002. Overview of primate lentiviruses and their evolution in non-human primates in Africa. In: HIV Sequence Compendium 2002. Kuiken CL, Foley B, Freed E, Hahn B, Marx PA, McCutchan F, Mellors JW, Wolinksy S and Korber B, Eds. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM, LA-UR 03-3564.

Pillia S, Good B, Richman D and Corbeil J. 2003. A new perspective on V3 phenotype prediction. Aids Res Hum Retroviruses 19 (2): 145-149

Piot P, Ouinn TC, Taelman H, Feinsod FM, Minlangu KB, Wobin O, Mbendi N, Mazebo P, Ndangi K and Stevens W. 1984. Acquired immunodeficiency syndrome in a heterosexual population in Zaire. Lancet 2:65-69

Posada D, Crandall KA and Hillis DM. 2001. Phylogenetics of HIV. In: Computational and Evolutionary analysis of HIV molecular sequences. Rodrigo AG and Learn GH (jnr). Eds. Kluwer Academic Publishers, Massachusetts (MA), USA

Preston BD, Poiesz BJ and Loeb LA. 1988. Fidelity of HIV-1 reverse transcriptase. Science 242:1168-1171

Puren AJ. 2002. The HIV-1 epidemic in South Africa. Oral Diseases 8 (Suppl2): 27-31

Nei M and Kumar S. 2000. Molecular evolution and phylogenetics. Chp5. Oxford University Press Publishers, New York, NY

Neilson JR, John GC, Carr JK, Lewis P, Kreiss JK, Jackson S, Nduati RW, Mbori-Ngacha D, Panteleeff DD, Bodrug S, Giachetti C, Bott MA, Richardson BA, Bwayo J, Ndinya-Achola J and Overbaugh J. 1999. Subtypes of Human Immunodeficiency Virus type 1 and disease stage among women in Nairobi, Kenya. J Virol 73(5): 4393-4403

Neuveut C and Jeang KT. 1996. Recombinant human immunodeficiency virus type 1 genomes with tat unconstrained by overlapping reading frames reveal residues in Tat important for replication in tissue culture. J Virol 70(8): 5572-81.

Novelli P, Vella C, Oxford J and Daniels RS. 2002. Construction and characterisation of a full-length HIV-1 (92UG001) subtype D infectious molecular clone. AIDS Res. Hum. Retroviruses 18 (1): 85-88

Novitsky VA, Montano MA, McLane MF, Renjifo B, Vannberg F, Foley BT, Ndung'u TP, Rahman M, Makhema MJ, Marlink R and Essex M. 1999. Molecular cloning and phylogenetic analysis of human immunodeficiency virus type 1 subtype C: a set of 23 full-length clones from Botswana. J Virol 73: 4427 - 4432

Ogert RA, Lee MK, Ross W, Buckler-White A, Martin MA and Cho MW. 2001. N-linked glycosylation sites adjacent to and within the V1/V2 and the V3 loops of dualtropic human immunodeficiency virus type 1 sequence DH12 gp120 affect coreceptor usage and cellular tropism. J Virol 75 (13): 5998-6006

O'Hagen A, Devitt A, Kunstman KJ, Gorry PR, Rose PP, Korber B, Taylor J, Levy R, Murphy RL, Wolinsky SM and Gabuzda D. 2003. Genetic and functional analysis of full-length human immunodeficiency virus type 1 env genes derived from brain and blood of patients with AIDS. J Virol 77 (22): 12336-12345

Oleske J, Muimefor A, Cooper R Jr, Thomas K, dela Cruz A, Ahdieh H, Guerrero I, Joshi VV and Desposito F. 1983. Immune deficiency syndrome in children. JAMA 249:2345-2351

Osmanov S, Pattou C, Walker N, Schwarlander B, Esparza J and the WHO-UNAIDS Network for HIV Isolation and Characterisation. 2002. Estimated global distribution and regional spread of HIV-1 genetic subtypes in the year 2002. JAIDS 29: 184-190

Pal R, Hoke GM and Sarngadharan MG. 1989. Role of oligosaccharides in the processing and maturation of envelope glycoproteins of human immunodeficiency virus type 1. Proc Natl Acad Sci U S A. 86 (9): 3384-8

Page RDM and Holmes EC. 1998. Molecular Evolution: A phylogenetic approach. p11-36. Blackwell Science Ltd, Oxford UK

Ras GJ, Simson IW, Anderson R, Prozesky OW and Hamersma T.1983. Acquired immunodeficiency syndrome: a report of 2 South African cases. S Afr Med J 64: 140-142

Rayfield MA, Downing RG, Baggs J, Hu DJ, Pieniazek D, Luo CC, Biryahwaho B, Otten RA, Sempala SD and Dondero TJ. 1998. A molecular epidemiologic survey of HIV in Uganda. HIV Variant Working Group. AIDS 12: 521 - 527

Renjifo B, Fawzi W, Mwakagile D, Hunter D, Msamanga G, Spiegelman DE, Garland M, Kagoma C, Kim A, Chaplin B, Hertzmark E and Essex M. 2001. Differences in perinatal transmission among human immunodeficiency virus type 1 genomes. J Hum Virol 4: 16-25

Renjifo B, Chaplin B, Msamanga G, Shah P, Vannberg F, Renjifo B and Essex M. 1999. Emerging recombinant human immunodeficiency viruses: uneven representation of the envelope V3 region. AIDS 13(13): 1613-21

Rey-Cuille MA, Berthier JL, Bomsel-Demontoy MC, Chaduc Y, Montagnier L, Hovanessian LA and Chakrabarati LA. 1998. Simian immunodeficiency virus replicates to high levels in sooty mangabeys without inducing disease. J Virol 72: 3872-3886

Robertson DL, Anderson JP and Bradac JA. 1999. HIV-1 nomenclature proposal. In: Human retroviruses and AIDS 1999: a compilation and analysis of nucleic acid and amino acid sequences. Kuiken CL, Foley B, Hahn B, Korber B, McCutchan F, Marx PA. Eds. Theoretical biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM

Robertson DL, Sharp PM, McCutchan FE and Hahn BH. 1995. Recombination in HIV-Nature 374: 124-126

Rodenburg CM, Li Y, Trask SA, Chen Y, Decker J, Robertson DL, Kalish ML, Shaw GM, Allen S, Hahn BH, Gao F and the UNAIDS and NIAID Networks for HIV isolation and Characterisation. 2001. Near-full length clones and reference sequences for subtype C sequences of HIV type 1 from three different continents. AIDS Res Hum Retroviruses 17(2): 161-168

Saitou N and Nei M. 1987. The neighbour-joining method: a new method for reconstructing phylo-genetic trees. Molecular Biology and Evolution 4(4):406-425.

Salemi M, De Oliveira T, Courgnaud V, Moulton V, Holland B, Cassol S, Switzer WM and Vandamme A. 2003. Mosaic genomes of the six major primate lentivirus lineages revealed by phylogenetic analyses. J Virol 77 (13) : 7202-7213

Salminen MO, Carr JK, Burke DS and McCutchan FE. 1995a. Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning. AIDS Res Hum Retroviruses 11: 1423-1425

Salminen MO, Koch C, Sander-Buell E, Ehrenberg PK, Michael NL, Carr JK, Burke DS and McCutchan FE. 1995b. Recovery of virtually full-length HIV-1 provirus of diverse subtypes from primary virus cultures using the polymerase chain reaction. Virology 213: 80-86

Sambrook S, Fritsch EF and Manniatis T, eds. 1989. Molecular cloning: A laboratory Manual, 2nd edition, pp 9.14 – 9.23, 1.25 – 1.28, E5 and A5. Published by Cold Spring Harbor Laboratory Press, New York, NY.

Sanders-Buell E, Salminen MO and McCutchan FE. 1995. Sequencing primers for HIV-1, pp III-15 to III-21. In Kuiken C, Foley B, Hahn B, Marx P, McCutchan F, Mellors J, Mullins J, Wolinsky S and Korber B, eds. Human Retroviruses and AIDS 1995: A Compilation and Analysis of Nucleic Acid and Amino Acid Sequences. Published by Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM

Sattentau QJ and Moore JP. 1991. Conformational changes induced in the human immunodeficiency virus envelope glycoprotein by soluble CD4 binding. J Exp Med. Aug 1;174(2):407-15

Schonning K, Jansson B, Olofsson S and Hansen JE. 1996. Rapid selection for an N-linked oligosaccharide by monoclonal antibodies directed against the V3 loop of human immunodeficiency virus type 1. J Gen Virol 77: 753-8

Schwartz S, Felber BK, Fenyo EM and Pavlakis GN. 1990. Env and Vpu proteins of human immunodeficiency virus type 1 are produced from multiple bicistronic mRNAs. J Virol 64(11): 5448-56

Scriba TJ, Treurnicht FK, Zeier M, Engelbrecht S and Janse van Rensburg E. 2001. Characterisation and phylogenetic analysis of South African HIV-1 subtype C accessory genes. AIDS Res Hum Retroviruses. 2001 May 20; 17(8): 775-81

Schubert U and Strebel K. 1994. Differential activities of the human immunodeficiency virus type 1-encoded Vpu protein are regulated by phosphorylation and occur in different cellular compartments. J Virol 68 (4): 2260-71

Schwartz S, Felber BK, Fenyo EM and Pavlakis GN. 1990. Env and Vpu proteins of human immunodeficiency virus type 1 are produced from multiple bicistronic mRNAs. J Virol 64(11): 5448-56

Sher R. 1989. HIV infection in South Africa, 1982-1988- a review. SAMT 76 (October): 314-318

Sherman MP, De Noronha CM, Williams SA and Greene WC. 2002. Insights into the biology of HIV-1 viral protein R. DNA Cell Biol 21(9): 679-88

Spire B, Sire J, Zachar V, Rey F, Barre-Sinoussi F, Galibert F, Hampe A and Chermann JC. 1989. Nucleotide sequence of HIV1-NDK: a highly cytopathic strain of the human immunodeficiency virus. Gene 81 (2): 275-284

Spira S, Wainberg MA, Loemba H, Turner D and Brenner BG. 2003. Impact of clade diversity on HIV-1 virulence, antiretroviral drug sensitivity and drug resistance. J Antimicrobial Chemotherapy 51: 229-240

Srinivasan A., Anand R, York D, Ranganathan P, Feorino P, Schochetman G, Curran J, Kalyanaraman WITH, Luciw PA and Sanchez-Pescador R. 1987. Molecular characterisation of human immunodeficiency virus from Zaire: nucleotide sequence analysis identifies conserved and variable domains in the envelope gene. Gene 52 (1): 71-82

Stanfield R, Cabezas E, Satterthwait A, Stura E, Profy A and Wilson I. 1999. Dual conformations for the HIV-1 gp120 V3 loop in complexes with different neutralizing fabs. Structure Fold Des. Feb 15;7(2):131-42

Starcich BR, Hahn BH, Shaw GM, McNeely PD, Modrow S, Wolf H, Parks ES, Parks WP, Josephs SF, Gallo RC and Wong-Staal F. 1986. Identification and characterisation of conserved and variable regions in the envelope gene of HTLV-III/LAV, the retrovirus of AIDS. Cell 45: 637-645

Strebel K. 1996. Structure and function of HIV-1 Vpu. In: Human Retroviruses and AIDS 1996: A Compilation and Analysis of Nucleic Acid and Amino Acid Sequences. Myers G, Korber BT, Foley BT, Jeang K-T, Mellors JW and Wain-Hobson S, Eds. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM

Strebel K, Klimkait T, Maldarelli F and Martin MA. 1989. Molecular and biochemical analyses of human immunodeficiency virus type 1 vpu protein. J Virol 63 (9):3784-91

Strebel K, Klimkait T and Martin MA. 1988. A novel gene of HIV-1, vpu, and its 16-kilodalton product. Science 241(4870):1221-3

Thomson MM, Pérez-Alvarez L and Najera R. 2002. Molecular epidemiology of HIV-1 genetic forms and its significance for vaccine development and therapy. Lancet Infect Dis 2:461-471

Thompson JD, Higgins DG and Gibson TJ. 1994. Clustal W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acid Research 22: 4673-4680

Thompson JD, Gibson TJ, Plewwick F, Jeanmongin F and Higgins DG. 1997. The Clustal X windows interface:Flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acid Research 25: 4876-4882

Tong-Starksen SE, Baur A, Lu XB, Peck E and Peterlin BM. 1993. Second exon of Tat of HIV-2 is required for optimal trans-activation of HIV-1 and HIV-2 LTRs. Virology 195(2): 826-30.

Treurnicht FK, Smith TL, Engelbrecht S, Claassen M, Robson BA, Zeier M and van Rensburg EJ. 2002. Genotypic and phenotypic analysis of the env gene from South African HIV-1 subtype B and C sequences. J Med Virol 68(2): 141-6

Tscherning C, Alaeus A, Fredriksonn R, Bjorndal A, Deng H, Littman DR, Fenyo EM and Albert J. 1998. Differences in chemokines co-receptor usage between genetic subtypes of HIV-1. Virology 241: 181-188

UNAIDS. 2003. AIDS epidemic update: December 2003. Geneva. www.unaids.org

UNAIDS. 2002. AIDS epidemic update: December 2002. Geneva. www.unaids.org

Valentine-Thon E. 2002. Quality control in nucleic acid testing—where do we stand? J Clin Virol. Dec; 25 Suppl 3:S13-21

Vandamme A. 2003. Basic concepts of molecular evolution. p1-23. In: The phylogenetic handbook: A practical approach to DNA and protein phylogeny. Eds: Salemi M and Vandamme A. Cambridge University Press, Cambridge, UK

Van de Peer Y and De Wachter R. 1994. TREECON for Windows: a software package for the construction and drawing of evolutionary trees for the Microsoft Windows environment. Comput. Appl. Biosci. 10:569-570

Van Harmelen JH, Van der Ryst E, Loubser AS, York D, Madurai S, Lyons S, Wood R and Williamson C. 1999. A Predominantly HIV Type 1 Subtype C-Restricted Epidemic in South African Urban Populations. AIDS Res Hum Retroviruses Mar 1;15(4):395-8.

van Harmelen J, Williamson C, Kim B, Morris L, Carr J, Karim SS and McCutchan F. 2001. Characterisation of full- length HIV type 1 subtype C sequences from South Africa. AIDS Res. Hum. Retroviruses 17(16): 1527-31

van Harmelen J, Wood R, Lambrick M, Rybicki EP and Williamson C. 1997. An association between HIV-1 subtypes and mode of transmission in Cape Town, South Africa. AIDS 11: 81-87

Vartanian JP, Meyehans A, Henry M and Wain-Hobson S. 1992. High resolution structure of HIV-1 quasispecies: identification of novel coding sequences. AIDS 6: 1095-1098

Vidal N, Koyalta D, Richard V, Lechiche C, Ndinaromtan T, Djimasngar A, Delapote E and Peeters M. 2003. Genetic variability of HIV-1 strains in Chad: high prevalence of recombinant strains and identification of a new D' variant. JAIDS 33: 239-246

Vidal N, Peeters M, Mulanga-Kabeya G, Nzilambi N, Robertson D, Ilunga W, Sema H, Tshimanga K, Bongo B and Delaporte E. 2000. Unprecedented degree of human

immunodeficiency virus type 1 (HIV-1) group M genetic diversity in the Democratic Republic of the Congo suggest that the HIV-1 pandemic originated in Central Africa. J Virol 74: 10498-10507

Weniger BG, Tabeke Y, Ou C-Y and Yamazaki S. 1994. The molecular epidemiology of HIV in Asia. AIDS 8 (Suppl 2): S13-S28

Williamson C, Engelbrecht S, Lambrick M, Janse van Rensburg E, Wood R, Bredell W and Williamson A. 1995. HIV-1 subtypes in different risk groups in South Africa. Lancet 346 (September 16): 782

Zhong P, Burda S, Konings F, Urbanski M, Ma L, Zekeng L, Ewane L, Agyingi L, Agwara M, Saa, Afane ZE, Kinge T, Zolla-Pazner S, and Nyambi P. 2003. Genetic and biological properties of HIV type 1 sequences prevalent in villagers of the Cameroon equatorial rain forests and grass fields: further evidence of broad HIV type 1 genetic diversity. AIDS Res Hum Retroviruses 19(12):1167-78

Zhu T, Wang N, Carr A, Wolinsky A and Ho DD. 1995. Evidence for coinfection by multiple strains of human immunodeficiency virus type 1 subtype B in an acute seroconverter. J.Virol 69: 1324-1327

zur Megede J, Engelbrecht S, de Oliveira T, Cassol S, Scriba TJ, Janse van Rensburg E and Barnett SW. 2002. Novel evolutionary analyses of full-length HIV type 1 subtype C molecular clones from Cape Town, South Africa. AIDS Res Hum Retroviruses 18(17): 1327-1332

Zwick MB, Labrijin AF, Wang M, Spenlehauer C, Saphire EO, Binley JM, Moore JP, Stiegler G, Katinger H, Burton DR and Parren PW. 2001. Broadly neutralizing antibodies targeted to the membrane proximal external region of human immunodeficieny virus type 1 glycoprotein gp 41. J Virol 75 (22): 10892- 10905

APPENDICES

Appendix A

Primers used to sequence the HIV-1 D plasmids

Appendix B: Full-length nucleotide and amino acid sequences for isolates

A1: R2

A2: R214

A3: R286

A4: R482

Appendix C: NCBI subtyping results

A: R2, B: R214, C: R286, D: R482

Appendix D: Genetic distances between HIV-1 subtype D isolates

E1: Full-length subtype D similarity matrix

E2: *gag* similarity matrix

E3: *pol* similarity matrix

E4: *env* similarity matrix

E5: *vif* similarity matrix

E6: *vpr* similarity matrix

E7: *vpu* similarity matrix

E8: *tat* similarity matrix

E9: *rev* similarity matrix

E10: *nef* similarity matrix

Appendix E

HIV-1 subtype A-K consensus sequence glycosylation graphs

Appendix A

The table gives the primers that were used to sequence the Tygerberg plasmids: pR2, pR214, pR286 and pR482. The primers for the *gag* and *env* genes was designed and described by Sanders-Buell *et al* (1995). The primers designed to sequence gaps in the Tygerberg plasmids: G05D, Pol 1D, Pol 2D, Pol 2Drev, Pol 3D, Pol 3Drev, Pol DF, Pol DR, Env DF and Env DR are described in thesis for the first time.

Appendix A – Primers used to sequence the Tygerberg HIV-1 D plasmids

Primer	Sequence (5'-3')	Primer	Sequence	Reference
G 00	GACTAGCGGAGGCTAGAAG	G01	AGGGGTCGTTGCCAAAGA	Sanders-Buell (1995)
G10	CAGTATTAAGCGGGGGAGAATT	G05	TGTTGGCTCTGGTCTGCTCT	
G20	GTATGGGCAAGCAGGGAGCTAGAA	G15	CTTTGCCACAATTGAAACACTT	
G30	CAGTAGCAACCCTCTATTGTGT	G25	ATTGCTTCAGCCAAAACTCTTGC	
G40	GACACCAAGGAAGCTTTAGA	G35	CATGCTGTCATCATTTCTTCTA	
G50	CACAGCAAGCAGCAGCTG	G45	TTGGACCAACAAGGTTTCTGTC	
G60	CAGCCAAAATTACCCTATAGTGCAG	G55	ATTTCTCCCACTGGGATAGGTGG	
G70	ATGAGGAAGCTGCAGAATGGG	G65	ATGCTGAAAACATGGGTA	
G80	ATGAGAGAACCAAGGGGAAGTGA	G75	CTTCTATTACTTTTACCCATGC	
G90	ATAATCCACCTATCCCAGTAGGAGAAAT	G85	TGCACTATAGGGTAATTTTG	
G100	TAGAAGAAATGATGACAG			
G110	AGGCTAATTTTTTAGGGA			
E0	TAGAGCCCTGGAAGCATCCAGGAAGTCAGCCTA	E01	TCCAGTCCCCCCTTTTCTTTTAAAAA	
E00	TAGAAAGAGCAGAAGACAGTGGCAATGA	E03	TAAGTCATTGGTCTTAAAGGTACCTG	
E10	TTGTGGGTCACAGTCTATTATGGGGT	E05	TATTTGAGGGCTTCCCACCCCC	
E20	GGGCCACACATGCCTGTGTACCCACAG	E15	CTCTCTCTCCACCTTCTTCTTC	
E30	GTGTACCCACAGACCCCAGCCCACAAG	E25	GGTGAGTATCCCTGCCTAAC	
E40	CATGTGGAAAAATGACATGGTGGATCA	E35	GGTGAGTATCCCTGCCTAACTCTATT	
E50	CATGGTAGAGCAGATGCAGGAGGATG	E45	CCTGCCTAACTCTATTCAC	
E60	TAATCAGTTTATGGGATCAAAGC	E55	GCCCCAGACTGTGAGTTGCAACAGATG	
E70	GGGATCAAAGCCTAAAGCCATGTGTAA	E65	AGTGCTTCCTGCTGCTCC	
E80	CCAATTCCCATACATTATTGTG	E75	GCGCCCATAGTGCTTCCTGCTGCTCCC	
E90	CACAGTACAATGTACACATGGAAT	E85	GTCCCTCATATCTCCTCCTCCAGGTCT	
E100	ACACATGGAATTAAGCCAGT	E95	GATGGGAGGGGCATACAT	
E110	CTGTTAAATGGCAGTCTAGCAGAA	E105	GCTTTTCCTACTTCCTGCCAC	
E120	GTAGAAATTAATTGTACAAGACCC	E115	AGAAAAATTCCCCTCCACAATTAA	
E130	ACAAATTATAAACATGTGGCAGG	E125	CAATTTCTGGGTCCCCTCCTGAGG	
E140	GTGAATTATATAAATATAAAGTAG	E135	AGCTGTACTATTATGGTTTTAGCATTGT	
E150	CCAGGGCAAAGAGAAGAGTGGTG	E145	CAGCAGTTGAGTTGATACTACTGG	
E160	GTGGGAATAGGAGCTGTGTTCCTTGGG	E155	CTGTTCTACCATGTTATTTTTCCACATGT	
E170	AGCAGGAAGCACTATGGG	E165	GGGGTCTGTGGGTACACAGGCATGTGT	
E180	GTCTGGTATAGTGCAACAGCA	E175	TTTAGCATCTGATGCACAAAATAG	
E190	CCTGGAACTCCACTTGGAG			
E200	GGGATAACATGACCTGGATGCAGTGGG			
E210	TAACAAATTGGCTGTGGTATATAA			
E220	TATCAAAATGGCTGTGGTATATAA			
E230	AATATTCATAATGATAGTAGGAGG			
E240	ATAATGATAGTAGGAGGCTTGATAGGC			
E250	GGAGGCTTGATAGGTTTAAGAATA			
E260	TTCAGCTACCACCGCTTGAGAGACT			
E270	GTGGAACTTCTGGGACGCAG			
G05D	ATG CAG AGA GGC AAT TTT AAG G			
Pol1D	TCC CTC AAA TCA CTC TTT GGC			
Pol2D	CTA TTG AAA CTG TAC C			
Pol2Drev	CCA TCC ATT CCT GGC			
Pol3D	CAG TAC TGG ATG TGG G			
Pol3Drev	CCC ACA TCC AGT ACT G			
Pol-DF	TTG TAC AGA TAT GGA AAA GGA AGG			
Pol-DR	AAT TTA GGA GTC TTT CCC			
Env-DF	GGT CAC AGT TTA TTA TGG G			
Env-DR	5'- GAA TTG CAA AAC CAG CTG G - 3'			

Appendix B

Full-length nucleotide and amino acid sequences for HIV-1 subtype D plasmids: PR2, pR214, pR286 and pR482

B1: Full-length sequence of pR2
B2: Full-length sequence of pR214
B3: Full-length sequence of pR286
B4: Full-length sequence of pR482

The full-length sequences in B1-B4 contain both the nucleotide and amino acid sequences of the plasmids. The start and end of the viral genes as well as the nucleotide positions are indicated.

B1: **pR2 Full-length sequence** (nucleotide and amino acid)

```
181         GACTGGTGAGTACGCTAAAAATTTTGACTAGCGGAGGCTAGAAGGAGAGAGATGGGTGCG
(gag start)                                                       M  G  A

241         AGAGCGTCAGTATTAAGCGGGGGAAAATTAGATGCATGGGAAAGAATTCGGTTAAGGCCA
(gag)        R  A  S  V  L  S  G  G  K  L  D  A  W  E  R  I  R  L  R  P

301         GGAGGGAAGAAAAAATATAAACTAAAACATATAGTATGGGCAAGCAGGGAGCTAGAACGA
(gag)        G  G  K  K  K  Y  K  L  K  H  I  V  W  A  S  R  E  L  E  R

361         TTTGCACTTAATCCTAGCCTTTTAGAAACAGCAGAAGGATGTAAACAAATAATAGGACAG
(gag)        F  A  L  N  P  S  L  L  E  T  A  E  G  C  K  Q  I  I  G  Q

421         CTACAACCAGCTGTTCAGACAGGATCAGAAGAACTTAAATCATTATATAATACAGTAATA
(gag)        L  Q  P  A  V  Q  T  G  S  E  E  L  K  S  L  Y  N  T  V  I

481         ACCCTCTATTGTGTACATGAAAGGATAGATGTAAAAGACACCAAGGAAGCTTTAGAAAAG
(gag)        T  L  Y  C  V  H  E  R  I  D  V  K  D  T  K  E  A  L  E  K

541         ATAGAGGAAGAACAAAACAAAAGTAAGAAAAAGAAGGCACAGCAAGCAGCAGCTGACACA
(gag)        I  E  E  E  Q  N  K  S  K  K  K  K  A  Q  Q  A  A  A  D  T

601         GGAAACAGCAGCCAGGTCAGCCAAAATTATCCTATAGTGCAGAACCTACAGGGGCAAATG
(gag)        G  N  S  S  Q  V  S  Q  N  Y  P  I  V  Q  N  L  Q  G  Q  M

661         GTACATCAGGCCATATCACCTAGAACTTTGAATGCATGGGTAAAAGTAATAGAAGAAAAG
(gag)        V  H  Q  A  I  S  P  R  T  L  N  A  W  V  K  V  I  E  E  K

721         GCCTTCAGCCCAGAAGTAATACCCATGTTTTCAGCATTATCAGAAGGAGCCACCCCACAA
(gag)        A  F  S  P  E  V  I  P  M  F  S  A  L  S  E  G  A  T  P  Q

781         GATTTAAACACCATGCTAAACACAGTGGGGGGACATCAAGCAGCCATGCAAATGCTAAAA
(gag)        D  L  N  T  M  L  N  T  V  G  G  H  Q  A  A  M  Q  M  L  K

841         GAGACCATCAATGAAGAAGCTGCAGAATGGGATAGGCTACATCCAGTGCATGCAGGGCCT
(gag)        E  T  I  N  E  E  A  A  E  W  D  R  L  H  P  V  H  A  G  P

901         ATTGCACCAGGCCAGATGAGAGAACCAAGGGGAAGTGATATAGCAGGAACTACTAGTACC
(gag)        I  A  P  G  Q  M  R  E  P  R  G  S  D  I  A  G  T  T  S  T

961         CTTCAGGAACAAATAGCATGGATGACAAGCAACCCACCTATCCCAGTAGGAGAAATCTAT
(gag)        L  Q  E  Q  I  A  W  M  T  S  N  P  P  I  P  V  G  E  I  Y
```

```
1021        AAAAGATGGATAATCCTGGGATTAAATAAAATAGTAAGAATGTATAGCCCTGTCAGCATT
(gag)        K  R  W  I  I  L  G  L  N  K  I  V  R  M  Y  S  P  V  S  I

1081        TTGGACATAAGACAGGGACCAAAGGAACCTTTTAGAGATTATGTAGACCGGTTCTATAAA
(gag)        L  D  I  R  Q  G  P  K  E  P  F  R  D  Y  V  D  R  F  Y  K

1141        ACTCTAAGAGCCGAGCAAGCTTCACAGGATGTAAAAAACTGGATGACAGAAACCTTGTTG
(gag)        T  L  R  A  E  Q  A  S  Q  D  V  K  N  W  M  T  E  T  L  L

1201        GTCCAAAATGCAAACCCAGATTGTAAAACCATCTTAAAAGCATTAGGACCACAGGCTACA
(gag)        V  Q  N  A  N  P  D  C  K  T  I  L  K  A  L  G  P  Q  A  T

1261        CTAGAAGAAATGATGACAGCGTGTCAGGGAGTGGGGGGGCCCAGCCATAAAGCAAGAGTT
(gag)        L  E  E  M  M  T  A  C  Q  G  V  G  G  P  S  H  K  A  R  V

1321        TTGGCTGAGGCAATGAGCCAAGCAACAAATTCAGCTACTGCAGTAATGATGCAGAGAGGC
(gag)        L  A  E  A  M  S  Q  A  T  N  S  A  T  A  V  M  M  Q  R  G

1381        AATTTTAAGGGCCAAAGAAAAATTATTAAGTGTTTCAACTGTGGCAAAGAAGGGCACATA
(gag)        N  F  K  G  Q  R  K  I  I  K  C  F  N  C  G  K  E  G  H  I

1441        GCAAAAAATTGCAGGGCCCCTAGGAAAAAGGGCTGTTGGAAATGTGGAAGGGAAGGACAC
(gag)        A  K  N  C  R  A  P  R  K  K  G  C  W  K  C  G  R  E  G  H

1501        CAAATGAAAGAGTGCACTGAAAGACAGGCTAATTTTTTAGGGAAAATTTGGCCTTCCCAC
(gag)        Q  M  K  E  C  T  E  R  Q  A  N  F  L  G  K  I  W  P  S  H
(pol start)                                  F  F  R  E  N  L  A  F  P  Q

1561        AAGGGAAGGCCGGGGAACTTTCTTCAGAGCAGACCAGAGCCAACAGCCCCACCATCAGAG
(gag)        K  G  R  P  G  N  F  L  Q  S  R  P  E  P  T  A  P  P  S  E
(pol)          G  K  A  G  E  L  S  S  E  Q  T  R  A  N  S  P  T  I  R  E

1621        AGCTTCGGGTTTGGGGAGGAGATAACCCCCTCTCAGAAACAGGAACAGAAAGACAAGGAA
(gag)        S  F  G  F  G  E  E  I  T  P  S  Q  K  Q  E  Q  K  D  K  E
(pol)          L  R  V  W  G  G  D  N  P  L  S  E  T  G  T  E  R  Q  G  T

1681        CTGTATCCTTTAACCTCCCTCAAATCACTCTTTGGGAGCGACCCCTTGTCACAATAAAAA
(gag end)    L  Y  P  L  T  S  L  K  S  L  F  G  S  D  P  L  S  Q  *
(pol)          V  S  F  N  L  P  Q  I  T  L  W  E  R  P  L  V  T  I  K  I

1741        TAGGGGGACAGCTAAAGGAAGCTCTATTAGATACAGGAGCAGATGATACAGTATTAGAAG
(pol)          G  G  Q  L  K  E  A  L  L  D  T  G  A  D  D  T  V  L  E  E

1801        AAATGAATTTGCCAGGAAAATGGAAACCAAAAATGATAGGGGGAATTGGAGGTTTTATCA
(pol)          M  N  L  P  G  K  W  K  P  K  M  I  G  G  I  G  G  F  I  K
```

```
1861        AAGTAAGACAGTATGATCAAATACCCCTAGAAATCTGTGGGCATAAAGCTATAGGTACAG
(pol)         V  R  Q  Y  D  Q  I  P  L  E  I  C  G  H  K  A  I  G  T  V

1921        TATTGATAGGACCTACACCTGTCAACATAATTGGAAGAAATTTGTTGACTCAGCTTGGCT
(pol)         L  I  G  P  T  P  V  N  I  I  G  R  N  L  L  T  Q  L  G  C

1981        GCACTTTAAATTTTCCAATTAGTCCTATTGAAACTGTACCAGTAAAATTAAAGCCAGGAA
(pol)         T  L  N  F  P  I  S  P  I  E  T  V  P  V  K  L  K  P  G  M

2041        TGGATGGCCCAAAAGTTAAACAATGGCCATTGACAGAAGAAAAAATAAAAGCATTAACAG
(pol)         D  G  P  K  V  K  Q  W  P  L  T  E  E  K  I  K  A  L  T  E

2101        AAATTTGTACAGATATGGAAAAGGAAGGAAAAATTTCAAGAATTGGGCCTGAAAATCCAT
(pol)         I  C  T  D  M  E  K  E  G  K  I  S  R  I  G  P  E  N  P  Y

2161        ACAATACTCCAATATTTGCCATAAAGAAAAAAGACAGTACTAAATGGAGAAAATTAGTAG
(pol)         N  T  P  I  F  A  I  K  K  K  D  S  T  K  W  R  K  L  V  D

2221        ATTTCAGAGAACTTAATAAGAGAACTCAAGATTTCTGGGAAGTACAATTAGGAATACCAC
(pol)         F  R  E  L  N  K  R  T  Q  D  F  W  E  V  Q  L  G  I  P  H

2281        ATCCTGCAGGGCTGAAAAAGAAAAAATCAGTAACAGTACTGGATGTGGGTGATGCATATT
(pol)         P  A  G  L  K  K  K  K  S  V  T  V  L  D  V  G  D  A  Y  F

2341        TTTCAGTTCCCTTATGTGAAGACTTTAGGAAATATACCGCATTTACCATACCTAGTATAA
(pol)         S  V  P  L  C  E  D  F  R  K  Y  T  A  F  T  I  P  S  I  N


2401        ACAATGAGACACCAGGGATTAGATATCAGTACAATGTGCTTCCACAGGGATGGAAAGGAT
(pol)         N  E  T  P  G  I  R  Y  Q  Y  N  V  L  P  Q  G  W  K  G  S

2461        CACCGGCAATATTCCAAAGTAGCATGACAAAAATCTTAGAGCCCTTTAGAAAACAAAATC
(pol)         P  A  I  F  Q  S  S  M  T  K  I  L  E  P  F  R  K  Q  N  P

2521        CAGAGATGGTTATCTATCAATACATGGATGATTTGTATGTAGGATCTGACTTAGAAATAG
(pol)         E  M  V  I  Y  Q  Y  M  D  D  L  Y  V  G  S  D  L  E  I  G

2581        GGCAACATAGAACAAAAATAGAGGAATTAAGAGAACATCTATTGAGGTGGGGATTTACCA
(pol)         Q  H  R  T  K  I  E  E  L  R  E  H  L  L  R  W  G  F  T  T

2641        CACCAGATAAAAAACATCAGAAGGAACCTCCATTTCTTTGGATGGGTTATGAACTCCATC
(pol)         P  D  K  K  H  Q  K  E  P  P  F  L  W  M  G  Y  E  L  H  P

2701        CTGATAAATGGACAGTACAGCCTATAATACTGCCAGACAAAAGAAAGGTTGGGACTGTCA
(pol)         D  K  W  T  V  Q  P  I  I  L  P  D  K  R  K  V  G  T  V  N
```

```
2761        ATGATATACAGAAGTTAGTAGGGAAATTAAACTGGGCAAGCCAGATTTATCCAGGAATTA
(pol)         D  I  Q  K  L  V  G  K  L  N  W  A  S  Q  I  Y  P  G  I  K

2821        AAGTAAAGCAATTATGTAAACTCCTTAGGGGAACCAAAGCACTAACAGAAGTAATATCAC
(pol)         V  K  Q  L  C  K  L  L  R  G  T  K  A  L  T  E  V  I  S  L

2881        TAACAGCAGAAGCAGAATTAGAACTGGCAGAAAACAGGGAAATTCTAAAAGAACCAGTAC
(pol)         T  A  E  A  E  L  E  L  A  E  N  R  E  I  L  K  E  P  V  H

2941        ATGGAGTGTATTATGACCCATCAAAAGACTTAATAGCAGAAATACAGAAACAAGGGAATG
(pol)      G  V  Y  Y  D  P  S  K  D  L  I  A  E  I  Q  K  Q  G  N  G

3001        GCCAATGGACATACCAAATTTATCAAGAACCATTTAAAAATCTGAAAACAGGAAAGTATG
(pol)         Q  W  T  Y  Q  I  Y  Q  E  P  F  K  N  L  K  T  G  K  Y  A

3061        CAAGAACGAGGGGTGCCCACACTAATGATGTAAAACAATTAGCAGAGGCAGTGCAAAAAA
(pol)         R  T  R  G  A  H  T  N  D  V  K  Q  L  A  E  A  V  Q  K  I

3121        TAGCCACAGAAGGCATAGTAATATGGGGAAAGACTCCTAAATTTAGACTGCCCATACAAA
(pol)         A  T  E  G  I  V  I  W  G  K  T  P  K  F  R  L  P  I  Q  K

3181        AGGAAACATGGAAAACATGGTGGATAGAGTATTGGCAAGCCACCTGGATTCCTGAGTGGG
(pol)         E  T  W  K  T  W  W  I  E  Y  W  Q  A  T  W  I  P  E  W  E

3241        AATTTGTCAATACCCCTCCTTTAGTAAAATTATGGTACCAATTAGAGAAGGAACCCATAA
(pol)         F  V  N  T  P  P  L  V  K  L  W  Y  Q  L  E  K  E  P  I  M

3301        TGGGAGCAGAAACTTTCTATGTAGATGGGGCAGCTAATAGAGAGACTAAAGTAGGAAAAG
(pol)         G  A  E  T  F  Y  V  D  G  A  A  N  R  E  T  K  V  G  K  A

3361        CAGGATATGTTACTGACAGAGGAAGACAGAAAGTTGTCCCTTTAACTGACACAACAAATC
(pol)         G  Y  V  T  D  R  G  R  Q  K  V  V  P  L  T  D  T  T  N  Q

3421        AGAAGACTGAGTTACAAGCAGTTAATCTAGCTTTGCAGGATTCGGGATTAGAAGTAAACA
(pol)         K  T  E  L  Q  A  V  N  L  A  L  Q  D  S  G  L  E  V  N  I

3481        TAGTAACAGATTCACAATATGTATTAGGAATCATTCAAGCACAACCAGATAAAAGTGAAT
(pol)         V  T  D  S  Q  Y  V  L  G  I  I  Q  A  Q  P  D  K  S  E  S

3541        CAGAGTTAGTCAGTCAAATAATAGAGCAGCTAATAAAAAAGGAAAAGGTTTACCTGGCAT
(pol)         E  L  V  S  Q  I  I  E  Q  L  I  K  K  E  K  V  Y  L  A  W

3601        GGGTACCAGCACACAAAGGAATTGGAGGAAATGAACAAGTAGATAAATTAGTCAGTCAGG
(pol)         V  P  A  H  K  G  I  G  G  N  E  Q  V  D  K  L  V  S  Q  G
```

```
3661        GAATCAGGAAAGTACTATTTTTGGATGGAATAGATAAGGCTCAAGAAGAACATGAGAAAT
(pol)         I  R  K  V  L  F  L  D  G  I  D  K  A  Q  E  E  H  E  K  Y

3721        ATCACAACAATTGGAGAGCAATGGCTAGTGATTTTAACCTACCACCTGTGGTAGCAAAAG
(pol)         H  N  N  W  R  A  M  A  S  D  F  N  L  P  P  V  V  A  K  E

3781        AAATAGTAGCTAGCTGTGATAAATGTCAGCTAAAAGGAGAAGCCATGCATGGACAAGTAG
(pol)         I  V  A  S  C  D  K  C  Q  L  K  G  E  A  M  H  G  Q  V  D

3841        ACTGTAGTCCAGGAATATGGCAATTAGATTGTACACATTTAGAAGGAAAAGTTATCATAG
(pol)         C  S  P  G  I  W  Q  L  D  C  T  H  L  E  G  K  V  I  I  V

3901        TAGCAGTTCATGTAGCCAGTGGCTATATAGAAGCAGAAGTTATTCCAGCAGAAACAGGGC
(pol)         A  V  H  V  A  S  G  Y  I  E  A  E  V  I  P  A  E  T  G  Q

3961        AGGAAACAGCATACTTTCTCTTAAAATTAGCAGGAAGATGGCCAGTAAAAGTAGTACATA
(pol)         E  T  A  Y  F  L  L  K  L  A  G  R  W  P  V  K  V  V  H  T

4021        CAGACAATGGCAGCAATTTCACCAGTGCTGCAGTTAAGGCCGCCTGCTGGTGGGCAGGTA
(pol)         D  N  G  S  N  F  T  S  A  A  V  K  A  A  C  W  W  A  G  I

4081        TCAAACAGGAATTTGGAATTCCCTACAATCCCCAAAGTCAAGGAGTAGTAGAATCTATGA
(pol)         K  Q  E  F  G  I  P  Y  N  P  Q  S  Q  G  V  V  E  S  M  N

4141        ATAAAGAATTAAAGAAAATACAAGGACAGGTTAGAGATCAAGCTGAACATCTTAAGACAG
(pol)         K  E  L  K  K  I  Q  G  Q  V  R  D  Q  A  E  H  L  K  T  A

4201        CAGTACAAATGGCAGTATTCATCCACAATTTTAAAAGAAAAGGGGGGATTGGGGGATACA
(pol)         V  Q  M  A  V  F  I  H  N  F  K  R  K  G  G  I  G  G  Y  S

4261        GTGCAGGGGAAAGAATAGTAGACATTAGAGCAACAGACATACAAACTAAAGAATTACAAA
(pol)         A  G  E  R  I  V  D  I  R  A  T  D  I  Q  T  K  E  L  Q  K

4321        AGCAAATCACAAAAATTCAAAATTTTCGGGTTTATTACAGGGACAGCAGAGATCCAATTT
(pol)         Q  I  T  K  I  Q  N  F  R  V  Y  Y  R  D  S  R  D  P  I  W

4381        GGAAAGGACCAGCAAAACTTCTCTGGAAAGGTGAAGGGGCAGGAGAAATACAAGACAATA
(pol)         K  G  P  A  K  L  L  W  K  G  E  G  A  G  E  I  Q  D  N  S

4441        GTGACATTAAGGTACTACCAAGAAGAAAAGTGCAAATCATTAGGGATTATGGAAAACAGA
(Vif start)                                                 M  E  N  R
(Pol)         D  I  K  V  L  P  R  R  K  V  Q  I  I  R  D  Y  G  K  Q  M

4501        TGGCAGGTGATGATTGTGTGGCAAGTAGACAGGATGAGGATTAGCACATGGAAAAGTTTA
(Vif)       W  Q  V  M  I  V  W  Q  V  D  R  M  R  I  S  T  W  K  S  L
(Pol end)     A  G  D  D  C  V  A  S  R  Q  D  E  D  *
```

```
4561        GTAAAACACCATATGTATGTTTCAAAAAAGGCTAAAGGATGGTTTTATAGACATCACTAT
(Vif)        V  K  H  H  M  Y  V  S  K  K  A  K  G  W  F  Y  R  H  H  Y

4621        GACAGCCCCCACCCAAAAATAAGCTCAGAAGTACACATTCCACTAGGAGAAGAAAGACTG
(Vif)        D  S  P  H  P  K  I  S  S  E  V  H  I  P  L  G  E  E  R  L

4681        ATAGTAAAAACATATTGGGGTCTGCATACAGGAGAAAGAGAATGGCATCTGGGTCAGGGA
(Vif)        I  V  K  T  Y  W  G  L  H  T  G  E  R  E  W  H  L  G  Q  G


4741        GTCTCCATAGAATGGAGGAAAAAGGAATATAGCACACAAGTTGACCCTGGCCTGGCAGAC
(Vif)        V  S  I  E  W  R  K  K  E  Y  S  T  Q  V  D  P  G  L  A  D

4801        CAACTCATTCATATATATTATTTTGATTGTTTTTCAGACTCTGCTATCAGAAAAGCCTTA
(Vif)        Q  L  I  H  I  Y  Y  F  D  C  F  S  D  S  A  I  R  K  A  L

4861        TTAGGACATATGGTTAGACCTAGGTGTGAATACCAAGCAGGACATAACAAGGTTGGATCC
(Vif)        L  G  H  M  V  R  P  R  C  E  Y  Q  A  G  H  N  K  V  G  S

4921        TTACAGTATTTGGCACGAACAGCATTATTACCACCAAAAAAGACAAAGCCACCTTTGCCT
(Vif)        L  Q  Y  L  A  R  T  A  L  L  P  P  K  K  T  K  P  P  L  P

4981        AGTGTTAGGAAGCTATCAGAAGATAGATGGAACAAGCCCCAGAAGACCAAGGGCCACAGA
(Vif)        S  V  R  K  L  S  E  D  R  W  N  K  P  Q  K  T  K  G  H  R
(Vpr start)                              M  E  Q  A  P  E  D  Q  G  P  Q  R

5041        GGGAGCCATACAACGAATGGACATTAGAACTTTTGGAGGAGCTTAAGAGTGAAGCTGTTA
(Vif end)    G  S  H  T  T  N  G  H  *
(Vpr)           E  P  Y  N  E  W  T  L  E  L  L  E  E  L  K  S  E  A  V  R

5101        GACACTTTCCTAGATTATGGCTCCATAGCTTAGGACAACATATCTATGAAACTTATGGGG
(Vpr)           H  F  P  R  L  W  L  H  S  L  G  Q  H  I  Y  E  T  Y  G  D

5161        ATTCCTGGGCAGGAGTTGAAGCTATAATAAGAATTCTGCAACAATTACTGTTTATTCATT
(Vpr)           S  W  A  G  V  E  A  I  I  R  I  L  Q  Q  L  L  F  I  H  F

5221        TCAGAATTGGGTGTCAACATACCAGAAGAGGTATTACTCGGCAGAGAAGAGCAAGAAATG
(Tatx1 start)                                                        M
(Vpr)           R  I  G  C  Q  H  T  R  R  G  I  T  R  Q  R  R  A  R  N  G

5281        GATCCAGTAGATCCTAGCCTAGAGCCCTGGAACCATCCAGGAAGTCAGCCTAAGACTGCT
(Tatx1)      D  P  V  D  P  S  L  E  P  W  N  H  P  G  S  Q  P  K  T  A
(Vpr end)    S  S  R  S  *
```

```
5341        TATAACAAGTGTCATTGTAAAAAGTGTTGCTTTCATTGTCAAGTTTGCTTCATAACGAAA
(Tatx1)      Y  N  K  C  H  C  K  K  C  C  F  H  C  Q  V  C  F  I  T  K

5401        GGCTTAGGCATCTCCTATGGCAGGAAGAAGCGGAGACAGCGACGAAAACCTCCTCACGGC
(Tatx1)      G  L  G  I  S  Y  G  R  K  K  R  R  Q  R  R  K  P  P  H  G
(Revx1 start)                  M  A  G  R  S  G  D  S  D  E  N  L  L  T  A

5461        GGTCAGGCTCATCAAGTTCCTATACCAGGGCAGTAAGTAGTTCATGTAATGCAACCTTTA
(Tatx1end)  G  Q  A  H  Q  V  P  I  P  G  Q  *                 M  Q  P  L
(Vpu start)
(Revx1 end)  V  R  L  I  K  F  L  Y  Q  G  S  K  *

5521        GTGATAATAGCAATAGCAGCATTAGTAGTAGCACTAATAATAGCAATAGTTGTGTGGACC
(Vpu)        V  I  I  A  I  A  A  L  V  V  A  L  I  I  A  I  V  V  W  T

5581        ATAGTATTCATAGAATATAGGAGAATAAAAAAGCAAGGAAAAATAGACTGTTTAATTGAT
(Vpu)        I  V  F  I  E  Y  R  R  I  K  K  Q  G  K  I  D  C  L  I  D

5641        AGAATAATAGAAAGAGCAGAAGACAGTGGCAATGAGAGCGAGGGGGATAGAGAGGAATCG
(Vpu)        R  I  I  E  R  A  E  D  S  G  N  E  S  E  G  D  R  E  E  S
(Env start)                                 M  R  A  R  G  I  E  R  N  R

5701        ACAAAACTTGTGGACATGGGGCATCATGCTCCTTGGGATGTTGATGATCTGTAATGCTGC
(Vpu end)    T  K  L  V  D  M  G  H  H  A  P  W  D  V  D  D  L  *
(Env)         Q  N  L  W  T  W  G  I  M  L  L  G  M  L  M  I  C  N  A  A

5761        AGAAAATTTGTGGGTCACAGTTTATTATGGGGTGCCTGTATGGAAGGAAGCAACCACTAC
(Env)         E  N  L  W  V  T  V  Y  Y  G  V  P  V  W  K  E  A  T  T  T

5821        TCTATTTTGTGCATCAGATGCTAAATCCTATGAAACAGAGGCACATAATATCTGGGCTAC
(Env)         L  F  C  A  S  D  A  K  S  Y  E  T  E  A  H  N  I  W  A  T

5881        ACATGCCTGTGTACCCACGGACCCCAGCCCACAAGAAATAGAACTGGAAAATGTGACCGA
(Env)         H  A  C  V  P  T  D  P  S  P  Q  E  I  E  L  E  N  V  T  E

5941        AAACTTTAATATGTGGAAAAATAACATGGTAGACCAGATGCATGAGGATATAATCAGTTT
(Env)         N  F  N  M  W  K  N  N  M  V  D  Q  M  H  E  D  I  I  S  L

6001        ATGGGATCAAAGCCTAAAACCATGTGTAAAATTAACCCCACTCTGTGTCACTTTAAACTG
(Env)         W  D  Q  S  L  K  P  C  V  K  L  T  P  L  C  V  T  L  N  C

6061        CAATAATAATGTTACCTTAAACAGCACTGGGGCCATCTGCAACAAGACTACGGGCAAAGC
(Env)         N  N  N  V  T  L  N  S  T  G  A  I  C  N  K  T  T  G  K  A

6121        CACTGTGGAGTCAGAACTGGAGGTAAAAAACTGCTCTTTCAATATAACTACAGTAGTAAG
(Env)         T  V  E  S  E  L  E  V  K  N  C  S  F  N  I  T  T  V  V  R
```

```
6181        AGATAAGAGAATGCAAGTACGTGCGCTTTTTTATAGACCTGATATAGTATCAATAGACAA
(Env)         D  K  R  M  Q  V  R  A  L  F  Y  R  P  D  I  V  S  I  D  N

6241        TGATAATACCAGTTATAGGTTAATAAATTGTAATACCTCAGCCATTACACAGGCTTGTCC
(Env)         D  N  T  S  Y  R  L  I  N  C  N  T  S  A  I  T  Q  A  C  P

6301        AAAGGTATCCTTTCAACCAATTCCAATACATTATTGTGCCCCAGCTGGTTTTGCAATTCT
(Env)         K  V  S  F  Q  P  I  P  I  H  Y  C  A  P  A  G  F  A  I  L

6361        TAAGTGTAGAGATAAGAAGTTCAATGGAACAGGCCCATGCACAAATGTCAGCACAGTACA
(Env)         K  C  R  D  K  K  F  N  G  T  G  P  C  T  N  V  S  T  V  Q

6421        ATGTACACATGGAATTAAGCCAGTGGTGTCAACTCAACTGCTGTTGAATGGCAGTCTAGC
(Env)         C  T  H  G  I  K  P  V  V  S  T  Q  L  L  L  N  G  S  L  A

6481        AGAAGAAGAGATCATAATTAGATCTGAAAATCTCACAAACAATGCTAAAAACATAATAGT
(Env)         E  E  E  I  I  I  R  S  E  N  L  T  N  N  A  K  N  I  I  V

6541        ACAGTTTAATGCATCTGTAGAAATTAATTGTACAAGGCCCTACAAATATACAATACAAAA
(Env)         Q  F  N  A  S  V  E  I  N  C  T  R  P  Y  K  Y  T  I  Q  K

6601        AACATCAATAGGACAAGGGCAAGCATTACATACAAGCAAGAGGATAATAGGAGACATAAG
(Env)         T  S  I  G  Q  G  Q  A  L  H  T  S  K  R  I  I  G  D  I  R

6661        ACAAGCACATTGTAACATTAGTGGAGAAAAATGGTATGAAACTCTACAACAGGTAGCTAT
(Env)         Q  A  H  C  N  I  S  G  E  K  W  Y  E  T  L  Q  Q  V  A  I

6721        AAAATTAGGAGACCTTCTTAACAAAACAACAATAACTTTTCGACCACCCTCAGGAGGGGA
(Env)         K  L  G  D  L  L  N  K  T  T  I  T  F  R  P  P  S  G  G  D

6781        CCCAGAAATTACAACACACAGTTTTAATTGTGGAGGGGAATTTTTCTACTGTAATACATC
(Env)         P  E  I  T  T  H  S  F  N  C  G  G  E  F  F  Y  C  N  T  S

6841        AAGGCTGTTTAATAATACATGGAATGGTACAACATGGTCAAATAAGACAGACACCAATGG
(Env)         R  L  F  N  N  T  W  N  G  T  T  W  S  N  K  T  D  T  N  G

6901        GACAGTCACACTCCCATGCAGAATAAAACAAATTATAAACATGTGGCAGGAAGTAGGAAA
(Env)         T  V  T  L  P  C  R  I  K  Q  I  I  N  M  W  Q  E  V  G  K

6961        AGCAATGTATGCCCCCCCCATAGAAGGACTACTTAGATGTTCATCAAATATTACAGGGTA
(Env)         A  M  Y  A  P  P  I  E  G  L  L  R  C  S  S  N  I  T  G  Y

7021        TATATTGACAAGAGATGGTGGTTATACCAGTTCTGGCAATGCGACCTTCAGACCTGGCGG
(Env)         I  L  T  R  D  G  G  Y  T  S  S  G  N  A  T  F  R  P  G  G
```

```
7861        AGGCAGAGACAAATCAATTCGATTGGTGAACGGATTCTCCGCACTTATCTGGGACGATCT
(Revx2)      R  Q  R  Q  I  N  S  I  G  E  R  I  L  R  T  Y  L  G  R  S
(Env)         G  R  D  K  S  I  R  L  V  N  G  F  S  A  L  I  W  D  D  L
(Tatx2 end)  A  E  T  N  Q  F  D  W  *

7921        GCGGAACCTGTGCCTCTTCAGCTACCACCGCTTCAGAGACTTACTCTTGTTTGCAGCGAG
(Revx2)      A  E  P  V  P  L  Q  L  P  P  L  Q  R  L  T  L  V  C  S  E
(Env)         R  N  L  C  L  F  S  Y  H  R  F  R  D  L  L  L  F  A  A  R

7981        GATTGTGGAACTTCTGGGACGCAGGGGGTGGGAAGCCCTCAAGTATCTGTGGAATCTCCT
(Revx2)      D  C  G  T  S  G  T  Q  G  V  G  S  P  Q  V  S  V  E  S  P
(Env)         I  V  E  L  L  G  R  R  G  W  E  A  L  K  Y  L  W  N  L  L

8041        GCAGTATTGGAGTCAGGAACTCAAGAATAGTGCTATTTACTTGCTTGCTACCATCGCAAT
(Revx2 end)A  V  L  E  S  G  T  Q  E  *
(Env)         Q  Y  W  S  Q  E  L  K  N  S  A  I  Y  L  L  A  T  I  A  I

8101        CGTACCAGCAGAGGGGACAGACAGGGTTATACAAGTTGTACGAAGAGCTTGCAGAGCTAT
(Env)         V  P  A  E  G  T  D  R  V  I  Q  V  V  R  R  A  C  R  A  I

8161        TCTTACCATACCCACAAGAATACGACAGGGCTTGGAAAGGCTTTTGCTATAAAATGGGTG
(Env end)    L  T  I  P  T  R  I  R  Q  G  L  E  R  L  L  L  *
(Nef start)                                                    M  G  G

8221        GCAAATGGTCAAAAAGTAGTATAGTTGGATGGTCTGCTATAAGGGAAAGAATAAGAAGAA
(Nef)         K  W  S  K  S  S  I  V  G  W  S  A  I  R  E  R  I  R  R  T

8281        CTGATCCAGCAGCAGATGGGGTGGGAGCAGTATCTCGAGACCTGGAAAAACATGGGGCAA
(Nef)         D  P  A  A  D  G  V  G  A  V  S  R  D  L  E  K  H  G  A  I

8341        TCACAAGTAGCAATACAGCAAGTACTAATGCTGACTGTGCCTGGCTAGAAGCACAAGAAG
(Nef)         T  S  S  N  T  A  S  T  N  A  D  C  A  W  L  E  A  Q  E  E

8401        AGAGTGAGGAGGTGGGCTTTCCAGTCAGACCTCAGGTACCTTTAAGACCAACGACTTACA
(Nef)         S  E  E  V  G  F  P  V  R  P  Q  V  P  L  R  P  T  T  Y  K

8461        AAGCAGCTGTAGATCTTAGCCACTTTTTAAAAGAAAAGGGGGGACTGGAAGGGCTTATTT
(Nef)         A  A  V  D  L  S  H  F  L  K  E  K  G  G  L  E  G  L  I  W

8521        GGTCCAAAAAGAGACAAGAGATCCTTGATCTTTGGGTCTACAACACACAAGGCTACTTCC
(Nef)         S  K  K  R  Q  E  I  L  D  L  W  V  Y  N  T  Q  G  Y  F  P

8581        CCGATTGGCAGAACTACACACCAGGGCCAGGGATCAGATATCCACTTACCTTTGGATGGT
(Nef)         D  W  Q  N  Y  T  P  G  P  G  I  R  Y  P  L  T  F  G  W  C
```

```
7081        AGGAGATATGAGGGACAATTGGAGAAGAGAATTATATACATACAAAGTAGTACAAATTGG
(Env)         G  D  M  R  D  N  W  R  R  E  L  Y  T  Y  K  V  V  Q  I  G

7141        ACCAATAGGAGTAGTGCCCACCAGGGCAAAGAGAAGAGTGGTGGAAAGGGAAAAAAGAGG
(Env)         P  I  G  V  V  P  T  R  A  K  R  R  V  V  E  R  E  K  R  G

7201        GGTTTTCTTGGGAGCAGCAGGAAGCACGATGGGCGCAGCGTCATTGTCGCTGCCGGTACA
(Env)         V  F  L  G  A  A  G  S  T  M  G  A  A  S  L  S  L  P  V  Q

7261        GGCCAGACAGGTATTGTCTGGTACAGTGCAACAGCAAAGCAATTTGCTCAGGGCTATATC
(Env)         A  R  Q  V  L  S  G  T  V  Q  Q  Q  S  N  L  L  R  A  I  S

7321        GGCGCAACAGCATCTGTTGCAACTCACGGTCTGGGGCATTAAACAGCTCCAGGCAAGAGT
(Env)         A  Q  Q  H  L  L  Q  L  T  V  W  G  I  K  Q  L  Q  A  R  V

7381        CCTGGCTGTGGAAAGATACCTTAAGGATCAACGGCTCCTGGGACTTTGGGGTTGCTCTGG
(Env)         L  A  V  E  R  Y  L  K  D  Q  R  L  L  G  L  W  G  C  S  G

7441        AAAACACATTTGCACCACTACTGTGCCCTGGAACTCTAGTTGGAGTAATAGAACTCAAGA
(Env)         K  H  I  C  T  T  T  V  P  W  N  S  S  W  S  N  R  T  Q  D

7501        TGAGATTTGGCATAACATGTCCTGGATGCAGTGGGAAAGAGAAATTGACAATTACACAGG
(Env)         E  I  W  H  N  M  S  W  M  Q  W  E  R  E  I  D  N  Y  T  G

7561        ACTATTATACACCTCAATTGAAAGTTCGCAGGTTCAGCAAGAAAAGAATGAACAAGAATT
(Env)         L  L  Y  T  S  I  E  S  S  Q  V  Q  Q  E  K  N  E  Q  E  L

7621        ATTGGAATTGGACAAGTGGGCAAGTCTGTGGAATTGGTTTAACATCACAAACTGGCTGTG
(Env)         L  E  L  D  K  W  A  S  L  W  N  W  F  N  I  T  N  W  L  W

7681        GTATACAAAAATATTCAGAATCATATGGGGAGGCTTACCAGGTTTTAGAATGGTTTTTGC
(Env)         Y  T  K  I  F  R  I  I  W  G  G  L  P  G  F  R  M  V  F  A

7741        TGTGCTTTCTGTGGTACATAGAGTTAGGCAGGGATACTCACCTCTGTCATTTCAGACCCT
(Revx2 start)                                                 S  D  P
(Env)         V  L  S  V  V  H  R  V  R  Q  G  Y  S  P  L  S  F  Q  T  L
(Tatx2 start)                                                    P  S

7801        CCTCCCAGCCCCGAGGGGACCCGACAGGCCCGAAGGAACAGAAGAAGAAGGTGGAGAGCG
(Revx2)     P  P  S  P  E  G  T  R  Q  A  R  R  N  R  R  R  R  W  R  A
(Env)         L  P  A  P  R  G  P  D  R  P  E  G  T  E  E  E  G  G  E  R
(Tatx2)       S  Q  P  R  G  D  P  T  G  P  K  E  Q  K  K  K  V  E  S  E
```

8641 GTTTCGAGCTATTACCAGTTGATCCACAGGAGGAAGAAGAGGCCACTGAGGGAGAGACCA

(Nef) F E L L P V D P Q E E E E A T E G E T N

8701 ACTGCTTGTTACACCCTATCAACCAGCATGGAATGGAGGACCCGGAGAGACAAGTGTTCA

(Nef) C L L H P I N Q H G M E D P E R Q V F K

8761 AGTGGAGATTTAACAGCAGACAAGCATTTGAGCACAAGGCCCGCCAGTTACATCCGGAGT

(Nef) W R F N S R Q A F E H K A R Q L H P E Y

8821 ACTACAAAGACTGCTGACACCGAGTTTTCTACAGGGGACTTTCCGCTGGGGACTTTCCAG

(Nef end) Y K D C *

B2: pR214 Full-length sequence (nucleotide and amino acid)

```
181         AAATTTTTGACTAGCGGAGGCTAGAGGAGAGAGATGGGTGCGCGAGCGTCGGTTTTAAGC
(Gag start)                                     M  G  A  R  A  S  V  L  S

241         GGGGGAGAATTAGATAGGTGGGAAAAAATTCGTTTAAGGCCGGGAGGGAAGAAAAAATAT
(Gag)       G  G  E  L  D  R  W  E  K  I  R  L  R  P  G  G  K  K  K  Y

301         AAACTTAAACATATACTATGGGCAAGCAGGGAGCTGGAACGATTTGCACTTAATCCTAGC
(Gag)        K  L  K  H  I  L  W  A  S  R  E  L  E  R  F  A  L  N  P  S

361         CTTCTAGAGTACAGCGAAGGATGTAAACAAATATTAGGACAGCTACAACCATCTCTTCAG
(Gag)        L  L  E  Y  S  E  G  C  K  Q  I  L  G  Q  L  Q  P  S  L  Q

421         ACAGGATCAGAAGAACTTAAATCATTATATATTACAGTAGTAACCCTCTATTGTGTACAA
(Gag)        T  G  S  E  E  L  K  S  L  Y  I  T  V  V  T  L  Y  C  V  Q

481         GAAAGGATAGAGGTAAAGGACACCAAGGAAGCTTTCAGAAAGATGGAGGAAGAACAAAAC
(Gag)        E  R  I  E  V  K  D  T  K  E  A  F  R  K  M  E  E  E  Q  N

541         AAATGTAAGAAAAAGAAGGCACAGCAAGCAGCGGCTGACACAGGGAACAGCAGCCAGGTC
(Gag)        K  C  K  K  K  K  A  Q  Q  A  A  A  D  T  G  N  S  S  Q  V

601         AGCCAAAATTATCTATATTGCAGAACTACAGGGCAAATGGTACATGGGGCCATATCACCT
(Gag)        S  Q  N  Y  L  Y  C  R  T  T  G  Q  M  V  H  G  A  I  S  P

661         AGAACTTTGAATGCATGGGTAAAAGTAATAGAGGAAAAGGCCTTCAGCCCAGAAGGAATA
(Gag)        R  T  L  N  A  W  V  K  V  I  E  E  K  A  F  S  P  E  G  I

721         CCCATGTTTTCAGCATATTCAGAAGGAGCCACCCCACAAGATTTAAACACCATGCTAAAC
(Gag)        P  M  F  S  A  Y  S  E  G  A  T  P  Q  D  L  N  T  M  L  N

781         ACAGTGGGGGGACATCAAGCAGCCATGCAAATGTACAAGGAGACCATCAATGAGGAAGCT
(Gag)        T  V  G  G  H  Q  A  A  M  Q  M  Y  K  E  T  I  N  E  E  A

841         GCAGAATGGGATAGGCTACATCCAGTGCATGCAGGGCCTATTGCACCAGGCCAGATCAGA
(Gag)        A  E  W  D  R  L  H  P  V  H  A  G  P  I  A  P  G  Q  I  R

901         GAACCAAGGGGAAGTGATATACCAGGAACTACTAGTACCCTTCAGGAACAAATAGGATGG
(Gag)        E  P  R  G  S  D  I  P  G  T  T  S  T  L  Q  E  Q  I  G  W

961         ATTACAAGCAACCCACCTATCCCAGTCGGAGAAATCTATAAAAGATGGATTATCCTGGGA
(Gag)        I  T  S  N  P  P  I  P  V  G  E  I  Y  K  R  W  I  I  L  G

1021        TTCAATAAAATACATAGAATGTATAGCCCTGTCAGCATTTTGGACATAAGACAGGGACCA
(Gag)        F  N  K  I  H  R  M  Y  S  P  V  S  I  L  D  I  R  Q  G  P

1081        AAGGAACCTTTTAGAGATTATGTATACCGGTTCTATAAAACTCAAAGAGCCGAGCAAGCT
(Gag)        K  E  P  F  R  D  Y  V  Y  R  F  Y  K  T  Q  R  A  E  Q  A

1141        TCACAGGATGGAAAAAACTGGATGCCAGAAACCTTGTTGGTCCAAAATGCAAACCCAGAT
(Gag)        S  Q  D  G  K  N  W  M  P  E  T  L  L  V  Q  N  A  N  P  D

1201        TGTAAAACCATCTTACAAGCATCAGGACCACAGGCTACACTAGAAGAAATGATGACAGCG
(Gag)        C  K  T  I  L  Q  A  S  G  P  Q  A  T  L  E  E  M  M  T  A

1261        TGTCAGGGAGTAGGAGGGCCCAGCCATAAAGCAAGAGTTTTGGCTGAGGCAATGAGCCAA
(Gag)        C  Q  G  V  G  G  P  S  H  K  A  R  V  L  A  E  A  M  S  Q

1321        GCAACAAATAGCGCAACGATCTACTGCCAGAGAGGCAATTTTAAGGGCCAAAGAAAAATT
(Gag)        A  T  N  S  A  T  I  Y  C  Q  R  G  N  F  K  G  Q  R  K  I

1381        GTAAAGTGTTTCAACTGTGGCAAGAAGGCACATAGCAAAAAATTGCAGGGCCCCAAGGAA
(Gag)        V  K  C  F  N  C  G  K  K  A  H  S  K  K  L  Q  G  P  K  E
```

```
1441        AAGGGCTGTTGGAAATGTGGAAGGGAAGGACACCAAATGAAAGATTGCACTGAAAGACAG
(Gag)        K  G  C  W  K  C  G  R  E  G  H  Q  M  K  D  C  T  E  R  Q

1501        GAAAATTTTTTAGAGAAAATTTTGCCTTCCCACAAGGGACGCCCGGGGAACTTTCTTCAG
(Gag)        E  N  F  L  E  K  I  L  P  S  H  K  G  R  P  G  N  F  L  Q
(Pol start)     F  F  R  E  N  F  A  F  P  Q  G  T  P  G  E  L  S  S  E

1561        AGCAGACCAGGGCCAACAGCCCCACCACTAGAGAGCTTCGGGTTTGGGGAGGAGATAACC
(Gag)        S  R  P  G  P  T  A  P  P  L  E  S  F  G  F  G  E  E  I  T
(Pol)         Q  T  R  A  N  S  P  T  T  R  E  L  R  V  W  G  G  D  N  P

1621        CCCTCTCAGAAACAGGAACAGAAAGACAAGGAACTGTATCCTTTAACCTCCCTCAAATCA
(Gag)        P  S  Q  K  Q  E  Q  K  D  K  E  L  Y  P  L  T  S  L  K  S
(Pol)          L  S  E  T  G  T  E  R  Q  G  T  V  S  F  N  L  P  Q  I  T

1681        CTCTTTGGGAGCGACCCCTTGTCACAATAGAGATACGGGGACAGCTCAAGGAAGCTCTAT
(Gag end)    L  F  G  S  D  P  L  S  Q  *
(Pol)          L  W  E  R  P  L  V  T  I  E  I  R  G  Q  L  K  E  A  L  L

1741        TATATACAGGAGCAGATGATACAGTATTTGAAGAAATTAATTTGCCAGGAAAATGGAAAC
(Pol)          Y  T  G  A  D  D  T  V  F  E  E  I  N  L  P  G  K  W  K  P

1801        CAAAAACGATAGGGGGAATTGGAGGTTTTATCAAAGTCAGACAGTATGATCAAATACCCC
(Pol)          K  T  I  G  G  I  G  G  F  I  K  V  R  Q  Y  D  Q  I  P  L

1861        TACAAATCTGTGGGCATAAAGCTAAAGGTACAGTACTCGTTGGGGCTACGCCTGTCAACA
(Pol)          Q  I  C  G  H  K  A  K  G  T  V  L  V  G  A  T  P  V  N  I

1921        TAATTGGAAGAAATTTGCTGACTCAGCTTGGTCGCACTTTAAATTCCCCAATCTCTGAAA
(Pol)          I  G  R  N  L  L  T  Q  L  G  R  T  L  N  S  P  I  S  E  T

1981        CTGTACCAGGAAAGTTAAAGCCAGGAATGGATGGCCCAAAAGTTTACCAATGGCCATTGA
(Pol)          V  P  G  K  L  K  P  G  M  D  G  P  K  V  Y  Q  W  P  L  T

2041        CAGAAGAAAAAATAAAAGCATTAACAGAAATTTGTACAGATATGGAAAAGGAAGGAAAAA
(Pol)          E  E  K  I  K  A  L  T  E  I  C  T  D  M  E  K  E  G  K  I

2101        TTTCAAGAATTGGGCCTGAAAATCCATACAATTACTTCCAATTTGCCATAAAGAAAAAAG
(Pol)          S  R  I  G  P  E  N  P  Y  N  Y  F  Q  F  A  I  K  K  K  D

2161        ACAGTACTAAATGGAGAAAATTAGTAGATTTCAGAGAACTTAATAAGAGAACTCAAGATT
(Pol)          S  T  K  W  R  K  L  V  D  F  R  E  L  N  K  R  T  Q  D  F

2221        TCTGGGAAGTACAATTAGGAATACCACATCCTGCAGGGCTGAAAAAGAAAAAATCAGTAA
(Pol)          W  E  V  Q  L  G  I  P  H  P  A  G  L  K  K  K  K  S  V  T

2281        CAGTACTGGATGTGGGTGATGCATATTTCTTCGTTCCCTTATGTGGAGCTTTTAGAAAAT
(Pol)          V  L  D  V  G  D  A  Y  F  F  V  P  L  C  G  A  F  R  K  Y

2341        ATACCGCATTTACCATACCTTCAATAACAAATGAGACACCAGGGATTAGATATCAGTACA
(Pol)          T  A  F  T  I  P  S  I  T  N  E  T  P  G  I  R  Y  Q  Y  N

2401        ATGTGCTTCCACAGGGATGGAAAGGATCACCGGCAATATTCCAAAGTAGCATGTCAAAAA
(Pol)          V  L  P  Q  G  W  K  G  S  P  A  I  F  Q  S  S  M  S  K  I

2461        TCTTACAGCCCTTTAGGAAACAAAATCCAGAGATGGTTATCTATCAATACATGGATGCTT
(Pol)          L  Q  P  F  R  K  Q  N  P  E  M  V  I  Y  Q  Y  M  D  A  L

2521        TGTATGTAGGATCTGCCTTAGAAATAGGGCAGCATAGAACAAAAATAGAGGAATTAAGAG
(Pol)          Y  V  G  S  A  L  E  I  G  Q  H  R  T  K  I  E  E  L  R  E

2581        AACATCTATTGAGATGGGGATTTACAACACCAATAAAAAAACATCAGAAAGAACCTCCAT
(Pol)          H  L  L  R  W  G  F  T  T  P  I  K  K  H  Q  K  E  P  P  F

2641        TTCTTTGGATGGGTTATGAACTCCATCCTGATAATTGGACAGTACAAGCCTATACTCTGC
(Pol)          L  W  M  G  Y  E  L  H  P  D  N  W  T  V  Q  A  Y  T  L  P
```

```
2701        CAGACAAAGAAAGCTGGACTGTCAATGATATTCAGAAGTTAGTAGGGAAATTAGTGGGAA
(Pol)          D  K  E  S  W  T  V  N  D  I  Q  K  L  V  G  K  L  V  G  S

2761        GCCAGATTTATCAGGAATTGAAAGTAAAGCAATTATGTAAACCCTTAGGGGAACCCAAAG
(Pol)          Q  I  Y  Q  E  L  K  V  K  Q  L  C  K  P  L  G  E  P  K  A

2821        CACTAACAGAAGTAATATCACTATCAGCAGAAGCAGAATTAGAACTGGCAGAAAACAGGG
(Pol)          L  T  E  V  I  S  L  S  A  E  A  E  L  E  L  A  E  N  R  E

2881        AAATTTATAAGGAACCAGTACATGGAGTGTATTATGACCCATCAAAAGACTTACTACCAG
(Pol)          I  Y  K  E  P  V  H  G  V  Y  Y  D  P  S  K  D  L  L  P  E

2941        AAATACAGAAACAAGGGAATGGCCAATGGACATACCAAATTTATCAAGAACCATTTAAAA
(Pol)          I  Q  K  Q  G  N  G  Q  W  T  Y  Q  I  Y  Q  E  P  F  K  N

3001        ATCTGAAAACAGGGAAGTATGCAAGAACGAGGGGTGCCCATACTAATGATGTAAAACAAT
(Pol)          L  K  T  G  K  Y  A  R  T  R  G  A  H  T  N  D  V  K  Q  L

3061        TACCAGAGGCAGTGCAAAAAATGGCCACAGAAAGGATAGTAATATGGGGAAAGACTCCTA
(Pol)          P  E  A  V  Q  K  M  A  T  E  R  I  V  I  W  G  K  T  P  K

3121        AATTTAGACTGCCCATACAAAAGGAAACATGGGAAACATGGTGGATAGAGTATTGGCAAG
(Pol)          F  R  L  P  I  Q  K  E  T  W  E  T  W  W  I  E  Y  W  Q  A

3181        CCACCTGGATTCCTGAGTGGGAATTTGTCAATACCCCTCCTTTGGTAAAATTATGGTACC
(Pol)          T  W  I  P  E  W  E  F  V  N  T  P  P  L  V  K  L  W  Y  Q

3241        AATTTAGAGGAACCCCCATAGTGGGAGCAGAAACTTTCTATGGAGATGGGGCAGCTAATA
(Pol)          F  R  G  T  P  I  V  G  A  E  T  F  Y  G  D  G  A  A  N  R

3301        GAGAGACTAGAGCAGGAAAAGCAGGATATGTTACTGACAGAGGAAGACAGAAAGTTGTCC
(Pol)          E  T  R  A  G  K  A  G  Y  V  T  D  R  G  R  Q  K  V  V  P

3361        CTTTTACTGACACAACAAATCAGAAGACTGAGTTACATGCAGTTAATCTACCTTTGCAGG
(Pol)          F  T  D  T  T  N  Q  K  T  E  L  H  A  V  N  L  P  L  Q  D

3421        ATTCGGGATTTGAAGTTAACAGCGTACCAGATTCACAATATGTATTTGGAATCATTCAAG
(Pol)          S  G  F  E  V  N  S  V  P  D  S  Q  Y  V  F  G  I  I  Q  A

3481        CACAACCAGATAAAAGTGAACCAGAGTTTGTCAGTCAAATATTATACCAGCGAATCAAAA
(Pol)          Q  P  D  K  S  E  P  E  F  V  S  Q  I  L  Y  Q  R  I  K  K

3541        AGGAAAAGGTTTACCTGGCATGGGTACCAGCACACAAAGGAATTGGAGGAAATGAACAAG
(Pol)          E  K  V  Y  L  A  W  V  P  A  H  K  G  I  G  G  N  E  Q  E

3601        AAGATACGTTTGTCAGTGCGGGAATCAGGAAAGTACTATTTTTGGATGGAATAGACAAGG
(Pol)          D  T  F  V  S  A  G  I  R  K  V  L  F  L  D  G  I  D  K  A
3661        CTCAAGAAGAACATGTGAAATATCACAACAATTGGAGAGCAATGGCTAGTGATTTTAGCC
(Pol)          Q  E  E  H  V  K  Y  H  N  N  W  R  A  M  A  S  D  F  S  L

3721        TACCACCTGTACTACCAAAAGAAATACTACCTACCTGTGATAAATGTCAGCTACAAGAAA
(Pol)          P  P  V  L  P  K  E  I  L  P  T  C  D  K  C  Q  L  Q  E  T

3781        CCATGCATGGACAAGTACACTGTAGTCCAGGAATATGGCAATTACATTGTACACATTTAG
(Pol)          M  H  G  Q  V  H  C  S  P  G  I  W  Q  L  H  C  T  H  L  E

3841        AAGGAAAAGTTATCATAGTAGCAGTTCATGTACCCAGTGGCTATATACAAGCAGAAGTTA
(Pol)          G  K  V  I  I  V  A  V  H  V  P  S  G  Y  I  Q  A  E  V  I

3901        TTCCGGCAGAAACAGGCCAGGAAACAGCATACTTTCTCTTTACATTATCAGGAAGATGGC
(Pol)          P  A  E  T  G  Q  E  T  A  Y  F  L  F  T  L  S  G  R  W  P

3961        CAGTTACAGTACTACATACAGACAATGGCAGCAATTTCACCAGTGCTGCAGTTATGGCCG
(Pol)          V  T  V  L  H  T  D  N  G  S  N  F  T  S  A  A  V  M  A  A

4021        CCTGCTGGTGGGCAGGCATCAAACAGGAATTTGGAATTCCCTACAATCCCCAAAGTCAAG
(Pol)          C  W  W  A  G  I  K  Q  E  F  G  I  P  Y  N  P  Q  S  Q  G
```

```
4081        GAAGTATTACAATCTATAATATAGAATTAAAGAAAATTATTGGGCAGGTAAGAGATCAAG
(Pol)         S  I  T  I  Y  N  I  E  L  K  K  I  I  G  Q  V  R  D  Q  A

4141        CTGAGCATCTAAAGACAGCAGTACAAATGGCAGTATTCATCCACAATTTTAAAAGAAAAG
(Pol)         E  H  L  K  T  A  V  Q  M  A  V  F  I  H  N  F  K  R  K  G

4201        GGGGGATTGGGGGATACAGTGCAGGGGAAAGAATATTACACATATTACCAACAGACATAC
(Pol)         G  I  G  G  Y  S  A  G  E  R  I  L  H  I  L  P  T  D  I  Q

4261        AAACTAAAGAATTACAAAAGCAAATCACAAAAATTCAAAATTTTCGGGTTTATTACAGGG
(Pol)         T  K  E  L  Q  K  Q  I  T  K  I  Q  N  F  R  V  Y  Y  R  D

4321        ACAGCAGAGATCCAATTTGGAAAGGACCAGCAAAACTTCTCTGGAAAGGTCAAGGGGCAG
(Pol)         S  R  D  P  I  W  K  G  P  A  K  L  L  W  K  G  Q  G  A  V

4381        TAGTAATACAAGACAATCGTTACATAAAGGTAGTACCAAGAAGAAAAGTGAAAATCATTC
(Pol)         V  I  Q  D  N  R  Y  I  K  V  V  P  R  R  K  V  K  I  I  R

4441        GGGATTATGGAAAACAGATGGCAGGAGACGATTGTGTGGCAAGTACACAGGACGAGGATT
(Vif start)      M  E  N  R  W  Q  E  T  I  V  W  Q  V  H  R  T  R  I
(Pol end)     D  Y  G  K  Q  M  A  G  D  D  C  V  A  S  T  Q  D  E  D  *

4501        AGCACATGGAAAAGTTTAGTAAAATACCATATGTATGTTTCAAAAAAGGCTAAAGGATGG
(Vif)        S  T  W  K  S  L  V  K  Y  H  M  Y  V  S  K  K  A  K  G  W

4561        TTTTATAGACACCATGGCAGCCCCCACCCAAAAATAAGCTCAGAAGTACACATTCCACTA
(Vif)        F  Y  R  H  H  G  S  P  H  P  K  I  S  S  E  V  H  I  P  L

4621        GGAGAAGAAAGACTGGTCGTACAAACATATTGGGGTCTGCATACAGGAGAAAGAGAATGG
(Vif)        G  E  E  R  L  V  V  Q  T  Y  W  G  L  H  T  G  E  R  E  W

4681        CATCTGGGTCAGGGAGTCTCCATAGAATGGAGGAAAAGGAAATATAGCACCCAAGTATAC
(Vif)        H  L  G  Q  G  V  S  I  E  W  R  K  R  K  Y  S  T  Q  V  Y

4741        CCTGGCCTGGCAGACCAACTAATTCATATATATTATTTTGATTGTTTTTCAGACTCTGCT
(Vif)        P  G  L  A  D  Q  L  I  H  I  Y  Y  F  D  C  F  S  D  S  A

4801        ATAAGAAAAGCCTTATTAGGACATATAGTTACACCTCGGTGTGAATATCAAGCAGGACAT
(Vif)        I  R  K  A  L  L  G  H  I  V  T  P  R  C  E  Y  Q  A  G  H

4861        CACAAGGTAGGATCCTTACAGTATTTGGCACTAACAGCATTAATAGCACCAAAAAAGACA
(Vif)        H  K  V  G  S  L  Q  Y  L  A  L  T  A  L  I  A  P  K  K  T

4921        AAGCCACCTTTGCCTATTGTTATGAAGCTAACAGAAGATACATGGAACAAGCCCCAGAAG
(Vif)        K  P  P  L  P  I  V  M  K  L  T  E  D  T  W  N  K  P  Q  K
(Vpr start)                                                M  E  Q  A  P  E  D

4981        ACCAAGGGCCACAGAGGGAGCCATACAATGAATGGACATTAGAACTTCTGGAGGAGCTTA
(Vif end)    T  K  G  H  R  G  S  H  T  M  N  G  H  *
(Vpr)          Q  G  P  Q  R  E  P  Y  N  E  W  T  L  E  L  L  E  E  L  K

5041        AGAGTGAAGCTGTTAGACACTTTCCTAGAATATGGCTCCATAGCTTAGGACAACATATCT
(Vpr)          S  E  A  V  R  H  F  P  R  I  W  L  H  S  L  G  Q  H  I  Y

5101        ATGAAACTTATGGGGATTCCTGGACAGGAGTTGAAGCTATAATAAGAATTCTGCAACAAT
(Vpr)          E  T  Y  G  D  S  W  T  G  V  E  A  I  I  R  I  L  Q  Q  L

5161        TACTGTTTATTCATTTCAGAATTGGGTGTCAACATCGCAGAATAGGTATTACTCGGCAGA
(Vpr)          L  F  I  H  F  R  I  G  C  Q  H  R  R  I  G  I  T  R  Q  R

5221        GAAGAGCAAGAAATGGATCCAGAAGATCCTAGCTTGAGCTGGAACCATCCAGGAAGTCAG
(Tatx1 start)          M  D  P  E  D  P  S  L  S  W  N  H  P  G  S  Q
(Vpr)          R  A  R  N  G  S  R  R  S  *

5281        CCTAAGACTGCTTGTAACAAGTGTCATTGTAAAAAGTGTTGCTTTCATTGTCAAGTTTGC
(Tatx1)      P  K  T  A  C  N  K  C  H  C  K  K  C  C  F  H  C  Q  V  C
```

```
5341        TTCATCACGAAAGGCTTTGGCATCTCCTATGGCAGGAAGAAGCGGAGACAGCGACGAAAA
(Tatx1)      F  I  T  K  G  F  G  I  S  Y  G  R  K  K  R  R  Q  R  R  K
(Revx1 start)                          M  A  G  R  S  G  D  S  D  E  N

5401        CCTCCTCACGGCGATCAGGCTCATCAAGTTCCTATACCAGAGCAGTAAGTAGTTCATGTA
(Tatx1 end)P  P  H  G  D  Q  A  H  Q  V  P  I  P  E  Q  *
(Revx1 end) L  L  T  A  I  R  L  I  K  F  L  Y  Q  S  S  K  *

5461        ATGCAGCCTTTAGTGATAATAGCAATAGCAGCATTAGTAGTAGCAATAATAATAGCAATA
(Vpu start)M  Q  P  L  V  I  I  A  I  A  A  L  V  V  A  I  I  I  A  I

5521        GTTGTGTGGACCATAGTATTCATAGAATATAGGAGAATAAAAAGGCAAAGAAAAATACAC
(Vpu)        V  V  W  T  I  V  F  I  E  Y  R  R  I  K  R  Q  R  K  I  H

5581        TGTTTACTTGATAGAATTATAGAAGACAGCAAGACCAGTGGCAATGAGAGCGAGGGGATA
(Vpu)        C  L  L  D  R  I  I  E  D  S  K  T  S  G  N  E  S  E  G  I
(Env start)                                                 M  R  A  R  G  Y

5641        CCAGAGGAATTGTCCACCAACTTGGTGGACATGGGGCATCATGCTCCTTGGGATGTTGAC
(Vpu)        P  E  E  L  S  T  N  L  V  D  M  G  H  H  A  P  W  D  V  D
(Env)        Q  R  N  C  P  P  T  W  W  T  W  G  I  M  L  L  G  M  L  T

5701        GATCTGTAGCGCTGCAAGAAATTTGTGGGTCACAGTTTATTATGGGGGTGCCTGTATTGG
(Vpu end)    D  L  *
(Env)        I  C  S  A  A  R  N  L  W  V  T  V  Y  Y  G  G  A  C  I  G

5761        ACTCTCTTTTTGTGATCAGATGTACTCTATTCAACAGAGGCCCATAATATTTGGGCTACA
(Env)        L  S  F  C  D  Q  M  Y  S  I  Q  Q  R  P  I  I  F  G  L  H

5821        CATGCCTGTGTACCCACGGACCCCAGCCCACAAGAAATATAACTGGAAAATGTGGCCGAA
(Env)        M  P  V  Y  P  R  T  P  A  H  K  K  Y  N  W  K  M  W  P  K

5881        AACTTTAATATGTGGAAAAATAACATGGGAGACCAGATGCATGAGGATAGAATCAGTTTA
(Env)        T  L  I  C  G  K  I  T  W  E  T  R  C  M  R  I  E  S  V  Y

5941        TGGGATCAAAGCCCTAAAGCCATGTGTAAAATTAACCCCACTCTGTGTCACTTTAAACTG
(Env)        G  I  K  A  L  K  P  C  V  K  L  T  P  L  C  V  T  L  N  C

6001        CAGTAATAATATTACCACCTTAAACAGCACTGGGAATGCCACCTTAAACAGCACTAGGAA
(Env)        S  N  N  I  T  T  L  N  S  T  G  N  A  T  L  N  S  T  R  N

6061        CGCCACTGTGGAGTCAGAACTGGAGATGAAAAACTGCTCTTTCAATATAACTACAGTAGT
(Env)        A  T  V  E  S  E  L  E  M  K  N  C  S  F  N  I  T  T  V  V

6121        AAGAGATAAGAAAATGCAAGTACATGCGCTTTTTTATAGACCTGATATAGTATCAATAAA
(Env)        R  D  K  K  M  Q  V  H  A  L  F  Y  R  P  D  I  V  S  I  N

6181        CAATGATAACACCAGTTATAGGTTAATAAATTGTAATACCTCATCCATTACACAGGCTTG
(Env)        N  D  N  T  S  Y  R  L  I  N  C  N  T  S  S  I  T  Q  A  C

6241        TCCAAAGGTATCCTTTGAACCAATTCCAATACATTATTGTGCCCCAGCTGGTTTTGCAAT
(Env)        P  K  V  S  F  E  P  I  P  I  H  Y  C  A  P  A  G  F  A  I

6301        TCTAAAGTGTAGAGATAAGAAGTTCAATGGAACAGGCCTATGCACAAATATCAGCACAGA
(Env)        L  K  C  R  D  K  K  F  N  G  T  G  L  C  T  N  I  S  T  E

6361        ACAATGTACACATGGAATTAAGCCAGTGGTGACAACTCAACTGCTGTTGAATGGCAGTCT
(Env)        Q  C  T  H  G  I  K  P  V  V  T  T  Q  L  L  L  N  G  S  L

6421        AGCAGAAGAAGAGATCATAATTAGATCTGAAAATCTCACAAACAATGCTAAAAACATAAT
(Env)        A  E  E  E  I  I  I  R  S  E  N  L  T  N  N  A  K  N  I  I

6481        AGTACAGTTTAATGCATCTGTAGAAATTAATTGTACAAGGCCCTACAGATATATAAGACA
(Env)        V  Q  F  N  A  S  V  E  I  N  C  T  R  P  Y  R  Y  I  R  Q

6541        AAAAACGTCAATAGGACAAGGGCAAACATTACATACAAGCAAGAGGATAATAGGAGACAT
(Env)        K  T  S  I  G  Q  G  Q  T  L  H  T  S  K  R  I  I  G  D  I
```

```
6601        AAGACAAGCACATTGTAACATTAGTGGAAGAAAATGGCATAAAACTTTACAACAGGTAGC
(Env)         R  Q  A  H  C  N  I  S  G  R  K  W  H  K  T  L  Q  Q  V  A

6661        TACAAAATTAAGAAACCTTCTTAATAAAACAACAATAATTTTTCGACCACCCCCAGGAGG
(Env)         T  K  L  R  N  L  L  N  K  T  T  I  I  F  R  P  P  P  G  G

6721        GGACCCAGAAATTACAACACACAGTTTTAATTGTGGAGGGGAATTTTTCTACTGTAATAC
(Env)         D  P  E  I  T  T  H  S  F  N  C  G  G  E  F  F  Y  C  N  T

6781        ATCTAGGCTGTTTAATAATACATGGAATGGTACACATGTCAATAAGACAGACACCAATGG
(Env)         S  R  L  F  N  N  T  W  N  G  T  H  V  N  K  T  D  T  N  G

6841        GGCAGTCACACTCCCATGCAGAATAAAACAAATTATAAACATGTGGCAGGGAGTGGGAAA
(Env)         A  V  T  L  P  C  R  I  K  Q  I  I  N  M  W  Q  G  V  G  K

6901        AGCAATGTATGCCCCTCCCATAGAAGGACTAATTAGATGTTCATCAAATATTACAGGGCT
(Env)         A  M  Y  A  P  P  I  E  G  L  I  R  C  S  S  N  I  T  G  L

6961        AATATTGACAAGAGATGGGGGTAATAGTAGTTCTGACAACGAGACCTTCAGACCTGGTGG
(Env)         I  L  T  R  D  G  G  N  S  S  S  D  N  E  T  F  R  P  G  G

7021        AGGAAATATGAGGGACAATTGGAGAAGTGAATTATATAAATACAAAGTAGTACAAATTGA
(Env)         G  N  M  R  D  N  W  R  S  E  L  Y  K  Y  K  V  V  Q  I  E

7081        ACCAATAGGAGTAGTGCCCACCAGGGCAAAGAGAAGAGTGGTGGAAAGGGAAAAAAGAGC
(Env)         P  I  G  V  V  P  T  R  A  K  R  R  V  V  E  R  E  K  R  A

7141        AATAGGACTAGGAGCCATGTTCCTTGGGTTCTTGGGAGCAGCAGGAAGCACGATGGGCGA
(Env)         I  G  L  G  A  M  F  L  G  F  L  G  A  A  G  S  T  M  G  E

7201        GTCATTGACGCTGACGGTACAGGCCAGACAGGTATTGTCTGGTATAGTGCAACAGCAAAG
(Env)         S  L  T  L  T  V  Q  A  R  Q  V  L  S  G  I  V  Q  Q  Q  S

7261        CAATTTGCTGAGGGCTATAGAGGCGCAACAGCATCTGTTGCAACTCACGGTCTGGGGCAT
(Env)         N  L  L  R  A  I  E  A  Q  Q  H  L  L  Q  L  T  V  W  G  I

7321        TATACAGCTCCAGGCAAGAATCCTGGCTGTGGAAAGATACCTAAAGGATCAACGGCTCCT
(Env)         I  Q  L  Q  A  R  I  L  A  V  E  R  Y  L  K  D  Q  R  L  L

7381        AGACTTGTGGGGTTGCTCTGGAAAACACATTTGCACCACTACTGTGCCCTGGAACTCTAG
(Env)         D  L  W  G  C  S  G  K  H  I  C  T  T  T  V  P  W  N  S  S

7441        TTGGAGTAATAAATCAAGATGCGATTTGCATACCATGACCTGGATGCGGGGAAAGAAAAT
(Env)         W  S  N  K  S  R  C  D  L  H  T  M  T  W  M  R  G  K  K  I

7501        TCACAATTACACGGACTATTATACAGCTTATTGCAGTTCGCAAATTCAGCAAGAAAAGAA
(Env)         H  N  Y  T  D  Y  Y  T  A  Y  C  S  S  Q  I  Q  Q  E  K  N

7561        TGACAAGGAATTATTGGAATTGGACAAGTGGGCAAGTCTGTGGAATTGGTTTACAATAAC
(Env)         D  K  E  L  L  E  L  D  K  W  A  S  L  W  N  W  F  T  I  T

7621        AAACTGGCTGTGGTATATAAGAATATTCATAATGATAGTAGGAGGCTTAATAGGTTTATG
(Env)         N  W  L  W  Y  I  R  I  F  I  M  I  V  G  G  L  I  G  L  C

7681        TATAGTTTTTTCTGTGCTTTCTGTACTACATAGAGTTAGGCAGGGATACTCACCTCTGTC
(Env)           I  V  F  S  V  L  S  V  L  H  R  V  R  Q  G  Y  S  P  L  S

7741        GTTTCAGACCCTCCTCCCGGCCCCGAGGGGACCCGACAGGCCCGAAGGAACAGAAGAAGA
(Revx2 start) S  D  P  P  P  G  P  E  G  T  R  Q  A  R  R  N  R  R  R
(Env)           F  Q  T  L  L  P  A  P  R  G  P  D  R  P  E  G  T  E  E  E
(Tatx2 start)      P  S  S  R  P  R  G  D  P  T  G  P  K  E  Q  K  K  K

7801        AGGTGGAGAGCGAGGCAGAGACAAATCAATTCATTTGGCGAACGGATTAGCAGCACTTAT
(Revx2)       R  W  R  A  R  Q  R  Q  I  N  S  F  G  E  R  I  S  S  T  Y
(Env)         G  G  E  R  G  R  D  K  S  I  H  L  A  N  G  L  A  A  L  I
(Tatx2 end)   V  E  S  E  A  E  T  N  Q  F  I  W  R  T  D  *
```

```
7861        CTGGGACGATCTGCGGAACCTGTGCCTCTTCAGCTACCACCGCTCGAGAGACTTACTCTT
(Revx2)      L  G  R  S  A  E  P  V  P  L  Q  L  P  P  L  E  R  L  T  L
(Env)         W  D  D  L  R  N  L  C  L  F  S  Y  H  R  S  R  D  L  L  F

7921        TATTGCAGCGAGGATTGTGGACCTTCTGGGACGCAGGGGGTGGGAATCAAGTATCTGTGG
(Revx2)      Y  C  S  E  D  C  G  P  S  G  T  Q  G  V  G  I  K  Y  L  W
(Env)         I  A  A  R  I  V  D  L  L  G  R  R  G  W  E  S  S  I  C  G

7981        ATCCTCCTGCAGTATTGGAGTCAGGAATGACGAAATAGAGCTATTAACTTGCTTGATACA
(Revx2 end)I  L  L  Q  Y  W  S  Q  E  *
(Env)         S  S  C  S  I  G  V  R  N  D  E  I  E  L  L  T  C  L  I  Q

8041        ATATCAATACTTACAGCTGCGGGGACAGATACGGTTACAGAAGTACTACAAAGAGCTTGC
(Env)         Y  Q  Y  L  Q  L  R  G  Q  I  R  L  Q  K  Y  Y  K  E  L  A

8101        AGAGCTAACCGTACCCACAAGAATACGACAGGGCTTGGAAAGGCTTTTGCTATAAAATGG
(Env end)    E  L  T  V  P  T  R  I  R  Q  G  L  E  R  L  L  L  *  N  G
(Nef start)                                                          M  G

8161        GTGGCAAATGGTCAAAAAGTACTATAGTTGGATGGTCTGCTATAAGGGAAAGAATAAGAA
(Nef)          G  K  W  S  K  S  T  I  V  G  W  S  A  I  R  E  R  I  R  R

8221        GAACTGATCCAGCAGCAGATGGGGTGGGAGCAGTATCTCGAGACCTGGAAAAACATGGGG
(Nef)          T  D  P  A  A  D  G  V  G  A  V  S  R  D  L  E  K  H  G  A

8281        CAATCACAAGTAGCAATACAGCAAGTACTAATGCTGACTGTGCCTGGCTAGAAGCACAAG
(Nef)          I  T  S  S  N  T  A  S  T  N  A  D  C  A  W  L  E  A  Q  E

8341        AAGAGAGTGAGGAGGTGGGCTTTCCAGTCAGACCTCAGGTACCTTTACGACCAATGTCTT
(Nef)          E  S  E  E  V  G  F  P  V  R  P  Q  V  P  L  R  P  M  S  Y

8401        ACAAAGCAGCTCTCGATCTTAGCCACTTTTTAAAAGAAAAGGGGGGACTGGAAGGGCAAA
(Nef)          K  A  A  L  D  L  S  H  F  L  K  E  K  G  G  L  E  G  Q  I

8461        TTTGGTCCAAAAAGAGACAGGAGATCCTTCATCTTTGGGTCTACCACACACAAGGCTACT
(Nef)          W  S  K  K  R  Q  E  I  L  H  L  W  V  Y  H  T  Q  G  Y  F

8521        TCCCCGATTGGCAGAACTACACACCAGGGCCAGGGATCAGATCTCCACTGACTTTTGGAT
(Nef)          P  D  W  Q  N  Y  T  P  G  P  G  I  R  S  P  L  T  F  G  W

8581        GGTGCTTCGAGCTACTACCAGTTGATCCACAGGAGGTAGAAGAGGCCACTGAGGGAGAGA
(Nef)          C  F  E  L  L  P  V  D  P  Q  E  V  E  E  A  T  E  G  E  T

8641        CCAACTGCTTGTTACACCCTATGAACCAGCATGGAATGGAGGACCCGGAGGGACAAGTGT
(Nef)          N  C  L  L  H  P  M  N  Q  H  G  M  E  D  P  E  G  Q  V  L

8701        TAAAGTGGAGATTTAACAGCAGACTAGCATTTGAGCACAAGGCCCGACAGCTACATCCGG
(Nef)          K  W  R  F  N  S  R  L  A  F  E  H  K  A  R  Q  L  H  P  E

8761        AGTACTACAAAGACTGCTGACACCGAGTTTTCTACAGGGGACTTTCCGCTGGGGACTTTC
(Nef end)      Y  Y  K  D  C  *
```

B3: pR286 Full-length sequence (nucleotide and amino acid)

```
181         AGCGACTGGTGAGTACGCTAAAATTTTTGACTAGCGGAGGCTAGAAGGAGAGAGATGGGT
(Gag start)                                                       M  G

241         GCGAGAGCGTCAGTATTAAGCGGGGGAAAATTAGATGCATGGGAAAGAATTCGGTTAAGG
(Gag)        A  R  A  S  V  L  S  G  G  K  L  D  A  W  E  R  I  R  L  R

301         CCAGGAGGAAAGAAACAATATAAACTAAAACATATAGTATGGGCAAGCAGGGAGCTAGAA
(Gag)        P  G  G  K  K  Q  Y  K  L  K  H  I  V  W  A  S  R  E  L  E

361         CGATTTGCACTTAATCCTGGCCTTTTAGAAACATCAGAAGGCTGTAAACAAATAATAGGA
(Gag)        R  F  A  L  N  P  G  L  L  E  T  S  E  G  C  K  Q  I  I  G

421         CAGCTCCAGCCATCTCTTCAGACAGGATCAGAAGAACTTAGATCATTATATCTAACAATA
(Gag)        Q  L  Q  P  S  L  Q  T  G  S  E  E  L  R  S  L  Y  L  T  I

481         GCAACCCTCTATTGTGTACATGCAAGGATAGATGTAAAAGACACCAAGGAAGCTTTAGAA
(Gag)        A  T  L  Y  C  V  H  A  R  I  D  V  K  D  T  K  E  A  L  E

541         AAGATAGAGGAAGAGCAAAACAAAAGTAAGAAAAAGAAGGCACAGCAAGCAGCGGCTGAC
(Gag)        K  I  E  E  E  Q  N  K  S  K  K  K  A  Q  Q  A  A  A  D

601         ACAGGAAACAGCAGCCAGGTCAGCCAAAATTATCCTATAGTGCAGAACCTACAGGGGCAA
(Gag)        T  G  N  S  S  Q  V  S  Q  N  Y  P  I  V  Q  N  L  Q  G  Q

661         ATGGTACATCAGGCCATATCACCAAGAACTTTAATCGCATGGGTAAAATATGTAGAAGAA
(Gag)        M  V  H  Q  A  I  S  P  R  T  L  I  A  W  V  K  Y  V  E  E

721         AAGGCCTTCAGCCCAGAAGTTATACCCATGTTTTCAGCATTATCAGAAGGAGCCACCCCA
(Gag)        K  A  F  S  P  E  V  I  P  M  F  S  A  L  S  E  G  A  T  P

781         CAAGATTTATACACCATGCTATACACAGTGGGGGGACATCAAGCAGCCATGCAAATGCTC
(Gag)        Q  D  L  Y  T  M  L  Y  T  V  G  G  H  Q  A  A  M  Q  M  L

841         AAAGAGACCATCAATGAGGAGGCTGCAGAATGGGATACGCTACATCCAGTGCATGCAGGG
(Gag)        K  E  T  I  N  E  E  A  A  E  W  D  T  L  H  P  V  H  A  G

901         CCTATTGCACCAGGCCAGATGAGAGAACCAAGGGGAAGTGCTATAGCAGGAACTACTATT
(Gag)        P  I  A  P  G  Q  M  R  E  P  R  G  S  A  I  A  G  T  T  I

961         ACCCTTCAGGAACAAATAGCATGGATGACAAGCAACCCACCTATCCCAGTAGGAGAAATC
(Gag)        T  L  Q  E  Q  I  A  W  M  T  S  N  P  P  I  P  V  G  E  I

1021        TATACAAGATGGATAATCCTGGGATTATATAAAATAGTAAGAATGTATATCCCTGTCAGC
(Gag)        Y  T  R  W  I  I  L  G  L  Y  K  I  V  R  M  Y  I  P  V  S

1081        ATTTTGGACATAAGACAGGGACCAAAGGAACCTTTTACAGATTATGTAGACCGGTTCTTA
(Gag)        I  L  D  I  R  Q  G  P  K  E  P  F  T  D  Y  V  D  R  F  L

1141        AAAACTCTACGAGCCGAGCAAGCTTCACAGGATGTATACAACTGGAAGACAGAAACCTTG
(Gag)        K  T  L  R  A  E  Q  A  S  Q  D  V  Y  N  W  K  T  E  T  L

1201        TTGGTCCAAAATGCAAACCCAGATTGTAAAACCATCTTACAAGCATTACGACCACAGGCT
(Gag)        L  V  Q  N  A  N  P  D  C  K  T  I  L  Q  A  L  R  P  Q  A

1261        ACACTAGAAGAAATGCTGCCAGCATGTCAGGGAGTGGGGGGGCCCAGCCATAAAGCAAGA
(Gag)        T  L  E  E  M  L  P  A  C  Q  G  V  G  G  P  S  H  K  A  R

1321        GTTTTGGCTGAGGCAATCAGCCAAGCAACAAATTCAGCTACTATAATGATGCTGCAGAGA
(Gag)        V  L  A  E  A  I  S  Q  A  T  N  S  A  T  I  M  M  L  Q  R

1381        GGCAATTTTTACGGCCAAAGAAAAATTGTTCAGTGTTTCAACTGTGGCAAAGAAGGGCCA
(Gag)        G  N  F  Y  G  Q  R  K  I  V  Q  C  F  N  C  G  K  E  G  P
```

```
1441        CATACCGCAAAAAATTGCAGGGCCCCTAGGAAAAAGGGCTGTTGGAAATGTGGAAGGGAA
(Gag)        H  T  A  K  N  C  R  A  P  R  K  K  G  C  W  K  C  G  R  E

1501        GGACACCAAATCAAAGAATGCACTGCAAGACAGGCTACTTTTTTTGGGAAGATTTGGCCT
(Gag)        G  H  Q  I  K  E  C  T  A  R  Q  A  T  F  F  F  G  K  I  W  P
(Pol start)                                              F  F  W  E  D  L  A  F

1561        TCCCAAAAGGGGAGGCCGGGGAACTTTCTTCAGAGCAGACCAGAGCCAACAGCCCCACCA
(Gag)        S  Q  K  G  R  P  G  N  F  L  Q  S  R  P  E  P  T  A  P  P
(Pol)          P  K  G  E  A  G  E  L  S  S  E  Q  T  R  A  N  S  P  T  S

1621        GCAGAGAGCTTCGGGTTTGGGGAGGAGATTACCCCCTCTCAGAAACAGGAACCAATAGAC
(Gag)        A  E  S  F  G  F  G  E  E  I  T  P  S  Q  K  Q  E  P  I  D
(Pol)          R  E  L  R  V  W  G  G  D  Y  P  L  S  E  T  G  T  N  R  Q

1681        AAGGAACTGTATCCTTTTACCTCCCTCAAATCACTCTTTGGGAACGACCCCTTGTCACAA
(Gag)        K  E  L  Y  P  F  T  S  L  K  S  L  F  G  N  D  P  L  S  Q
(Pol)          G  T  V  S  F  Y  L  P  Q  I  T  L  W  E  R  P  L  V  T  I

1741        TAAAGATAGGGGGACAGCTAAAGGAAGCTCTATTAGATACAGGAGCAGATGTTACAGTAT
(Gag end)    *
(Pol)          K  I  G  G  Q  L  K  E  A  L  L  D  T  G  A  D  V  T  V  L

1801        TAGAAGAAATGAATTTGCCAGGAAAATGGAAACCAAAAATGATAGGGGGAATTGGAGGTT
(Pol)          E  E  M  N  L  P  G  K  W  K  P  K  M  I  G  G  I  G  G  F

1861        TTATCAAAGTAAGACAGTCATGTTCAAATACCCCCTTAGAAATCTGTGGGCATAAAGCTA
(Pol)          I  K  V  R  Q  S  C  S  N  T  P  L  E  I  C  G  H  K  A  I

1921        TTGGTACAGTATTCATAGGACCTACACCCGTCAACATAATTGGAAGAAATTTGTTGACTC
(Pol)          G  T  V  F  I  G  P  T  P  V  N  I  I  G  R  N  L  L  T  Q

1981        AGCCTGGCTGCACTTTACATTTTCCAATTAGTCCTAGTGAAACTGTACCAGTTAAATTCA
(Pol)          P  G  C  T  L  H  F  P  I  S  P  S  E  T  V  P  V  K  F  K

2041        AGCCAGGAATGGATGGCCCAAAAGTTAAGCAATGGCCATTGCCAGAAGAAAAATACAAGG
(Pol)          P  G  M  D  G  P  K  V  K  Q  W  P  L  P  E  E  K  Y  K  A

2101        CATTACCAGAAATTTGTACAGAAATGGAAAAGGAAGGAAAAATTTCAAGAATTGGGCCTG
(Pol)          L  P  E  I  C  T  E  M  E  K  E  G  K  I  S  R  I  G  P  E

2161        AAAATCCATACAATACTCCAATATTTGCCATAAAGAAAAAAGACAGTACTATATGGAGAA
(Pol)          N  P  Y  N  T  P  I  F  A  I  K  K  K  D  S  T  I  W  R  K

2221        AATTACTATACTTCAGAGAACTTAATCAGAGAACTCAAGATTTCTGGGAAGTACAATTAG
(Pol)          L  L  Y  F  R  E  L  N  Q  R  T  Q  D  F  W  E  V  Q  L  G

2281        GAATACCGCATCCTGCAGGGCTGAAAAAGAAAAAATCAGGAACAGTACTGGATGTGGGTG
(Pol)          I  P  H  P  A  G  L  K  K  K  K  S  G  T  V  L  D  V  G  D

2341        ATGCATATTTTTCAGTTCCCTTATGTGAAGACTTTAGAAAATATACTGCATTTACCATAC
(Pol)          A  Y  F  S  V  P  L  C  E  D  F  R  K  Y  T  A  F  T  I  P

2401        CGAGTATAAACAATGCGACACCGGGAATTAGATATCAGTACAATGTGCTTCCACAGGGAT
(Pol)          S  I  N  N  A  T  P  G  I  R  Y  Q  Y  N  V  L  P  Q  G  W

2461        GGAAAGGATCACCGGCAATATTCCAAAGTAGCATTACAAAAATCTTTGAGCCCTTTAGAA
(Pol)          K  G  S  P  A  I  F  Q  S  S  I  T  K  I  F  E  P  F  R  K

2521        AACAAAATCCAGAGAAAGCTATCTATCAATACATGGATGATTTGTATGTACGATCTGACT
(Pol)          Q  N  P  E  K  A  I  Y  Q  Y  M  D  D  L  Y  V  R  S  D  S

2581        CAAAATATGGCCAGCATACAACAAAAATAGAGGAATTACGAGAACATCTATTGCGGTGGG
(Pol)          K  Y  G  Q  H  T  T  K  I  E  E  L  R  E  H  L  L  R  W  G

2641        GATTTACTACACCAGAAAAAAAACATCAGAAAGAACCTCCATTTCTTTGGATGGGTTATG
(Pol)          F  T  T  P  E  K  K  H  Q  K  E  P  P  F  L  W  M  G  Y  E
```

```
2701        AACTCCATCCTGTCAAATGGACAGTACAGCCTATACAACTGCCAGAAAAAGAAGACTGGA
(Pol)         L  H  P  V  K  W  T  V  Q  P  I  Q  L  P  E  K  E  D  W  T

2761        CTGTCAATGCTATACAGAAGTTATTACGGAAATTATACTGGGCAAGCCAGATTTATCCAG
(Pol)         V  N  A  I  Q  K  L  L  R  K  L  Y  W  A  S  Q  I  Y  P  G

2821        GAATCAAAGTATGGCAATTATGGAAACTCCTTATGGGAACCAAAGCACTACCAGAAGTAC
(Pol)         I  K  V  W  Q  L  W  K  L  L  M  G  T  K  A  L  P  E  V  L

2881        TACCACTATCAGAAGAAGCAGAATTAGAACTGGCAGAAAACAGGGAAATTCTACAAGAAC
(Pol)         P  L  S  E  E  A  E  L  E  L  A  E  N  R  E  I  L  Q  E  P

2941        CAGTACATGGGGTGTATTATGCCCCATCAAAAGACTTAATAGCGGAAATACAGAAACAAG
(Pol)         V  H  G  V  Y  Y  A  P  S  K  D  L  I  A  E  I  Q  K  Q  G

3001        GGCAAGGACAATGGACATACCAAATTTATCAAGAACCATTTATACATCTGCAAACAGGAA
(Pol)         Q  G  Q  W  T  Y  Q  I  Y  Q  E  P  F  I  H  L  Q  T  G  K

3061        AGTATGCAAGAACGAGGGGTGCCCACACTATTCATGTACAACAATTATCAGAGGCAGTGC
(Pol)         Y  A  R  T  R  G  A  H  T  I  H  V  Q  Q  L  S  E  A  V  Q

3121        AAAAAATATCCACAGAAGGCATAGTGATATGGGGAAAGACTCCTAAATTTAGACTGCCCA
(Pol)         K  I  S  T  E  G  I  V  I  W  G  K  T  P  K  F  R  L  P  I

3181        TACAAAAGGAAACATGGGAAACATGGTGGATAGAGTATTGGCAAGCCACCTGGATTCCTG
(Pol)         Q  K  E  T  W  E  T  W  W  I  E  Y  W  Q  A  T  W  I  P  A

3241        CGTGGGAATTTGTCAATACCCCTCCTTTAGTAAAATTATGGTCCATTACAAAAGGACCCA
(Pol)         W  E  F  V  N  T  P  P  L  V  K  L  W  S  I  T  K  G  P  I

3301        TAATAGGAGCAGAAACTTTCTATGTAGATGGGGCAGCTAATAGAGAAACTAAAATAGGAA
(Pol)         I  G  A  E  T  F  Y  V  D  G  A  A  N  R  E  T  K  I  G  K

3361        AAGCAGGATATGTTACTGACAGGGGAAGACAGAAAGTTGTCCCTTTAACTGCCACAACAA
(Pol)         A  G  Y  V  T  D  R  G  R  Q  K  V  V  P  L  T  A  T  T  N

3421        ATCAGAAGACCGAGTTACAAGCAGTTTATCTAGCTTTGCAGGATTCGGGATTAGAAGTAA
(Pol)         Q  K  T  E  L  Q  A  V  Y  L  A  L  Q  D  S  G  L  E  V  N

3481        ACATAGTAACAGATTCACAATATGTATTGGGAATCATTCAAGCACAACCAGATCAAAGTC
(Pol)         I  V  T  D  S  Q  Y  V  L  G  I  I  Q  A  Q  P  D  Q  S  Q

3541        AATCAGAGTTAGTCAGTCAAATAATAGAGCAGCTAATAAAAAAGGAAAGGGTTTACCTGG
(Pol)         S  E  L  V  S  Q  I  I  E  Q  L  I  K  K  E  R  V  Y  L  A

3601        CATGGGTACCAGCACACAAAGGAATTGGAGGAAATGCACAAGTAGATAAGTTAGTCAGTC
(Pol)         W  V  P  A  H  K  G  I  G  G  N  A  Q  V  D  K  L  V  S  Q

3661        AGGGAATTCGAAAAGTACTATTTTTGGATGGAATAGATCAGGCTCAAGAAGAACATGCGA
(Pol)         G  I  R  K  V  L  F  L  D  G  I  D  Q  A  Q  E  E  H  A  K

3721        AATATCACAACAATTGGAGAGCAATGGCTACTGCTTTTATCCTACCACCTGTAGTAGCCA
(Pol)         Y  H  N  N  W  R  A  M  A  T  A  F  I  L  P  P  V  V  A  K

3781        AAGAAATACTATCTAGCTGTGATAAATGTCAGCTACAAGGAGAAGCCATGCATGGACAAG
(Pol)         E  I  L  S  S  C  D  K  C  Q  L  Q  G  E  A  M  H  G  Q  V

3841        TATACTGTAGTCCAGGAATATGGCAATTAGATTGTACACATCTAGAAGGAAAAGTTATCA
(Pol)         Y  C  S  P  G  I  W  Q  L  D  C  T  H  L  E  G  K  V  I  I

3901        TAGTAGCAGTTCATGTAGCCAGTGGCTATATAGAAGCAGAAGTTATTTCAGCAGAAACAG
(Pol)         V  A  V  H  V  A  S  G  Y  I  E  A  E  V  I  S  A  E  T  G

3961        GGCAGGAAACAGCATACTTTCTCTTAAAATTAGCAGGAAGATGGCCAGTAAAAGTAGTAC
(Pol)         Q  E  T  A  Y  F  L  L  K  L  A  G  R  W  P  V  K  V  V  H

4021        ATACAGACAATGGCAGAAATTTCACCAGTGCTGCAGTCAAGGCCGCCTGCTGGTGGGCAG
(Pol)         T  D  N  G  R  N  F  T  S  A  A  V  K  A  A  C  W  W  A  G
```

```
4081        GTATTTATCAGGAATTTGGAATTCCCTACAATCCCCAAAGTCAAGGAGTACTACAATCTA
(Pol)         I  Y  Q  E  F  G  I  P  Y  N  P  Q  S  Q  G  V  L  Q  S  M

4141        TGCATAAAGAATTACAGAAAATTATTGGACAGGTTACAGATCAAGCTGCACATCTTACGA
(Pol)         H  K  E  L  Q  K  I  I  G  Q  V  T  D  Q  A  A  H  L  T  T

4201        CAGCAGTACAAATGGCAGTATTCATCCACAATTTTACAAGAAAAGGGGGGATTGGGGGAT
(Pol)         A  V  Q  M  A  V  F  I  H  N  F  T  R  K  G  G  I  G  G  Y

4261        ACAGTGCAGGGGAAAGAATACTATACATATTACCAACAGACATACAAACTAAAGAATTAC
(Pol)         S  A  G  E  R  I  L  Y  I  L  P  T  D  I  Q  T  K  E  L  Q

4321        AAAAACAAATCACAAAAATTCAAAATTTTCGGGTTTATTACAGGGACAGCAGAGATCCAA
(Pol)         K  Q  I  T  K  I  Q  N  F  R  V  Y  Y  R  D  S  R  D  P  I

4381        TTTGGAAAGGACCAGCAAAACTTCTTTGGAAAGGTGCAGGGGCAGTATTATTACAAGACA
(Pol)         W  K  G  P  A  K  L  L  W  K  G  A  G  A  V  L  L  Q  D  N

4441        ATACTGTCATACAGGTTGTACCAAGAAGAAAAGTCAAATCATTACGGGACTATGGAAAAC
(Vif start)                                                   M  E  N
(Pol)         T  V  I  Q  V  V  P  R  R  K  V  K  S  L  R  D  Y  G  K  Q

4501        AGATGGCAGGTCATCATTGTGTGGCAAGCAGACAGGATGAGGATTAGCACATGGAAAAGT
(Vif)        R  W  Q  V  I  I  V  W  Q  A  D  R  M  R  I  S  T  W  K  S
(Pol end)    M  A  G  H  H  C  V  A  S  R  Q  D  E  D  *

4561        TTAGTAAAATACCATATGCATGTTTCAAAGAAGGCTAAAGGATGGTTTTATAGACATCAC
(Vif)        L  V  K  Y  H  M  H  V  S  K  K  A  K  G  W  F  Y  R  H  H

4621        TATGACAGCCCCCACCCAAAAATAAGTTCAGAAGTACACATTCCACTAGGAGAAGCTAGA
(Vif)        Y  D  S  P  H  P  K  I  S  S  E  V  H  I  P  L  G  E  A  R


4681        CTGGTAGTAAAAACATATTGGGGTCTGCATACAGGAGAAAGAGAATACCATCTGGGTCAG
(Vif)        L  V  V  K  T  Y  W  G  L  H  T  G  E  R  E  Y  H  L  G  Q

4741        GGAGTCTCCATACAATGGAGGAAAAGGAGATATAGCACACAAGTAGACCCTGGCCTGGCA
(Vif)        G  V  S  I  Q  W  R  K  R  R  Y  S  T  Q  V  D  P  G  L  A

4801        GACCAACTAATTCATATATATTATTTTGTTTGTTTTTCAGACTCTGCTATAAGAAAAGCC
(Vif)        D  Q  L  I  H  I  Y  Y  F  V  C  F  S  D  S  A  I  R  K  A

4861        ACATTAGGACATATAGTTAGCCCTACGTGTGAATATCAAGCAGGACATAACAAGGTCGGA
(Vif)        T  L  G  H  I  V  S  P  T  C  E  Y  Q  A  G  H  N  K  V  G

4921        TCCTTACAGTATTTGGCACTACCAGCATTATTACCACCAAAAAAGACAAAGCCACCCTTG
(Vif)        S  L  Q  Y  L  A  L  P  A  L  L  P  P  K  K  T  K  P  P  L

4981        CCTAGTGTTAGGAAGCTACCAGAAGATAGATGGAACAAGCCCCAGAAGACCAAGGGCCAC
(Vif)        P  S  V  R  K  L  P  E  D  R  W  N  K  P  Q  K  T  K  G  H
(Vpr start)                              M  E  Q  A  P  E  D  Q  G  P  Q

5041        AGCGGGAGCCATACAATGAATGGACATTAGAACTTTTGGAGGAGCTTATGAGTCAAGCTG
(Vif end)    S  G  S  H  T  M  N  G  H  *
(Vpr)          R  E  P  Y  N  E  W  T  L  E  L  L  E  E  L  M  S  Q  A  V

5101        TTAGACACTTTCCTACAATATGGCTCCAAAGCTTAGGACAATATATCTATGCAACTTATG
(Vpr)          R  H  F  P  T  I  W  L  Q  S  L  G  Q  Y  I  Y  A  T  Y  G

5161        GGGATACCTGGGCAGGAGTTCAAGCTTATTACAGAATTCTGCAACAACTACTGTTTATTC
(Vpr)          D  T  W  A  G  V  Q  A  Y  Y  R  I  L  Q  Q  L  L  F  I  H

5221        ATTTCAGAATTGGGTGTCAACATAGCAGAATAGGTATTACTCGCCAGAGAAGAGCAAGAA
(Vpr)          F  R  I  G  C  Q  H  S  R  I  G  I  T  R  Q  R  R  A  R  N
```

```
5281        ATGGATCCAGTAGATCCTAGCCTAGAGCCCTGGAACCATCCAGGAAGTCAGCCTAAGACT
(Tatx1)      M  D  P  V  D  P  S  L  E  P  W  N  H  P  G  S  Q  P  K  T
(Vpr end)       G  S  S  R  S  *

5341        GCTTGTAACAAATGTCATTGTAAAAAGTGTTGCTATCATTGCCAAGTTTGCTTCATAACG
(Tatx1)      A  C  N  K  C  H  C  K  K  C  C  Y  H  C  Q  V  C  F  I  T

5401        AAAGGCTTTGGCATCTCCTATGGCAGGAAGAAGCGGAGACAGCGACGAAAACCTTCTCAC
(Tatx1)      K  G  F  G  I  S  Y  G  R  K  K  R  R  Q  R  R  K  P  S  H
(Revx1 start)                  M  A  G  R  S  G  D  S  D  E  N  L  L  T

5461        GGCGATCAGGATCATCAAGTTCCTATACCAGAGCAGTAAGTAGTTTAATGTAATGCAACC
(Tatx1)      G  D  Q  D  H  Q  V  P  I  P  E  Q  *           (Vpu start)
(Revx1 end) A  I  R  I  I  K  F  L  Y  Q  S  S  K  *              M  Q  P

5521        TTTAGTGATAATAGCAATAGCAGCATTAGTAGTAGCACTAATAATAGCAATAGTTGTGTG
(Vpu)         L  V  I  I  A  I  A  A  L  V  V  A  L  I  I  A  I  V  V  W

5581        GACCATAGTATTCATAGAATATAGGAGAATAAAAAGGCAAAGAAAAATAGACTGTTTAAT
(Vpu)         T  I  V  F  I  E  Y  R  R  I  K  R  Q  R  K  I  D  C  L  I

5641        TGATAGAATAAGAGAAAGAGCAGAAGACAGTGGCAATGAGAGCGAGGGGGATGAAGAGGA
(Vpu)         D  R  I  R  E  R  A  E  D  S  G  N  E  S  E  G  D  E  E  E
(Env start)                                        M  R  A  R  G  M  K  R  N

5701        ATTGTCAAAACTTGTGGACAAGGGGCATCATGCTCCTTGGGATGTTGATGATCTGTAGTG
(Vpu end)     L  S  K  L  V  D  K  G  H  H  A  P  W  D  V  D  D  L  *
(Env)         C  Q  N  L  W  T  R  G  I  M  L  L  G  M  L  M  I  C  S  V

5761        TTGCAGAAAATTTGTGGGTCACAGTTTATTATGGGGTGCCTGTATGGAAGGAAGCAACCA
(Env)         A  E  N  L  W  V  T  V  Y  Y  G  V  P  V  W  K  E  A  T  T

5821        CCACTCTATTTTGTGCATCAGATGCTAAAGCATATAAAACAGAGGCACATAACATCTGGG
(Env)         T  L  F  C  A  S  D  A  K  A  Y  K  T  E  A  H  N  I  W  A

5881        CTACACATGCCTGTGTACCCACGGACCCCAGCCCACAAGAAATAGAACTGGAAAATGTGT
(Env)         T  H  A  C  V  P  T  D  P  S  P  Q  E  I  E  L  E  N  V  S

5941        CCGAAAACTTTAATATGTGGAAAAATAACGTGGTATACCAGATGCAGGAGGATATTATCA
(Env)         E  N  F  N  M  W  K  N  N  V  V  Y  Q  M  Q  E  D  I  I  S

6001        GTTTATGGGATGAAAGCCTACAACCATGTGCAAAATTAACCCCACTCTGTGTCACTTTAA
(Env)         L  W  D  E  S  L  Q  P  C  A  K  L  T  P  L  C  V  T  L  N

6061        ACTGCACTAATGCCATCTTACATAATGTCACCTCAAACAGCATTGTGGAGCCAAAACTGG
(Env)         C  T  N  A  I  L  H  N  V  T  S  N  S  I  V  E  P  K  L  E

6121        AGGTGAAAAACTGCTCTTTCAGGAAAACTACAGAAGGAAGAGAGAAGAAAAAGAAAGCAA
(Env)         V  K  N  C  S  F  R  K  T  T  E  G  R  E  K  K  K  K  A  N

6181        ATGCGCTTTTTTATAGACCTGATATACTACCAACAAACAATGATAATAGTAGTACTAATT
(Env)         A  L  F  Y  R  P  D  I  L  P  T  N  N  D  N  S  S  T  N  Y

6241        ATACCAAGTATAGGTTATTATATTGTAATACCTCAGCCATTACACAGGCTTGTCCAAAGG
(Env)         T  K  Y  R  L  L  Y  C  N  T  S  A  I  T  Q  A  C  P  K  V

6301        TATCCTTTGAACCAATTCCAATACATTATTGTGCCCCAGCTGGTTTTGCAATTCTCAAGT
(Env)         S  F  E  P  I  P  I  H  Y  C  A  P  A  G  F  A  I  L  K  C

6361        GTAGAGATAAGAAGTTCAATGGAACAGGCCCATGCACAGATGTCAGCACAATACAATGTA
(Env)         R  D  K  K  F  N  G  T  G  P  C  T  D  V  S  T  I  Q  C  T

6421        CACATGGAATTAAGCCAGTGGTGTCAACTCAACTGCTGTTCAATGGCAGTCTCGCAGAAG
(Env)         H  G  I  K  P  V  V  S  T  Q  L  L  F  N  G  S  L  A  E  E

6481        AAGAGATCATCATTAGATCTGAAAATCTCACAAACAATGCTAAAAACATATTATTACAGT
(Env)         E  I  I  I  R  S  E  N  L  T  N  N  A  K  N  I  L  L  Q  F
```

```
6541        TTAATGCATCTGAAGAAATTAATTGTACAAGGCCCTACCAATATGCAAGACAAAAGACAT
(Env)         N  A  S  E  E  I  N  C  T  R  P  Y  Q  Y  A  R  Q  K  T  S

6601        CAAGAGGACAAGGGCAAACACTCTATACAAGCAAGAAGATTATTGGAGACATAAGACAAG
(Env)         R  G  Q  G  Q  T  L  Y  T  S  K  K  I  I  G  D  I  R  Q  A

6661        CATATTGTAACATTAGTGGAGAAAAATGGAATAAAACTTTACAACAGGGAGCTATACAAT
(Env)         Y  C  N  I  S  G  E  K  W  N  K  T  L  Q  Q  G  A  I  Q  S

6721        CAGGAAAACTTCTTAACAAAACAACAATATTTTTTCAACCACCCTCAGGAGGGGACTCAG
(Env)         G  K  L  L  N  K  T  T  I  F  F  Q  P  P  S  G  G  D  S  E

6781        AAATTACAACACACAGTTTTAATTGTGGAGGGGAATTTTTCTACTGTAATACATCAAGGC
(Env)         I  T  T  H  S  F  N  C  G  G  E  F  F  Y  C  N  T  S  R  L

6841        TGTTTAGTAATACATGGATGGTACATGGGATAATAATACATGGTTCAAATCAGACAGTCA
(Env)         F  S  N  T  W  M  V  H  G  I  I  I  H  G  S  N  Q  T  V  R

6901        GACTCCCATGCAGAATAAAACAAATTATAAACATGTGGCAGGAAGTAGGAAAAGCAATGT
(Env)         L  P  C  R  I  K  Q  I  I  N  M  W  Q  E  V  G  K  A  M  Y

6961        ATGCCCCTCCCATAGAAGGAACAATTAGGTGTTCATCAAATATTACAGGGCTAATATTGA
(Env)         A  P  P  I  E  G  T  I  R  C  S  S  N  I  T  G  L  I  L  T
7021        CAAGAGATGGTGGTAATAATAGTTCTAACAACGAGACCTTCAGACCTGGCGGAGGAGATA
(Env)         R  D  G  G  N  N  S  S  N  N  E  T  F  R  P  G  G  G  D  M

7081        TGAGGGACAATTGGAGAAGTGAATTATATAAATACAAAGTAATACAAATTGAACCAATAG
(Env)         R  D  N  W  R  S  E  L  Y  K  Y  K  V  I  Q  I  E  P  I  G

7141        GAGTAGCGCCCACCAAGGCAAAGAGAAGAGTGGTGGAAAGGGAAAAAAGAGCAATAGGAC
(Env)         V  A  P  T  K  A  K  R  R  V  V  E  R  E  K  R  A  I  G  L

7201        TAGGAGCTATGTTCCTTGGGTTCTTGGGAGCAGCAGGAAGCACAATGGGCGCAGCGTCAG
(Env)         G  A  M  F  L  G  F  L  G  A  A  G  S  T  M  G  A  A  S  V

7261        TGACGCTGACGGTACAGGCCAGACAGGTATTGTCTGGTAGAGTGCAACAGCAAAACAATT
(Env)         T  L  T  V  Q  A  R  Q  V  L  S  G  R  V  Q  Q  Q  N  N  L

7321        TGGCCAGGGCTATAGAGGCGCAACAGCATCTGTTGCAACTCACGGTCTGGGGCATTAAAC
(Env)         A  R  A  I  E  A  Q  Q  H  L  L  Q  L  T  V  W  G  I  K  Q

7381        AGCTCCAGGCAAGAATCCTGGCTGTGGAAAGATACCTAAAGGATCAACGGCTCCTAGGCA
(Env)         L  Q  A  R  I  L  A  V  E  R  Y  L  K  D  Q  R  L  L  G  I

7441        TTACGGGTTGCTCTGGAAAACATATTTGCACCACTAATGTGCCCTGGAACTCTTCTTGGA
(Env)         T  G  C  S  G  K  H  I  C  T  T  N  V  P  W  N  S  S  W  S

7501        GTAATAAATCCTTAGATGAGATTTGGCAAAACTTGCCCTGGAAGAAAGTGGGAAGAGAAA
(Env)         N  K  S  L  D  E  I  W  Q  N  L  P  W  K  K  V  G  R  E  I

7561        TCGACAATTACACAGGACTAATATACAACTTAATTGAAGAATCGCAGATCCAGCAGGAGA
(Env)         D  N  Y  T  G  L  I  Y  N  L  I  E  E  S  Q  I  Q  Q  E  K

7621        AGAATAAGACAGAATTATTGGAATTGGACAAGTGGGCAAGCCTGTGGAATTGGTTTGACA
(Env)         N  K  T  E  L  L  E  L  D  K  W  A  S  L  W  N  W  F  D  I

7681        TAACAAACTGGCTGTGGTATATAAAAATATTCATAATGATTGTAGGAGGCTTAATAGGTT
(Env)         T  N  W  L  W  Y  I  K  I  F  I  M  I  V  G  G  L  I  G  L

7741        TAAGAATACTTTTTGCTGTGCTTTCTGTAGTAAACAGAGTTTGGCAGGGATACTCACCTC
(Env)         R  I  L  F  A  V  L  S  V  V  N  R  V  W  Q  G  Y  S  P  L

7801        TGTCATTTCAGACCCTCCTCCCAGCCCCGAGGGGACCCGACAGGCCCGAAGGAACAGAAG
(Tatx2 start)              P  S  S  Q  P  R  G  D  P  T  G  P  K  E  Q  K
(Revx2 start)        S  D  P  P  P  S  P  E  G  T  R  Q  A  R  R  N  R  R
(Env)         S  F  Q  T  L  L  P  A  P  R  G  P  D  R  P  E  G  T  E  E
```

```
7861        AAGAAGGTGGAGAGCGAGGCGGAGACAGATCCATTCGATTGATGAACGGATTCTCAGCCT
(Tatx2 end)K  K  V  E  S  E  A  E  T  D  P  F  D  *
(Revx2)      R  R  W  R  A  R  R  R  Q  I  H  S  I  D  E  R  I  L  S  L
(Env)         E  G  G  E  R  G  G  D  R  S  I  R  L  M  N  G  F  S  A  L

7921        TATTCTGGGACGATCTGCGGAACCTGTGCCTCTTCAGCTACCACCGCTTGAGAGACTTAC
(Revx2)      I  L  G  R  S  A  E  P  V  P  L  Q  L  P  P  L  E  R  L  T
(Env)         F  W  D  D  L  R  N  L  C  L  F  S  Y  H  R  L  R  D  L  L

7981        TCTTGATTGCAGCGAGGATTGTGGAACTTCTGGGACGCCGGGGGTGGGAAGCCCTCAAGT
(Revx2)      L  D  C  S  E  D  C  G  T  S  G  T  P  G  V  G  S  P  Q  V
(Env)         L  I  A  A  R  I  V  E  L  L  G  R  R  G  W  E  A  L  K  Y


8041        ATCTGTGGAATTTCCTGCAGTATTGGAGTCAGGAACTCAGGAATAGTGCTTCTTCCTTGC
(Revx2 end) S  V  E  F  P  A  V  L  E  S  G  T  Q  E  *
(Env)         L  W  N  F  L  Q  Y  W  S  Q  E  L  R  N  S  A  S  S  L  L

8101        TTGCTACCATAGCAATAGCAACAGCTGCGGGGACAGAAAGGGTTATAGAAGTAGTACTAA
(Env)         A  T  I  A  I  A  T  A  A  G  T  E  R  V  I  E  V  V  L  R

8161        GAGCTTGCAGAGCTCTTAACATACCCACAAGAATAAGACAGGGCTTGGAAAGGCTTTTGC
(Env)         A  C  R  A  L  N  I  P  T  R  I  R  Q  G  L  E  R  L  L  L

8221        TATAAAATGGGTGGCAAATGGTCAAAAAGTAGTATAGTTGGATGGCCTGCTATAAGGGAA
(Nef start)      M  G  G  K  W  S  K  S  S  I  V  G  W  P  A  I  R  E
(Env end)     *

8281        AGAATAAGAAGAACTGATCCAGCAGCAGATGGGGTGGGAGCAGTATCTCGAGACCTGGAA
(Nef)        R  I  R  R  T  D  P  A  A  D  G  V  G  A  V  S  R  D  L  E

8341        AGACATGGGGCAATCACAAGTAGTAATACAGCAAGTACTAATGCTGACCTTGCCTGGCTA
(Nef)        R  H  G  A  I  T  S  S  N  T  A  S  T  N  A  D  L  A  W  L

8401        GAAGCACAAGAGAAAGGTGAGGAGGTGGGCTTTCCAGTCAGACCTCAGGTACCTTTAAGA
(Nef)        E  A  Q  E  K  G  E  E  V  G  F  P  V  R  P  Q  V  P  L  R

8461        CCAATGACTTTCAAAGGAGCTGTAGATCTTAGCCACTTTTTAAAAGAAAAGGGGGGACTG
(Nef)        P  M  T  F  K  G  A  V  D  L  S  H  F  L  K  E  K  G  G  L

8521        GATGGGATAATTTGGTCCAAAAGGAGACAAGAGATCCTTGATCTTTGGGTCTACAACACA
(Nef)        D  G  I  I  W  S  K  R  R  Q  E  I  L  D  L  W  V  Y  N  T

8581        CAAGGCTACTTCCCTGATTGGCAGAACTACACACCAGGGCCAGGGACCAGATATCCACTG
(Nef)        Q  G  Y  F  P  D  W  Q  N  Y  T  P  G  P  G  T  R  Y  P  L

8641        ACCTTTGGATGGTGCTTCGAGCTAGTACCAGTTGATCCACAGGAGGTAGAAGAGGCCACT
(Nef)        T  F  G  W  C  F  E  L  V  P  V  D  P  Q  E  V  E  E  A  T

8701        GGGGGAGAGACCAACTGCTTGTTACACCCTATGAACCAGCATGGAATGGATGACCCGGAG
(Nef)        G  G  E  T  N  C  L  L  H  P  M  N  Q  H  G  M  D  D  P  E

8761        AGACAAGTGCTAAAGTGGAGATTTAACAGCAGACTAGCATTTGAGCACAAGGCCCGACAG
(Nef)        R  Q  V  L  K  W  R  F  N  S  R  L  A  F  E  H  K  A  R  Q

8821        CTACATCCGGAGTACTACAAAGACTGCTGA
(Nef end)    L  H  P  E  Y  Y  K  D  C  *
```

B4: pR482 Full-length sequence (nucleotide and amino acid)

```
181         ACTGGTGAGTACGCTAAAATTTTTGACTAGCGGAGGCTAGAAGGAGAGAGATGGGTGCGA
(Gag start)                                                   M  G  A  R

241         GAGCGTCAGTATTAAGCGGGGGAAAATTAGATGCATGGGAAAGAATTCGGTTAAGGCCAG
(Gag)         A  S  V  L  S  G  G  K  L  D  A  W  E  R  I  R  L  R  P  G

301         GAGGAAAGAAAAAATATCAACTAAAGCATATAGTATGGGCAAGCAGGGAGCTAGAACGAT
(Gag)         G  K  K  K  Y  Q  L  K  H  I  V  W  A  S  R  E  L  E  R  F

361         TTGCACTTAACCCTGGCCTTTTAGAAACATCAGAAGGCTGTAAACAAATAATAGAACAGC
(Gag)         A  L  N  P  G  L  L  E  T  S  E  G  C  K  Q  I  I  E  Q  L

421         TACAGCCATCCATTCAGACAGGATCAGAAGAACTTAAATCATTATATAATACAGTAGCAA
(Gag)         Q  P  S  I  Q  T  G  S  E  E  L  K  S  L  Y  N  T  V  A  T

481         CCCTCTATTGTGTACATGAAAGGATAGATGTAAAAGACACCAAGGAAGCTTTAGAAAAAA
(Gag)         L  Y  C  V  H  E  R  I  D  V  K  D  T  K  E  A  L  E  K  I

541         TAGAGGAAGAACAAAACAAAAGTAAGAAAAAGAAGGCACAGCAAGCAGAGGCTGACACAG
(Gag)         E  E  E  Q  N  K  S  K  K  K  K  A  Q  Q  A  E  A  D  T  G

601         GGAACAGCAGTCAGGTCAGCCAAAATTATCCTATAGTGCAGAACCTACAGGGGCAAATGG
(Gag)         N  S  S  Q  V  S  Q  N  Y  P  I  V  Q  N  L  Q  G  Q  M  V

661         TACATCAGGCCATATCACCTAGAACTTTGAATGCATGGGTAAAAGTAATAGAAGAAAAGG
(Gag)         H  Q  A  I  S  P  R  T  L  N  A  W  V  K  V  I  E  E  K  A

721         CCTTCAGCCCAGAAATAATACCCATGTTTTCAGCATTATCAGAAGGAGCCACCCCACAAG
(Gag)         F  S  P  E  I  I  P  M  F  S  A  L  S  E  G  A  T  P  Q  D

781         ATTTAAACACCATGCTAAACACAGTGGGGGGACATCAAGCAGCCATGCAAATGCTAAAAG
(Gag)         L  N  T  M  L  N  T  V  G  G  H  Q  A  A  M  Q  M  L  K  E

841         AGACCATCAATGAGGAAGCTGCAGACTGGGATAGGCTACATCCAGTGCATGTAGGGCCTA
(Gag)         T  I  N  E  E  A  A  D  W  D  R  L  H  P  V  H  V  G  P  I

901         TTGCACCAGGCCAGATGAGAGAACCAAGGGGAAGTGATATAGCAGGAACTACTAGTACCC
(Gag)         A  P  G  Q  M  R  E  P  R  G  S  D  I  A  G  T  T  S  T  L

961         TTCAGGAACAAATAGCATGGATGACAAGTAACCCATCTGTCCCAGTAGGAGAAATCTATA
(Gag)         Q  E  Q  I  A  W  M  T  S  N  P  S  V  P  V  G  E  I  Y  K

1021        AAAGATGGATAATCCTGGGATTAAATAAAATTGTAAGAATGTATAGCCCTGTCAGCATTT
(Gag)         R  W  I  I  L  G  L  N  K  I  V  R  M  Y  S  P  V  S  I  L

1081        TGGACATAAGACAGGGACCAAAGGAACCTTTTAGAGATTATGTAGACCGGTTCTATAAAA
(Gag)         D  I  R  Q  G  P  K  E  P  F  R  D  Y  V  D  R  F  Y  K  T

1141        CTCTAAGAGCCGAGCAAGCTTCACAGGATGTAAAAAACTGGATGACAGAAACCTTGTTGG
(Gag)         L  R  A  E  Q  A  S  Q  D  V  K  N  W  M  T  E  T  L  L  V

1201        TCCAAAATGCAAACCCAGGTTGTAAAACCATCTTAAAAGCATTAGGACCACAGGCTACAC
(Gag)         Q  N  A  N  P  G  C  K  T  I  L  K  A  L  G  P  Q  A  T  L

1261        TAGAAGAAATGATGACAGCATGTCAGGGAGTGGGGGGGCCCGGCCATAAAGCAAGAGTTT
(Gag)         E  E  M  M  T  A  C  Q  G  V  G  G  P  G  H  K  A  R  V  L
1321        TGGCTGAGGCAATGAGCCAAGCAACAAATTTAGCTACTGCAGTAATGATGCAGAGAGGCA
(Gag)         A  E  A  M  S  Q  A  T  N  L  A  T  A  V  M  M  Q  R  G  N

1381        ATTTTAAGGGCCAAAGAAGAATTATTAAGTGTTTCAACTGTGGCAAAGAAGGGCACGTAG
(Gag)         F  K  G  Q  R  R  I  I  K  C  F  N  C  G  K  E  G  H  V  A
```

```
1441        CAAAAAATTGCAGGGCCCCTAAAAAAAAGGGCTGTTGGAAATGTGGAAGGGAAGGACACC
(Gag)         K  N  C  R  A  P  K  K  K  G  C  W  K  C  G  R  E  G  H  Q

1501        AAATGAAAGATTGCACTGAAAGACAGGCTAATTTTTTACGGAAGATTTGGCCTTCCCACA
(Pol start)                                   F  F  T  E  D  L  A  F  P  Q
(Gag)         M  K  D  C  T  E  R  Q  A  N  F  L  R  K  I  W  P  S  H  K

1561        AGGGAAGGCCGGGGAATTTTCTTCAGAGCAGACCAGAGCCAACAGCCCCACCAGAAGAGA
(Pol)         G  K  A  G  E  F  S  S  E  Q  T  R  A  N  S  P  T  R  R  E
(Gag)         G  R  P  G  N  F  L  Q  S  R  P  E  P  T  A  P  P  E  E  S

1621        GCGTCGGGTTTGGGGTGGAGACAACCCCCTCTCAGAAACAGGAACCCATAGACAAGGAAC
(Pol)         R  R  V  W  G  G  D  N  P  L  S  E  T  G  T  H  R  Q  G  T
(Gag)         V  G  F  G  V  E  T  T  P  S  Q  K  Q  E  P  I  D  K  E  L

1681        TGTATCCTTTATCCTCCCTCAAATCACTCTTTGGGAGCGACCCCTTGTCACAATAAAGAT
(Pol)         V  S  F  I  L  P  Q  I  T  L  W  E  R  P  L  V  T  I  K  I
(Gag end)     Y  P  L  S  S  L  K  S  L  F  G  S  D  P  L  S  Q  *

1741        AGGGGGACAGCTAAAGGAAGCTCTATTAGATACAGGAGCAGATGATACAGTATTAGAAGA
(Pol)         G  G  Q  L  K  E  A  L  L  D  T  G  A  D  D  T  V  L  E  E

1801        AATGAATTTGCCAGGAAAATGGAAACCAAAAATGATAGGGGGAATTGGGGGTTTTATCAA
(Pol)         M  N  L  P  G  K  W  K  P  K  M  I  G  G  I  G  G  F  I  K

1861        AGTAAGACAGTATGATCAAATACCCTTAGAAATCTGTGGGCATAAAGCTATAGGTACAGT
(Pol)         V  R  Q  Y  D  Q  I  P  L  E  I  C  G  H  K  A  I  G  T  V

1921        ATTAATAGGACCTACACCTGTCAACATAATTGGAAGAAATTTGTTGACTCAGCTTGGCTG
(Pol)         L  I  G  P  T  P  V  N  I  I  G  R  N  L  L  T  Q  L  G  C

1981        CACTTTAAATTTCCCAATTAGTCCTATTGAAACTGTACCAGTAAAATTAAAGCCAGGAAT
(Pol)         T  L  N  F  P  I  S  P  I  E  T  V  P  V  K  L  K  P  G  M

2041        GGATGGCCCAAAAGTTAAACAATGGCCATTGACAGAAGAAAAAATAAAAGCATTAACAGA
(Pol)         D  G  P  K  V  K  Q  W  P  L  T  E  E  K  I  K  A  L  T  E

2101        AATTTGTCTAGAAATGGAAAAGGAAGAAAAAATTTCAAGAATTGGGCCTGAAAATCCATA
(Pol)         I  C  L  E  M  E  K  E  E  K  I  S  R  I  G  P  E  N  P  Y

2161        CAATACTCCAATATTTGCCATAAAGAAAAAAGACAGTACTAAATGCAGAAAATTAGTAGA
(Pol)         N  T  P  I  F  A  I  K  K  K  D  S  T  K  C  R  K  L  V  D

2221        TTTCAGAGAACTTAATAAGAGAACTCAAGATTTCTGGGAAGTACAATTAGGAATACCGCA
(Pol)         F  R  E  L  N  K  R  T  Q  D  F  W  E  V  Q  L  G  I  P  H

2281        CCCTGCAGGGCTGAAAAAAAAAAAAATCAGTAACAGTACTGGATGTGGGTGATGCATATTT
(Pol)         P  A  G  L  K  K  K  K  S  V  T  V  L  D  V  G  D  A  Y  F

2341        TTCAGTTCCCGTATGTGAAGACTTTAGGAAATATACCGCATTTACCATACCTAGTACAAA
(Pol)         S  V  P  V  C  E  D  F  R  K  Y  T  A  F  T  I  P  S  T  N

2401        CAATGAGACACCAGGGATTATATATCAGTACAATGTGCTTCCACAGGGATGGAAAGGATC
(Pol)         N  E  T  P  G  I  I  Y  Q  Y  N  V  L  P  Q  G  W  K  G  S

2461        ACCGGCAATATTCCAATCAAGCATGACAAAAATCTTAGAGCCCTTTCAAAAACAAAATCC
(Pol)         P  A  I  F  Q  S  S  M  T  K  I  L  E  P  F  Q  K  Q  N  P

2521        AGATATAGTTATCTATCAATACATGGAAGATTTGTATGTAGGATCCGATTTAGAAATAGG
(Pol)         D  I  V  I  Y  Q  Y  M  E  D  L  Y  V  G  S  D  L  E  I  G

2581        GCAGCATCGAACAAAAATAGAGGAATTAAGAGAACATCTATTGAGATGGGGATTTACTAC
(Pol)         Q  H  R  T  K  I  E  E  L  R  E  H  L  L  R  W  G  F  T  T

2641        ACCAGATCAAAAACATCAGAAAGAACCTCCATTTCTTTGGATGGGTTATGAACTCCATCC
(Pol)         P  D  Q  K  H  Q  K  E  P  P  F  L  W  M  G  Y  E  L  H  P
```

```
2701        TGATAAATGGACAGTACAGCCTATAGTACTGCCAGAAAAAGAAAACTGGACTGTCAATGA
(Pol)         D  K  W  T  V  Q  P  I  V  L  P  E  K  E  N  W  T  V  N  D

2761        TATACAGAAGTTAGTAGGGAAATTAAACTGGGCAAGCCAGATTTATCCAGGAATTAAAGT
(Pol)         I  Q  K  L  V  G  K  L  N  W  A  S  Q  I  Y  P  G  I  K  V

2821        AAAGCAATTATGTAAACTCCTTAGGGGAACCAAAGCACTAACAGAAGTAATACCACTAAC
(Pol)         K  Q  L  C  K  L  L  R  G  T  K  A  L  T  E  V  I  P  L  T

2881        AGCAGAAGCAGAATTAGAACTGGCAGAAAACAGGGAAATTCTAAAAGAACCAGTACATGG
(Pol)         A  E  A  E  L  E  L  A  E  N  R  E  I  L  K  E  P  V  H  G

2941        AGTGTATTATGACCCATCAAAAGACTTAATAGCAGAAATACAGAAACAAGGGAATGGCCA
(Pol)         V  Y  Y  D  P  S  K  D  L  I  A  E  I  Q  K  Q  G  N  G  Q

3001        ATGGACATATCAAATTTATCAAGAACCATTTAAAAATCTGAAAACAGGAAAGTATGCAAG
(Pol)         W  T  Y  Q  I  Y  Q  E  P  F  K  N  L  K  T  G  K  Y  A  R

3061        AACGAGGGGTGCCCACACTAATGATGTAAAACAATTAGCAGAGGCAGTGCAAAAAATAGC
(Pol)         T  R  G  A  H  T  N  D  V  K  Q  L  A  E  A  V  Q  K  I  A

3121        CACAGAAGGCATAGTGATATGGGGAAAGACTCCTAAATTTAGACTGCCCATACAAAAGGA
(Pol)         T  E  G  I  V  I  W  G  K  T  P  K  F  R  L  P  I  Q  K  E

3181        AACATGGGAAACATGGTGGATAGAGTATTGGCAAGCCACCTGGATTCCAGAGTGGGAATT
(Pol)         T  W  E  T  W  W  I  E  Y  W  Q  A  T  W  I  P  E  W  E  F

3241        TGTCAATACCCCTCCCTTAGTAAAATTATGGTACCAATTAGAGAGGGAACCCATAGTAGG
(Pol)         V  N  T  P  P  L  V  K  L  W  Y  Q  L  E  R  E  P  I  V  G

3301        AGCAGAAACTTTCTATGTAGATGGGGCAGCTAATACAGAAACCAGACTACAAAAAGCAGG
(Pol)         A  E  T  F  Y  V  D  G  A  A  N  T  E  T  R  L  Q  K  A  G

3361        ATATGTTACTTACAGAGGAAGACAGAAAGTTGTCCCTTTAACTGCCACAACAAATCAGAA
(Pol)         Y  V  T  Y  R  G  R  Q  K  V  V  P  L  T  A  T  T  N  Q  K

3421        GACTGCATTACAAGCAGTTATTCTAGCTTTGCAAGATTCGGGATTAGAAGTAAACATAGT
(Pol)         T  A  L  Q  A  V  I  L  A  L  Q  D  S  G  L  E  V  N  I  V

3481        AACAGATTCACAATATGTATTAGGAATCATTCAAGCACAACCAGAGAAGAGTCAATCAGA
(Pol)         T  D  S  Q  Y  V  L  G  I  I  Q  A  Q  P  E  K  S  Q  S  E

3541        GTTAGTCAGTCAAATAATAGAGCAGCTAATAAAAAAGGAAAAGGTTTACCTGGCATGGGT
(Pol)         L  V  S  Q  I  I  E  Q  L  I  K  K  E  K  V  Y  L  A  W  V

3601        ACCAGCACACAAAGGAATTGGAGGAAATGTACAAGTAGATATATTAGTCAGTCAGGGAAT
(Pol)         P  A  H  K  G  I  G  G  N  V  Q  V  D  I  L  V  S  Q  G  I

3661        CAGGAAAGTACTATTTTTGGATGGAATAGATATGGCTCAAAAAGAACATGTGAAATATCA
(Pol)         R  K  V  L  F  L  D  G  I  D  M  A  Q  K  E  H  V  K  Y  H

3721        CAACAATTGGAGAGCAATGGCTATTGCTTTTACCCTACCACCTGTGGTAGCAAAAAAAAT
(Pol)         N  N  W  R  A  M  A  I  A  F  T  L  P  P  V  V  A  K  K  I

3781        AGTAGCAAGCTGCGATATATGTCAGCTAAAAGGAGAAGCCATGCATGGACAAGTAGACTG
(Pol)         V  A  S  C  D  I  C  Q  L  K  G  E  A  M  H  G  Q  V  D  C

3841        TTGTCCAGGAATATGGCAATTAGATTGTACACATTTAGAAGGAAAAGTTATCATAGTAGC
(Pol)         C  P  G  I  W  Q  L  D  C  T  H  L  E  G  K  V  I  I  V  A

3901        AGTTCATGTAGCTACTGGCTATATAGAAGCAGAAGTTATTTCAGCAGAAACAGGGCAGGA
(Pol)         V  H  V  A  T  G  Y  I  E  A  E  V  I  S  A  E  T  G  Q  E

3961        AACAGCATACTTTCTCTTAAAATTAGCAGGAAGATGGCCAGTAAAAGTAGTACATACAGA
(Pol)         T  A  Y  F  L  L  K  L  A  G  R  W  P  V  K  V  V  H  T  D

4021        CAATGGCAGCAACTTCACCAGTGCTGCAGTCAAGGCCGCCTGCTGGTGGGCAGGTATCAA
(Pol)         N  G  S  N  F  T  S  A  A  V  K  A  A  C  W  W  A  G  I  K
```

```
4081        ACAGGAATTTGGAATTCCCTACAATCCCCAAAGTCAAGGAGTAGTAGAATCTATAAATAC
(Pol)         Q  E  F  G  I  P  Y  N  P  Q  S  Q  G  V  V  E  S  I  N  T

4141        AAAATTAAAGAAAATTATAGGACAGGTAAGAGACCAAGCTGAACATCTTAAGACAGCAGT
(Pol)         K  L  K  K  I  I  G  Q  V  R  D  Q  A  E  H  L  K  T  A  V

4201        ACAAATGGCAGTATTCATCCACAATTTTAAAAGAAAAGGGGGGATTGGGGGTTACAGTGC
(Pol)         Q  M  A  V  F  I  H  N  F  K  R  K  G  G  I  G  G  Y  S  A

4261        AGGGGAAAGAATACTACACATACTATCAACAGACATACAAACTAAAGAATTACAAAAACA
(Pol)         G  E  R  I  L  H  I  L  S  T  D  I  Q  T  K  E  L  Q  K  Q

4321        AATTACAAAAATTCAAAATTTTCGGGTTTATTACAGGGACAGCAGAGATCCAATTTGGAA
(Pol)         I  T  K  I  Q  N  F  R  V  Y  Y  R  D  S  R  D  P  I  W  K

4381        AGGACCAGCAAAACTTCTCTGGAAAGGTTACGGGGCAGTAGTAATACAAGACAATACTGC
(Pol)         G  P  A  K  L  L  W  K  G  Y  G  A  V  V  I  Q  D  N  T  A

4441        CATAAAGGTAGTACCAAGAAGCAAAGTGAAAATCATTACGGATTATGGAAAACAGATGGC
(Pol)         I  K  V  V  P  R  S  K  V  K  I  I  T  D  Y  G  K  Q  M  A
(Vif start)                                                  M  E  N  R  W  Q

4501        AGGTGTTGTTTGTGTGGCAAGTATACAGGATGGGGATTAACACATGGAAAAGCCTTGTAA
(Pol end)     G  V  V  C  V  A  S  I  Q  D  G  D  *
(Vif)           V  L  F  V  W  Q  V  Y  R  M  G  I  N  T  W  K  S  L  V  K

4561        AAAAATACCATATGCATGTTTCAAAGAAAGCTAATCGATGGTTTTATAAACATCACTATG
(Vif)         K  Y  H  M  H  V  S  K  K  A  N  R  W  F  Y  K  H  H  Y  D

4621        ACAGCCCCCACCCAAAAATAAGTTCAGAAGTGCACATTCCACTAGGAGAAGCTAGACTGG
(Vif)         S  P  H  P  K  I  S  S  E  V  H  I  P  L  G  E  A  R  L  V

4681        TAGTAAAAACATATTGGGGTCTGCATACAGGAGAAAAGGAATGGCATCTGGGTCAGGGAG
(Vif)         V  K  T  Y  W  G  L  H  T  G  E  K  E  W  H  L  G  Q  G  V

4741        TCTCCATAGAACCCTGGAGGAAAAGGAGATATACCACACAAGTAGACCCAGGCCTGGCAG
(Vif)         S  I  E  P  W  R  K  R  R  Y  T  T  Q  V  D  P  G  L  A  D

4801        ACCAACTAATTCATATATATTATTTTGATTGTTTTTCAGACTCTGCTATAAGAAAAGCCA
(Vif)         Q  L  I  H  I  Y  Y  F  D  C  F  S  D  S  A  I  R  K  A  I

4861        TATTAGGACATATAGTTAGACCTAGGTGTGAATATCAAGCAGGACATAACCAGGTAGGAT
(Vif)         L  G  H  I  V  R  P  R  C  E  Y  Q  A  G  H  N  Q  V  G  S

4921        CCTTACAGTATTTGGCACTAACAGCATTAATAGCACCAAAAAGGACAAAGCCACCTTTAC
(Vif)         L  Q  Y  L  A  L  T  A  L  I  A  P  K  R  T  K  P  P  L  P

4981        CTAGTGTTAGGAAGCTAACAGAAGACAGATGGAACAAGCCCCAGAAGAACAAGGGCCACA
(Vpr start)                                 M  E  Q  A  P  E  E  Q  G  P  Q
(Vif)         S  V  R  K  L  T  E  D  R  W  N  K  P  Q  K  N  K  G  H  R

5041        GAGGAAGCCACACAACGAATGGACATTAGAACTTTTGGAAGAGCTTCAGAAGGAAGCTGT
(Vpr)         R  K  P  H  N  E  W  T  L  E  L  L  E  E  L  Q  K  E  A  V
(Vif end)     G  S  H  T  T  N  G  H  *

5101        TACACACTTTCCAAGCATATGGCTCCTCAGCTTAGGACACTATATCGAAACTTATGGGGA
(Vpr)         T  H  F  P  S  I  W  L  L  S  L  G  H  Y  I  E  T  Y  G  D

5161        TACCAGGGCAGGAGTCGAAGCTATAAGAATTCTGCAACAACTACTGTTTATTCATTTCAG
(Vpr)         T  R  A  G  V  E  A  I  R  I  L  Q  Q  L  L  F  I  H  F  R

5221        AATTGGGTGTCAACATACCAGAATAGGTATTACTCGACAGAGAAGAGCAAGAAATGGATC
(Vpr)         I  G  C  Q  H  T  R  I  G  I  T  R  Q  R  R  A  R  N  G  S
(Tatx1 start)                                                  M  D  P
```

```
5281       CAGTACATCCTAGCCTACAGCCCTGGAACCATCCAGGAAGTCAGCCTAAGACTGCTTGTA
(Vpr end)    S  T  S  *
(Tatx1)       V  H  P  S  L  Q  P  W  N  H  P  G  S  Q  P  K  T  A  C  N

5341       ACAAATGTCATTGTAAAAAGTGTTGCTATCATTGCCAAGTTTGCTTCATCACGAAAGGCT
(Tatx1)       K  C  H  C  K  K  C  C  Y  H  C  Q  V  C  F  I  T  K  G  F

5401       TCGGCATCTCCTATGGCAGGAAGAAGCGGAGACAGCGACGAAGATCTCCTCAAGGCGATC
(Revx1 start)          M  A  G  R  S  G  D  S  D  E  D  L  L  K  A  I
(Tatx1)       G  I  S  Y  G  R  K  K  R  R  Q  R  R  R  S  P  Q  G  D  Q

5461       AGGCTCATCAAGTTCCTATACCAGAGCAGTAAGTAGTTCATGAAATGCAACCTTTACAGA
(Revx1 end)R  L  I  K  F  L  Y  Q  S  S  K  *          (Vpu start)
(Tatx1 end)  A  H  Q  V  P  I  P  E  Q  *             M  Q  P  L  Q  I

5521       TATTATCAATATTAGCATTAGTAGTAGCAGCAATACTAGCAATAGTTGTGTACACCATAG
(Vpu)         L  S  I  L  A  L  V  V  A  A  I  L  A  I  V  V  Y  T  I  V

5581       TATTCATAGAATATAGGAAAATAAAAAGGCAAAGAACAATAGACTGTTTAATTGATAGAA
(Vpu)         F  I  E  Y  R  K  I  K  R  Q  R  T  I  D  C  L  I  D  R  I

5641       TAAGAGAAAGAGCAGAAGACAGTGGCAATGAGAGCGAGGGGGATAGAGAGGAATTGTCAA
(Env start)                            M  R  A  R  G  I  E  R  N  C  Q
(Vpu)         R  E  R  A  E  D  S  G  N  E  S  E  G  D  R  E  E  L  S  K

5701       AACTTGTGGAAATGGGGCATCATGCTCCTGGGGATGTTGATGATCTGTAGTGCTGCAGGA
(Env)        N  L  W  K  W  G  I  M  L  L  G  M  L  M  I  C  S  A  A  G
(Vpu end)    L  V  E  M  G  H  H  A  P  G  D  V  D  D  L  *

5761       AATTTGTGGGTCACAGTTTATTATGGGGTGCCTGTCTGGAGGGAAGCAACCACTACTCTA
(Env)        N  L  W  V  T  V  Y  Y  G  V  P  V  W  R  E  A  T  T  T  L


5821       TTTTGTGCATCAGATGCTAAAGCATATAAAACAGAGGCACATAATATCTGGGCTACACAT
(Env)        F  C  A  S  D  A  K  A  Y  K  T  E  A  H  N  I  W  A  T  H

5881       GCCTGTGTACCCACGGACCCCAGCCCACAAGAAATAGAACTGGTAAATGTGACCGAAAAC
(Env)        A  C  V  P  T  D  P  S  P  Q  E  I  E  L  V  N  V  T  E  N

5941       TTTAACATGTGGAAAAATAACATGGTAGACCAGATGCATGAGGATATAATCAGTTTATGG
(Env)        F  N  M  W  K  N  N  M  V  D  Q  M  H  E  D  I  I  S  L  W

6001       GATCAAAGTCTAAAACCATGTGTAAAATTAACCCCACTCTGTGTCACCTTAAACTGCACT
(Env)        D  Q  S  L  K  P  C  V  K  L  T  P  L  C  V  T  L  N  C  T

6061       AATGCCAACATAAACAGCACTGGGAGCAACGCCCTATGGGAGCCAACAAAGGAGGTGAAA
(Env)        N  A  N  I  N  S  T  G  S  N  A  L  W  E  P  T  K  E  V  K

6121       AACTGCTCTTTCAATGTAACTACAGTAGTAAGAGATAAGAAAAAGCAAGTATATGCGCTT
(Env)        N  C  S  F  N  V  T  T  V  V  R  D  K  K  K  Q  V  Y  A  L

6181       TTTTATAAACCTGATATCGTACCAAAAGACAATGATAATAATAGGACCAATTATAGGTTT
(Env)        F  Y  K  P  D  I  V  P  K  D  N  D  N  N  R  T  N  Y  R  F

6241       ATATGTTGTAATACCTCAGCCATTACGCAGGCTTGTCCAAAGATATCCTTTGAGCCAATT
(Env)        I  C  C  N  T  S  A  I  T  Q  A  C  P  K  I  S  F  E  P  I

6301       CCAATACATTATTGTGCCCCAGCTGGTTTTGCGATTCTTAAGTGTAGAAATAAGAAGTTC
(Env)        P  I  H  Y  C  A  P  A  G  F  A  I  L  K  C  R  N  K  K  F

6361       AATGGAACAGGCCCATGCAAAAATGTCAGCACAGTACAATGTACACATGGAATTAAGCCA
(Env)        N  G  T  G  P  C  K  N  V  S  T  V  Q  C  T  H  G  I  K  P

6421       GTGGTGTCAACTCAACTGCTGTTCAATGGCAGTCTACCAGAAGAAGAGATCATTATTAGA
(Env)        V  V  S  T  Q  L  L  F  N  G  S  L  P  E  E  E  I  I  I  R

6481       TCTGAAAATCTCACAAACAATGCTAAAAACATTATACTACAGTTTAATGCATCTGTTAAA
(Env)        S  E  N  L  T  N  N  A  K  N  I  I  L  Q  F  N  A  S  V  K
```

```
6541        ATTAATTGTACAAGGCCCTACGAAATTAGAATACAAAAGACATCAATAGGACAAGGGCAA
(Env)        I  N  C  T  R  P  Y  E  I  R  I  Q  K  T  S  I  G  Q  G  Q

6601        GCACTCAATACAAACAAGAGGATTATACGAGACAATAGACAAGCAAATTGTACCATTAGT
(Env)        A  L  N  T  N  K  R  I  I  R  D  N  R  Q  A  N  C  T  I  S

6661        GGAGAAAAATGGAATAAAACTTTACAACAGGCAGCTATACAATTGGGAAACCTTCTTAAC
(Env)        G  E  K  W  N  K  T  L  Q  Q  A  A  I  Q  L  G  N  L  L  N

6721        AAAACAACAATACCTTTTCGACCACCCTCAGGAGGGGACCCAGAAATTACAACACACAGT
(Env)        K  T  T  I  P  F  R  P  P  S  G  G  D  P  E  I  T  T  H  S

6781        GTTAATTGTGGAGGGGAATTTTTCTACTGTAATACATCAGGGCTGTTTAATAATACATGG
(Env)        V  N  C  G  G  E  F  F  Y  C  N  T  S  G  L  F  N  N  T  W

6841        GATAATAGTAATAGGACATGGTCAAATAAGGGAGCATGGTCAAATCAGACAGTCACACTC
(Env)        D  N  S  N  R  T  W  S  N  K  G  A  W  S  N  Q  T  V  T  L

6901        CCATGCAGAATACGACAAATTATATACATGTGGCAGAAAGTTGGAAAAGCAATGTATGCC
(Env)        P  C  R  I  R  Q  I  I  Y  M  W  Q  K  V  G  K  A  M  Y  A

6961        CCTCCCATACAAGGAACACTTAGATGCTCATCAAATATTACAGGACTACTATTCACAAGA
(Env)        P  P  I  Q  G  T  L  R  C  S  S  N  I  T  G  L  L  F  T  R

7021        GATGGTGGTAATAATAGTTCTAACAACGAGACCTTCAGACCTGGCGGAGGAGATACGAGG
(Env)        D  G  G  N  N  S  S  N  N  E  T  F  R  P  G  G  G  D  T  R

7081        GACAATTGGAGAAGTGAATTATATAAATACAAAGTACTACAAATTGAACCAAGAGGAGCA
(Env)        D  N  W  R  S  E  L  Y  K  Y  K  V  L  Q  I  E  P  R  G  A

7141        GCGCCCACCAAGGCAAAGAGAAGAGTGGTGGAAAGGGAAAAAAGAGCAATACGACTCGGA
(Env)        A  P  T  K  A  K  R  R  V  V  E  R  E  K  R  A  I  R  L  G

7201        GCTATGTTCCTTGGGTTCTTGGGAGCAGCAGGAAGCACAATGGGCGCAGCGTCAGAGACG
(Env)        A  M  F  L  G  F  L  G  A  A  G  S  T  M  G  A  A  S  E  T

7261        CGGACGGTACAGGCCAGACAGGTATTGTCTGGTATACTGCAACAGCAAAACAATTTGCTC
(Env)        R  T  V  Q  A  R  Q  V  L  S  G  I  L  Q  Q  Q  N  N  L  L

7321        AGGGCTATCGAGGCGCAACAGCATCTGTTGCAACTCACGGTCTGGGGCATTAAACAGCTC
(Env)        R  A  I  E  A  Q  Q  H  L  L  Q  L  T  V  W  G  I  K  Q  L

7381        CAGGCAAGAATCCTGGCTGTGGAAAGATACCTCAAGGATCGACGGCTCCTATGCCTTTGG
(Env)        Q  A  R  I  L  A  V  E  R  Y  L  K  D  R  R  L  L  C  L  W

7441        GGTTGCTCTGGAAAACACATTTGCACCACTACTGTGCCCTGGAACTCTAGTTGGAGTAAT
(Env)        G  C  S  G  K  H  I  C  T  T  T  V  P  W  N  S  S  W  S  N

7501        AAAACTCAAACTGAGATTTGGCAGAACATTACCTGGGTGCAGTGGGAAAGAGAAATTGAA
(Env)        K  T  Q  T  E  I  W  Q  N  I  T  W  V  Q  W  E  R  E  I  E

7561        AATTACACAGGACTATTATACAACTTATTTGAGGAATCGCAGATCCAGCAAGAAAAGAAT
(Env)        N  Y  T  G  L  L  Y  N  L  F  E  E  S  Q  I  Q  Q  E  K  N

7621        GAACAAGAATTATTGGAATTGGACAAGTGGGCAAGTCTGTGGAATTGGTTTGACAAAACA
(Env)        E  Q  E  L  L  E  L  D  K  W  A  S  L  W  N  W  F  D  K  T

7681        AGCTGGCTGTGGTATAGAAAAATATTCATTATGCTACTACGAGGTTTGCTACGTTTTAGA
(Env)        S  W  L  W  Y  R  K  I  F  I  M  L  L  R  G  L  L  R  F  R

7741        ATATTTTTTGCTGTGCTTTCTGTATTATACAGAGTTAGGCAGGGATACTCACCTCTGTCG
(Env)        I  F  F  A  V  L  S  V  L  Y  R  V  R  Q  G  Y  S  P  L  S

7801        TTTCAGACCCTCTTCCCAGCCCCGAGGGGACCCGACAGGCCCGAAGGAACAGAAGAAGAA
(Env)        F  Q  T  L  F  P  A  P  R  G  P  D  R  P  E  G  T  E  E  E
(Tatx2 start)     P  S  S  Q  P  R  G  D  P  T  G  P  K  E  Q  K  K  K
(Revx2 start)S  D  P  L  P  S  P  E  G  T  R  Q  A  R  R  N  R  R  R  R
```

```
7861        GGTGGAGAGCAAGGCAGAGACAAATACATTCGATTGATGCGCGGATTCTCCGCACTTATC
(Env)        G  G  E  Q  G  R  D  K  Y  I  R  L  M  R  G  F  S  A  L  I
(Tatx2 end) V  E  S  K  A  E  T  N  T  F  D  *
(Revx2)        W  R  A  R  Q  R  Q  I  H  S  I  D  A  R  I  L  R  T  Y  L

7921        TGGGACGATCTGCGGAACCTGTGCCTCTTCGGCTACCACCGCTCGAGAGACTTACTCTTG
(Env)        W  D  D  L  R  N  L  C  L  F  G  Y  H  R  S  R  D  L  L  L
(Revx2)        G  R  S  A  E  P  V  P  L  R  L  P  P  L  E  R  L  T  L  A

7981        CTTGCAGCGAGGATTGTGGAACTTCTGGGACGCAGGGGGTGGGAAGCCCTCAAGTATCTG
(Env)        L  A  A  R  I  V  E  L  L  G  R  R  G  W  E  A  L  K  Y  L
(Revx2)        C  S  E  D  C  G  T  S  G  T  Q  G  V  G  S  P  Q  V  S  V

8041        TGGAATCTCCTGCAGTATTGGAGTCAGGAACTCAAGAATAGTGTTATTAGCTTGCTTGAT
(Env)        W  N  L  L  Q  Y  W  S  Q  E  L  K  N  S  V  I  S  L  L  D
(Revx2 end)  E  S  P  A  V  L  E  S  G  T  Q  E  *

8101        ACCATCGCAATCGCAACAGCTGAGGGGACAGATAGGGTTACAGAAGTACTACTACGAGCT
(Env)        T  I  A  I  A  T  A  E  G  T  D  R  V  T  E  V  L  L  R  A

8161        TGCAGAGCTATTCTTAACGTACCCAGAAGAATCAGACAGGGCTTTGAAAGGATTTTGCTA
(Env)        C  R  A  I  L  N  V  P  R  R  I  R  Q  G  F  E  R  I  L  L

8221        TAAAATGGGTGGCAAATGGTCAAAAAGTAGTATAGTTGGATGGCCTGCTATAAGGGAAAG
(Env end)    *
(Nef start)     M  G  G  K  W  S  K  S  S  I  V  G  W  P  A  I  R  E  R

8281        AATAAGAAGAACTAATCCAGCAGCAGATGGGGTGGGAGCAGTATCTCGAGACCTAGAAAA
(Nef)         I  R  R  T  N  P  A  A  D  G  V  G  A  V  S  R  D  L  E  K

8341        ACATGGGGCAATCACAAGTAGCAATACAGCAAGTACTAATGCTGACTGTGCCTGGCTAGA
(Nef)         H  G  A  I  T  S  S  N  T  A  S  T  N  A  D  C  A  W  L  E

8401        AGCACAAGAAGAGAGTGAGGAAGTGGGCTTTCCAGTCAAACCTCAGGTACCTTTAAGACC
(Nef)         A  Q  E  E  S  E  E  V  G  F  P  V  K  P  Q  V  P  L  R  P

8461        AATGACTTACAAAGCAGCTGTAGATCTTAGCCACTTTTTAAAAGAAAAGGGGGGACTGGA
(Nef)         M  T  Y  K  A  A  V  D  L  S  H  F  L  K  E  K  G  G  L  E

8521        AGGGCTAATTTGGTCCAAAGAGAGACAAGACATCCTTGATCTTTGGGTCTACAACACACA
(Nef)         G  L  I  W  S  K  E  R  Q  D  I  L  D  L  W  V  Y  N  T  Q

8581        AGGCTACTTCCCCGATTGGCAGAACTACACACCAGGGCCAGGGATCAGATATCCAATAAC
(Nef)         G  Y  F  P  D  W  Q  N  Y  T  P  G  P  G  I  R  Y  P  I  T

8641        CTTTGGATGGTGCTTCGAGCTAGTACCAGTTGACCCACAGGAAGTAGAAGAGGCCACTGA
(Nef)         F  G  W  C  F  E  L  V  P  V  D  P  Q  E  V  E  E  A  T  E

8701        GGGAGAGAACAACTGCTTGTTACACCCTATGAACCAGCATGGAATAGAGGACACGGAGAG
(Nef)         G  E  N  N  C  L  L  H  P  M  N  Q  H  G  I  E  D  T  E  R

8761        ACAAGTGTTAAAGTGGAGATTTAACAGCAGACTAGCATTTGAGCACAAGGCCCGAGAGAA
(Nef)         Q  V  L  K  W  R  F  N  S  R  L  A  F  E  H  K  A  R  E  K

8821        ACATCCGGAGTACTACAAAGACTGCTGA
(Nef end)     H  P  E  Y  Y  K  D  C  *
```

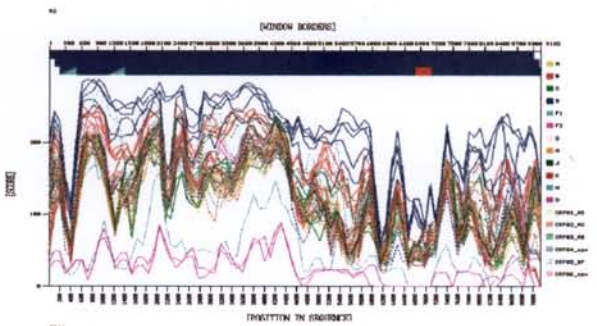
Appendix C

The similarity plots of the NCBI HIV-1 subtyping database for the different Tygerberg plasmids are shown in figures C1-C4. The right hand legend indicates the different reference subtypes, which the query sequence was compared to. The solid bar at the top of the graph represents the subtype most similar to the query sequence. The x-axis represents the nucleotide position, while the y-axis represents the score of identity obtained when comparing the query sequence to the individual reference sequences. The similarity between the query and the reference sequences increases with an increase in score.

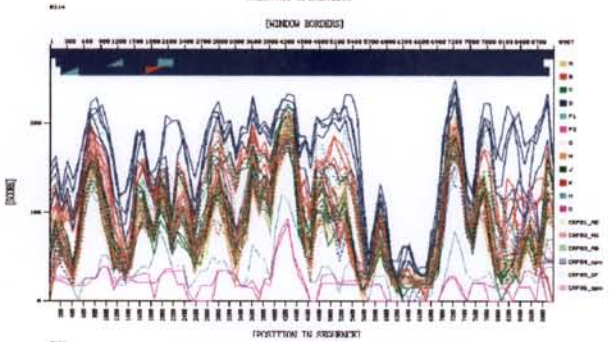
Appendix C

NCBI Subtyping Results of the Tygerberg plasmid sequences (pR2-pR482)

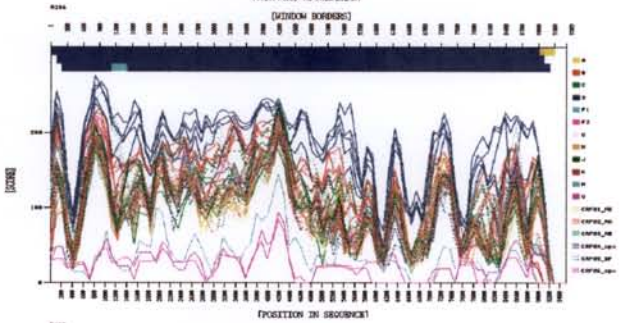
C1: pR2



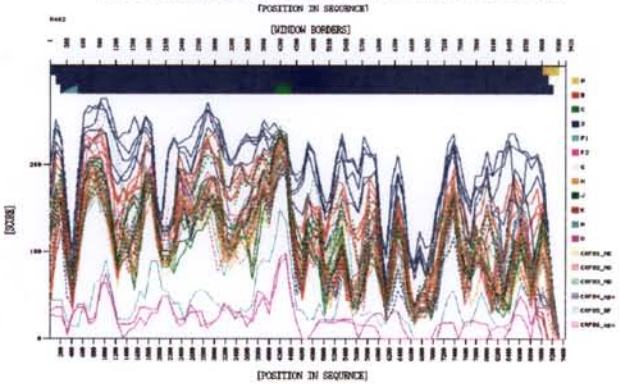
C2: pR214



C3: pR286



C4:pR482



Appendix D

Distance Matrices

The distance matrices are generated by BioEdit and is a function describing the relationship between pairs of sequences. The similarity between sequences are given as percentage similarity. The distance matrices are given for the full-length dataset (D1), *gag* (D2), *pol* (D3), *env* (D4), *vif* (D5), *vpr* (D6), *vpu* (D7), *tat* (D8), *rev* (D9) and *nef* (D10).

D1: Sequence Identity Matrix for the Full-length HIV-1 subtype D Isolates

Seq->	G3509	57140	57132	57146	MB205	J2759	G1055	D2683	UG114	E2343	B2187	K0995	B2564	J3222	E1361	B3239	F0372	K0925	57131	A0848	E0836	F0573	A0741	D2355	57128	A0334	NKU30	TZBFL	57130	1	57143	R2a	R214a	R286a	R482a	ELI	Z226	NDK	MN011
G3509	100.0%	92.0%	91.9%	91.5%	91.6%	91.8%	91.7%	91.0%	91.6%	91.4%	91.8%	91.4%	91.6%	92.4%	91.9%	91.2%	91.1%	91.1%	91.2%	91.0%	91.6%	91.3%	91.3%	90.7%	90.8%	90.6%	90.8%	90.0%	90.3%	90.2%	90.0%	77.6%	74.9%	72.8%	74.4%	89.2%	89.6%	89.2%	88.4%
57140	—	100.0%	92.1%	92.2%	91.9%	92.0%	91.9%	91.1%	91.7%	91.7%	92.1%	91.8%	92.0%	92.6%	91.9%	91.5%	91.3%	91.1%	91.1%	91.1%	91.4%	91.3%	91.5%	90.6%	90.9%	90.7%	90.8%	89.6%	90.8%	90.6%	90.7%	77.6%	74.8%	72.8%	74.2%	89.6%	89.8%	89.6%	88.5%
57132	—	—	100.0%	93.3%	92.5%	92.8%	92.4%	91.6%	92.6%	92.6%	92.6%	92.0%	92.4%	92.9%	92.4%	92.1%	91.5%	91.6%	92.0%	92.0%	92.2%	91.8%	91.8%	91.2%	91.3%	91.6%	91.6%	90.4%	91.0%	90.9%	90.8%	77.5%	74.6%	72.7%	74.3%	89.8%	90.1%	89.7%	88.6%
57146	—	—	—	100.0%	92.7%	92.7%	92.6%	91.4%	92.5%	92.5%	92.4%	92.1%	92.1%	92.7%	92.1%	92.1%	91.2%	91.2%	91.2%	91.4%	91.5%	91.7%	91.8%	90.8%	91.2%	91.0%	91.1%	90.1%	90.9%	90.8%	90.8%	77.1%	74.4%	72.6%	74.3%	89.6%	90.0%	89.6%	88.5%
MB205	—	—	—	—	100.0%	93.8%	93.4%	92.0%	93.1%	93.0%	93.2%	92.9%	92.6%	92.8%	92.5%	92.4%	91.6%	91.6%	91.5%	91.6%	91.7%	92.0%	92.0%	91.3%	91.7%	91.5%	91.6%	90.6%	91.2%	91.0%	91.0%	77.6%	74.8%	73.2%	74.6%	90.3%	90.4%	90.2%	89.1%
J2759	—	—	—	—	—	100.0%	93.6%	92.5%	93.1%	93.4%	93.3%	92.9%	93.0%	93.4%	92.8%	92.7%	91.9%	91.7%	91.9%	92.2%	92.3%	92.0%	92.4%	91.7%	91.9%	91.8%	91.9%	90.5%	91.6%	91.5%	91.2%	77.8%	74.7%	73.1%	74.6%	90.2%	90.4%	90.1%	89.0%
G1055	—	—	—	—	—	—	100.0%	92.2%	92.8%	93.1%	93.0%	92.9%	92.5%	93.3%	92.6%	92.4%	91.6%	91.7%	92.0%	91.7%	91.9%	91.8%	92.1%	91.5%	91.7%	91.9%	91.4%	90.4%	91.5%	91.2%	91.1%	77.8%	74.9%	73.3%	74.7%	90.4%	90.5%	90.2%	88.8%
D2683	—	—	—	—	—	—	—	100.0%	92.7%	91.9%	92.5%	91.8%	91.5%	92.0%	91.7%	91.6%	90.8%	91.0%	91.0%	90.8%	90.8%	90.9%	91.0%	90.6%	90.8%	90.7%	90.6%	89.7%	90.8%	90.6%	90.5%	77.4%	74.6%	72.9%	74.3%	89.6%	90.0%	89.6%	88.2%
UG114	—	—	—	—	—	—	—	—	100.0%	92.9%	93.1%	92.5%	92.6%	93.0%	92.7%	92.6%	91.5%	91.7%	92.0%	91.8%	91.7%	92.0%	92.0%	91.6%	92.0%	91.8%	91.6%	90.7%	91.6%	91.6%	91.4%	77.7%	74.8%	73.2%	74.7%	90.2%	90.4%	90.0%	88.8%
EZ343	—	—	—	—	—	—	—	—	—	100.0%	92.9%	93.1%	92.6%	92.9%	92.1%	92.2%	91.4%	91.6%	91.5%	91.7%	92.0%	91.9%	92.2%	91.0%	91.4%	91.5%	91.4%	90.2%	91.2%	91.0%	90.9%	77.5%	74.8%	73.1%	74.5%	90.2%	90.3%	89.9%	88.6%
B2187	—	—	—	—	—	—	—	—	—	—	100.0%	92.5%	92.5%	92.9%	92.5%	92.3%	91.8%	91.8%	91.7%	92.1%	92.2%	92.0%	92.2%	91.7%	91.6%	91.7%	91.6%	90.8%	91.3%	91.1%	90.9%	77.8%	74.9%	73.1%	74.6%	90.3%	90.4%	90.1%	89.1%
K0995	—	—	—	—	—	—	—	—	—	—	—	100.0%	92.2%	92.7%	92.3%	92.1%	91.5%	91.5%	91.2%	91.5%	91.7%	91.8%	91.8%	90.8%	91.3%	91.5%	91.1%	90.1%	91.1%	90.9%	90.8%	77.6%	74.6%	73.0%	74.3%	89.7%	89.8%	89.7%	88.7%
B2564	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	94.1%	93.2%	92.8%	92.2%	92.6%	92.3%	92.5%	92.3%	92.4%	92.0%	91.8%	91.8%	91.5%	91.7%	90.7%	91.0%	90.9%	90.4%	77.8%	74.9%	73.0%	74.8%	90.0%	90.1%	89.8%	88.8%
J3222	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	94.0%	93.4%	92.5%	92.5%	92.8%	92.6%	92.7%	92.4%	92.5%	92.0%	92.0%	92.1%	92.1%	91.1%	91.5%	91.4%	91.2%	78.2%	75.3%	73.5%	74.9%	90.4%	90.5%	90.6%	89.5%
E1361	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	93.0%	92.1%	92.1%	92.1%	92.3%	91.9%	92.0%	91.9%	91.4%	91.7%	91.7%	91.6%	90.6%	91.0%	90.9%	90.7%	77.8%	74.8%	73.1%	74.4%	90.0%	90.2%	89.8%	88.8%
B3239	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	92.5%	92.4%	92.1%	92.1%	91.7%	91.8%	91.8%	91.2%	91.6%	91.2%	91.4%	90.5%	90.8%	90.8%	90.8%	77.6%	74.9%	73.0%	74.4%	89.9%	89.9%	89.7%	88.9%
F0372	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	91.6%	91.8%	91.8%	91.5%	91.7%	91.1%	90.4%	90.9%	90.8%	90.8%	90.1%	90.4%	90.3%	90.4%	77.1%	74.3%	72.3%	73.8%	89.4%	89.5%	89.2%	88.4%
K0925	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	92.0%	92.1%	91.8%	91.5%	90.9%	90.8%	91.3%	91.1%	91.0%	90.1%	90.4%	90.3%	90.0%	77.5%	74.6%	72.7%	74.2%	89.3%	89.7%	89.4%	88.2%
57131	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	91.6%	91.6%	91.5%	91.2%	90.8%	91.1%	91.1%	90.8%	89.8%	90.4%	90.3%	90.1%	77.2%	74.2%	72.5%	73.9%	89.1%	89.4%	89.2%	88.1%
A0848	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	91.4%	91.4%	91.2%	91.3%	91.2%	91.2%	91.1%	91.0%	90.0%	90.3%	90.3%	89.9%	77.3%	74.6%	72.5%	73.9%	88.9%	89.3%	89.1%	88.1%
E0836	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	91.6%	91.8%	91.5%	91.2%	91.2%	91.1%	91.1%	90.1%	90.5%	90.4%	90.4%	77.2%	74.4%	72.3%	73.8%	89.2%	89.5%	89.2%	88.2%
F0573	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	91.4%	91.4%	91.2%	91.8%	90.9%	91.2%	91.2%	91.0%	90.0%	90.4%	90.3%	90.4%	77.5%	74.7%	72.9%	74.4%	89.6%	89.9%	89.6%	88.7%
A0741	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	91.5%	90.8%	91.0%	91.0%	91.0%	90.1%	90.5%	90.4%	90.4%	77.2%	74.3%	72.6%	74.1%	89.4%	89.7%	89.3%	88.7%
D2355	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	90.7%	91.9%	91.8%	91.3%	89.7%	89.8%	89.7%	89.9%	77.0%	74.1%	72.2%	73.9%	89.1%	89.1%	88.8%	87.7%
57128	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	91.9%	91.4%	89.8%	90.3%	90.2%	89.9%	77.1%	74.2%	72.8%	74.0%	89.4%	89.4%	89.2%	88.2%
A0334	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	91.2%	89.7%	90.1%	90.0%	90.1%	77.1%	74.1%	72.4%	73.8%	89.0%	89.3%	89.3%	88.0%
NKU30	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	89.7%	90.2%	90.1%	89.8%	76.9%	74.1%	72.4%	73.7%	89.2%	89.3%	88.9%	88.0%
TZBFL	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	89.2%	89.2%	88.6%	78.2%	73.4%	71.5%	73.0%	88.2%	88.2%	88.0%	87.4%
57130	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	99.1%	92.7%	78.0%	75.2%	73.6%	75.0%	90.6%	91.0%	90.5%	89.4%
1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	92.4%	78.1%	75.2%	73.6%	75.0%	90.6%	91.1%	90.5%	89.5%
57143	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	77.8%	74.9%	73.4%	74.7%	90.3%	90.4%	90.1%	88.8%
R2a	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	92.4%	88.8%	90.5%	80.4%	80.9%	80.7%	77.8%
R214a	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	86.4%	87.3%	77.2%	77.7%	77.7%	75.0%
R286a	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	91.0%	75.6%	78.2%	78.2%	73.4%
R482a	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	77.2%	77.5%	77.5%	74.6%
ELI	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	95.5%	94.2%	90.4%
Z226	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	94.4%	90.6%
NDK	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	90.3%
MN011	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%
MN012	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
94ZR0	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—

D2: Sequence Identity Matrix for the HIV-1 subtype D gag

Seq->	G3509	57140	57132	57146	MB205	J2759	G1055	D2683	UG114	E2343	B2187	K0995	B2564	J3222	E1361	B3239	F0372	K0925	57131	A0848	E0836	F0573	A0741	D2355	57128	A0334	NKU30	TZBFL	57130	1	57143	R2a	R214a	R286a	R482a	ELI	Z226	NDK	MN011	MN012
G3509	100.0%	92.0%	91.9%	91.5%	91.8%	91.8%	91.7%	91.0%	91.8%	91.4%	91.8%	91.4%	91.6%	92.4%	91.9%	91.2%	91.1%	91.1%	91.2%	91.0%	91.6%	91.3%	91.3%	90.7%	90.8%	90.6%	90.6%	90.0%	90.3%	90.2%	90.0%	77.6%	74.9%	72.8%	74.4%	89.2%	89.6%	89.2%	88.4%	88.0%
57140	—	100.0%	92.1%	92.2%	91.9%	92.0%	91.9%	91.1%	91.7%	91.7%	92.1%	91.8%	92.0%	92.6%	91.9%	91.5%	91.3%	91.1%	91.1%	91.1%	91.4%	91.3%	91.5%	90.8%	90.9%	90.7%	90.6%	89.6%	90.8%	90.6%	90.7%	77.6%	74.8%	72.8%	74.2%	89.6%	89.8%	89.6%	88.5%	88.2%
57132	—	—	100.0%	93.3%	92.5%	92.8%	92.4%	91.8%	92.6%	92.6%	92.8%	92.0%	92.4%	92.9%	92.4%	92.1%	91.5%	91.8%	92.0%	92.0%	92.2%	91.9%	91.8%	91.2%	91.3%	91.8%	91.6%	90.4%	91.0%	90.9%	90.8%	77.5%	74.6%	72.7%	74.3%	89.8%	90.1%	89.7%	88.6%	88.2%
57146	—	—	—	100.0%	92.7%	92.7%	92.8%	91.4%	92.5%	92.5%	92.4%	92.1%	92.1%	92.7%	92.1%	92.1%	91.2%	91.2%	91.2%	91.4%	91.5%	91.7%	91.8%	90.8%	91.2%	91.0%	91.1%	90.1%	90.9%	90.8%	90.8%	77.1%	74.4%	72.8%	74.3%	89.6%	90.0%	89.8%	88.5%	88.4%
MB205	—	—	—	—	100.0%	93.8%	93.4%	92.0%	93.1%	93.0%	93.2%	92.9%	92.8%	92.8%	92.5%	92.4%	91.6%	91.8%	91.5%	91.8%	91.7%	92.0%	92.0%	91.3%	91.7%	91.5%	91.6%	90.8%	91.2%	91.0%	91.0%	77.8%	74.8%	73.2%	74.6%	90.3%	90.4%	90.2%	89.1%	88.6%
J2759	—	—	—	—	—	100.0%	93.8%	92.5%	93.1%	93.4%	93.3%	92.9%	93.0%	93.4%	92.8%	92.7%	91.9%	91.7%	91.9%	92.2%	92.3%	92.0%	92.4%	91.7%	91.9%	91.8%	91.9%	90.5%	91.6%	91.5%	91.2%	77.8%	74.7%	73.1%	74.6%	90.2%	90.4%	90.1%	89.0%	88.6%
G1055	—	—	—	—	—	—	100.0%	92.2%	92.8%	93.1%	93.0%	92.9%	92.5%	93.3%	92.6%	92.4%	91.8%	91.7%	92.0%	91.7%	91.9%	91.8%	92.1%	91.5%	91.7%	91.9%	91.4%	90.4%	91.5%	91.2%	91.1%	77.8%	74.9%	73.3%	74.7%	90.4%	90.5%	90.2%	88.8%	88.6%
D2683	—	—	—	—	—	—	—	100.0%	92.7%	91.9%	92.5%	91.8%	91.5%	92.0%	91.7%	91.6%	90.8%	91.0%	91.0%	90.8%	90.8%	90.9%	91.0%	90.6%	90.8%	90.7%	90.6%	89.7%	90.8%	90.6%	90.5%	77.4%	74.6%	72.9%	74.3%	89.8%	90.0%	89.8%	88.2%	88.1%
UG114	—	—	—	—	—	—	—	—	100.0%	92.9%	93.1%	92.5%	92.0%	93.0%	92.7%	92.6%	91.5%	91.7%	92.0%	91.8%	91.7%	92.0%	92.0%	91.6%	92.0%	91.8%	91.8%	90.7%	91.6%	91.6%	91.4%	77.7%	74.8%	73.2%	74.7%	90.2%	90.4%	90.0%	88.8%	88.5%
E2343	—	—	—	—	—	—	—	—	—	100.0%	92.9%	93.1%	92.6%	92.9%	92.1%	92.2%	91.4%	91.6%	91.5%	91.7%	92.0%	91.9%	92.2%	91.0%	91.4%	91.5%	91.4%	90.2%	91.2%	91.0%	90.9%	77.5%	74.6%	73.1%	74.5%	90.2%	90.3%	89.9%	88.6%	88.3%
B2187	—	—	—	—	—	—	—	—	—	—	100.0%	92.5%	92.5%	92.9%	92.5%	92.3%	91.8%	91.8%	91.7%	92.1%	92.2%	92.0%	92.2%	91.7%	91.8%	91.7%	91.6%	90.8%	91.3%	91.1%	90.9%	77.6%	74.9%	73.1%	74.6%	90.3%	90.4%	90.1%	89.1%	88.7%
K0995	—	—	—	—	—	—	—	—	—	—	—	100.0%	92.2%	92.7%	92.3%	92.1%	91.5%	91.5%	91.2%	91.5%	91.7%	91.8%	91.8%	90.8%	91.3%	91.5%	91.1%	90.1%	91.1%	90.9%	90.8%	77.6%	74.6%	73.0%	74.3%	89.7%	89.8%	89.7%	88.7%	88.2%
B2564	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	94.1%	93.2%	92.8%	92.2%	92.8%	92.3%	92.5%	92.3%	92.4%	92.0%	91.8%	91.8%	91.5%	91.7%	90.7%	91.0%	90.9%	90.4%	77.8%	74.9%	73.0%	74.6%	90.0%	90.1%	89.8%	88.8%	88.3%
J3222	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	94.0%	93.4%	92.5%	92.5%	92.8%	92.8%	92.8%	92.7%	92.4%	92.5%	92.0%	92.1%	92.1%	91.1%	91.5%	91.4%	91.2%	78.2%	75.3%	73.5%	74.9%	90.4%	90.5%	90.8%	89.5%	89.1%
E1361	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	93.0%	92.1%	92.1%	92.1%	92.3%	91.9%	92.0%	91.9%	91.4%	91.7%	91.7%	91.6%	90.7%	91.5%	91.4%	90.7%	77.8%	74.8%	73.1%	74.4%	90.0%	90.2%	89.8%	88.8%	88.4%
B3239	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	92.5%	92.4%	92.1%	92.1%	91.7%	91.8%	91.6%	91.2%	91.6%	91.2%	91.4%	90.5%	90.8%	90.8%	90.6%	77.6%	74.9%	73.0%	74.4%	89.9%	89.9%	89.7%	88.9%	88.4%
F0372	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	91.8%	91.8%	91.8%	91.8%	91.5%	91.7%	91.1%	90.9%	90.6%	90.8%	90.1%	90.4%	90.3%	90.0%	77.5%	74.6%	72.7%	74.2%	89.3%	89.7%	89.4%	88.2%	87.8%
K0925	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	92.0%	92.1%	91.8%	91.5%	90.9%	90.6%	91.3%	91.1%	91.0%	90.1%	90.4%	90.3%	90.1%	77.2%	74.2%	72.5%	73.9%	89.1%	89.4%	89.2%	88.1%	87.9%
57131	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	91.8%	91.8%	91.5%	91.2%	90.8%	91.1%	91.1%	91.0%	90.1%	90.4%	90.3%	90.1%	77.2%	74.2%	72.5%	73.9%	89.1%	89.4%	89.2%	88.1%	87.9%
A0848	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	91.4%	91.4%	91.2%	91.3%	91.2%	91.1%	91.0%	90.0%	90.3%	90.3%	89.9%	77.3%	74.6%	72.5%	73.9%	88.9%	89.3%	89.1%	88.1%	87.8%
E0836	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	92.2%	91.8%	90.8%	91.2%	91.2%	91.0%	90.0%	90.4%	90.3%	90.4%	77.5%	74.7%	72.9%	74.4%	89.6%	89.9%	89.6%	88.7%	88.2%
F0573	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	91.8%	90.9%	91.2%	91.2%	91.0%	90.0%	90.4%	90.3%	90.4%	77.2%	74.3%	72.6%	74.1%	89.4%	89.7%	89.3%	88.7%	88.4%
A0741	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	90.8%	90.9%	91.0%	91.0%	90.1%	90.5%	90.4%	90.4%	77.2%	74.3%	72.6%	74.1%	89.4%	89.7%	89.3%	88.7%	88.4%
D2355	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	91.9%	91.8%	91.3%	89.7%	89.8%	89.7%	89.9%	77.0%	74.1%	72.2%	73.9%	89.1%	89.1%	88.8%	87.7%	87.3%
57128	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	91.9%	91.4%	89.8%	90.3%	90.2%	89.9%	77.1%	74.2%	72.6%	74.0%	89.4%	89.4%	89.2%	88.2%	87.9%
A0334	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	91.2%	89.7%	90.1%	90.0%	90.1%	77.1%	74.1%	72.4%	73.8%	89.0%	89.3%	89.3%	88.0%	87.6%
NKU30	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	89.7%	90.2%	90.1%	89.8%	76.9%	74.1%	72.4%	73.7%	89.2%	89.3%	88.9%	88.0%	87.6%
TZBFL	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	89.2%	89.2%	88.8%	76.2%	73.4%	71.5%	73.0%	88.2%	88.2%	88.0%	87.4%	87.2%
57130	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	99.1%	92.7%	78.0%	75.2%	73.6%	75.0%	90.6%	91.0%	90.5%	89.4%	89.2%
1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	92.4%	78.1%	75.2%	73.8%	75.0%	90.8%	91.1%	90.5%	89.5%	89.2%
57143	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	77.8%	74.9%	73.4%	74.7%	90.3%	90.4%	90.1%	88.8%	88.7%
R2a	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	92.4%	88.8%	90.5%	80.4%	80.9%	80.7%	77.8%	77.4%
R214a	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	86.4%	87.3%	77.2%	77.7%	77.7%	75.0%	74.5%
R286a	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	91.0%	75.8%	76.2%	76.2%	73.4%	72.9%
R482a	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	77.2%	77.5%	77.5%	74.8%	74.3%
ELI	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	95.5%	94.2%	90.4%	90.0%
Z226	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	94.4%	90.8%	90.2%
NDK	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	90.3%	89.9%
MN011	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	93.8%
MN012	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%
84ZR0	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—

D3:Sequence Identity Matrix for the HIV-1 subtype D *pol*

Seq->	57130	1	57143	MN011	MN012	84ZR0	ELI	Z2Z6	NDK	R 2	R 214	R 286	R 482	MB205	57128	J2759	B2187	G1055	E1361	J3222	D2355	A0848	B2564	A0741	F0573	B3239	57132	57146	57140	D2683	94UG1	E2343	A0334	E0836	57131	K0995	G3509	K0925	NKU30	TZBFL
57130	100.0%	99.4%	95.5%	93.0%	93.1%	94.1%	94.4%	94.7%	95.0%	94.2%	89.2%	90.6%	92.2%	94.3%	94.2%	94.6%	94.8%	94.9%	94.6%	94.8%	94.0%	94.1%	94.4%	94.0%	94.1%	94.6%	94.9%	94.7%	94.6%	93.8%	94.2%	94.6%	94.1%	95.2%	94.6%	94.4%	94.3%	94.0%	94.8%	92.1%
1	—	100.0%	95.3%	93.0%	93.1%	94.1%	94.4%	94.6%	95.0%	94.2%	89.2%	90.6%	92.2%	94.3%	94.1%	94.6%	94.8%	94.7%	94.5%	94.7%	93.9%	94.0%	94.3%	93.9%	94.0%	94.7%	94.8%	94.6%	94.5%	93.8%	94.1%	94.6%	94.1%	95.1%	94.5%	94.4%	94.3%	93.8%	94.7%	92.1%
57143	—	—	100.0%	92.5%	92.4%	93.8%	94.3%	94.5%	94.5%	94.2%	89.3%	90.6%	92.2%	93.6%	93.5%	94.1%	94.1%	94.1%	94.1%	93.8%	93.8%	93.6%	93.8%	93.6%	93.5%	94.1%	94.4%	94.0%	94.1%	93.5%	93.7%	93.9%	93.5%	94.3%	93.6%	93.8%	93.6%	93.6%	94.3%	91.5%
MN011	—	—	—	100.0%	95.3%	94.7%	93.8%	93.8%	94.4%	93.1%	88.4%	89.7%	91.1%	91.9%	92.2%	92.4%	92.4%	92.4%	92.1%	92.4%	92.0%	92.0%	92.4%	92.3%	92.1%	92.3%	92.4%	92.4%	92.3%	91.4%	91.7%	92.3%	91.7%	92.3%	91.9%	92.0%	92.4%	91.8%	92.5%	90.3%
MN012	—	—	—	—	100.0%	94.7%	93.7%	93.8%	94.1%	93.0%	88.5%	89.5%	91.1%	91.8%	92.0%	92.2%	92.4%	92.3%	92.1%	92.4%	91.9%	92.1%	92.2%	92.2%	91.9%	92.1%	92.2%	92.6%	92.2%	91.4%	91.7%	92.1%	91.5%	92.2%	91.6%	91.7%	92.3%	91.7%	92.5%	89.9%
84ZR0	—	—	—	—	—	100.0%	95.2%	95.3%	95.8%	95.0%	90.1%	91.1%	93.0%	93.2%	93.3%	93.6%	93.7%	93.9%	93.5%	93.8%	93.2%	93.3%	93.6%	93.3%	92.8%	93.3%	93.5%	93.5%	93.5%	92.7%	93.1%	93.5%	93.0%	93.5%	93.1%	93.4%	93.3%	93.0%	93.6%	91.1%
ELI	—	—	—	—	—	—	100.0%	97.5%	97.1%	96.2%	91.1%	92.2%	94.0%	93.8%	93.6%	94.1%	94.0%	94.2%	94.0%	94.1%	93.9%	93.6%	93.9%	93.6%	93.5%	94.3%	94.0%	93.9%	94.2%	93.2%	93.7%	94.0%	93.5%	94.0%	93.4%	93.8%	93.6%	93.4%	94.1%	91.6%
Z2Z6	—	—	—	—	—	—	—	100.0%	97.1%	96.3%	91.1%	92.5%	94.1%	93.7%	93.6%	94.0%	94.1%	94.3%	94.1%	94.0%	93.8%	93.9%	94.0%	93.5%	93.5%	94.1%	94.1%	94.0%	94.1%	93.3%	93.6%	94.0%	93.6%	94.2%	93.3%	93.6%	93.6%	93.4%	94.3%	91.5%
NDK	—	—	—	—	—	—	—	—	100.0%	96.6%	91.6%	92.8%	94.7%	94.2%	93.9%	94.6%	94.5%	94.9%	94.3%	94.8%	94.0%	94.0%	94.5%	93.9%	93.9%	94.5%	94.3%	94.3%	94.4%	93.8%	93.8%	94.4%	93.8%	94.5%	93.7%	94.0%	94.2%	93.9%	94.5%	91.8%
R 2	—	—	—	—	—	—	—	—	—	100.0%	93.1%	93.7%	95.7%	93.5%	93.1%	93.7%	93.8%	94.1%	93.6%	93.8%	93.4%	93.4%	93.7%	93.0%	93.2%	93.7%	93.9%	93.5%	93.9%	93.2%	93.3%	93.6%	93.3%	93.7%	93.1%	93.3%	93.7%	93.3%	93.6%	91.6%
R 214	—	—	—	—	—	—	—	—	—	—	100.0%	89.6%	91.1%	88.8%	88.4%	88.8%	89.1%	89.5%	88.9%	89.1%	88.6%	88.6%	88.9%	88.5%	88.5%	89.0%	89.0%	88.6%	89.1%	88.5%	88.8%	88.9%	88.4%	89.0%	88.2%	88.7%	89.2%	88.7%	88.9%	86.9%
R 286	—	—	—	—	—	—	—	—	—	—	—	100.0%	92.6%	89.8%	89.6%	90.0%	90.1%	90.5%	90.1%	90.3%	89.8%	89.8%	90.2%	89.5%	90.1%	90.1%	90.3%	90.2%	90.2%	89.6%	89.7%	90.4%	89.8%	90.1%	89.4%	89.8%	90.0%	89.7%	90.0%	88.1%
R 482	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	91.5%	91.2%	91.7%	91.8%	92.1%	91.6%	92.0%	91.4%	91.4%	92.0%	91.1%	91.5%	91.7%	91.8%	91.7%	91.9%	91.4%	91.6%	91.8%	91.4%	91.8%	90.9%	91.4%	91.8%	91.2%	91.7%	89.4%
MB205	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	95.7%	96.2%	96.1%	96.0%	95.8%	95.9%	95.0%	95.8%	95.5%	95.2%	94.9%	95.4%	95.5%	95.3%	95.5%	94.7%	95.2%	95.6%	95.1%	95.5%	94.7%	95.5%	95.2%	94.9%	95.3%	93.4%
57128	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	95.6%	95.8%	95.4%	95.8%	96.0%	95.1%	95.2%	95.4%	95.0%	94.7%	95.7%	95.3%	95.4%	95.3%	94.2%	95.1%	95.3%	95.0%	95.3%	95.0%	95.1%	94.9%	94.7%	95.3%	93.2%
J2759	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	96.2%	96.4%	96.1%	96.2%	95.4%	95.9%	96.1%	95.8%	95.3%	96.0%	95.9%	95.6%	95.8%	95.1%	95.7%	96.0%	95.5%	95.9%	95.2%	95.9%	95.5%	95.3%	95.9%	93.5%
B2187	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	96.5%	96.2%	96.5%	95.7%	95.9%	96.9%	95.8%	95.2%	95.7%	96.0%	95.5%	95.5%	95.2%	95.6%	96.0%	95.4%	95.9%	95.2%	95.7%	95.5%	95.0%	95.8%	93.7%
G1055	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	96.3%	96.4%	95.5%	95.9%	96.1%	95.6%	95.1%	95.9%	95.8%	95.7%	95.9%	95.1%	95.5%	96.2%	95.4%	95.9%	95.5%	96.0%	95.4%	95.3%	95.8%	93.7%
E1361	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	96.5%	95.4%	95.7%	96.0%	95.3%	95.3%	95.9%	95.6%	95.6%	95.8%	95.0%	95.4%	95.8%	95.4%	95.9%	95.3%	95.8%	95.4%	95.2%	95.6%	93.4%
J3222	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	95.5%	95.6%	96.0%	95.7%	95.4%	96.3%	95.9%	95.9%	95.9%	94.9%	95.4%	96.0%	95.4%	95.9%	95.3%	95.6%	95.5%	95.1%	95.6%	93.7%
D2355	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	95.4%	95.4%	94.8%	94.8%	95.2%	95.1%	94.8%	95.0%	94.3%	95.0%	95.1%	94.5%	94.8%	94.7%	94.9%	94.5%	94.3%	95.2%	93.2%
A0848	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	95.7%	95.2%	94.9%	95.4%	95.4%	95.2%	95.4%	94.4%	95.1%	95.6%	94.8%	95.5%	94.9%	95.2%	94.9%	94.8%	95.1%	93.2%
B2564	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	95.2%	94.9%	95.6%	95.4%	95.5%	95.4%	94.6%	95.2%	95.7%	95.1%	95.6%	94.9%	95.5%	95.1%	95.1%	95.3%	93.3%
A0741	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	95.0%	95.6%	95.0%	94.9%	95.1%	94.2%	94.9%	95.5%	95.0%	95.3%	94.6%	95.0%	95.0%	94.8%	95.3%	93.3%
F0573	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	95.2%	94.8%	94.8%	94.8%	94.0%	94.6%	95.2%	94.9%	95.2%	94.6%	94.8%	94.4%	94.3%	95.0%	92.8%
B3239	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	95.6%	95.2%	95.6%	94.9%	95.3%	95.9%	95.3%	95.5%	95.2%	95.5%	95.1%	95.1%	95.4%	93.3%
57132	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	96.1%	95.5%	94.9%	95.3%	95.9%	95.3%	95.6%	95.2%	95.4%	95.3%	94.9%	95.6%	93.6%
57146	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	96.0%	94.7%	95.1%	95.5%	95.1%	96.4%	94.9%	95.4%	95.0%	94.7%	95.6%	93.2%
57140	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	94.9%	95.2%	95.7%	94.9%	95.3%	95.0%	95.3%	95.1%	94.7%	95.5%	93.4%
D2683	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	95.1%	95.0%	94.4%	94.5%	94.2%	94.7%	94.5%	94.1%	94.6%	92.2%
94UG1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	95.6%	94.9%	95.3%	94.9%	95.2%	95.0%	94.6%	95.0%	93.1%
E2343	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	95.6%	95.8%	95.2%	95.6%	95.3%	94.9%	95.7%	93.3%
A0334	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	95.4%	94.7%	95.3%	94.7%	94.6%	95.1%	93.1%
E0836	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	95.6%	95.9%	95.4%	95.2%	95.7%	93.2%
57131	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	95.1%	95.0%	94.6%	94.9%	92.8%
K0995	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	95.2%	94.9%	95.3%	93.1%
G3509	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	94.9%	95.1%	92.9%
K0925	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	95.0%	93.0%
NKU30	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	93.3%
TZBFL	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%
F0372	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—

D4:Sequence Identity Matrix for the HIV-1 subtype D env

Seq->	R2	R214	R286	R482	.ZZ6	ELI	NDK	84ZR0	MN011	MN012	57143	57130	G3509	57140	A0741	E2343	K0995	57132	57146	MB205	B2187	D2355	A0334	57128	E0836	F0573	J2759	G1055	57131	94UG1	J3222	NKU30	E1361	K0925	A0848	F0372	B3239	TZBFL	D2683	B2564
R2	100%	89%	88%	88%	86%	85%	85%	84%	83%	81%	84%	83%	82%	81%	80%	80%	81%	81%	80%	81%	81%	80%	80%	80%	80%	80%	81%	80%	80%	82%	82%	79%	81%	81%	80%	80%	80%	79%	80%	81%
R214	—	100%	86%	86%	84%	83%	83%	82%	81%	80%	82%	81%	81%	80%	78%	79%	80%	79%	79%	79%	80%	78%	79%	79%	78%	79%	79%	79%	78%	80%	80%	78%	79%	79%	79%	79%	79%	78%	79%	80%
R286	—	—	100%	90%	87%	86%	86%	85%	84%	82%	85%	84%	82%	82%	81%	81%	82%	81%	81%	81%	82%	80%	81%	81%	80%	81%	82%	82%	80%	83%	82%	80%	82%	81%	80%	80%	81%	80%	81%	81%
R482	—	—	—	100%	86%	86%	85%	85%	83%	82%	85%	84%	82%	81%	80%	81%	81%	81%	81%	81%	82%	81%	81%	82%	80%	82%	82%	82%	81%	83%	82%	80%	81%	81%	80%	80%	80%	79%	81%	81%
.ZZ6	—	—	—	—	100%	93%	90%	90%	87%	86%	88%	88%	85%	85%	84%	84%	85%	85%	85%	85%	85%	84%	84%	85%	84%	85%	86%	85%	84%	86%	86%	83%	85%	84%	83%	84%	84%	83%	85%	85%
ELI	—	—	—	—	—	100%	90%	88%	86%	85%	86%	87%	84%	84%	83%	84%	84%	83%	83%	84%	84%	82%	83%	83%	85%	83%	84%	84%	84%	83%	85%	85%	82%	84%	83%	82%	83%	82%	84%	84%
NDK	—	—	—	—	—	—	100%	87%	85%	84%	87%	86%	83%	83%	82%	83%	84%	83%	83%	84%	84%	82%	83%	83%	83%	83%	84%	84%	84%	83%	85%	84%	81%	84%	82%	82%	82%	83%	82%	83%
84ZR0	—	—	—	—	—	—	—	100%	87%	85%	87%	87%	83%	83%	83%	83%	84%	84%	84%	84%	85%	83%	84%	84%	83%	84%	85%	84%	83%	85%	84%	82%	84%	83%	83%	83%	83%	82%	84%	84%
MN011	—	—	—	—	—	—	—	—	100%	91%	86%	85%	84%	84%	83%	82%	84%	83%	83%	84%	84%	83%	84%	84%	83%	83%	84%	83%	82%	84%	85%	82%	83%	83%	83%	82%	84%	82%	84%	84%
MN012	—	—	—	—	—	—	—	—	—	100%	85%	84%	83%	83%	82%	81%	82%	82%	82%	82%	83%	81%	83%	82%	82%	82%	83%	83%	81%	83%	84%	80%	82%	82%	81%	81%	82%	82%	82%	82%
57143	—	—	—	—	—	—	—	—	—	—	100%	87%	85%	85%	84%	84%	84%	84%	84%	85%	84%	83%	84%	84%	83%	83%	85%	84%	83%	85%	86%	82%	84%	84%	83%	83%	83%	82%	84%	84%
57130	—	—	—	—	—	—	—	—	—	—	—	100%	85%	86%	84%	85%	85%	85%	85%	85%	86%	84%	85%	85%	84%	85%	87%	86%	84%	87%	86%	84%	86%	85%	84%	85%	85%	83%	85%	85%
G3509	—	—	—	—	—	—	—	—	—	—	—	—	100%	87%	84%	84%	83%	84%	84%	85%	85%	83%	84%	84%	84%	84%	84%	84%	84%	85%	85%	83%	85%	84%	83%	83%	83%	82%	84%	84%
57140	—	—	—	—	—	—	—	—	—	—	—	—	—	100%	84%	84%	84%	85%	85%	85%	86%	83%	84%	84%	84%	84%	85%	85%	84%	86%	86%	83%	86%	84%	84%	84%	84%	83%	85%	85%
A0741	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100%	89%	88%	88%	87%	87%	87%	84%	85%	85%	86%	86%	87%	87%	85%	87%	87%	84%	86%	85%	83%	84%	85%	84%	85%	86%
E2343	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100%	89%	89%	88%	89%	89%	85%	86%	86%	86%	87%	88%	87%	85%	88%	87%	85%	87%	86%	84%	84%	86%	84%	86%	87%
K0995	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100%	88%	88%	89%	88%	86%	87%	87%	86%	87%	88%	87%	85%	88%	88%	87%	86%	86%	85%	86%	87%	85%	86%	88%
57132	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100%	90%	89%	89%	85%	87%	86%	86%	88%	88%	87%	86%	88%	87%	86%	87%	86%	85%	86%	87%	85%	86%	88%
57146	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100%	90%	89%	86%	86%	86%	86%	88%	89%	88%	86%	88%	88%	85%	87%	87%	85%	86%	88%	85%	87%	87%
MB205	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100%	91%	86%	87%	87%	87%	88%	90%	89%	87%	89%	88%	86%	87%	87%	87%	88%	89%	87%	88%	89%
B2187	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100%	87%	87%	87%	87%	88%	89%	89%	87%	89%	88%	86%	86%	88%	87%	87%	87%	86%	87%	87%
D2355	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100%	89%	88%	87%	87%	87%	86%	85%	86%	88%	85%	86%	86%	86%	85%	86%	84%	85%	87%
A0334	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100%	88%	87%	89%	88%	88%	87%	88%	89%	85%	88%	87%	86%	86%	87%	85%	86%	87%
57128	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100%	86%	88%	88%	87%	86%	86%	88%	85%	88%	87%	86%	86%	87%	85%	86%	87%
E0836	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100%	88%	87%	86%	86%	86%	88%	85%	87%	87%	86%	86%	86%	85%	85%	87%
F0573	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100%	88%	87%	87%	89%	88%	86%	88%	88%	87%	87%	87%	86%	86%	88%
J2759	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100%	90%	88%	90%	90%	87%	88%	88%	87%	87%	88%	85%	88%	88%
G1055	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100%	88%	90%	90%	85%	88%	87%	86%	87%	87%	85%	87%	87%
57131	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100%	89%	89%	86%	88%	87%	86%	86%	88%	85%	87%	87%
94UG1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100%	90%	87%	89%	88%	86%	87%	88%	86%	88%	88%
J3222	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100%	89%	91%	88%	88%	87%	88%	86%	87%	89%
NKU30	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100%	87%	86%	86%	85%	86%	84%	84%	86%
E1361	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100%	87%	87%	87%	88%	85%	86%	86%
K0925	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100%	88%	88%	87%	85%	86%	87%
A0848	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100%	88%	86%	85%	85%	86%
F0372	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100%	86%	85%	86%	86%
B3239	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100%	87%	86%	87%
TZBFL	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100%	84%	85%
D2683	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100%	86%
B2564	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100%

DS:Sequence Identity Matrix for the HIV-1 subtype D viF

Seq->	R 2	R 214	R 266	R 482	Z22B	ELI	NDK	MN011	MN012	84ZR0	57130	57143	B3239	F0372	57131	NKU30	MB205	57146	J2759	57132	E0836	G1055	E2343	57128	D2683	K0925	B2564	57140	A0848	G3509	J3222	A0741	D2355	K0995	E1361	F0573	B2187	A0334	94UG1	TZBFL
R 2	100.0%	94.1%	94.3%	91.2%	92.9%	93.2%	93.4%	91.1%	90.1%	93.0%	92.0%	92.7%	89.6%	89.2%	91.1%	91.1%	90.8%	89.9%	91.3%	89.8%	89.8%	91.3%	91.8%	90.5%	91.1%	89.6%	91.3%	89.9%	91.0%	88.9%	91.0%	91.7%	90.5%	88.9%	91.0%	89.4%	89.6%	90.3%	89.6%	88.7%
R 214	—	100.0%	92.5%	90.4%	91.7%	92.2%	92.7%	90.3%	89.2%	92.4%	90.8%	91.7%	88.7%	88.6%	90.1%	90.6%	89.6%	88.9%	90.3%	88.9%	88.9%	90.1%	91.0%	89.9%	90.1%	88.9%	90.3%	89.1%	90.1%	88.2%	90.3%	90.6%	89.6%	88.2%	90.3%	88.4%	88.6%	89.2%	88.4%	87.9%
R 266	—	—	100.0%	91.7%	93.2%	93.6%	94.3%	91.8%	90.5%	93.7%	92.4%	93.0%	89.9%	89.6%	91.1%	91.5%	91.1%	90.1%	91.3%	90.5%	90.1%	91.3%	92.0%	91.5%	91.0%	90.3%	91.8%	90.3%	91.3%	89.4%	91.3%	92.0%	90.8%	89.8%	91.3%	90.3%	89.2%	90.6%	89.9%	88.1%
R 482	—	—	—	100.0%	91.2%	92.1%	92.9%	90.4%	89.4%	91.9%	90.2%	90.9%	88.5%	88.0%	89.4%	90.2%	88.5%	88.3%	89.4%	88.7%	88.2%	89.5%	90.0%	89.9%	88.7%	88.3%	89.9%	88.8%	89.0%	88.3%	89.7%	89.9%	89.0%	87.1%	89.4%	88.2%	87.3%	89.0%	88.0%	88.3%
Z22B	—	—	—	—	100.0%	96.0%	95.8%	92.5%	91.0%	93.8%	92.9%	94.4%	91.1%	91.3%	92.0%	92.4%	92.0%	91.7%	92.4%	91.1%	91.1%	92.7%	93.2%	91.3%	92.5%	91.3%	92.9%	91.5%	92.7%	90.1%	92.0%	93.2%	92.0%	90.8%	92.7%	90.6%	91.0%	91.7%	91.5%	90.6%
ELI	—	—	—	—	—	100.0%	97.0%	93.7%	92.9%	95.1%	93.7%	95.3%	92.7%	92.0%	92.5%	93.9%	93.4%	92.5%	93.2%	92.5%	92.0%	93.6%	94.4%	92.4%	92.0%	91.8%	93.4%	92.7%	93.6%	91.0%	93.2%	93.7%	93.0%	91.3%	93.2%	91.7%	91.1%	92.5%	91.7%	91.8%
NDK	—	—	—	—	—	—	100.0%	93.7%	92.9%	96.3%	94.3%	95.5%	92.4%	91.8%	93.4%	93.7%	92.8%	92.4%	93.6%	92.2%	92.5%	93.7%	94.1%	92.9%	92.0%	92.2%	93.7%	93.0%	94.1%	91.7%	94.1%	94.1%	92.7%	91.7%	93.4%	92.0%	91.5%	92.4%	91.8%	91.8%
MN011	—	—	—	—	—	—	—	100.0%	95.3%	94.1%	92.7%	93.4%	90.5%	89.8%	91.8%	91.8%	91.1%	90.3%	91.5%	90.6%	90.3%	91.5%	91.5%	91.5%	91.0%	89.9%	91.3%	90.5%	91.1%	89.8%	91.1%	91.7%	91.7%	90.1%	91.5%	91.1%	90.3%	91.3%	90.5%	90.3%
MN012	—	—	—	—	—	—	—	—	100.0%	94.3%	91.3%	92.0%	90.5%	89.2%	91.0%	90.3%	90.1%	89.9%	90.5%	89.8%	89.4%	90.1%	90.3%	90.1%	89.8%	88.9%	90.8%	90.1%	90.8%	89.2%	89.9%	90.3%	91.3%	89.4%	90.5%	89.6%	88.9%	89.8%	89.6%	89.1%
84ZR0	—	—	—	—	—	—	—	—	—	100.0%	94.1%	95.1%	91.1%	90.3%	92.7%	92.4%	91.7%	91.5%	92.2%	91.3%	91.5%	92.2%	92.7%	92.2%	92.0%	91.7%	92.9%	91.7%	92.9%	90.6%	92.7%	92.7%	92.2%	90.6%	92.5%	91.3%	90.8%	91.8%	91.3%	90.1%
57130	—	—	—	—	—	—	—	—	—	—	100.0%	96.8%	92.0%	90.6%	92.4%	92.7%	92.2%	91.3%	93.0%	92.0%	91.8%	92.9%	93.2%	91.7%	91.1%	90.8%	92.9%	91.5%	93.0%	90.3%	93.0%	93.2%	92.0%	91.8%	93.4%	90.8%	92.0%	92.7%	92.7%	90.5%
57143	—	—	—	—	—	—	—	—	—	—	—	100.0%	92.7%	92.0%	93.9%	94.6%	94.1%	93.2%	94.6%	93.4%	93.9%	94.6%	95.1%	92.5%	93.2%	92.9%	94.8%	93.4%	94.8%	91.7%	93.9%	95.1%	93.6%	92.9%	94.6%	92.5%	92.9%	94.1%	93.0%	91.8%
B3239	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	93.9%	94.8%	95.1%	94.9%	94.1%	95.3%	94.1%	94.8%	95.5%	94.8%	93.6%	92.9%	93.9%	95.3%	93.4%	94.9%	93.2%	93.2%	94.8%	93.6%	93.2%	94.8%	92.7%	93.4%	93.6%	93.6%	92.5%
F0372	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	93.9%	93.6%	92.9%	92.0%	93.7%	92.7%	92.7%	93.8%	93.0%	91.5%	91.8%	91.8%	93.0%	92.9%	93.4%	91.7%	92.0%	93.6%	93.4%	91.1%	93.2%	91.3%	92.2%	91.8%	91.0%	91.0%
57131	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	95.1%	94.9%	94.8%	95.8%	95.1%	95.3%	95.8%	95.3%	94.1%	93.9%	94.4%	95.5%	94.4%	95.6%	93.2%	93.9%	95.1%	95.1%	94.3%	95.1%	93.4%	94.3%	94.6%	93.7%	92.2%
NKU30	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	95.5%	94.4%	96.0%	94.6%	94.8%	96.0%	96.2%	94.3%	93.6%	93.7%	96.0%	94.6%	95.8%	93.7%	94.4%	96.0%	94.8%	93.9%	95.3%	93.2%	94.3%	94.6%	93.6%	92.7%
MB205	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	98.2%	98.3%	95.5%	95.3%	96.7%	96.5%	94.3%	93.7%	94.4%	96.3%	94.6%	96.7%	93.4%	94.1%	95.9%	94.4%	94.8%	96.0%	94.1%	93.7%	94.1%	93.7%	92.7%
57146	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	96.0%	95.5%	95.1%	95.6%	95.6%	94.4%	93.2%	93.7%	96.0%	94.9%	95.3%	93.0%	93.7%	95.3%	93.9%	93.4%	94.9%	93.2%	93.4%	93.4%	94.1%	92.0%
J2759	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	96.5%	96.3%	97.2%	97.4%	95.5%	94.9%	95.1%	96.8%	95.5%	96.5%	93.6%	94.6%	96.8%	95.1%	94.6%	96.2%	93.7%	94.9%	95.3%	94.4%	93.4%
57132	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	94.9%	95.8%	95.8%	93.7%	93.2%	94.1%	95.8%	94.1%	95.5%	92.7%	93.7%	95.6%	93.9%	93.0%	94.8%	92.7%	93.7%	93.6%	93.0%	92.4%
E0836	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	96.0%	95.5%	93.9%	93.0%	94.3%	96.0%	94.3%	95.3%	93.0%	93.7%	95.8%	94.1%	93.7%	95.3%	93.0%	93.7%	94.4%	93.6%	93.0%
G1055	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	97.0%	94.8%	94.3%	94.9%	97.0%	95.1%	96.5%	93.7%	95.6%	96.5%	95.1%	94.4%	96.5%	94.1%	94.9%	94.9%	94.4%	93.2%
E2343	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	95.5%	95.1%	94.9%	96.7%	94.9%	96.7%	93.4%	95.5%	96.7%	94.9%	94.6%	96.0%	93.9%	94.1%	94.8%	93.9%	92.7%
57128	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	93.6%	93.4%	95.5%	93.7%	94.1%	92.9%	93.2%	94.3%	93.7%	93.6%	94.3%	92.9%	92.4%	93.0%	92.7%	92.2%
D2683	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	93.6%	94.8%	92.5%	93.9%	91.5%	92.0%	93.7%	93.0%	92.4%	93.2%	91.5%	92.2%	93.0%	92.4%	90.8%
K0925	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	95.8%	93.0%	94.8%	92.7%	93.2%	94.8%	93.4%	93.2%	94.3%	92.0%	92.7%	93.0%	92.9%	91.8%
B2564	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	95.5%	96.8%	93.9%	94.9%	96.3%	95.3%	94.6%	96.2%	94.1%	94.8%	94.6%	94.6%	93.2%
57140	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	95.6%	93.2%	94.3%	95.1%	94.4%	92.9%	94.4%	92.7%	93.2%	93.4%	93.6%	91.3%
A0848	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	93.9%	94.8%	96.2%	94.8%	94.6%	96.2%	93.4%	94.3%	94.4%	93.7%	92.2%
G3509	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	93.4%	93.6%	92.7%	92.4%	93.9%	91.3%	92.0%	91.7%	91.5%	89.9%
J3222	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	94.9%	93.6%	92.7%	94.1%	92.5%	92.4%	93.6%	92.7%	91.3%
A0741	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	95.8%	94.3%	95.8%	93.7%	94.8%	95.1%	94.3%	93.0%
D2355	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	92.9%	94.8%	93.2%	93.4%	94.1%	93.0%	92.2%
K0995	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	94.8%	92.2%	92.7%	93.2%	92.7%	91.1%
E1361	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	94.1%	94.1%	93.9%	94.8%	92.4%
F0573	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	91.7%	92.4%	92.4%	91.5%
B2187	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	94.1%	92.7%	91.3%
A0334	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	93.7%	92.4%
94UG1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	91.5%
TZBFL	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%

D6:Sequence Identity Matrix for the HIV-1 subtype D *vpr*

Seq->	MN011	MN012	R 2	R 214	R 286	R 482	Z228	ELI	NDK	84ZR0	57130	57143	94UG1	E1361	E2343	57140	57146	A0848	J2759	B3239	A0741	J3222	F0573	G1055	E0636	G3509	57132	D2683	K0995	A0334	D2355	57128	K0925	TZBFL	NKU30	B2187	MB205	B2564	57131	F0372
MN011	100.0%	92.4%	87.9%	87.9%	85.5%	81.4%	90.7%	90.7%	90.0%	90.3%	89.0%	87.2%	87.8%	89.6%	85.2%	86.5%	87.2%	87.2%	86.9%	88.6%	87.8%	87.9%	87.9%	87.9%	88.9%	83.5%	86.5%	89.3%	87.8%	87.2%	86.2%	86.2%	87.9%	86.5%	87.6%	87.6%	87.6%	87.9%	89.3%	87.9%
MN012	—	100.0%	91.0%	91.0%	89.0%	84.5%	94.1%	93.8%	93.1%	94.5%	91.4%	90.0%	88.8%	90.3%	87.2%	89.3%	90.3%	90.3%	89.0%	90.7%	90.3%	90.3%	89.3%	90.0%	88.3%	86.9%	89.6%	91.7%	89.8%	89.0%	88.6%	88.6%	90.3%	89.0%	88.8%	89.3%	90.3%	90.0%	91.7%	88.6%
R 2	—	—	100.0%	97.2%	92.7%	89.3%	94.5%	93.4%	94.5%	93.1%	90.7%	90.0%	89.8%	90.0%	88.3%	89.8%	90.3%	89.3%	89.3%	91.0%	91.4%	90.7%	88.6%	90.3%	88.8%	86.9%	88.6%	92.0%	90.3%	90.7%	90.7%	89.3%	90.3%	89.0%	89.0%	91.0%	89.6%	90.7%	90.0%	87.9%
R 214	—	—	—	100.0%	92.7%	87.9%	94.5%	93.4%	94.5%	93.1%	90.7%	89.6%	90.0%	90.7%	88.6%	90.0%	91.0%	89.6%	89.6%	91.4%	91.7%	91.0%	89.0%	90.7%	89.0%	86.9%	89.3%	92.4%	90.7%	91.0%	91.4%	90.0%	91.0%	89.0%	89.6%	91.4%	89.6%	90.3%	90.0%	87.6%
R 286	—	—	—	—	100.0%	89.0%	92.0%	90.0%	91.7%	91.0%	89.0%	88.3%	87.9%	88.3%	85.9%	87.2%	87.9%	86.5%	86.9%	87.9%	89.0%	88.3%	85.5%	87.6%	85.2%	84.8%	85.9%	89.0%	87.9%	87.9%	87.2%	85.9%	88.6%	88.2%	86.5%	87.9%	87.6%	88.8%	87.6%	85.2%
R 482	—	—	—	—	—	100.0%	87.9%	85.9%	87.6%	86.9%	85.2%	84.5%	83.5%	84.1%	82.4%	83.1%	84.1%	82.1%	82.4%	83.8%	84.5%	84.8%	82.4%	84.5%	81.4%	82.4%	82.4%	84.8%	84.5%	84.1%	84.5%	81.7%	83.8%	82.4%	82.4%	83.5%	83.5%	84.8%	83.5%	81.0%
Z228	—	—	—	—	—	—	100.0%	98.2%	96.5%	95.8%	94.1%	92.4%	91.4%	92.0%	90.0%	91.4%	91.7%	90.0%	90.3%	91.4%	92.4%	92.4%	89.6%	91.4%	89.8%	88.3%	90.0%	93.8%	91.4%	91.7%	91.4%	90.3%	92.0%	90.3%	91.0%	91.4%	91.7%	93.1%	92.7%	90.3%
ELI	—	—	—	—	—	—	—	100.0%	95.5%	95.6%	93.4%	91.7%	91.0%	91.7%	89.3%	92.0%	92.4%	90.0%	89.6%	92.7%	91.7%	91.7%	90.7%	92.0%	90.3%	88.3%	91.4%	93.1%	90.7%	91.7%	91.4%	90.0%	91.7%	90.7%	90.7%	92.0%	92.4%	91.7%	92.7%	91.0%
NDK	—	—	—	—	—	—	—	—	100.0%	97.2%	93.4%	92.0%	91.4%	92.0%	90.0%	91.4%	91.7%	90.7%	90.3%	92.0%	92.4%	92.4%	90.3%	92.0%	89.8%	88.3%	90.7%	93.8%	90.7%	91.7%	91.4%	90.3%	92.4%	91.0%	91.0%	92.0%	91.7%	92.4%	93.4%	89.6%
84ZR0	—	—	—	—	—	—	—	—	—	100.0%	93.8%	92.7%	92.0%	93.1%	90.7%	92.0%	92.4%	91.0%	91.7%	92.0%	93.1%	93.1%	90.7%	92.0%	90.3%	89.0%	91.0%	93.8%	91.7%	92.4%	91.4%	90.0%	91.7%	91.0%	90.7%	92.0%	92.4%	93.1%	93.4%	89.6%
57130	—	—	—	—	—	—	—	—	—	—	100.0%	93.4%	90.7%	91.4%	89.0%	90.7%	91.0%	90.3%	90.7%	90.7%	91.0%	91.4%	89.0%	91.0%	89.3%	87.2%	90.0%	91.4%	90.3%	91.0%	89.6%	87.9%	91.0%	89.6%	90.3%	91.0%	92.0%	92.7%	93.1%	90.0%
57143	—	—	—	—	—	—	—	—	—	—	—	100.0%	90.3%	91.4%	90.0%	91.4%	91.4%	90.0%	89.6%	90.7%	91.7%	91.0%	88.3%	91.0%	89.0%	86.2%	88.3%	91.0%	90.0%	91.0%	89.0%	87.2%	90.0%	88.3%	90.0%	90.3%	90.3%	91.7%	91.0%	89.0%
94UG1	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	94.8%	93.4%	94.5%	94.5%	92.4%	93.8%	95.1%	95.1%	95.8%	92.7%	95.1%	92.7%	89.6%	91.7%	93.8%	94.1%	93.8%	92.4%	90.3%	92.7%	92.0%	93.1%	94.8%	94.1%	95.5%	92.4%	93.1%
E1361	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	92.0%	92.7%	94.1%	91.4%	93.4%	93.8%	95.1%	95.1%	92.0%	94.1%	93.1%	90.3%	92.0%	94.1%	93.1%	93.1%	91.7%	89.3%	92.7%	92.4%	92.0%	93.4%	93.1%	93.1%	93.1%	91.0%
E2343	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	94.5%	93.8%	91.4%	92.4%	93.1%	93.4%	93.8%	90.7%	93.8%	91.4%	89.6%	91.0%	92.0%	92.7%	92.0%	91.7%	89.3%	92.0%	90.0%	91.4%	92.7%	92.0%	93.8%	91.0%	90.3%
57140	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	95.8%	92.4%	93.4%	94.8%	94.8%	94.1%	92.0%	94.5%	92.0%	90.0%	92.0%	93.4%	92.7%	93.8%	92.7%	91.0%	93.1%	90.3%	92.7%	94.1%	94.1%	94.5%	92.4%	93.1%
57146	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	92.7%	93.4%	95.8%	95.8%	95.8%	93.1%	96.2%	93.1%	91.7%	94.5%	94.5%	93.8%	94.1%	93.4%	92.0%	93.8%	91.7%	93.4%	94.1%	94.1%	94.5%	93.8%	92.4%
A0848	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	92.7%	93.8%	93.4%	93.1%	90.3%	93.4%	91.4%	90.0%	91.7%	92.7%	92.0%	92.4%	91.7%	89.0%	92.0%	89.6%	91.4%	93.1%	92.0%	93.4%	92.4%	91.4%
J2759	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	94.5%	95.1%	94.5%	90.7%	93.4%	91.0%	89.6%	91.0%	93.1%	92.0%	93.4%	92.4%	89.6%	91.4%	90.7%	90.7%	93.1%	91.7%	93.4%	93.1%	91.7%
B3239	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	95.8%	95.8%	94.1%	96.5%	93.4%	91.4%	94.1%	95.5%	94.5%	94.1%	93.8%	91.4%	93.4%	92.7%	93.4%	94.8%	93.8%	94.5%	94.1%	93.8%
A0741	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	97.2%	93.1%	95.5%	93.8%	92.4%	93.4%	95.8%	94.5%	94.8%	94.1%	91.4%	94.5%	93.4%	94.1%	95.1%	93.4%	94.8%	94.5%	92.4%
J3222	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	94.1%	97.2%	94.8%	92.4%	94.1%	95.8%	95.1%	95.5%	94.5%	91.0%	94.1%	93.4%	93.1%	94.8%	93.8%	95.1%	93.8%	92.4%
F0573	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	95.5%	93.1%	89.6%	92.4%	93.8%	93.4%	92.0%	91.4%	89.0%	91.7%	90.3%	91.7%	93.1%	92.7%	92.0%	92.0%	90.7%
G1055	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	95.5%	91.7%	94.1%	95.5%	94.5%	95.8%	93.8%	91.0%	94.1%	92.7%	93.8%	95.5%	93.8%	95.1%	94.1%	93.1%
E0636	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	90.3%	93.4%	93.4%	93.1%	93.4%	92.7%	89.0%	93.4%	91.4%	92.4%	93.4%	92.0%	92.0%	93.1%	90.3%
G3509	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	91.7%	91.7%	90.0%	90.7%	91.0%	85.9%	89.6%	87.9%	90.0%	90.3%	89.3%	90.0%	90.0%	88.3%
57132	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	93.4%	92.0%	91.7%	93.1%	89.0%	92.4%	91.0%	93.1%	93.4%	92.7%	92.4%	93.4%	90.7%
D2683	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	93.4%	93.8%	93.1%	90.3%	93.8%	92.4%	92.7%	95.1%	92.7%	94.1%	94.1%	91.4%
K0995	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	94.1%	93.8%	90.0%	92.0%	90.7%	92.4%	93.4%	93.8%	93.8%	91.7%	91.0%
A0334	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	93.1%	91.0%	92.0%	90.7%	92.7%	94.1%	93.1%	93.8%	92.4%	91.0%
D2355	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	91.7%	93.4%	92.0%	93.1%	93.4%	92.4%	92.7%	92.4%	90.3%
57128	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	91.7%	88.3%	90.3%	90.3%	89.0%	90.7%	90.0%	88.6%
K0925	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	92.0%	93.1%	93.4%	92.0%	93.4%	93.4%	91.0%
TZBFL	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	91.7%	92.4%	92.0%	92.4%	92.0%	90.7%
NKU30	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	95.5%	93.4%	94.1%	93.4%	92.4%
B2187	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	95.1%	94.8%	93.4%	93.4%
MB205	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	94.1%	92.7%	92.0%
B2564	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	95.1%	93.8%
57131	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	92.0%
F0372	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%

D7:Sequence identity Matrix for the HIV-1 subtype D *vpu*

Seq->	A0334	TZBFL	E1361	K0995	NKU30	57143	D2683	R 286	R 2	R 214	Z226	NDK	ELI	R 482	84ZR0	MN011	MN012	57130	F0372	J2759	MB205	E0636	57140	57146	D2355	57132	A0741	B2187	B3239	E2343	B2584	94UG1	K0925	57131	F0573	A0848	57128	J3222	G3509	G1055
A0334	100.0%	90.8%	92.7%	93.2%	88.1%	91.9%	88.9%	85.7%	84.5%	77.5%	88.1%	88.1%	86.8%	83.3%	89.8%	85.3%	87.7%	89.8%	87.7%	91.9%	90.6%	88.5%	88.1%	90.8%	78.8%	91.1%	89.8%	88.9%	91.5%	90.2%	89.4%	91.5%	89.8%	92.7%	87.7%	87.8%	90.6%	92.3%	89.4%	88.9%
TZBFL	---	100.0%	90.8%	89.4%	86.0%	87.2%	88.1%	80.8%	79.6%	72.8%	82.6%	83.4%	83.0%	78.4%	86.4%	82.8%	84.3%	86.0%	86.0%	89.8%	89.8%	86.4%	86.0%	88.5%	77.3%	88.9%	88.1%	89.4%	90.6%	89.4%	86.5%	89.8%	87.2%	89.4%	84.3%	87.0%	88.9%	89.4%	86.0%	85.1%
E1361	---	---	100.0%	94.0%	90.6%	90.6%	91.5%	84.5%	83.7%	75.5%	86.4%	86.4%	86.8%	82.5%	89.4%	85.7%	89.8%	90.2%	89.8%	94.4%	92.7%	92.3%	88.9%	91.1%	78.4%	93.2%	90.6%	91.9%	94.4%	92.7%	91.9%	93.6%	91.1%	92.3%	87.2%	89.5%	92.7%	94.0%	88.9%	88.1%
K0995	---	---	---	100.0%	88.1%	91.1%	89.4%	85.3%	85.3%	78.7%	88.9%	88.9%	88.5%	83.3%	88.9%	85.3%	88.9%	92.3%	89.4%	94.4%	92.7%	91.1%	89.4%	93.2%	78.8%	92.7%	90.2%	90.2%	92.7%	91.9%	90.6%	91.9%	89.4%	91.5%	87.2%	88.7%	90.2%	94.0%	90.2%	87.7%
NKU30	---	---	---	---	100.0%	90.2%	88.9%	81.7%	80.8%	73.4%	84.3%	85.1%	84.3%	80.4%	87.7%	82.0%	86.4%	88.5%	85.5%	90.6%	88.5%	89.4%	85.1%	88.9%	78.0%	90.8%	89.8%	87.2%	89.4%	89.4%	87.7%	90.2%	88.1%	87.7%	85.1%	86.1%	88.9%	88.5%	86.8%	85.1%
57143	---	---	---	---	---	100.0%	89.4%	84.9%	84.9%	77.1%	87.2%	88.5%	85.5%	82.5%	88.5%	83.6%	86.4%	89.8%	87.2%	91.9%	90.2%	86.5%	88.9%	92.3%	79.2%	91.1%	90.6%	88.5%	90.2%	89.4%	88.9%	90.2%	88.5%	90.6%	86.4%	86.1%	88.8%	91.1%	88.5%	86.8%
D2683	---	---	---	---	---	---	100.0%	82.9%	80.4%	74.2%	83.0%	84.7%	83.0%	79.8%	86.8%	83.2%	85.5%	88.5%	86.4%	90.2%	89.8%	88.1%	86.0%	88.9%	77.6%	90.6%	89.4%	91.1%	91.1%	89.4%	88.1%	89.8%	89.8%	89.4%	83.0%	84.9%	89.8%	89.8%	90.2%	86.4%
R 286	---	---	---	---	---	---	---	100.0%	96.3%	89.1%	90.6%	92.2%	89.0%	92.6%	87.3%	87.3%	84.9%	85.3%	82.5%	85.7%	84.9%	84.1%	82.1%	85.3%	75.7%	85.7%	83.3%	84.9%	85.7%	85.3%	84.5%	85.3%	85.3%	86.5%	83.3%	81.9%	82.5%	85.7%	81.7%	82.1%
R 2	---	---	---	---	---	---	---	---	100.0%	87.9%	89.8%	91.4%	89.4%	91.4%	85.7%	86.5%	84.9%	84.5%	82.9%	85.7%	83.3%	84.1%	82.1%	84.5%	74.6%	84.5%	82.9%	82.9%	84.1%	84.5%	82.9%	83.7%	83.7%	84.9%	82.9%	81.5%	82.5%	85.3%	80.4%	82.1%
R 214	---	---	---	---	---	---	---	---	---	100.0%	82.3%	85.1%	82.3%	83.9%	78.7%	78.7%	76.3%	77.1%	74.2%	77.1%	76.3%	75.5%	75.1%	77.5%	68.4%	77.9%	75.5%	75.9%	76.7%	77.1%	75.9%	76.7%	76.3%	77.5%	76.3%	74.8%	73.8%	76.7%	75.5%	74.2%
Z226	---	---	---	---	---	---	---	---	---	---	100.0%	95.3%	93.2%	89.8%	90.8%	86.1%	87.7%	88.1%	85.1%	87.7%	86.8%	86.0%	85.1%	86.5%	73.8%	88.1%	85.5%	86.0%	87.7%	88.1%	85.5%	87.7%	87.7%	88.1%	85.5%	84.5%	86.0%	87.2%	84.3%	83.8%
NDK	---	---	---	---	---	---	---	---	---	---	---	100.0%	94.0%	88.6%	89.8%	88.2%	88.9%	88.9%	85.1%	89.4%	87.7%	87.7%	85.1%	89.8%	74.2%	89.4%	87.2%	86.8%	88.5%	88.9%	87.2%	88.9%	87.7%	89.8%	87.7%	85.3%	85.1%	88.9%	86.0%	85.5%
ELI	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	86.9%	87.2%	84.5%	88.1%	86.0%	84.7%	86.8%	85.1%	85.5%	83.4%	87.7%	73.8%	87.2%	86.4%	86.0%	86.4%	86.8%	85.1%	86.4%	86.0%	86.4%	86.0%	83.8%	84.7%	86.4%	83.8%	83.0%
R 482	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	85.7%	84.1%	83.3%	84.9%	80.8%	82.9%	81.3%	81.3%	79.8%	84.1%	75.3%	83.3%	82.1%	81.7%	83.7%	84.1%	81.7%	84.5%	82.5%	82.9%	81.7%	81.1%	81.3%	82.5%	78.8%	79.2%
84ZR0	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	87.4%	88.5%	89.8%	85.5%	89.8%	88.9%	87.2%	85.1%	88.5%	75.0%	88.1%	86.8%	87.2%	89.8%	90.2%	87.7%	88.9%	89.8%	89.4%	85.5%	85.7%	86.4%	88.5%	85.1%	86.8%
MN011	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	92.0%	85.7%	83.2%	87.4%	86.6%	87.4%	82.4%	86.1%	74.6%	87.0%	84.1%	85.3%	88.2%	87.8%	86.6%	87.0%	86.6%	86.6%	83.6%	84.7%	84.1%	86.1%	81.5%	82.4%
MN012	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	86.8%	86.4%	90.6%	87.7%	88.5%	83.8%	87.7%	74.6%	88.5%	85.1%	86.8%	89.4%	89.8%	87.2%	87.7%	87.2%	87.2%	84.7%	84.9%	86.8%	88.9%	84.7%	84.7%
57130	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	88.5%	92.3%	90.6%	88.9%	86.8%	91.5%	76.9%	90.2%	88.9%	88.9%	91.5%	91.1%	89.4%	91.5%	88.5%	88.5%	87.2%	89.5%	88.1%	90.6%	87.2%	87.7%
F0372	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	90.6%	88.9%	88.5%	89.2%	90.6%	76.1%	88.9%	86.8%	89.4%	89.4%	89.4%	87.2%	89.4%	87.2%	88.1%	83.8%	86.1%	87.7%	90.6%	86.0%	88.3%
J2759	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	95.7%	94.0%	89.4%	93.6%	80.0%	94.0%	91.5%	92.3%	94.9%	94.0%	93.6%	94.0%	90.8%	93.6%	88.5%	91.6%	91.5%	96.1%	88.9%	90.2%
MB205	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	92.3%	88.5%	92.7%	79.6%	93.6%	89.8%	92.7%	94.4%	92.7%	92.3%	92.7%	90.2%	92.7%	86.4%	89.5%	91.1%	93.6%	88.5%	87.7%
E0636	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	87.7%	91.1%	77.6%	92.7%	88.9%	89.8%	92.7%	92.3%	91.5%	90.6%	90.6%	91.1%	87.2%	89.1%	91.1%	92.7%	86.4%	86.0%
57140	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	90.6%	75.3%	89.4%	87.7%	86.9%	87.7%	86.8%	86.8%	88.5%	86.8%	86.1%	84.3%	84.9%	86.1%	89.4%	87.2%	85.4%
57146	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	80.0%	94.0%	91.5%	91.5%	92.3%	91.5%	92.3%	93.2%	90.2%	91.5%	89.4%	89.5%	86.8%	92.3%	90.8%	88.1%
D2355	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	81.9%	79.2%	78.4%	79.2%	78.0%	78.8%	80.7%	76.5%	78.8%	75.7%	77.1%	79.2%	78.0%	76.1%	73.4%
57132	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	92.3%	92.7%	94.4%	94.0%	93.2%	94.4%	91.1%	93.2%	88.9%	89.9%	92.3%	92.7%	90.6%	86.8%
A0741	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	93.2%	91.9%	91.1%	90.2%	92.7%	88.9%	91.9%	87.2%	89.9%	90.6%	91.1%	87.2%	86.4%
B2187	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	93.2%	91.5%	91.1%	93.2%	89.8%	91.9%	88.0%	90.3%	91.5%	91.9%	88.5%	88.8%
B3239	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	96.6%	94.4%	94.9%	92.7%	93.6%	87.7%	91.2%	93.2%	93.8%	87.7%	87.7%
E2343	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	93.6%	93.2%	91.9%	92.3%	88.1%	91.6%	93.2%	92.7%	88.4%	86.8%
B2584	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	93.6%	89.4%	92.3%	87.2%	89.5%	89.4%	91.5%	85.5%	86.4%
94UG1	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	91.5%	94.4%	89.4%	91.2%	91.5%	92.7%	88.1%	86.8%
K0925	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	92.7%	86.8%	87.4%	91.1%	90.2%	87.2%	85.5%
57131	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	88.1%	89.9%	91.5%	93.2%	87.7%	88.9%
F0573	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	89.1%	88.1%	88.4%	83.4%	82.2%
A0848	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	90.3%	89.5%	83.2%	84.5%
57128	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	91.1%	86.8%	86.0%
J3222	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	90.2%	89.4%
G3509	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	85.5%
G1055	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%

D5:Sequence Identity Matrix for the HIV-1 subtype D set

Seq->	B2187	D2683	57140	K0995	B2564	E1361	57131	1	57130	A0848	F0372	TZBFL	D2355	B3239	NKU30	94UG1	57146	57132	J3222	Z226	R 2	R 214	R 286	R 482	NDK	ELI	57128	A0334	G3509	E0636	F0573	G1055	JZ759	E2343	A0741	84ZR0	57143	MN012	MN011	K0925
B2187	100.0%	94.1%	93.1%	92.1%	93.4%	92.4%	91.8%	90.1%	90.5%	91.1%	91.5%	93.4%	90.1%	93.7%	93.4%	92.4%	95.4%	92.8%	93.4%	93.7%	90.2%	83.9%	89.9%	90.5%	92.8%	92.8%	90.8%	91.1%	94.1%	93.7%	95.0%	93.7%	95.7%	93.4%	94.1%	93.1%	91.8%	90.8%	88.8%	91.8%
D2683	---	100.0%	91.1%	91.1%	91.5%	91.1%	91.1%	90.1%	91.1%	91.1%	89.2%	91.5%	88.5%	91.5%	89.8%	90.8%	92.4%	90.8%	90.8%	92.4%	88.9%	82.7%	88.9%	89.5%	92.4%	91.5%	88.8%	88.5%	91.1%	90.1%	90.5%	91.5%	93.4%	91.8%	92.8%	92.4%	91.1%	89.2%	85.9%	89.2%
57140	—	—	100.0%	94.1%	94.1%	93.4%	93.1%	90.8%	91.5%	91.5%	90.5%	91.5%	90.5%	92.8%	92.1%	91.8%	94.7%	93.1%	93.1%	93.7%	90.2%	83.9%	89.9%	90.8%	92.1%	91.8%	90.8%	91.1%	93.1%	91.5%	93.7%	92.8%	95.0%	92.8%	94.7%	93.1%	92.1%	89.2%	86.2%	90.1%
K0995	—	—	—	100.0%	93.1%	92.1%	91.1%	89.8%	90.5%	90.5%	91.1%	92.1%	92.1%	93.1%	92.4%	91.1%	93.1%	92.8%	93.1%	91.8%	90.6%	84.2%	90.2%	90.8%	91.1%	90.5%	92.4%	90.8%	92.1%	92.1%	92.8%	92.8%	94.1%	92.6%	93.4%	91.5%	91.8%	88.9%	84.3%	89.5%
B2564	—	—	—	---	100.0%	92.8%	92.4%	92.4%	92.4%	90.1%	90.1%	93.7%	90.5%	93.4%	92.4%	91.1%	92.8%	92.4%	93.1%	92.8%	89.9%	84.2%	89.6%	91.5%	91.8%	91.5%	90.5%	91.1%	92.1%	90.8%	93.1%	92.8%	95.0%	92.4%	94.4%	92.1%	91.5%	88.8%	86.9%	90.5%
E1361	—	—	—	---	---	100.0%	94.1%	90.8%	91.1%	92.1%	90.8%	91.8%	91.1%	92.8%	91.8%	91.1%	92.8%	93.1%	92.1%	93.4%	90.6%	85.2%	90.2%	90.5%	92.4%	90.8%	90.5%	91.5%	92.8%	90.5%	92.1%	93.1%	95.0%	92.8%	93.7%	92.1%	90.8%	88.2%	86.2%	90.8%
57131	—	---	—	—	—	—	100.0%	90.5%	90.8%	91.8%	90.8%	90.8%	90.1%	91.8%	89.8%	90.1%	92.1%	91.8%	90.8%	92.1%	89.3%	83.0%	88.3%	89.8%	90.5%	90.5%	88.8%	90.1%	92.4%	89.8%	91.5%	91.5%	94.1%	93.4%	92.4%	91.1%	90.8%	88.8%	85.2%	91.8%
1	—	—	—	---	—	—	—	100.0%	98.6%	90.8%	89.5%	91.1%	90.1%	90.8%	90.8%	89.5%	92.1%	89.5%	91.8%	90.5%	86.7%	81.7%	87.7%	88.5%	90.1%	88.8%	88.5%	88.5%	90.1%	87.9%	89.5%	90.8%	93.1%	90.8%	91.5%	90.5%	90.1%	86.6%	85.6%	88.8%
57130	—	—	—	—	—	—	---	---	100.0%	91.1%	89.8%	91.5%	90.5%	91.1%	91.5%	90.1%	92.1%	89.8%	92.4%	90.1%	86.7%	81.7%	87.3%	88.5%	90.1%	89.2%	89.2%	88.5%	90.5%	88.2%	89.8%	91.1%	93.4%	91.1%	91.5%	90.5%	90.5%	86.9%	85.2%	88.8%
A0848	—	—	—	---	—	—	—	—	—	100.0%	89.5%	89.8%	90.5%	91.5%	90.5%	90.8%	91.8%	91.5%	90.8%	90.8%	88.0%	81.7%	87.7%	88.5%	89.8%	89.8%	88.5%	90.1%	91.1%	89.5%	89.8%	91.8%	93.7%	92.4%	93.1%	91.1%	90.1%	87.5%	84.6%	89.8%
F0372	—	—	---	---	—	—	—	—	—	—	100.0%	91.5%	88.8%	91.5%	89.8%	87.5%	90.5%	89.8%	91.5%	90.1%	87.0%	80.8%	86.4%	86.9%	89.8%	88.8%	88.8%	89.5%	90.8%	90.1%	90.8%	90.8%	92.8%	91.1%	92.4%	89.8%	89.2%	86.6%	84.6%	89.5%
TZBFL	—	—	---	—	—	—	—	—	—	—	—	100.0%	92.1%	93.7%	91.5%	89.5%	92.1%	91.8%	92.4%	91.5%	89.3%	83.6%	88.9%	90.1%	91.1%	90.8%	90.5%	90.5%	93.1%	91.5%	91.8%	92.4%	95.0%	92.4%	93.4%	90.5%	90.8%	88.2%	85.6%	90.5%
D2355	—	---	---	—	—	—	—	—	—	---	---	---	100.0%	92.1%	91.5%	89.2%	91.8%	91.8%	92.8%	90.5%	88.0%	82.3%	88.6%	89.8%	90.1%	89.2%	91.1%	88.8%	90.8%	89.2%	89.5%	91.5%	92.8%	90.8%	91.5%	89.5%	90.5%	86.2%	83.9%	89.2%
B3239	—	—	—	—	—	---	—	—	—	—	—	—	—	100.0%	92.1%	91.1%	92.8%	91.5%	93.1%	94.1%	90.9%	85.2%	90.6%	90.8%	92.4%	92.4%	91.5%	90.8%	92.1%	91.8%	93.4%	92.4%	94.7%	93.7%	93.7%	92.4%	92.4%	88.8%	86.6%	92.1%
NKU30	—	—	—	—	—	---	—	—	—	—	—	—	—	—	100.0%	93.4%	94.7%	92.8%	94.7%	91.5%	88.3%	83.0%	89.3%	90.1%	91.8%	90.5%	91.5%	90.1%	92.4%	91.8%	92.4%	93.4%	95.0%	92.8%	93.7%	91.8%	92.1%	88.2%	86.9%	89.8%
94UG1	---	---	—	—	—	—	—	—	—	---	---	—	—	—	—	100.0%	93.4%	92.8%	91.5%	91.1%	88.3%	82.0%	88.0%	88.2%	90.8%	90.1%	88.8%	88.5%	91.5%	89.2%	90.1%	92.4%	92.1%	92.1%	91.5%	90.8%	91.5%	87.9%	85.6%	89.8%
57146	—	—	—	—	—	—	—	—	—	—	—	—	—	—	---	---	100.0%	93.7%	94.1%	94.4%	90.9%	84.5%	90.9%	91.8%	93.4%	93.1%	92.4%	91.5%	95.4%	93.4%	94.4%	94.7%	95.4%	93.7%	95.4%	93.7%	93.1%	90.1%	87.9%	90.5%
57132	---	---	—	—	—	—	—	—	—	---	---	—	—	—	—	—	—	100.0%	93.4%	91.8%	88.6%	82.3%	88.6%	89.2%	90.5%	90.1%	90.8%	90.5%	94.1%	91.5%	91.8%	93.4%	94.4%	93.4%	92.4%	91.8%	90.8%	88.2%	87.2%	89.8%
J3222	—	—	—	---	---	—	—	—	—	—	—	—	---	---	---	—	—	—	100.0%	92.1%	89.9%	84.2%	90.9%	91.8%	93.1%	91.1%	92.1%	91.5%	93.1%	91.5%	92.4%	94.1%	95.4%	92.8%	93.7%	92.1%	92.4%	90.8%	88.2%	91.1%
Z226	—	—	—	---	---	—	—	—	—	—	—	—	---	---	---	—	—	—	—	100.0%	93.2%	87.1%	93.5%	92.8%	95.0%	94.1%	91.1%	92.8%	92.8%	91.5%	93.7%	93.1%	94.7%	93.4%	95.0%	94.7%	91.8%	90.8%	87.2%	90.8%
R 2	—	—	—	—	—	—	---	---	—	—	—	—	—	—	—	—	—	---	---	---	100.0%	91.8%	95.7%	93.2%	93.5%	91.9%	90.2%	89.9%	89.3%	88.0%	89.6%	90.9%	91.2%	90.9%	92.2%	90.9%	88.6%	88.3%	83.8%	87.3%
R 214	—	—	---	---	—	—	—	—	—	—	---	---	---	—	—	—	—	—	—	—	—	100.0%	90.2%	88.6%	88.0%	85.5%	84.5%	84.2%	83.6%	82.0%	83.3%	85.5%	85.8%	85.5%	86.4%	85.2%	82.7%	82.7%	78.9%	81.1%
R 286	—	—	—	—	—	---	---	—	—	—	—	—	—	—	---	---	---	---	—	—	—	—	100.0%	94.4%	93.8%	91.9%	90.2%	89.6%	89.6%	87.7%	89.6%	91.2%	91.5%	89.3%	91.9%	90.8%	88.9%	88.6%	84.7%	87.7%
R 482	—	—	---	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	---	---	100.0%	93.1%	91.8%	90.5%	90.1%	90.5%	88.8%	90.1%	92.4%	92.1%	90.5%	92.4%	90.8%	89.8%	89.2%	85.2%	88.2%
NDK	—	—	---	—	—	—	—	—	—	---	—	—	—	—	—	—	—	—	—	—	—	---	---	---	100.0%	93.7%	90.1%	91.5%	92.4%	91.1%	92.1%	93.4%	94.4%	92.1%	95.4%	92.4%	91.1%	90.8%	86.9%	88.5%
ELI	—	—	—	—	---	---	—	—	—	—	—	—	---	---	---	---	—	—	—	—	—	—	—	—	—	100.0%	90.8%	91.1%	92.1%	91.8%	91.5%	93.7%	93.4%	92.8%	93.4%	91.5%	91.5%	89.2%	85.9%	88.8%
57128	—	---	—	—	—	—	—	—	---	---	—	—	—	—	—	—	—	—	—	—	---	---	---	---	—	—	100.0%	91.8%	91.1%	90.1%	90.8%	92.1%	92.8%	91.1%	91.5%	90.1%	89.8%	86.9%	84.6%	88.5%
A0334	—	---	---	—	—	—	—	—	—	---	—	—	—	—	—	—	—	—	—	—	---	---	---	---	—	—	—	100.0%	91.1%	90.1%	91.1%	91.8%	94.1%	91.6%	92.8%	90.5%	88.8%	88.2%	86.6%	88.5%
G3509	—	—	—	—	---	—	—	—	—	—	—	—	---	---	---	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	94.1%	93.1%	93.7%	95.0%	93.1%	93.7%	91.5%	92.1%	90.8%	88.5%	90.1%
E0636	—	—	—	---	---	—	—	—	—	—	—	—	---	---	---	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	93.4%	93.7%	93.4%	91.8%	93.1%	90.1%	91.1%	88.2%	85.2%	89.8%
F0573	---	—	—	—	—	—	—	---	---	---	—	—	—	—	—	—	---	---	---	---	---	—	—	—	—	—	—	—	—	—	100.0%	94.1%	94.4%	92.8%	93.4%	91.8%	91.5%	89.2%	86.6%	91.5%
G1055	—	—	—	---	—	—	—	—	—	—	---	---	---	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	---	---	100.0%	95.4%	94.4%	94.1%	92.1%	92.1%	90.1%	86.9%	90.8%
J2759	---	—	—	—	—	—	---	—	—	—	—	—	—	—	---	---	---	---	---	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	95.7%	96.4%	94.4%	93.1%	90.8%	88.5%	92.4%
E2343	—	—	---	—	—	—	—	—	—	—	---	—	—	—	—	—	—	—	—	—	—	---	---	---	---	---	---	---	—	—	—	—	—	100.0%	94.1%	93.7%	91.8%	89.5%	86.9%	91.5%
A0741	—	—	---	—	—	—	—	—	—	—	---	—	—	—	—	—	—	—	—	—	—	—	—	—	---	---	---	---	—	—	—	—	—	—	100.0%	95.0%	93.4%	90.8%	88.2%	90.1%
84ZR0	—	—	—	—	---	---	---	—	—	—	—	—	—	—	---	---	---	---	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	91.8%	89.8%	87.9%	89.2%
57143	—	---	---	—	—	—	—	—	---	---	---	—	—	—	—	—	—	—	—	—	—	---	---	---	---	---	—	—	—	—	—	—	—	—	—	—	100.0%	88.2%	85.9%	89.2%
MN012	—	—	—	---	---	—	—	—	—	—	—	---	---	---	---	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	92.1%	86.6%
MN011	---	—	—	—	—	—	—	---	---	---	—	—	—	—	—	—	—	---	---	---	---	---	---	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	83.3%
K0925	---	—	—	—	—	—	—	---	---	---	—	—	—	—	—	—	—	---	—	—	---	---	---	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%

D9:Sequence Identity Matrix for the HIV-1 subtype D rev

Seq->	NDK	ELI	Z226	MN012	MN011	84ZR0	B2187	D2683	94UG1	57143	B2564	TZBFL	57130	NKU30	J2759	K0995	MB205	B3239	A0648	A0334	E1361	57132	57146	57140	G3509	G1055	E0836	F0573	K0925	E2343	F0372	57131	O2355	J3222	57128	A0741	R 2	R 286	R 482
NDK	100.0%	92.3%	93.7%	91.4%	88.6%	91.7%	92.3%	92.0%	92.3%	91.4%	90.5%	89.1%	92.5%	92.8%	94.5%	89.4%	91.7%	90.5%	89.4%	90.3%	91.1%	90.8%	92.3%	90.0%	91.4%	92.8%	91.7%	92.0%	89.7%	92.8%	89.1%	88.1%	90.5%	91.7%	89.1%	92.3%	90.8%	90.5%	89.4%
ELI	---	100.0%	93.7%	90.3%	88.0%	90.5%	91.1%	90.8%	90.5%	90.5%	89.4%	87.4%	90.3%	90.0%	91.4%	88.8%	89.4%	89.4%	87.7%	87.7%	89.4%	88.8%	91.1%	88.3%	90.3%	92.0%	91.7%	89.4%	88.8%	91.7%	88.8%	88.3%	88.8%	90.0%	89.4%	91.4%	89.7%	88.6%	88.6%
Z226	---	---	100.0%	94.0%	91.1%	95.1%	94.0%	94.3%	94.0%	93.7%	91.7%	89.7%	93.7%	93.4%	94.8%	91.1%	93.4%	93.1%	91.1%	90.3%	92.3%	92.5%	94.5%	91.7%	93.1%	94.3%	94.5%	93.4%	92.0%	94.5%	90.5%	91.4%	91.1%	93.1%	90.8%	94.5%	91.1%	90.0%	89.7%
MN012	---	---	---	100.0%	93.4%	92.0%	91.7%	91.4%	91.1%	90.0%	88.8%	86.8%	90.6%	90.5%	92.0%	88.3%	90.3%	89.1%	88.8%	88.0%	89.4%	86.8%	91.7%	89.1%	92.0%	91.7%	91.1%	90.5%	89.1%	91.7%	87.7%	89.1%	87.7%	91.7%	88.0%	92.3%	89.7%	88.0%	88.6%
MN011	---	---	---	---	100.0%	90.5%	89.4%	88.8%	88.8%	88.3%	87.7%	84.3%	89.1%	88.8%	89.7%	86.8%	88.7%	88.8%	86.0%	85.7%	88.3%	88.3%	89.1%	88.8%	89.4%	88.8%	88.6%	88.0%	85.7%	88.8%	88.0%	88.3%	85.4%	89.4%	88.6%	89.4%	88.1%	85.0%	85.8%
84ZR0	---	---	---	---	---	100.0%	93.1%	93.4%	93.1%	93.1%	92.5%	89.7%	93.1%	93.4%	94.0%	90.5%	92.5%	90.5%	90.8%	88.8%	90.8%	92.0%	93.1%	91.4%	91.1%	92.3%	92.5%	91.4%	90.5%	94.0%	88.7%	90.8%	90.3%	92.5%	89.7%	93.7%	90.0%	88.0%	88.6%
B2187	---	---	---	---	---	---	100.0%	95.4%	95.1%	92.5%	92.0%	88.8%	92.5%	93.7%	94.5%	91.1%	94.0%	90.8%	90.6%	90.0%	91.1%	93.4%	94.5%	91.4%	92.5%	93.7%	94.0%	92.3%	91.4%	94.8%	88.1%	91.4%	90.5%	92.6%	90.0%	93.7%	88.4%	86.3%	88.6%
D2683	---	---	---	---	---	---	---	100.0%	95.1%	94.3%	93.4%	90.0%	93.7%	94.0%	94.8%	91.1%	94.3%	91.7%	92.0%	90.0%	92.0%	93.4%	94.5%	91.1%	92.6%	94.3%	95.1%	93.1%	92.0%	96.1%	88.7%	92.0%	91.4%	93.1%	90.8%	94.5%	90.2%	86.8%	89.4%
94UG1	---	---	---	---	---	---	---	---	100.0%	94.5%	92.8%	89.4%	93.7%	94.5%	96.0%	92.0%	94.3%	91.4%	90.8%	90.3%	93.1%	94.3%	95.4%	91.7%	93.7%	94.3%	94.3%	93.7%	93.4%	95.4%	90.3%	93.1%	91.1%	93.7%	90.5%	94.3%	89.4%	87.7%	88.3%
57143	---	---	---	---	---	---	---	---	---	100.0%	92.3%	90.3%	94.3%	94.0%	94.3%	91.1%	92.6%	92.0%	91.1%	89.4%	92.0%	92.0%	93.1%	91.7%	92.6%	93.1%	94.0%	91.7%	90.5%	94.0%	90.3%	92.5%	91.4%	93.1%	90.8%	94.0%	89.1%	87.5%	88.6%
B2564	---	---	---	---	---	---	---	---	---	---	100.0%	92.0%	94.0%	93.7%	94.0%	90.5%	91.7%	90.5%	89.7%	89.4%	92.0%	91.7%	92.6%	90.3%	91.7%	92.8%	92.8%	93.1%	91.4%	94.3%	89.7%	92.5%	91.4%	95.4%	90.5%	94.6%	89.0%	88.3%	88.3%
TZBFL	---	---	---	---	---	---	---	---	---	---	---	100.0%	91.1%	90.8%	90.8%	87.4%	89.1%	89.7%	87.1%	86.8%	89.1%	88.8%	89.7%	89.1%	88.8%	90.0%	90.3%	88.8%	87.4%	90.0%	88.3%	89.7%	90.0%	88.6%	85.2%	89.1%	88.0%	89.4%	
57130	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	94.8%	94.8%	91.4%	93.1%	92.0%	91.7%	90.0%	92.0%	92.0%	93.1%	91.4%	92.3%	94.3%	93.7%	92.5%	91.7%	94.3%	90.8%	92.3%	92.5%	94.3%	90.8%	94.5%	89.1%	88.8%	89.7%
NKU30	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	97.1%	93.4%	96.2%	92.6%	92.0%	91.7%	93.4%	93.4%	95.1%	92.8%	94.5%	95.1%	94.5%	93.4%	92.8%	94.8%	90.8%	93.1%	92.8%	95.4%	89.7%	96.8%	89.7%	88.8%	89.7%
J2759	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	94.0%	98.0%	93.7%	92.8%	92.8%	94.0%	94.0%	95.4%	92.3%	94.0%	96.0%	95.1%	94.3%	94.0%	96.8%	92.0%	93.7%	93.4%	95.4%	92.5%	96.2%	90.2%	89.1%	90.0%
K0995	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	93.7%	90.8%	88.1%	88.8%	90.8%	91.4%	92.0%	90.8%	90.5%	91.1%	92.0%	91.7%	91.1%	93.1%	90.5%	90.5%	90.3%	92.0%	90.5%	92.5%	87.5%	86.8%	87.2%
MB205	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	93.4%	91.4%	91.1%	92.5%	93.1%	94.3%	92.0%	93.7%	94.8%	94.5%	94.3%	92.5%	94.3%	90.8%	92.3%	91.1%	93.1%	90.5%	94.3%	88.6%	87.2%	88.8%
B3239	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	89.4%	90.5%	91.1%	91.1%	91.4%	91.1%	90.5%	92.0%	92.0%	91.7%	90.3%	92.3%	90.3%	91.4%	90.0%	91.1%	88.5%	94.1%	88.0%	86.1%	87.5%
A0648	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	89.4%	90.5%	91.1%	91.7%	90.8%	90.3%	91.1%	92.0%	90.3%	89.7%	92.8%	88.8%	90.5%	91.1%	93.1%	87.7%	93.1%	87.2%	85.5%	86.6%
A0334	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	89.4%	89.7%	91.1%	88.8%	90.3%	91.4%	90.5%	90.8%	88.3%	91.4%	88.3%	89.4%	89.7%	90.3%	88.8%	91.7%	85.2%	84.1%	85.5%
E1361	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	93.4%	92.8%	91.7%	93.1%	93.4%	93.4%	92.5%	91.1%	93.4%	90.8%	93.1%	91.4%	93.1%	90.3%	94.5%	88.0%	86.9%	87.5%
57132	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	93.7%	91.4%	91.7%	92.8%	93.7%	92.8%	91.4%	94.3%	89.7%	91.4%	90.3%	92.3%	88.8%	93.1%	88.3%	87.2%	87.7%
57146	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	94.3%	94.8%	94.8%	96.0%	94.0%	92.3%	95.4%	90.5%	92.3%	91.7%	94.3%	92.3%	94.8%	90.0%	88.3%	89.4%
57140	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	92.5%	91.7%	92.5%	90.5%	89.4%	92.5%	88.8%	90.8%	89.7%	92.0%	89.4%	93.4%	87.2%	86.1%	88.6%
G3509	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	94.0%	94.8%	93.4%	91.7%	93.1%	89.7%	91.7%	90.3%	92.8%	91.1%	94.0%	89.4%	88.3%	89.7%
G1055	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	95.7%	94.3%	91.4%	94.5%	90.0%	92.5%	92.0%	94.3%	91.4%	95.4%	89.7%	88.8%	89.7%
E0836	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	94.0%	92.8%	95.4%	90.8%	92.8%	92.3%	94.3%	92.5%	95.4%	90.0%	88.3%	89.4%
F0573	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	91.4%	94.5%	90.5%	92.3%	90.5%	92.5%	90.5%	93.7%	88.0%	86.9%	87.5%
K0925	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	94.8%	90.8%	92.0%	90.8%	92.0%	89.1%	92.0%	89.1%	87.5%	88.3%
E2343	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	92.3%	94.3%	92.8%	94.5%	92.3%	95.4%	91.1%	88.8%	89.2%
F0372	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	89.4%	89.4%	90.8%	89.2%	91.4%	87.2%	85.2%	86.1%
57131	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	92.0%	92.8%	89.7%	94.0%	87.5%	85.2%	86.9%
O2355	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	92.5%	90.5%	92.5%	87.5%	86.3%	87.5%
J3222	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	92.5%	95.1%	86.1%	86.0%	86.1%
57128	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	92.3%	88.8%	86.3%	86.3%
A0741	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	90.0%	88.3%	89.1%
R 2	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	95.0%	95.5%
R 286	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%	83.6%
R 482	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	100.0%
R 214	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

D10: Sequence Identity Matrix for the HIV-1 subtype D *nef*

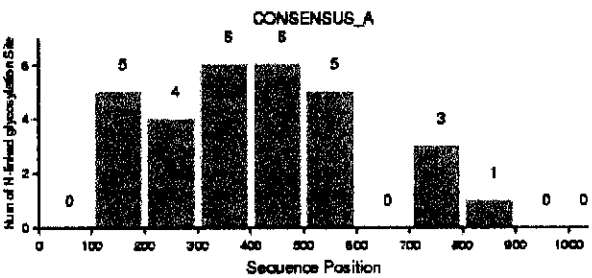
Seq->	R 2	R 214	R 482	R 286	ELI	NDK	MN011	MN012	84ZR0	E0836	A0334	E2343	F0573	57132	57130	1	57143	B2187	E1361	K0995	57146	J3222	J2759	D2355	F0372	94UG1	K0925	57128	NKU30	G3509	57140	A0741	D2683	57131	G1055
R 2	100.0%	96.3%	95.8%	95.1%	93.9%	93.1%	89.7%	84.6%	86.5%	88.7%	82.6%	88.3%	87.8%	87.9%	84.2%	83.7%	85.9%	87.9%	86.3%	82.5%	87.7%	85.2%	87.9%	88.2%	82.3%	86.3%	83.9%	82.1%	87.7%	88.4%	89.0%	88.8%	87.4%	85.8%	87.3%
R 214	---	100.0%	95.1%	94.8%	93.4%	92.9%	89.5%	84.4%	85.9%	88.9%	81.9%	88.7%	87.5%	87.3%	84.5%	84.4%	85.4%	87.4%	86.3%	82.5%	87.3%	84.9%	87.7%	88.2%	82.1%	86.0%	83.9%	81.8%	87.7%	88.2%	89.3%	88.5%	87.1%	85.7%	87.3%
R 482	—	—	100.0%	94.5%	94.0%	93.1%	90.7%	85.0%	87.3%	89.7%	82.8%	89.7%	88.4%	88.5%	84.7%	84.2%	86.0%	89.0%	87.3%	83.1%	88.8%	85.6%	88.2%	88.4%	83.5%	87.1%	84.1%	82.7%	88.4%	89.3%	89.6%	89.5%	89.0%	87.2%	88.2%
R 286	—	—	—	100.0%	92.9%	92.8%	90.8%	85.0%	85.9%	89.4%	83.5%	89.1%	87.6%	87.7%	84.0%	83.8%	86.0%	87.9%	86.6%	83.7%	88.2%	85.0%	87.6%	89.0%	82.3%	86.0%	83.8%	82.2%	88.4%	88.7%	88.8%	88.8%	87.9%	85.5%	87.1%
ELI	---	---	—	—	100.0%	93.9%	89.5%	85.4%	87.1%	90.3%	83.7%	89.9%	87.6%	88.4%	84.2%	83.7%	85.7%	88.2%	86.1%	81.7%	87.7%	85.8%	87.9%	88.8%	82.3%	86.5%	82.9%	83.4%	88.7%	88.4%	90.3%	89.8%	87.9%	85.7%	88.7%
NDK	—	—	—	---	—	100.0%	89.5%	84.9%	87.3%	89.5%	81.9%	88.4%	86.8%	87.9%	85.0%	84.7%	85.2%	87.6%	85.7%	81.7%	86.8%	85.2%	87.3%	88.0%	82.4%	86.5%	82.5%	81.6%	88.0%	87.7%	89.0%	87.9%	88.4%	85.1%	88.7%
MN011	—	—	—	—	---	---	100.0%	91.1%	88.5%	89.8%	82.0%	88.6%	87.5%	87.3%	84.0%	83.4%	85.4%	87.9%	87.1%	83.1%	86.5%	84.9%	87.6%	87.4%	83.0%	87.6%	83.3%	83.1%	87.3%	87.7%	87.1%	87.4%	86.5%	85.4%	87.1%
MN012	—	---	---	—	—	—	---	100.0%	84.2%	86.2%	82.3%	84.9%	82.8%	82.9%	83.9%	83.3%	81.6%	83.6%	82.9%	83.1%	83.2%	84.9%	84.0%	83.6%	82.0%	83.3%	81.4%	82.6%	83.2%	83.2%	83.6%	83.5%	82.3%	81.5%	83.5%
84ZR0	—	—	—	---	---	—	—	—	100.0%	85.4%	83.3%	85.3%	84.9%	84.8%	83.0%	82.2%	82.8%	85.9%	84.2%	80.1%	83.9%	82.9%	85.4%	86.4%	79.6%	85.8%	81.5%	80.9%	85.3%	85.8%	84.7%	85.0%	84.2%	86.7%	85.8%
E0836	---	—	—	—	—	---	—	—	—	100.0%	83.8%	91.1%	89.0%	89.3%	85.4%	84.8%	86.5%	89.3%	88.4%	84.5%	88.2%	87.0%	90.7%	89.5%	84.2%	89.4%	84.1%	85.5%	90.0%	89.8%	90.0%	90.1%	88.8%	87.7%	90.3%
A0334	—	—	---	—	—	—	—	---	---	—	100.0%	85.0%	84.0%	84.3%	85.9%	85.6%	81.7%	84.1%	82.6%	83.5%	82.9%	85.3%	84.1%	83.1%	81.6%	81.9%	81.3%	83.3%	84.9%	84.1%	83.8%	84.1%	83.7%	85.0%	83.7%
E2343	—	—	—	—	—	---	—	—	—	—	—	100.0%	89.0%	90.1%	85.1%	84.8%	86.5%	89.5%	88.8%	84.8%	89.2%	86.8%	90.1%	90.7%	84.7%	88.4%	84.5%	84.8%	90.7%	90.9%	89.8%	89.8%	89.2%	88.3%	90.3%
F0573	---	---	—	—	—	---	—	—	—	—	—	—	100.0%	92.8%	87.2%	87.0%	89.1%	91.7%	90.1%	86.7%	89.7%	89.1%	91.2%	90.7%	85.4%	90.7%	86.9%	86.1%	91.2%	92.3%	91.3%	90.5%	90.1%	88.2%	91.0%
57132	—	—	—	---	—	—	—	—	---	---	—	—	—	100.0%	87.1%	86.7%	89.0%	92.6%	89.1%	85.7%	90.5%	88.6%	92.3%	91.5%	84.3%	89.9%	86.4%	86.7%	92.1%	92.7%	92.7%	91.1%	91.6%	88.2%	91.6%
57130	—	—	—	—	---	—	—	—	—	—	—	—	---	—	100.0%	98.7%	86.1%	88.4%	85.4%	87.0%	86.2%	88.7%	86.7%	87.3%	86.1%	86.4%	86.2%	84.7%	86.5%	86.8%	87.1%	86.7%	87.3%	83.1%	87.3%
1	—	---	---	—	—	—	—	---	---	—	—	—	—	—	---	100.0%	86.3%	88.1%	85.3%	86.8%	85.9%	88.4%	86.4%	86.7%	86.7%	86.1%	86.2%	84.4%	86.5%	86.5%	86.5%	86.1%	87.0%	82.8%	87.0%
57143	—	—	—	---	---	—	—	—	---	---	---	—	—	—	—	—	100.0%	89.0%	86.0%	84.1%	87.5%	86.0%	88.1%	86.8%	82.4%	87.3%	84.6%	82.6%	88.3%	88.9%	88.1%	87.2%	88.1%	84.5%	87.5%
B2187	---	—	—	—	---	---	—	—	—	—	—	—	---	---	—	—	—	100.0%	90.0%	86.3%	91.9%	89.0%	91.8%	92.7%	86.4%	91.3%	87.6%	86.8%	92.4%	91.8%	91.6%	91.5%	91.1%	87.2%	91.6%
E1361	—	---	---	—	—	—	—	---	—	—	—	—	—	—	—	---	---	—	—	100.0%	86.6%	90.5%	88.1%	91.1%	91.3%	85.3%	89.1%	86.2%	85.0%	90.8%	91.1%	90.2%	89.9%	89.5%	87.4%	91.3%
K0995	—	—	—	---	---	—	—	—	---	---	---	—	—	—	—	—	—	—	---	---	100.0%	85.4%	88.1%	86.2%	86.0%	84.4%	85.9%	85.1%	84.9%	85.9%	86.0%	85.3%	85.9%	85.6%	82.7%	85.4%
57146	---	—	—	—	—	---	—	—	—	—	—	—	---	---	—	—	—	—	—	—	—	100.0%	89.0%	91.1%	91.8%	85.0%	90.3%	85.9%	86.2%	91.0%	90.7%	91.9%	89.5%	90.8%	87.9%	91.1%
J3222	—	—	---	—	—	—	—	---	---	---	—	—	—	—	—	---	---	---	—	—	—	—	100.0%	89.2%	89.3%	88.3%	88.3%	88.6%	89.6%	89.3%	89.8%	88.9%	89.0%	88.0%	85.2%	90.0%
J2759	—	—	—	—	---	—	—	—	—	—	---	---	---	—	—	—	—	—	—	—	—	—	—	100.0%	93.1%	86.3%	91.3%	86.5%	86.8%	92.7%	91.3%	91.8%	92.3%	91.9%	88.0%	91.9%
D2355	---	---	—	—	—	—	---	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	86.7%	90.5%	87.2%	87.3%	92.4%	92.3%	92.3%	92.8%	91.5%	88.5%	92.6%
F0372	—	—	—	---	—	—	—	—	---	---	—	—	—	—	—	—	—	—	---	---	—	—	—	—	—	100.0%	86.5%	87.4%	87.3%	86.0%	86.0%	84.4%	86.0%	84.6%	82.7%	86.1%
94UG1	---	—	—	—	---	—	—	—	—	—	—	---	---	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	86.7%	86.6%	90.8%	91.6%	89.4%	89.4%	88.6%	86.5%	90.7%
K0925	—	—	—	—	---	—	—	—	—	—	—	---	---	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	87.8%	85.9%	87.0%	85.9%	86.4%	85.3%	83.1%	86.4%
57128	—	---	—	—	—	—	—	---	—	—	—	—	—	---	---	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	86.8%	86.5%	86.7%	85.8%	85.4%	82.9%	86.1%
NKU30	—	—	—	---	—	—	—	—	---	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	93.2%	92.7%	91.8%	90.5%	88.2%	91.5%
G3509	—	—	—	—	---	—	—	—	—	—	—	---	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	---	100.0%	92.6%	93.1%	91.3%	89.4%	92.6%
57140	—	---	---	—	—	—	—	---	---	—	—	—	—	—	---	---	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	92.9%	91.9%	88.3%	92.6%
A0741	—	—	—	---	—	—	—	—	---	—	—	—	—	—	—	—	—	—	---	---	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	91.1%	87.5%	92.1%
D2683	---	—	—	—	—	---	—	—	—	—	—	---	—	—	—	—	—	—	—	—	—	—	—	—	—	---	---	---	---	---	---	—	—	100.0%	88.5%	91.8%
57131	—	---	---	—	—	—	---	---	—	—	—	—	—	---	---	---	---	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100.0%	89.1%
G1055	—	—	—	---	---	—	—	—	---	---	—	—	—	—	—	—	—	—	---	---	---	---	---	—	—	—	---	—	—	—	—	—	—	---	---	100.0%
B2564	—	—	—	---	—	—	—	—	---	---	—	—	—	—	—	—	—	—	—	---	—	---	---	---	---	---	---	---	—	—	—	---	---	---	---	---
B3239	---	—	—	—	—	---	—	—	—	—	—	---	---	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—

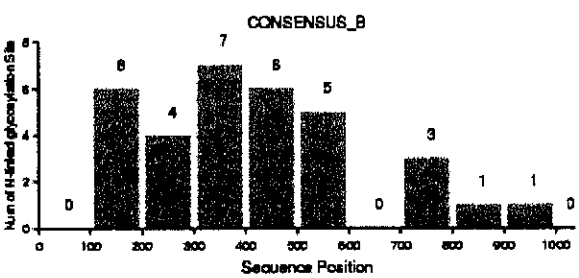
Appendix E

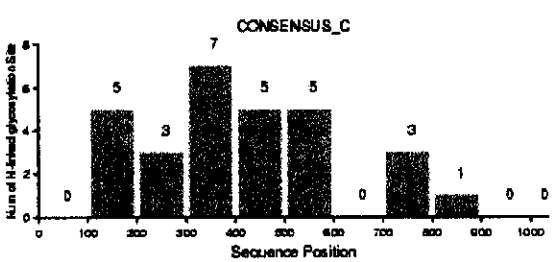
The glycosylation graphs of the HIV-1 subtype A-K consensus *env* sequences are given. The x-axis indicates the position of the glycosylation sites in the sequence and the y-axis represents the number of N-linked glycosylations sites.

Appendix E

HIV-1 subtype A-K consensus glycosylation patterns

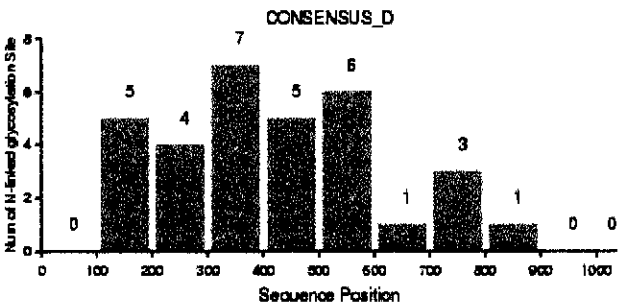


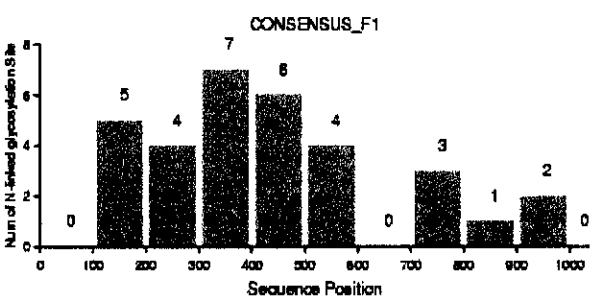


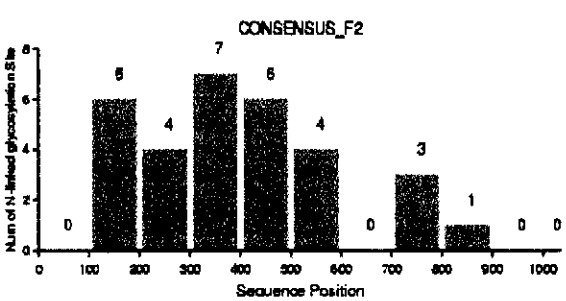


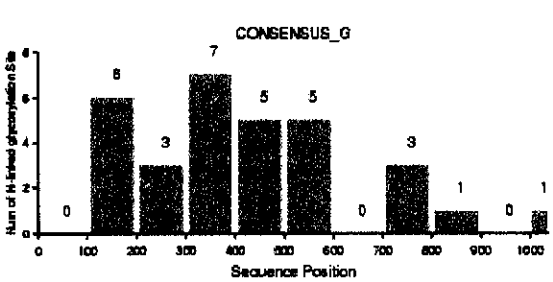
Appendix E:

HIV-1 subtype A-K consensus glycosylation patterns









Appendix E:

HIV-1 subtype A-K consensus glycosylation patterns