Molecular characterization of non-subtype C and recombinant HIV-1 viruses from Cape Town, South Africa

by

Eduan Wilkinson

Submitted in fulfilment for the degree

MSc in BioMedical Science

at

Stellenbosch University

Division of Medical Virology, Department of Pathology

Faculty of Health Science

Supervisor: Prof Susan Engelbrecht

Date: 1 December 2008

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

_____

Name in full

_____/_____/_____

Date

Acknowledgements

I would hereby like to extend my fullest gratitude to the following people and institutions; without which this thesis would not have been possible:

Prof. Susan Engelbrecht, the promoter of this study, for her expert advice, guidance and assistance throughout the course of this project.

Annette Laten for all her assistance in the laboratory and with the sequences.

To my parents, Edward and Elizabeth Wilkinson, for always believing in me and supporting me with my studies.

To all the people in the Division of Medical Virology, Department of Pathology, Faculty of Health Sciences, University of Stellenbosch for their moral support throughout the course of the two years.

To my friends for all their support and encouragement throughout my studies.

To the Poliomyelitis Research Foundation (PRF) and the National Research Foundation (NRF) of South Africa for the generous bursary money.

To the Harry Crossly and the Poliomyelitis Research Foundation for funding the project.

Abstract

HIV-1 was first diagnosed within South Africa in 1982. In the 1980's homosexual transmission dominated the HIV-1 epidemic within the country. In the late 1980's the second HIV-1 epidemic was recognized amongst heterosexual individuals. Today heterosexual transmission of HIV-1 dominates the epidemic in South Africa. Subtype C HIV-1 is responsible for the overwhelming majority of heterosexual infections. An estimated 95% of all infections in the country are thought to be subtype C related. To date only a few papers have been published on non-subtype C HIV within the country. This study characterized subgenomic and near full-length sequences of non-subtype C HIV-1 viruses from the Cape Town area.

The *gag* p24, *pol-integrase,* and *env* gp41 regions of 11 of the 12 samples were characterized by amplification and direct sequencing. Phylogenetic analysis of the sequenced data, with online subtyping tools (REGA and jpHMM) and the drawing of NJ-trees revealed the presence of subtype A1, B, F1 and recombinant viral forms such as AD, AG and AC. One of the isolates was classified as a subtype C and was included for control purposes.

Near full-length characterization of four of the samples were attempted, through full genome PCR amplification and sequencing. Analysis of sequenced data with the use of subtyping-, recombination identification, and tree drawing tools revealed a subtype B, and A1 isolate. The other two isolates were identified as possible AC and AD recombinants.

The data that was generated will greatly improve our knowledge of non-subtype C isolates circulating within South Africa. Due to the possible impact that the high degree of genetic variation that HIV may have on vaccine design and development and ARV treatment and HIV diagnosis, ongoing research of the epidemiology and spread of HIV within South Africa are needed.

Opsomming

MIV was in 1982 vir die eerste keer in Suid Afrika gediagnoseer en was hoofsaaklik deur homoseksuele kontak oorgedra. Aan die begin van die 1990's is `n tweede MIV epidemie gewaar onder heteroseksuele individue. Heteroseksuele oordrag van die virus domineer tans die MIV epidemie in Suid Afrika en is meestal subtipe C verwant. Subtipe C, MIV-1 is verantwoordelik vir 95 persent van alle infeksies in die land. Tot hede is slegs `n paar publikasies oor die nie-subtipe C epidemie in die land gepubliseer. Die huidige studie was gemik op die karakterisering van subgenomiese en vollengte genome van nie-subtipe C MIV isolate van die Kaapstad omgewing.

Die *gag* p24, *pol-integrase* en *env* gp41 subgenomiese fragmente van 12 monsters was gekarakteriseer deur amplifikasie en DNS nukleotied volgorde bepaling. Filogenetiese analise deur middel van subtipering (REGA en jpHMM aanlyn subtiperings programme) asook NJ-filogenetiese bome van die data het die teenwoordigheid van subtipe A1, B, en F1, asook verskeie rekombinante viruse insluitende AG, AD en AC vorme aangedui. Een van die isolate was geklassifiseer as `n subtipe C maar is in die studie ingevoeg vir kontrole doeleindes.

Vollengte karakterisering van 4 uit die 12 isolate was ook gedoen deur vollengte genoom amplifikasie en DNS nukleotied volgorde bepaling. Tydens die analisering van die DNS volgorde data, deur middel van aanlyn subtipering, rekombinasie identifikasie (Simplot en RIP), en filogenetiese boom konstruksie programme is twee isolate geidentifiseer as subtipe B en A1 MIV-1 viruse. Die ander twee isolate was as moontlike AC en AD rekombinante geklassifiseer.

Die data van nie-subtipe C MIV isolate sal ons kennis van die nie-subtipe C epidemie in Suid Afrika versterk. As gevolg van die impak wat die hoë graad van genetisie variasie van MIV op die ontwikkeling van entstowwe, sowel as die diagnose en behandeling van pasiente kan hê, is verdere navorsing in die epidemiologie van die MI-virus in Suid Afrika nodig.

Table of Contents

List of abbreviations

˚C – degrees Celsius

µl – micro liters

A – Adenine

AIDS – Acquired Immunodeficiency syndrome

ARV – Antiretroviral

BEAST – Bayesian Evolutionary Analysis Sampling Trees

BLAST – Basic Local Alignment Search Tool

bp – base pairs

C – Cytosine

CA – California

CCR5 – Chemokine (C-C motif) receptor 5

CD4 – cluster of differentiation 4

CDC – Center for Disease Control and Prevention

cDNA – complimentary deoxyribonucleic acid

CRF – Circulating recombinant form

CTL – Cytotoxic T lymphocytes

DDBJ – DNA Database of Japan

DNA – Deoxyribonucleic acid

dNTP – Deoxyribonucleotide triphosphates

DOE – Department of Energy

DRC – Democratic Republic of the Congo

EMBL – European Molecular Biology Laboratory

F – Forward Primer

g – Gravitational constant

G – Guanine

gp – glycoprotein

GTR – General Time Reversible Model

HIV – Human Immunodeficiency Virus

HIV-1 – Human Immunodeficiency Virus type 1

HIV-2 – Human Immunodeficiency Virus type 2

HKY – Hasegawa, Kishino, Yano Model

HMA – Hetroduplex Mobility Assay

HSRV – Human Spuma Retrovirus

HTLV-I – Human T-cell leukemia virus type I

HTLV-II – Human T-cell leukemia virus type II

HTLV-III – Human T-cell leukemia virus type III

JAMA – Journal of American Medical Assosiation

jpHMM – jumping profile Hidden Markov Model

kbp – kilo base pairs

KS – Kaposi's sarcoma

LANL – Los Alamos National Laboratory

LAS – Lymphadenopathy syndrome

LAV – Lymphadenopathy virus

LAV-2 – Lymphadenopathy virus type 2

LTR – Long Terminal Repeats

M – Molecular marker

MEGA – Molecular Evolutionary Genetics Analysis

$MgCl_2$ – Magnesium Chloride

MHC – Major Histocompatibility Complex

mM – millimoles

MMWR – Morbitity and Mortality Weekly

mRNA – messenger RNA

NCBI – National Center for BioInformatics

NFLG – Near full-length genome

NIAID –National Institute of Allergy and Infectious Diseases

nm – nanometers

NM – New Mexico

NNI – Nearest Neighbor Interchange

NRF – National Research Foundation

NY – New York

OTU – Operating Taxonomic Units

p – Protein

PAUP – Phylogenetic analysis using Parsimony* and other methods

PBMC – Peripheral Blood Mononuclear Cell

PCP – *Pneumocystis carinii* pneumonia

PCR – Polymerase Chain Reaction

RDP – Recombination Detection Program

PHYLIP – PHYLogeny Inference Package

RIP – Recombination Identification Program

R – Reverse Primer

PR – Protease

PRF – Poliomyelitis Research Foundation

rev – regulator gene

RNA – Ribonucleic acid

RSA – Republic of South Africa

RT – Reverse Transcriptase

SA – South Africa

Simplot – Similarity Plot

SIV – Simian Immunodeficiency Virus

SPR – Subtree Pruning and Regrafting

ssRNA – single stranded RNA

T – Thymine

TAC – Treatment Action Campaign

tat – transactivation gene

TB – Tuberculosis

T_A – Annealing temperature

T_M – Melting temperature

UK – United Kingdom

UNAIDS – Joint United Nations program on HIV and AIDS

UPGMA – Unweighted Pair Group Method with Arithmetic Means

URF – Unique Recombinant Form

US – United States

USA – United States of America

UV – Ultraviolet light

vif – Viral infectivity factor

vpr – viral protein R

vpu – viral protein U

WHO – World Health Organisation

WI – Wisconsin

List of Figures

Page

Page

List of Tables

Page

CHAPTER ONE – Introduction and Literature Review

Table of Content **Page**

CHAPTER ONE

INTRODUCTION

The clinical condition known as acquired immunodeficiency syndrome (AIDS) is caused by the human immunodeficiency virus (HIV). The virus exists in two distinct forms, HIV-type 1 (HIV-1) and HIV-type 2 (HIV-2). With the help of epidemiological and molecular tools closer analysis has revealed that HIV-1 has spread all over the world, whereas HIV-2 is mostly confined to areas of West Africa [Essex and Mboup, 2002; Levy, 2007]. HIV-1 can be divided into three groups: group M (major), group N (non-M or non-O) and group O (outlier) [Peeters, 2001]. Group M HIV-1 is responsible for the majority of infections worldwide and can be subdivided into a large variety of subtypes and circulating recombinant forms (CRF's) [http://www.hiv.lanl.gov/]. The genetic similarity between HIV-1 and other closely related viruses, such as HIV-2 and SIV, can be used as genetic markers in the study of the virus with the help of molecular and phylogenetic techniques.

The July 2008 UNAIDS report, estimates that 34 million adults and 2.3 million children are living with HIV/AIDS, with 2.7 million new infections each year. The report also estimates that 2 million people die, due to AIDS related deaths each year [UNAIDS, 2008]. Although HIV and AIDS are found in all parts of the world, some areas are more severely affected than others. Of the 36 million people infected worldwide, an estimated 24.1 million infected individuals live in sub-Saharan Africa where, in some places, more than one in five adults can be infected [UNAIDS, 2008]. Though sub-Saharan Africa is most severally affected, the epidemic is spreading more rapidly in Eastern Europe and Central Asia, where the rate of new infections increased by as much as 50% between 2004 and 2008 [UNAIDS, 2008].

In South Africa (SA) the HIV-I epidemic was initially associated with the homosexual population [Sher, 1989]. In the 1980's, HIV-1 was isolated form homosexual men, who introduced the virus into SA from other countries [Becker *et al,* 1985]. These initial isolates were later identified as HIV-1

subtype B and D viruses [Becker *et al,* 1995; Engelbrecht *et al,* 1995]. Today the HIV-1 epidemic has spread widely amongst the heterosexual population of the region with as many as 6.2 million infected people in SA alone [UNAIDS, 2008]. Subtype C HIV-1 is responsible for nearly 52% of HIV infections worldwide and is commonly found in parts of the Indian subcontinent and eastern and southern Africa. In SA, it can account for as much as 95% of all infections [Ariën *et al*, 2007]. Though HIV-I subtype C still holds a dominant position within southern Africa, non-subtype C HIV-I infections have been growing in importance over the past couple of years, which may impact a wide spectrum of fields e.g. antiretroviral treatment [Spira *et al*, 2003], diagnostics and the development of an effective vaccine [Buonaguro *et al,* 2007; Peeters *et al,* 2003]. In the following section the history of the HIV pandemic, the genetic diversity, the virus structure, as well as phylogenetic methods used in analysis of HIV will be reviewed.

LITERATURE REVIEW

1.1. History

1.1.1 The HIV pandemic

Kaposi's Sarcoma (KS) was a very rare form of a relatively benign cancer that mostly affected the elderly [Gange and Jones, 1978; Penn, 1979]. By March 1981 eight cases of a more aggressive form of KS was reported amongst young homosexual men from New York [Hymes *et al*, 1981, Friedman-Kien *et al*, 1981]. At this time there were increases all over the United States (US) in the number of cases of a rare lung infection, *Pneumocystis carinii* pneumonia (PCP) [Gottlieb *et al*, 1981; Masur *et al,* 1981]. In April of 1981 these increases were noticed by a young drug technician working at the Center of Disease Control and Prevention (CDC) in Atlanta, when a doctor treating a young patient with PCP requested a refill for the drug used in the treatment of PCP patients [CDC, 1981]. This was very unusual as most patients responded to medication within 10 days of treatment or they passed away. Afterwards a number of theories were developed about the possible cause of these opportunistic infections and cancers.

By the end of 1981 it was clear that the disease was affecting other population groups, when the first cases of PCP infection were reported in intravenous drug users [Shilts, 1987]. Shortly afterwards, the first reported cases of these opportunistic infections were documented in continental Europe and the UK. By the beginning of the next year the disease still did not have a name. By June 1982 a report suggested that the disease might be caused by an infectious agent that was sexually transmitted, after a small cluster of cases amongst homosexual men in southern California was reported. By October 1982, 452 cases from 23 states had been reported to the CDC [CDC, 1982]. It soon became apparent that the disease was also occurring amongst heterosexuals. By this time reported cases had been observed in homosexual and heterosexual people, intravenous drug users, certain blood transfusion recipients [Curran *et al,* 1984; Jaffe *et al,* 1984], some organ transplant recipients [Curran *et al*, 1984], a few newborn infants [Rubinstein *et al,* 1983; Oleske *et al,* 1983] and international travelers from African decent [Clumeck *et al,* 1983].

The acronym AIDS was beginning to be used on government level, the press and in scientific journals. Doctors thought AIDS was an appropriate name because people acquired the condition rather than inherited it, resulting in a deficiency within the immune system, with a number of manifestations. By this time still very little was known about the transmission of the infectious agent and public anxiety started to grow. By the end of 1982 a 20-month old child who received multiple transfusions of blood and blood products died from infections related to AIDS [Curran *et al*, 1984]. Meanwhile in Uganda, doctors were beginning to see the first cases of a new, fatal wasting disease which came to be known locally as “slimming-disease” [Serwadda *et al,* 1985].

In May 1983 doctors and scientist at the Pasteur Institute in France reported the isolation of a new retrovirus from the lymph node of a man with persistent Lymphadenopathy Syndrome (LAS), which they suggested might be the causative agent of AIDS and was subsequently named Lymphadenopathy virus (LAV) [Barre-Sinoussi *et al,* 1983]. At the same time, reports from Europe suggested that two rather separate AIDS epidemics were occurring. In

the UK, Germany and Denmark, the majority of people with AIDS were homosexual men with a history of sexual encounters with American nationals. In France and Belgium however, AIDS occurred mainly in migrants from former African colonies or those with links to areas in Africa [Clumeck *et al,* 1983]. Due to the fact that these patients had no history of blood transfusion, homosexuality or intravenous drug use, European and American scientists set out to discover more about the occurrence of AIDS in Africa [Weller *et al*, 1984]. By the end of 1983 the World Health Organization (WHO) reported that the disease were present in the United States, Canada, fifteen European countries, Haiti, Zaire (now the Democratic Republic of the Congo), seven Latin American countries, Australia and Japan [WHO, 1983].

A year after the team of Barre-Sinoussi [Barre-Sinoussi *et al,* 1983] isolated the Lymphadenopathy virus at the Pasteur Institute, Robert Gallo and his colleagues postulated that a variant T-lymphotropic retrovirus, which they called HTLV-III, might be the causative agent of AIDS [Gallo *et al,* 1984]. By 1984, Levy and co-workers, found the same viral agent that Barre-Sinoussi [Barre-Sinoussi *et al,* 1983] discovered an year earlier in one of the samples of an infected patient [Levy *et al*, 1984]. Ratner and co-workers independently confirmed that the new variant of HTLV-III was the causative agent of AIDS and also published the first full sequenced genome of the virus [Ratner *et al*, 1985a; Ratner *et al*, 1985b]. Medical data and records of patients suffering from opportunistic infections, from Central African countries, suggested that the disease was already present within the region in the 1970's and asserted the claims that it might have arisen from the region [Quinn *et al,* 1986]. Though HIV/AIDS was initially associated with people from particular risk groups, such as homosexuals and intravenous drug users, heterosexual transmission is now responsible for the majority of new infections [Osmanov *et al*, 2002; Esparza and Bhamaraparvati, 2000].

By 1986 two new retroviruses were isolated from patients with AIDS like symptoms from West Africa. Closer analysis of these viruses (LAV-2 or later called HIV-2), showed that although both caused immunodeficiency and AIDS in infected individuals, HIV-1 and HIV-2 differ in their natural history of

infection and pathogenicity [De Cock *et al,* 1993; Pepin *et al,* 1991]. Since then it has been shown that both HIV-1 and HIV-2 shares structural, genetic and biological properties and cause CD4 cell depletion in infected individuals [Markovitz, 1993].

1.1.2 The origin of HIV

HIV is a member of the Lentivirus subfamily of retroviruses (Retroviridae). Since HIV was established as the etiological agent of AIDS an estimated 60 million people have been infected with HIV-1 worldwide [UNAIDS, 2008]. Though HIV/AIDS only came under the attention of humans in the early 1980's recent scientific discoveries dates the existence of the virus in humans as far back as the 1930's [Hahn *et al*, 2000; Korber *et al*, 2000]. A plasma sample from 1959 that was taken from a patient from the Central African country of Zairë (now the DRC), gives credible evidence that the disease has been in humans for some time longer than we first thought and also suggests that the epidemic might have originated in Africa [Nahmias *et al*, 1986].

Considerable evidence for a simian ancestor for HIV exists today. Simian immunodeficiency virus (SIV), which is also a member of the lentivirus family, is found in a large group of species of non-human primates and is related to HIV on a genomic level (Figure 1.1) [Myers *et al*, 1992]. All factors indicate that HIV originated through cross-species transmission from naturally infected primates to humans in Africa, a process commonly known as zoonosis [Hahn *et al,* 2000]. Phylogenetic analysis indicated that multiple zoonotic events, from simian species to humans, lead to the formation of two genetically distinct types of HIV (HIV-1 and HIV-2) and three main groups of HIV-1 [Gao *et al*, 1999; Papathanasopoulos *et al*, 2003a].

Molecular studies showed that HIV-1 is more closely related to primate lentiviruses from chimpanzees (SIV_{cpz}), mainly from the subspecies *Pan troglodytes troglodytes* [Gao *et al,* 1999]. Similarly HIV-2 is more closely related to SIV sequences commonly found in sooty mangabeys, *Caeoceus atys,* [Gao *et al,* 1992; Hirsch *et al,* 1989]. Chimpanzees and other non-human primates are commonly hunted for food in certain regions in Central and

Western Africa and represent an easy source for zoonotic transmission of SIV to humans [Hahn *et al,* 2000; Papathanasopoulos *et al*, 2003a; Essex and Kanki, 1998]. Sooty mangabeys are also commonly used as a food source and in some cases domesticated as house pets in parts of Western Africa [Peeters *et al,* 2003]. Further support for the zoonotic infections of humans is that natural SIV infections in their simian hosts fail to induce a state of disease, which indicates that the virus has adapted itself to the host or that they co-exist in a symbiotic way [Cichutek and Norley, 1993; Rey-Cuille *et al*, 1998; Silvestri *et al*, 2003].

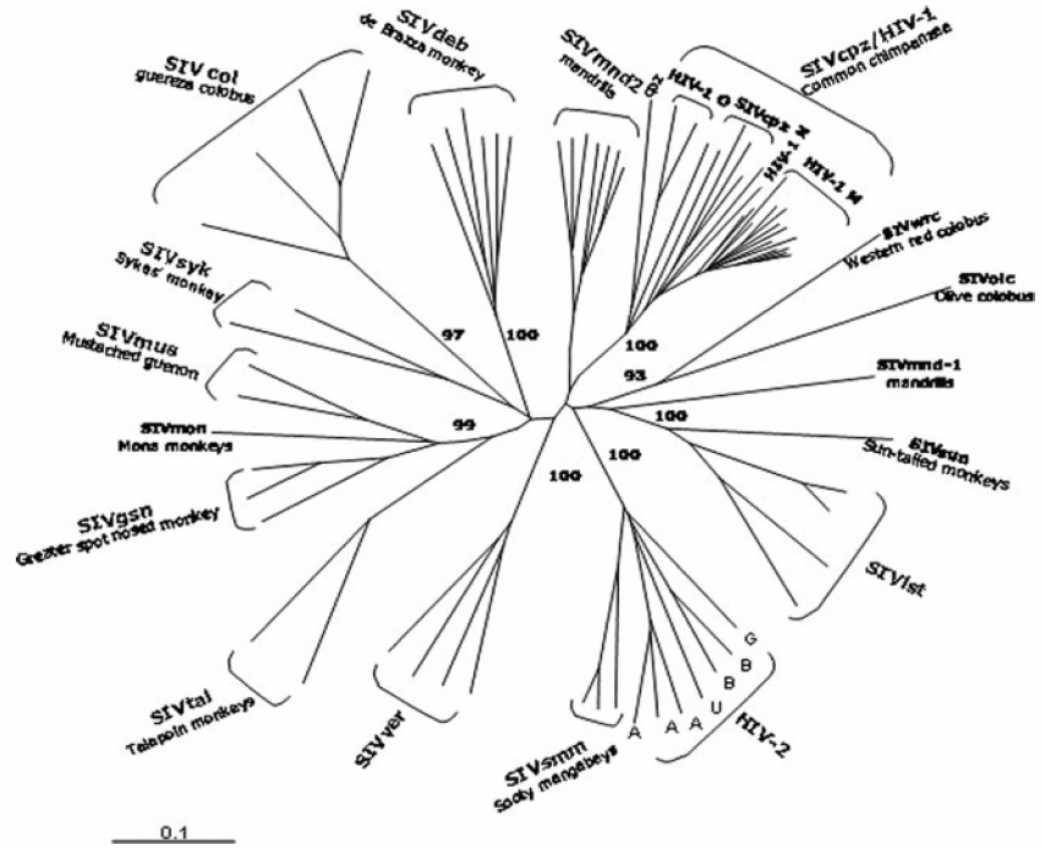


Figure 1.1: A phylogenetic tree depicting the evolutionary relationship between: HIV-1 and its major groups; HIV-2; and several different SI-viruses isolated from different primates on the African continent. Bootstrap values greater than 90 percent are included in the tree [Adapted from Buonaguro *et al,* 2007].

The timing of the zoonotic events leading to the rise of HIV in man has been a question for debate. Advanced phylogenetic methods have estimated 1930 as the time of the last common ancestor of the HIV-1 group of viruses [Korber *et*

al, 2000; Salemi *et al,* 2001]. These estimations are based on the assumption of a molecular clock, with genetic change in a linear function and substitution occurring according to a Poisson distribution. Similarly the best estimate of the most recent ancestor of the HIV-2 group was estimated to be 1940 for HIV-2 group A and 1945 for HIV-2 group B, plus or minus 16 years [Lemey *et al*, 2003].

1.2 The HIV-1 virus

1.2.1 Retroviruses

The first discovery of a retrovirus was made as far back as 1910 at the Rockefeller Institute of Medical Research in New York. It was called avian sarcoma virus and induced tumors in muscle, bone and other tissues of chickens [Huebner and Todaro, 1969]. Later other retroviruses were identified that could induce the same symptoms in mice and other mammals [Essex *et al,* 1975; Hardy *et al,* 1973]. Retroviruses possess a unique enzyme called reverse transcriptase (RT), which uses the viral RNA as a template for making a DNA copy, which is then integrated into the host cell nucleic acid [Bebenek *et al,* 1989; Boyer *et al,* 1992]. The discovery of reverse transcriptase were profound in the field of biology as it overturned the central belief of molecular biology, that genetic information flows only in one direction (which is from DNA to RNA and on to protein synthesis). Retroviruses are also diploid, which means that there is a constant opportunity for recombination to occur when different parental genomes are packaged in the same virus particle [Robertson *et al,* 1995]. Initially these retroviruses were regarded to be of little importance because they were only associated with organisms other than humans. That all changed with the discovery of human associated retroviruses.

To date several different human retroviruses have been identified which can be divided into three families including; Lentivirinae (HIV-1 and HIV-2), Oncovirinae (HTLV-I and HTLV-II) and Spumavirinae. The human spumavirus (HSRV) was the first isolated human retrovirus [Achong *et al,* 1971]. Spumaviruses are found worldwide and can be isolated from a wide range of

species, from monkeys to cattle [Flügel, 1991]. To date no known form of pathology of HSRV in humans has been found. The Oncovirinae family all shares the ability to induce cancerous conditions in its hosts, including lymphomas, leukemia, carcinoma and other forms of cancer. Two species of this family of retroviruses have been identified in humans to date, HTLV-I and HTLV-II. In 1976 a retrovirus was isolated from the lymph node of an ATL-patient (adult T-cell leukemia-lymphoma patient) in Japan [Uchiyama *et al,* 1977]. The etiological agent of ATL was named human T-cell leukemia virus Type - I or HTLV-I [Ammann *et al,* 1983; Fahey *et al,* 1984]. HTLV-I proved to be endemic in Japan, Central and South America, the Caribbean and Africa. The origin and the transmission of HTLV-I are thought to be the same as for HIV-1. Later a second virus, HTLV-II was isolated from the cells of a patient with a rare form of leukemia [Gallo, 2005]. An Africa origin was later hypothesized for HTLV-II, which then spread to other areas of the world where it is mostly associated with intravenous drug users [Vandamme *et al,* 1998]. The Lentivirinae family of retroviruses includes HIV-1, HIV-2 and other lentiviruses commonly found in non-human primates which are termed SIV. Later the fifth human retrovirus, HIV-2, was isolated from mildly immune suppressed patients in West Africa which appeared to be less pathogenic [Marlink *et al,* 1994; Kanki *et al,* 1994].

1.2.2 The Virion Structure

HIV-1 and HIV-2 belong to the group of lentiviruses [http://www.hiv.lanl.gov/] which can be distinguished from other retroviruses (such as HTLV-I and -II) by the presence of a cone-shaped nucleoid, absence of oncogenicity and the length and slow onset of clinical symptoms. The HIV particle is approximately 100 nm in size with an outer envelope of a lipid bilayer, which arises from the virus budding from the host cell. This lipid bilayer is penetrated by 72 glycoprotein spikes, the envelope (*env*) protein. The env polypeptide is composed of two subunits: the outer glycoprotein knob (gp120) and a transmembrane portion (gp41) which connects the knob to the virus lipid envelope [Levy, 2007] (Figure 1.2).

On the inside of the lipid bilayer the envelope is lined with a matrix protein (p17). Also present within the lipid envelope are cellular proteins such as MHC class 1 and class 2 antigens. In HIV-1 the lipid envelope encloses an icosahedral shell of protein (p24) which in turn encloses a cone-shaped protein core (p7 and p9 proteins). Within the cone-shaped core are two molecules of ssRNA in the form of a ribonucleoprotein. Bound to the diploid, positive-sense ssRNA are multiple copies of reverse transcriptase (RT), integrase (for genome integration into the host cell nucleic acid) and protease enzymes [Levy, 2007].

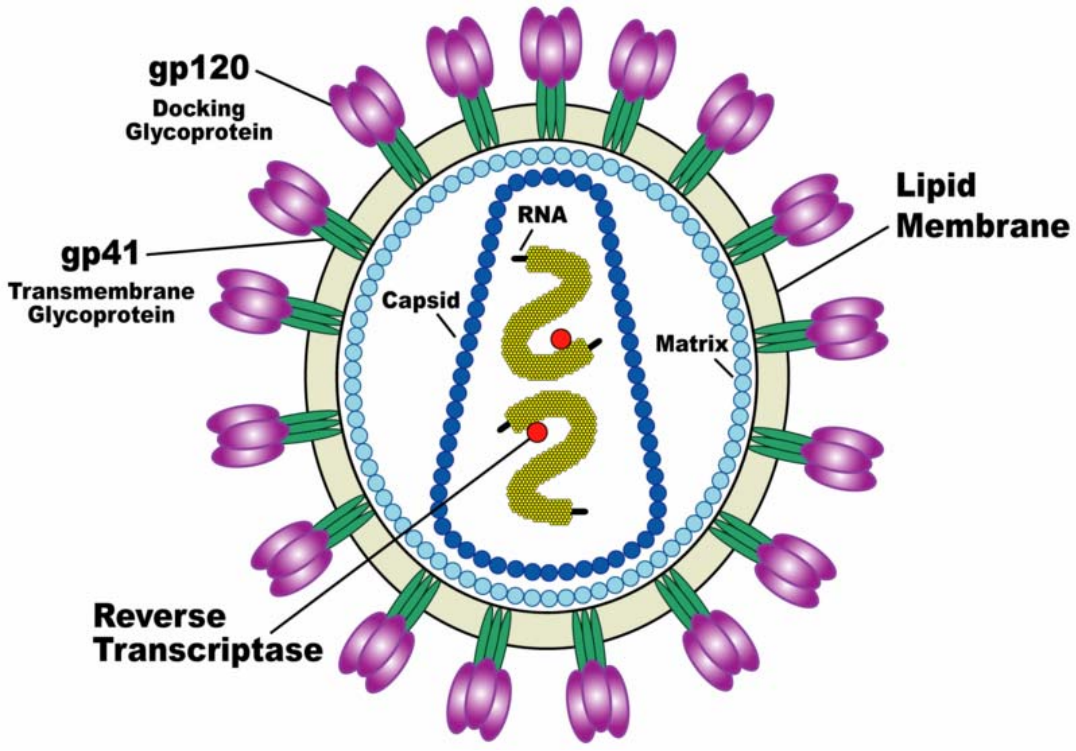


Figure 1.2: A schematic diagram of the HIV virion. The diagram displays all the major proteins and the ssRNA [http://en.wikipedia.org/wiki/HIV].

1.2.3 Genomic organization of HIV-1

The genome of the HIV-1 virus is approximately 9.7 kb long with long terminal repeats (LTR's) on both sides (Figure 1.3). Within the 9.7 kb fragment of the genome there are several open reading frames coding for a multitude of functional virus proteins namely; structural genes (*gag*, *pol*, and *env*),

regulatory genes (*tat*, and *rev*), and accessory genes (*vif*, *vpr, nef,* and *vpu*). The *gag* gene provides the structural elements of the virus. The p24 part of the gene makes up the viral capsid whereas the p6 and p7 parts provide the nucleocapsid and p17 provides a protective matrix. The *pol* gene is a common feature of all retroviruses. It encodes for the reverse transcriptase enzyme that is responsible for the transcription of viral RNA into double-stranded DNA [Levy, 2007].

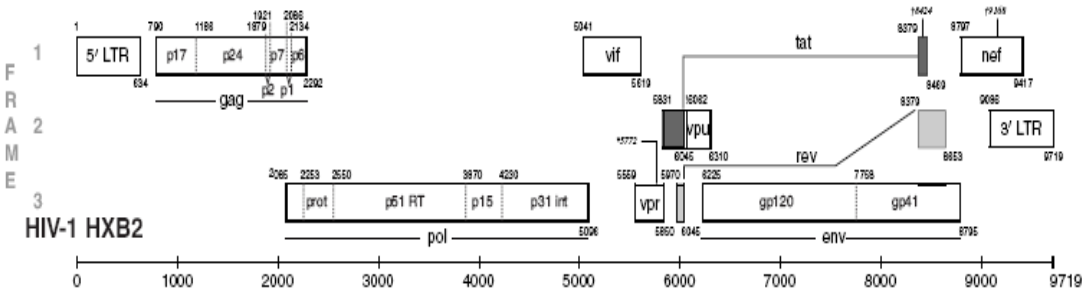


Figure 1.3: A diagrammatical representation of the genome layout of HIV-1. All three reading frames with all the most important genes are shown. All start and stop coordinates of genes on the diagram corresponds to that of the HXB2 reference strain. Exons 1 and 2 of tat and rev are indicated on the diagram with dark and light gray respectively [Adapted from http://www.hiv.lanl.gov/].

The *pol* gene also codes for the *integrase* and *protease* enzymes. *Integrase* is responsible for the integration of the double-stranded DNA into the host cells genome [Dicker *et al,* 2007]. The *gag* and *pol* genes do not produce their proteins in their final form, but as a large polypeptide. HIV *protease* then cleaves these large protein segments into separate functional units. The *env* gene encodes for a precursor protein of gp120 and gp41 called gp160, which is cleaved into the two functional proteins by the host cell's own enzymes. Env gp120 is exposed on the surface of the viral envelope and binds the virus to the CD4 receptors on the surface of any target cells. The glycoprotein gp41 is non-covalently bound to gp120, and facilitates the second step of viral entering into the target cells. The gp41 is originally found inside the viral envelope, but when gp120 binds to the CD4 receptor, gp120 undergoes a conformational change causing gp41 to become exposed on the viral envelope, where it can assist in the fusion of the virus with the host cell [Chan

et al, 1997]. A summary of all the HIV genes and proteins are listed in Table 1.1.

Table 1.1: The HIV-1 genes and proteins with their functions. [Adapted from Levy, 2007]

HIV proteins and their function			
Proteins	**Size (kDa)**	**Function**	**Abbreviation**
Gag	p24	Capsid (CA), structural protein	CA
	p17	Matrix (MA) protein, myristoylated	MA
	p7	Nucleocapsid (NC) protein, helps in reverse transcription	NC
	p6	Role in viral budding (L domain)	-
Polymerase	p66, p51	Reverse Transcription (RT): RNase H - inside core	RT
Protease	p10	Gag/Pol cleavage and maturation	PR
Integrase	p32	Viral cDNA integration	IN
Env	gp120	Envelope surface protein	SU
	gp41	Envelope transmembrane protein	TM
Tat	p14	Transactivation	Tat
Rev	p19	Regulation of viral mRNA expression	Rev
Nef	p27	Pleiotropic, can increase or decrease virus replication	Nef
Vif	p23	Promotes virion maturation and infectivity	Vif
Vpr	p15	Helps in virus replication, transactivation	Vpr
Vpu	p16	Helps in virus release, disrupts gp160:CD4 complexes	Vpu
Vpx	p15	Helps in entry and infectivity (Only in HIV-2 and SIV)	Vpx

Key: p (protein), gp (glycoprotein), - (no abbreviation), cDNA (complimentary DNA), RT (Reverse Transcriptase), mRNA (messenger RNA), and CD4 (cluster of differentiation 4)

Unlike certain oncogenic retroviruses, such as HTLV-I and -II in the retrovirus family, HIV-1 has no *onc* gene, but possess other unique genes such as; *rev* (facilitates the exportation of mRNA from the nucleus to the cytoplasm), *tat* (transactivation of HIV gene expression), *vif* (inhibits the cellular protein, APOBEC3G, from entering the virion at the time of budding), *vpr* (plays an important role in regulating nuclear importation of the HIV-1 pre-integrated complex and is required for virus replication in non-dividing cells), *vpu* (facilitates viral budding), and *nef* (ensuring T cell activation and the establishment of a persistent state of infection) [Levy, 2007].

1.2.4 The life cycle of HIV-1 (Replication)

The human immunodeficiency virus (HIV) can infect a variety of immune cells such as CD4+ T-cells, Cytotoxic T-lymphocytes (CTLs), CD4+ monocytes, macrophages and CD4+ dendritic cells of the host's immune system [Chan *et al,* 1998]. The virus life cycle can be divided into two distinctive stages: the early establishment of infection and the later viral replication stage (Figure 1.4).

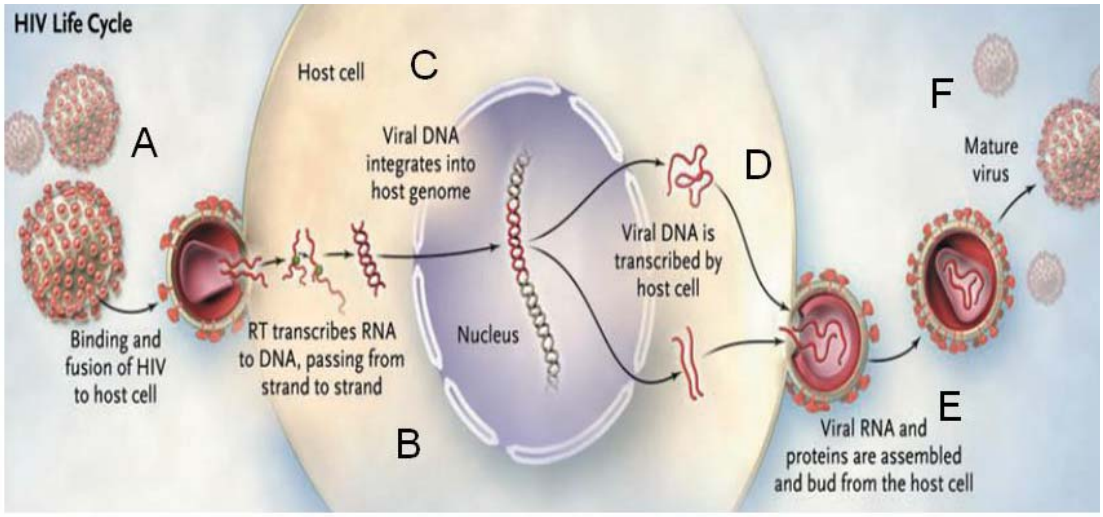


Figure 1.4: The HIV-1 life cycle. Stages A – C represents the viral infection and establishment of the virus in the host cell and stages D – F the viral replication [Adapted from Taylor *et al*, 2008].

Successful infection of a target cell is accomplished through a series of virus-host cell interactions. These include the binding of the virus to the cell surface, the fusion of the virus and the cell membrane, entry of the virus capsid into the cytoplasm of the host cell, the reverse transcription of RNA to DNA and the incorporation of the viral DNA into the nucleus of the host cell [Chan *et al,* 1998]. Briefly, entry to the cell begins through interaction of the trimeric envelope complex (gp160 spike) and both CD4 and a chemokine receptor (generally either CCR5 or CXCR4) on the cell surface [Levy, 2007]. The first step in fusion involves the high-affinity attachment of the CD4 binding domains of gp120 to CD4. Once gp120 is bound with the CD4 protein, the envelope complex undergoes a structural change, exposing the chemokine binding domains of gp120 and allowing them to interact with the target

chemokine receptor. This allows for a more stable two-pronged attachment, which allows the N-terminal fusion peptide gp41 to penetrate the cell membrane [Wyatt and Sodroski, 1998].

After successful fusion p24 is released into the cytoplasm, which uncoats to release the viral RNA into the host cells cytoplasm. The RNA is then reverse transcribed using the host cells reverse transcriptase enzyme to synthesis a double-stranded DNA copy of its RNA. With the help of the viral *integrase* the newly synthesized double-stranded copy of DNA is then incorporated into the host cells genome.

Once integration has occurred in a host cell the cells progeny will also be infected [Levy, 2007]. The provirus establishes latency in the infected cell. New virions are then synthesized from the integrated provirus. When a host cell is infected by two or more viruses there is a possibility that two different genomes may be transcribed into the same virion. The successful infection, integration and translation, of these virions into new host cells will then lead to the production of proviral genomes that are recombinants of the two viral genomes [Rodrigo and Learn, 2000]. Viral recombination in HIV-1 is made possible by the diploid (two RNA molecules) nature of the viral genome. Viral recombination of HIV-1 occurs during the reverse-transcription step, before viral integration and dependence on the co-packaging of two different viral genomes. This requires simultaneous infection of the same cell by two different strains and subsequent integration of two different parental-generation proviruses in the same nucleus (Figure1.5) [Taylor *et al,* 2008].

The later viral replication stage starts as soon as the viral DNA is transcribed into RNA by the host cells DNA polymerase. Spliced and unspliced viral mRNA is then exported to the cytoplasm of the host cell for translation and virus formation. Spliced viral mRNA is translated into viral proteins in the cytoplasm of the host cell. The virus capsid incorporates an unspliced copy of viral mRNA into a newly formed particle on the inner surface of the host membrane [Gelderblom, 1997]. New virions are produced as the virus buds through a region of the host's cell membrane. As the virion buds through the host cells membrane, outer surface proteins of the host cell become

associated with the virus. The host cell dies from the effect of continuous immune activation that occurs in HIV-1 infected patients [Goldberg and Stricker, 1999]. Today it is widely hypothesized that the apoptosis of immune cells in infected patients is the major cause of the severe depletion of CD4+ T-cells and the eventual paralysis of an individual's immune system.

The intense study of the viral life cycle of HIV has been a key to developing valuable antiretroviral drugs against HIV-1. Today a wide range of antiretroviral drugs can be used to target up to six different stages of the viral life cycle including viral entry, reverse transcription, integration, expression of viral proteins, viral assembly and viral release [van Rossum AMC *et al,* 2002].

1.3 Diversity of HIV-1

HIV-1 is characterized by a high degree of genetic variation driven by a wide range of factors, such as the lack of a proofreading ability by its relatively weak reverse transcriptase [Op de Coul *et al,* 1997; Roberts *et al,* 1988], the rapid turnover time of HIV-1 *in vivo* [Ho *et al,* 1995], host selective pressures [Michael, 1999], and recombination events in dually infected patients [Temin, 1993]. The rate of sequence variation across the genome of HIV varies, with the highest degree of sequence variation in the *env* gene, intermediate amounts in the *gag* and a low degree in the *pol* gene [Shankarappa *et al,* 1999]. Thus, no two strains are identical and even in infected individuals; HIV is present in a swarm of micro variants (quasispecies) that are highly related, but genetically distinct.

The family of human immunodeficiency viruses consists of groups of genetically distinct retroviruses. The most basic division of the viruses that cause AIDS in humans is the human immunodeficiency virus type 1 (HIV-1) and type 2 (HIV-2). HIV-1 can be divided into three subgroups; Group M, N and O and each of them have arisen due to a single zoonotic transmission from non-human primates. Group M in itself is made up of nine subtypes, several sub-subtypes and 43 circulating recombinant forms (CRF's) [De Leys *et al,* 1990; Gurtler *et al,* 1994]. New data also support the idea that a subtype pattern can be found within group O sequences [Lemey *et al,* 2004].

The subtype classification that is used for HIV is based on a phylogenetic system, which means that subtypes are grouped on their inferred evolutionary relationship, rather than on other characteristics such as serological reactivity, phenotype, co-receptor usage and many other possible biological characteristics. This sets HIV viral subtype classification apart from other older viral pathogens where serological subtyping is the norm. HIV was discovered during the times in which PCR and sequencing technologies were discovered, which lead to the use of such techniques for the subtype classification instead of the older method of serology which was widely used for other older viral pathogens [Rodrigo and Learn, 2000]. The HIV-1 classification system was first used in 1992. By the end of 1992, five subtypes were known for *env* (A through E) and four for *gag* (A through D) [Myers *et al,* 1992; Louwagie *et al,* 1993]. Since then the classification system has been continually updated as new viral isolates were sequenced and new data became available. In 1993, subtypes F, G and H were added. Since then there has been many additions to the classification system [http://www.hiv.lanl.gov./].

Due to all these changes in the field of subtype classification, new criteria for assigning a sequence to an existing subtype and for creating new subtypes or sub-subtypes were decided on in September of 1999. These criteria set out to give more order to the classification system and set out strict rules for the creation of new groups, subtypes and sub-subtypes [Robertson *et al,* 2000]. The classification scheme of HIV-1 into subtypes has proven useful in the phylogenetic analysis for clarifying epidemiological relationships and possible ancestry of HIV and the classification of new sequences. It has given us a clear picture of the worldwide spread of the epidemic and provided us with information on how HIV has entered different countries, where patterns differ depending on epidemiological factors. In some cases there have been isolates that did not fit the subtype classification system very well. Some were found to be part of newly discovered subtypes, such as the subtype F which was then sub-divided into sub-subtypes [Potts *et al,* 1993]. Others were later found to be recombinants that could not be assigned to a subtype unambiguously, but were more likely a mix of a wide variety of subtypes [Carr *et al,* 1996; Gao *et al,* 1996].

Advances in full-genome amplification and sequencing of HIV made the identification of circulating and unique recombinant forms (CRFs and URFs respectively) much easier. These isolates are the result of recombination between subtypes within a dually infected person (Figure 1.5), from which the recombinant forms are then passed on to other people. The progeny of such a recombinant virus are classified as a CRF if they have been identified in three or more epidemiologically unlinked people. If a particular recombinant form has only been identified in one or two cases and thus are of little epidemiological value then such viruses are classified as unique recombinant forms [Rodrigo and Learn, 2000].

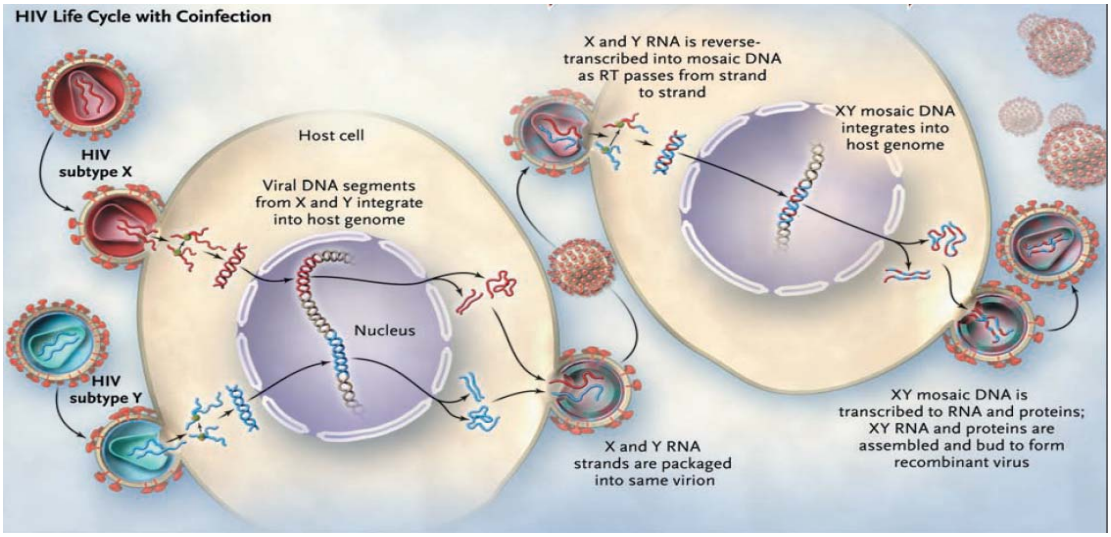


Figure 1.5: Creation of a URF in a coinfected individual. The individual is coinfected with two different subtypes X and Y, where the genome of the URF represents part of both parental strands [Taylor *et al,* 2008].

1.3.1 Global diversity of HIV-1

Globally group M is the predominant circulating HIV-1 group. Group M HIV-1 can be divided into 9 subtypes (A, B, C, D, F, G, H, J and K), sub-subtypes (A1-A3 and F1-F2), circulating recombinant forms (a total of 43 have been identified to date) and several unique recombinant forms [Gurtler *et al,* 1994; Janssens *et al,* 1994; Taylor *et al,* 2008]. In some cases, subtypes can be linked to a specific geographical region or epidemiological group [Hemelaar *et al,* 2006]. These distribution patterns are either the consequence of accidental trafficking (due to international travel) or due to a prevalent route of

transmission, which results in a strong advantage for a specific subtype to become dominant within a specific region or country [Myers, 1994]. Recent academic data indicates that the most prevalent HIV-1 genetic forms are subtypes A, B, and C, with subtype C accounting for as much as 50% of all infections worldwide [Ariën *et al*, 2007].

Molecular epidemiology studies have shown that, with the exception of sub-Saharan Africa where most subtypes and circulating recombinant forms can be found, there is a specific geographic distribution pattern for HIV-1 subtypes (Figure 1.6).

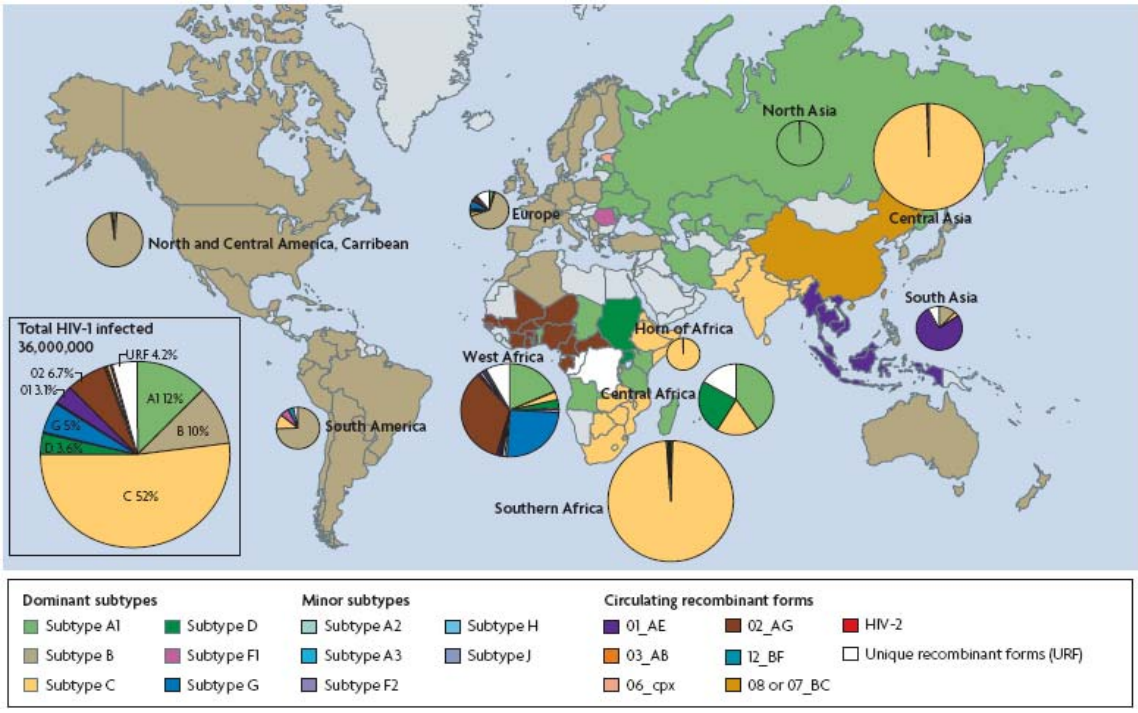


Figure 1.6: Global distribution of different HIV-1 subtypes and circulating recombinant forms. [Adapted from Ariën *et al*, 2007]

Subtype A viruses are most prevalent in areas of central and eastern Africa (Kenya & Tanzania) and in the East European countries formerly part of the Soviet Union [Buonaguro *et al,* 2007]. These areas where subtype A viruses co-circulate with other viral subtypes have also seen the rise of recombinant viruses. Some areas have seen the rise of A recombinant viral forms due to the co-circulation of other viral subtypes in the same geographical area e.g.

the prevalence of subtypes A and B which lead to the rise of AB recombinant viruses in Eastern Europe [Liitsola *et al,* 1998] and the co-circulating of subtype A and D viruses in East African countries such as Kenya which lead to the rise of AD recombinants [Songok *et al,* 2004; Dowling *et al,* 2002].

Subtype B is the major genetic clade in the rest of Europe, the Americas, Japan and Australia, but is also found in large numbers in the countries of Southeast Asia, Northern Africa and the Middle East and amongst homosexual men from South Africa and Russia. Subtype C HIV-1 is predominant in the countries of Southern Africa and the Indian subcontinent [Buonaguro *et al,* 2007].

Of the 43 CRF's known to us, some 20 reports have been made of CRF's been isolated in areas where the parental strains of the recombinant viral forms are cocirculating. The existence of multiple subtypes and CRF's within the same region increases the probability that individuals will become infected with different HIV-1 genetic forms, which can exchange parts of their genetic material, and result in the formation of recombinant viruses [McCutchan, 2006]. The role that circulating recombinant forms or CRF's are playing in the global HIV-1 pandemic is increasingly being recognized, with CRF's accounting for more than 18% of global infections [Osmanov *et al,* 2002; Hemelaar *et al,* 2006; Peeters, 2000]. In some areas of the world, CRF's represent the dominant form of HIV-1 such as in Southeast Asia (CRF01_AE) [Menu *et al,* 1996] and in West and West Central Africa (CRF02_AG) [McCutchan *et al,* 1999; Njai *et al,* 2006].

CRF01_AE (Figure 1.7) plays a major role in the HIV-1 epidemic in Southeast Asia. This subtype is responsible for more than 75% of the infections within the region and accounts for an estimated 4.7% worldwide infections [Hemelaar *et al,* 2006]. This subtype was first identified in Thailand in the late 1980's [McCutchan *et al,* 1992; Carr *et al,* 1996]. It was first classified as a new clade called subtype "E" but after full-length sequence analysis of these isolates it became apparent that the virus appeared to have a mosaic-like structure, with the *gag* gene clustering with other subtype A isolates and the

env genes from clade E [Leitner *et al,* 2005; Carr *et al,* 1996; Gao *et al,* 1996]. To date the parental clade E strain has not been found. Extensive studies of CRF01_AE have shown that the isolate was introduced into Thailand from Africa and from there the isolate has spread to other counties in Southeast Asia where it has become the most prevalent form of HIV-I where it is linked to heterosexual sex workers and intravenous drug users within the region [Anderson *et al,* 2000].

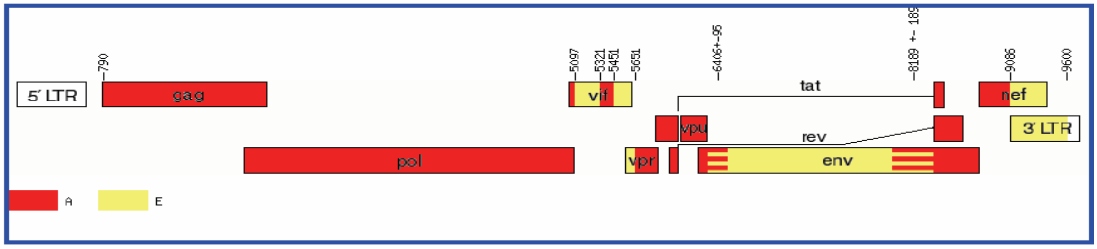


Figure 1.7: The genomic structure of a CRF01_AE isolate. The yellow areas correspond with HIV-I subtype E and the red areas with HIV-1 subtype A. The breakpoints of the recombination events are also shown (breakpoint coordinates relative to the reference strain HXB2) [Adapted from LANL database, http://www.hiv.lanl.gov/].

1.3.2 HIV-1 diversity in Africa

The African continent is home to the largest portion of all HIV infections in the world with as many as 23.6 million people living with HIV/AIDS [UNAIDS, 2008]. A multitude of subtypes and circulating recombinant forms can be found in Africa which is possibly due to the African origin of HIV-1 [Torques *et al,* 1999; Vidal *et al,* 2000]. The distribution and occurrence of different subtypes within African populations are not generally linked to a particular lifestyle habit as it would be in other parts of the world. In Central and Eastern Africa subtype A1 and D HIV-1 is the most prevalent form of HIV. Subtype C is dominating the epidemic in Southern Africa where it can account for as many as 95 percent of all HIV infections (Figure 1.8). Due to the importance of subtype C HIV within the Southern African region this subtype has been extensively studied in the past [Bell *et al,* 2007; Engelbrecht *et al,* 2001; Scriba *et al,* 2002; Bessong *et al,* 2005; Hunt *et al,* 2003; Papathanasopoulos *et al,* 2003b; zur Megede *et al,* 2002].

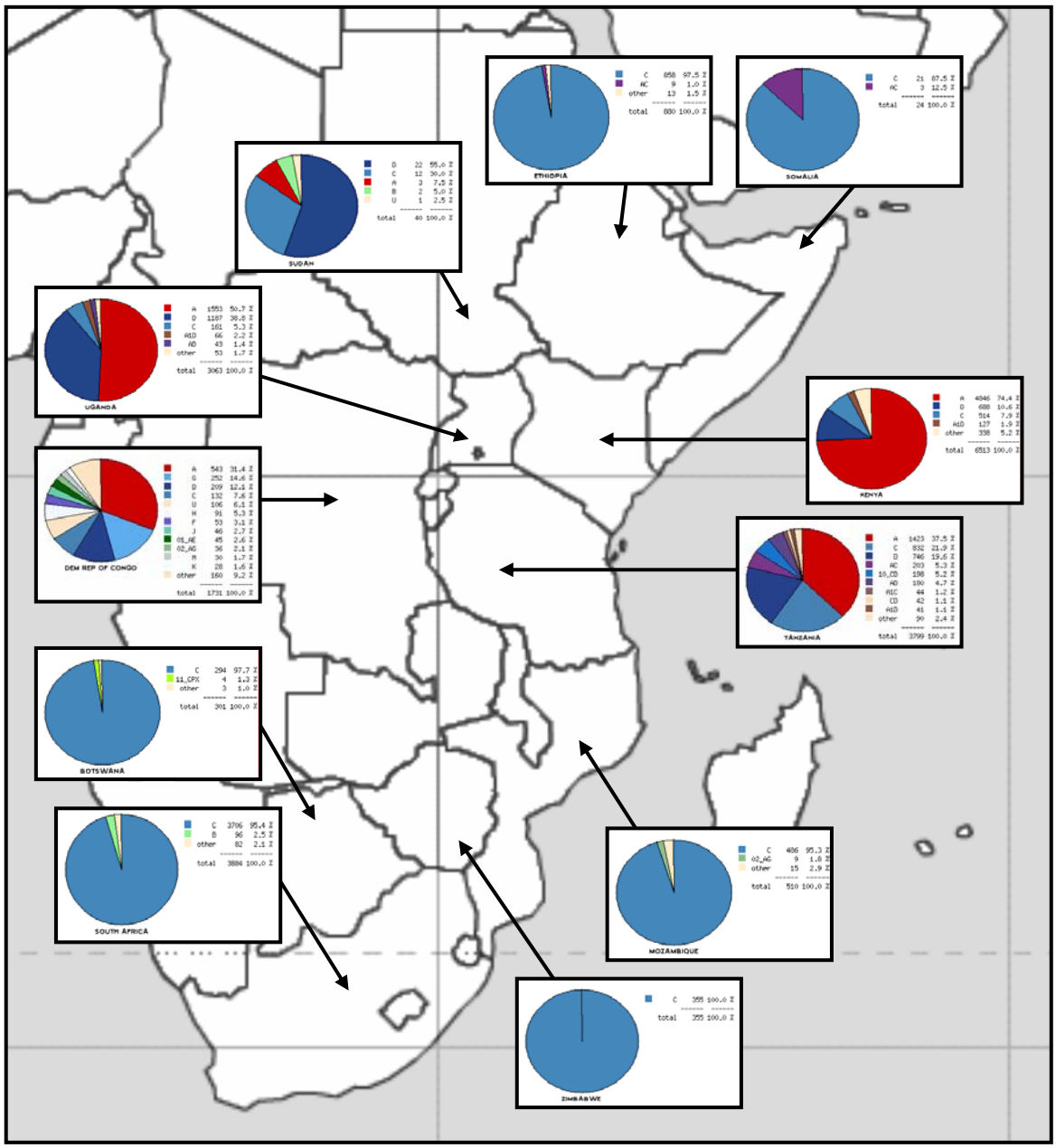


Figure 1.8: Subtype distribution of HIV-1 in a handful of sub-Sahara African countries [Adapted from: http:www.hiv.lanl.gov./].

CRF02_AG (Figure 1.9) which is the dominant form of HIV in Western and Central Africa also contributes considerably to the global epidemiology of the virus accounting for an estimated 4.6% of worldwide infections [Hemelaar *et al,* 2006].

Initially the CRF02_AG viral subtype was described as a divergent lineage within HIV-1 subtype A, based on partial *gag* and *env* sequences [Howard and

Rasheed, 1996]. After full length sequences of these isolates were obtained, it was recognized as a complex mosaic virus of alternating subtype A and subtype G regions [Carr *et al*, 1998]. This viral form of HIV-1 have been isolated in various regions ranging from West Africa to East African countries and a CRF02_AG isolate have also been identified in South Africa. [Bredell *et al,* 2002, Jacobs *et al,* submitted] In countries of West and Central Africa, such as Nigeria and Cameroon, this recombinant form is responsible for between 50 – 70% of new infections [Andersson *et al,* 1999; McCutchan *et al,* 1999; Fischetti *et al,* 2004].

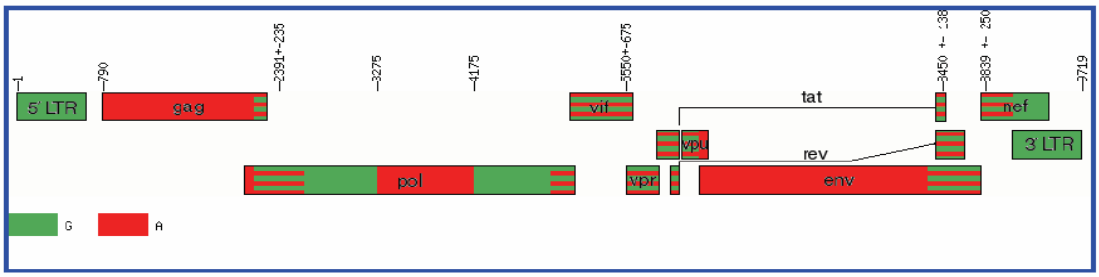


Figure 1.9: The genomic structure of CRF02_AG. The green areas correspond with HIV-I subtype G and the red areas with HIV-1 subtype A. The breakpoints of the recombination events are also shown (breakpoint coordinates relative to the reference strain HXB2) [Adapted from LANL database, http://www.hiv.lanl.gov/].

1.3.3 HIV-1 diversity in South Africa

South Africa is currently in the grip of one of the most devastating AIDS epidemics in the world. By the end of 2007, 6.2 million people were living with HIV in South Africa, with countless more being affected by the lost of family and loved ones. UNAIDS estimated that there were 1.4 million South African children orphaned by AIDS in 2007 [UNAIDS, 2008]. The UNAIDS report of 2008 estimate that almost 1,000 AIDS related deaths occurring every day within the country [UNAIDS, 2008].

In 1982 the first reported case of AIDS in South Africa was documented, with the first virus isolation done in 1984 [Ras *et al,* 1983; Becker *et al,* 1985]. Initially HIV infections seemed to be occurring mainly amongst homosexual white men [Sher, 1989]. By the start of the 1990's it became apparent that a

second HIV pandemic, amongst the heterosexual indigenous black population, was occurring within the country [Williamson *et al,* 1995]. By 1990 an estimated 74,000 people were living with HIV in South Africa [Department of Health, 2005]. The following year the number of diagnosed heterosexually-transmitted HIV infections equaled the number of infections transmitted through men having sex with other men. Since then heterosexual transmission has dominated the epidemic in the country. The most rapid increase in South Africa's HIV prevalence took place between 1993 and 2000, during which time the country was distracted by major political changes. While the attention of the South African people and the world's media was focused on the political and social changes occurring in the country, HIV was rapidly becoming more widespread. Subtype C HIV-1 is responsible for nearly 95 percent of all infections within the country.

Over the past years subtype C HIV-1 has been extensively studied in South Africa due to its immense importance [Jacobs *et al,* 2006; Hunt *et al,* 2003; Papathanasopoulos *et al*, 2002; Rousseau *et al,* 2006; de Oliveira *et al,* 2003; Gordon *et al,* 2003; zur Megede *et al,* 2002]. To date only a few papers have been published on complete genomes of non-subtype C HIV-1 isolates in South Africa [Papathanasopoulos *et al*, 2001; Loxton *et al*, 2005]. A lot of research has been conducted on HIV-1 over the years within the country. To date several papers on a wide range of HIV-1 subtypes have been published on subgenomic fragments (Table 1.2) or near full-length sequences (Table 1.3). These viruses were isolated from several geographical locations within the country. The majority of HIV research has been conducted within the three biggest urban areas, Johannesburg, Cape Town, and Durban, but smaller studies has also been conducted in other areas of the country, such as Limpopo and Mpumalanga.

Table 1.2: Summary of published data of subgenomic fragments of HIV-1 isolates from South Africa.

Characterization of subgenomic regions				
Publication	**Number of samples**	**Method of Subtyping**	**Region of Genome**	**Subtypes**
Engelbrecht *et al,* 1994	17 Samples	Serology, Sequencing & Phylogenetics	env gp41	Subtype B
				Subtype D
				Subtype C
Becker *et al,* 1995		Sequencing and Phylogenetics	*gag & env*	Subtype B
				Subtype D
				Subtype C
van Harmelen *et al,* 1997	61 Samples	HMA & partial sequencing	V3 - V5 or *gag*	Subtype B
				Subtype C
				Subtype D (n = 1)
				CRF01_AE (n = 1)
Bredell *et al,* 1998	44 Samples	HMA & sequencing	V3 - V5 *env* region	All subtype C
Engelbrecht *et al,* 1999	81 Samples	Serology (cPEIA) and Phylogenetics	V3 region of env	Subtype C
				Subtype B
van Harmelen *et al,* 1999	87 Samples	RFLP	*gag*	Subtype A (n = 2)
				Subtype B (n = 28)
				Subtype C (n = 56)
				Subtype D (n = 1)
Hunt *et al,* 2001	60 Samples	HMA & Phylogenetics	*gag* & *env*	Subtype C (n = 43)
				Subtype A (n = 2)
				Subtype B (n = 3)
				Several Recombinants
Engelbrecht *et al,* 2001	13 Samples	Phylogenetics	*gag* & *env*	All subtype C
Bredell *et al,* 2002	10 Samples	HMA & partial sequencing	*gag* & *env*	Subtype A (n = 2)
				C CA, CD, G, AG and D

Characterization of subgenomic regions - Continued				
Publication	**Number of samples**	**Method of Subtyping**	**Region of Genome**	**Subtypes**
Scriba *et al,* 2002	14 Samples	Phylogenetics	5' LTR, *nef, tat* and *rev*	All subtype C
Gordon *et al,* 2003	72 Samples	Phylogenetics	*pol* and *env* C2V5	Subtype C (n = 71)
				CD Recombinants (n = 1)
Bessong *et al,* 2005	42 Samples	HMA & sequencing	*gag* & *env*	All Subtype C
Bell *et al,* 2007	20 Samples	Phylogenetics	*vif, vpr* & *vpu*	All subtype C
Jacobs *et al,* 2008a	50 Samples	Phylogenetics	*vif*	Subtype C (n = 48)
				Subtype B (n = 2)
Jacobs *et al,* 2008b	140 Samples	Phylogenetics	*pol* sequences	Subtype C (n = 133)
				Subtype B (n = 5)
				CRF02_AG (n = 1)
				Subtype G (n = 1)
Jacobs *et al,* submitted	410 Samples	Serology (cPEIA) and Phylogenetics	V3 region of *env*	Subtype C (n = 341)
				Subtype B (n = 36)
				Subtype A (n = 7)
				Subtype D (n = 3)
				Several Recombinants

Key: n (number)

Table 1.3: Published data of near-full length HIV-1 sequences from South Africa.

Full length-genome characterization			
Publication	**Number of samples**	**Method of Subtyping**	**Subtypes**
van Harmelen *et al,* 2001	4 Samples	Sequencing and Phylogenetics	Subtype C (n = 4)
Papathanasopoulos *et al,* 2002	4 Samples	Sequencing and Phylogenetics	Subtype C (n = 2)
			A2C Recombinant (n = 1)
			ACDGK Recombinant (n = 1)
zur Megede *et al,* 2002	3 Samples	Sequencing and Phylogenetics	All subtype C
Papathanasopoulos *et al,* 2003b	2 Samples	Sequencing and Phylogenetics	All subtype C
Hunt *et al,* 2003	3 Samples	Sequencing and Phylogenetics	All subtype C
Loxton *et al,* 2005	4 Samples	Sequencing and Phylogenetics	All subtype D
Rousseau *et al,* 2006	244 Samples	Sequencing and Phylogenetics	Subtype C (n = 241)
			Subtype B (n = 1)
			CA Recombinants (n = 2)
Jacobs *et al,* 2007	1 Sample	Sequencing and Phylogenetics	Subtype D (n = 1)

Key: n (number)

1.4 Diversity within the HIV – 2 genome

Since the discovery of HIV-2 in Western Africa in the mid-1980's a great deal has been learned about the epidemiology, spread and pathogenesis of the virus [Markovitz, 1993]. Unlike HIV-1, HIV-2 has not spread all over the world, but is mostly confined to areas of West Africa. Analyses of HIV-2 have revealed the existence of five major lineages within the cluster, which are termed HIV-2 groups A – E [Gao *et al,* 1994; Hasegawa *et al,* 1989].

1.5 Methods used in the phylogenetic analysis of HIV

Phylogenetics is the science of estimating the evolutionary relationship, which is based on the comparison of DNA or protein sequences with the phylogeny ultimately depicted in the form of an evolutionary tree [Graur and Li, 1999; Salemi and Vandamme, 2003]. The use of trees to depict such hypotheses goes back to the start of the field of evolution in Charles Darwin's days. It is only until recently that methods have been developed to numerically calculate trees through quantitative methods. In the modern age of rapid gene sequencing, digitization of sequences and creation of sequence databases, molecular phylogeny has become a powerful tool for making sense of these vast amounts of data.

Before one can start with phylogenetic analysis of data, one needs to generate DNA sequence data. These procedures include the designing and testing of primers, the amplification of target genes or genomes, and the sequencing of amplified products [McCormack and Clewley, 2002]. These processes include a wide range of procedures and technical expertise on its own. The following section will take a look at some of the methods commonly used in modern phylogenetic analysis.

1.5.1 Searching for homologous sequences

The ultimate goal of phylogenetic analysis is to compare the evolutionary relationship between new sequences and other known sequences. One of the first things to do before starting analysis of data is to obtain homologous

sequences. Today most genetic data, either in nucleic acid or amino acid formats, is stored in online accessible sequence databases with the three largest of these being; EMBL (European Molecular Biology Laboratory), GenBank (at NCBI), and the DDBJ (DNA Data Bank of Japan) [Salemi and Vandamme, 2003]. The BLAST (Basic Local Alignment Search Tool) is the most widely used method in modern molecular evolutionary biology to search for homologous sequences and can be performed on most of the genetic databases [Altschul *et al,* 1990]. Some databases were developed for specific uses such as the LANL (Los Alamos National Laboratory, Los Alamos, New Mexico, USA) which only contains HIV and SIV related sequences [http://www.hiv.lanl.gov/].

The HIV sequence database at Los Alamos was created in 1986 under the guidance of the AIDS Program of the US National Institute of Allergy and Infectious Diseases (NIAID) in association with the US Department of Energy (DOE) [Rodrigo and Learn, 2000]. It is a specialized molecular-sequence database that provides a wide range of services, free of charge to the global HIV research community. The HIV database, unlike other more general genetic databases such as Genbank, contains only sequences that relate to primate lentiviruses and a few primate genes and proteins that interact with HIV. Any data in the database is shared with the general databases, so that all the data also appears in the GenBank, EMBL, and DDBJ databases [Learn *et al,* 1996]. The HIV database also contains useful tools for working with sequences. One can search the database of sequences in a number ways: by genome region, country of origin, subtype, viral phenotype, year of sampling, similarity to a given sequence (BLAST) and health of patients (with respect to ARV therapies, CD4 counts and several other factors) [Holmes, 2000]. The database also contains several pre-built sequence alignments of nucleotides or amino acids of partial fragments or complete genomes. Other tools include the jpHMM (for subtype and recombinant identification) [Zhang *et al,* 2006; Schultz *et al,* 2006], sequence locator (useful for assigning a DNA fragment to a particular region of the HIV genome relative to HXB2) [http://www.hiv.lanl.gov/], RIP or Recombinant Identification Program (for

recombinant virus identification) [Siepel *et al,* 1995] and gene cutter [http://www.hiv.lanl.gov/].

1.5.2 Aligning homologous sequences

Before phylogenetic trees can be draw one must first construct a multiple alignment containing the sequences of interest with homologous sequences that were obtained from a specific sequence database [Hogeweg and Hesper, 1984]. Accurate alignment of sequences is extremely important in the analysis of datasets. Only homologous sequences can be aligned with one another. Homologous sequences are any two sequences that share a common recent ancestor. One should distinguish between sequence similarity and sequence homology. Any sequences have some measurable similarity, but homology implies that this similarity is the result of a share common recent ancestor [Abecasis *et al,* 2007].

The most common method of constructing multiple alignments is by progressive alignment such as Clustal W or Clustal X [Salemi and Vandamme, 2003]. When a multiple alignment can be used to construct a phylogenetic tree then the converse is also true. With progressive alignment methods sequences are aligned in pairs to create a distance matrix based on their alignment scores. These scores are then downweighted according to how closely related the sequences are. This distance matrix is used to construct a Neighbor Joining (NJ) guide tree. The guide tree is used to cluster the sequences during the stepwise alignment, with the isolates that were clustered closest together being aligned with one another first. For further alignment these two sequences are treated as one with other sequences being aligned to them one by one. Clustal W [Thompson *et al,* 1994] and Clustal X [Thompson *et al,* 1997] are the most widely used programs for carrying out multiple alignments. These two programs are identical in terms of alignment methods, but Clustal W offers a simple text-based interface whereas Clustal X has a more graphical interface which might be more user friendly [Salemi and Vandamme, 2003]. Before proceeding with further analysis one should always manually check the alignment with a sequence

editing program such as BioEdit [Hall, 2001]. Often one can easily improve the alignment and in some cases it might be required to delete blocks containing gaps [Abecasis *et al,* 2007].

1.5.3 Choosing a model of evolution

Genetic sequences are not very informative regarding their evolutionary history. When we compare homologous sites in sequences, we only observe that the sequences are the same or different [Page and Holmes, 1998]. Evolution is caused by mutations which spread through populations by genetic drift or natural selection. These mutations can be caused by a number of events such as nucleotide substitutions, insertions, deletions or recombination events [Graur and Li, 1999]. Phylogenetic analysis makes certain assumptions about the process and rate of DNA substitutions or amino acid replacements in the model of evolution they employ [Salemi and Vandamme, 2003]. Point mutations can either by be due to transitions, when a purine nucleic base (A, G) replaces another purine base or a pyrimidine base (C, T) replaces another pyrimidine base, or transversions, when a purine is replaced by a pyrimidine base or *vice versa* [Graur and Li, 1999].

To study the dynamics of these changes in the sequences, one needs to use mathematical models that take into account different rates of nucleotide substitution. To date, a vast array of these models has been developed over the years by scientists [Li, 1997; Graur and Li, 1999; Salemi and Vandamme, 2003]. The first of these models, the Jukes and Cantor method, was developed as far back as the late 1960's [Jukes and Cantor, 1969]. The three most commonly used methods for the analysis of HIV datasets are the Kimura two-parameter model [Kimura, 1980]; the Hasegawa, Kishino, Yano (HKY) method [Hasegawa *et al,* 1985] and the General time-reversible (GTR) model [Rodriguez *et al,* 1990; Yang *et al,* 1994].

In most cases, as for HIV, the number of transitions is often higher than the rate of transversions. In 1980, Kimura developed an algorithm for estimating the number of nucleotide substitutions per site, which took into account the higher probability of transitional change [Kimura, 1980]. The model assumes a

total rate of nucleotide substitution of: $\alpha + 2\beta$, where rate of transitions per site is α and the rate of transversions is β. At any particular site the nucleotide base can undergo three possible changes, one being a transition and the other two being transversions [Li, 1997].

The Hasegawa, Kishino, Yano (HKY) model was first described in 1985 by Hasegawa and co-workers [Hasegawa *et al,* 1985]. As in the case of the Kimura 2-parameter model, the HKY-model also allows for a transition/transversion bias, but unlike the Kimura 2-parameter model that estimates equal base frequencies, the HKY-model allows base frequencies to vary. Theoretically in a given sequence each nucleotide base (Adenine, Cytosine, Guanine, or Thymine) has an equal probability (0.25) of appearing, however it often does not hold true. Some organisms have a higher composition of guanine and cytosine which makes their DNA much more thermodynamically stable due to the higher concentration of triple hydrogen bonds in the DNA molecules. Thus it is clear when working with sequences that the use of a model which allows for base frequencies to vary would be much more useful and accurate than other models that do not allow for this [Page and Holmes, 1998].

The GTR or general time-reversible model is the most general, unbiased, independent, finite-sites, time-reversible model possible, and was first described in 1986 by Simon Tavaré [Tavaré, 1986]. The probability matrix of the GTR-model has six parameters so that each possible substitution has its own probability [Page and Holmes, 1998]. Thus this model allows not only for different base frequencies, but also for different rates for all six substitutions. For a time-reversible model, there is no assumption that substitutions preferentially change in a certain direction over time.

1.5.4 Drawing phylogenetic trees

A phylogenetic tree, much like a real tree, is made up of branches and nodes. The branches are connected to the nodes and the nodes are the point of branch divergence. Nodes can also be internal or external. External nodes represent the sequences from which the tree was constructed or operational

taxonomic units (OUT's) whereas the internal nodes represent a common ancestor between two or more taxa (Figure 1.10).

Phylogenetic trees are usually drawn so that the branch lengths correspond to the amount of evolution between the two nodes they connect and such trees are termed additive trees. That means the longer the branches, the more divergent the sequences are. At the base of a phylogenetic tree is normally a root, which represents the oldest common ancestor of all the sequences in the tree [Hall, 2004]. Trees are rooted by using outgroups or an external point of reference. An outgroup may be anything that is not a natural member of the sequences or group of interest [Baldauf 2003; Salemi and Vandamme 2003].

Two main methods of calculating phylogenetic trees exist today: the distance-matrix method, also known as clustering or the algorithmic method (e.g. UPGMA, neighbor-joining, or Fitch - Margoliash) and the discrete data method which is also known as the tree searching method (e.g. parsimony, maximum likelihood, or the Bayesian method) [Baldauf 2003].

The distance method is extremely easy and fast to use, but does not involve an evolutionary model. The distance (or percentage sequence difference calculated by pairing up two sequences in a matrix), is calculated for all pairwise combinations of all OTU's and then the distances are assembled into a tree [Baldauf 2003]. Thus sequences with the closest distances are grouped close together on the representative tree. The UPGMA or Unweighted Pair Group Method with Armetric Means, searches for the smallest value in the pairwise distance matrix to construct a phylogenetic tree [Sneath and Sokal, 1973] The neighbor-joining method sequentially finds its closest neighbors based on the internal branch lengths of the tree [Saitou and Nei, 1987], and the Fitch-Margoliash method evaluates all possible trees to find the one with the shortest overall branch lengths [Fitch and Margoliash, 1967].

The discrete data method, such as the maximum parsimony, maximum likelihood and Bayesian methods, examines each column of the multiple alignments separately and then searches for the tree that best accommodates all of this information. Discrete data analyses are rich in information because it

creates a hypothesis for every column in the alignment and one can thus trace the evolution at a specific site in a DNA molecule (e.g. regulatory regions) [Baldauf 2003].

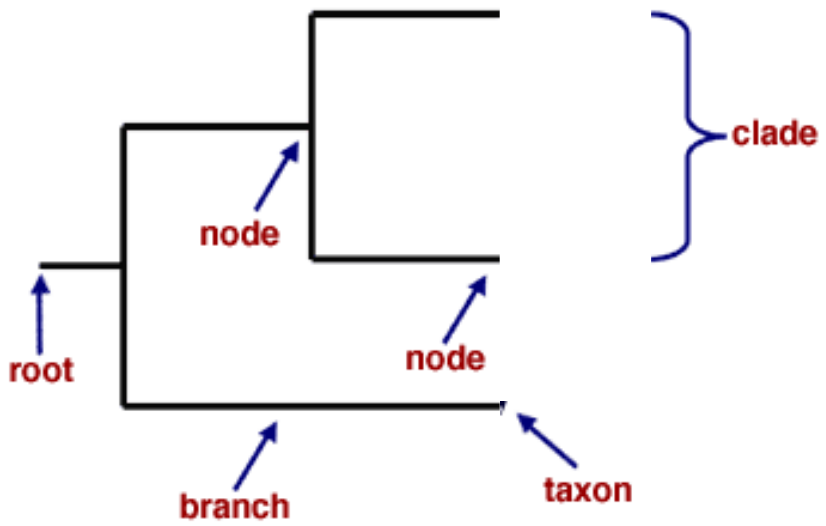


Figure 1.10: A simplistic representation of a phylogenetic tree. The tree indicates the root, branches, and nodes of the tree. A clade of are shown as well as a taxon (OTU) [Adapted from: http://www.talkorigins.org/].

The maximum parsimony and maximum likelihood methods use the theory of stepwise addition and branch swapping to search for the most representative phylogenetic tree. Through stepwise addition, branches are added in succession at different levels on the tree. Each level is then evaluated and the best tree is chosen before the addition of the next branch. Branch swapping techniques allow for the pre-defined rearrangement of the branches. The most common branch swapping techniques are tree bisection and reconstruction; nearest-neighbour interchange; and subtree pruning and regrafting [Salemi and Vandamme, 2003]. Maximum parsimony and maximum likelihood can both be performed with the PHYLIP software package [Felsenstein, 1982]. Baysian analysis is very much like that of maximum likelihood [Mau *et al*, 1999; Rannala and Yang, 1996]. Instead of seeking the tree that maximizes the likelihood of observing data, Bayesian analysis seeks those trees with the greatest likelihood of the given data. Baysian models search for the best tree

which is consistent with both the evolutionary model of choice and the data in an alignment. Baysian analysis of datasets is the most commonly used method today, with new software, such as BEAST (Bayesian Evolutionary Analysis Sampling Trees) [Drummond and Rambaut, 2007].

1.5.5 Detection of recombinant viruses

Recombination within viral genomes, and especially within HIV's genome, is extremely common. Recombination within the HIV genome occurs when an individual is co-infected with multiple strains of the virus. To date, several methods have been developed for the identification of recombinant viruses. In some cases one can characterize several small subgenomic regions throughout the viral genome [Swanson *et al,* 2003]. This method, though widely used does have some drawbacks. One can miss small regions of recombination in-between the regions that were characterized. Recent advances in the field of viral DNA or RNA amplification (Figure 1.11), which makes the amplification of large fragments (6 – 36 kbp) possible, has made the identification of viral recombinants much easier [Salminen *et al,* 1995a; Nadai *et al,* 2008].

The importance of full genome characterization of samples has been described on several occasions in the past [Choi *et al,* 1997; Carr *et al,* 1996; Gao *et al,* 1996]. To successfully identify viral recombinants there must be enough genetic variation between the different lineages of the particular virus in order to confirm that genetic exchange has occurred. The basic strategy of recombination identification is to construct a multiple alignment containing the query sequence and several different isolates from the different lineages. When these alignments of full or near full-length genome sequences are analyzed with the use of software packages such as; Simplot (Similarity Plot) [Lole *et al,* 1999; Salminen *et al,* 1995b], RIP (Recombination Identification Program) [Siepel *et al,* 1995], or RDP3 (Recombination Detection Program) [Martin *et al,* 2004], one can see the full extent of viral recombination within a single isolate.

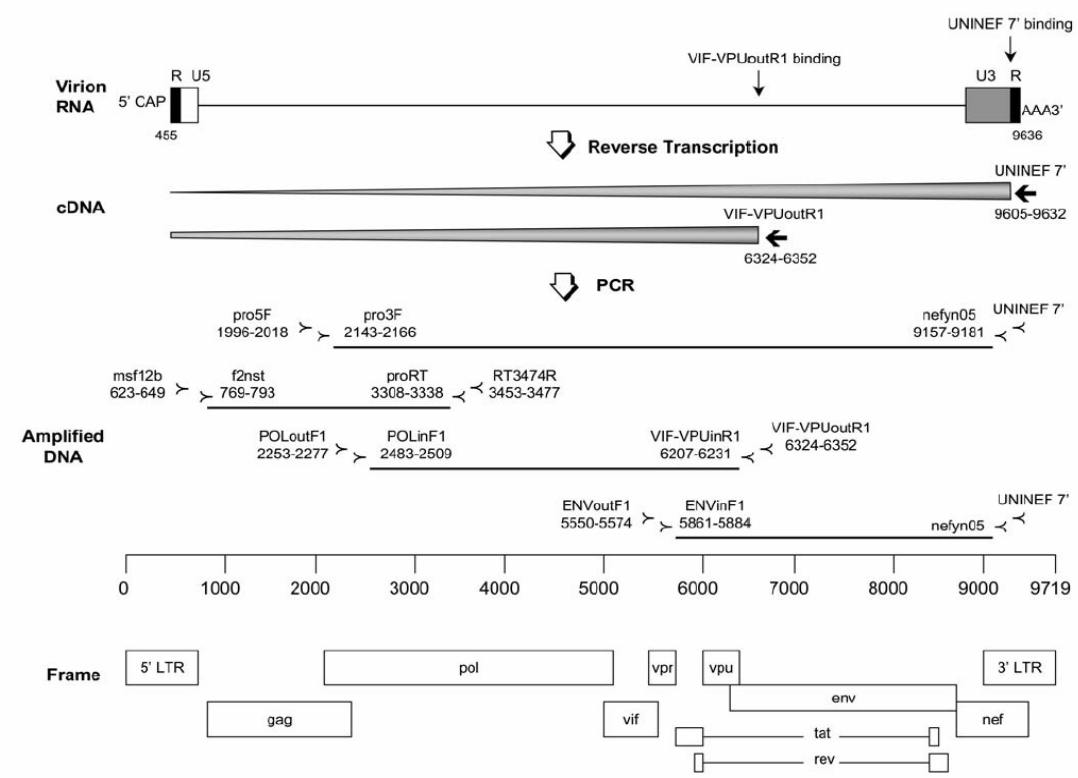


Figure 1.11: A graphical representation of a full-length amplification assay, for the characterization of a near full-length HIV-1 genome. Viral RNA is reverse transcribed into cDNA. PCR's are then performed in four overlapping fragments to amplify a near full-length genome of the virus. The methodology that was used here was developed as a standard protocol for the characterization of near full-length HIV-1 genomes from RNA [Nadai *et al,* 2008].

Simplot uses a sliding window approach moving across a multiple alignment in small increment steps to generate a similarity plot [Lole *et al,* 1999; Salminen *et al,* 1995b]. The program allows the user to query any sequence within the alignment and adjust the window and step sizes. The program is based on the Kimura 2-parameter substitution model. The recombination identification program (RIP) is one of the many online tools which are accessible from the LANL database [http://www.hiv.lanl.gov/; Siepel *et al,* 1995]. The program gives the user the opportunity to compare the query sequence against their own alignment or the pre-made alignment of the program itself. The pre-made alignment consists of several isolates from the most prominent viral subtypes (A1, B, C, D, F1, F2, G, H and CRF01_AE). Other subtypes such as A2, J and K are only represented by a single sequence in the alignment. The large size of the pre-made alignment makes

the identification of recombination events much easier. The program also allows the user to define the appropriate window size. It further allows for the simplifying of the results, with the use of the rerun application, and exports the graphs in a convenient manner [http://www.hiv.lanl.gov/].

AIM OF THE STUDY

The large variety of subtypes, sub-subtypes and circulating recombinant forms of HIV-1 complicates the development of effective vaccination strategies and has major implications for diagnostic assays and the effective treatment of infected people with anti-retroviral drugs. It is thus of the utmost importance that we continue monitoring the epidemiology of the HIV-1 epidemic. To date very few papers have been published on near full-length non-subtype C HIV type 1 viruses in South Africa [Papathanasopoulos *et al,* 2001; Loxton *et al*, 2005]. This study was aimed at characterizing non-subtype C HIV type 1 viruses from Cape Town, South Africa. This will greatly enrich our knowledge of other viral subtypes that are becoming more and more important in the South African setting.

CHAPTER TWO – MATERIALS AND METHODS

Table of Contents **Page**

CHAPTER TWO – MATERIALS AND METHODS

The experimental procedures used in this study are illustrated in this chapter. The materials and sampling methodology used will be described, followed by the experimental procedures used for the characterization of subgenomic and near full-length sequences (Figure 2.1).

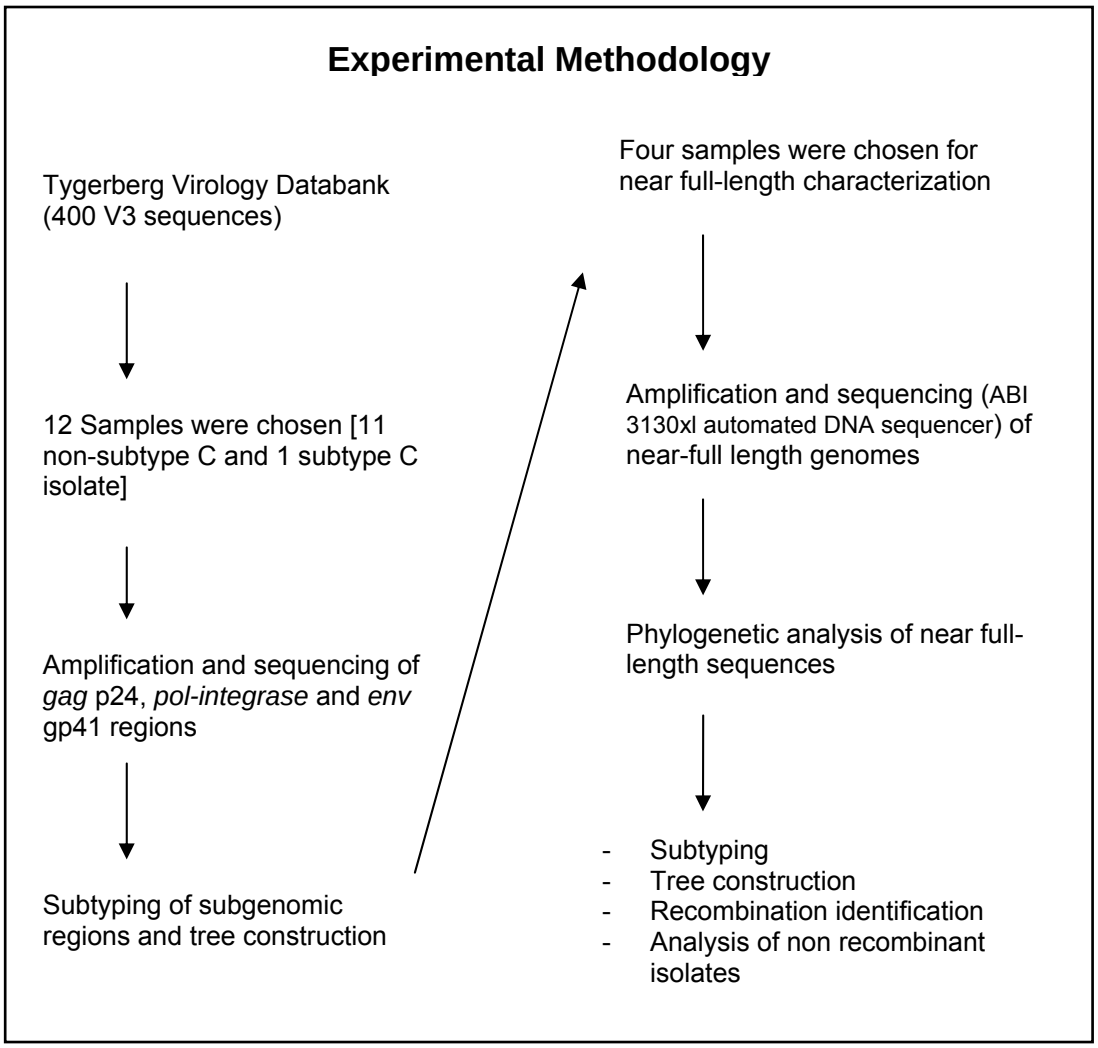


Figure 2.1: A flow diagram summarizing the methodology used in the study.

Briefly, twelve samples were selected based on previous data that was generated within the department. Subgenomic regions of these 12 samples were amplified and directly sequenced. HIV subtyping and phylogenetic analysis was then performed on the sequenced data. From the data that was gathered, four samples were chosen for full or near full-length genomic characterization. After amplification and sequencing the sequenced data were

analyzed with the use of phylogenetic methods, such as tree construction, subtyping, and recombination identification.

2.1 Reagents and equipment used in the study

All the materials and equipment that were used for the characterization of samples are described. A brief summary of all the equipment that was used are summarized in Table 2.1. The symbols ® and ™ indicate that the particular product are either a registered trademark or trademark of the suppliers.

Table 2.1: Equipment used to perform sample analysis.

Equipment	Supplier	Location
Eppendorf Centrifuge 5417C	Eppendorf	Hamburg, Germany
Eppendorf Centrifuge 5415D	Eppendorf	Hamburg, Germany
GeneAMP® 9700 PCR system	Applied BioSystems	California, USA
Hoefer EPS 2 A 200, Power Pack	Pharmacla Biotechnologies	San Francisco, CA, USA
NanodropTM ND 1000	Nanodrop Technologies Inc.	Delaware, USA
SyngeneTM GeneGenius Computer System	Synoptics Ltd.	Cambridge, UK
Vortex Mixer VM 300	Gemmy Industrial Corp.	Taipei, Taiwan
ABI 3130 xl Automated DNA sequencer	Applied BioSystems	California, USA

The chemicals and commercial products used in the study, and the various software packages used for sequence analysis are summarized in Tables 2.2 and 2.3 respectively.

Table 2.2: List of chemicals and commercial products used in the study.

Product	Company	Location	Catalogue Number
Big Dye® Terminator v 3.1 Cycle Sequencing Kit	Applied BioSystems	Foster City, CA, USA	0 211 005
Half Dye Mix	Bioline	London, UK	BIO-36026
GoTaq® DNA Polymerase	Promega	Madison, WI, USA	M 8305
Expand dNTP pack	Roche Diagnostics	Mannheim, Germany	11 681 834 001
Wizard SV Gel & PCR Clean-up kit	Promega	Madison, WI, USA	A 9281
dNTP's	Roche Diagnostics	Mannheim, Germany	11 636 103 001
Ethidium bromide	Promega	Madison, WI, USA	H 5041
Nuclease Free Water	Promega	Madison, WI, USA	P 1193
Molecular grade Agarose	Whitehead Scientific (Pty) Ltd.	Brackenfell, Cape Town, RSA	# D1 - LE
I kbp DNA Bench Top Ladder	Promega	Madison, WI, USA	G7541
Blue/Orange Loading Dye	Promega	Madison, WI, USA	G1881
PCR Product Pre-Sequencing Kit	USB Corporation	Cleveland, Ohio, USA	70995

Table 2.3: Software packages used in the analysis of sequenced data.

Software package	Reference / Licensed Company
Sequencher 4.8	Gene Codes Corporation, Ann Arbor, MI, USA
Sequence Scanner v 1.0	Applied Biosystems, Foster City, CA, USA
Clustal X	Thompson© *et al,* 1997
DNAMAN v 4.0	Lynnon BioSoft© 1994 – 1997
BioEdit v 5.0.9	Hall©, 2001
Tree view 1.6.6	Page©, 2001
MEGA v 4.1	Tamura *et al,* 2007
Simplot v 3.5.1	Stuart C Ray, Copyright© 1997 - 2003
PAUP* 4.0b10	Swofford, 2002
PhyML 3.0	Guindon and Gascuel, 2003
Geneious 4	Biomatters Ltd., Auckland, New Zealand
FigTree v 1.1.2	Rambaut A, Institute of Evolutionary Biology, University of Edinburgh

2.2 Patient samples

From the data of a previous study that was done within the department, that characterized the V3 region of 410 samples [Jacobs *et al,* submitted], 11 samples were chosen. The study identified 36 out of 410 samples (8.8%) as non-subtype C HIV-1 isolates but due to the limited amount of DNA available only 12 samples were chosen for further characterization. Ten of these samples were suspected to be non-subtype C HIV-1. The other sample was a known subtype C viral isolate, which was included for control purposes. A known subtype B isolate which was isolated from a homosexual male in the mid 1980's was also included for control purposes. A brief summary of the patient information are listed in Table 2.4.

Table 2.4: Patient samples and demographics.

Sample	Race and Gender	Year of Birth	Country of infection	Symptoms	CD4 Cells/µl	ARV treatment
R 84	Caucasian male	ND	SA	Asymptomatic	NA	No
TV 86	African Male	1965	SA	Cryptococcal meningitis	207	No
TV 101	African Female	1977	SA	Asymptomatic	2,000	No
TV 218	African Female	1975	SA	Asymptomatic	NA	ND
TV 239	African Male	1973	SA	TB	64	No
TV 314	African Male	1965	SA	Asymptomatic	229	No
TV 340	African Male	1963	DRC	TB abdomen severe thrush	3	No
TV 412	African Male	1955	Kenya	Chronic staph skin sepses	71	No
TV 441	African Female	1976	SA	Asymptomatic	178	Yes
TV 480	Coloured Female	1968	SA	Pheumonia	NA	No
TV 515	Coloured Female	1970	SA	NA	NA	No
TV 546	African Female	1970	SA	NA	NA	ND

Key: SA (South Africa), DRC (Democratic Republic of the Congo), TV (Tygerberg Virology), NA (Not available), TB (Tuberculosis), staph (staphylococcal), and ND (No data)

For DNA isolation, EDTA blood was centrifuged at 2,500 x g for 10 min, to separate blood plasma and buffy coat cells. The buffy coat cells were then used to extract DNA with the use of the QIAamp® DNA Mini Kit (QIAGEN, Hilden, Germany), according to manufacturer's specifications. The DNA was eluted in AE buffer and the concentrations were determined with the Nanodrop™ ND 1000 (Nanodrop Technologies Inc., Delaware, USA)

2.3 Amplification and sequencing of partial *gag*, *pol* and *env* fragments

2.3.1 PCR amplification of partial *gag, pol* and *env* fragments

PCR's were performed on the *gag* p24, *pol-integrase* and *env* gp41 regions of the 12 samples. All the primers used and other relevant data are summarized in Table 2.5. All PCR's that was used for the amplification of subgenomic regions were performed with the GeneAmp PCR System 9700 thermal cycler (Applied BioSystems, CA, USA) with GoTaq DNA polymerase (Promega, Madison, WI, USA).

Briefly, the PCR methods and primers were adapted from Swanson *et al,* 2003. Both the prenested and nested PCR reactions for the *gag, pol* and *env* regions contained 0.2mM of dNTP's, 20 µM of each primer, 1.5 mM of $MgCl_2$, and 1U of *Taq* polymerase in a total volume of 50 µl. The following cycling conditions were used for both the *gag*, *pol* and *env* PCR's: One cycle of denaturation at 94°C for 2 minutes; followed by 40 cycles of: denaturing at 94°C for 30 seconds, primer annealing for 30 seconds (Table 2.5), and elongation at 68°C for 1 minute; one final step of elongation at 68°C for 10 minutes and afterwards the samples were cooled down to 4°C until the PCR tubes were removed and stored at 4°C. Two and a half micro liters of the prenested product was carried over to the nested reaction.

2.3.2 Gel electrophoresis and sample clean-up of PCR fragments

PCR products of the *gag* p24, *pol-integrase* and *env* gp41 PCR's were run on 0.8% agarose gels (10 cm long) at 50 Volts for 45 minutes in TAE buffer (0.04 M TRIS-acetate & 0.001 M EDTA). After the samples migrated through the gels, the gels were stained with Ethidium Bromide (0.5µg/ml) and exposed to UV light and photographs were taken.

Samples were then cleaned up with the use of the PCR Product Pre-Sequencing Kit (USB Corporation, Cleveland, Ohio, USA). The kit uses two enzymes, Exonuclease I, which is responsible for removing any residual single-stranded primers or any other extraneous single-stranded DNA, and Shrimp Alkaline Phosphatase, which is responsible for the removing of any remaining dNTP's. The two enzymes were added to 8 µl of amplified product and incubated at 37°C for 15 minutes. Samples were then heated at 80°C for 15 min to inactivate the enzymes.

The concentrations of the cleaned up DNA products were determined with the Nanodrop^{TM} ND 1000 (Nanodrop Technologies Inc., Delaware, USA).

Table 2.5: List of different primers used in the amplification of the *gag* (p24), *pol-integrase* and the *env* (gp41) fragments.

	Primers	Oligonucleotide sequences	T_A (^oC)	HXB2 Position	F/R	[$MgCl_2$]	Size
PN *gag* p24	p24-1	AGYCAAAATTAYCCYATAGT	45	1174 - 1194	F	1.5 mM	671 bp
	p24-7	CCCTGRCATGCTGTCATCA	45	1844 - 1826	R	1.5 mM	
N *gag* p24	p24-2	AGRACYTTRAAYGCATGGGT	50	1237 - 1256	F	1.5 mM	485 bp
	p24-6	TGTGWAGCTTGYTCRGCTC	50	1721 - 1703	R	1.5 mM	
PN *pol-IN*	poli 5	CACACAAAGGRATTGGAGGAAATG	50	4162 - 4185	F	1.5 mM	1056 bp
	poli 8	TAGTGGGATGTGTACTTCTGAAC	50	5217 - 5195	R	1.5 mM	
N *pol-IN*	poli 6	ATACATATGRTGTTTTACTAARCT	45	5130 - 5107	R	1.5 mM	945 bp
	poli 7	AACAAGTAGATAAATTAGTCAGT	45	4186 - 4208	F	1.5 mM	
PN *env* gp41	JH38	GGTGARTATCCCTKCCTAAC	50	8365 - 8346	R	1.5 mM	568 bp
	JH41	TTATATAATTCACTTCTCCAATT	50	7775 - 7797	F	1.5 mM	
N *env* gp41	env 27F	CTGGYATAGTGCARCARCA	45	7861 - 7879	F	1.5 mM	439 bp
	Menv 19R	AARCCTCCTACTATCATTATRA	45	8299 - 8278	R	1.5 mM	

Key: PN (Prenested PCR), N (Nested PCR), TA (annealing temperature), F (Forward primer), R (Reverse primer), ^oC (degrees Celsius), bp (base pairs), IN (Integrase), and mM (millimoles per liter)

2.3.3 Sequencing of partial *gag*, *pol* and *env* PCR fragments

Amplified *gag* p24, *pol-integrase* and *env* gp41 fragments were all directly sequenced with the use of the primers listed in Table 2.6. All primers were described by Swanson *et al,* 2003 except for primer FGF 46 [Fong *et al,* 1996]. Samples TV101 and TV218 have already been characterized and was not sequenced with the other samples [Personal Communication, S Engelbrecht]. The BigDye^{TM} Terminator cycle sequencing ready reaction kit (Applied Biosystems, Foster City, California, USA) was used for the sequencing reactions.

Every sequencing reaction contained: approximately 50 ng of the purified PCR product, 5 pmol of sequencing primer, 1.3 µl of Big Dye terminator enzyme mix, and 2.7 µl of Half Dye (Bioline, London, United Kingdom). Nuclease free water was added to the reaction mix to give a final volume of 10 µl. Each sequencing reaction was performed under the following conditions: 25 cycles of denaturation at 96°C for 10 seconds, primer annealing for 5 seconds and an elongation step at 60°C for 4 minutes. Afterwards the samples were cooled down to 4°C and sent to the Central Analytical Facility of the University of Stellenbosch where they could be cleaned-up and run on the ABI 3130xl automated DNA sequencer.

The trace data files were received from the Central Analytical Facility and were imported into Sequencer 4.7 (Gene Codes Corporation, Ann Arbor., Michigan, USA) and assembled into contiguous fragments. After the assembled fragments were proofread they were exported in a text file (.txt) and labeled for later use.

Table 2.6: List of different sequencing primers for the *gag* p24, *pol-integrase* and the *env* gp41 fragments.

***gag* p24 Sequences**					
Primer	**Oligo nucleotide sequence**	**Bases**	**F / R**	**HXB2 Position**	**T_A (^oC)**
p24-2	AGRACYTTRAAYGCATGGGT	20	F	1237 - 1257	50
p24-6	TGTGWAGCTTGYTCRGCTC	19	R	1703 - 1721	50
***pol-int*egrase Sequences**					
Primer	**Oligo nucleotide sequence**	**Bases**	**F / R**	**HXB2 Position**	**T_A (^oC)**
poli7	AACAAGTAGATAAATTAGTCAGT	23	F	4186 - 4209	45
FGF-46	GCATTCCCTACAATCCCCAAAG	22	F	4648 - 4670	55
poli10b	TATTCATAGATTCYACTACTCCTTG	25	R	4695 - 4670	45
poli6	ATACATATGRTGTTTTACTAARCT	24	R	5130 - 5106	45
***env* gp41 Sequences**					
Primer	**Oligo nucleotide sequence**	**Bases**	**F / R**	**HXB2 Position**	**T_A (^oC)**
env 27 F	CTGGYATAGTGCARCARCA	19	F	7861 - 7880	50
Menv19R	AARCCTCCTACTATCATTATRA	22	R	8299 - 8277	45

Key: T_A (annealing temperature), F (Forward primer), R (Reverse Primer), and ^oC (degrees Celsius)

2.4 HIV-1 subtyping of partial *gag*, *pol* and *env* sequences using REGA and jpHMM online tools

HIV subtyping was performed on all the sequenced samples to establish the genomic variation of our cohort of 12 samples. Two different online subtyping tools were used for this purpose: the REGA subtyping tool (http://www.bioafrica.net/virus-genotype/html/subtyping.html) and the jumping profile Hidden Markov Model or jpHMM (http://jphmm.gobics.de/) which is also accessible form the LANL webpage (http://www.hiv.lanl.gov/). The REGA subtyping tool is an easy online method of subtyping full-length or subgenomic fragments by combining different phylogenetic analytical approaches with bootscanning methods [de Oliveira *et al,* 2005].

The jpHMM method uses a jumping alignment approach, first proposed by Spang and co-workers [Spang *et al,* 2002], for the subtyping of sequence fragments or the identification of recombinant viruses. Instead of a query sequence (X) being compared with a multiple alignment (Y), the query sequence (X) is compared and aligned to individual sequences from the alignment. In the case of recombination events the query sequence (X) can then jump between different sequences of the multiple alignment as a sliding window moves over the alignment. This approach also makes the identification of particular breakpoint within a recombinant isolate much easier [Schultz *et al,* 2006; Zhang *et al,* 2006].

2.5 Multiple alignments of the partial *gag*, *pol* and *env* sequences

Before multiple alignments could be constructed, reference sequences were obtained from the LANL database (http://www.hiv.lanl.gov/). For the alignment of *gag* p24 fragments, sequences of 485 bp and stretching from 1237 – 1721 (relative to HXB2) were obtained. Reference sequences corresponding to the same length and location of the *pol-integrase* and *env* gp41 PCR fragments, stretching from 4186 – 5130 (944 bp) and 7861 – 8299 (438 bp) relative to HXB2 respectively, were acquired. After the reference sequences were obtained multiple alignments were constructed with the use of Clustal X v 1.81

[Thompson© *et al,* 1997]. All alignments were manually checked and improved, where possible, with BioEdit v 5.09 [Hall©, 2001].

The isolate `N.CM.95.YBF30.AJ006022` was included as an outgroup in the alignments. Every name in the LANL database is specifically classified and annotated. For example in strain `N.CM.95.YBF30.AJ006022`, the `N` stands for Group N, `CM` the country of origin (Cameroon), `95` the year the sample was collected (1995), `YBF30` is the name of the virus strain and `AJ006022` is the GenBank Accession number.

2.6 Construction of NJ phylogenetic trees using MEGA

After the alignments were done and manually inspected, Neighbor-Joining phylogenetic trees were constructed. Adjusted multiple alignment files were imported into MEGA v 4.1 [Tamura *et al,* 2007] where the alignment files (.aln or .pir) were converted into MEGA format (.meg). The MEGA files (.meg) were then opened in MEGA and NJ-trees [Saitou and Nei, 1987] were constructed with the use of the Kimura 2 parameter method of nucleotide substitution [Kimura, 1980]. Bootstrap analysis [Felsenstein, 1985] was also performed on the trees to confer statistical significance, with a total of a 1,000 bootstrap replicates for each dataset.

2.7 Amplification and sequencing of NFLG's from 4 samples

After data analysis of the *gag pol* and *env* regions, full- or near full-length genome characterization of four of the twelve samples (R84, TV239, TV314 and TV412) were attempted. The entire genomes of the four different samples were first amplified in a single amplification assay to obtain 9.2 kbp products.

2.7.1 PCR amplification of the 9.2 kbp HIV-1 genome

Briefly, a prenested and nested PCR were carried out with the Expand long range dNTP pack (Roche Diagnostics, Mannheim, Germany) to obtain near-full length genome amplification of HIV samples (Figure 2.2).

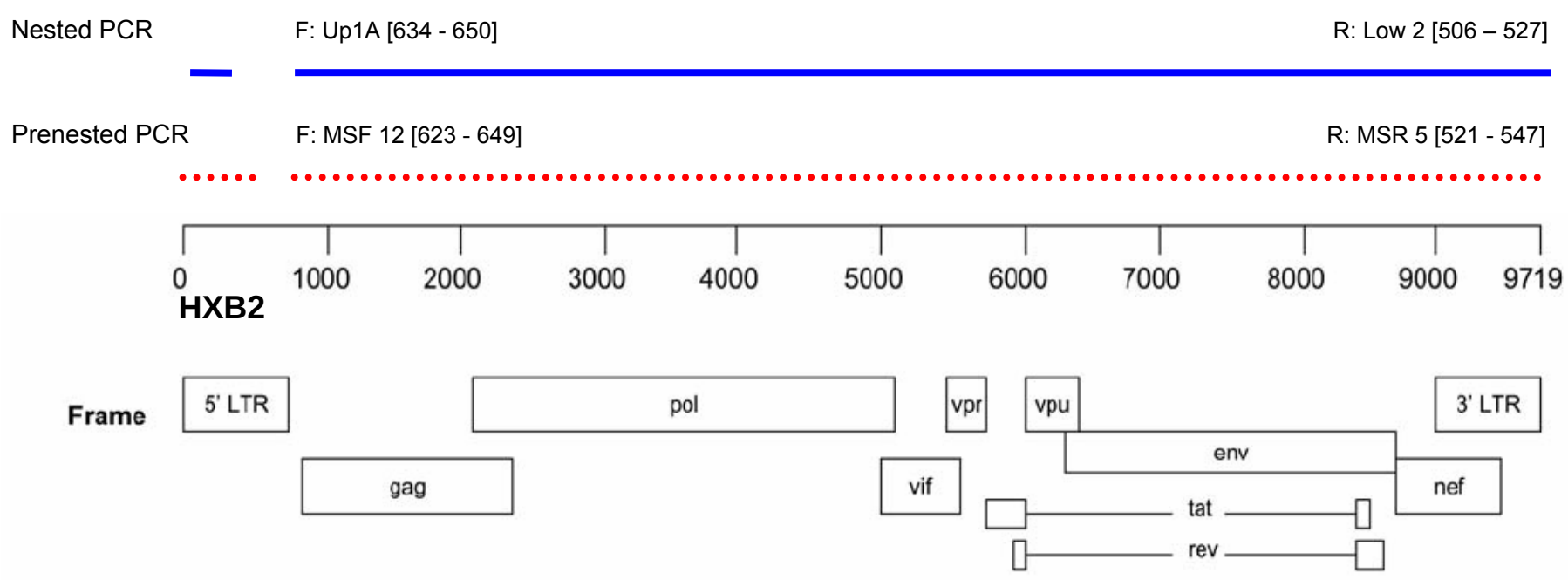


Figure 2.2: Schematic diagrams of the full genome amplification of an HIV isolate using a long PCR amplification assay. The prenested and the nested PCR are indicated on the diagram. Both forward and reverse primers as well as their coordinates (relative to HXB2) are indicated.

Each PCR reaction contained 0.2mM of dNTP's, 20 µM of each primer, 12 % DMSO solution and 5U of DNA polymerase in a total volume of 50 µl. The buffer that was used already contained $MgCl_2$. The rest of the volume was made up to 50 µl with nuclease free water (Promega, Madison, WI, USA).

The following cycling conditions were used for the prenested and nested amplification assays: one cycle of denaturation at 92°C for 2 minutes; followed by 10 cycles of denaturation at 92°C for 10 seconds, primer annealing for 30 seconds (Table 2.7), and elongation at 68°C for 10 minutes; which was followed by 30 cycles of; denaturation at 92°C for 10 seconds, primer annealing for 30 seconds, and elongation at 68°C for 10 minutes plus 20 seconds for each consecutive cycle, final elongation at 68°C for 10 minutes and afterwards the samples were cooled down to 4°C until the PCR reactions were collected. Five micro liters of the prenested product were carried over to the nested reaction.

Table 2.7: Primers that were used for the amplification of 9.2 kbp fragments.

	Full genome Primers					
	Primer	**Oligo nucleotide sequence**	**F / R**	**HXB2 Position**	**T_A (°C)**	**Size**
PN	MSF12[A]	AAATCTCTAGCAGTGGCGCCCCGAACAG	F	623 - 649	55	9.17 kbp
	MSR5[A]	GCACTCAAGGCAAGCTTTATTGAGGCT	R	9797- 9823	55	
N	UP1A[B]	AGTGGCGCCCGAACAGG	F	634 - 650	60	9.09 kbp
	LOW2[B]	TGAGGCTTAAGCAGTGGGTTTC	R	9706 - 9727	60	

Key: PN (Prenested PCR), N (Nested PCR), F (Forward primer), R (Reverse primer), °C (degrees Celsius), and kbp (kilo base pairs) [[A] Rodenburg *et al,* 2001; [B] Mwaengo and Novembre, 1998].

The following cycling conditions were used for the prenested and nested amplification assays: one cycle of denaturation at 92°C for 2 minutes; followed by 10 cycles of denaturation at 92°C for 10 seconds, primer annealing for 30 seconds (Table 2.7), and elongation at 68°C for 10 minutes; which was followed by 30 cycles of; denaturation at 92°C for 10 seconds, primer annealing for 30 seconds, and elongation at 68°C for 10 minutes plus 20 seconds for each consecutive cycle, final elongation at 68°C for 10 minutes and afterwards the samples were cooled down to 4°C until the PCR reactions

were collected. Five micro liters of the prenested product were carried over to the nested reaction.

2.7.2 Amplification of the HIV genome in four overlapping fragments

The original proviral DNA was then used for the amplification of a near-full length genome consisting of four overlapping PCR fragments (Figure 2.3) spanning the entire genome of the four isolates.

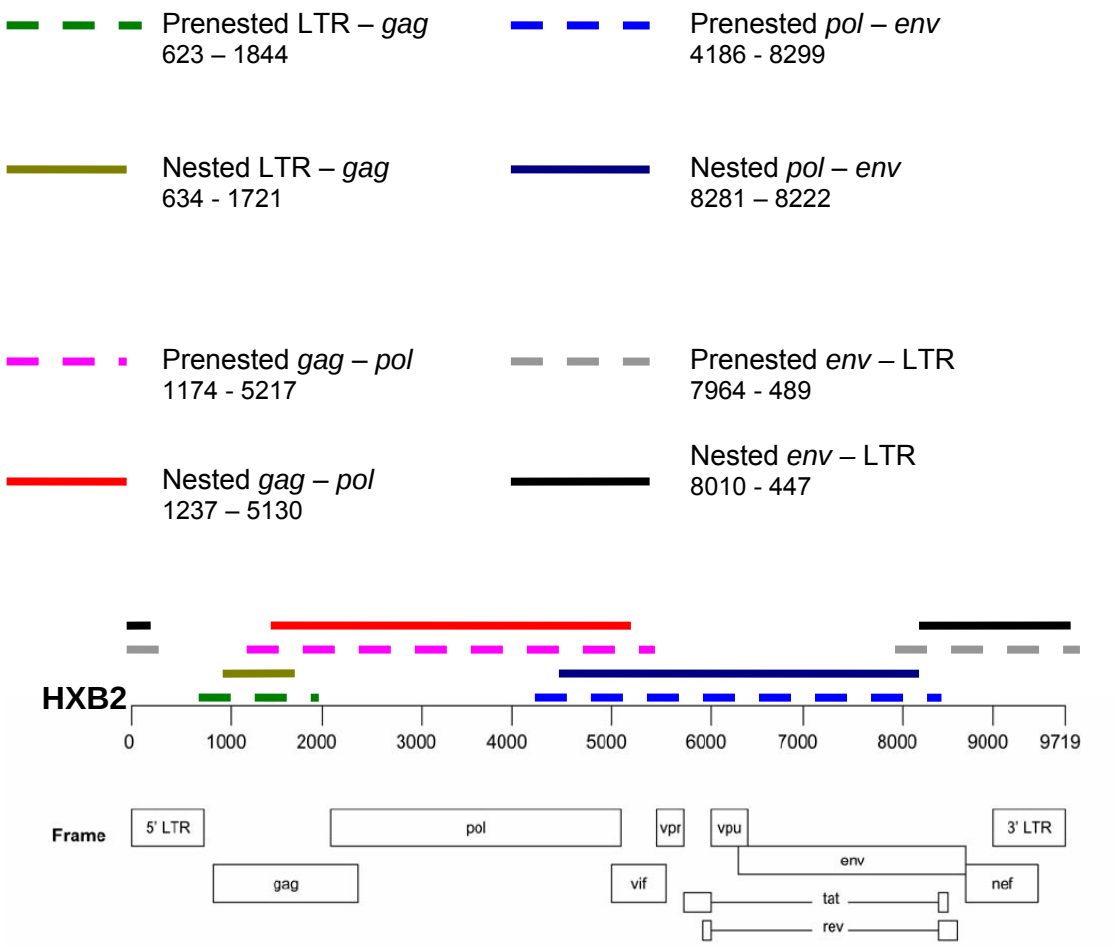


Figure 2.3: Schematic diagrams of the full genome amplification of an HIV isolate in four overlapping fragments. The prenested and nested amplifications and their approximate location (relative to HXB2) are indicated. All dotted lines indicate prenested reactions and all solid lines nested reactions.

Table 2.8: Primers that were used to PCR the overlapping fragments (LTR-*gag*, *gag-pol*, *pol-env*, and *env*-LTR).

	LTR-*gag* Primers						
	Primer	**Oligo nucleotide sequence**	**F / R**	**HXB2 Position**	**T_A (^oC)**	**Size**	**Ref**
PN	MSF12	AAATCTCTAGCAGTGGCGCCCCGAACAG	F	623 - 649	52	1.21 kbp	Rodenburg *et al,* 2001
	P24-7	CCCTGRCATGCTGTCATCA	R	1826 - 1844	52		Swanson *et al,* 2003
N	UP1A	AGTGGCGCCCGAACAGG	F	634 - 650	50	1.09 kbp	Mwaengo and Novembre, 1998
	P24-6	TGTGWAGCTTGYTCRGCTC	R	1703 - 1721	50		Swanson *et al,* 2003
	***gag-pol* Primers**						
	Primer	**Oligo nucleotide sequence**	**F / R**	**HXB2 Position**	**T_A (^oC)**	**Size**	**Ref**
PN	p24-1	AGYCAAAATTAYCCYATAGT	F	1174 - 1193	45	4.04 kbp	Swanson *et al,* 2003
	poli 8	TAGTGGGATGTGTACTTCTGAAC	R	5195 - 5217	45		
N	p24-2	AGRACYTTRAAYGCATGGGT	F	1237 - 1256	50	3.89 kbp	
	poli 6	ATACATATGRTGTTTTACTAARCT	R	5107 - 5130	50		

Key: PN (Prenested PCR), N (Nested PCR), LTR (Long terminal repeat), ^oC (degrees Celsius), T_A (annealing temperature), F (Forward primer), R (Reverse primer), Ref (Reference) and kbp (kilo base pairs)

Table 2.8 continued: Primers that were used to PCR the overlapping fragments (LTR-*gag*, *gag-pol*, *pol-env*, and *env*-LTR).

	***pol-env* Primers**						
	Primer	**Oligo nucleotide sequence**	**F / R**	**HXB2 Position**	**T_A (^oC)**	**Size**	**Ref**
PN	poli 7	AACAAGTAGATAAATTAGTCAGT	F	4186 - 4208	45	4.11 kbp	Swanson *et al,* 2003
	Menv 19	AARCCTCCTACTATCATTATRA	R	8278 - 8299	45		
N	PPF17	AATTGGAGAGCAATGGCTAGTGA	F	4281 - 4303	50	3.94 kbp	
	LP 7728	CCACTTGTCCAATGCCAATAAGTCTTGT	R	8195 - 8222	50		
	***env*-LTR Primers**						
	Primer	**Oligo nucleotide sequence**	**F / R**	**HXB2 Position**	**T_A (^oC)**	**Size**	**Ref**
PN	7496 F	CCTKGCYCTGGAAAGATACCTA	F	7964 - 7985	52	1.70 kbp	Personal Communication John Hackett
	9131R-2	CTCYCAGGCTCARATCTGGTC	R	468 - 489	52		
N	7542 F	TGGGGCTGCTCTGGAAAACT	F	8010 - 8029	50	1.64 kbp	
	9110R-2	CAAGAGAGACCCAGTACAG	R	447 - 465	50		

Key: PN (Prenested PCR), N (Nested PCR), LTR (Long terminal repeat), ^oC (degrees Celsius), T_A (annealing temperature), F (Forward primer), R (Reverse primer), Ref (Reference) and kbp (kilo base pairs)

Table 2.9: PCR cycling condition of the four overlapping fragments.

Cycle(s)	Reaction	Temperatures and Time							
		LTR-*gag* PCR		*gag-pol* PCR		*pol-env* PCR		*env*-LTR PCR	
		Prenested PCR	Nested PCR	Prenested PCR	Nested PCR	Prenested PCR	Nested PCR	Prenested PCR	Nested PCR
1 x	Template denaturing	94°C, 2 m	94°C, 2 m	94°C, 2 m	94°C, 2 m	94°C, 2 m	94°C, 2 m	94°C, 2 m	94°C, 2 m
40 x	Template denaturing	94°C , 30 s	94°C , 30 s	94°C , 30 s	94°C , 30 s	94°C , 30 s	94°C , 30 s	94°C , 30 s	94°C , 30 s
	Primer Annealing	52°C, 30 s	51°C, 30 s	45°C, 30 s	45°C, 30 s	45°C, 30 s	50°C, 30s	52°C, 30 s	50°C, 30 s
	Elongation	68°C, 90 s	68°C, 90 s	68°C, 4 m	68°C, 4 m	68°C, 4 m	68°C, 4 m	68°C, 2 m	68°C, 2 m
1 x	Final Elongation	68°C, 10 m	68°C, 10 m	68°C, 10 m	68°C, 10 m	68°C, 10 m	68°C, 10 m	68°C, 10 m	68°C, 10 m
1 x	Storing	4°C, Indef	4°C, Indef	4°C, Indef	4°C, Indef	4°C, Indef	4°C, Indef	4°C, Indef	4°C, Indef

Key: x (times), PCR (polymerase chain reaction),°C (degrees Celsius), m (minutes), s (seconds), and Indef (indefinitely).

Each PCR reaction contained 0.2mM of dNTP's, 20 µM of each primer, 1.5 mM of $MgCl_2$, and 1U of *Taq* polymerase in a total volume of 50 µl. Five micro liters of the prenested product (ranging between 10 and 50 µg/µl) was carried over to each of the nested reaction. Table 2.9 gives a brief summary of the cycling conditions that was used.

2.7.3 Gel electrophoresis and clean-up of NFLG PCR fragments

PCR products were run on 0.8% agarose gels (10 cm in length) at 50 Volts for 45 minutes in TAE buffer (0.04 M TRIS-acetate & 0.001 M EDTA). A 1kbp molecular marker was run in parallel with all samples. After the samples migrated through the gels, the gels were stained with Ethidium Bromide (0.5 µg/ml) and exposed to UV light before photographs were taken.

The Wizard SV gel and PCR clean-up kit from Promega, Madison, Wisconsin, USA were used to purify the amplified products of any unwanted dNTP's or oligonucleotides. The concentrations of the cleaned up products were determined with the Nanodrop™ ND 1000 (Nanodrop Technologies Inc., Delaware, USA) after they were eluted from the spin column.

2.7.4 Sequencing of NFLG PCR fragments

The various amplified products were directly sequenced by employing primer walking techniques. Appropriate primers were chosen for every 400 – 500 base pairs in both the forward and reverse directions (John Hackettt, personal communication). All the sequencing primers use in the sequencing of each isolate is listed in Tables 6.4 – 6.7 in Appendix C, Chapter 6.

The BigDye™ Terminator cycle sequencing ready reaction kit (Applied Biosystems, Foster City, California, USA) was used for the PCR based sequencing reactions. Approximately 50 ng of the purified PCR product were used with, 5 pmol of sequencing primer, 1.3 µl of Big Dye terminator enzyme

mix, and 2.7 µl of Half Dye (Bioline, London, United Kingdom). Nuclease free water was added to the mix to give a final volume of 10 µl of reaction mix. Each sequencing reaction were performed under the following sequencing cycling reaction conditions: 25 cycles of denaturization at 96°C for 10 seconds, primer annealing for 5 seconds and an elongation step at 60°C for 4 minutes. Afterwards the samples were cooled down to 4°C and sequenced at the Central Analytical Facility at the University of Stellenbosch, as described before. After the trace data files were recovered from the Central Analytical Facility they were imported into Sequencer 4.7 (Gene Codes Corporation, Ann Arbor., Michigan, USA) were they were assembled into contiguous fragments. After the assembled fragments were proofread they were exported as text files (.txt).

The contiguous fragments were also analyzed with the use of sequence tools such as the Gene Cutter tool of the LANL database [http://www.hiv.lanl.gov/]. The Gene Cutter tool [Goldman and Yang, 1994] is an online tool from the Los Alamos National Laboratory which can be used to identify the different genes within sequences, produce information on the amino acid sequence, and identify premature stop codons and sites with multi-state characters. Sequences can be submitted in an aligned or unaligned format.

2.8 Phylogenetic analysis of near full-length genome sequences

After all sequences were proof-read, quality controlled phylogenetic analysis could be conducted. Briefly, the DNA sequence fragments were used to perform subtyping to establish viral subtypes and/or recombinants. Multiple alignments were constructed and phylogenetic trees were drawn. Bootstrap analysis was performed on all trees for statistical purposes. Recombination identification was also performed to identify recombination events within isolates. Non-recombinant HIV isolates were also compared against samples of the same subtype for further in-depth analysis.

2.8.1 Subtyping with REGA and jpHMM tools

All the near full-length fragments were submitted to two online subtyping tools; the jpHMM (http://jphmm.gobics.de/ which is also accessible at the LANL website http://hiv.lanl.gov./) and the REGA subtyping tools (http://www.bioafrica.net/virus-genotype/html/subtyping.html).

2.8.2 Construction of a multiple alignment of NFLG sequences

Reference sequences were obtained from the LANL database before multiple alignments could be constructed. All non-recombinant subtypes of the reference set, as well as the most prominent recombinant viruses, were included in the datasets that was used for the various alignments. Due to the varying length of the different fragments reference sequences ranging from 1230 - 8700 (relative to the coordinates of the reference strain HXB2) were downloaded. Because no contiguous fragment could be obtained for TV239, reference sequences in two fragments stretching from 1246 - 5534 and 5908 - 9106 (relative to the coordinates of the reference strain HXB2) were downloaded. The reference strain N.CM.95.YBF30.AJ006022 was included in each dataset for the use as an outgroup.

Multiple alignments were then constructed with the use of Clustal X v 1.81 [Thompson© *et al,* 1997]. The three isolates; R84, TV314, and TV412 were all aligned in a single alignment. Due to a gap of nearly 660 bp in between the two fragments of TV239, a separate alignment was performed for each of these fragments. After the alignments were created in Clustal X, they were manually checked with BioEdit v 5.09 [Hall©, 2001].

2.8.3 Construction of phylogenetic trees

Neighbor-Joining trees [Saitou and Nei, 1987] were drawn with MEGA v 4.1 [Tamura *et al,* 2007] of all the data. Three trees were drawn in total, one containing the NFLG fragments of R84, TV314 and TV412 with reference sequences, and two trees for the two fragments (*gag-pol* and *env-nef*) of

TV239 also with reference sequences. The Kimura 2 parameter method of nucleotide substitution [Kimura, 1980] was used and a total of a 1000 bootstrap replicates was performed on each of the trees. The datasets were also analyzed with PAUP* version 4.0b10 (Phylogenetic Analysis using Parsimony* and other methods) [Swofford, 2002] for the construction of maximum likelihood trees.

The methods used by PAUP*, though far more thorough [Salemi and Vandamme, 2003], are extremely computationally intensive and the output format of the data needs far more additional work than with the use of other tree drawing software such as MEGA.

2.9 Detection of recombinant viruses using RIP and Simplot

Recombinant identification was performed with two widely used methods: Simplot [Lole *et al,* 1999] and the Recombinant Identification Program, RIP [Siepel *et al,* 1995]. The consensus alignment (excluding CRF01_AE) of RIP was used to query the sequences of interest. All NFLG fragments (R84, TV314 and TV 412) and the *gag-pol* and *env-nef* fragments of isolate TV239 were queried. The raw text files (.txt) of each of the fragments were uploaded into the program and a window size of 300 bp was chosen.

The multiple alignments were also imported into Simplot version 3.5 [Lole *et al,* 1999; Salminen *et al,* 1995b] to identify recombination events within the NFLG samples and the two fragments of TV239. For the analysis of the NFLG fragments all three isolates (R84, TV314 and TV 412) were queried. A window size of 350 bp and a step size of 50 bp were selected. The *gag-pol* and *env-nef* fragments of isolate TV239 were also queried with the use of the same window and step size as with the analysis of the NFLG fragments.

All sequences (R84, TV314 and TV 412) were first queried against all subtypes in the alignments for both the Simplot and the RIP analysis before each of the sequence analysis were rerun to simplify the output format.

After the recombinant identification was completed Neighbor-Joining trees of the different recombinant sections were drawn. Reference sequences of pure subtype were obtained from the LANL database. Breakpoint coordinates which corresponded with the breakpoints that were obtained from the subtype and recombination identification done by the jpHMM analysis was used to obtain genome segments which corresponded with the same area of suspected viral recombination. Only a small number (10-15) of reference samples, which represented the most prominent viral subtypes were included in these datasets. After the various reference sequences were obtained multiple alignments of the different recombinant fragments were constructed with the use of Clustal X v 1.81 [Thompson© *et al,* 1997]. After the alignments were completed and was manually checked with BioEdit v 5.09 [Hall©, 2001] the different alignment files (.aln) were converted to MEGA format (.meg). These files were imported into MEGA v 4.1 [Tamura *et al,* 2007] to construct Neighbor-joining phylogenetic trees [Saitou *et al,* 1987]. The Kimura 2-parameter [Kimura, 1980] was employed for the construction of the NJ-trees. A 100 bootstrap replicates were performed on all these datasets to confer statistical significance.

2.10 Phylogenetic analysis of non-recombinant NFLG sequences

The two isolates of a non-recombinant nature were also compared to other non-recombinant subtypes of A and B. From the previous analysis it could be concluded that sample R84 was an HIV-1 subtype B virus and TV 314 a subtype A1 virus. Subtype B and A1 sequences were obtained from the LANL database to compare with the sequences of interest. A BLAST was performed with R84 and TV314 and full-length sequences which were most closely related to the isolates were downloaded. Because R84 was collected in the mid 1980's more subtype B viruses from the 1980's and early 1990's were included in the subtype B dataset. After the full-length sequences were retrieved from the database, multiple alignments of the reference sequences with the sequences of interest were constructed with Clustal X v 1.81 [Thompson© *et al,* 1997]. The isolate K.CM.96.MP535.AJ249239 was included for the use as an outgroup and all possible output formats were selected for

each of the alignments. Each alignment were manually checked with BioEdit v 5.09 [Hall©, 2001] after the alignments were done.

After the alignments were completed the alignment files (.aln) were converted to MEGA format (.meg). These files were then imported into MEGA v 4.1 [Tamura *et al,* 2007] and Neighbor-Joining phylogenetic trees [Saitou *et al,* 1987] were constructed with the use of the Kimura 2-parameter [Kimura, 1980]. Bootstrap analysis, with a total of a 1000 bootstrap replicates, were also performed on each of the datasets to infer statistical significance.

CHAPTER THREE – RESULTS

Table of Content Page

CHAPTER THREE - RESULTS

3.1 PCR amplification of partial *gag*, *pol* and *env* fragments

The PCR results of the three subgenomic regions of the 12 samples are summarized in Table 3.1. Sample TV546 could not be amplified with the *gag* p24, *pol-integrase* and *env* gp41 amplification assays. TV546 was previously characterized as a subtype G sample [Personal communication, Susan Engelbrecht] and thus may require the use of subtype specific primer sets. Other than TV 546 only the *env* gp41 amplification of sample TV 480 was unsuccessful. An example of the PCR results is illustrated in Figure 3.1

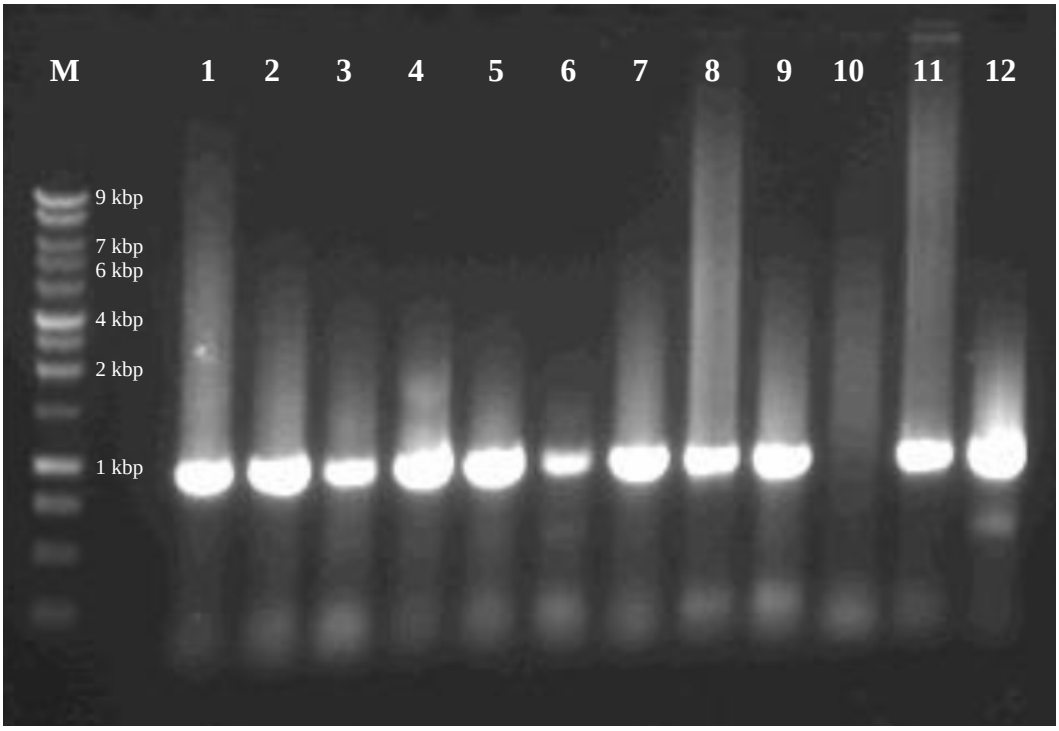


Figure 3.1: Agarose gel electrophoresis of the nested *pol-integrase* PCR products. Lanes: Lane M – 1kb marker, Lane 1 – TV412, Lane 2 – TV314, Lane 3 – TV340, Lane 4 – TV218, Lane 5 – TV239, Lane 6 – TV 101, Lane 7 – TV480, Lane 8 – TV 86, Lane 9 – TV515, Lane 10 – TV546, Lane 11 – TV 412, and Lane 12 – R84.

Table 3.1 PCR amplification of the subgenomic regions of the 12 samples.

PCR Results			
Sample	***gag* p24**	***pol-integrase***	***env* gp41**
R84	Positive	Positive	Positive
TV 86	Positive	Positive	Positive
TV 101	Positive	Positive	Positive
TV 218	Positive	Positive	Positive
TV 239	Positive	Positive	Positive
TV 314	Positive	Positive	Positive
TV 340	Positive	Positive	Positive
TV 412	Positive	Positive	Positive
TV 441	Positive	Positive	Positive
TV 480	Positive	Positive	Negative
TV 515	Positive	Positive	Positive
TV 546	Negative	Negative	Negative

The sequencing of positive PCR products was successful with one exception. The *gag* p24 sequence of TV340 could not be used due to multiple peaks in the electropherogram. The sequencing results of the three subgenomic fragments are summarized in Table 3.2. The text files of the sequences are presented in Appendix A. The sequences were also submitted to Genbank with accession numbers FJ647150 - FJ647168.

Table 3.2: Sequencing results for the subgenomic regions.

Sequencing Results			
Sample	***gag* p24**	***pol-integrase***	***env* gp41**
R84	Positive	Positive	Positive
TV 86	Positive	Positive	Positive
TV 101	Previous data*	Previous data*	Previous data*
TV 218	Previous data*	Previous data*	Previous data*
TV 239	Positive	Positive	Positive
TV 314	Positive	Positive	Positive
TV 340	Negative	Positive	Positive
TV 412	Positive	Positive	Positive
TV 441	Positive	Positive	Positive
TV 480	Positive	Positive	Negative
TV 515	Positive	Positive	Positive

Key: * Previous data (Personal communication, Susan Engelbrecht)

3.2 Subtyping of the partial *gag*, *pol* and *env* regions

The results of the viral subtyping done with the REGA and jpHMM viral subtyping tools as well as the NJ-trees of the three genomic fragments are

presented in sections 3.3.1 (*gag* p24), 3.3.2 (*pol integrase*) and 3.3.3 (*env* gp41) respectively. The phylogenetic data and subtyping results are summarized in Table 3.3.

The results of the online viral subtyping will be presented in Appendix B Figures 6.1– 6.3 and Tables 6.1 – 6.3, illustrates the REGA and jpHMM results respectively.

3.3 NJ phylogenetic tree analysis of partial *gag*, *pol* and *env* fragments.

The results of the NJ phylogenetic analysis for the *gag, pol* and *env* fragments will be discussed in the following sections (3.3.1 – 3.3.3).

3.3.1 The *gag* p24 region

For the NJ-tree of the 10 *gag* p24 sequences (Figure 3.2) with the reference sequences from the LANL database [http://www.hiv.lanl.gov/], four samples (TV314, TV412, TV239, and TV101) clustered with subtype A1 sequences. One sample, TV515, clustered with subtype F1 sequences in the tree. Sample R84, clustered with other subtype B strains in the tree. The other four samples (TV86, TV218, TV441, and TV480) clustered amongst other subtype C isolates in the tree. All these subtypes were confirmed with high bootstrap values. The clustering pattern of the *gag* sequences of the 10 samples in the NJ-tree corresponded with the subtyping results that were obtained with the REGA and jpHMM analysis.

Table 3.3: Subtyping analysis performed on the *gag* p24, *pol-integrase* and *env* gp41 fragments.

Sample	***gag* p24**			***pol-integrase***			***env* gp41**			**Assumed subtype**
	jpHMM viral subtyping	REGA viral subtyping	Clustering pattern in NJ-tree	jpHMM viral subtyping	REGA viral subtyping	Clustering pattern in NJ-tree	jpHMM viral subtyping	REGA viral subtyping	Clustering pattern in NJ-tree	
R 84	B	B	B	B	B	B	B	B	B	B
TV86	C	C	C	C	C	C	C	C	C	C
TV101	A1	A1	A1	A1D	A1*	Unclass	A1	A1	A1	A1 Recombinant
TV218	C	C	C	C	C	C	A1	A1*	A2	CA Recombinant
TV239	A1	A1	A1	A1H	A1*	A1	A1	A1*	A1 / A2	A1 Recombinant
TV314	A1	A1	A1	A1J	A1*	A1	A1	A1*	A1 / A2	A1
TV340	-	-	-	GB	G*	G	A1	A1	A1	ABG Recombinant
TV412	A1	A1	A1	A1DJ	A1*	A1	A1	A1	A1	A1 Recombinant
TV441	C	C	C	C	C*	C	A1	A1*	A2	CA Recombinant
TV480	C	C	C	CJ	C*	C	-	-	-	C Recombinant
TV515	F1	F1	F1	F1	F1	F1	F1	F1	F1	F1

Key: * (the bootstrap values of the analysis was not supportive or < 70%), multi state characters e.g. A1D (indicates possible recombination events), / (sample clustered with both the subtypes or sub-subtypes), Unclass (Unclassified), and – (PCR or sequencing reactions were not successful)

3.3.2 The *pol-integrase* region

In the *pol-integrase* tree (Figure 3.3), TV515 and R84, clustered with other F1 and B sequences respectively. Three samples (TV412, TV314, and TV239) clustered with other A1 isolates. These three samples were identified as A1 recombinants with the REGA and jpHMM subtyping and recombination analysis (Table 3.2). TV101 did not cluster with any of the 9 subtypes in the tree, which indicates possible viral recombination in this sequence. jpHMM analysis indicated that TV101 might be an AD recombinant within the *pol-integrase* sequenced fragment. Sample TV441 was an outlier of the C cluster, but with the online subtyping results was continuously identified as a subtype C isolate. TV340 was an outlier of the subtype G cluster which might indicate possible recombination. The REGA subtyping tool identified TV340 as an subtype G sample but with a very low bootstrap support. jpHMM analysis of TV340 however identified the sample as AG recombinant form.

3.3.3 The *env* gp41 region

In the *env* NJ-tree three (Figure 3.4) the following samples clustered with other A1 samples: TV101, TV340 and TV412 which corresponds well with the subtyping results of the REGA and jpHMM analysis. TV218 and TV441 clustered with other A2 sequences, with TV239 and TV314 outliers of the A1/A2 cluster, indicating possible recombination. REGA and jpHMM analysis of these four samples revealed that all these samples (TV218, TV239, TV314 and TV412) were subtype A1 in the env gp41 region. As with the *gag* and *pol* trees TV86, TV515, and R84 clustered with other C, F1, and B sequences respectively and corresponded well with the online subtyping results.

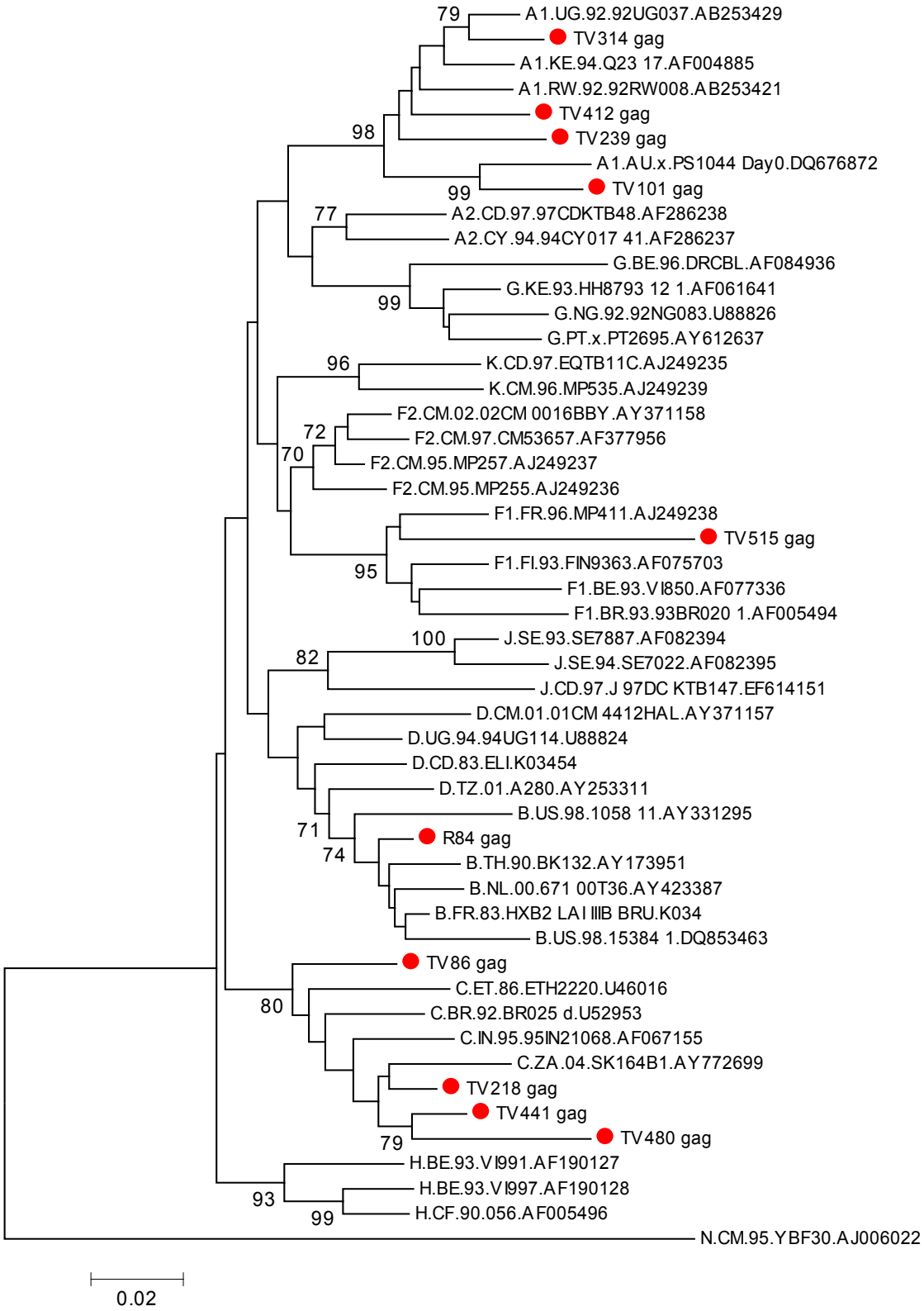


Figure 3.2: A Neighbor-joining tree of *gag* sequences (485 bp) indicating reference sequences and TV sequences. The TV sequences are indicated with a red dot. In the bottom line the genetic distance, which corresponds to the length of the branches, is shown. Only bootstrap values greater than 70 percent are included.

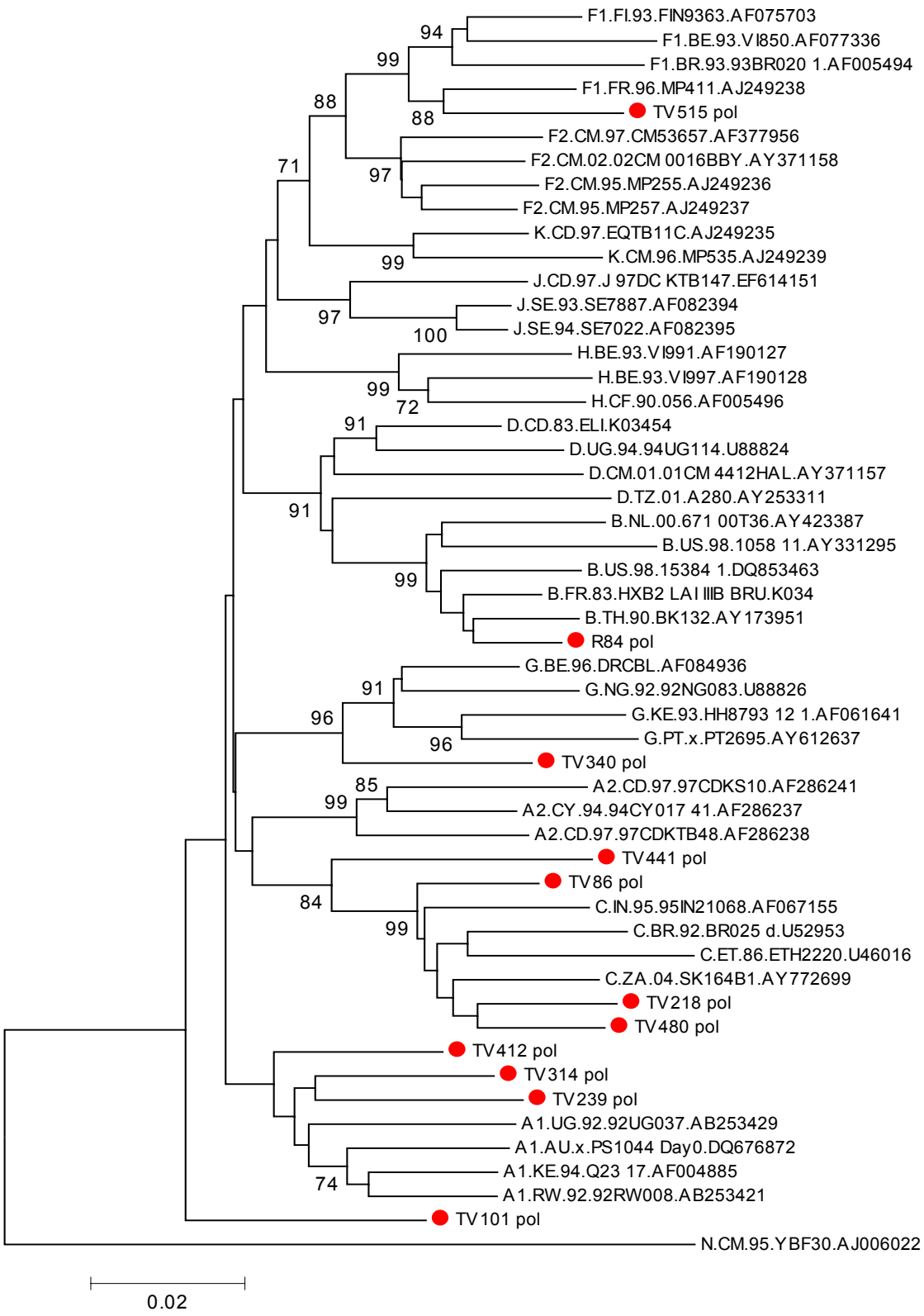


Figure 3.3: A Neighbor-joining tree of *pol* sequences (944 bp) indicating reference sequences and TV sequences. Each TV sequences are indicated with a red dot. The genetic distance, which corresponds to the length of the branches, is shown in the bottom line. Only bootstrap values greater than 70 percent are included.

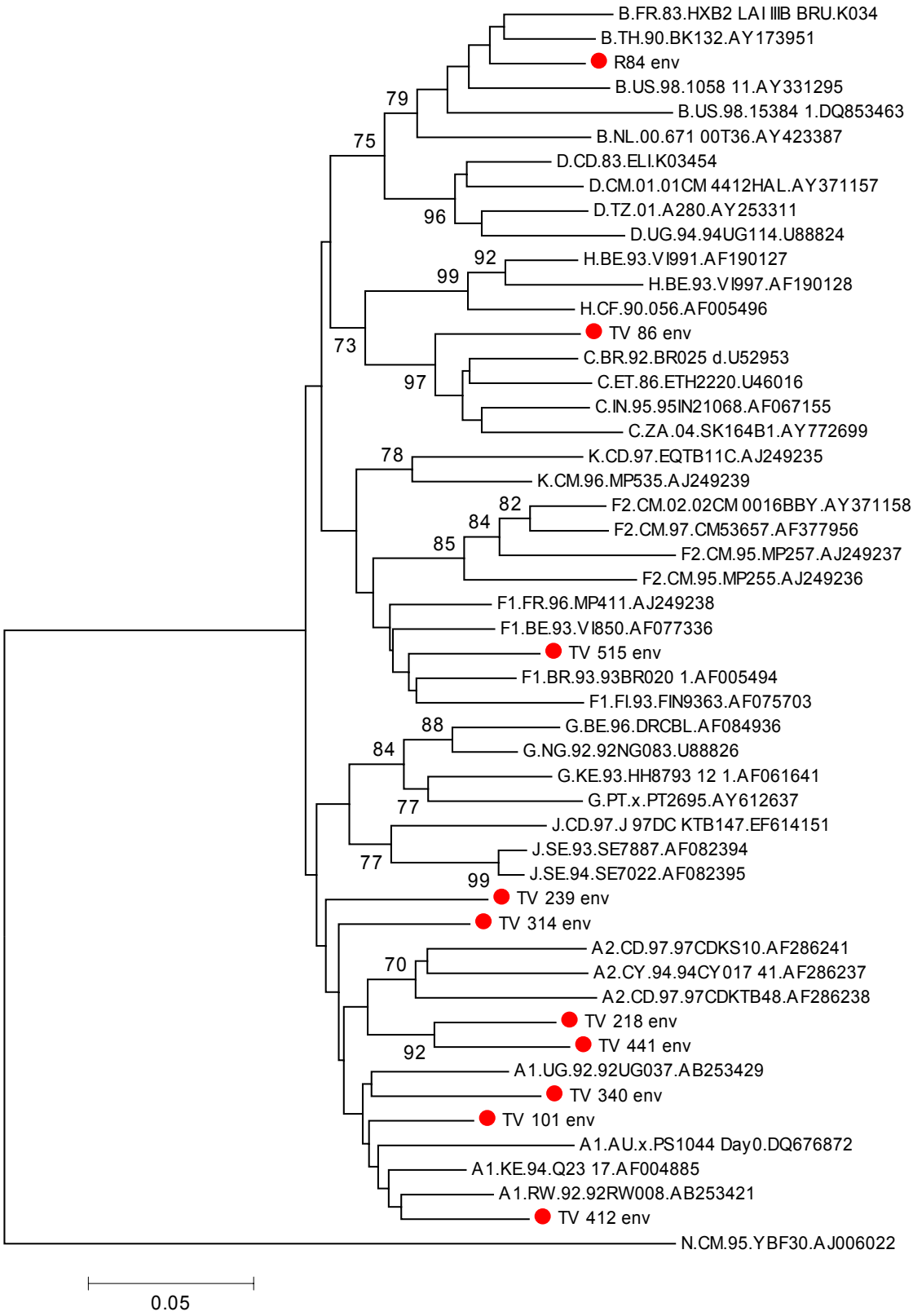


Figure 3.4: A Neighbor-joining tree of *env* sequences (438 bp) indicating reference sequences and TV sequences. Each TV sequences are indicated with a red dot. The genetic distance, which corresponds to the length of the branches, is shown in the bottom line. Only bootstrap values greater than 70 percent are included.

3.4 Amplification and sequencing of NFLG of four samples

3.4.1 PCR amplification assays of 9.2 kbp fragments of four samples

The long amplification assays (9.2 kbp) of the four samples (R84, TV239, TV314, and TV412) were not successful. Due to limited template DNA, a large amount of 9.2 kbp PCR products could not be obtained for sequencing. An example of the results of long PCR is shown in Figure 3.5.

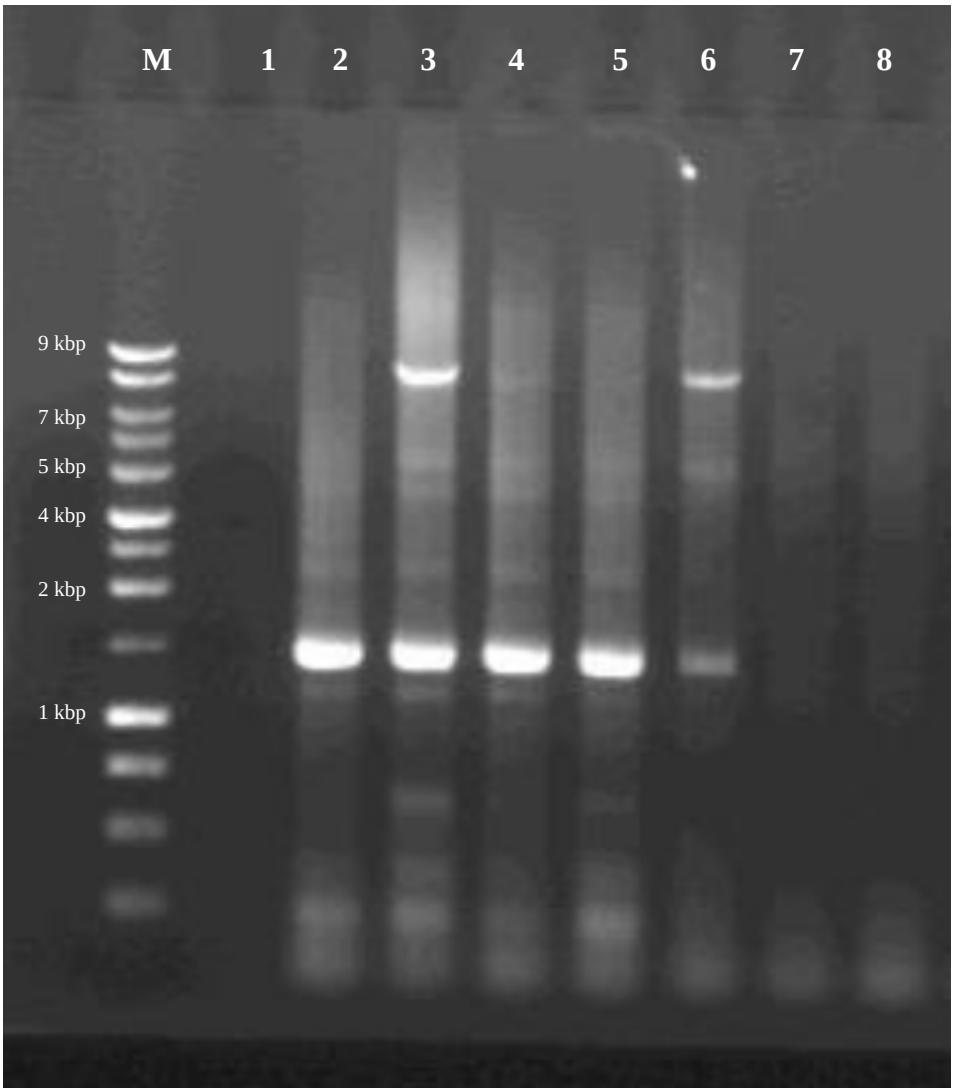


Figure 3.5: Agarose gel of the unsuccessful 9.2 kbp PCR products of sample R84. Lanes: lane M – 1 kbp Molecular Marker, lane 1 - Blank, lane 2 – 500 ng/µl of template DNA, lane 3 – 400 ng/µl of template DNA, lane 4 - 300 ng/µl of template DNA, lane 5 - 200 ng/µl of template DNA, lane 6 – 100 ng/µl of template DNA, lanes 7 & 8 – Negative Control

3.4.2 PCR amplification of four overlapping fragments

Sample R84 were amplifiable for all four fragments (LTR-*gag*, *gag-pol*, *pol-env*, and *env*-LTR). For the other three samples (TV239, TV314, and TV412), only the amplification of the *gag-pol*, *pol-env*, and *env*-LTR fragments were successful (data not shown). The results of the PCR amplification assays are summarized in Table 3.4.

Table 3.4: PCR amplification of the four overlapping fragments.

Sample	Fragment			
	LTR-*gag*	*gag-pol*	*pol-env*	*env*-LTR
R 84	Positive	Positive	Positive	Positive
TV239	Negative	Positive	Positive	Positive
TV314	Negative	Positive	Positive	Positive
TV412	Negative	Positive	Positive	Positive

3.4.3 Sequences results of the NFLG fragments

A continues fragment of sample R84, stretching from the start of the *gag* region up to the 3'LTR region (position 601-9514 relative to HXB2) were obtain from the sequenced data. The sequencing of the other samples was obtained with mixed results. All attempts to sequence the 5'LTR-*gag* regions of the other three isolates (TV239, TV314, and TV412) as well as the 660 bp gap between the two fragments of sample TV239 from the 9.2 kbp PCR fragment were unsuccessful. This resulted in; two fragments for TV 239 stretching from position 1245 - 5534 and 6195 - 9146 (relative to HXB2 coordinates), a single continues fragment for TV314 stretching from 1235 - 9551 and a single fragment, stretching from position 1246 – 8254, for sample TV412.

Each sequence fragment was analyzed with the Gene Cutter tool from LANL database. No pre-mature stop codons could be found in any of the open reading frames of the various genes. The results of the Gene Cutter analysis for each of the four isolates sequences are presented in Appendix D. All the near full-length sequences were submitted to GenBank with accession numbers FJ647145 - FJ647149.

3.5 Subtype identification of NFLG fragments with REGA and jpHMM online tools.

The online subtyping and recombination results, which was performed with the jpHMM and REGA viral subtyping tools, of the three near full-length sequences as well as the two fragments of sample TV239 are summarized in Table 3.5. For the full results of the analysis please refer to Appendix E, Figure 6.4 and Table 6.8, for the REGA and jpHMM results respectively.

Table 3.5: jpHMM and REGA subtyping tools.

Sample	jpHMM subtyping tool	REGA subtyping tool
R 84	B	B*
TV 239 *gag-pol*	A1	A1*
TV 239 *env-nef*	C / A1 / C / A1 / C	C / A1 / C *
TV314	A1	A1*
TV 412	A1 / D / A1 / D / A1	A1 / D / A1*

Key: * (The REGA reports of each of the analysis were checked and all subtyping and recombination patterns are with bootstrap support - >70 percent), and / (indicate possible recombination events).

From the analysis of the four isolates with the two different subtyping and recombination tools one can conclude that: R84 is subtype B isolate, TV239 is a AC recombinant virus (with the *gag-pol* fragment belonging to subtype A1 and the *env-nef* fragment showing signs of AC recombination breaking repeatedly throughout the fragment). Similarly, TV314 was classified as a subtype A1 isolate and TV412 is an AD recombinant virus (also breaking repeatedly throughout the fragment) by both the subtyping methods.

Though both the REGA and the jpHMM analysis were able to correctly identify the four isolates, the jpHMM subtyping tool was found to be much more accurate due to the phylogenetic approach used by the program.

3.6 Construction of NFLG phylogenetic trees

Three NJ-trees (one containing the three sequences of the NFLG fragments and two for each of the TV239 fragments) were constructed with the use of MEGA v 4.1. Data sets were also analyzed with PAUP v 4.0b10 and PhyML v 3.0 and maximum likelihood trees were constructed with the use of these programs as well as the HKY and GTR models of nucleotide substitution (data not shown).

Three NJ-trees were constructed with the use of MEGA v 4.1. The Neighbor-Joining tree of the three NFLG's containing samples R84, TV314 and TV412 as well as several reference strains from the LANL database can be seen in Figure 3.6. In the tree (Figure 3.6) two sample TV314 and R84 cluster within subtype clusters. R84 clustered with other subtype B viruses and was most closely related to `B.FR.83.HXB2 LAI IIIB BRU.K034` with a strong bootstrap support of 94%. TV314 clustered with A1 sequences and was more closely related to the reference strain `A1.UG.92.92UG037.AB253429`, than to any of the other samples in the cluster, but with a very low bootstrap support of only 42%. Sample TV412 was an outlier of the A1 cluster which possibly indicates viral recombination between subtype A1 and another HIV-1 subtype. The outgroup, `N.CM.95.YBF30.AJ006022`, rooted the tree.

Figures 3.7 and 3.8 shows the NJ-trees of sample TV239's two fragments (*gag-pol* and *env-nef*) respectively.

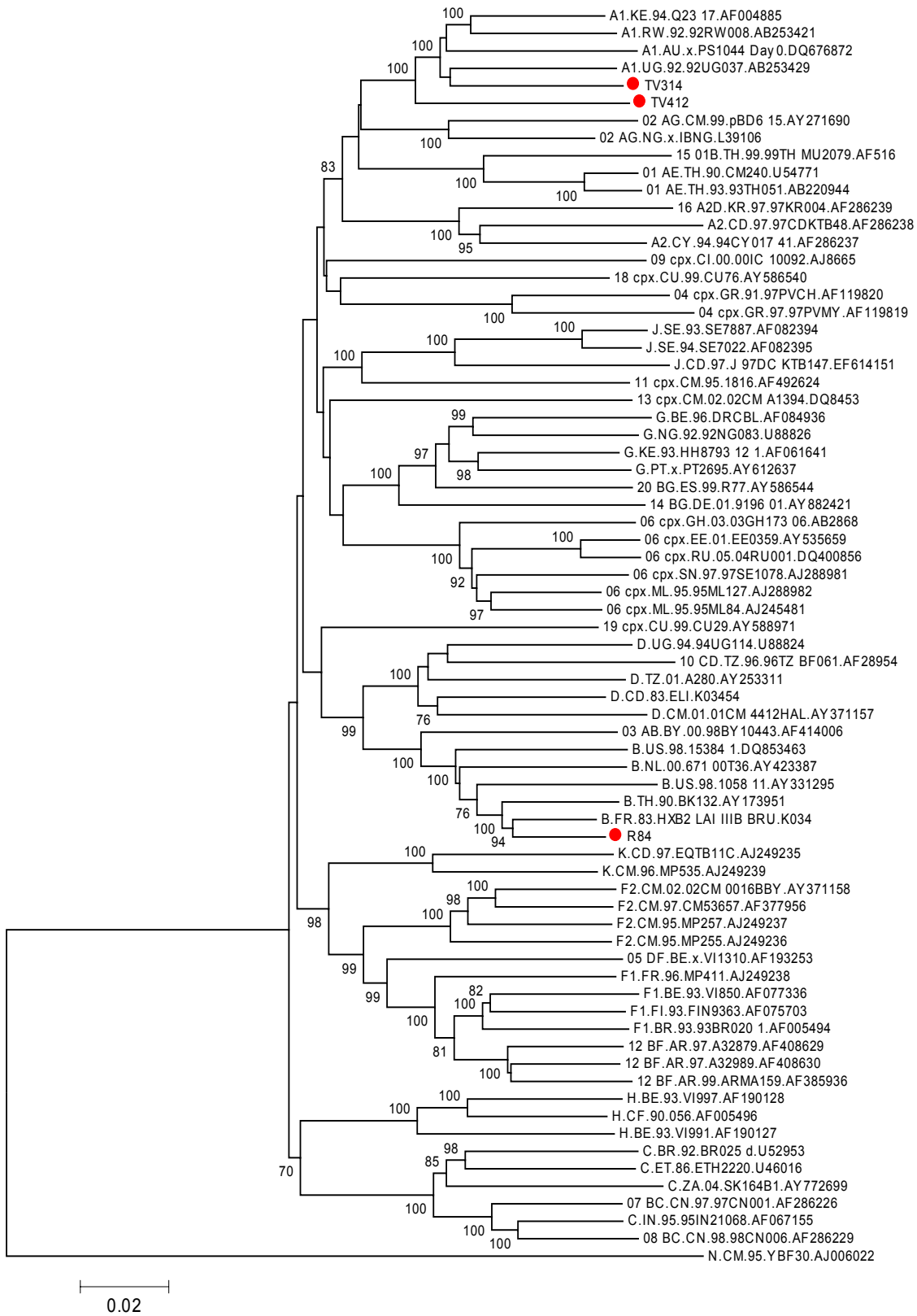


Figure 3.6: A Neighbor-joining tree containing the NFLG sequences of the three TV samples and reference samples. Each TV sequences are indicated with a red dot. The genetic distance is shown in the bottom line and corresponds to the length of the branches. Bootstrap values greater than 70 percent are shown.

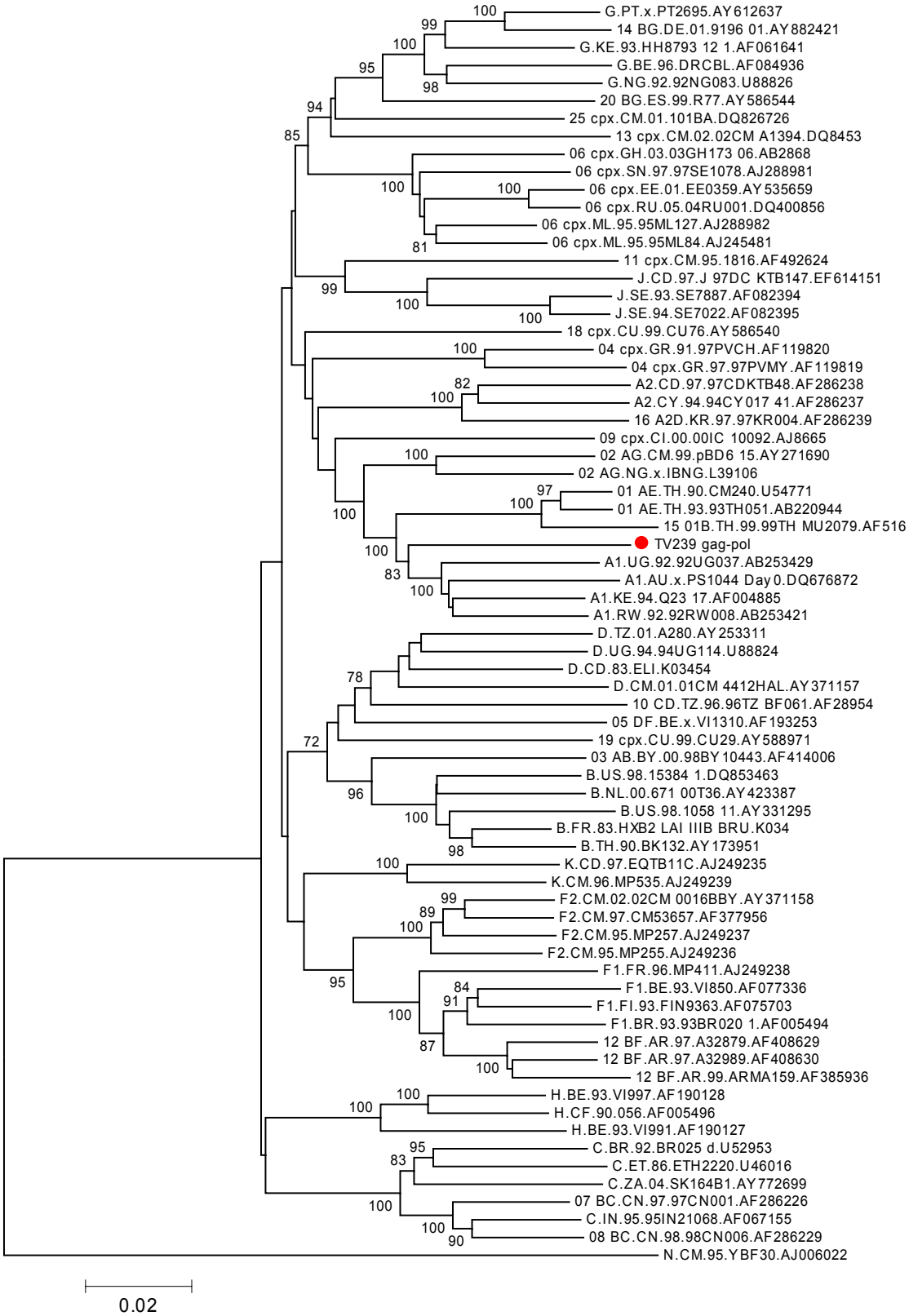


Figure 3.7: A Neighbor-joining tree containing reference sequences and the sequence of the *gag-pol* fragment of TV239. The TV sequence is marked with a red dot. The genetic distance, which corresponds to the branch lengths, is shown at the bottom. Bootstrap values greater than 70 percent are shown.

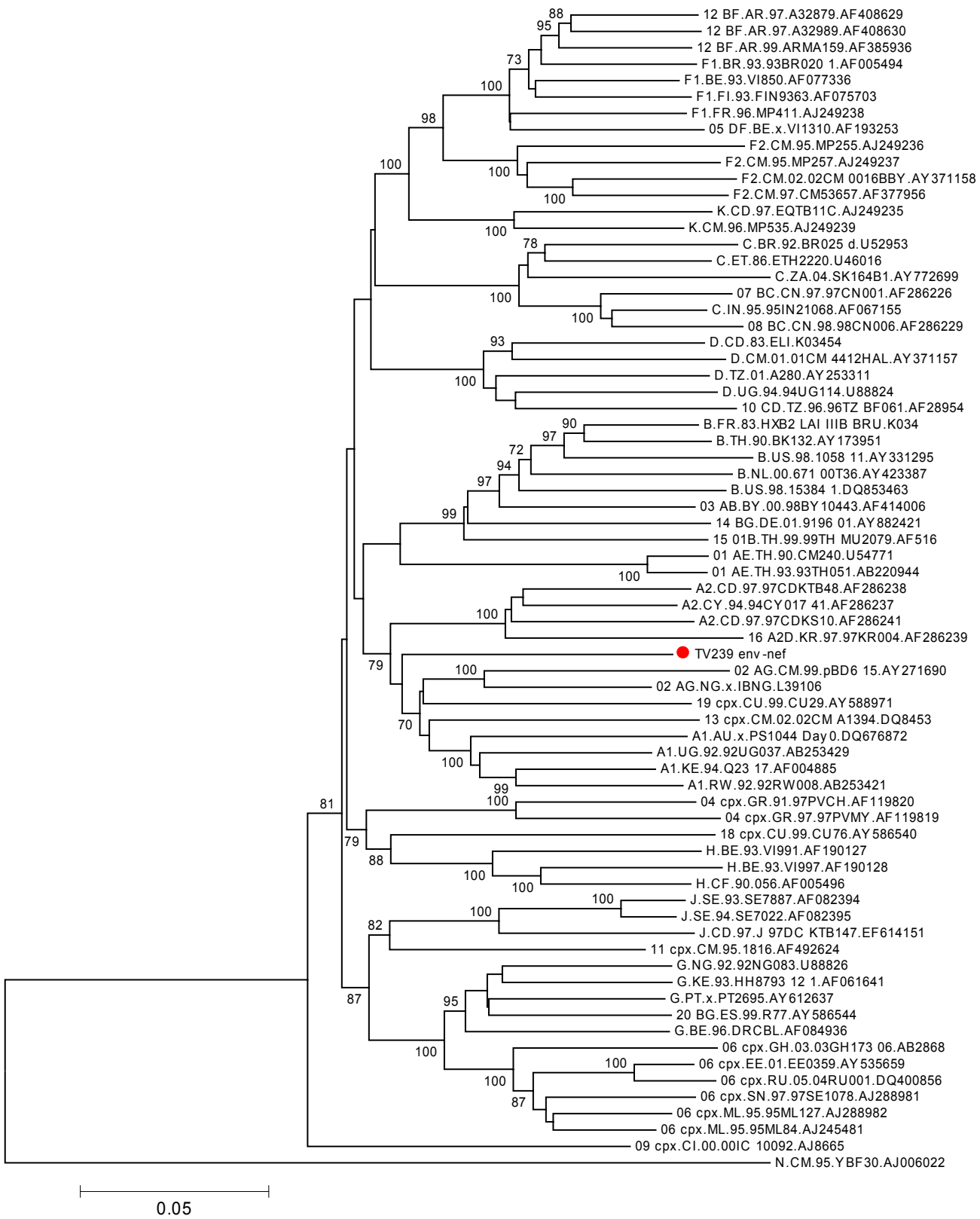


Figure 3.8: A Neighbor-joining tree containing reference sequences and the sequence of the *env-nef* fragment of TV239. The TV sequence is marked with a red dot. The genetic distance, which corresponds to the branch lengths, is shown at the bottom. Bootstrap values greater than 70 percent are shown.

Inspection of these trees revealed that out of the 71 taxa in the trees: the *gag-pol* fragment clearly clustered with other subtype A1 isolates in the tree with a strong bootstrap support of 83% and that the *env-nef* fragment was an outlier with a bootstrap support of 53%, containing A1 isolates as well as CRF02, cpx 19 and cpx 13 reference strains. CRF13_cpx is a complex recombinant form containing viral segments from subtypes A, G, and J as well as fragments of CRF01_AE. The CRF19_cpx sequence on the other hand contains viral segments from subtypes A1, D and G. From the analysis of the two fragments of sample TV239 it is clear that this isolate can be considered as an A1 recombinant virus, with the *gag-pol* fragment as a subtype A1 and the *env-nef* fragment as the breakpoint of viral recombination.

3.7 Results of recombination identification with RIP and Simplot

The analysis of the four samples revealed recombination within two of the isolates genomes. TV412 and TV239 were identified by both the REGA and jpHMM subtype and recombination identification programs as an AD and AC recombinant respectively.

Apart from the REGA and jpHMM analysis the near-full length sequences were also analyzed with the use of two widely used recombination identification programs to observe potential recombination events: RIP [Siepel *et al,* 1995] and Simplot. For the RIP analysis, the consensus alignment of the LANL database [http://www.hiv.lanl.gov/] was used with a window size of 300 bp to obtain the results.

Only two of the four isolates (TV239 and TV412) showed signs of viral recombination. Similarity plots for these two isolates were downloaded and are presented in Figures 3.9 and 3.10. The similarity plot analysis in RIP of samples R84 and TV314 indicated that: R84 had a high similarity with other subtype B viral subtypes and that TV314 had a high similarity with other subtype A1 isolates in the LANL reference alignment. Similarly the analysis of the *gag-pol* fragment of TV239 also showed a high similarity with other A1 subtypes in the LANL reference alignment. Only the *env-nef* fragment of

sample TV239 (Figure 3.10) and the fragment of sample TV412 (Figure 3.11) showed signs of viral recombination.

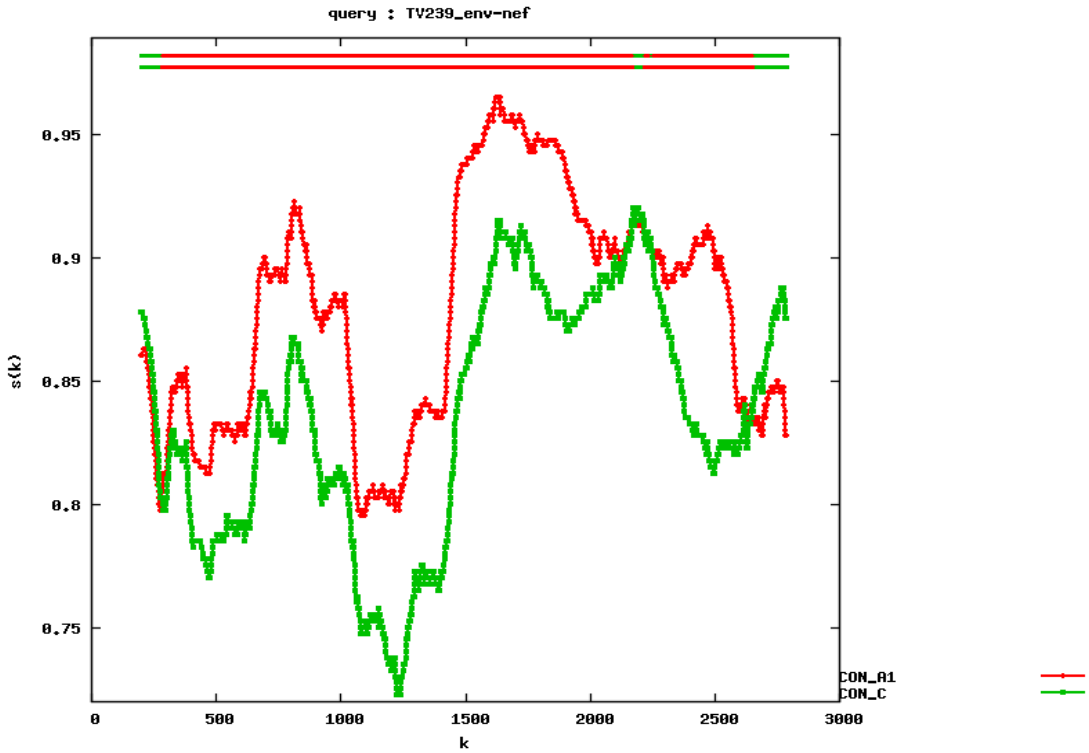


Figure 3.9: The analysis of the TV239 *env-nef* fragment with RIP. The s and the k symbols on the vertical and horizontal axis represent similarity and distance (in nucleotides or base pairs) respectively.

The RIP similarity analysis of the *env-nef* fragment of TV239 indicates AC recombination within the fragment, with multiple breakpoints. The analysis of TV412 also indicated recombination within the fragment with the genome breaking between subtype A and D viruses throughout the fragment. The results of the RIP analysis roughly corresponds to the breakpoints and recombination pattern that was obtained from the jpHMM analysis.

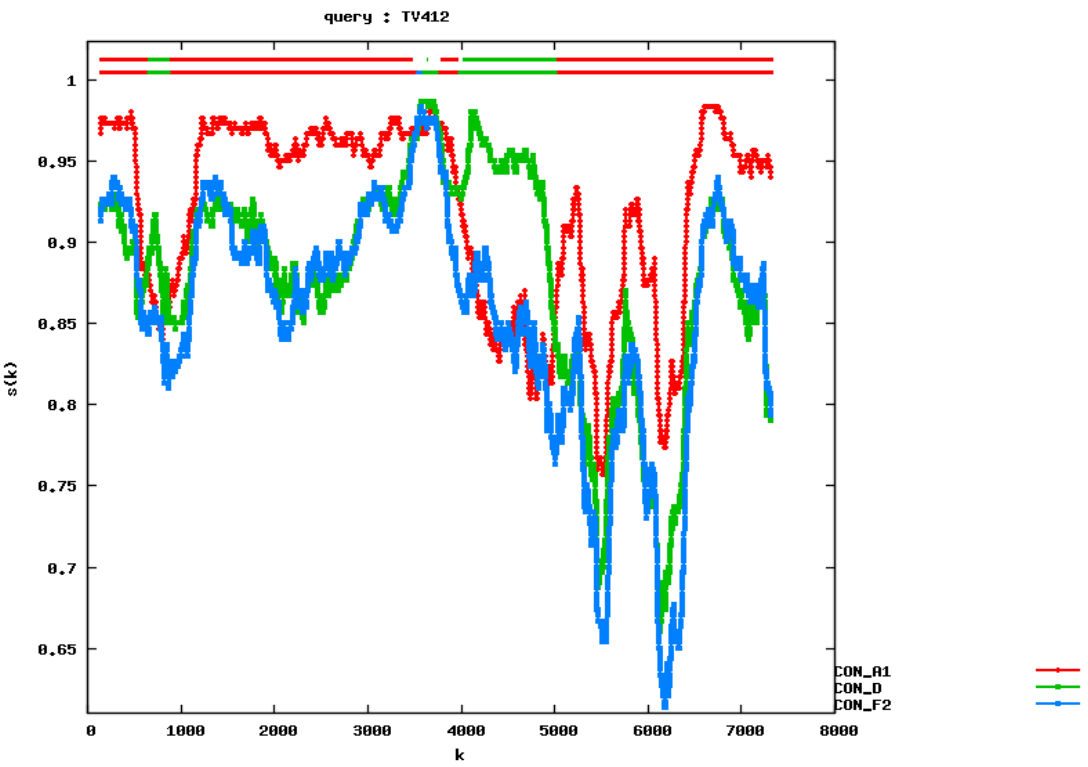


Figure 3.10: The analysis of TV412 with RIP. The s and the k symbols on the vertical and horizontal axis represent similarity and distance (in nucleotides or base pairs) respectively.

The four samples were also analyzed with Simplot v 3.5 [Lole *et al,* 1999; Salminen *et al,* 1995]. As with the RIP analysis samples R84, TV314 did not show signs of viral recombination and had high similarities with subtype B and A1 isolates respectively. The analysis of the TV239 fragments showed that the *gag-pol* fragment was a subtype A1 isolate whereas the *env-nef* (Figure 3.11) fragment showed signs of AC recombination. The analysis of TV412 (Figure 3.12) indicated viral recombination, between subtypes A and D, within the fragment.

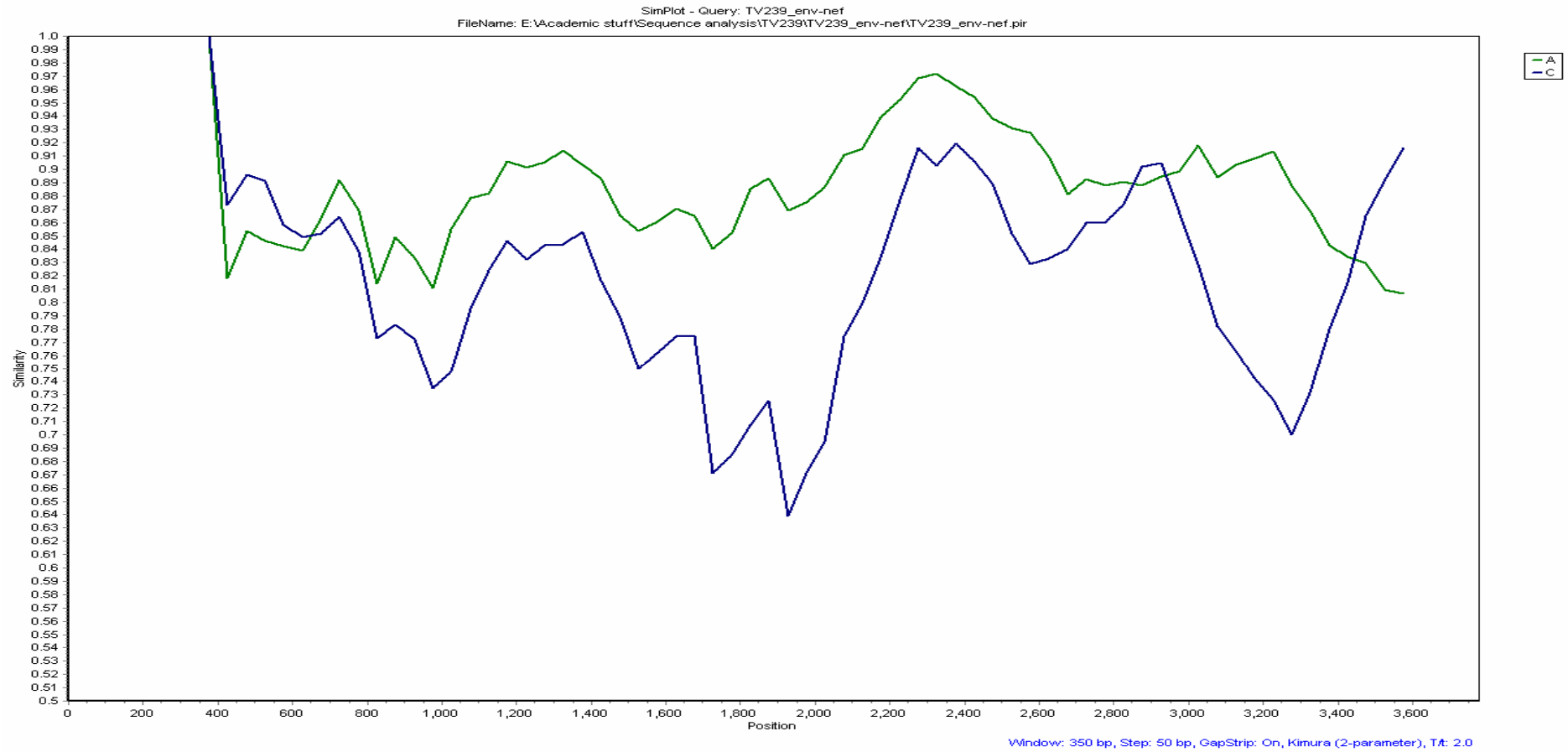


Figure 3.11: Simplot of TV239 *env-nef* fragment. The horizontal axis represents the position in the fragment and the vertical axis the similarity with viral subtypes against which the sequence of interest (TV239 *env-nef*) was queried. A window size of 350 bp and a step size of 50 bp were used. The Kimura 2-parameter of nucleotide substitution was used for the analysis.

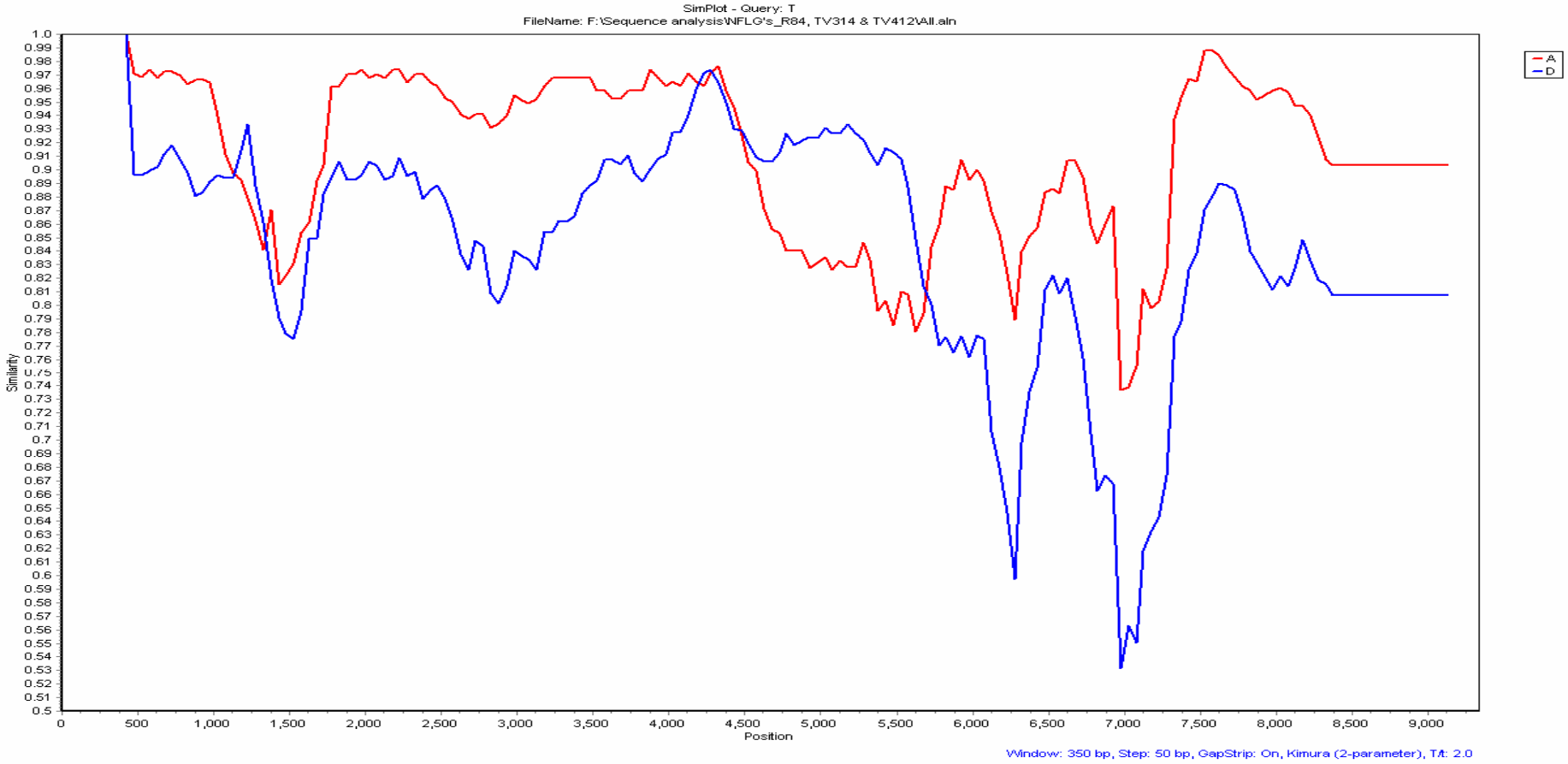


Figure 3.12: Simplot of TV412. The horizontal axis represents the position in the fragment and the vertical axis the similarity with viral subtypes against which the sequence of interest (TV412) was queried. A window size of 350 bp and a step size of 50 bp were used. The Kimura 2-parameter of nucleotide substitution was used for the analysis.

3.7.1 Phylogenetic analysis of recombinant breakpoints

NJ-trees were also drawn of each of the recombinant segments corresponding to the jpHMM breakpoint coordinates (obtained from the subtyping analysis). The breakpoint coordinates of the jpHMM were used, instead of the REGA breakpoints, due to the higher accuracy of the jpHMM analysis ability to calculate breakpoints. The jpHMM uses a jumping approach were the queried sequence can jump between samples in a multiple alignment as a sliding window moves across the fragment, which makes the identification of breakpoints to within a few base pairs much easier. The jpHMM breakpoints of samples TV239 and TV412 are summarized in Table 3.6.

Table 3.6: Breakpoint coordinates of the four samples as was determined by jpHMM analysis.

Sample	Coordinates (Relative to HXB2)	Sample coordinates	Subtype
TV239	6195 - 6335	1 - 140	C
	6336 - 8246	141 - 2050	A1
	8247 - 8532	2051 - 2337	C
	8533 - 8846	2338 - 2651	A1
	8847 - 9146	2652 - 2951	C
TV412	1264 - 1858	1 - 594	A1
	1859 - 2095	595 - 831	D
	2096 - 5217	832 - 3953	A1
	5218 - 6130	3954 - 4866	D
	6131 - 8254	4867 - 6990	A1

A schematic diagram was constructed for each of the recombinant fragments and each recombinant segment in a diagram was assigned a roman numeral (e.g. I, II, and III). See Figure 3.13 A and 3.13 B for the recombinant diagram and trees for the TV239 *env-nef* fragment. A total of a 100 bootstrap replicates

was performed on each of the trees. Only bootstrap values greater than 70 percent was included on the trees.

The NJ-tree analysis of the TV239 *env-nef* fragment revealed that there are five recombinant breakpoints within the sequence/fragment which clusters two different subtypes. The first half of the fragment (6195 – 6335) clustered with subtype C viruses. The second part (6336-8246) with subtype A1 and the third part with (8247-8532) subtype C again. The fourth (8533-8846) and fifth (8847-9146) segments similarly clustered with subtype A1 and C respectively.

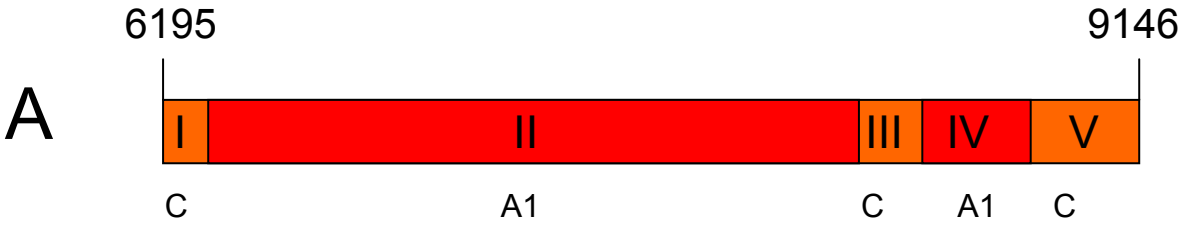


Figure 3.13A: A schematic diagram of viral recombination within the TV 239 *env-nef* fragment. The Roman numerals correspond with the trees in Figure 3.13B. The coordinates on the diagram corresponds with HXB2 coordinates.

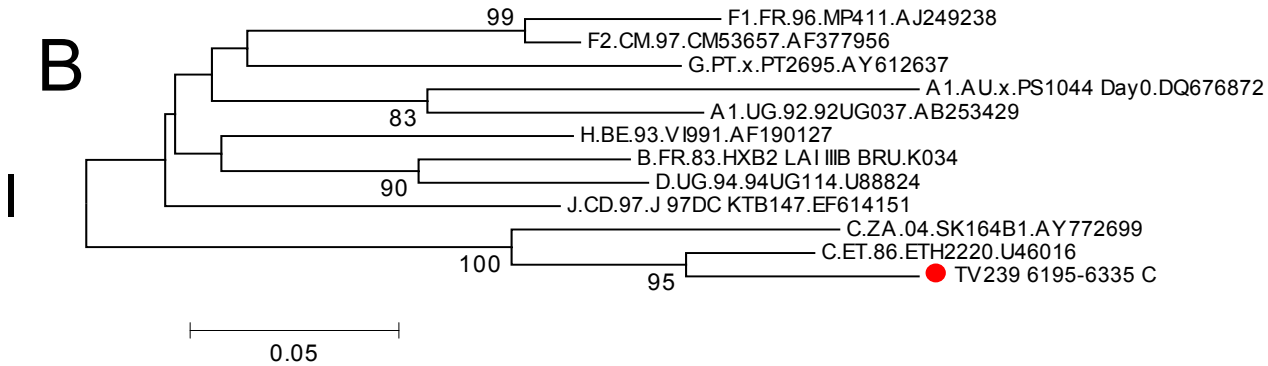


Figure 3.13B: NJ-trees of the different viral recombinant regions of the TV 239 *env-nef* fragment. The Roman numeral in front of each of the trees corresponds with the numerals in the schematic diagram in Figure 3.13A. The sequence of interest is marked with a red dot. The genetic distance, which corresponds to the length of the branches, is shown in the bottom left hand corner. Only bootstrap values greater than 70 percent are shown on the trees. Figure 3.13B (II - V) continues on page 90.

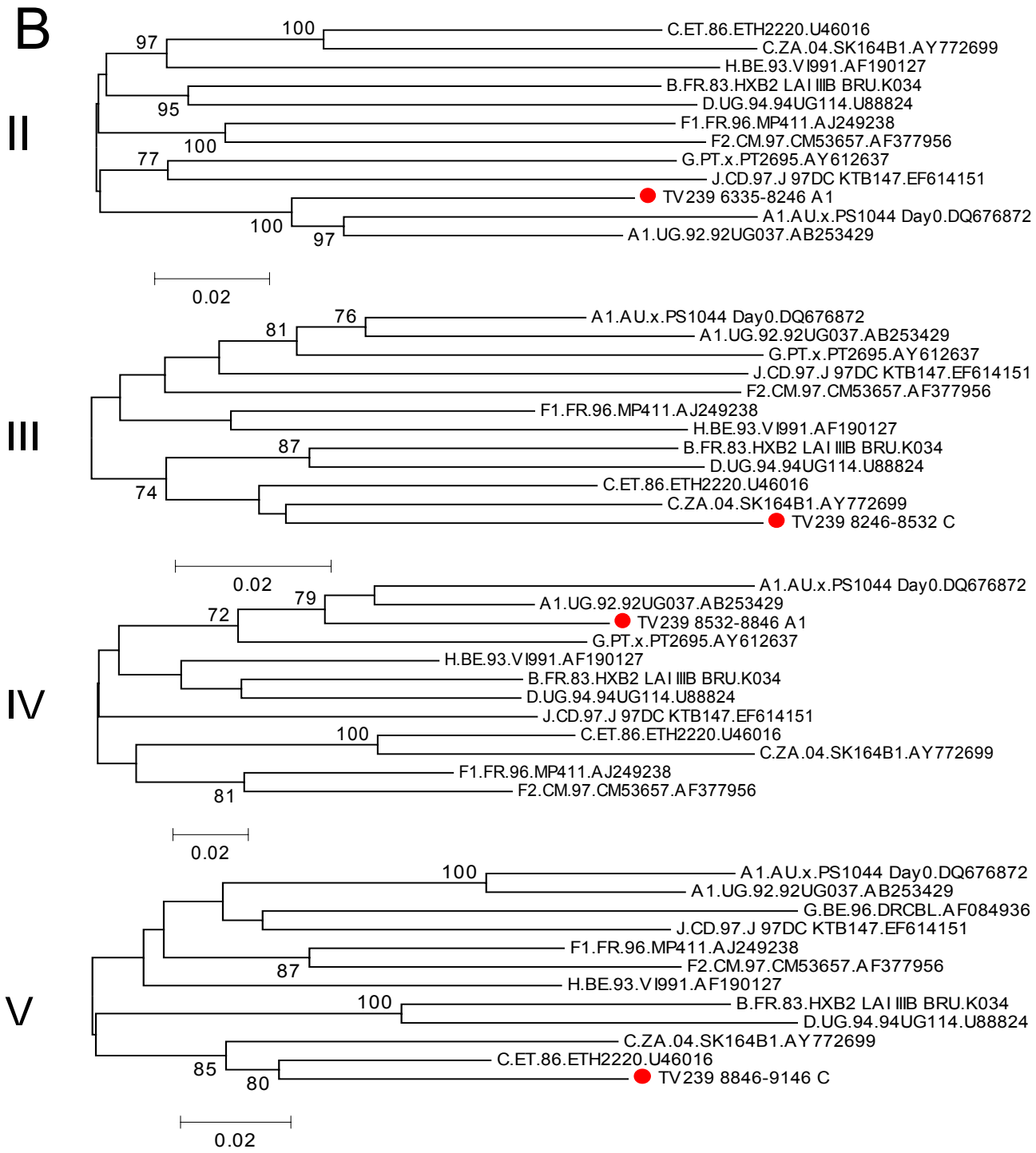


Figure 3.13B (continued): NJ-trees of the different viral recombinant regions of the TV 239 *env-nef* fragment. The Roman numeral in front of each of the trees corresponds with the numerals in the schematic diagram in Figure 3.13A. The sequences of interest are marked with a red dot. The genetic distance, which corresponds to the length of the branches, is shown in the bottom left hand corner. Only bootstrap values greater than 70 percent are shown on the trees. Figure continued from page 89.

The results of the viral recombination pattern of TV412 are demonstrated in a similar manner (Figures 3.14A and Figure 3.14B). Similar analysis of the TV412 NJ-trees revealed that this fragment contained five viral segments within the sequenced fragment, which corresponds with two different subtypes. The first (1246 - 1858), third (2096 - 5217) and fifth (6131 - 8254) segments of TV412 clearly clustered with subtype A1 samples in the NJ-trees. The second (1859 - 2095) and fourth (5218 - 6130) segments on the other hand clustered with subtype D viruses in the tree, in particular with `D.UG94.94UG114.U88824` from Uganda.

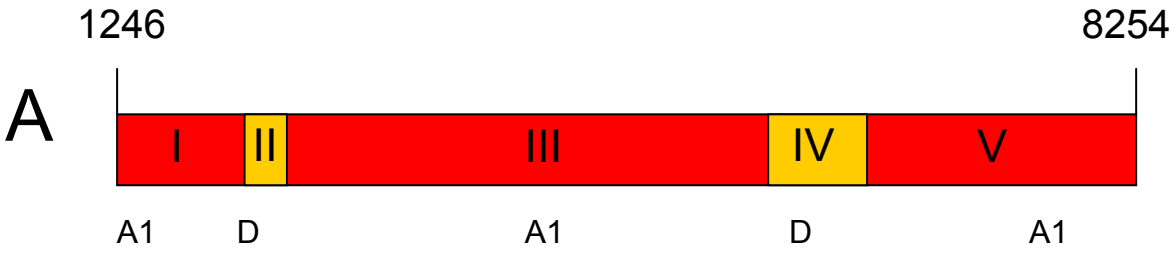


Figure 3.14A: A schematic diagram of viral recombination within sample TV412. The Roman numerals correspond with the trees in Figure 3.15B. The coordinates on the schematic corresponds with that of HXB2.

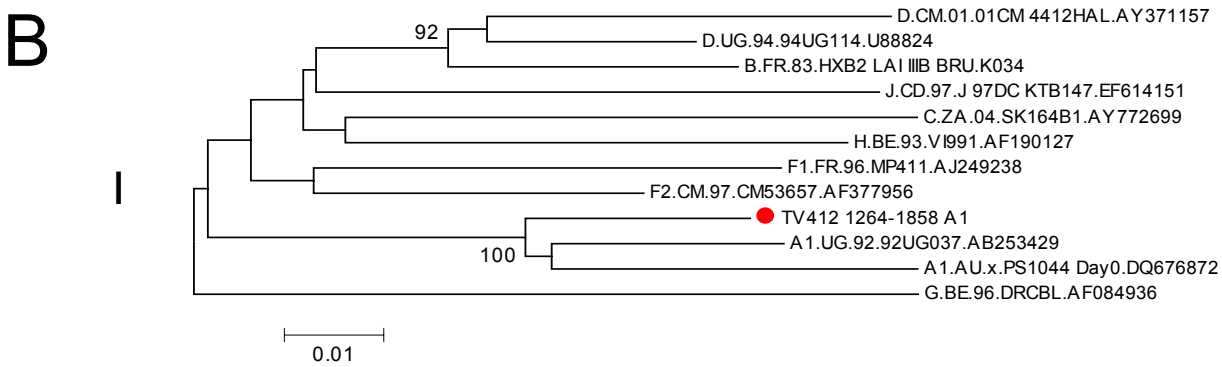


Figure 3.14B: NJ-trees of the different viral recombinant regions of sample TV412. Each Roman numeral corresponds with the numerals in the schematic diagram in Figure 3.14A. Sequences of interest are indicated with a red dot. Only bootstrap values greater than 70 percent are shown on the trees. The genetic distance, corresponding to the branch lengths are shown in the bottom left hand corner. Figure continues on page 92.

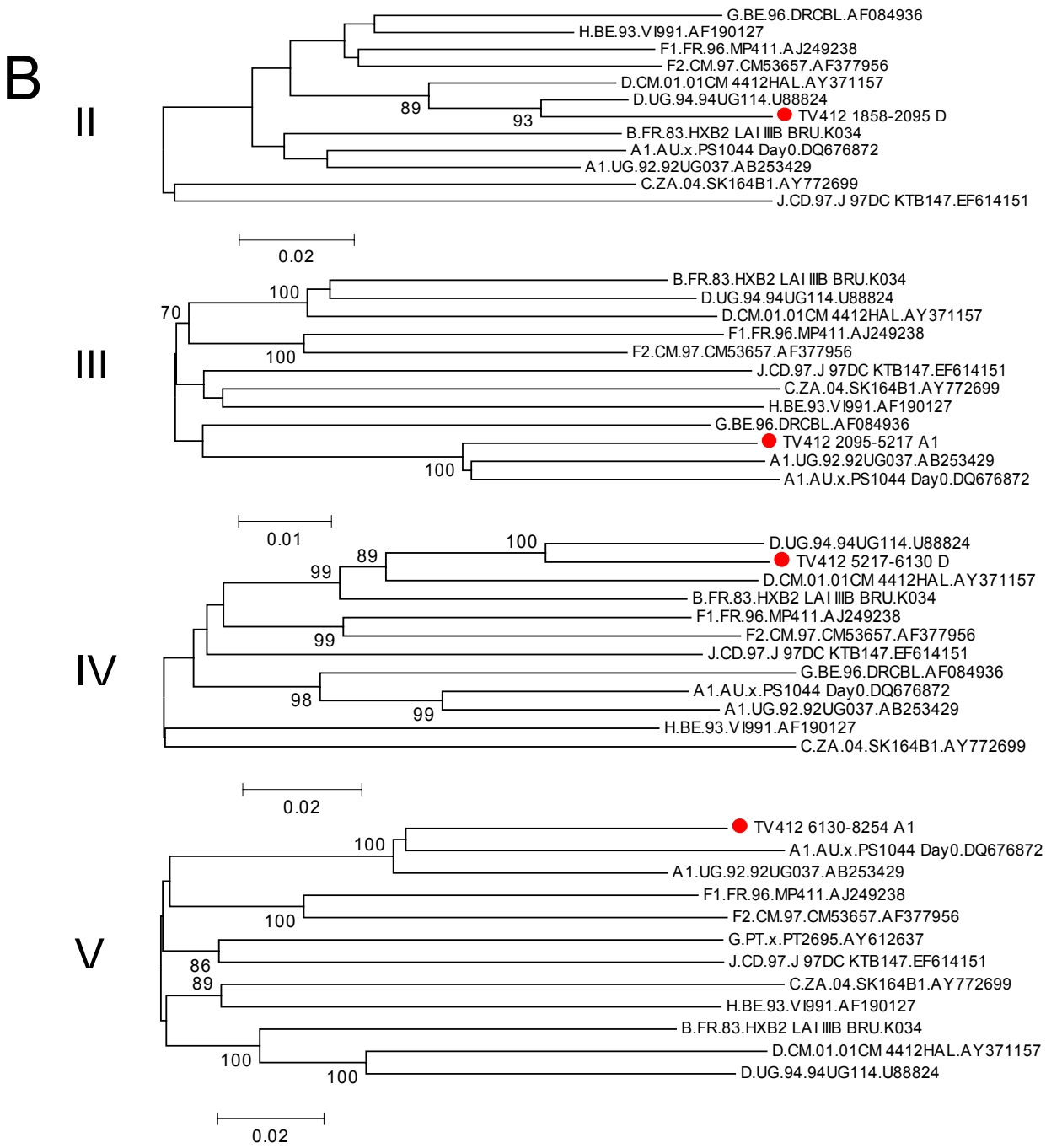


Figure 3.14B continued: NJ-trees of the different viral recombinant regions of sample TV412. Each Roman numeral corresponds with the numerals in the schematic diagram in Figure 3.14A. Sequences of interest are indicated with a red dot. Only bootstrap values greater than 70 percent are shown on the trees. The genetic distance, corresponding to the branch lengths are shown in the bottom left hand corner. Figure continued from page 91.

3.8 Analysis of non-recombinant isolates with reference sequences

The two non-recombinant viral isolates (R84 and TV314) were compared with reference sequences of the same subtype(s). Sample R84 was compared with other subtype B viruses, mostly HIV-1 subtype B isolates from the 1980's and early 1990's. R 84 was aligned with these reference sequences, which was obtained from the LANL database, and a Neighbor-Joining tree (Figure 3.15) was constructed in MEGA v 4.1. TV 314 was compared to other subtype A1 isolates from the LANL database in a similar fashion (Figure 3.16).

Analysis of the phylogenetic analysis of the pure viral fragments revealed that R84 clustered with other subtype B isolates from the United States from the mid 1980's, in particular with sample B.US.83.5082 83, but with a vary low bootstrap support of 50 percent. This sample was isolated in the mid 1980's an thus are suspected to be more closely related to other subtype B viruses from the same period.

The tree with TV314 contains two major clusters, one containing samples of African origin as well as A1 sequences from Sweden and another containing A1 samples that was sampled outside of Africa (mostly in the former Soviet countries or Australia). TV314 clustered with A1 sequences, of African origin, form the mid 1990's till the year 2000. From the analysis it appears that TV314 is more closely related to sample A1.SE.95.UGSE8131.AF107771, with a very high bootstrap support of 87%. This sample was isolated from a patient in Sweden, but clearly clusters amongst other A1 sequences from East African origin, which suggests the introduction and spread of East African A1 isolates into Northern Europe in the distant past. All these assumptions were made basted on the raw phylogenetic data that was obtained from the analysis and more in-depth phylogenetic analysis of TV314 will be needed to verify these claims.

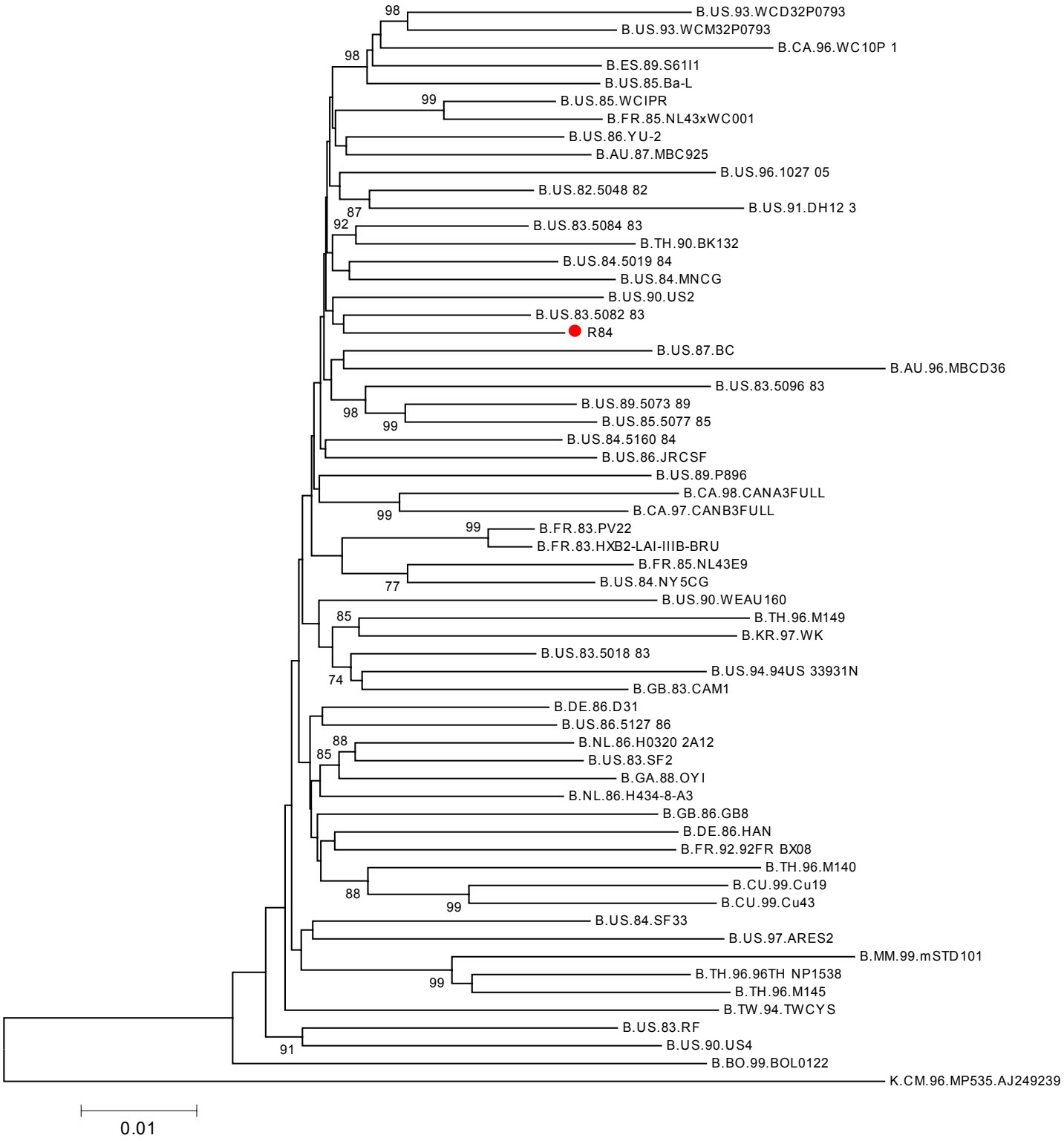


Figure 3.15: A NJ tree exploring the relationship of subtype B HIV-1 isolates with sample R84. The TV sequence is marked with a red dot. The genetic distance, which corresponds to the branch lengths, is shown at the bottom. Bootstrap values greater than 70 percent are shown.

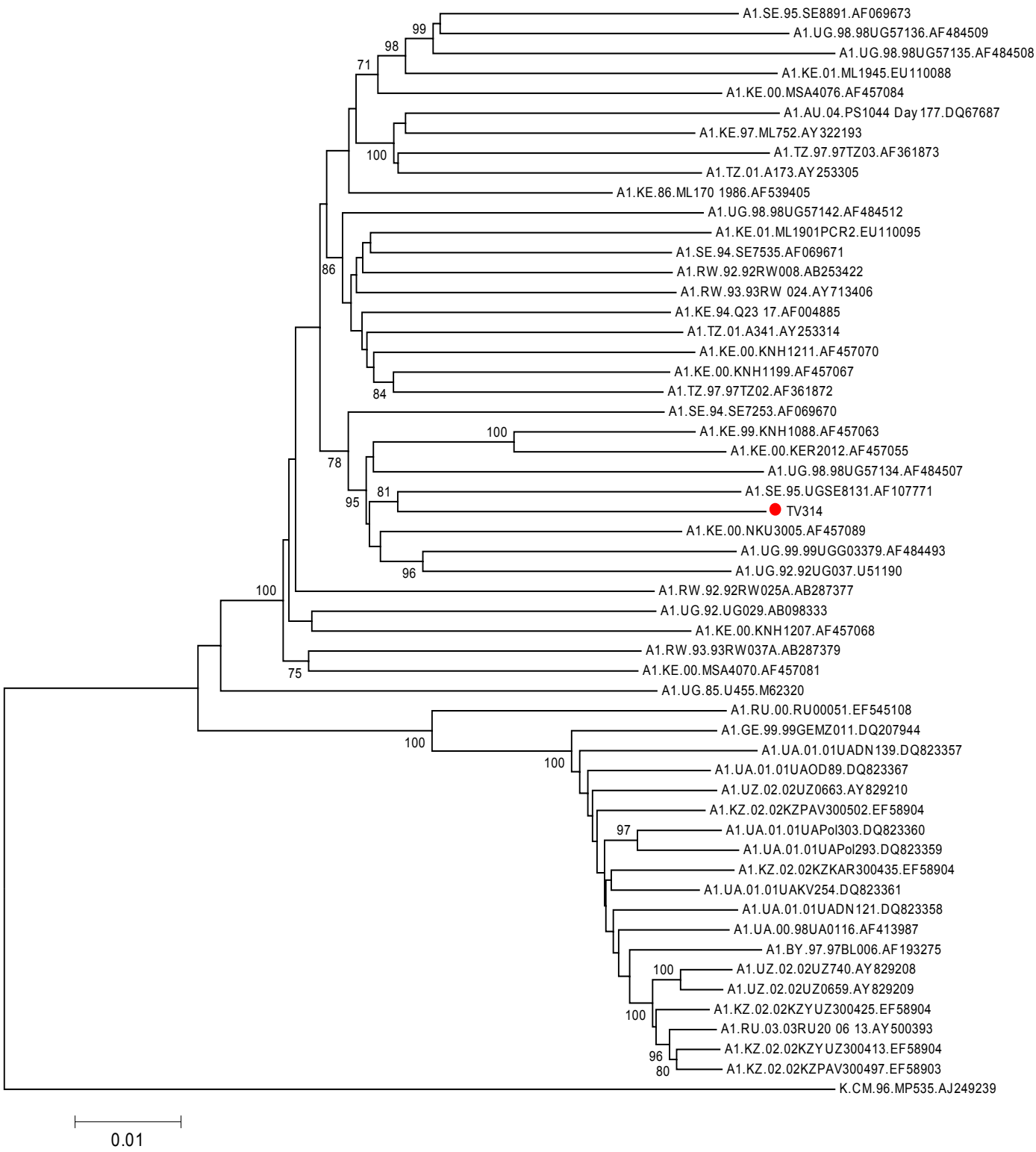


Figure 3.16: A Neighbor-joining tree looking at the relationship between subtypes A1 HIV-1 isolates and sample TV314. The TV sequence is marked with a red dot. The genetic distance, which corresponds to the branch lengths, is shown at the bottom. Bootstrap values greater than 70 percent are shown.

CHAPTER FOUR – DISCUSSION AND CONCLUSION

TABLE OF CONTENTS **Page**

CHAPTER FOUR – DISCUSSION AND CONCLUSION

DISCUSSION

The objective of this study was to characterize non-subtype C isolates, from the greater Cape Town Metropolitan area. This study indicated that non-subtype C HIV-1 isolates, including subtypes B, A1 and A1 recombinant viruses, are circulating within the Cape Town area.

4.1 HIV-1 and HIV characterization in South Africa

The first documented cases of HIV-1 infection in South Africa occurred in 1982 [Ras *et al,* 1983]. Unlike the rest of sub-Sahara Africa, the HIV-1 epidemic in South Africa was mostly associated with the homosexual population [Sher, 1989]. The isolates were later characterized as subtype B and D viruses [Becker *et al,* 1995; Engelbrecht *et al,* 1995]. Subtype B and D viruses were probably introduced into the country by homosexual men through international travel. By the start of the 1990's a new epidemic of HIV started to occur within the country amongst heterosexual individuals, which was largely associated with members of the indigenous black population [Williamson *et al,* 1995]. Viruses that were isolated from heterosexual individuals of the second epidemic mostly belonged to HIV-1 subtype C [van Harmelen *et al,* 1997]. The second epidemic quickly overtook the first initial epidemic in the early 1990's and today the heterosexually transmitted dominates the HIV epidemic within the country [McCutchan *et al,* 1996], with as many as 6.2 million people infected [UNAIDS, 2008].

To date several papers on a wide range of HIV-1 subtypes have been published within South Africa on subgenomic fragments or near full-length sequences (Tables 1.2 and 1.3). These viruses were isolated from several geographical locations within the country. The following section will take a look at the publications.

Subtype C HIV-1 accounts for nearly 95% of all infections within the country and subtype C viruses have been characterized on several occasions. Of the 3884 South African full or partial sequences in the LANL database, 3706 or 95.4% are subtype C (Figure 1.8). One should note that these figures may be misleading as several sequences may be from the same patient sample. The rest of the sequences represents subtype B and D isolates, which has been described in the past in association with homosexual mode of transmission [Engelbrecht *et al,* 1994; Becker *et al,* 1995], and other subtypes and recombinant forms.

Two hundred and sixty five full-length HIV-1 sequences from 8 independent studies have been described within the country. The majority of these sequences represent subtype C isolates [van Harmelen *et al,* 2001; Papathanasopoulos *et al,* 2002; zur Megede *et al,* 2002; Papathanasopoulos *et al,* 2003; Hunt *et al,* 2003; Rousseau *et al,* 2006]. Only 11 full-length sequences of non-subtype C isolates have been described to date within South Africa. These include 5 subtype D isolates [Loxton *et al,* 2005; Jacobs *et al,* 2007], one subtype B [Rousseau *et al,* 2006], one subtype A1 [Rousseau *et al,* 2006], three AC recombinants [Papathanasopoulos *et al,* 2002; Rousseau *et al,* 2006] and one complex viral form containing subtypes A, C, D, G and K segments [Papathanasopoulos *et al,* 2002].

As with the characterization of full-length genomes from South Africa, most partial sequence fragments also represents subtype C isolates [Bredell *et al,* 1998; Engelbrecht *et al,* 2001; Scriba *et al,* 2002; Gordon *et al,* 2003; Bessong *et al,* 2005; Bell *et al,* 2007]. Several non subtype C viruses, apart from the subtype B and D viruses that was mentioned earlier [Engelbrecht *et al,* 1994; Becker *et al,* 1995], have been identified. These include, an CRF01_AE isolate [van Harmelen *et al,* 1997], subtype A isolates [Jacobs et al, submitted; Bredell *et al,* 2002], a G and one CRF02_AG isolate [Bredell *et al,* 2002; Jacobs et al, 2008], CD recombinants [Bredell et al, 2002; Gordon *et al,* 2003], an F1 isolate [Jacobs *et al,* submitted], CA recombinants [Bredell *et al,* 2002] and a CH recombinant [Jacobs *et al,* submitted]. Some of these isolates are extremely rare, such as F1, CH recombinants and subtype G, and

are mostly confined to areas of Central Africa e.g. the DRC (Figure 1.8) [Geretti, 2006].

4.2 Full genome characterization of HIV

HIV sequences can be grouped into one of the 9 subtype by either characterization of small gene fragments throughout the genome or by characterizing the full HIV genome. The use of the first method has been well established over the years [Swanson *et al,* 2003]. This method does however have its own drawbacks as some recombination events are so small and can be missed in such an approach. With these limitations of partial genome sequencing the focus was shifted to the characterization of full-length HIV-1 genomes. Improvements in nucleic acid amplification technology, which permits the amplification of large gene fragments of (9 – 36 kbp) now enables researchers to amplify and characterize full-genomes of HIV much easier. This procedure was first described in HIV-1 research by Mika Salminen and co-workers in 1995 [Salminen *et al,* 1995a]. Since then the value of full-genome sequencing of HIV has been repeatedly demonstrated. The subtype E viruses, which were isolated and sequenced in the early 1990 in Southeast Asia, were initially termed as subtype E based on sequences from the *env* gene [Carr *et al,* 1996; Gao *et al,* 1996]. Full genome characterization of these isolates reveled that the isolates were recombinant in nature with subtype A in the *gag* and subtype E in the *env* regions. The Z321 strain from Zaire, which was one of the earliest identified strains HIV-1, was initially thought to have been subtype A based on *env* sequences, but with full genome amplification it was found to be an AG recombinant [Choi *et al,* 1997]. This emphasizes the importance of full genome amplification and sequencing of HIV isolates, especially when viral recombination is suspected.

Full genome characterization of HIV-1 can either be performed on, proviral DNA or RNA (isolated from PBMC) or viral RNA from blood plasma. In the HIV database of LANL there are only 1404 HIV-1 sequences which are greater than 8 kbp in length (any continues HIV fragment larger than 8000 bp is considered to be a near full-length genome) [http://www.hiv.lanl.gov/]. The

majority of these near full-length sequences were generated by using proviral DNA from either peripheral blood mononuclear cells (PBMC) or either cultured cells. A standard protocol for the characterization of near full-length sequences from viral RNA was recently developed [Nadai *et al,* 2008]. Viral RNA represents the current replicating viral population within a patient and thus may offer more pathogenic and phylogenetic significance than with DNA isolates from PBMC's [Wei *et al,* 1995].

4.3 Success of the various amplification assays and sequencing reactions

The amplification of the near-full length genomes were all unsuccessful and non specific products were obtained. This might be due to un-optimized PCR assays, old DNA or the use of unspecific primer pairs.

All of the other PCR's that were performed in this study were successful, except for the three subgenomic fragments of TV546, the *env* fragments of TV480 and the LTR-*gag* amplification assays of TV239, TV314 and TV412. PCR failures might be due to the high degree of sequence variation in HIV-1. Subtype specific primers could have been design but due to sample limitations was abandoned. TV480 and TV546 have been identified as possible CJ and CG recombinant viruses in a previous study [Jacobs *et al,* submitted]. The primer sets that were used by Swanson and co-workers [Swanson *et al,* 2003] might not have been specific enough for these two samples. When doing PCR's it is also important to optimize each assay for greater yield in product.

The sequencing of the positive PCR products was all successful with the exception of TV340 *gag* p24 and TV239. Multiple peaks in the electrophenogram made it impossible to obtain a reliable sequence for TV340 *gag* p24. The failure of the sequencing of TV239 which produced a 600 bp gap in the *pol-vif* region of TV239 might be due to frequent recombination events within this region of the genome.

4.4 HIV subtype results of the present study

This section will discuss the HIV subtype results that were generated in this study. In the first section the results of the characterization of the various subgenomic regions will be discussed. The second section will discuss the results that were obtained in the near full-length characterization of four samples.

4.4.1 Discussion of the various subgenomic fragments

The characterization of small subgenomic regions throughout the HIV genome to subtype an isolate or identify possible recombinant has been widely used in the past [Swanson *et al,* 2003]. The method that was developed by Swanson and co-workers targeted specific regions that is important in HIV antibody/antigen diagnostic assays and commercially available viral load assays. Theses regions are generally highly conserved and the PCR's that was developed for this study was able to detect most HIV-1 groups and subtypes.

In the present study subtype C, F1, A1, and B isolates were identified with this method of genome characterization. Several recombinant HIV viruses were also identified, including AD, CA, AG and other possible A1 recombinant forms. Sample TV546 was not amplifiable with the primer sets that were used. This sample was identified as a subtype G or subtype C isolate on previous occasion [Jacobs *et al,* submitted]. This illustrates the need for subtype specific primers when working with rare viral subtypes or recombinants forms.

The V3 region of TV515 was characterized on a previous occasion [Jacobs *et al,* submitted]. These *gag* p24, *pol-integrase* and *env* gp41 sequences of TV515 will represent the second sequenced set of F1 sequences from South Africa, though from the same patient. Sample TV340 were previously classified [Jacobs *et al,* submitted] as a subtype A isolate. This sample was isolated from a patient in the Cape Town, which became infected in the DRC (Table 2.4). Through analysis of the *pol* and *env* subgenomic regions this

isolate was classified as an AG recombinant, with the *pol* fragment belonging to subtype G and the *env* fragment to subtype A1.

Similarly analysis of TV441 was identified as a subtype A1 isolate in the same study. The analysis of the present study indicated that this sample is a possible AC recombinant with the *gag* p24 and *pol-integrase* regions clustering with subtype C viruses and the *env* gp41 region with subtype A viruses. It might be concluded that subgenomic characterization of isolates may yield more accurate results when several regions throughout the genome of HIV-1 is targeted.

Though these methods of HIV genome characterization may give an indication of the viral subtype of a particular isolate or identify viral recombination, only full genome analysis of samples will allow one to make a safe and accurate assumption on the subtype and recombination of particular isolates, as recombination can occur in the regions between the three subgenomic fragments characterized.

4.4.2 Discussion of near full-length sequence results

4.4.2.1 R84 and other subtype B viruses in South Africa

Subtype B HIV are typically found amongst homosexual men in North America and Europe. Since the outbreak of the epidemic, subtype B has also established itself in South America, parts of Asia, South Africa, and in Australia [Leitner *et al*, 1996]. In South Africa subtype B, along with subtype D, viruses were isolated and later sequenced from homosexual men in the mid 1980's and early 1990's [Becker *et al,* 1995; Engelbrecht *et al,* 1995]. To date only a few cases have been reported where subtype B viruses were isolated from heterosexual individuals [van Harmelen *et al*, 1997; van Harmelen *et al*, 1999]. Heterosexual transmittion of subtype B has also been reported in other areas of the world [Lara *et al*, 1997; Hayman *et al*, 2001; Cleghorn *et al*, 2000]. Subtype B HIV has also been reported in MTCT cohorts from Cape Town [Jacobs *et al,* submitted]. Only one full-length sequence of a

South African subtype B virus, which was isolated from a heterosexual female, have been characterized to date [Rousseau *et al*, 2006].

A full-length subtype B sequence, which was isolated from a homosexual caucasian male in the mid 1980's, were characterized in this study. Amplification and sequencing of the isolate produced an 8913 bp fragment, which represents the second near full-length sequence of a subtype B isolate from South Africa. This subtype B isolate along with the subtype D viruses, which has been characterized in the past [Loxton *et al,* 2005; Jacobs *et al,* 2007], represent the only near full-length sequences of the first initial homosexual HIV epidemic within the country. This near full-length sequence will greatly improve our knowledge and understanding of the initial HIV epidemic within the country.

As was mentioned previously, heterosexual transmittion of subtype B have been reported in the past [van Harmelen *et al*, 1997; van Harmelen *et al*, 1999; Jacobs et al, submitted]. It would be reasonable to assume that if the circulation of subtype B viruses continues to increase and co-circulate with subtype C amongst the heterosexual population of South Africa that this could create a good opportunity for viral recombination to occur and the possible rise of CB recombinants within the region.

4.4.2.2 TV239 and other HIV recombinants from South Africa

Sample TV239 were isolated from a South African male, which presented with tuberculosis and a low CD4 cell count of 64. TV239 was characterized previously, through amplification and sequencing of a small (300 bp) fragment of the *env* gene, as a subtype A HIV-1 isolate. With the full genome analysis of this isolate it became apparent that the TV239 isolate was an AC recombinant. This only stresses the importance of full-genome characterization of HIV-1 isolates, as such recombination events can be missed with the amplification of smaller fragments.

Only three full-length sequences of AC recombinant viruses have been described in South Africa to date [Papathanasopoulos *et al,* 2002; Rousseau

et al, 2006]. Through phylogenetic analysis of the two AC recombinants that was described by Rousseau and colleagues revealed that the A fragments of the two recombinants were more closely related to sub-subtype A1. The one AC recombinant of Papathanasopoulos and co-workers was classified with the A recombinant fragments clustering with subtype A2 isolates as with A1 sub-subtypes of subtype A HIV-1 isolates. All the subtype A recombinant fragments as well as those of TV239 was more closely related to other subtype A sequences from East African. TV239 thus represents the fourth near full-length AC recombinant that has been described in South Africa to date, and only the third A1C recombinant.

As subtype A continue to spread within the region of Southern Africa and co-circulate with subtype C and other viral subtype one can expect and increase in the prevalence of these recombinant forms in the future.

4.4.2.3 Subtype A's in South Africa and TV314

TV314 was obtained from an asymptomatic South African male. Near full-length genome characterization revealed that the isolate was subtype A1 with no trace of recombination within the sequenced fragment. A total of 37 sequences of subtype A viruses from South Africa can be found in the LANL database, but only one full-length subtype A1 sample have been described to date [Rousseau *et al,* 2006]. The near full-length subtype A1 sequence that was described previously within the country was isolated form a heterosexually-infected female of Zulu or Xhosa ethnicity. The *gag* portion of this isolate was more closely related to the subtype A *gag* part that is characteristic of CRF01_AE isolates.

TV314 was more closely related to an A1 sequence that was characterized from Sweden, `A1.SE.95.UGSE8131.AF107771`, but these two sequences were grouped together with other subtype A1 sequences from East African origin. The dataset that was used for the analysis contained four sequences from Sweden, all of whom clustered amongst the other East African sequences. It may be concluded that this strain was introduced into Northern Europe via international travel or emigration.

4.4.2.4 TV412 and other AD recombinants from Kenya

Sample TV412 were isolated from an African male, which presented with chronic staphylococcal skin sepses. The patient's immune system was severely depleted with a CD4 count of 71. According to Hospital records the patient became infected with HIV in Kenya. Full-genome amplification was performed, but was insufficient for sequencing. Amplification was performed in four overlapping fragments, three of which were successful. This resulted in a 7008 bp sequenced fragment, stretching from 1246 – 8254 (relative to HXB2 coordinates). TV412 was classified, with the use of REGA and jpHMM subtyping, as a subtype A1 isolate based on the 300 bp V3 sequence that was characterized within the department [Jacobs *et al,* submitted]. In this study, through the characterization of a larger fragment (7 kbp) the sample classified as an AD recombinant. This stresses the importance of full-genome characterization of samples in order to make an accurate and informed conclusion concerning an isolates subtype status.

Subtype A and D HIV co-circulate in large numbers within the East African country of Kenya. Subtype A and D respectively represents 74.4 and 10.6 percent of the subtypes circulating within the country (Figure 1.8). This co-circulation of these two subtype, which are also found in large numbers in other East African countries (e.g. Uganda and Tanzania), presents a good opportunity for viral recombination. Today, AD recombinant viruses comprise nearly 1.9% of the viruses that are sampled. Similarly, Tanzania and Uganda have reported AD recombinant levels of 1.1 and 2.2 percent respectively (Figure 1.8). Recent data on co-infection of individuals with subtypes A and D showed that there are an ongoing generation and selection for A/D recombinant forms [Songok *et al,* 2004].

CONCLUSION

This work represents partial *gag*, *pol* and *env* and near-full length sequences of non-subtype C HIV-1 sequences from the Tygerberg Hospital, Cape Town, South Africa. Phylogenetic analysis of the sequenced data revealed the presence of subtype A, B, F1 as well as AC and AD recombinant viruses within the region. The two near full-length sequences of the recombinant viruses constitute two new unique recombinant HIV-1 forms. The data that was gathered in this study will greatly improve our knowledge of subtype distributions within the country. Due to the impact that HIV genetic diversity might have on vaccine design and development, as well as HIV diagnosis and the treatment of patients with antiretroviral therapeutic drug, ongoing research into the epidemiology and spread of HIV subtypes and recombinants within South Africa are needed.

CHAPTER FIVE - REFERENCE LIST

Abecasis A, Vandamme A-M, Lemey P. Sequence Alignment in HIV Computational Analysis pp. 2-16 in HIV Sequence Compendium 2006/2007. Edited by: Thomas Leitner T, Foley B, Hahn B, Marx P, McCutchan F, Mellors J, Wolinsky S, Korber B. Published by: Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM. LA-UR 07-4826. 2007

Achong G, Mansell PWA, Epstein MA, and Clifford P. An unusual virus in cultures from a human nasopharyngeal carcinoma. NJCI. 1971. 42: 299-307

Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ. Basic Local Alignment Search Tool. Journal of Molecular Biology. 1990. 215: 403-410

Ammann AJ, Abrams D, Conant M, Chadwin D, Cowan M, Volberding P, Lewis B, and Casavant C. Acquired immune dysfunction in homosexual men: immunologic profiles. Clinical Immunology and Immunopathology. 1983. 27: 315-325

Andersson S, Norrgren H, Dias F, Biberfeld G, and Albert J. Molecular characterization of human immunodeficiency virus (HIV)-1 and -2 in individuals from Guinea-Bissau with single or dual infections: predominance of a distinct HIV-1 subtype A/G recombinant in West Africa. Virology. 1999. 262: 312-320

Anderson JP, Rodrigo AG, Learn GH, Madan A, Delahuntry C, Coon M, Girard M, Osmanov S, Hood L, and Mullins JI. Testing the Hypothesis of a Recombinant Origin of Human Immunodeficiency Virus Type 1 Subtype E. Journal of Virology. 2000. 74: 10752-10765

Ariën KK, Vanham G, and Arts EJ. Is HIV-1 evolving to a less virulent form in Humans? Nature Reviews Microbiology. 2007. 5: 141-151

Bachmann MH, Delwart EL, Shpaer EG, Ligenfelter P, Singal R, and Mullins JI. Rapid genetic characterization of HIV type 1 strains from four World Health Organization-sponsored vaccine evaluation sites using a heteroduplex mobility assay. WHO Network for HIV Isolation and Characterization. AIDS Research and Human Retroviruses. 1994. 10: 1345-1353

Baldauf SL. Phylogeny for the faint of heart: a tutorial. TRENDS in Genetics. 2003. 19: 345-351

Barre-Sinoussi F, Chermann JC, Rey F, Nugeyre MT, Chamaret S, Gruest J, Dauguet C, Axler-Blin C, Brun-Vezinet F, Rouzioux C, Rozenbaum W, and Montagnier L. Isolation of a T-Lymphotropic retrovirus from a patient at risk for Acquired Immune Deficiency Syndrome (AIDS). Science. 1983. 220: 868-871

Bebenek K, Abbotts J, Roberts JD, Wilson SH, and Kunkel TA. Specificity and mechanism of error-prone replication by human immunodeficiency virus-1 reverse transcriptase. Journal of Biological Chemistry. 1989. 264: 16948-16956

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

Becker MLB, Spracklen FHN, and Becker WB. Isolation of a Lymphadenopathy associated virus from a patient with acquired immune deficiency syndrome. South African Medical Journal. 1985. 68: 144-147

Bell CM, Connell BJ, Capovilla A, Venter WDF, Stenves WS, and Papathanasopoulos MA. Molecular Characterization of the HIV Type 1 Subtype C Accessory Genes *vif*, *vpr*, and *vpu*. AIDS Research and Human Retroviruses. 2007. 23: 322-330

Bessong PO, Obi CL, Cilliers T, Choge I, Phoswa M, Pillay C, Papathanasopoulos M, and Morris L. Characterization of Human Immunodeficiency Virus Type 1 from a Previously Unexplored Region of

South Africa with a High HIV Prevalence. AIDS Research and Human Retroviruses. 2005. 21: 103-109

Boyer JC, Bebenek K, Kunkel TA. Unequal human immunodeficiency virus type 1 reverse transcriptase error rates with RNA and DNA templates. Proceedings of the National Academy of Sciences. 1992. 89: 6919-6923

Bredell H, Hunt G, Casterling A, Cilliers T, Rademeyer C, Coetzer M, Miller S, Johnson D, Williamson C, and Morris L. HIV-1 Subtype A, D, G, AG and Unclassified Sequences identified in South Africa. Aids Research and Human Retroviruses. 2002. 18; 9: 681-683

Bredell H, Williamson C, Sonnenberg P, Martin DJ, and Morris L. Genetic characterization of HIV type 1 from migrant workers in three South African gold miners. AIDS Research and Human Retroviruses. 1998. 14: 677-684

Buonaguro L, Tornesello ML, and Buonaguro FM. Human Immunodeficiency Virus Type 1 Subtype Distribution in the Worldwide Epidemic: Pathogenetic and Therapeutic Implications. Journal of Virology. 2007. 81: 10209-10219

Carr JK, Salminen MO, Albert J, Sanders – Buell E, Gotte D, Birx DL, and McCutchan FE. Full genome sequences of human immunodeficiency virus type 1 subtype G and A/G intersubtype recombinants. Virology. 1998. 247: 22-31

Carr JK, Salminen MO, Koch C, et al. Full-length sequence and mosaic structure of a human immunodeficiency virus type 1 isolate from Thailand. Journal of Virology. 1996. 70: 5935-5943

CDC. 'Kaposi's Sarcoma (KS), Pneumocystis Carinii Pneumonia, and Other Opportunistic Infections (01): Cases Reported to CDC as of July 8'. 1982 – www.cdc.gov/hiv/topics/surveillance/resources/reports/surveillance82.pdf

CDC. Kaposi's sarcoma and Pnerumocycsits pneumonia among homosexual men – New York City and California. Morbitity and Mortality Weekly Report. 1981. 30: 305-308

Chan D and Kim P. HIV entry and its inhibition. Cell. 1998. 93: 681-684

Chan DC, Fass D, Berger JM, and Kim PC. Core Structure of gp41 from the HIV Envelope Glycoprotein. Cell. 1997. 89: 263-273

Choi DJ, Dube S, Spicer TP, Slade HB, Jensen FC, and Poiesz BJ. HIV type 1 isolate Z321, the strain used to make a therapeautic HIV type 1 immunogen, is intersubtype recombinant. AIDS Research and Human Retrovirus. 1997. 13: 357-361

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

Cleghorn FR, Jack N, Carr JK, Edwards J, Mahabir B, Sill A, McDanal CB, Connolly SM, Goodman D, Bennetts RQ, O'Brien TR, Weinhold KJ, Bartholomew C, Blattner WA and Greenberg ML. 2000. A distinctive clade B HIV type 1 is heterosexually transmitted in Trinidad and Tobago. Proceedings of the National Academy of Science. 97: 10532-10537.

Clumeck N, Mascart-Lemone F, de Maubeuge J, Brenez D, and Marcelis L. Acquired immune deficiency syndrome in Black Africans. Lancet. 1983.1: 642

Curran JW, Lawrence DN, Jaffe H, Kaplan JE, Zyla LD, Chamberland M, Weinstein R, Lui KJ, Schonberger LB, and Spira TJ. Acquired immunodeficiency syndrome (AIDS) associated with transfusions. New England Journal of Medicine. 1984. 310; 69-74

De Cock KM, Adjorlolo KMG, Ekpini E, Sibailly T, Kouadio J, Maran M, Brattegaard K, Vetter KM, Doorly R, and Gayle HD. Epidemiology and transmission of HIV-2: why there is no HIV-2 pandemic. Journal of American Medical Association. 1993. 270: 2083-2086

De Leys R, Vanderborght B, Vanden Haesevelde M, Heyndrickx L, van Geel A, Wauters C, Bernaerts R, Saman E, Nijs P, Willems B, Taelman H, Van der Groen G, Piot P, Tersmette T, Huisman JG, and Van Heuverswyn H. Isolation and partial characterization of an unusual human immunodeficiency retrovirus

from two persons of west-central African origin. Journal of Virology. 1990. 64: 1207-1216

Delwart EL, Shpaer EG, McCutchan FE, Louwagie J, Grez M, Rübsamen-Waigmann H, Mullins JI.. Genetic relationships determined by a DNA heteroduplex mobility assay: Analysis of HIV-1 *env* genes. Science. 1993. 262:1257–1261

de Oliveira T, Engelbrecht S, Janse van Rensburg E, Gordon M, Bishop K, zur Magede J, Barnett SW, and Cassol S. Variability at Human Immunodeficiency Virus Type 1 Subtype C Protease Cleavage Sites: an Indication of Viral Fitness? Journal of Virology. 2003. 77: 9422-9430

de Oliveira T, Deforche K, Cassol S, Salminen M, Paraskevis D, Seebregts C, Snoeck J, Janse van Rensburg E, Wensing AMJ, van de Vijver DA, Boucher CA, Camacho R, and Vandamme A-M. An automated genotyping system for analysis of HIV-1 and other microbial sequences. Bioinformatics Applications Note. 2005. 21: 3797-3800

Derdeyn CA, Becker JM, Bibollet-Ruche F, Mokili JL, Muldoon M, Denham SA, Heil ML, Kasolo F, Musonda R, Hahn BH, Shaw GM, Korber BT, Allen S, and Hunter E. Envelope-constrained nuetralizationsensitive HIV-1 after heterosexual transmission. Science. 2004. 303: 2019-2022

Department of Health, Republic of South Africa. National HIV and Syphilus Antenatal sero-prevalence survey in South Africa. 2004. Pretoria, South Africa: Directorate Health systems Research, Department of Health; 2005.

Dicker IB, Samanta HK, Li Z, Hong Y, Tian Y, Banville J, Remillard RR, Walker MA, Langley DR, and Krystal M. Changes to the HIV Long Terminal repeat and to HIV Integrase Differentially Impact HIV Integrase Assembly, Activity, and the Binding of Strand Transfer Inhibitors. Journal of Biological Chemistry. 2007. 43: 31186-1196

Dowling WE, Kim B, Mason CJ, Birx DL, Robb ML, McCutchan FE, and Carr JK. Forty-one near full-length HIV-1 sequences from Kenya reveal an

epidemic of subtype A and A-containing recombinants. AIDS. 2002. 16: 1809-1820

Drummond AJ and Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. BioMed Central Evolutionary Biology. 2007. 7; 214

Engelbrecht S, de Villiers T, Sampson CC, zur Megede J, Barnett SW, and Janse van Rensburg E. Genetic Analysis of the Complete gag and env Genes of HIV Type 1 Subtype C Primary Isolates from South Africa. AIDS Research and Human Retroviruses. 2001. 17: 1533-1547

Engelbrecht S, Laten JD, Smith T-L, and van Rensburg EJ. Identification of env subtypes in fourteen HIV type 1 isolates from South Africa. AIDS Research and Human Retroviruses. 1995. 11: 1269-1271

Engelbrecht S, Smith T-L, Kasper P, Faatz E, Zeier M, Moodley D, Clay CG, and van Rensburg EJ. HIV Type 1 V3 Domain Serotyping and Genotyping in Gauteng, Mpumalanga, KwaZulu-Natal, and the Western Cape Provinces of South Africa. AIDS Research and Human Retroviruses. 1999. 15: 325-328

Engelbrecht S, de Jager GJ, and van Rensburg EJ. Evaluation of commercially available assays fro antibodies to HIV-1 in serum obtained from South African patients infected with HIV-1 subtypes B, C, and D. Journal of Medical Virology. 1994. 44: 223-228

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

Essex M, Hardy WD Jr, Cotter SM, Jakowski RM, and Sliski A. Naturally occurring persistent feline oncornavirus infections in the absence of disease. Infection and Immunity. 1975.11(3): 470-475

Essex M and Kanki PJ. The origins of the AIDS virus. Science America. 1998. 259: 64-71

Essex M and Mboup S. AIDS in Africa (2nd edition). Edited by Essex M, Mboup S, Kanki PJ, Marlink RG and Tlou SD. Kluwer Academic/Plenum Publishers, New York (NY), United States of America. 2002.

Fahey JL, Prince H, Weaver M, Groopman J, Visscher B, Schwartz K, and Detels R. Quantitative changes in T helper or T suppressor/cytotoxic lymphocyte subsets that distinguish acquired immune deficiency syndrome from other immune subset disorders. American Journal of Medicine. 1984. 76: 95-100

Fang G, Weiser B, Visosky AA, Townsend L, and Burger H. Molecular Cloning of Full-length HIV-1 Genomes Directly from Plasma Viral RNA. AIDS and Human Retroviruses. 1996. 12: 352-357

Felsenstein J. Numerical methods for inferring evolutionary trees. Quarterly Review of Biology. 1982. 57: 379-404.

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

Fischetti L, Opare-Serri O, Candotti D, Sarkodie F, Lee H, and Allaini JP. Molecular Epidemiology of HIV in Ghana: Dominance of CRF02_AG. Journal of Medical Virology. 2004. 73: 158-166

Fitch WM and Margoliash E. Construction of phylogenetic trees. Science. 1967. 20: 279-284

Flügel RM. Spumaviruses: a group of complex retroviruses. AIDS. 1991. 4: 739-750

Friedman-Kien AE, Laubenstain L, Marmor M, Hymes K, Green J, Ragaz A, Gottleib J, Muggia F, Demopoulos R, Weintraub M, Williams D, Oliveri R, Marmer J, Wallace J, Halperin I, Gillooley JF, Prose N, Klein E, Vogel J, Safai B, Myskowski P, Urmacher C, Koziner B, Nisce L, Kris M, Armstrong D, Gold J, Mildran D, Tapper M, Weissman JB, Rothenberg R, Friedman SM, Siegal FP, Groundwater J, Gilmore J, Coleman D, Follansbee S, Gullett J, Stegman

SJ, Wofsy C, Bush D, Drew L, Braff E, Dritz S, Klein M, Preiksaitis JK, Gottlieb MS, Jung R, Chin Jand Goedert J. Kaposi's sarcoma and Pneumocystis pneumonia among homosexual men - New York City and California. 1981. Morbitity and Mortality Weekly Report – Center for Disease Control and Prevention. 30: 305-308

Gallo RC. The discovery of the first human retrovirus: HTLV-1 and HTLV-2. Retrovirology. 2005. 2: 17

Gallo RC, Salahuddin SZ, Popvic M, Shearer GM, Kaplan M, Haynes BF, Palker TJ, Redfield R, Olseke J, and Safai B. Frequenct detection and isolation of cytopathic retroviruses (HTLV-III) from patients with AIDS and at risk of AIDS. Science. 1984. 224(4648): 500-503

Gange RW, and Jones EW. Kaposi's sarcoma and immunosuppressive therapy: an appraisel. Clinical Experimental Dermatology. 1978. 3: 135-146

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. Origin of HIV-1 in the chimpanzee *Pan troglodytes troglodytes*. Nature. 1999. 97: 436–441.

Gao F, Robertson DL, Morrison SG, et al. The heterosexual human immunodeficiency virus type 1 epidemic in Thailand is caused by an intersubtype (A/E) recombinant of African origin. Journal of Virology. 1996. 70: 7013-7029

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

Gao F, Yue L, Robertson DL, Hill SC, Hui H, Biggar RJ, Neequaye AE, Whelan TM, Ho DD, Shaw GM, Sharp PM, and Hahn BH. Genetic diversity of Human Immunodeficiency Virus Type2: Evidence for Distinct Sequence Subtypes with Differences in Virus Biology. Journal of Virology. 1994. 68: 7433-7447

Gelderblom H (1997).Fine Structure of HIV and SIV. pp. IV-37-50 in HIV Molecular Immunology Database 1997. Edited by: Korber B, Brander C, Haynes B, Koup R, Moore J, Walker B. Published by: Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM. 2007

Geretti AM. HIV-1 subtypes: epidemiology and significance of HIV management. Current Opinions in Infectious Diseases. 2006. 19: 1-7

Goldberg B, and Stricker RB. Apoptosis and HIV infection: T-cells fiddle while the immune system burns. Immunology Letters. 1999. 70: 5-8

Goldman N and Yang Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Molecular Biological Evolution. 1994. 11: 725-736

Gordon M, de Oliveira T, Bishop K, Coovadia HM, Madurai L, Engelbrecht S, Janse van Rensburg E, Mosam A, Smith A, and Cassol S. Molecular Characteristics of Human Immunodeficiency Virus Type 1 Subtype C Viruses from KwaZulu-Natal, South Africa: Implications for Vaccine and Antiretroviral Control Strategies. Journal of Virology. 2003. 77: 2587-2599

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

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

Guindon S, and Gascuel O. A simple, fast and accurate algorithm to estimate larges phylogenies by maximum likelihood. Syst. Biol. 2003. 696-704

Gurtler LG, Hauser PH, Eberle J, von Brunn A, Knapp S, Zekeng L, Tsague JM, and Kaptue L. A new subtype of human immunodeficiency virus type 1 (MVP-5180) from Cameroon. Journal of Virology. 1994. 68: 1581-1585

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

Hall BG. Phylogenetic Trees Made Easy: A How-To Manual. Sunderland Associates, Inc. Massachusetts, United States of America. 2004

Hall T. BioEdit, Biological sequence alignment for Windows 95/98/NT. 2001

Hardy WD Jr, Old LJ, Hess PW, Essex M, and Cotter S. Horizontal transmission of feline leukaemia virus. Nature. 1973. 244: 266-269

Hasegawa M, Kishino K, and Yano T. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. Journal of Molecular Evolution. 1985. 22: 160-174

Hasegawa A, Tsujimoto H, Maki N, Ishikawa KI, Miura T, Fukasawa M, Miki K, and Hayami M. Genomic divergence of HIV-2 from Ghana. AIDS Research Human Retroviruses. 1989. 5: 593-604

Hayman A, Moss T, Simmons G, Arnold C, Holmes EC, Naylor-Adamson L, Hawkswell J, Allen K, Radford J, Nguyen-Van-Tam J and Balfe P. 2001. Phylogenetic analyses of multiple heterosexual transmission events involving 10 subtype B of HIV type 1. Aids Research and Human Retroviruses. 17: 689-695.

Hemelaar J, Gouws E, Ghys PD, and Osmanov S. Global and regional distribution of HIV-1 genetic subtypes and recombinants in 2004. AIDS. 2006. 20(16): W13 - W23

Hirsch V, Olmsted R, Murphy-Corb M, Pyrcell A and Johnson P. An African primate lentivirus (SIVsm) closely related to HIV-2. Nature. 339: 389-392

Ho DD, Neumann AU, Perelson SA, Chen J, Leonard JM, and Markowits M. Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infections. Nature. 1995. 373: 123-126

Hogeweg P and Hesper B. The alignment of sets of sequences and the construction of phylogenetic trees: An integrated method. Journal of Molecular Evolution. 1984. 20: 175-186

Holmes I. Sequence homology search tools on the World Wide Web. pp. 44-53 in HIV Sequence Compendium 2000. Edited by: Kuiken C, McCutchan F, Foley B, Mellors JW, Hahn B, Mullins J, Marx P, Wolinsky S. Published by: Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM. 2000.

Howard TM, and Rasheed S. Genomic structure and nucleotide sequence analysis of a new HIV type 1 subtype A strain from Nigeria. AIDS Research and Human Retroviruses. 1996. 12: 1413-1425

http://en.wikipedia.org/wiki/HIV

http://www.hiv.lanl.gov/ - HIV Sequence Compendium, 06/07. Leitner T, McCutchan F, Foley B, Mellors JW, Hahn B, Wolinsky S, Marx P and Korber B. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM, LA-UR 03-3564

http://www.talkorigins.org/

Huebner RJ, and Todaro GJ. Oncogenes of RNA tumor viruses as determinants of cancer. Proceedings of the National Academy of Science USA. 1969. 64: 1087-1094

Hunt GM, Johnson D, and Tiemessen CT. Characterisation of the Long Terminal Repeat Regions of South African Human Immunodeficiency Virus Type 1 Isolates. Virus Genes. 2001. 23: 27-34

Hunt GM, Papathanasopoulos MA, Gray GE, and Tiemessen CT. Characterization of Near-full Length Genome Sequences of Three South African Human Immunodeficiency Virus Type 1 Subtype C Isolates. Virus Genes. 2003. 26: 49-56

Hymes KB, Cheung T, Greene JB, Prose NS, Marcus A, Ballard H, and Williams DC. 'Kaposi's sarcoma in homosexual men: A report of eight cases', Lancet. 1981. 2: 598-600

Jacobs GB, de Beer C, Fincham JE, Adams V, Dansay MA, Janse Van Rensburg E, and Engelbrecht S. Serotyping and Genotyping of HIV-1 Infection in Residents of Khayelitsha, Cape Town, South Africa. Journal of Medical Virology. 2006. 78: 1529-1536

Jacobs GB, Loxton AG, Laten A, and Engelbrecht S. Complete Genome Sequencing of a Non-syncytium-Inducting HIV Type 1 Subtype D Strain from Cape Town, South Africa. AIDS Research and Human Retroviruses. 2007. 23: 1575-1578

Jacobs GB, Nistal M, Laten A, Janse van Rensburg E, Rethwilm A, Preiser W, Bodem J, and Engelbrecht S. Molecular Analysis of HIV Type 1 vif Sequences from Cape Town, South Africa. AIDS Research and Human Retroviruses. 2008a. 24: (991-994)

Jacobs GB, Loxton AG, Laten JD, Brenda R, van Rensburg EJ, and Engelbrecht S. Emergence and diversity of different HIV-1 subtypes in Cape Town, South Africa. Journal of Medical Virology. Submitted

Jacobs GB, Laten AD, van Rensburg EJ, Bodem J, Weissbrich B, Rethwilm A, Preiser W, and Engelbrecht S. Phylogenetic Diversity and Low Level Antiretroviral Resistance Mutations in HIV Type 1 Treatment-Naive Patients from Cape Town, South Africa. AIDS Research and Human Retroviruses. 2008b. 24: (1009-1012)

Jaffe HW, Francis DP, McLane MF, Cabradilla C, Curran JW, Kilbourne BW, Lawrence DN, Haverkos HW, Spira TJ, and Dodd RY. Transfusion-associated AIDS: serologic evidence of human T-cell leukemia virus infection of donors. Science. 1984. 223: 1309-1312

Janssens W. Heyndrickx L, Fransen K, Motte J, Peeters M, Nkengasong JN, Ndumbe PM, Delaporte E, Perret JL, Atende C, Piot P, and van der Groen G.

Gentic and phylogentic analysis of env subtypes G and H in Central Africa. AIDS Research and Human Retroviruses. 1994. 10: 877-879

Jukes T, and Cantor CR. Evolution of protein molecules. In: Mammalian protein Metabolism, edited by Munro HN, pp 21-132. New York, Academic Press (1969).

Kanki PJ, Travers KU, Mboup S, Hsieh CC, Marlink RG, and Gueye-NDiaye A. Slower heterosexual spread of HIV-2 than HIV-1. Lancet. 1994. 343: 943-946

Kemp DJ, Smith DB, Foote SJ, Samaras N, and Petersen MG. Colorimetric detection of specific DNA segments amplified by polymerase chain reactions. Proceedings of the National Academy of Sciences. 1989. 86: 2423-2427.

Kimura M. The Neutral Theory of Molecular Evolution. Cambridge: Cambridge University Press, UK. 1980

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

Lara C, Sällberg M, Johasen B, De Rivera I and Sönnerborg A. 1 1997. The Honduran human immunodeficiency virus type 1 (HIV-1) epidemic is dominated by HIV-1 subtype B determined by V3 domain sero and genotyping. Journal of Clinical Microbiology 35: 783-784.

Learn GH, Korber BTM, Foley B, Hahn BH, Wolinsky SM, and Mullins JI. Maintaining the integrity of human immunodeficiency virus database. 1996. Journal of Virology. 70: 5720-5730

Leitner T, Foley B, Hahn B, *et al.* HIV sequence compendium 2005. Los Alamos, NM: Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, 2005.

Leitner T. Genetic subtypes of HIV-1. In Human Retroviruses and AIDS 1996. A compilation and analysis of nucleic acid and amino acid sequences. Myers G, Korber BT, Foley BT. Theoretical Biology and Biophysics Group. Los Alamos National Laboratory, Los Alamos, New Mexico. 1996

Lemey P, Pybus OG, Rambaut A, Drummond AJ, Robertson DL, Roques P, Worobey M, and Vandamme A-M. The Molecular population Genetics of HIV-1 Group O. Genetics. 2004. 167: 1059-1068

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

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

Levy JA. HIV and the Pathogenesis of AIDS. 3rd Edition. ASM Press, Washington D.C. (2007).

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

Lole KS, Bollinger RC, Paranjape RS, Gadkari D, Kulkarni SS, Novak NG, Ingersoll R, Sheppard HW, Ray SC. Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. Journal of Virology. 1999. 73: 152-160

Louwagie J, McCutchan FE, Peeters M, Brenamn TP, Sanders-Buell E, Eddy GA, van der Groen G, Fransen K, Gershy-Damet GM, Deleys R, and Burke DS. Phylogenetic analysis of gag genes from 70 international HIV-1 isolates provides evidence for multiple genotypes. AIDS. 1993. 7: 769-780

Loxton AG, Treurnicht F, Laten A, Janse Van Rensburg E, and Engelbrecht S. Sequence analysis of near full-length HIV type 1 subtype D primary strains

isolated in Cape Town, South Africa, from 1984 to 1986. AIDS Research and Human Retroviruses. 2005. 21: 410-413

Markovitz D. Infection with the human immunodeficiency virus type 2. Annuals of Internal Medicine. 1993. 188: 211-218

Marlink R, Kanki PJ, Thior I, Travers K, Eisen G, Siby T, Traore I, Hsieh CC, Dia MC, and Gueye EH. Reduced rate of disease development after HIV-2 infection as compared to HIV-1. Science. 1994. 265: 1587-1590

Martin DP, Williamson C, and Posada D. RDP2: recombination detection and analysis from sequence alignments. Bioinformatics. 2004. 21: 260-262

Masur H, Michelis MA, Greene JB, Onorato I, Stouwe RA, Holzman RS, Wormser G, Brettman L, Lange M, Murray HW, and Cunningham-Rundles S. An outbreak of community-acquired Pneumocystis carinii pneumonia: initial manifestation of cellular immune dysfunction. New England Journal of Medicine. 1981. 305: 1431-1438

Mau B, Newton MA and Larget B. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Biometrics. 1999. 55: 1-12

McCormack GP and Clewley JP. The application of molecular phylogenetics to the analysis of viral genome diversity and evolution. Review in Medical Virology. 2002. 12: 221-238

McCutchan FE, Carr JK, Bajani M, Sanders-Buell E, Harry TO, Stoeckli TC, Robbins KE, Gashau W, Nasidi A, Janssens W, and Kalish ML. Subtype G and multiple forms of A/G intersubtype recombinant human immunodeficiency virus type 1 in Nigeria. Virology. 1999. 254: 226-234

McCutchan FE. Global epidemiology of HIV. Journal of Medical Virology. 2006. 78(suppl. 1): S7-S12

McCutchan FE, Hegerich PA, Brennan TP, et al. Genetic variants of HIV-1 in Thailand. AIDS Research Human Retroviruses. 1992. 8: 1887-1895

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

Menu E, Truong TX, Lafon ME, Nguyen TH, Muller-Trutwin MC, Nguyen TT, Deslandres A, Chaouat G, Duong QT, Ha BK, Fleury HJ, and Barre-Sinoussi F. HIV type 1 Thai subtype E is predominant in South Vietnam. AIDS Research and Human Retroviruses. 1996. 12: 629-633

Michael NL. Host genetic influences on HIV-1 pathogenesis. Current Opinions in Immunology. 1999. 11: 466-474

Mwaengo DM, and Novembre FJ. Molecular cloning and characterization of viruses isolated from chimpanzees with pathogenic human immunodeficiency virus type 1 infections. Journal of Virology. 1998. 72: 8976-8987

Myers G. HIV-1 and HIV-2 sequence subtypes. Human Retroviruses and AIDS. 1992. (Myers G, Korber B, Berzofsky JA, and Smith RA., eds) Los Alamos National Laboratory, Los Alamos, NM. 1992.

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

Myers G. Tenth anniversary perspectives on AIDS. HIV: between the past and future. AIDS Research and Human Retroviruses. 1994. 10: 1317-1324

Nadai Y, Eyzaguirre ML, Constantine NT, Sill AM, Cleghorn F, Blattner WA, and Carr JK. Protocol for Nearly Full-Length Sequencing of HIV-1 RNA from Plasma. PLOS one. 2008. 1: e1420

Nahmias AJ, Weiss J, Yao X, Lee F, Kodsi R, Schanfield M, Matthews T, Bolognesi D, Durack D, and Motusky A. Evidence for human infection with an HTLV-III/LAV like virus in central Africa, 1952. Lancet. 1986. I: 373-386

Njai HF, Gali Y, Vanham G, Clybergh C, Jennes W, Vidal N, Butel C, Mpoudi-Ngolle E, Peeters M, and Ariên KK. The predominance of Human

Immunodeficiency Virus type 1 (HIV-1) circulating recombinant form 02 (CRF02_AG) in West Central Africa may be related to its replicative fitness. Retrovirology. 2006. 3:40 4690-4701

Olseke J, Minnefor A, and Cooper R Jr. Immune deficiency syndrome in children. Journal of American Medical Association. 1983. 249: 2345-2349

Op de Coul ELM, Prins M, Cornelissen M, van der Schoot A, Boufassa F, Brettle RP, Hernandez-Aguado I, Schiffer V, McMenamin J, Rezza G, Robertson R, Zangerie R, Goudsmit J, Coutinho RA, and Lukashov V. Using phylogenetic analysis to trace HIV-1 migration among western European infecting drug users seroconverting from 1984 to 1997. AIDS. 2001. 15: 257-266

Osmanov S, Pattou C, Walker N, Schwarlander B, Esparza J and the WHO-AIDS Network for HIV isolation and Characterisation. Estimated global distribution and regional spread of HIV-1 genetic subtypes in the year 2002. Journal of Acquired Immune Deficiency Syndrome. 2002. 29: 184-190

Ou C-Y, Kwok S, Mitchell SW, Mach DH, Sninsky JJ, Krebs JW et al,. DNA amplification for direct detection of HIV-1 in DNA of peripheral blood mononuclear cells. Science. 1988. 239: 295-297

Page RDM and Holmes EC. Molecular evolution: A phylogenetic approach. Blackwell Science Ltd. Oxford, United Kingdom. 1998.

Papathanasopoulos MA, Cilliers T, Morris L, Mokili JL, Dowling W, Brix DL, and McCutchan FE. Full-Length Genome Analysis of HIV-1 Subtype C Utilizing CXCR4 and Intersubtype Recombinants Isolated in South Africa. AIDS Research and Human Retroviruses. 2002. 18: 879-886

Papathanasopoulos MA, Cilliers T, Morris L, Mokili JL, Dowling W, Carr JK, McCutchan F. First Full-Genome Sequences of Intersubtype Recombinants Circulating in South Africa. AIDS Vaccine 2001, Sep 5-8; National Institute of Virology, Johannesburg, South Africa. 2001.

Papathanasopoulos MA, Hunt GM, and Tiemessen CT. Evolution and Diversity of HIV in Africa – a Review. Virus Genes. 2003a. 26: 151-163

Papathanasopoulos MA, Patience T, Meyers TM, McCutchan FE, and Morris L. Full-Length Genome Characterization of HIV Type 1 Subtype C Isolates from Two Slow-Progressing Perinatally Infected Siblings in South Africa. AIDS Research and Human Retroviruses. 2003b. 19: 1033-1037

Peeters M. Recombinant HIV sequences: Their role in the global epidemic. pp. I-39-54 in HIV Sequence Compendium. 2000. Edited by: Kuiken CL, and Foley B.

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

Peeters M, Toure-Kane C, and Nkengasong JN. Genetic diversity of HIV in Africa: impact on diagnosis, treatment, vaccine development and trails [EDITORIAL REVIEW]. AIDS. 2003. 17: 2547-256

Penn I. Kaposi's sarcoma in organ transplant recipients: report of 20 cases. Transplantation. 1979. 27: 8-11

Pepin J, Morgan G, Dunn D, Gevao S, Mendy M, Gaye I, Scollen N, Tedder R, and Whttle H. HIV-2-induced immunosuppression among asymptomatic West African prostitutes: evidence that HIV-2 is pathogenic, but less so than HIV-1. AIDS. 1991. 5: 1165-1172

Pieniazek D, Yang C, and Lal RB. Phylogenetic Analysis of gp41 Envelope of HIV-1 Groups M, N and O in Human Retroviruses and AIDS 1998: A Compilation and Analysis of Nucleic Acid and Amino Acid Sequences. Korber B, Kuiken CL, Foley B, Hahn B, McCutchan F, Mellors JW, and Sodroski J, eds. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM. 1998.

Plantier JC, Dachraoui R, Lemee V, Gueudin M, Borsa-Lebas F, Caron F, and Simon F. HIV-1 resistance genotyping on dried serum spots. AIDS. 2005. 19(4): 391-397

Potts KE, Kalish ML, Bandea CI, Orloff GM, St. Louis M, Brown C, Malanda N, Kavuka M, Schochetman G, Ou CY, and Heyward WL. Genetic diversity of human immunodeficiency virus type 1 strain in Kinshasa, Zaire. AIDS Research Human Retroviruses. 1993. 9: 613-618

Quinn TC, Mann JM, Curran JW, and Piot P. AIDS in Africa: An Epidemiologic Paradigm. Science. 1986. 234: 955-963

Rannala B and Yang Z. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. Journal of Molecular Evolution. 1996. 43: 304-311

Ras GJ, Simson IW, Anderson R, Prozeksy OW and Hamersma T. Acquired immunodeficiency syndrome: a report of 2 South African cases. South African Medical Journal. 1983. 64: 140-142

Ratner L, Haseltine W, Patarca R, Livak KJ, Starcich B, Josephs SF, Doran ER, Rafalski A, Whitehorn EA, Baumeister K, Ivanoff L, Petteway SR, Pearson ML, Lauenberger JA, Papas TS, Ghrayeb J, Chang NT, Callo RC and Wong-Staal F. Complete nucleotide sequence of the AIDS virus, HTLV-III. Nature 1985a. 313: 227-284

Ratner L, Gallo RC, and Wong-Staal F. HTLV-III, LAV, ARV are variants of same AIDS virus. Nature. 1985b 313: 636-637

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

Robertson DL, Anderson JP, Bradac JA, Carr JK, Foley B, Funkhouser RK, Gao F, Hahn BH, Kalish ML, Kuiken C, Learn GH, Leither T, McCutchan F,

Osmanov S, Peeters M, Pieniazek D, Salminen M, Sharp PM, Wolinsky S, and Korber B. HIV-1 nomenclature proposal. Science. 2000. 288: 55-56

Roberts JD, Bebenek K, and Kunkel TA. The accuracy of reverse transcriptase from HIV-1. Science. 1988. 242: 1171-1173

Robertson DL, Sharp PM, McCutchan FE, ad Hahn BH. Recombination in HIV-1. Nature. 1995. 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, UNAIDS and NIAID Networks for HIV Isolation and Characterization. Near full-length clones and reference sequences for subtype C isolates of HIV type 1 from three different continents. AIDS Research and Human Retroviruses. 2001. 17: 161-168

Rodriguez F, Oliver JF, Marin A, and Medina JR. The general stochastic model of nucleotide substitution. Journal of Theoretical Biology. 1990. 142: 485-501

Rodrigo AG and Learn GH Jr. Computational and Evolutionary analysis of HIV molecular sequences. Kluwer Academic Publishers. Dordrecht, The Netherlands. 2001.

Rousseau CM, Birditt BA, McKay AR, Stoddard JN, Lee TC, McLauglin S, Moore SW, Shindo N, Learn GH, Korber BT, Brander C, Goulder PJR, Kiepiela P, Walker BD, and Mullins JI. Large-scale amplification, cloning and sequencing of near full-length HIV-1 subtype C genomes. Journal of Virological Methods. 2006. 136: 118-125

Rubinstein A, Sticklick M, Gupta A, Bernstein L, Klein N, Rubinstein E, Spigland I, Fruchter L, Litman N, Lee H, and Hollander M. Acquired immunodeficiency with reversed T4/T8 ratios in infants born to promiscuous and drug-addicted mothers. Journal of American Medical Assosiation. 1983. 249: 2350-2356

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

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

Salminen MO, Carr JK, Burke DS, and McCutchan FE. Identification of breakpoints in intergenotypic recombinats of HIV type 1 by booscanning. AIDS Research and Human Retroviruses. 1995b. 11: 1423-1425

Salemi M and Vandamme A-M. The Phylogenetic Handbook: A practical approach to DNA and protein phylogeny. Cambridge University Press. Cambridge, United Kingdom. 2003

Salemi M, Strimmer K, Hall WW, Duffy M, Delaporte E, Mboup S, Peeters M and Vandamme A-M. Dating the common ancestor of SIVcpz and HIV-1 group M and the origin of HIV-1 subtypes using a new method to uncover clock-like molecular evolution. The FASEB Journal. 2001. 15: 276-278

Sanders-Buell E, Salminen MO, and McCutchan FE. Sequencing primers for HIV-1 IN: Human Retroviruses and AIDS 1995: Kuiken C, Foley B, Hahn B, Marx P, McCutchan F, Mellors J, Wolinsky S, and Korber B. A compilation and analysis on nucleic acid and amino acid sequences. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM, USA. 1995.

Schultz AK, Zhang M, Leitner T, Kuiken C, Korber B, Morgenstern B, and Stanke M. A jumping profile Hidden Markov Model and applications to recombination sites in HIV and HCV genomes. BMC Bioinformatics. 2006. 7: 265-280

Scriba TJ, de Villiers T, Treurnicht FK, zur Megede J, Barnett SW, Engelbrecht S, and van Rensburg EJ. Characterization of the South African

HIV Type 1 C Complete 5' Long Terminal Repeat, *nef*, and Regulatory genes. AIDS Research and Human Retroviruses. 2002. 18: 149-159

Serwadda D, Mugerwa RD, Sewankambo NK, Lwegaba A, Carswell JW, Kirya GB, Bayley AC, Downing RG, Tedder RS, and Clayden SA. 1985. Slim disease: a new disease in Uganda and its association with HTLV-III infection. Lancet. 2: 849-852

Shankarappa R, Margolick JB, Gange SJ, Rodrigo AG, Upchurch D, Fazadegan H, Gupta P, Rinaldo CR, Learn GH, He XL and Mullins JI. Consistent viral evolutionary dynamics associated with the progression of HIV-1 infection. Journal of Virology. 1999. 73: 10489-10502

Sher R. HIV infection in South Africa, 1982 - 1988 - a review. South African Medical Journal. 1989. 76(Oct): 314-318

Shilts RM. And the band Played on: Politics, people and the AIDS epidemic. St. Martin's Press. New York, USA. 1987.

Siepel AC, Halpern AL, Macken C, and Korber BT. A computer program designed to screen rapidly for HIV type 1 intersubtype recombinant sequences. AIDS Research Human Retroviruses. 1995. 11: 1413-1416

Silvestri G, Sodora D, Koup R, Paiardini M, O'Neil S, McClure S, Staprans S, and Feinberg M. Nonpathogenic SIV infection of Sooty Mangabeys is characterized by limited bystander immunopathology despite chronic high-level viremia. Immunity. 2003. 18: 441-452

Sneath PHA and Sokal RR. Numerical Taxonomy. Freeman, San Francisco, USA. 1973.

Songok EM, Lwembe RM, Kibaya R, Kobayashi K, Ndembi N, Kita K, Vulule J, Oishi I, Okoth F, Kageyama S, and Icimura H. Active Generation and Selection of HIV Intersubtype A/D Recombinant Forms in a Coinfected Patient in Kenya. AIDS Research and Human Retroviruses. 2004. 20: 255-258

Spang R, Rehmsmeier M, and Stoye J. A Novel Approach to Remote Homology Detection: Jumping Alignments. Journal of Computational Biology. 2002. 9: 747-760

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

Swanson P, Devare SG, and Hackettt J Jr. Molecular Characterization of 39 HIV-1 Isolates Representing Group M (Subtype A-G) and Group O: Sequence Analysis of gag p24, pol Integrase, and env gp41. AIDS Research and Human Retroviruses. 2003. 19: 625-629

Swofford DL. PAUP*. *Phylogenetic Analysis Using Parsimony (* and other methods). Version 4.0b10.* Sunderland, Massachusetts (USA): Sinauer Associates, Inc. 2002.

Tamura K, Dudley J, Nei M, and Kumar S. *MEGA4*: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Molecular Biology and Evolution. 2007. 24:1596-1599.

Tavaré S. Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences. American Mathematical Society: Lectures on Mathematics in the Life Sciences. 1986. 17: 57–86.

Taylor BS, Sobieszcczyk ME, McCutchan FE, and Hammer SM. The Challenge of HIV-1 Subtype Diversity. New England Journal of Medicine. 2008. 358: 1590-1602

Temin HM. Retrovirus variation and reverse transcription: abnormal strand transfers result in retrovirus genetic variation. Proc. Natl. Acad. Sci USA. 1993. 22: 6900-6903

Thompson JD, Higgins DG and Gibson TJ. CLUSTAL W: Improving the sensitivity of progressive multiple-sequence alignment through sequence

weighting, position-specific gap penalties, and weight matrix choice. Nucleic Acids Research. 1994. 22: 4673-4680

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

Torques K, Bourgeois A, Saragosti S, Vidal N, Mpoudi-Ngolle E, Nzilambi N, Apetrei C, Ekwalanga M, Delaporte E, and Peeters M. High diversity of HIV-1 subtype F strains in Central Africa. Virology. 1999. 259: 99-109

Treurnicht FK, Smith T-L, Engelbrecht S, Claassen M, Robson BA, Zeier M, and van Rensburg EJ. Genotypic and Phenotypic Analysis of the Env Gene From South African HIV-1 Subtype B and C Isolates. Journal of Medical Virology. 2002. 68: 141-146

UNAIDS 2008. 2008 Report on the Global AIDS epidemic. - http://www.unaids.org/en/

Uchiyama T, Yodoi J, Sagawa K, Takatsuki K, and Uchino H. Adult T-cell leukemia: clinical and hematologic features of 16 cases. 1977. Blood. 50: 481-492

Vandamme A-M, Salemi M, van Brussel M, Liu HF, Van Laethem K, Van Ranst M, Michels L, Desmyter J, and Boubau P. African Origin of Human T-Lymphotropic Virus Type 2 (HTLV-2) Supported by a Potential New HTLV-2d Subtype in Congolese *Bambuti Efe* Pygmies. Journal of Virology. 1998. 72: 4327-4340

van Harmelen JH, van der Ryst E, Loubser AS, York D, Madurai S, Lyons S, Wood R, and Williamson C. A Predominantly HIV-1 Subtype C-Restricted Epidemic in South African urban Populations. AIDS Research and Human Retroviruses. 1999. 15: 395-398

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

van Harmelen J, Williamson C, Kim B, Morris L, Carr J, Karim SS, and McCutchan F. Characterization of full-length HIV type 1 subtype C sequences from South Africa. AIDS Research and Human Retroviruses. 2001. 17: 1527-1531

van Rossum AMC, Fraaii PLA, and de Groot R. Efficacy of highly active antiretroviral therapy in HIV-1 infected children. The lancet Infectious Diseases. 2002. 2: 93-102

Vidal N, Peeters M, Mulanga-Kabeya C, Nzilambi N, Robertson D, Ilunga W, Sema H, Tshimanga K, Bongo B, and Delaporte E. Unprecedented Degree of Human Immunodeficiency Virus Type 1 (HIV-1) Group M Genetic Diversity in the Democratic Republic of Congo Suggests that the HIV-1 Pandemic Originated in Central Africa. Journal of Virology. 2000. 74: 10498-10507

Wyatt R and Sodroski J. The HIV-1 envelope glycoproteins: fusogens, antigens, and immunogens. Science. 1998. 280(5371): 1884-1888

Wei X, Ghosh SK, Taylor ME, Johnson VA, Emini EA, et al. Viral dynamics in human immunodeficiency virus type 1 infection. Nature. 1995. 373: 117-112

Weller I, Crawford DH, Iliescu V, MacLennan K, Sutherland S, Tedder RS, and Adler MW. Homosexual men in London: Lymphadenopathy, immune status, and Epstein-Barr virus infection. Edited by Selikoff IJ, Teirstein AS, and Hirschman SZ. Annals of the New York Academy of Science. 1984. 437: 248-249

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

WHO / World Health Organization (1983), Acquired Immune Deficiency Syndrome Emergencies. Report of a WHO Meeting, Geneva. 22-25, November

Yang Z, Goldman N, and Friday A. Comparison of models for nucleotide substitution used in maximum likelihood phylogenetic estimation. Molecular Biology and Evolution. 1994. 11: 316-324

Zhang M, Schultz AK , Calef C, Kuiken C, Leitner T, Korber B, Morgenstern B, and Stanke M. jpHMM at GOBICS: a web server to detect genomic recombinations in HIV-1. Nucleic Acid Research. 2006. 34: 463-5

zur Megede J, Engelbrecht S, de Oliveira T, Cassol S, Scriba TJ, Janse van Rensburg E, and Barnett SW. Novel Evolutionary Analyses of Full-Length HIV Type 1 Subtype C Molecular Clones from Cape Town, South Africa. AIDS Research and Human Retroviruses. 2002. 18: 1327-1332

CHAPTER SIX – APPENDICES

Table of Content **Page**

Appendix A

Partial *gag* sequences

>R84_gag
GCATGGGTAAAAGTAGTAGAAGAGAAGGCTTTCAGCCCAGAAGTGATACCCATGTTTTCAGCATTATCA
GAAGGAGCCACCCCACAAGATTTAAACACCATGCTAAACACAGTGGGGGGACATCAAGCAGCCATGCAA
ATGTTAAAAGAGACCATCAATGAGGAAGCTGCAGAATGGGATAGATTGCATCCAGTGCATGCAGGGCCT
ATTGCACCAGGCCAGATGAGAGAACCAAGGGGAAGTGACATAGCAGGAACTACTAGTACCCTTCAGGAA
CAAATAGGATGGATGACAAATAATCCACCTATCCCAGTAGGAGAAATCTATAAAAGATGGATAATCCTA
GGATTAAATAAAATAGTAAGGATGTATAGCCCTGTCAGCATTCTGGACATAAGACAAGGACCAAAGGAA
CCCTTTAGAGACTATGTAGACCGGTTCTATAAAACTCTAAGAGCCGAACAAGCTTCACACAAGGT

>TV86_gag
TGGGTAAAAGTAATAGAGGAGAAGGCTTTCAGCCCAGAAGTAATACCCATGTTTACAGCATTATCAGAA
GGAGCCACCCCACAAGATTTAAACACCATGTTAAATACAGTGGGGGGACATCAAGCAGCCATGCAAATG
TTAAAAGATACCATCAATGAAGAGGCTGCAGAATGGGATAGGTTACATCCAGTGCAGGCAGGGCCTATT
GCACCAGGCCAGATGAGAGAACCAAGGGGAAGTGACATAGCAGGAACTACTAGTAACCTTCAGGAACAA
ATAGCATGGATGACAGGTAATCCACCTATTCCAGTAGGAGACATCTATAAAAGATGGATAATTATAGGG
TTAAATAAAATAGTAAGAATGTATAGCCCTGTCAGCATTTTGGACATAAAACAAGGGCCAAAGGAACCC
TTTAGAGACTATGTAGATCGGTTCTTTAAAACTTTAAGAGCCGAAC

>TV101_gag
AGGACTTTAAATGCATGGGTAAAAGTAATAGAAGAAAAGGCTTTCAGCCCAGAAGTAATACCCATGTTC
TCAGCATTATCAGAAGGAGCCACCCCACAAGATCTAAATACAATGCTGAACGTAGTGGGGGGACACCAG
GCAGCTATGCAGATGTTAAAAGACACCATCAATGAGGAAGCTGCAGAGTGGGACAGGTTACATCCACAA
CATGCAGGGCCTATTCCACCAGGCCAGATAAGGGAACCAAGGGGAAGTGAYATAGCAGGAACTACCAGT
ACCCCTCAAGAACAATTGCAATGGATGACAGGCAACCCACCTATCCCAGTGGGAGACATCTATAAAAGA
TGGATAATCCTGGGATTAAATAAAATAGTAAGAATGTATAGCCCTGTTAGCATTTTGGATATAAGACAA
GGGCCAAAAGAACCCTTCAGAGACTATGTAGATAGGTTCTTTAAAACTCTTAGAGCTGAGCAAGCTACA
CA

>TV218_gag
AGAACCTTGAATGCATGGGTAAAAGTAGTAGAGGAGAAGGCTTTTAGCCCAGAGATAATACCCATGTTT
ACAGCATTATCAGAAGGAGCCACCCCACAAGRTTTAAACACCATGCTAAATACGGTGGGGGGACATCAA
GCAGCCATGCAGATGTTAAAAGATACAATCAATGAAGAGGCTGCAGAATGGGATAGATTACATCCAGTG
CATGCAGGGCCTATTGCACCAGGCCAAATGAGAGAACCAAGGGGAAGTGACATAGCAGGAACTACTAGT
ACCCTTCAGGAACAAATAGCATGGATGACAAGTAACCCACCTATTCCAGTGGGAGACATCTATAAAAGA
TGGATAATTCTGGGATTAAATAAAATAGTGAGAATGTATAGCCCGGTCAGCATTTTGGACATAAGACAA

GGACCAAAGGAACCCTTTAGAGACTATGTAGATCGGTTCTTTAAAACTCTAAGAGCTGAACAAGCTACA
CA

>TV239_gag
TTGAATGCATGGGTAAAAGTAATAGAAGAAAAGGCTTTCAGCCCAGAAGTAATACCCATGTTCTCAGCA
TTATCAGAAGGAGCCACCCCGCAAGATTTAAATATGATGCTAAACATAGTGGGGGGACACCAGGCAGCT
ATGCAAAAGTTAAAAGATACCATCAATGAGGAAGCTATAGAATGGGACAGGACACATCCAGTACATGCA
GGGCCTATCCCACCAGGCCAGATGAGAGAACCAAGTGGAAGTGATATAGCAGGAACTACTAGTACCCTT
CAAGAACAGATAGGATGGATGACAAGTAACCCACCTATCCCAGTGGGAGACATCTATAAAAGGTGGATA
ATTCTGGGATTAAATAAAATAGTAAGAATGTATAGCCCTGTCAGCATTTTGGACATAAGACAAGGGCCA
AAAGAACCCTTCAGAGACTATGTAGATAGGTTCTTTAAAGCTCTCAGAGC

>TV314_gag
TTGAATGCATGGGTAAAAGTAATAGAAGAAAAGGCTTTCAGCCCTGAAGTAATACCCATGTTCTCAGCA
TTATCAGAAGGAGCCACCCCACAAGATTTAAATATGATGCTGAACATAGTGGGGGGACACCAGGCAGCT
ATGCAAATAAGTAACAGTGATACCATCAATGAGGAAGCTGCAGAATGGGATAGGCTACATCCAGTACAT
GCAGGGCCAGTTGCACCAGGCCAGATGAGAGAACCAAGTGGAAGTGATATAGCAGGAACTACTAGTACC
CCTCAAGAACAAATAGCATGGATGACAGGCAACCCACCTATCCCAGTGGGAGACATCTATAAAAGATGG
ATAATCCTAGGGTTAAATAAAATAGTAAGAATGTATAGCCCTGTTAGCATTTTGGATATAAAACAAGGG
CCAAAAGAACCCTTCAGAGACTATGTAGATAGGTTCTTTAAAACTCTCAGAGCCGA

>TV412_gag
CTTTGAATGCATGGGTAAAAGTAATAGAAGAAAAGGCTTTCAGCCCAGAAGTAATACCCATGTTCTCAG
CATTATCAGAAGGAGCCACCCCACAAGATTTAAATATGATGCTGAACATAGTGGGGGGACACCAGGCAG
CCATGCAAAAGTTACCAAGATACCATTAATGAGGAAGCTGCAGAATGGGACAGAGTACATCCAGTACAT
GCAGGGCCTATTCCACCAGGCCAGATGAGAGAACCAAGAGGAAGTGACATAGCAGGAACTACTAGTACC
ATTCAAGAACAAATAGGATGGATGACAAGCAACCCACCTGTCCCAGTGGGAGAAATCTATAAAAGATGG
ATAATCCTGGGATTAAATAAAATAGTAAGAATGTATAGCCCTGTTAGCATTTTGGATATAAAACAAGGG
CCAAAAGAACCCTTCAGAGATTATGTAGATAGGTTCTTTAAAACTCTCAGAGCCGAACAAGCTTCACAA
AAGGTTCTA

>TV441_gag
GCATGGGTAAAAGTAGTAGAGGAGAAGGCTTTTAGCCCAGAGGTAATACCCATGTTTACAGCATTATCA
GAAGGAGCCACCCCACAAGATTTAAACTCCATGCTAAATGCGGTGGGGGGGCATCAAGCAGCCATGCAA
ATGTTAAAAGATACCATCAATGAAGAGGCTGCAGAATGGGATAGATTACATCCAGTACATGCAGGGCCC
ATTGCACCAGGCCAAATGAGAGAACCAAGGGGAAGTGACATAGCAGGAACTACTAGTACCCTTCAGGAA
CAAGTAGCATGGATGACAAGTAACCCACCTATTCCAGTGGGAGACATCTATAAAAGATGGATAATTCTG
GGGTTAAATAAAATAGTAAGAATGTATAGCCCTGTCAGTATTTTGGACATAAGACAAGGGCCAAAGGAA
CCCTTTAGAGACTATGTAGATCGGTTCTTTAAAACTTTAAGAGC

>TV480_gag
GCATGGGTAAAAGTAGTAGAGGAGAAGGCCTTTAGCCCAGAGGTAATACCCATGTTTACAGCATTATCA
GAAGGAGCCACCCCCTCTGATTTAAACTCCATGTTAAATGCGGTGGGGGGACATCAAGCAGCCATGCAA
AAGTTAAAAGATACCATCAATGAAGAGGCTGCAGAATGGGATAGATTACATCCAGTACATGCGGGGCCT
GTTGCACCAGGCCAAATGAGAGAACCAAGGGGAAGTGACATAGCAGGAACTACTAGTACCCTTCAGGAA
CAAATAGCATGGAGTACAGCTAACCCAGCTATTCCAGTGGGAGAAATCTATAAAAGATGGATAATTCTG
GGGCTAAATAAAATAGTGAGAATGTATAGCCCTGTCAGCATTTTGGACATTAGACAAGGGCCAAAGGAA
CCCTTTAGAGACTATGTAGATCGGTTCTTTAAAACTTTAAGAGCCGAACAAGCTTCACA

>TV515_gag
TTTGAATGCATGGGTAAAGGTGATAGAAGAGAAGGCTTTTAGCCCAGAAGTGATACCCATGTTCTCAGC
ATTATCAGAAGGGGCCACCCCACAAGATTTAAACACCATGCTAAATACAGTAGGAGGACATCAAGCAGC
CATGCAAATTTTAAAAGACACCATCAATGAGGAAGCTGCAGAATAGGACAGAGTACATCCACCACAGGC
AAGGCCTCACCCACCAGGCCAGATAAAGGAACCTAGAGGAAGTGACATAGCTGGAACTACTAGTACCCT
TCATGAACAGATACAATAGATGACAAGCAACCCACCTATCCCAGTAAGAGACATCTATAAAAGATAGAT
CATCCTAAGATTAAATAAAATAGTAAGAATGTATAGCCCTGTCAGCATTCTGGACATAAAACAAAGGCC
AAAGGAACCCTTTAGAGACTATGTAGATAGGTTCTTCAAAACTCTAAGAGCCGAACAA

Partial *pol* sequences

>R84_pol
AGTAGATAAATTAGTCAGTGCTGGAATCAGGAGAGTACTATTTTTAGATGGGATAGATAAGGCCCAAGA
AGAACATGAGAAATATCACAGTAATTGGAGAGCAATGGCTAGTGATTTTAACCTGCCACCTATAGTAGC
AAAAGAGATAGTAGCCAGCTGTGATAAATGTCAGTTAAAAGGAGAAGCCATACATGGACAAGTAGACTG
TAGTCCAGGAATATGGCAACTAGATTGTACACATTTAGAAGGAAAAGTTATCCTGGTAGCAGTTCATGT
AGCCAGTGGATATATAGAAGCAGAAGTTATTCCAGCAGAGACAGGGCAGGAAACAGCATACTTTCTCTT
AAAATTAGCAGGAAGATGGCCAGTAAAAACAATACATACAGACAATGGCAGCAATTTCACCAGTACTAC
GGTTAAGGCCCCCTGTTGGTGGGCGGGGATCAAGCAGGAATTTGGCATTCCCTACAATCCCCAAAGTCA
AGGAGTAGTAGAATCTATGAATAAAGAATTAAAGAAAATTATAGGACAGGTAAGAGATCAGGCTGAACA
TCTTAAGACAGCAGTACAAATGGCAGTATTCATCCACAATTTTAAAAGAAAAGGGGGGATTGGGGGGTA
CAGTGCAGGGGAAAGAATAGTAGACATAATAGCAACAGACATACAAACTAAAGAATTACAAAAACAAAT
TACAAAAATTCAAAATTTTCGGGTTTATTACAGGGACAGCAGAGAGCCACTTTGGAAAGGACCAGCAAA
GCTTCTCTGGAAAGGTGAAGGGGCAGTAGTAATACAAGATAATAGTGACATAAAAGTAGTGCCAAGAAG
AAAAGTAAAAATCATTAGGGATTATGGAAAACAGATGGCAGGTGATGATTGTGTGGCAAGTAGACAGGA
TGAGGATTAGAACATGGAACAGCTTAGTAAAACACCATATGTAT

>TV86_pol
GTCAGTAAAGGAATCAGGAAAGTGCTGTTTCTAGATGGAATAGATAAGGCTCAAGAAGAGCATGAAAAG
TATCACAGCAATTGGAGAGCAATGGCTAGTGAGTTTAATATGCCACCCGTAGTAGCAAAAGAAATAGTA
GCTAGCTGTGATAAATGTCAGCTAAAAGGGGAGGCCATGCATGGACAAGTAGACTGTAGTCCAGGAATA

TGGCAATTAGATTGTACACATTTAGAAGGAAAAGTCATCCTGGTAGCAGTCCATGTAGCTAGTGGCTAC
ATAGAAGCAGAGGTTATCCCAGCAGAAACAGGACAAGAAACAGCATACTATATACTAAAATTAGCAGGA
AGATGGCCAGTCAAAGTAATACATACAGACAATGGCAGTAATTTTACCAGTACTCCAGTTAAGGCAACC
TGTTGGTGGGCAGGTATCCAACAGGAATTTGGAATTCCCTACAATCCCCAAAGTCAGGGAGTAGTAGAA
TCCATGAATAAAGAATTAAAGAAAATAATAGGACAAGTAAGAGATCAAGCTGAGCACCTTAAGACAGCA
GTACAAATGGCAGTATTCATTCACAATTTTAAAAGAAAAGGGGGGATTGGGGGGTACAGTGCAGGGGAA
AGAATAATAGACATAATAGCAACAGACATACAAACTAAAGAATTACAAAAACAAATTACAAAAATTCAA
AATTTTCGGGTTTATTACAGAGACAGCAGAGACCCTATTTGGAAAGGACCAGCCAAACTACTCTGGAAA
GGTGAAGGGGCAGTAGTAATACAAGATAACAGTGACATAAAGGTAGTACCAAGGAGGAAAGCAAAAATC
ATTAGGGATTATGGAAAACAGATGGCAGGTGATGATTGTGTGGCAGGTAGACAGGATGAAGATTAGAAC
ATGGAATAGCTTAGTAAAACACCATA

>TV101_pol
AACAARTAGATAAATTAGTTAGCTCTGGGATCAGGAAGGTACTATTTTTAGATGGGATAGACAAGGCTC
AAGAAGACCATGAGAGRTATCACAGCAATTGGAGAACAATGGCTAGTGATTTTAATCTRCCACCTATAG
TAGCAAAGGAAATAGTAGCCAGCTGTGATAAATGTCAGCTAAAAGGAGAAGCCATGCATGGACAAGTAG
ACTGTAGTCCAGGAATATGGCAATTAGACTGTACACATTTAGAGGGAAAAGTTATCCTGGTAGCAGTCC
ATGTAGCCAGTGGCTATATAGAAGCAGAAGTTATTCCAGCAGAAACAGGACAGGAAACAGCCTATTTTA
TCTTAAAATTAGCAGGAAGATGGCCAGTAAAAGTAGTACATACAGACAATGGCAGCAATTTCACTAGCA
CTGCAGTTAAAGCAGCCTGTTGGTGGGCAAATGTCCAACAGGAATTTGGGATTCCCTACAATCCCCAAA
GTCAGGGAGTAGTAGAATCTATGAATAAAGAATTAAAGAAAATYATAGGGCAGGTAAGAGATCAAGCTG
AACATCTTAAAACAGCAGTACAAATGGCAGTGTTCATTCACAATTTTAAAAGAAAAGGGGGGATTGGGG
GATACAGTGCAGGGGAAAGAATAATAGACATAATAGCAACAGACATACAAACTAAAGAATTACAAAAAC
AYATTTCAAAAATTCAAAATTTTCGGGTTTATTACAGGGACAGCAGAGATCCAATTTGGAAAGGACCAG
CAAARCTKCTCTGGAAAGGTGAAGGGGCAGTGGTAATACAAGACARTAGTGAAATAAAGGTAGTGCCAA
GAAGGAAAGCAAAGATCATTAGGGATTATGGAAAACAGATGGCAGGTGATGATTGTGTGGCAGGTAGAC
AGGATGAGGATTAGAACATGGCACAGTTTAATAAAACACCATATGTAT

>TV218_pol
AACAAGTAGATAAATTAGTAAGTAAAGGGATCAGRAAAGTGCTGTTTCTAGATGGAATAGATAAGGCTC
AAGAAGATCATGAAAGATATCACAGCAATTGGAGRGCAATGGCTAGTGAGTTTAATCTGCCACCCATAG
TAGCAAAAGAAATAGTAGCTAGCTGTGATAAATGTCAGYTAAAAGGGGAAGCCATACATGGACAGGTAG
ACTGTAGCCCGGGGATATGGCAATTAGACTGTACACATTTAGAAGGAAAAATYATCCTGGTAGCAGTCC
ATGTAGCCAGTGGCTACATAGAAGCAGAGGTTATCCCAGCAGAAACAGGACAAGAAACAGCATACTTTA
TACTAAAATTAGCAGGAAGATGGCCAGTCAAAGTAATACATACAGACAATGGCAGTAATTTCACCAGTG
CTGCAGTTAAGGCAGCCTGTTGGTGGGCAGGTATCCAACAGGAATTTGGGATTCCCTACAATCCCCAAA
GTCAGGGAGTGGTAGAATCCATGAATAAAGAATTAAAGAAAATCATAGGGCAGGTAAGAGAYCAAGCTG
AGCACCTTAAGACAGCAGTGCAAATGGCAGTATTCATTCACAATTTTAAAAGAAAAGGGGGGATTGGGG
GGTACAGTGCAGGGGAAAGAATAATAGACATAATAGCAACAGACATACAAACTAAAGAATTACAAAAAC
AAATTACAAAAATTCAAAATTTTCGGGTTTATTACAGAGACAGCAGAGACCCTATTTGGAAAGGACCAG

CCAAACTACTCTGGAAAGGTGAAGGAGCAGTAGTAATACAAGATAACAGTGACATCAAGGTAGTACCAA
GGAGGAAAGCAAAAATCATTAA
GGACTATGGAAAACAGATGGCAGGTGCTGATTGTGTGGCAGGTAGACAGGATGAAGATTAGAACATGGA
ATAGTTTGGTAAAGCACCATATGCAT

>TV239_pol
AAGTAGATAAATTAGTCAGTTCTGGAATCAGGAAGGTGCTATTTTTAGATGGGATAGATAAGGCTCAAG
AAGAGCATGAAAGGTATCACAGCAATTGGAGAGCAATGGCTAGTGATTTCAACCTGCCACCAGTAGTAG
CAAAGGAAATAGTAGCCAGCTGTGATAAATGTCAACTAAAAGGGGAAGCCATGCATGGACAAGTAGACT
GTAGTCCAGGAATATGGCAAGTAGATTGCACACATCTAGAAGGAAAAGTAATCATAGTAGCAGTCCATG
TAGCCAGTGGCTATATAGAGGCAGAAGTTATCCCAGCAGAAACAGGACAGGAGGCAGCATACTTTCTGT
TAAAATTAGCAGCAAGATGGCCAGTAAAGGTAATACACACAGACAATGGCAGTAATTTCACCAGTGCCG
CAGTTAAAGCAGCCTGTTGGTGGGCAAATATCCAACAGGAATTTGGAATTCCCTACAATCCCCAAAGTC
AAGGAGTAGTGGAATCTATGAATAAAGAATTAAAGAAAATCATAGGGCAGGTAAGAGAGCAAGCTGAGC
ACCTTAAGACAGCAGTACAAATGGCAGTATTCATTCACAATTTTAAAAGAAAAGGGGGGATTGGGGGGT
ACAGTGCAGGGGAAAGAATAATAGACATAATAGCAACAGACATACRAACTAAAGAATTACAAAAACAAA
TTACAAAAATTCAAAATTTCCGGGTTTATTACAGGGACAGCAGAGACCCAATTTGGAAAGGACCAGCAA
GACTACTCTGGAAAGGTGAAGGGGCAGTAGTAATACAGGACAATAGTGATATAAAGGTAGTACCAAGAA
GAAAAGCAAAAATCATTAGGGACTATGGAAAACAGATGGCAGGTGATGATTGTGTGGCAGGTAGACAGG
ATGAGGATTAGAACATGGCACAGCTTAGTAAAACACCATATGTAT

>TV314_pol
AGTAGATAAATTAGTCAGTTCTGGAATCAGGAAGGTGCTATTTCTAGATGGGATAGATAAGGCTCAAGA
AGAACATGAAAGATATCACAGCAACTGGAGAGCTATGGCTAGTGATTTTAATCTGCCACCTATAGTAGC
AAAGGAGATAGTAGCCAGCTGTGATAAATGCCAGCTAAAAGGGGAAGCCATGCATGGACAAGTAGACTG
CAGTCCAGGAATATGGCAATTAGACTGCACACATCTAGAAGGAAAAGTAATTCTGGTAGCAGTCCATGT
AGCCAGTGGCTATATAGAAGCAGAAGTTATCCCAGCAGAAACAGGACAAGAGACAGCATACTTTCTATT
AAAATTAGCAGGAAGATGGCCAGTAAAAACAGTACACACAGACAATGGCAGCAATTTCACCAGTGCTGC
AGTTAAAGCAGCCTGTTGGTGGGCAGGTATCAAACAGGAATTTGGAATTCCCTACAATCCCCAAAGTCA
AGGAGTAGTGGAATCTATGAATAAGGAATTAAAGAAAATCATAGGGCAGGTAAGAGAGCAAGCTGAACA
CCTTAAGACAGCAGTACAAATGGCAGTATTCATTCACAATTTTAAAAGAAAAGGGGGGATTGGGGGGTA
CAGTGCAGGGGAAAGAATAATAGACATAATAGCAACAGACATACAAACTAAAGAATTACAAAAACAAAT
TACAAAAATTCAAAAATTTCGGGTTTATTACAGGGACAGCAGAGATCCAATTTGGAAAGGACCAGCAAA
ACTACTCTGGAAAGGTGAAGGGGCAGTGGTAATACAGGACAATAGTGATATCAAAGTAGTACCAAGAAG
AAAAGCAAAGATCCTTAGGGAT
TATGGACAACAGATGGCAGGTGATGATTGTGTGGCAGGTAGACAGGATGAGGATTAGAACATGGCACAG
CTTAGTAAAACACCATATGTAT

>TV340_pol
AGTAGATAAATTAGTCAGTAGTGGAATCAGAAAAGTACTATTTCTAGATGGCATAGATAAAGCCCAAGA
AGAGCATGAAAGATATCACAGCAATTGGAGGGCAATGGCTAATGACTTTAATCTGCCACCTATAGTAGC
AAAAGAAATAGTGGCCAGCTGTGATAAGTGCCAGCTAAAAGGGGAAGCCATGCATGGACAAGTAGACTG
TAGTCCAGGAATATGGCAATTAGATTGTACACATTTAGAAGGAAAAATTATCCTGGTAGCAGTCCATGT
AGCCAGTGGCTATATAGAAGCAGAAGTTATCCCAGCAGAAACAGGACAGGAAACAGCATACTTTATATT
AAAATTAGCAGGAAGATGGCCAGTAAAAGTAATACACACAGATAATGGCAGCAATTTCACCAGCAGTGC
AGTAAAGGCAGCATGTTGGTGGGCAAATATCACACAAGAATTTGGAATTCCCTACAATCCCCAAAGCCA
AGGAGTAGTAGAGTCTATGAATAAAGAATTAAAGAAAATTATAGGACAGGTCAGAGATCAAGCTGAACA
CCTTAAGACAGCAGTACAGATGGCAGTATTCATTCATAATTTTAAAAGAAAAGGGGGGATTGGGGGGTA
CAGTGCAGGGGAAAGAATAGTAGACATAATAGCATCAGATATACAAACTAAAGAACTACAAAAACAAAT
TATAAAAATTCAAAATTTTCGGGTTTATTACAGGGACAGCAGAGACCCAATTTGGAAAGGACCAGCAAA
ACTACTCTGGAAAGGTGAAGGGGCAGTAGTAATACAGGACAATAGTGATATAAAAGTAGTACCAAGAAG
AAAAGCAAAGATCATTAGGGATTATGGAAAACAGATGGCAGGTGATGATTGTGTGGCAGGTAGACAGGA
TGAGGATTAATACATGGAAAAGCTTAGTAAAACACCATATG

>TV412_pol
AGTAGATAAATTAGTCAGTAATGGAATCAGGAAGATACTATTTTTAGATGGGATAGATAAGGCTCAAGA
AGAACATGAAAGATATCATAGCAATTGGAGAGCAATGGCTAATGATTTTAACCTGCCACCTGTGGTAGC
AAAGGAAATAGTAGCCAGCTGTGATAAATGTCAGCTAAAAGGGGAAGCCATGCATGGACAGGTAGACTG
TAGTCCAGGAATATGGCAATTAGATTGCACACATCTAGAAGGAAAAGTAATTCTGGTAGCAGTTCATGT
AGCCAGTGGCTATATAGAAGCAGAAGTTATCCCAGCAGAAACAGGACAGGAAACAGCATACTTTCTGCT
AAAATTAGCAGGGAGGTGGCCAGTAAAAGTAGTTCACACAGACAATGGCAGCAATTTCACCAGTGCTGC
AGTTAAAGCAGCCTGTTGGTGGGCAAATATCCAACAGGAATTTGGGATTCCCTACAATCCCCAAAGTCA
AGGAGTAGTAGAATCTATGAATAAAGAATTAAAGAAAATTATAGGACAGGTAAGAGATCAAGCTGAACA
TCTTAAGACAGCAGTACAAATGGCAGTATTCATTCACAATTTTAAAAGAAAAGGGGGGATTGGGGGGTA
CAGTGCAGGGGAAAGAATAATAGACATAATAGCAACAGACATACAAACTAAAGAACTACAAAAACAAAT
TACAAAAATTCAAAATTTTCGGGTTTATTACAGGGACAGCAGAGATCCAGTTTGGAAAGGACCAGCAAA
GCTTCTCTGGAAAGGTGAAGGGGCAGTAGTAATACAAGACAATAGTGAAATAAAGGTAGTACCAAGAAG
AAAAGCAAAGATCATTAGGGATTATGGAAAACAGATGGCAGGTGATGATTGTGTGGCAGGTAGACAGGA
TGAGGATTAGAACATGGAACAGCTTAGTAAAACACCATATGTAT

>TV441_pol
CAAGTAGATAAATTAGTCAGTAGTGGAATCAGGAAAGTATTGTTCCTAGATGGAATAGATAAGGCTCAA
GAAGAGCATGAAAAATATCACAATAATTGGAGAGCAATGGCTGCTGATTTTAATCTGCCACCTGTAGTA
GCAAAAGAAATAGTAGCTAGCTGTGATAAGTGTCAGCTAAAAGGGGAAGCCATGCATGGACAAGTAGAC
TGTAGTCCAGGAATATGGCAATTAGATTGCACTCATCTAGAAGGGAAAGTTATCCTGGTAGCAGTCCAT
GTAGCCAGTGGCTATATGGAAGCAGAAGTTATCCCAGCAGAAACAGGACAGGAGACAGCATACTTTATA
TTAAAGCTAGCAGGAAGGTGGCCAGTAAGAGTAATACATACAGACAACGGCCCCAATTTCACCAGTGCA
GCAGTTAAGGCAGCCTGTTGGTGGGCAAATATCCAACAGGAATTTGGGATTCCCTACAACCCCCAAAGT
CAAGGAGTAGTGGAATCTATGAATAAAGAATTAAAGAAAATCATAGGGCAGGTAAGAGATCAAGCTGAG

CACYTTAAGACAGCAGTACAAATGGCAGTATTCATTCACAATTTTAAAAGAAAAGGGGGGATTGGGGGG
TACAGTGCAGGGGAAAGAATAATAKACATAATAGCAACAGACATACAAACTAAAGAATTACAAAAACAA
ATTATAAAAATTCAAAATTTTCGGGTTTATTACAGAGACAGCAGAGACCCTATTTGGAAAGGACCAGCC
AAACTACTCTGGAAAGGTGAAGGGGCAGTAGTAATACAAGACAACAGTGACATAAAGGTAGTACCAAGG
AGGAAAGTAAAAATCATTAAGGACTATGGAAAACAGATGGCAGGTGCTGATTGTGTGGCAGGTAGACAG
GATGAGATTAGAACATGGAATAGCTTAGTAAAACACCATATGTAA

>TV480_pol
AACAAGTAGATAAATTAGTCAGTAATGGAATCAGGAAGGTGCTGTTTCTAGATGGAATAGATAAGGCTC
AAGAAGAGCATGAAAAATATCACAGCAATTGGAGAGCAATGGCTAGTGAGTTTAACCTGCCACCCATAG
TAGCAAAAGAAATAGTAGCTAGCTGTGATAAATGTCAGCTAAAAGGGGAAGCCATACATGGACAAGTAG
ACTGTAGTCCAGGGATATGGCAATTAGATTGTACACATTTAGAAGGAAAAGTCATCCTGGTAGCAGTCC
ATGTAGCCAGTGGCTACATAGAGGCAGAGGTTATCCCAGCAGAAACAGGACAAGAAACAGCATACTATA
TACTGAAATTAGCAGGAAGATGGCCAGTCAAAGTGATACATACAGACAATGGCAGTAATTTCACCAGTG
CTGCAGTTAAGGCAGCCTGTTGGTGGGCAGGTATCCAACAGGAATTCGGGATTCCCTACAATCCCCAAA
GTCAGGGAGTAGTAGAATCCATGAATAAAGAATTAAAGAAAATCATAGGGCAGGTAAGAGAGCAAGCAG
AGCACCTTAAGACAGCAGTACAAATGGCAGTGTTCATTCACAATTTTAAAAGAAGAGGGGGGATTGGGG
GGTACAGTGCAGGGGAAAGAATAATAGACATAATAGCAACAGACATACAAACTAAAGAATTACAAAAAC
AAATTATAAAAATTCAAAATTTCCGGGTTTATTACAGAGACAGCAGAGACCCTATTTGGAAAGGACCAG
CCAAACTACTCTGGAAAGGTGAAGGGGCAGTAGTAATACAAGATAACAGTGACATAAAGGTAGTACCAA
GGAGGAAAGTAAAAATCATTAAGGACTATGGAAAACAGATGGCAGGTGCTGATTGTGTGGCAGGTAGAC
AGGATGAGGATTAGAACTTGGAATAGCTTAGTAAAACACCATATGTAT

>TV515_pol
AGTAGATAAATTAGTCAGTGCTGGAATCAGAAAAGTACTGTTTTTAGATGGGATAGATAAGGCACAAGC
AGAACATGAAAAATATCACAACAATTGGAGAGCAATGGCTAGTGACTTTAATCTGCCCCCTGTAATAGC
AAAAGAAATAGTAGCTAGCTGTGATAAATGTCAGCTAAAAGGGGAAGCTATGCATGGACAAGTAGACTG
TAGCCCAGGGATATGGCAATTAGATTGCACACATTTAGAAGGAAAAATTATCCTAGTAGCTGTTCATGT
AGCTAGTGGATATATAGAAGCAGAAGTCATTCCAGCAGAAACAGGACAGGAAACTGCCTACTACATACT
AAAGTTAGCAGGAAGATGGCCAGTAAAAATAATACATACAGACAATGGCAGCAATTTCACCAGTACTGC
GGTTAAGGCAGCCTGTTGGTGGGCAGGTATCCAGCAAGAATTTGGAATTCCCTACAATCCCCAAAGTCA
AGGAGTAGTAGAATCTATGAATAAAGAATTAAAGAAAATCATAGGACAAGTAAGAGATCAAGCTGAACA
TCTTAAAACAGCAGTACAAATGGCAGTATTCATTCATAATTTTAAAAGAAAAGGGGGGATTGGGGGGTA
CAGTGCAGGGGAAAGAATAATAGACATAATATCAACAGACATACAAACTAAAGAACTACAAAAACAAAT
TATAAAAATTCAAAATTTCCGGGTTTATTACAGGGACAGCAGAGACCCAGTTTGGAAAGGACCAGCAAA
GCTACTGTGGAAAGGTGAAGGGGCAGTAGTCATACAAGACAATAGTGAAATAAAAGTAGTGCCAAGAAG
GAAAGTAAAGATCATTAGGGATTATGGAAAACAGATGGCAGGTGATGATTGTGTGGCAAGTAGACAGGA
TGAGGATTAACACATGGAAAAGCTTAGTAAAACACCATATG

Partial *env* sequences

>R84_env
CAAAACAATTTGCTGAAGGCTATTGAGGCGCAACAGCACCTGTTGCAACTCACAGTCTGGGGCATCAAG
CAGCTCCAGGCAAGAGTCCTGGCTATAGAAAGATACCTAAAGGATCAACAGCTCCTGGGGATTTGGGGT
TGCTCTGGAAAACTCATTTGCACCACTGCTGTGCCTTGGAATGCTAGTTGGAGTAATAAATCTCAGGAT
AAGATTTGGCATAACATGACCTGGATGGAGTGGGAAAGAGAAATTAGCAATTACACAAGCTTAATATAC
ACCTTAATTGAAGAATCGCAGAACCAACAAGAAAAGAATGAACAAGAATTGTTGGAATTGGATCAATGG
GCAAATTTGTGGAATTGGTTTGACATAACAAAATGGCTGTGGTATATAAAAATATTYATAATGATA

>TV_86_env
CAAAGCAATTTGCTGAAGGCTATAGAGGCGCAACAGCATATGTTGCAACTCACGGTCTGGGGCATTAAG
CAGCTCCAGGCAAGAGTCCTGGCTATAGAAAGATACCTAAAGGATCAACAGCTCCTAGGGATTTGGGGC
TGCTCTGGAAAACTCATCTGCACCACTAATGTGCCTTGGAACACAAGTTGGAGTAGTAAATCTGAAGAT
GAGATTTGGAATAACATGACTTGGATGCAGTGGGATAGAGAAATTAGTAATTACACAGGCATAATATAC
CATTTGCTTGAAGACTCGCAAAACCAGCAGGAAAGGAATGAAAAAGACTTGTTTGCATTGGACAGTTGG
AAAAATCTGTGGAATTGGTTTGACATATCAAAATGGCTGTGGTATATAAAAATATTYATAATGATA

>TV_101_env
CTGGCATAGTGCAACAGCAAAGCAATTTGCTGAGGGCTATAGAGGCTCAACAGCATCTGTTGAAGCTCA
CGGTCTGGGGCATTAAACAGCTCCAGGCAAGAGTCCTGGCTSTGGAAAGATACCTAAAGGATCAACAGC
TCCTAGGAATTTGGGGCTGCTCTGGAAAACTCATCTGCACCACTACTGTGCCCTGGAACTCTAGTTGGA
GTAAYAAATCCCAGAATGAAATATGGGACAACATGACCTGGATGCAATGGGATAAAGAAATTAGCAATT
ACACACAGATAATATATAGTCTAATTGAAGAATCACAAAACCAGCAGGAAAAGAATGAACAAGAGTTAC
TGGCATTGGACAAGTGGGCAAATCTGTGGAATTGGTTTGATATATCAAATTGGCTGTGGTACATAAAGA
TATTTATAATGATAGTAGGAGGCTTAATAGGATT

>TV_218_env
CTGGCATAGTGCGGCAGCAAAGCAATTTGCTGCAGGCTATAGAGGCTCAACAACATCTGTTGARACTCA
CAGTCTGGGGCATTAAACAGCTCCAGGCAAGAGTCCTGGCTCTGGAAAGATACCTACAGGATCAACAGC
TCCTAGGAATTTGGGGCTGCTCTGGAAAACTTATCTGCACCACTACTGTGCCTTGGAACTCTAGTTGGA
GTAACAAATCTTATRAKGACATTTGGGRAAACATGACCTGGTTGCAGTGGGATAGAGAAATTAGCAATT
ACACAAACACAATATACAGGCTACTTGAGGAGTCACAGAACCAGCAGGAAATTAATGAACAAGATTTAT
TGGCCTTGGACAAGTGGGCAGGTCTGTGGAGTTGGTTTAGYATATCAAATTGGCTGTGGTATATAAAAA
TRTTTATAATGATAGTAGGAGGCTT

>TV_239_env
TTGCTGAGGGCTATAGAGGCTCAACAGCATCTGTTGAAACTCACAGTCTGGGGCATTAAACAGCTCCAG
GCAAGAATCCTGGCTGTGGAAAGATACCTAAAGGATCAACAGCTTCTAGGAATTTGGGGCTGCTCAGGA
AAGTTAATCTGCACCACTGCTGTGCCCTGGAACTCTAGTTGGAGTAATAAATCTTATAATGAAATATGG
GATAACATGACCTGGATGCAATGGGAAAAGGAAATTGACAATTACACAGGCATAATATATACTCTAATT

GAAGAATCGCAGAACCAACAGGAAAAGAATGAACAAGATTTATTGGCATTGGACAAGTGGGCAAGTCTG
TGGAATTGGTTTGACATATCAAATTGGCTATGGTATATAAAAAT

>TV_314_env
CAAAGCAATTTGCTGAGGGCTATAGAGGCTCAACAGCATCTGTTGAAACTCACAGTCTGGGGCATTAAA
CAGCTCCAGGCAAGAGTCCTGGCTCTAGAGAGATACCTAAGGGATCAACAGCTCCTAGGAATTTGGGGC
TGCTCTGGAAAACTCATTTGCGCCACTAATGTGCCTTGGAACTCTAGTTGGAGTAATAAATCTTATAAT
GAAATATGGGATAACATGACCTGGCTGCAGTGGGATAAAGAAATTGACAATTACACAGAAACAATATAT
AGGCTAATTGAAGAATCGCAAAACCAGCAGGAAAAGAATGAACAAGACTTATTGGCATTGGACAAGTGG
ACAAATCTGTGGAGTTGGTTTGACATATCGAACTGGCTGTGGTATATAAAAATATTYATAATGATA

>TV_340_env
CAGAGCAATCTGCTGAGGGCTATAGAGGCTCAACAGCATTTGTTGAAACTCACAGTCTGGGGCATTAAA
CAGCTCCAGGCAAGAGTCCTAGCTATAGAAAGATACCTAAGGGATCAACAGCTCCTAGGAATCTGGGGA
TGCTCTGGAAAACTCATCTGCCCCACTAATGTGCCCTGGAACTCCAGCTGGAGTAATAAGTCTCAGAGT
GAAATATGGGATAACATGACCTGGGTGCAATGGGATAAAGAAATTAGCAATTACACACAATTAATATAT
GGTCTACTTGAGGAATCGCAGAACCAGCAGGAAAAGAATGAACAGGACTTATTGGCATTGGACAAGTGG
GCAAGCCTGTGGAATTGGTTTGATATATCAAATTGGCTGTGGTACATAAAAATATTYATAATGATA

>TV_412_env
CAAAGCAATTTGCTGAGGGCTATAGAGGCTCAACAACATCTGTTGAAACTCACGGTCTGGGGCATTAAA
CAGCTCCGGGCAAGAGTCCTGGCTGTGGAAAGATACCTAAAGGATCAACAGCTCCTAGGAATTTGGGGC
TGCTCTGGAAAACTCATCTGCACCACTAATGTGCCCTGGAATTCTAGTTGGAGTAATAAATCTCAGGAG
GAGATATGGGGGAACATGACCTGGCTGCAATGGGATAAAGAAGTTAACAATTATACAGAATTAATATAC
TCCCTAATTGAAGAATCGCAGATCCAGCAGGAAAAGAATGAACAAGACTTATTGGCATTGGACAAATGG
GCAAATCTGTGGAGTTGGTTTAGCATATCAAATTGGCTGTGGTATATAAGAATATTTATAATGATA

>TV_441_env
CAAAGCAATTTGCTGCAGGCTATAGAGGCTCAACAACATCTGTTGAAACTCACTGTCTGGGGCATTAAA
CAGCTCCAGGCAAGAGTCCTGGCTCTGGAAAGATACCTAAAGGATCAACAGCTCCTAGGAATTTGGGGC
TGCTCTGGAAAACTTATCTGCACCACTACTGTGCCTTGGAACTCTAGTTGGAGTAATAAATCTTATAAT
GAGATTTGGGATAACATGACTTGGTTGCAGTGGGATAGAGAAATTAGCAATTACACAGAAACAATATAC
AGGCTACTCCAAGACTCACAAATCCAGCAGGAACAGAATGAAAARGAGTTATTGGAATTGGACAAGTGG
GCAAATCTGTGGAATTGGTTTGACATATCAAAGTGGCTATGGTACATAAAAATATTYATAATGATA

>TV_515_env
CAGAACAATCTGCTGAGGGCTATTGAAGCGCAACAGCATCTGTTGCAGCTCACAGTCTGGGGCATTAAA
CAGCTCCAGGCAAGAGTCCTGGCTGTGGAAAGATACCTAAAGGATCAACAGCTCCTAGGGATTTGGGGC
TGCTCTGGAAAACTCATCTGCACCACTAATGTGCCGTGGAACTCTAGTTGGAGTAATAGATCTCTGGAA
GACATTTGGGAAAACATGACCTGGAGGGAGTGGGAAAAAGAGATTGGTAATTACTCAAACATAATATAT

AGGTTAATTGAACAATCGCAGAACCAGCAGGAAATAAATGAAAAAGACTTATTGGCATTGGACAAGTGG
GCAAGTCTGTGGAATTGGTTTGACATAACAAGCTGGCTGTGGTATATAAAAATATTYATAATGATA

Appendix B

Name	Length	Report	Assignment	Support	Genome
R84_gag	479bp	Report	HIV-1 Subtype B	90.0	
TV86_gag	460bp	Report	HIV-1 Subtype C	100.0	
TV101_gag	485bp	Report	HIV-1 Subtype A (A1)	99.0	
TV218_gag	485bp	Report	HIV-1 Subtype C	100.0	
TV239_gag	464bp	Report	HIV-1 Subtype A (A1)	98.0	
TV314_gag	470bp	Report	HIV-1 Subtype A (A1)	100.0	
TV412_gag	492bp	Report	HIV-1 Subtype A (A1)	99.0	
TV441_gag	458bp	Report	HIV-1 Subtype C	100.0	
TV480_gag	473bp	Report	HIV-1 Subtype C	100.0	
TV515_gag	472bp	Report	HIV-1 Subtype F (F1)	98.0	

Figure 6.1: REGA subtyping results of the *gag* p24 sequences. All data are included in the report including the size of each fragment, the subtype and the bootstrap values.

Name	Length	Report	Assignment	Support	Genome
R84_pol	1286bp	Report	Check the bootscan	*NA*	
TV86_pol	923bp	Report	HIV-1 Subtype C	100.0	
TV101_pol	945bp	Report	HIV-1 Subtype A (A1)	79.0	
TV218_pol	945bp	Report	HIV-1 Subtype C	100.0	
TV239_pol	1313bp	Report	Check the bootscan	*NA*	
TV314_pol	1392bp	Report	Check the bootscan	*NA*	
TV340_pol	938bp	Report	Check the bootscan	*NA*	
TV412_pol	1392bp	Report	Check the bootscan	*NA*	
TV441_pol	955bp	Report	Check the bootscan	*NA*	
TV480_pol	1572bp	Report	Check the bootscan	*NA*	
TV515_pol	938bp	Report	HIV-1 Subtype F (F1)	100.0	

Figure 6.2: REGA subtyping results of the *pol - integrase* sequences. The subtype and the bootstrap values are included.

Name	Length	Report	Assignment	Support	Genome
R84_env	411bp	Report	HIV-1 Subtype B	100.0	
TV_86_env	411bp	Report	Check the report	*NA*	
TV_101_env	448bp	Report	HIV-1 Subtype A (A1)	96.0	
TV_218_env	439bp	Report	Check the report	*NA*	
TV_239_env	389bp	Report	Check the report	*NA*	
TV_314_env	411bp	Report	Check the report	*NA*	
TV_340_env	411bp	Report	HIV-1 Subtype A (A1)	98.0	
TV_412_env	411bp	Report	HIV-1 Subtype A (A1)	94.0	
TV_441_env	411bp	Report	Check the report	*NA*	
TV_515_env	411bp	Report	Check the report	*NA*	

Figure 6.3: REGA subtyping results of the *env* gp41 sequences. The subtype and the bootstrap values are included.

Table 6.1: jpHMM subtyping results of gag p24 subgenomic regions.

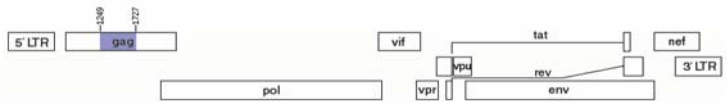
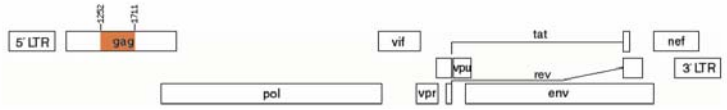
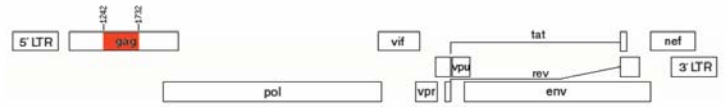
Sample	Subtype	*gag* p24
R84	**B**	
TV86	**C**	
TV101	**A1**	
TV218	**C**	
TV239	**A1**	
TV314	**A1**	
TV412	**A1**	
TV441	**C**	
TV480	**C**	
TV515	**F1**	

Table 6.2: jpHMM subtyping results of *pol - integrase* subgenomic regions.

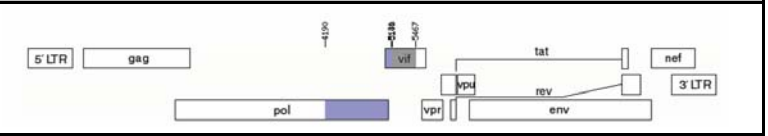
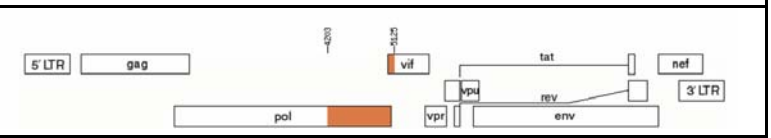
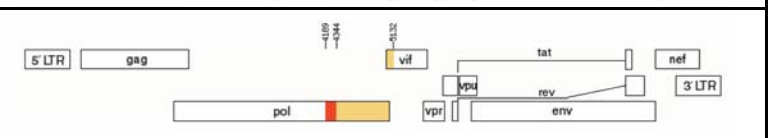
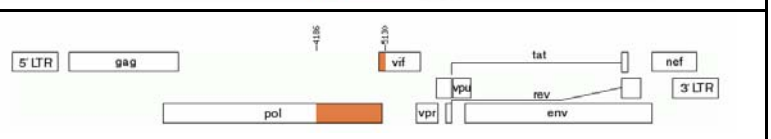
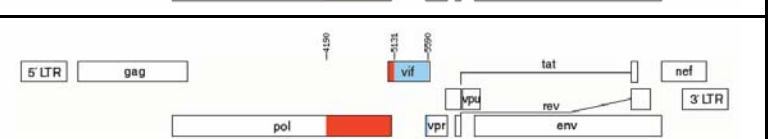
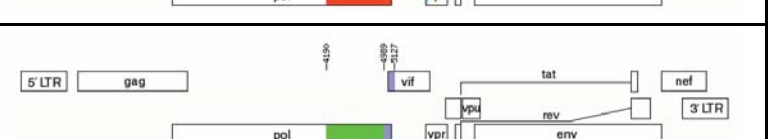
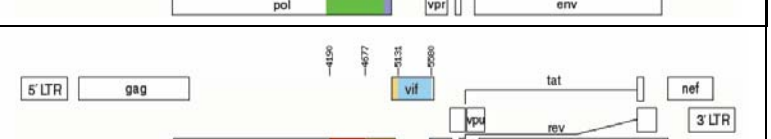
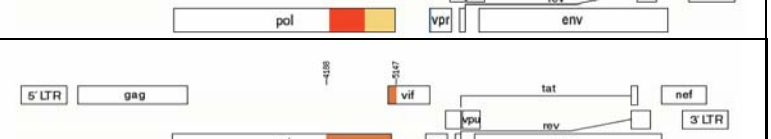
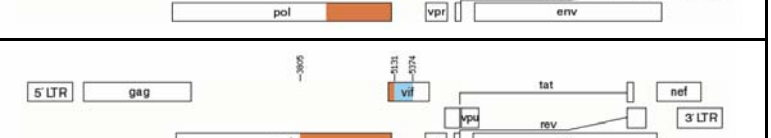
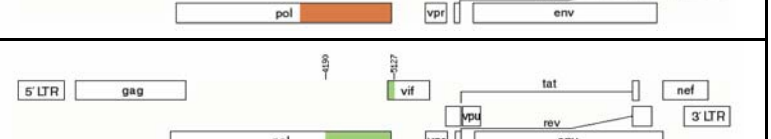
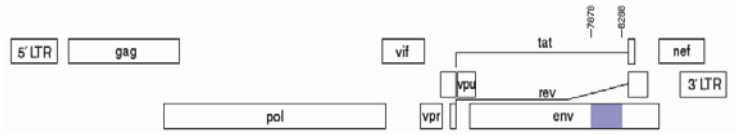
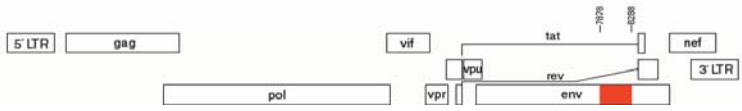
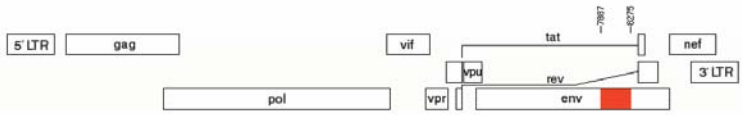
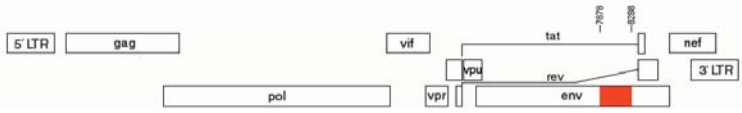
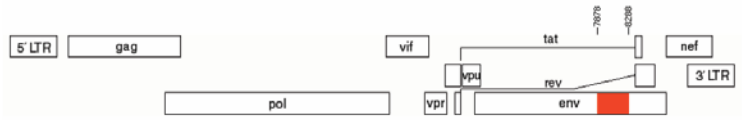
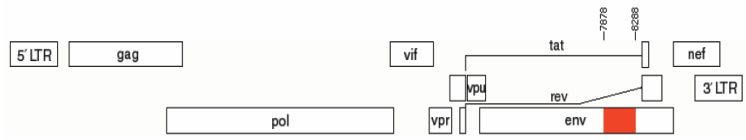
Sample	Subtype	*pol - integrase*
R84	**B**	
TV86	**C**	
TV101	**A1 / D**	
TV218	**C**	
TV239	**A1 / H**	
TV314	**A1 / J**	
TV340	**G / B**	
TV412	**A1 / D / J**	
TV441	**C**	
TV480	**C / J**	
TV515	**F1**	

Table 6.3: jpHMM subtyping results of *env* gp41 subgenomic regions.

Sample	Subtype	*env* gp 41
R84	**B**	
TV86	**C**	
TV101	**A1**	
TV218	**A1**	
TV239	**A1**	
TV314	**A1**	
TV340	**A1**	
TV412	**A1**	
TV441	**A1**	
TV515	**F1**	

Appendix C

Table 6.4: Sequencing primers used for the characterization of R84.

Primer	Oligo Nucleotide Sequence	HXB2 Position	T_A	Ref
9110R	GCAAGAGAGACCCAGTACAG	10166 - 10147	50	Personal Communication, John Hackett
R749 LTR-gag	ATTTTGCACTATAGGGTAAT	1181 - 1200	45	Personal Communication, John Hackett
p24-1	AGYCAAAATTAYCCYATAGT	1174 - 1193	45	Swanson *et al,* 2003
p24-2	AGRACYTTRAAYGCATGGGT	1237 - 1256	50	Swanson *et al,* 2003
GagA	AGAGAACCAAGGGGAAGTGA	1474 - 1493	50	Kemp *et al,* 1989
GagB	TCTCTAAAGGGTTCCTTTGG	1654 - 1673	45	Kemp *et al,* 1989
p24-6	TGTGWAGCTTGYTCRGCTC	1703 - 1721	50	Swanson *et al,* 2003
cm237R2020	GGTGGGGCTGTTGGCTCTGG	2146 - 2165	60	Personal Communication, John Hackett
cm237F2030	GGAAACCAAAAATGATAGGGGG	2377 - 2398	50	Personal Communication, John Hackett
JA217	CTTTTATTTTTTCTTCTGTCAATGG	2622 - 2646	50	Plantier *et al,* 2005
ABB20-3F	ATCAGTACAATGTGCTTCCA	2980 - 2999	45	Personal Communication, John Hackett
RTUG F3	GAAGCAGAATTAGAAYTGGCAGA	3441 - 3463	50	Personal Communication, John Hackett
cm237F3000	TGTGRGTCTGTTACTATRTTTACTTC	4023 - 4048	50	Personal Communication, John Hackett
rtin seq F2	CAACCAGAYARRAGTGAATCAGA	4074 - 4096	50	Personal Communication, S Engelbrecht
poli7	AACAAGTAGATAAATTAGTCAGT	4186 - 4208	45	Swanson *et al,* 2003
PPF17	AATTGGAGAGCAATGGCTAGTGA	4281 - 4303	50	Swanson *et al,* 2003
ppr15	CCTTCTAAATGTGTACAATC	4419 - 4438	45	Personal Communication, John Hackett
PPF2b	GCAGTCCATGTAGCCAGTGG	4455 - 4474	50	Personal Communication, John Hackett
FGF46	GCATTCCCTACAATCCCCAAAG	4648 - 4669	55	Fong *et al,* 1996

Key: T_A, (Annealing temperature), and Ref (Reference)

Table 6.4 continued: Sequencing primers used for the characterization of R84.

Primer	Oligo Nucleotide Sequence	HXB2 Position	T_A	Ref
poli10b	TATTCATAGATTCYACTACTCCTTG	4671 - 4695	50	Personal Communication, John Hackett
poli2	TAAARACARYAGTACWAATGGCA	4744 - 4766	50	Personal Communication, John Hackett
PPR5b	ACTACTGCCCCTTCACCTTTCCA	4956 - 4978	55	Personal Communication, John Hackett
poli6	ATACATATGRTGTTTTACTAARCT	5107 - 5130	45	Personal Communication, John Hackett
PPR6	CCTGCCATCTGTTTTCCATA	5040 - 5059	45	Personal Communication, John Hackett
55env272R	TCTGGGGCTTGTTCCATCTATCTT	5552 - 5575	55	Personal Communication, John Hackett
55env-5820R	ATCAAAGTCCCCCATCTCCACAAG	6251 - 6273	55	Personal Communication, John Hackett
20env5'2287F	CTTTGAGCCCATTCCCATACATTA	6851 - 6874	50	Personal Communication, John Hackett
55env222R	AATCGCAAAACCAGCTGGAGCAC	6877 - 6899	55	Personal Communication, John Hackett
ES7X	CTGTTAAATGGCAGTCTAGC	7002 - 7021	55	Bachmann *et al,* 1994
ES125	CAATTTCTGGGTCCCCTCCTGAG	7316 - 7338	60	Bachmann *et al,* 1994
ES8X	CACTTCTCCAATTGTCCCTCA	7648 - 7668	55	Bachmann *et al,* 1994
SK68	AGCAGCAGGAAGCACTATGG	7796 - 7815	55	Ou *et al,* 1988
env 27F	CTGGYATAGTGCARCARCA	7861 - 7879	50	Swanson *et al,* 2003
7496F	CCTKGCYCTGGAAAGATACCTA	7964 - 7985	45	Personal Communication, John Hackett
7542F.1	TGGGGCTGCTCTGGAAAACT	8010 - 8029	55	Personal Communication, John Hackett
LP7728R	CCACTTGTCCAATGCCAATAAGTCTTT	8222 - 8195	55	Personal Communication, John Hackett
Menv 19R	AARCCTCCTACTATCATTATRA	8278 - 8299	45	Swanson *et al,* 2003
GP41R1	AACGACAAAGGTGAGTATCCCTGCCTA	8347 - 8374	60	Pieniazek *et al*, 1998
20LTR-3825F	TGGGTGGCAAGTGGTCAAAAAGTA	8798 - 8821	55	Personal Communication, John Hackett
20env 4038R	GTACCTGCGGCCTGACTGGA	9000 - 9019	55	Personal Communication, John Hackett

Key: T_A, (Annealing temperature), and Ref (Reference)

Table 6.5: Sequencing primers used for the sequencing of sample TV239

Primer	Oligo Nucleotide Sequence	HXB2 Position	T_A	Ref
env N	CTGCCAATCAGGGAAGTAGCCTTGTGT	60 - 86	55	Derdeyn *et al,* 2004
p24-2	AGRACYTTRAAYGCATGGGT	1237 - 1245	50	Swanson *et al,* 2003
p24-6	TGTGWAGCTTGYTCRGCTC	1703 - 1721	50	Swanson *et al,* 2003
F1432	AGGCAATGAGTCAAGTACAACA	1883 - 1904	50	Personal Communication, John Hackett
cm237R2020	GGTGGGGCTGTTGGCTCTGG	2146 - 2165	60	Personal Communication, John Hackett
cm237F2030	GGAAACCAAAAATGATAGGGGG	2377 - 2398	50	Personal Communication, John Hackett
JA217	CTTTTATTTTTTCTTCTGTCAATGG	2622 - 2646	50	Plantier *et al,* 2005
ABB20-3F	ATCAGTACAATGTGCTTCCA	2980 - 2999	45	Personal Communication, John Hackett
ABB20-11R	TATGTCCATTGGTCTTGCCC	3546 - 3565	50	Personal Communication, John Hackett
RTUG F3	GAAGCAGAATTAGAAYTGGCAGA	3441 - 3463	50	Personal Communication, John Hackett
cm237F3000	TGTGRGTCTGTTACTATRTTTACTTC	4023 - 4048	50	Personal Communication, John Hackett
rtin seq F2	CAACCAGAYARRAGTGAATCAGA	4074 - 4096	50	Personal Communication, S Engelbrecht
poli7	AACAAGTAGATAAATTAGTCAGT	4186 - 4208	45	Swanson *et al,* 2003
PPF17	AATTGGAGAGCAATGGCTAGTGA	4281 - 4303	50	Personal Communication, John Hackett
ppr15	CCTTCTAAATGTGTACAATC	4419 - 4438	45	Personal Communication, John Hackett
PPF2b	GCAGTCCATGTAGCCAGTGG	4455 - 4474	50	Personal Communication, John Hackett
FGF46	GCATTCCCTACAATCCCCAAAG	4648 - 4669	55	Fong *et al,* 1996
poli10b	TATTCATAGATTCYACTACTCCTTG	4671 - 4695	50	Personal Communication, John Hackett
poli2	TAAARACARYAGTACWAATGGCA	4744 - 4766	50	Swanson *et al,* 2003
PPR6	CCTGCCATCTGTTTTCCATA	5040 - 5059	45	Personal Communication, John Hackett
55env272R	TCTGGGGCTTGTTCCATCTATCTT	5552 - 5575	55	Personal Communication, John Hackett

Key: T_A, (Annealing temperature), and Ref (Reference)

Table 6.5 continued: Sequencing primers used for the sequencing of sample TV239

Primer	Oligo Nucleotide Sequence	HXB2 Position	T_A	Ref
55env-5521R	GCTTCCGCTTCTTCCTGCCATAG	5968 - 5990	55	Personal Communication, John Hackett
20env5'2287F	CTTTGAGCCCATTCCCATACATTA	6851 - 6874	50	Personal Communication, John Hackett
55env222R	AATCGCAAAACCAGCTGGAGCAC	6877 - 6899	55	Personal Communication, John Hackett
ES7X	CTGTTAAATGGCAGTCTAGC	7002 - 7021	55	Bachmann *et al,* 1994
ES8X	CACTTCTCCAATTGTCCCTCA	7648 - 7668	55	Bachmann *et al,* 1994
SK68	AGCAGCAGGAAGCACTATGG	7796 - 7815	55	Ou *et al,* 1988
ED12	AGTGCTTCCTGCTGCTCCCAAGAACCCA	7782 - 7811	60	Delwart *et al,* 1993
env 27F	CTGGYATAGTGCARCARCA	7861 - 7879	50	Personal Communication, John Hackett
LP7725R	GTCCAATGCCAATAAGTCTTGTTC	8193 - 8216	50	Personal Communication, John Hackett
Menv 19R	AARCCTCCTACTATCATTATRA	8278 - 8299	45	Swanson *et al,* 2003
69env-3848F	TGCTGCGAGGGGTGTGGAACTT	8555 - 8576	60	Personal Communication, John Hackett
20env 4038R	GTACCTGCGGCCTGACTGGA	9000 - 9019	55	Personal Communication, John Hackett

Key: T_A, (Annealing temperature), and Ref (Reference)

Table 6.6: Sequencing primers used for the sequencing of TV314.

Primer	Oligo Nucleotide Sequence	HXB2 Position	T_A	Ref
p24-6	TGTGWAGCTTGYTCRGCTC	1703 - 1721	50	Swanson *et al,* 2003
F1432	AGGCAATGAGTCAAGTACAACA	1883 - 1904	50	Personal Communication, John Hackett
cm237R2020	GGTGGGGCTGTTGGCTCTGG	2146 - 2165	60	Personal Communication, John Hackett
cm237F2030	GGAAACCAAAAATGATAGGGGG	2377 - 2398	50	Personal Communication, John Hackett
JA217	CTTTTATTTTTTCTTCTGTCAATGG	2622 - 2646	50	Plantier et al, 2005
ABB20-3F	ATCAGTACAATGTGCTTCCA	2980 - 2999	45	Personal Communication, John Hackett
ABB20-11R	TATGTCCATTGGTCTTGCCC	3546 - 3565	50	Personal Communication, John Hackett
RTUG F3	GAAGCAGAATTAGAAYTGGCAGA	3441 - 3463	50	Personal Communication, John Hackett
cm237F3000	TGTGRGTCTGTTACTATRTTTACTTC	4023 - 4048	50	Personal Communication, John Hackett
rtin seq F2	CAACCAGAYARRAGTGAATCAGA	4074 - 4096	50	Personal Communication, S Engelbrecht
PPF2b	GCAGTCCATGTAGCCAGTGG	4455 - 4474	50	Personal Communication, John Hackett
poli10b	TATTCATAGATTCYACTACTCCTTG	4671 - 4695	50	Personal Communication, John Hackett
poli2	TAAARACARYAGTACWAATGGCA	4744 - 4766	50	Swanson *et al,* 2003
PPR6	CCTGCCATCTGTTTTCCATA	5040 - 5059	45	Personal Communication, John Hackett
55env4F	GGTGGTGGGGCCTCCATAGAAT	5284 - 5305	55	Personal Communication, John Hackett
55env272R	TCTGGGGCTTGTTCCATCTATCTT	5552 - 5575	55	Personal Communication, John Hackett
20env-1301F	ACTATGGGGTACCGGTGTGGAGA	6340 - 6362	55	Personal Communication, John Hackett
55env222R	AATCGCAAAACCAGCTGGAGCAC	6877 - 6899	55	Personal Communication, John Hackett
ES7X	CTGTTAAATGGCAGTCTAGC	7002 - 7021	55	Bachmann *et al,* 1994
E120	GTAGAAATTAATTGTACAAGACCC	7098 - 7121	45	Bachmann *et al,* 1994

Key: T_A (Annealing temperature), F (Forward primer) and R (Reverse primer)

Table 6.6 continued: Sequencing primers used for the sequencing of TV314.

Primer	Oligo Nucleotide Sequence	HXB2 Position	T_A	Ref
SK68	AGCAGCAGGAAGCACTATGG	7796 - 7815	55	Ou *et al,* 1988
7542F.1	TGGGGCTGCTCTGGAAAACT	8010 - 8029	55	Personal Communication, John Hackett
LP7725R	GTCCAATGCCAATAAGTCTTGTTC	8193 - 8216	50	Personal Communication, John Hackett
20env-3481R	CAGGCAAGCGCGAAGAATCC	8475 - 8494	55	Personal Communication, John Hackett
69env-3848F	TGCTGCGAGGGGTGTGGAACTT	8555 - 8576	60	Personal Communication, John Hackett
20LTR-3825F	TGGGTGGCAAGTGGTCAAAAAGTA	8798 - 8821	55	Personal Communication, John Hackett
20env 4038R	GTACCTGCGGCCTGACTGGA	9000 - 9019	55	Personal Communication, John Hackett
ABB55-4-1186	CCAGGGCCAGGGGTTAGAT	9181 - 9199	55	Personal Communication, John Hackett

Key: T_A (Annealing temperature), F (Forward primer) and R (Reverse primer)

Table 6.7: Sequencing primers used for the sequencing of TV 412.

Primer	Oligo Nucleotide Sequence	HXB2 Position	T_A	Ref
p24-2	AGRACYTTRAAYGCATGGGT	1237 - 1256	50	Swanson et al, 2003
p24-6	TGTGWAGCTTGYTCRGCTC	1703 - 1721	50	Swanson et al, 2003
cm237F2030	GGAAACCAAAAATGATAGGGGG	2377 - 2398	50	Personal Communication, John Hackett
JA217	CTTTTATTTTTTCTTCTGTCAATGG	2622 - 2646	50	Plantier et al, 2005
ABB20-3F	ATCAGTACAATGTGCTTCCA	2980 - 2999	45	Personal Communication, John Hackett
ABB20-11R	TATGTCCATTGGTCTTGCCC	3546 - 3565	50	Personal Communication, John Hackett
RTUG F3	GAAGCAGAATTAGAAYTGGCAGA	3441 - 3463	50	Personal Communication, John Hackett
cm237F3000	TGTGRGTCTGTTACTATRTTTACTTC	4023 - 4048	50	Personal Communication, John Hackett
rtin seq F2	CAACCAGAYARRAGTGAATCAGA	4074 - 4096	50	Personal Communication, S Engelbrecht
PPF17	AATTGGAGAGCAATGGCTAGTGA	4281 - 4303	50	Personal Communication, John Hackett
ppr15	CCTTCTAAATGTGTACAATC	4419 - 4438	45	Personal Communication, John Hackett
PPF2b	GCAGTCCATGTAGCCAGTGG	4455 - 4474	50	Personal Communication, John Hackett
poli10b	TATTCATAGATTCYACTACTCCTTG	4671 - 4695	50	Swanson et al, 2003
poli2	TAAARACARYAGTACWAATGGCA	4744 - 4766	50	Personal Communication, John Hackett
PPR6	CCTGCCATCTGTTTTCCATA	5040 - 5059	45	Personal Communication, John Hackett
55env272R	TCTGGGGCTTGTTCCATCTATCTT	5552 - 5575	55	Personal Communication, John Hackett
55env-5820R	ATCAAAGTCCCCCATCTCCACAAG	6251 - 6273	55	Personal Communication, John Hackett
20env-1301F	ACTATGGGGTACCGGTGTGGAGA	6340 - 6362	55	Personal Communication, John Hackett
20env5'2287F	CTTTGAGCCCATTCCCATACATTA	6851 - 6874	50	Personal Communication, John Hackett

Key: T_A (Annealing temperature), and Ref (Reference)

Table 6.7 continued: Sequencing primers used for the sequencing of TV 412.

Primer	Oligo Nucleotide Sequence	HXB2 Position	T_A	Ref
55env222R	AATCGCAAAACCAGCTGGAGCAC	6877 - 6899	55	Personal Communication, John Hackett
ES7X	CTGTTAAATGGCAGTCTAGC	7002 - 7021	55	Bachmann *et al,* 1994
20env-2513F	GCAGAAAGTAGGGCAAGCAATGTA	7505 - 7528	55	Personal Communication, John Hackett
SK68	AGCAGCAGGAAGCACTATGG	7796 - 7815	55	Ou *et al,* 1988
LP7725R	GTCCAATGCCAATAAGTCTTGTTC	8193 - 8216	50	Personal Communication, John Hackett
GP41R1	AACGACAAAGGTGAGTATCCCTGCCTAA	8374 - 8374	60	Pieniazek *et al,* 1998
JH38	GGTGARTATCCCTKCCTAAC	8346 - 8365	45	Swanson *et al,* 2003
GP47R2	TTAAACCTATCAAGCCTCCTACTATCATTA	8281 - 8310	50	Pieniazek *et al,* 1998
env-end	CTTTTTGACCACTTGCCACCCAT	8797 - 8819	55	Personal Communication, S Engelbrecht

Key: T_A (Annealing temperature), and Ref (Reference)

Appendix D

6.4 Gene Cutter Results of NFLG's

Sample R84 from 601 – 9514 (coordinates relative to HXB2)

Nucleotide and Amino Acid composition.

gag – Red

pol - Blue

vif - Orange

vpr - Green

tat - Pink

rev – Dark Red

vpu – Sky Blue

env – Sea Green

nef – Dark Teal

```
1       TTTTGACTAGCGGAGGCTAGAAGGAGAGAGATGGGTGCGAGAGCGTCAGTATTAAGCGGG
1                                         M  G  A  R  A  S  V  L  S  G

61      GGAGAATTAGATAGATGGGAAAAAATTCGGTTAAGGCCAGGGGGAAAGAAAAAATATAGA
21       G  E  L  D  R  W  E  K  I  R  L  R  P  G  G  K  K  K  Y  R

121     TTAAAACATATAGTATGGGCAAGCAGGGAGCTAGAACGATTCGCAGTTAACCCTGGCCTG
41       L  K  H  I  V  W  A  S  R  E  L  E  R  F  A  V  N  P  G  L

181     TTAGAAACATCAGAAGGCTGTAGACAAATACTGGGACAGCTACAACCAGCCCTTCAGACA
61       L  E  T  S  E  G  C  R  Q  I  L  G  Q  L  Q  P  A  L  Q  T

241     GGATCAGAAGAACTTAGATCATTATATAATACAGTAGCAACCCTCTATTGTGTACATCAA
81       G  S  E  E  L  R  S  L  Y  N  T  V  A  T  L  Y  C  V  H  Q

301     AAGATAGAGGTAAGAGACACCAAGGAAGCTTTAGAAAAGATAGAGGAAGAGCAAAACAAA
101      K  I  E  V  R  D  T  K  E  A  L  E  K  I  E  E  E  Q  N  K

361     AGTAAGAAAAAAGCACAGCAAGCAGCAGCTGACACAGGAAACAGCAGTCAGGTCAGCCAA
121      S  K  K  K  A  Q  Q  A  A  A  D  T  G  N  S  S  Q  V  S  Q

421     AATTACCCTATAGTGCAGAATATCCAGGGGCAAATGGTACATCAGGCCATATCACCTAGA
141      N  Y  P  I  V  Q  N  I  Q  G  Q  M  V  H  Q  A  I  S  P  R

481     ACTTTAAATGCATGGGTAAAAGTAGTAGAAGAGAAGGCTTTCAGCCCAGAAGTGATACCC
161      T  L  N  A  W  V  K  V  V  E  E  K  A  F  S  P  E  V  I  P
```

```
541     ATGTTTTCAGCATTATCAGAAGGAGCCACCCCACAAGATTTAAACACCATGCTAAACACA
181      M  F  S  A  L  S  E  G  A  T  P  Q  D  L  N  T  M  L  N  T

601     GTGGGGGGACATCAAGCAGCCATGCAAATGTTAAAAGAGACCATCAATGAGGAAGCTGCA
201      V  G  G  H  Q  A  A  M  Q  M  L  K  E  T  I  N  E  E  A  A

661     GAATGGGATAGATTGCATCCAGTGCATGCAGGGCCTATTGCACCAGGCCAGATGAGAGAA
221      E  W  D  R  L  H  P  V  H  A  G  P  I  A  P  G  Q  M  R  E

721     CCAAGGGGAAGTGACATAGCAGGAACTACTAGTACCCTTCAGGAACAAATAGGATGGATG
241      P  R  G  S  D  I  A  G  T  T  S  T  L  Q  E  Q  I  G  W  M

781     ACAAATAATCCACCTATCCCAGTAGGAGAAATCTATAAAAGATGGATAATCCTAGGATTA
261      T  N  N  P  P  I  P  V  G  E  I  Y  K  R  W  I  I  L  G  L

841     AATAAAATAGTAAGGATGTATAGCCCTGTCAGCATTCTGGACATAAGACAAGGACCAAAG
281      N  K  I  V  R  M  Y  S  P  V  S  I  L  D  I  R  Q  G  P  K

901     GAACCCTTTAGAGACTATGTAGACCGGTTCTATAAAACTCTAAGAGCCGAACAAGCTTCA
301      E  P  F  R  D  Y  V  D  R  F  Y  K  T  L  R  A  E  Q  A  S

961     CAGGATGTAAAAAATTGGATGACAGAAACCTTGTTGGTCCAAAATGCGAACCCAGATTGT
321      Q  D  V  K  N  W  M  T  E  T  L  L  V  Q  N  A  N  P  D  C

1021    AAGACAATTTTAAAAGCATTGGGACCAGCAGCTACCCTAGAAGAAATGATGACAGCATGT
341      K  T  I  L  K  A  L  G  P  A  A  T  L  E  E  M  M  T  A  C

1081    CAGGGAGTAGGAGGACCCGGCCATAAAGCAAGAGTTTTGGCGGAAGCAATGGGCCAAGTA
361      Q  G  V  G  G  P  G  H  K  A  R  V  L  A  E  A  M  G  Q  V

1141    ACAAATGCAGCTACCATAATGATGCAGAAAGGCAATTTTAGGAACCAAAGAAAAAATGTT
381      T  N  A  A  T  I  M  M  Q  K  G  N  F  R  N  Q  R  K  N  V

1201    AAGTGTTTCAATTGTGGCAAAGAAGGGCACATAGCCAGAAATTGCAGGGCCCCTAGGAAA
401      K  C  F  N  C  G  K  E  G  H  I  A  R  N  C  R  A  P  R  K

1261    AGGGGCTGTTGGAAATGTGGAAAGGAAGGACACCAAATGAAAGATTGTACTGAGAGACAG
421      R  G  C  W  K  C  G  K  E  G  H  Q  M  K  D  C  T  E  R  Q

1321    GCTAATTTTTTAGGGAAGATCTGGCCTTCCCACAAGGGAAGGCCAGGGAATTTTCTTCAG
441      A  N  F  L  G  K  I  W  P  S  H  K  G  R  P  G  N  F  L  Q
441           F  F  R  E  D  L  A  F  P  Q  G  K  A  R  E  F  S  S  E
```

```
1381    AGCAGACCAGAGCCAACAGCCCCACCAGAAGAGAGCTTCAGGTTTGGGGAAGAGACAACA
461      S  R  P  E  P  T  A  P  P  E  E  S  F  R  F  G  E  E  T  T
461        Q  T  R  A  N  S  P  T  R  R  E  L  Q  V  W  G  R  D  N  N

1441    ACTCCCTCTCAGAAGCAGGAGCCGATAGACAAGGAACTATATCCTTTAGCTTCCCTCAGA
481      T  P  S  Q  K  Q  E  P  I  D  K  E  L  Y  P  L  A  S  L  R
481        S  L  S  E  A  G  A  D  R  Q  G  T  I  S  F  S  F  P  Q  I

1501    TCACTCTTTGGCAACGACCCCTCGTCACAATAAAGATAGGGGGGCACCTAAAGGAAGCTC
501      S  L  F  G  N  D  P  S  S  Q
501        T  L  W  Q  R  P  L  V  T  I  K  I  G  G  H  L  K  E  A  L

1561    TATTAGATACAGGAGCAGATGATACAGTATTAGAAGAAATGAGTTTGCCAGGAAGATGGA
521        L  D  T  G  A  D  D  T  V  L  E  E  M  S  L  P  G  R  W  K

1621    AACCAAAAATGATAGGGGGAATTGGAGGTTTTATCAAAGTAAGACAGTATGATCAGATAC
541        P  K  M  I  G  G  I  G  G  F  I  K  V  R  Q  Y  D  Q  I  P

1681    CCATAGAAATCTGTGGACATAAAGCTATAGGTACAGTATTAATAGGACCTACACCTGTCA
561        I  E  I  C  G  H  K  A  I  G  T  V  L  I  G  P  T  P  V  N

1741    ACATAATTGGAAGAAATCTGTTGACTCAGCTTGGTTGCACTTTAAATTTTCCCATTAGTC
581        I  I  G  R  N  L  L  T  Q  L  G  C  T  L  N  F  P  I  S  P

1801    CTATTGAAACTGTACCAGTAAAATTAAAGCCAGGAATGGATGGCCCAAAAGTTAAACAAT
601        I  E  T  V  P  V  K  L  K  P  G  M  D  G  P  K  V  K  Q  W

1861    GGCCATTGACAGAAGAAAAAATAAAAGCATTAGTAGAAATTTGTACAGAAATGGAAAAGG
621        P  L  T  E  E  K  I  K  A  L  V  E  I  C  T  E  M  E  K  E

1921    AAGGGAAAATTTCAAAAATTGGGCCTGAAAATCCATACAATACTCCAGTATTTGCCATAA
641        G  K  I  S  K  I  G  P  E  N  P  Y  N  T  P  V  F  A  I  K

1981    AGAAAAAAGACAGTACTAAATGGAGAAAATTAGTAGATTTCAGAGAACTTAATAAGAAAA
661        K  K  D  S  T  K  W  R  K  L  V  D  F  R  E  L  N  K  K  T

2041    CTCAAGACTTCTGGGAAGTTCAATTAGGAATACCACATCCCGCAGGGTTAAAAAAGAAAA
681        Q  D  F  W  E  V  Q  L  G  I  P  H  P  A  G  L  K  K  K  K

2101    AATCAGTAACAGTACTGGATGTGGGTGATGCATATTTTTCAGTTCCCTTAGATAAAGACT
701        S  V  T  V  L  D  V  G  D  A  Y  F  S  V  P  L  D  K  D  F

2161    TCAGGAAGTATACTGCATTTACCATACCTAGTATAAACAATGAGACACCAGGGATTAGAT
721        R  K  Y  T  A  F  T  I  P  S  I  N  N  E  T  P  G  I  R  Y
```

2221 ATCAGTACAATGTGCTTCCACAGGGATGGAAAGGATCACCAGCAATATTCCAAAGTAGCA
741 Q Y N V L P Q G W K G S P A I F Q S S M

2281 TGACAAAAATCTTAGAGCCTTTTAGAAAGCAAAATCCAGACATAGTTATCTATCAATACA
761 T K I L E P F R K Q N P D I V I Y Q Y M

2341 TGGATGATTTGTATGTAGGATCTGATTTAGAAATAGGGCAGCATAGAACAAAAATAGAAG
781 D D L Y V G S D L E I G Q H R T K I E E

2401 AATTGAGACAACATCTGTTGAGGTGGGGATTCACCACACCAGACAAAAAACATCAGAAAG
801 L R Q H L L R W G F T T P D K K H Q K E

2461 AACCTCCATTCCTTTGGATGGGTTATGAACTCCATCCTGATAAATGGACAGTACAGCCTA
821 P P F L W M G Y E L H P D K W T V Q P I

2521 TAGTGCTGCCAGAAAAAGACAGCTGGACTGTCAATGACATACAGAAGTTAGTGGGAAAAT
841 V L P E K D S W T V N D I Q K L V G K L

2581 TGAATTGGGCAAGTCAGATTTACGCAGGGATTAAAGTAAGGCAATTATGTAAACTCCTTA
861 N W A S Q I Y A G I K V R Q L C K L L R

2641 GGGGAGCCAAAGCACTAACAGAAGTAATACCACTAACAGAAGAAGCAGAGCTAGAACTGG
881 G A K A L T E V I P L T E E A E L E L A

2701 CAGAAAACAGGGAAATTCTAAAAGAACCAGTACATGGAGTGTATTATGACCCATCAAAAA
901 E N R E I L K E P V H G V Y Y D P S K N

2761 ACTTAATAGCAGAAATACAGAAGCAGGGGCAAGGCCAATGGACATATCAAATTTATCAAG
921 L I A E I Q K Q G Q G Q W T Y Q I Y Q E

2821 AGCCATTTAAAAATCTGAAAACAGGAAAATATGCAAGAATGAGGGGTGCCCACACTAATG
941 P F K N L K T G K Y A R M R G A H T N D

2881 ATGTAAAACAATTAACAGAGGCAGTGCAAAAAATAGCCATAGAAAGCATAGTAATATGGG
961 V K Q L T E A V Q K I A I E S I V I W G

2941 GAAAGACTCCTAAATTTAAACTACCCATACAAAAAGAAACATGGGAAGCATGGTGGACAG
981 K T P K F K L P I Q K E T W E A W W T E

3001 AGTATTGGCAAGCCACCTGGATTCCTGAGTGGGAGTTTGTCAATACCCCTCCCTTAGTGA
1001 Y W Q A T W I P E W E F V N T P P L V K

3061 AATTATGGTACCAGTTAGAGAAAGAACCCATAGTAGGAGCAGAAACTTTCTATGTAGATG
1021 L W Y Q L E K E P I V G A E T F Y V D G

3121 GGGCAGCTAACAGGGAGACTAAATTAGGAAAAGCAGGATATGTTACTAACAAAGGAAGAC
1041 A A N R E T K L G K A G Y V T N K G R Q

3181 AAAAAGTTATCACCCTAACTGACACAACAAATCAGAAGACTGAGTTACAAGCAATTCATC
1061 K V I T L T D T T N Q K T E L Q A I H L

3241 TAGCGTTGCAGGATTCGGGATTAGAAGTAAACATAGTAACAGACTCACAATATGCATTAG
1081 A L Q D S G L E V N I V T D S Q Y A L G

3301 GAATCATTCAAGCACAACCAGATAAAAGTGAATCAGAGTTAGTCAGTCAAATAATAGAGC
1101 I I Q A Q P D K S E S E L V S Q I I E Q

3361 AGTTAATAAAAAAGGAAAAGGTCTACCTGGCATGGGTACCAGCACACAAAGGAATTGGAG
1121 L I K K E K V Y L A W V P A H K G I G G

3421 GAAATGAACAAGTAGATAAATTAGTCAGTGCTGGAATCAGGAAAGTACTATTTTTAGATG
1141 N E Q V D K L V S A G I R K V L F L D G

3481 GGATAGATAAGGCCCAAGAAGAACATGAGAAATATCACAGTAATTGGAGAGCAATGGCTA
1161 I D K A Q E E H E K Y H S N W R A M A S

3541 GTGATTTTAACCTGCCACCTATAGTAGCAAAAGAGATAGTAGCCAGCTGTGATAAATGTC
1181 D F N L P P I V A K E I V A S C D K C Q

3601 AGTTAAAAGGAGAAGCCATACATGGACAAGTAGACTGTAGTCCAGGAATATGGCAACTAG
1201 L K G E A I H G Q V D C S P G I W Q L D

3661 ATTGTACACATTTAGAAGGAAAAGTTATCCTGGTAGCAGTTCATGTAGCCAGTGGATATA
1221 C T H L E G K V I L V A V H V A S G Y I

3721 TAGAAGCAGAAGTTATTCCAGCAGAGACAGGGCAGGAAACAGCATACTTTCTCTTAAAAT
1241 E A E V I P A E T G Q E T A Y F L L K L

3781 TAGCAGGAAGATGGCCAGTAAAAACAATACATACAGACAATGGCAGCAATTTCACCAGTA
1261 A G R W P V K T I H T D N G S N F T S T

3841 CTACGGTTAAGGCCGCCTGTTGGTGGGCGGGGATCAAGCAGGAATTTGGCATTCCCTACA
1281 T V K A A C W W A G I K Q E F G I P Y N

3901 ATCCCCAAAGTCAAGGAGTAGTAGAATCTATGAATAAAGAATTAAAGAAAATTATAGGAC
1301 P Q S Q G V V E S M N K E L K K I I G Q

3961 AGGTAAGAGATCAGGCTGAACATCTTAAGACAGCAGTACAAATGGCAGTATTCATCCACA
1321 V R D Q A E H L K T A V Q M A V F I H N

```
4021    ATTTTAAAAGAAAAGGGGGGATTGGGGGGTACAGTGCAGGGGAAAGAATAGTAGACATAA
1341      F  K  R  K  G  G  I  G  G  Y  S  A  G  E  R  I  V  D  I  I

4081    TAGCAACAGACATACAAACTAAAGAATTACAAAAACAAATTACAAAAATTCAAAATTTTC
1361      A  T  D  I  Q  T  K  E  L  Q  K  Q  I  T  K  I  Q  N  F  R

4141    GGGTTTATTACAGGGACAGCAGAGAGCCACTTTGGAAAGGACCAGCAAAGCTTCTCTGGA
1381      V  Y  Y  R  D  S  R  E  P  L  W  K  G  P  A  K  L  L  W  K

4201    AAGGTGAAGGGGCAGTAGTAATACAAGATAATAGTGACATAAAAGTAGTGCCAAGAAGAA
1401      G  E  G  A  V  V  I  Q  D  N  S  D  I  K  V  V  P  R  R  K

4261    AAGTAAAAATCATTAGGGATTATGGAAAACAGATGGCAGGTGATGATTGTGTGGCAAGTA
1421                          M  E  N  R  W  Q  V  M  I  V  W  Q  V
1421      V  K  I  I  R  D  Y  G  K  Q  M  A  G  D  D  C  V  A  S  R

4321    GACAGGATGAGGATTAGAACATGGAACAGTTTAGTAAAACACCATATGTATAGGTCAGGG
1441     D  R  M  R  I  R  T  W  N  S  L  V  K  H  H  M  Y  R  S  G
1441      Q  D  E  D

4381    AAAGCTAGGGGATGGGTTTATAGACATCACTATGAAAGCACTCATCCAAGAATAAGTTCA
1461     K  A  R  G  W  V  Y  R  H  H  Y  E  S  T  H  P  R  I  S  S

4441    GAAGTACACATCCCACTAGGGGACGCTAGATTGATAATAACAACATATTGGGGTCTGCAT
1481     E  V  H  I  P  L  G  D  A  R  L  I  I  T  T  Y  W  G  L  H

4501    ACAGGAGAAAGAGACTGGCATTTGGGTCAGGGAGTCTCCATAGAATGGAGGAAAGAGAGA
1501     T  G  E  R  D  W  H  L  G  Q  G  V  S  I  E  W  R  K  E  R

4561    TATAGCACACAAGTAGACCCTAGCCTAGCAGACCAACTAATTCATATGTATTACTTTAAT
1521     Y  S  T  Q  V  D  P  S  L  A  D  Q  L  I  H  M  Y  Y  F  N

4621    TGTTTTTCAGAATCTGCTATAAGAAATGCCATATTAGGACATAGAGTTAGTCCTAGTTGT
1541     C  F  S  E  S  A  I  R  N  A  I  L  G  H  R  V  S  P  S  C

4681    GAATATCAAGCAGGACATAACAAGGTAGGATCTCTACAGTACTTGGCACTAGCAGCATTA
1561     E  Y  Q  A  G  H  N  K  V  G  S  L  Q  Y  L  A  L  A  A  L

4741    ATAACACCAAAAAGGATAAAGCCACCTTTGCCTAGTGTTACGAAACTGACAGAGGATAGA
1581     I  T  P  K  R  I  K  P  P  L  P  S  V  T  K  L  T  E  D  R
1581                                                                M

4801    TGGAACAAGCCCCAGAAGACCAAGGGCCACAGAGGGAGCCATACAATGAATGGACACTAG
1601     W  N  K  P  Q  K  T  K  G  H  R  G  S  H  T  M  N  G  H
```

```
1601        E  Q  A  P  E  D  Q  G  P  Q  R  E  P  Y  N  E  W  T  L  E

4861     AGCTTTTAGAGGAGCTTAAGAATGAAGCTGTTAGACACTTTCCTAGGATCTGGCTCCATG
1621        L  L  E  E  L  K  N  E  A  V  R  H  F  P  R  I  W  L  H  G

4921     GATTAGGGCAACATATCTATGAAACATATGGGGATACTTGGGCAGGAGTGGAAGCCATAA
1641        L  G  Q  H  I  Y  E  T  Y  G  D  T  W  A  G  V  E  A  I  I

4981     TAAGAATTTTGCAACAACTGCTGTTTATTCATTTCAGAATTGGGTGTCGCCATAGCAGAA
1661        R  I  L  Q  Q  L  L  F  I  H  F  R  I  G  C  R  H  S  R  I

5041     TAGGCATTACTCTACAGAGAAGAGCAAGAAATGGAGCCAGTAGATCCTAGACTAGAGCCC
1681                                   M  E  P  V  D  P  R  L  E  P
1681        G  I  T  L  Q  R  R  A  R  N  G  A  S  R  S

5101     TGGAAGCATCCAGGAAGTCAGCCTAAGACTGCTTGTACCCCTTGCTATTGTAAAAGGTGT
1701      W  K  H  P  G  S  Q  P  K  T  A  C  T  P  C  Y  C  K  R  C

5161     TGCTTTCATTGCCAAGTTTGTTTCATAAAAAAAGCTTTAGGCATCTCCTATGGCAGGAAG
1721      C  F  H  C  Q  V  C  F  I  K  K  A  L  G  I  S  Y  G  R  K
                                                          M  A  G  R

5221     AAGCGGAGACAGCGACGAAGAGCTCCTCAAAGCAGTCAGACTCATCAAGTTTCTCTATCA
1741      K  R  R  Q  R  R  R  A  P  Q  S  S  Q  T  H  Q  V  S  L  S
1741       S  G  D  S  D  E  E  L  L  K  A  V  R  L  I  K  F  L  Y  Q

5281     AAGCAGTAAGTAGTACATGTAATGCAATCTTTACACATATTAGCAATAGTAGCATTAGTA
1761      K  Q                 M  Q  S  L  H  I  L  A  I  V  A  L  V
1761       S

5341     GTAGCAATAATAATAGCAATAGTTGTGTGGTCCATAGTATTCATAGAATATAGGAAAATA
1781      V  A  I  I  I  A  I  V  V  W  S  I  V  F  I  E  Y  R  K  I

5401     TTAAGACAAAGAAAAATAGACAGGTTAATTGATAGAATAAGAGAAAGAGCAGAAGACAGT
1801      L  R  Q  R  K  I  D  R  L  I  D  R  I  R  E  R  A  E  D  S

5461     GGCAATGAAAGTGAAGGGGATCAGGAAGAATTATCAGCACTTGTGGAGATGGGGCACCAT
1821      G  N  E  S  E  G  D  Q  E  E  L  S  A  L  V  E  M  G  H  H
1821         M  K  V  K  G  I  R  K  N  Y  Q  H  L  W  R  W  G  T  M

5521     GCTCCTTGGGATGTTAATGATCTGTAGTGCTGCAGAACAATTGTGGGTCACAGTCTATTA
1841      A  P  W  D  V  N  D  L
1841       L  L  G  M  L  M  I  C  S  A  A  E  Q  L  W  V  T  V  Y  Y
```

```
5581    TGGGGTACCTGTGTGGAAAGAAGCAACCACCACTCTATTTTGTGCCTCAGATGCTAAAGC
1861     G  V  P  V  W  K  E  A  T  T  T  L  F  C  A  S  D  A  K  A

5641    ATATGATACAGAGGTACATAATGTTTGGGCCACACATGCCTGTGTACCCACAGACCCCAA
1881     Y  D  T  E  V  H  N  V  W  A  T  H  A  C  V  P  T  D  P  N

5701    CCCACAAGAAGTAGTATTGGAAAATGTGACAGAATATTTTAACATGTGGAAAAATGACAT
1901     P  Q  E  V  V  L  E  N  V  T  E  Y  F  N  M  W  K  N  D  M

5761    GGTAGAACAGATGCATGAGGATATAATCAGTTTATGGGATCAAAGCCTAAAGCCATGTGT
1921     V  E  Q  M  H  E  D  I  I  S  L  W  D  Q  S  L  K  P  C  V

5821    AAAATTAACCCCACTCTGTGTTACTTTAAATTGCACTGATTGGAAGAATACTACTAATAC
1941     K  L  T  P  L  C  V  T  L  N  C  T  D  W  K  N  T  T  N  T

5881    CACTAGTAGTGGCGGGGAAATGATGGGGGAAGGAGAAATGAAAAACTGCTCTTTCAACAT
1961     T  S  S  G  G  E  M  M  G  E  G  E  M  K  N  C  S  F  N  I

5941    CACCACAAGCTTAAAAGATAAGGTGCAGAGAGAATATGCACTTCTTTATAAACTTGATGT
1981     T  T  S  L  K  D  K  V  Q  R  E  Y  A  L  L  Y  K  L  D  V

6001    AGTGCCAATAGAGAATGATGATAGTAATTCCAGATATAGGTTGATAAGTTGTAACACCTC
2001     V  P  I  E  N  D  D  S  N  S  R  Y  R  L  I  S  C  N  T  S

6061    AGTCATTACACAGGCCTGTCCAAAAGTATCCTTTCASCCAATTCCCATACATTATTGTGC
2021     V  I  T  Q  A  C  P  K  V  S  F  X  P  I  P  I  H  Y  C  A

6121    CCCGGCTGGTTTTGCGATTCTAAAGTGTAACGATAAGAGGTTCGAGGGAAAAGGGCCGTG
2041     P  A  G  F  A  I  L  K  C  N  D  K  R  F  E  G  K  G  P  C

6181    TACAAATGTCAGCACAGTACAATGTACACATGGAATTAAGCCAGTAGTATCAACTCAACT
2061     T  N  V  S  T  V  Q  C  T  H  G  I  K  P  V  V  S  T  Q  L

6241    GYTGTTAAATGGCAGTCTAGCAGAAGAAGAGGTAGTAATTAGATCTGACAATTTCACGAA
2081     X  L  N  G  S  L  A  E  E  E  V  V  I  R  S  D  N  F  T  N

6301    CAATGCTAAAACCATAATAGTACAACTGAAAGAATCTGTAGAAATTAATTGTACAAGGCC
2101     N  A  K  T  I  I  V  Q  L  K  E  S  V  E  I  N  C  T  R  P

6361    CAACAACAATACAAGAAAAAGTATACATATAGGACCAGGGAGAGCATTTTATACAACAGG
2121     N  N  N  T  R  K  S  I  H  I  G  P  G  R  A  F  Y  T  T  G

6421    AGAAATAATAGGAGATATAAGACAAGCATATTGTAACATTAGTAGTACAAAATGGAATAA
2141     E  I  I  G  D  I  R  Q  A  Y  C  N  I  S  S  T  K  W  N  N
```

6481 CACTTTAAGACAGATAGTTGGAAAATTAAGAGAAAAATTTGGGAATAAAACAATAGTTTT
2161 T L R Q I V G K L R E K F G N K T I V F

6541 TAATCAATCCTCAGGAGGGGACATAGAAATTGTAATGCACAGTTTTAATTGTGGAGGGGA
2181 N Q S S G G D I E I V M H S F N C G G E

6601 ATTTTTCTACTGTAACTTAACACAACTGTTTAATAGTACTTGGAATGTAAATGGTACTGA
2201 F F Y C N L T Q L F N S T W N V N G T E

6661 AGGGTCAAATAACACTGACAAAAATATCACACTCCCATGCAGGATAAAACAAATTATAAA
2221 G S N N T D K N I T L P C R I K Q I I N

6721 CATGTGGCAGACAGTAGGAAAAGCAATGTATGCCCCTCCCATCAGAGGACAAATTAGATG
2241 M W Q T V G K A M Y A P P I R G Q I R C

6781 TTCATCAAATATTACAGGGCTGCTATTAACAAGAGATGGTGGTAATAACGAGAACGATCC
2261 S S N I T G L L L T R D G G N N E N D P

6841 CGAAAATTCACCCGGAGGAGGAGATATGAGGGACAATTGGAGAAGTGAATTATATAAATA
2281 E N S P G G G D M R D N W R S E L Y K Y

6901 TAAAGTATTAAAAATTAAAGCATTAGGAGTAGCACCCCCCAAGGCCAAGAGAAGAGTGGT
2301 K V L K I K A L G V A P P K A K R R V V

6961 GCAAAGAGTAAAAAGAGCAGTGGGAATAGGAGCTGTGTTCTTTGGGTTCTTGGGAGCAGC
2321 Q R V K R A V G I G A V F F G F L G A A

7021 AGGAAGCACTATGGGCGCAGCGTCAATGACGCTGACGGTACAGGCCAGACTATTATTGTC
2341 G S T M G A A S M T L T V Q A R L L L S

7081 TGGTATAGTGCAACAGCAAAACAATTTGCTGAAGGCTATTGAGGCGCAACAGCACCTGTT
2361 G I V Q Q Q N N L L K A I E A Q Q H L L

7141 GCAACTCACAGTCTGGGGCATCAAGCAGCTCCAGGCAAGAGTCCTGGCTATAGAAAGATA
2381 Q L T V W G I K Q L Q A R V L A I E R Y

7201 CCTAAAGGATCAACAGCTCCTGGGGATTTGGGGTTGCTCTGGAAAACTCATTTGCACCAC
2401 L K D Q Q L L G I W G C S G K L I C T T

7261 TGCTGTGCCTTGGAATGCTAGTTGGAGTAATAAATCTCAGGATAAGATTTGGCATAACAT
2421 A V P W N A S W S N K S Q D K I W H N M

7321 GACCTGGATGGAGTGGGAAAGAGAAATTAGCAATTACACAAGCTTAATATACAACTTACT
2441 T W M E W E R E I S N Y T S L I Y N L L

```
7381    TGAAGAATCGCAAAACCAACAAGAAAAGAATGAACAAGAATTATTGGAATTAGATAAATG
2461     E  E  S  Q  N  Q  Q  E  K  N  E  Q  E  L  L  E  L  D  K  W

7441    GGCAAATTTGTGGAATTGGTTTAGCATATCAAACTGGCTGTGGTATATAAAAATATTCAT
2481     A  N  L  W  N  W  F  S  I  S  N  W  L  W  Y  I  K  I  F  I

7501    AATGATAATAGGAGGCTTGGTAGGTTTAAGAATAGTTTTTGCTGTACTTTCTATAGTGAA
2501     M  I  I  G  G  L  V  G  L  R  I  V  F  A  V  L  S  I  V  N

7561    TAGAGTTAGGCAGGGATACTCACCATTATCGTTGCAGACCCGCCTGCCAGCAGCCTCGAG
2521                                            P  P  A  S  S  L  E
2521     R  V  R  Q  G  Y  S  P  L  S  L  Q  T  R  L  P  A  A  S  R
2521                                         P  A  C  Q  Q  P  R  G

7621    GGGACCCGACAGGCCCGAAGGAATCGAAGAAGAAGGTGGAGAGAGAGACAGAGACAGATC
2541    G  T  R  Q  A  R  R  N  R  R  R  R  W  R  E  R  Q  R  Q  I
2541     G  P  D  R  P  E  G  I  E  E  E  G  G  E  R  D  R  D  R  S
2541      D  P  T  G  P  K  E  S  K  K  K  V  E  R  E  T  E  T  D  P

7681    CGGTGGATTAGTGAACGGATTCTTAGCACTTATCTGGATCGACCTACGGAGCCTGTGCCT
2561    R  W  I  S  E  R  I  L  S  T  Y  L  D  R  P  T  E  P  V  P
2561     G  G  L  V  N  G  F  L  A  L  I  W  I  D  L  R  S  L  C  L
2561      V  D

7741    CTTCAGCTACCACCGCTTGAGAGACTTACTCTTGATTGTAACGAGGATTGTGGAACTTCT
2581    L  Q  L  P  P  L  E  R  L  T  L  D  C  N  E  D  C  G  T  S
2581     F  S  Y  H  R  L  R  D  L  L  L  I  V  T  R  I  V  E  L  L

7801    GGGACGCAGGGGGTGGGAACCCCTCAAATATTGGTGGAATCTCCTACAGTATTGGAGTCA
2601    G  T  Q  G  V  G  T  P  Q  I  L  V  E  S  P  T  V  L  E  S
2601     G  R  R  G  W  E  P  L  K  Y  W  W  N  L  L  Q  Y  W  S  Q

7861    GGAACTAAAGAATAGTGCTATTAGCTTGCTCAATGCCACAGCCATAGCAGTAGCTGAGGG
2621    G  T  K  E
2621     E  L  K  N  S  A  I  S  L  L  N  A  T  A  I  A  V  A  E  G

7921    GACAGATAGGGTTATAGAAGTATTACAAAGAGCTTATAGAGTTATTCTCCACATACCTAG
2641     T  D  R  V  I  E  V  L  Q  R  A  Y  R  V  I  L  H  I  P  R

7981    AAGAATAAGACAGGGCGCGGAAAGGGCTTTGGTATAAGATGGGTGGCAAGTGGTCAAAAA
2661     R  I  R  Q  G  A  E  R  A  L  V
2661                                         M  G  G  K  W  S  K  S

8041    GTAGTGTGGTTGGATGGCCTACTGTAAGGGAAAGAATGAGACGAGCACGAGCTGAGCCAG
2681      S  V  V  G  W  P  T  V  R  E  R  M  R  R  A  R  A  E  P  A
```

8101 CAGCAGAGCCAGCAGCATGTGGGGTGGGAGCAGCATCTCGAGACCTGGAAAAACATGGAG
2701 A E P A A C G V G A A S R D L E K H G A

8161 CACTCACAAGTAGCAATACAGCAACTAACAATGCTGATTGTGCCTGGCTAAAAGCACAAG
2721 L T S S N T A T N N A D C A W L K A Q E

8221 AGGAGGAGGTGGTGGTTTTTCCAGTCAGACCTCAGGTACCTTTAAGACCAATGACTTACA
2741 E E V V V F P V R P Q V P L R P M T Y K

8281 AGGCAGCTTTAGATCTTAGCCACTTTTTAAAAGAAAAGGGGGGACTGGAAGGGCTAATTC
2761 A A L D L S H F L K E K G G L E G L I H

8341 ACTCCCAAAAAAGACAAGATATCCTTGATCTGTGGGTCTACCACACACAAGGCTACTTCC
2781 S Q K R Q D I L D L W V Y H T Q G Y F P

8401 CTGATTGGCAGAACTACACACCAGGGCCAGGGATCAGGTACCCACTGACCTTCGGATGGT
2801 D W Q N Y T P G P G I R Y P L T F G W C

8461 GCTTCAAGCTAGTACCTGTTGAACCAGAGAAGATAGAAGAAGCCAATGAAGGAGAGAACA
2821 F K L V P V E P E K I E E A N E G E N N

8521 ACAGATTGTTACACCCTATGAGCCTGCATGGGATGGAGGACCCGGAGAGAGAAGTGTTAG
2841 R L L H P M S L H G M E D P E R E V L E

8581 AGTGGAGGTTTGACAGTCGCCTAGCATATCATCACTTGGCCCGAGAGATACATCCGGAGT
2861 W R F D S R L A Y H H L A R E I H P E Y

8641 ACTACAAGGACTGCTGACATCGAGCTTTCTACAAGGGACTTTCCCCTGGGGGACTTTCCA
2881 Y K D C

Sample TV 239 *gag-pol* from 1245 – 5534 (coordinates relative to HSB2)

Nucleotide and Amino Acid composition

gag **– Red**

pol **- Blue**

vif **- Orange**

```
1       TTGAATGCATGGGTAAAAGTAATAGAAGAAAAGGCTTTCAGCCCAGAAGTAATACCCATG
1        L  N  A  W  V  K  V  I  E  E  K  A  F  S  P  E  V  I  P  M

61      TTCTCAGCATTATCAGAAGGAGCCACCCCGCAAGATTTAAATATGATGCTAAACATAGTG
21       F  S  A  L  S  E  G  A  T  P  Q  D  L  N  M  M  L  N  I  V

121     GGGGGACACCAGGCAGCAATGCAAATGTTAAAAGATACCATCAATGAGGAGGCTATAGAA
41       G  G  H  Q  A  A  M  Q  M  L  K  D  T  I  N  E  E  A  I  E

181     TGGGACAGGACACATCCAGTACATGCAGGACCTATCCCACCAGGCCAGATGAGAGAACCA
61       W  D  R  T  H  P  V  H  A  G  P  I  P  P  G  Q  M  R  E  P

241     AGGGGAAGTGATATAGCAGGAACTACTAGTACCCTTCAAGAACAGATAGGATGGATGACA
81       R  G  S  D  I  A  G  T  T  S  T  L  Q  E  Q  I  G  W  M  T

301     AGTAACCCACCTATCCCAGTGGGAGACATCTATAAAAGGTGGATAATCCTGGGATTAAAT
101      S  N  P  P  I  P  V  G  D  I  Y  K  R  W  I  I  L  G  L  N

361     AAAATAGTAAGAATGTATAGCCCTGTCAGCATTTTGGACATAAGACAAGGGCCAAAAGAA
121      K  I  V  R  M  Y  S  P  V  S  I  L  D  I  R  Q  G  P  K  E

421     CCCTTCAGAGACTATGTAGATAGGTTCTTTAAAGCTCTCAGAGCTGAGCAAGCTACACAA
141      P  F  R  D  Y  V  D  R  F  F  K  A  L  R  A  E  Q  A  T  Q

481     GAAGTAAAAAACTGGATGACAGAAACTTTACTGGTCCAAAATGCAAATCCAGACTGTAAG
161      E  V  K  N  W  M  T  E  T  L  L  V  Q  N  A  N  P  D  C  K

541     TCTATTTTAAAAGCATTAGGACAGGGGCCTACATTAGAAGAAATGATGACAGCATGCCAG
181      S  I  L  K  A  L  G  Q  G  P  T  L  E  E  M  M  T  A  C  Q

601     GGAGTGGGAGGACCCAGCCATAAGGCAAGGGTTTTAGCAGAAGCAATGAGTCAAGTACAA
201      G  V  G  G  P  S  H  K  A  R  V  L  A  E  A  M  S  Q  V  Q

661     AACACAAACATAATGATGCAGAGAGGCAATTTTAAGGGCCAGAAAAGAACTATTAAGTGT
221      N  T  N  I  M  M  Q  R  G  N  F  K  G  Q  K  R  T  I  K  C

721     TTCAATTGTGGCAAAGAAGGACACCTAGCCAGAAATTGCAGGGCCCCTAGAAAAAAGGGC
241      F  N  C  G  K  E  G  H  L  A  R  N  C  R  A  P  R  K  K  G
```

```
781     TGTTGGAAATGTGGAAAAGAAGGGCACCAGATGAAGGACTGTACTGAGAGACAGGCTAAT
261      C  W  K  C  G  K  E  G  H  Q  M  K  D  C  T  E  R  Q  A  N

841     TTTTTAGGGAAAATCTGGCCTTCCAGCAAAGGGAGGCCAGGGAATTTTCCTCAGAACAGA
281      F  L  G  K  I  W  P  S  S  K  G  R  P  G  N  F  P  Q  N  R
281        F  R  E  N  L  A  F  Q  Q  R  E  A  R  E  F  S  S  E  Q  T

901     CTGGAGCCAACAGCCCCACCAGCAGAGAGCTTTGGGATGGGAGAAGGAATAACCTCCCCT
301      L  E  P  T  A  P  P  A  E  S  F  G  M  G  E  G  I  T  S  P
301        G  A  N  S  P  T  S  R  E  L  W  D  G  R  R  N  N  L  P  S

961     CCGAAGCAGGAGCAGAGAGACAGGGAACAGCCCCCTCCCTTAGTTTCCCTCAAATCACTC
321      P  K  Q  E  Q  R  D  R  E  Q  P  P  P  L  V  S  L  K  S  L
321        E  A  G  A  E  R  Q  G  T  A  P  S  L  S  F  P  Q  I  T  L

1021    TTTGGCAACGACCCCTTGTCACAGTAAAAATAGGGGGACAACTAAGAGAAGCCCTATTAG
341      F  G  N  D  P  L  S  Q
341        W  Q  R  P  L  V  T  V  K  I  G  G  Q  L  R  E  A  L  L  D

1081    ATACAGGGGCAGATGATACAGTATTAGAAGAAATAAATTTGCCAGGAAAATGGAAACCAA
361        T  G  A  D  D  T  V  L  E  E  I  N  L  P  G  K  W  K  P  K

1141    AAATGATAGGGGGAATTGGAGGTTTTATTAAGGTAAGACAATATGATCAGATACTTATAG
381        M  I  G  G  I  G  G  F  I  K  V  R  Q  Y  D  Q  I  L  I  E

1201    AAATTTGTGGAAAGAAGGCAATAGGTACAGTATTGGTAGGACCTACACCTGTCAACATAA
401        I  C  G  K  K  A  I  G  T  V  L  V  G  P  T  P  V  N  I  I

1261    TTGGAAGAAATATGTTGACCCAGATTGGTTGTACCTTAAATTTTCCAATTAGTCCTATTG
421        G  R  N  M  L  T  Q  I  G  C  T  L  N  F  P  I  S  P  I  E

1321    AAACTGTACCAGTAACATTAAAGCCAGGAATGGATGGCCCAAAGGTTAAACAATGGCCAT
441        T  V  P  V  T  L  K  P  G  M  D  G  P  K  V  K  Q  W  P  L

1381    TGACAGAAGAAAAAATAAAAGCATTAACAGAAATTTGTACAGAAATGGAAAAGGAAGGAA
461        T  E  E  K  I  K  A  L  T  E  I  C  T  E  M  E  K  E  G  K

1441    AAATTTCAAAAATTGGGCCTGAAAATCCATACAATACTCCAGTATTTGCTATAAAGAAGA
481        I  S  K  I  G  P  E  N  P  Y  N  T  P  V  F  A  I  K  K  K

1501    AGGACAGCACTAAATGGAGGAAGTTAGTAGATTTCAGAGAACTCAATAAAAGAACTCAGG
501        D  S  T  K  W  R  K  L  V  D  F  R  E  L  N  K  R  T  Q  D
```

```
1561    ACTTCTGGGAAGTTCAATTAGGAATACCACATCCAGCAGGTTTAAAAAAGAAAAAATCAG
521       F  W  E  V  Q  L  G  I  P  H  P  A  G  L  K  K  K  K  S  V

1621    TAACAGTACTAGATGTGGGGGACGCATATTTTTCAGTTCCTTTAGATGAAAACTTTAGAA
541       T  V  L  D  V  G  D  A  Y  F  S  V  P  L  D  E  N  F  R  K

1681    AGTATACTGCGTTCACCATACCTAGTACAAACAATGAGACACCAGGAGTCAGGTATCAAT
561       Y  T  A  F  T  I  P  S  T  N  N  E  T  P  G  V  R  Y  Q  Y

1741    ACAATGTGCTTCCACAGGGATGGAAAGGATCACCAGCAATATTCCAAAGTAGCATGACAA
581       N  V  L  P  Q  G  W  K  G  S  P  A  I  F  Q  S  S  M  T  K

1801    AAATCTTAGAACCCTTTAGATCACAAAATCCAGAAATAGTTATCTATCAATACATGGATG
601       I  L  E  P  F  R  S  Q  N  P  E  I  V  I  Y  Q  Y  M  D  D

1861    ACTTATATGTAGGATCTGATTTAGAAATAGGGCAGCATAGAGAAAAGGTGGAAGAGTTAA
621       L  Y  V  G  S  D  L  E  I  G  Q  H  R  E  K  V  E  E  L  R

1921    GAAAGCATCTATTGAGCTGGGGATTAACTACACCAGACAAAAAGCACCAGAAAGAACCTC
641       K  H  L  L  S  W  G  L  T  T  P  D  K  K  H  Q  K  E  P  P

1981    CATTTCTTTGGATGGGGTATGAACTCCATCCTGACAAATGGACAGTCCAGCCTATACAGC
661       F  L  W  M  G  Y  E  L  H  P  D  K  W  T  V  Q  P  I  Q  L

2041    TGCCAGACAAGGACAGCTGGACTGTTAATGATATACAGAAATTAGTGGGAAAACTAAATT
681       P  D  K  D  S  W  T  V  N  D  I  Q  K  L  V  G  K  L  N  W

2101    GGGCAAGTCAGATTTATCCAGGGATTAGAGTAAAACAACTGTGTAAACTCCTCAGGGGAG
701       A  S  Q  I  Y  P  G  I  R  V  K  Q  L  C  K  L  L  R  G  A

2161    CCAAAGCACTAACAGATGTAGTAACACTGACAGAGGAAGCAGAATTAGAATTGGCAGAGA
721       K  A  L  T  D  V  V  T  L  T  E  E  A  E  L  E  L  A  E  N

2221    ACAGGGAAATTCTAAAAGACCCTGTGCATGGGGTATATTATGACCCATCAAAAGACCTAG
741       R  E  I  L  K  D  P  V  H  G  V  Y  Y  D  P  S  K  D  L  V

2281    TAGCAGAAATACAGAAACAGGGACAAGACCAATGGACATATCAAATTTATCAAGAGCCAT
761       A  E  I  Q  K  Q  G  Q  D  Q  W  T  Y  Q  I  Y  Q  E  P  F

2341    TTAAAAATCTAAAGACAGGAAAATATGCAAAAAAGAAGTCTGCTCACACTAATGATGTAA
781       K  N  L  K  T  G  K  Y  A  K  K  K  S  A  H  T  N  D  V  K

2401    AACAGTTAACAGAAGTGGTGCAAAAAGTGGTTACAGAAAGCATAGTAATCTGGGGAAAGA
801       Q  L  T  E  V  V  Q  K  V  V  T  E  S  I  V  I  W  G  K  T
```

2461 CCCCTAAATTTAAGTTACCTATACAAAAAGAAACATGGGAAACATGGTGGACGGAGTATT
821 P K F K L P I Q K E T W E T W W T E Y W

2521 GGCAGGCTACTTGGATTCCTGAATGGGAGTTTGTCAATACCCCTCCTCTAGTAAAATTAT
841 Q A T W I P E W E F V N T P P L V K L W

2581 GGTATCAGTTAGAAAAAGACCCCATATTAGGAGTAGAGACTTTTTATGTAGATGGGGCAG
861 Y Q L E K D P I L G V E T F Y V D G A A

2641 CTAACAGGGAGACTAAGCTAGGAAAAGCAGGGTATGTCACTGATAGGGGAAGACAAAAGG
881 N R E T K L G K A G Y V T D R G R Q K V

2701 TTGTTTCCCTAACTGAGACAACAAATCAAAAGACTGAATTACATGCAATCTATCTAGCCT
901 V S L T E T T N Q K T E L H A I Y L A L

2761 TGCAGGATTCAGGACCAGAAGTAAACATAGTAACAGACTCACAGTATGCATTAGGAATCA
921 Q D S G P E V N I V T D S Q Y A L G I I

2821 TTCAGGCACAACCAGACAGGAGTGAAACAGAAATAGTCAATCAAATAATAGAGAAGCTAA
941 Q A Q P D R S E T E I V N Q I I E K L I

2881 TAGAAAAAGAAAAAGTCTACCTGTCATGGGTACCAGCACATAAGGGAATTGGAGGAAATG
961 E K E K V Y L S W V P A H K G I G G N E

2941 AACAAGTAGATAAATTAGTCAGTTCTGGAATCAGGAAGGTGCTATTTTTAGATGGGATAG
981 Q V D K L V S S G I R K V L F L D G I D

3001 ATAAAGCTCAAGAAGAGCATGAAAGGTATCACAGCAATTGGAGAGCAATGGCTAGTGATT
1001 K A Q E E H E R Y H S N W R A M A S D F

3061 TCAACCTGCCACCAGTAGTAGCAAAGGAAATAGTAGCCAGCTGTGATAAATGTCAACTAA
1021 N L P P V V A K E I V A S C D K C Q L K

3121 AAGGGGAAGCCATGCATGGACAAGTAGACTGTAGTCCAGGAATATGGCAAGTAGATTGCA
1041 G E A M H G Q V D C S P G I W Q V D C T

3181 CACATCTAGAAGGAAAAGTAATCATAGTAGCAGTCCATGTAGCCAGTGGCTATATAGAGG
1061 H L E G K V I I V A V H V A S G Y I E A

3241 CAGAAGTTATCCCAGCAGAAACAGGACAGGAGGCAGCATACTTTCTGTTAAAATTAGCAG
1081 E V I P A E T G Q E A A Y F L L K L A A

3301 CAAGATGGCCAGTAAAGGTAATACACACAGACAATGGCAGTAATTTCACCAGTGCCGCAG
1101 R W P V K V I H T D N G S N F T S A A V

3361 TTAAAGCAGCCTGTTGGTGGGCAAATATCCAACAGGAATTTGGAATTCCCTACAATCCCC
1121 K A A C W W A N I Q Q E F G I P Y N P Q

3421 AAAGTCAAGGAGTAGTGGAATCTATGAATAAAGAATTAAAGAAAATCATAGGTCAGGTAA
1141 S Q G V V E S M N K E L K K I I G Q V R

3481 GAGAACAAGCTGAGCACCTTAAGACAGCAGTACAAATGGCAGTATTCATTCACAATTTTA
1161 E Q A E H L K T A V Q M A V F I H N F K

3541 AAAGAAAAGGGGGGATTGGGGGGTACAGTGCAGGGGAAAGAATAATAGACATAATAGCAA
1181 R K G G I G G Y S A G E R I I D I I A T

3601 CAGACATACAAACTAAAGAATTACAAAAACAAATTACAAAAATTCAAAATTTCCGGGTTT
1201 D I Q T K E L Q K Q I T K I Q N F R V Y

3661 ATTACAGGGACAGCAGAGACCCAATTTGGAAAGGACCAGCAAGACTGCTCTGGAAAGGTG
1221 Y R D S R D P I W K G P A R L L W K G E

3721 AAGGGGCAGTAGTAATACAGGACAATAGTGATATAAAGGTAGTACCAAGAAGAAAAGCAA
1241 G A V V I Q D N S D I K V V P R R K A K

3781 AAATCATTAGGGACTATGGAAAACAGATGGCAGGTGATGATTGTGTGGCAGGTAGACAGG
1261 M E N R W Q V M I V W Q V D R
1261 I I R D Y G K Q M A G D D C V A G R Q D

3841 ATGAGGATTAGAACATGGCACAGTTTAGTAAAATATCATATATATGTCTCTAAGAAAGCT
1281 M R I R T W H S L V K Y H I Y V S K K A
1281 E D

3901 AAAGATTGGTTTTATAGACACCACTATGAAAGCCTGCATCCAAAAGTGAGTTCAGAAATA
1301 K D W F Y R H H Y E S L H P K V S S E I

3961 CACATCCCACTAGGGGATGCTAGATTAGTAGTAAAAACATATTGGGGTCTGCAGACAGGA
1321 H I P L G D A R L V V K T Y W G L Q T G

4021 GAAAAAGACTGGCAATTGGGTCATGGGGTCTCCATAGAATGGAGACAGAACAGATATGTT
1341 E K D W Q L G H G V S I E W R Q N R Y V

4081 ACACAAATAGATCCTGATCTAGCTGACCAACTAATTCACCTTCATCCTTTTAACTGTTTT
1361 T Q I D P D L A D Q L I H L H P F N C F

4141 TCAGAATCTGCCATAAGGAAAGTCATATTAGGACAAGTAGTTAGTCCTAGGTGTGAATAT
1381 S E S A I R K V I L G Q V V S P R C E Y

```
4201    CCAACAGGACATAATAAGGTAGGATCTCTACAATATTTGGCTCTGAAAGCATTAGTAGCA
1401     P  T  G  H  N  K  V  G  S  L  Q  Y  L  A  L  K  A  L  V  A

4261    CCAAGAAAGCCAAAGCCACCTGCCCAG
1421     P  R  K  P  K  P  P
```

Sample TV 239 *env-nef* from 6195-9146 (coordinates relative to HXB2)
Nucleotide and amino acid composition.

vpr - Green
tat - Pink
rev – Dark Red
vpu – Sky Blue
env – Sea Green
nef – Dark Teal

```
1       GAATTAGGGAAAGAGCAGAAGACAGTGGCAATGAGAGTGATGGGGATACAGAGGAATTGT
1                                       M  R  V  M  G  I  Q  R  N  C
1          I  R  E  R  A  E  D  S  G  N  E  S  D  G  D  T  E  E  L  S

61      CAACAATGGTGGATATGGGGCATCTTAGGCTTTTGGATGATATATAATGGGATGGGGGCG
21       Q  Q  W  W  I  W  G  I  L  G  F  W  M  I  Y  N  G  M  G  A
21         T  M  V  D  M  G  H  L  R  L  L  D  D  I

121     GGCTTGTGGGTCACGGTCTATTATGGAGTACCTGTGTGGAAAGACGCAGATACCACCCTA
41       G  L  W  V  T  V  Y  Y  G  V  P  V  W  K  D  A  D  T  T  L

181     TTTTGTGCATCAGATGCTAAGGCATATGATACAGAAGTGCATAATGTCTGGGCTACACAT
61       F  C  A  S  D  A  K  A  Y  D  T  E  V  H  N  V  W  A  T  H

241     GCCTGTGTACCCACAGACCCCAACCCACAAGAAATGACTTTAATGAATGTAACAGAAAAG
81       A  C  V  P  T  D  P  N  P  Q  E  M  T  L  M  N  V  T  E  K

301     TTTAACATGTGGAAAAATAACATGGTAGAACAAATGCACACAGATATAATCAGTTTATGG
101      F  N  M  W  K  N  N  M  V  E  Q  M  H  T  D  I  I  S  L  W

361     GACCAAAGCCTAAAACCATGTGTAAGCTTAACCCCTCTCTGTGTTACTTTAAATTGCAGA
121      D  Q  S  L  K  P  C  V  S  L  T  P  L  C  V  T  L  N  C  R

421     AATGTCACTATTAATGACACTATTAGAAACAGCAGTGTTATTGGTGACATGAAAGAAGAA
141      N  V  T  I  N  D  T  I  R  N  S  S  V  I  G  D  M  K  E  E

481     GTAACAAATTGCTCTTTCAATATAACCACAGAACTAAGAGATAAGAGACAAAAAGTATAT
161      V  T  N  C  S  F  N  I  T  T  E  L  R  D  K  R  Q  K  V  Y

541     TCACTTTTTTATAAACTTGATGTAGTACAAATTAATCCTGCTGATAAGAATAGTACCCAA
181      S  L  F  Y  K  L  D  V  V  Q  I  N  P  A  D  K  N  S  T  Q

601     TATAGACTAATAAATTGTAATACCTCAACCATTACACAGGCTTGTCCAAAGGTATCCTTT
201      Y  R  L  I  N  C  N  T  S  T  I  T  Q  A  C  P  K  V  S  F

661     GAGCCAATTCCCATACATTATTGTGCTCCAGCTGGTTTTGCGATTCTAAAGTGTAATGAT
221      E  P  I  P  I  H  Y  C  A  P  A  G  F  A  I  L  K  C  N  D
```

```
721     AAGAGGTTCAATGGAACAGGGACATGCAATAATGTCAGTACAGTACAGTGCACACATGGA
241      K  R  F  N  G  T  G  T  C  N  N  V  S  T  V  Q  C  T  H  G

781     ATCAGGCCAGTAGTATCAACTCAATTGTTGTTAAATGGCAGTTTAGCAGAAGAAAAGATA
261      I  R  P  V  V  S  T  Q  L  L  L  N  G  S  L  A  E  E  K  I

841     ATGATTAGATCTGAAAATATCACAAACAATGCCAAAATCATAATAGTACAGCTTAATGAG
281      M  I  R  S  E  N  I  T  N  N  A  K  I  I  I  V  Q  L  N  E

901     ACTGTACAAATTAATTGTACCAGGCCTAACAACAATACAAGAACAAGTGTACGTATAGGT
301      T  V  Q  I  N  C  T  R  P  N  N  N  T  R  T  S  V  R  I  G

961     CCAGGACAAACATTCTATGCAACAGGGGAAATCATAGGGGATATAAGAAAAGCACATTGT
321      P  G  Q  T  F  Y  A  T  G  E  I  I  G  D  I  R  K  A  H  C

1021    AATGTCAGTGAAAGAGAATGGATGAAAACTTTATATCATGTAGCTAAAAGATTAAGAGAG
341      N  V  S  E  R  E  W  M  K  T  L  Y  H  V  A  K  R  L  R  E

1081    GTACACTTTGAAAACAAGACAATAATCTTTAATAAGTCCTCAGGAGGGGATTTAGAAATT
361      V  H  F  E  N  K  T  I  I  F  N  K  S  S  G  G  D  L  E  I

1141    ACAACACATAGTTTTAATTGTGGAGGAGAATTCTTCTATTGCAATACATCAGGCCTGTTT
381      T  T  H  S  F  N  C  G  G  E  F  F  Y  C  N  T  S  G  L  F

1201    AACAGCACTTGGGAGTTTAACAACCCTTTTAATGAAACTGAGTGGCCACAAAATAAAACT
401      N  S  T  W  E  F  N  N  P  F  N  E  T  E  W  P  Q  N  K  T

1261    ATAATTCTCCAATGCAGAATAAAGCAAATTGTAAATATGTGGCAGAGAGTAGGACAGGCG
421      I  I  L  Q  C  R  I  K  Q  I  V  N  M  W  Q  R  V  G  Q  A

1321    ATGTATGCCCCTCCCATCAAAGGAGTAATAAGGTGTAATTCAACCATTACAGGACTATTC
441      M  Y  A  P  P  I  K  G  V  I  R  C  N  S  T  I  T  G  L  F

1381    TTAACAAGAGATGGTGGAAATACTAGCAGTACAAATGAGACCTTCAGGCCTGGAGGAGGA
461      L  T  R  D  G  G  N  T  S  S  T  N  E  T  F  R  P  G  G  G

1441    GATATGAGGGACAATTGGAGAAGTGAATTGTATAAATATAAAGTAATAAAAATTGAACCA
481      D  M  R  D  N  W  R  S  E  L  Y  K  Y  K  V  I  K  I  E  P

1501    ATAGGAGTAGCACCCACCAGGGCAAAAAGAAGAGTGGTGGAAAGAGAAAAAAGAGCAGTT
501      I  G  V  A  P  T  R  A  K  R  R  V  V  E  R  E  K  R  A  V

1561    GGACTGGGAGCTGTTTTCCTTGGGTTCTTAGGAGCAGCAGGAAGCACTATGGGCGCAGCG
521      G  L  G  A  V  F  L  G  F  L  G  A  A  G  S  T  M  G  A  A
```

```
1621    TCAATAACGCTGACGGTACAGGCCAGACAATTATTGTCTGGCATAGTGCAACAGCAAAGC
541      S  I  T  L  T  V  Q  A  R  Q  L  L  S  G  I  V  Q  Q  Q  S

1681    AATTTGCTGAGGGCTATAGAGGCTCAACAGCATCTGTTGAAACTCACAGTCTGGGGCATT
561      N  L  L  R  A  I  E  A  Q  Q  H  L  L  K  L  T  V  W  G  I

1741    AAACAGCTCCAGGCAAGAGTCCTGGCTGTGGAAAGATACCTAAAGGATCAACAGCTTCTA
581      K  Q  L  Q  A  R  V  L  A  V  E  R  Y  L  K  D  Q  Q  L  L

1801    GGAATTTGGGGCTGCTCAGGAAAGTTAATCTGCACCACTGCTGTGCCCTGGAACTCTAGT
601      G  I  W  G  C  S  G  K  L  I  C  T  T  A  V  P  W  N  S  S

1861    TGGAGTAATAAATCTTACAATGAAATATGGGATAACATGACCTGGATGCAATGGGAAAAG
621      W  S  N  K  S  Y  N  E  I  W  D  N  M  T  W  M  Q  W  E  K

1921    GAAATTGACAATTACACAGGCATAATATATACTCTAATTGAAGAATCGCAGAACCAACAG
641      E  I  D  N  Y  T  G  I  I  Y  T  L  I  E  E  S  Q  N  Q  Q

1981    GAAAAGAATGAACAAGATTTATTGGCATTGGACAAGTGGGCAAGTCTGTGGAATTGGTTT
661      E  K  N  E  Q  D  L  L  A  L  D  K  W  A  S  L  W  N  W  F

2041    GACATAACAAATTGGCTATGGTATATAAAAATATTCATAATGAAGGTAGGGGGCTTGATA
681      D  I  T  N  W  L  W  Y  I  K  I  F  I  M  K  V  G  G  L  I

2101    GGTTTAAGAATAATTTTTACTGTACTCTCTATAGTGAATAGAGTTAGGCAGGGATACTCA
701      G  L  R  I  I  F  T  V  L  S  I  V  N  R  V  R  Q  G  Y  S

2161    CCTTTGTCGTTTCAGACCCTTACCCCAAACCCGAGGGAACTCCACAGGCTCGGAAGAATC
721      P  L  S  F  Q  T  L  T  P  N  P  R  E  L  H  R  L  G  R  I
721                        P  L  P  Q  T  R  G  N  S  T  G  S  E  E  S
721                         P  Y  P  K  P  E  G  T  P  Q  A  R  K  N  R

2221    GAAGAAGAAGGTGGAGAGCCAGACAGAGACAGATCAGTTCGCTTAGTGAGCGGATTCTTA
741      E  E  E  G  G  E  P  D  R  D  R  S  V  R  L  V  S  G  F  L
741       K  K  K  V  E  S  Q  T  E  T  D  Q  F  A
741         R  R  R  W  R  A  R  Q  R  Q  I  S  S  L  S  E  R  I  L  S

2281    GCACTTTTCTGGGACGACCTACGGAACCTGTGCCTCTTCAGTTACCACCGCTTGAGAGAC
761      A  L  F  W  D  D  L  R  N  L  C  L  F  S  Y  H  R  L  R  D
761         T  F  L  G  R  P  T  E  P  V  P  L  Q  L  P  P  L  E  R  L
```

```
2341    TTCATCTTGATTGCAGCGAGGACTGTGGAACTTCTGGGACACAACAGTCTCAAGGGACTG
781      F  I  L  I  A  A  R  T  V  E  L  L  G  H  N  S  L  K  G  L
781        H  L  D  C  S  E  D  C  G  T  S  G  T  Q  Q  S  Q  G  T  E

2401    AGACTGGGGTGGGAAGGAATCAAGTATCTGTGGAATCTCCTGTTATATTGGGGTCAGGAA
801      R  L  G  W  E  G  I  K  Y  L  W  N  L  L  L  Y  W  G  Q  E
801        T  G  V  G  R  N  Q  V  S  V  E  S  P  V  I  L  G  S  G  T

2461    CTAAAGAATAGTGCTATCTCTCTGTTTGATGCTACAGCAATAACAGTAGCTGGGTGGACA
821      L  K  N  S  A  I  S  L  F  D  A  T  A  I  T  V  A  G  W  T
821        K  E

2521    GACAGGGTTATAGAACTAGGACAAAGAATTGTTAGAGCTTTTCTCCACATACCTAGAAGA
841      D  R  V  I  E  L  G  Q  R  I  V  R  A  F  L  H  I  P  R  R

2581    ATCAGACAGGGCTTCGAAAGAGCTTTGCTATAGCATGGGGGGCAAGTGGTCAAAAAGTAG
861      I  R  Q  G  F  E  R  A  L  L
861                                   M  G  G  K  W  S  K  S  S

2641    CATAGTGGGGTGGCCTGCGATTAGGGAGAGAATAAGAAGGACTGAGCCAGCAGCAGAGGG
881        I  V  G  W  P  A  I  R  E  R  I  R  R  T  E  P  A  A  E  G

2701    AGTAGGAGCAGCGTCTCGAGACTTGGATAAACATGGGGCACTTACAACCAGCAACACAGT
901        V  G  A  A  S  R  D  L  D  K  H  G  A  L  T  T  S  N  T  V

2761    CGCCAACAATGCTGCTTGTGCCTGGCTGGAAGCACAAGAGGAAGAAGGAGAGGTAGGCTT
921        A  N  N  A  A  C  A  W  L  E  A  Q  E  E  E  G  E  V  G  F

2821    TCCAGTCAGACCCCAGGTACCTTTAAGACCAATGACTTTTAAGGCAGCATTTGATCTCAG
941        P  V  R  P  Q  V  P  L  R  P  M  T  F  K  A  A  F  D  L  S

2881    CTTCTTTTTAAAAGAAAAGGGGGGACTGGAAGGGTTAATTTACTCCAGGAAAAGGCAAGA
961        F  F  L  K  E  K  G  G  L  E  G  L  I  Y  S  R  K  R  Q  E

2941    GATCCTTGATTGTGGTCCCTGAGCGC
981        I  L  D
```

Sample TV 314 from 1235 – 9551 (coordinate relative to HXB2)
Nucleotide and amino acid composition

gag – Red
pol - Blue
vif - Orange
vpr - Green
tat - Pink
rev – Dark Red
vpu – Sky Blue
env – Sea Green
nef – Dark Teal

```
1       TCAGGACTTTGGATGCATGGGTAAAAGTAATAGAAGAAAAGGCTTTCAGCCCTGAAGTAA
1          R  T  L  D  A  W  V  K  V  I  E  E  K  A  F  S  P  E  V  I

61      TACCCATGTTCTCAGCATTATCAGAAGGAGCCACCCCACAAGATTTAAATATGATGCTGA
21         P  M  F  S  A  L  S  E  G  A  T  P  Q  D  L  N  M  M  L  N

121     ACATAGTGGGGGGACACCAGGCAGCTATGCAAATGTTAAAGGATACCATCAATGAGGAAG
41         I  V  G  G  H  Q  A  A  M  Q  M  L  K  D  T  I  N  E  E  A

181     CTGCAGAATGGGATAGGCTACATCCAGTACATGCAGGGCCAGTTGCACCAGGCCAGATGA
61         A  E  W  D  R  L  H  P  V  H  A  G  P  V  A  P  G  Q  M  R

241     GAGAACCAAGGGGAAGTGATATAGCAGGAACTACTAGTACCCCTCAAGAACAAATAGCAT
81         E  P  R  G  S  D  I  A  G  T  T  S  T  P  Q  E  Q  I  A  W

301     GGATGACAGGCAACCCACCTATCCCAGTGGGAGACATCTATAAAAGATGGATAATCCTAG
101        M  T  G  N  P  P  I  P  V  G  D  I  Y  K  R  W  I  I  L  G

361     GGTTAAATAAAATAGTAAGAATGTATAGCCCTGTTAGCATTTTGGATATAAAACAAGGGC
121        L  N  K  I  V  R  M  Y  S  P  V  S  I  L  D  I  K  Q  G  P

421     CAAAAGAACCCTTCAGAGACTATGTAGATAGGTTCTTTAAAACTCTCAGGGCTGAGCAAG
141        K  E  P  F  R  D  Y  V  D  R  F  F  K  T  L  R  A  E  Q  A

481     CTACACAGGAAGTAAAAAATTGGATGACAGAAACATTATTAGTACAAAATGCAAATCCAG
161        T  Q  E  V  K  N  W  M  T  E  T  L  L  V  Q  N  A  N  P  D

541     ATTGTAAGTCCATTTTAAGAGCATTAGGACCAGGGGCTACATTAGAAGAAATGATGACAG
181        C  K  S  I  L  R  A  L  G  P  G  A  T  L  E  E  M  M  T  A

601     CATGCCAGGGAGTGGGAGGACCTAGCCATAAAGCAAGGGTTTTAGCCGAGGCAATGAGTC
201        C  Q  G  V  G  G  P  S  H  K  A  R  V  L  A  E  A  M  S  Q
```

```
661     AAGCACAACAAACAAACATACTGGTGCAGAGAGGCAATTTTGGGGGTCATAAAAGGATTA
221       A  Q  Q  T  N  I  L  V  Q  R  G  N  F  G  G  H  K  R  I  K

721     AGTGTTTCAACTGTGGCAAAGAAGGACACCTAGCCAGAAATTGCAGGGCCCCTAGGAAAA
241       C  F  N  C  G  K  E  G  H  L  A  R  N  C  R  A  P  R  K  K

781     AGGGCTGTTGGAAATGTGGGAAAGAGGGACACCAAATGAAAGACTGCACTGAAAGACAGG
261       G  C  W  K  C  G  K  E  G  H  Q  M  K  D  C  T  E  R  Q  A

841     CTAATTTTTTAGGGAAAATTTGGCCTTCCCACAAGGGGAGGCCAGGGAACTTCCCTCAGA
281         F  F  R  E  N  L  A  F  P  Q  G  E  A  R  E  L  P  S  E
281       N  F  L  G  K  I  W  P  S  H  K  G  R  P  G  N  F  P  Q  S

901     GCAGACCAGAGCCAACAGCCCCACCAGCAGAGATGTTTGGGATGAGGGAAGAGATAGCCT
301     Q  T  R  A  N  S  P  T  S  R  D  V  W  D  E  G  R  D  S  L
301       R  P  E  P  T  A  P  P  A  E  M  F  G  M  R  E  E  I  A  S

961     CCCCTCCGAAGCAGGAGCAGAACAGCAGGGACCAGAACCCACCTTCAATTTCCCTCAAAT
321     P  S  E  A  G  A  E  Q  Q  G  P  E  P  T  F  N  F  P  Q  I
321       P  P  K  Q  E  Q  N  S  R  D  Q  N  P  P  S  I  S  L  K  S

1021    CACTCTTTGGCAACGACCTATTGTCACAGTAAGAATAGAGGGACAGCTAAAGGAAGCTCT
341     T  L  W  Q  R  P  I  V  T  V  R  I  E  G  Q  L  K  E  A  L
341       L  F  G  N  D  L  L  S  Q

1081    ATTAGATACAGGAGCAGATGATACAGTATTAGAAGACATAAATTTGCCAGGGAAATGGAA
361     L  D  T  G  A  D  D  T  V  L  E  D  I  N  L  P  G  K  W  K

1141    ACCAAAAATGATAGGGGGAATTGGAGGTTTCATCAAAGTAAGACAGTATGATCAAATACT
381     P  K  M  I  G  G  I  G  G  F  I  K  V  R  Q  Y  D  Q  I  L

1201    TATAGAAATTTGTGGAAAAAAGGCTATGGGTACAGTATTGGTAGGACCTACACCTGTCAA
401     I  E  I  C  G  K  K  A  M  G  T  V  L  V  G  P  T  P  V  N

1261    CATAATTGGAAGAAATATGTTGACCCAGATTGGTTGTACTTTAAATTTCCCAATTAGCCC
421     I  I  G  R  N  M  L  T  Q  I  G  C  T  L  N  F  P  I  S  P

1321    TATCGATACTGTACCAGTAAAATTAAAGCCAGGAATGGATGGCCCAAAGGTTAAACAATG
441     I  D  T  V  P  V  K  L  K  P  G  M  D  G  P  K  V  K  Q  W

1381    GCCATTGACAGAAGAAAAAATAAAAGCATTAACAGAAATTTGTACAGAAATGGAAAAGGA
461     P  L  T  E  E  K  I  K  A  L  T  E  I  C  T  E  M  E  K  E
```

```
1441    AGGAAAAATTTCAAAAATTGGGCCTGAAAATCCATACAATACTCCAATATTTGCTATAAA
481      G  K  I  S  K  I  G  P  E  N  P  Y  N  T  P  I  F  A  I  K

1501    GAAAAAAGACAGCACTAAATGGAGAAAATTAGTAGATTTCAGAGAGCTCAATAAAAGAAC
501      K  K  D  S  T  K  W  R  K  L  V  D  F  R  E  L  N  K  R  T

1561    TCAAGACTTTTGGGAAGTTCAATTAGGAATACCGCATCCAGCGGGCTTAAAAAAGAAAAA
521      Q  D  F  W  E  V  Q  L  G  I  P  H  P  A  G  L  K  K  K  K

1621    ATCAGTAACAGTACTAGATGTGGGGGACGCATATTTTTCAGTTCCCTTAGATGAAAGTTT
541      S  V  T  V  L  D  V  G  D  A  Y  F  S  V  P  L  D  E  S  F

1681    TAGAAAGTATACTGCATTCACCATACCTAGTATAAACAATGAGACACCAGGAATCAGGTA
561      R  K  Y  T  A  F  T  I  P  S  I  N  N  E  T  P  G  I  R  Y

1741    TCAGTACAATGTGCTTCCACAGGGATGGAAAGGATCACCAGCAATATTCCAGAGTAGCAT
581      Q  Y  N  V  L  P  Q  G  W  K  G  S  P  A  I  F  Q  S  S  M

1801    GACAAAAATCTTAGATCCCTTTAGGTCAAAAAATCCAGAACTAATTATCTATCAATACAT
601      T  K  I  L  D  P  F  R  S  K  N  P  E  L  I  I  Y  Q  Y  M

1861    GGATGACTTGTATGTAGGATCTGATTTAGAAATAGGGCAGCATAGAGCAAAAATAGAAGA
621      D  D  L  Y  V  G  S  D  L  E  I  G  Q  H  R  A  K  I  E  E

1921    GTTGAGAGCTCATCTATTAAGCTGGGGATTTACTACACCAGACAAAAAGCATCAGAAAGA
641      L  R  A  H  L  L  S  W  G  F  T  T  P  D  K  K  H  Q  K  E

1981    GCCTCCATTCCTTTGGATGGGATATGAACTCCATCCTGACAAGTGGACAGTCCAACCTAT
661      P  P  F  L  W  M  G  Y  E  L  H  P  D  K  W  T  V  Q  P  I

2041    ACAGCTGCCAGAAAAAGACAGTTGGACTGTCAATGATATACAGAAGCTAGTGGGGAAACT
681      Q  L  P  E  K  D  S  W  T  V  N  D  I  Q  K  L  V  G  K  L

2101    AAATTGGGCAAGTCAGATTTACCCAGGGATTCAAGTAAGACAATTGTGTAAACTCCTCAG
701      N  W  A  S  Q  I  Y  P  G  I  Q  V  R  Q  L  C  K  L  L  R

2161    GGGAGCCAAAGCACTAACAGATATAGTAACATTGACTGAGGAAGCAGAATTAGAATTGGC
721      G  A  K  A  L  T  D  I  V  T  L  T  E  E  A  E  L  E  L  A

2221    AGAGAACAGGGAAATTCTAAAAGACCCTGTGCATGGAGTCTATTATGACCCATCAAAAGA
741      E  N  R  E  I  L  K  D  P  V  H  G  V  Y  Y  D  P  S  K  D

2281    CTTAATAACAGAAATACAGAAACAAGGGCAAGACCAATGGACATATCAAATTTATCAAGA
761      L  I  T  E  I  Q  K  Q  G  Q  D  Q  W  T  Y  Q  I  Y  Q  E
```

2341 ACCATTTAAAAATCTAAAAACAGGAAAATATGCAAGAAGGAGGTCTGCTCACACTAATGA
781 P F K N L K T G K Y A R R R S A H T N D

2401 TGTAAAACAGTTAACAGAAGTGGTGCAAAAAGTGGCCACGGAAAGTATAGTAATATGGGG
801 V K Q L T E V V Q K V A T E S I V I W G

2461 AAAGACTCCTAAATTTAGACTACCCATACAAAAAGAAACATGGGAAACATGGTGGATGGA
821 K T P K F R L P I Q K E T W E T W W M D

2521 CTATTGGCAGGCTACCTGGATCCCTGAATGGGAATTTGTCAATACCCCTCCCCTAGTAAA
841 Y W Q A T W I P E W E F V N T P P L V K

2581 ATTATGGTACCAGTTAGAAAAAGACCCCATAGCAGGAGCAGAGACTTTCTATGTAGATGG
861 L W Y Q L E K D P I A G A E T F Y V D G

2641 GGCATCCAGTAGGGAGACTAAGTTAGGAAAAGCAGGGTATGTCACTGACAGAGGAAGACA
881 A S S R E T K L G K A G Y V T D R G R Q

2701 AAAGGTTGTTTCCCTAACTGAAACAACAAATCAAAAAGCTGAATTACATGCAATCCATCT
901 K V V S L T E T T N Q K A E L H A I H L

2761 AGCCTTGCAGGATTCAGGATCAGAAGTAAACATAGTAACAGACTCACAGTATGCATTAGG
921 A L Q D S G S E V N I V T D S Q Y A L G

2821 CATCATTCAGGCACAACCAGACAGGAGTGAGTCAGAATTAGTCAATCAAATAATAGAGAA
941 I I Q A Q P D R S E S E L V N Q I I E K

2881 GCTAATAGGAAAAGATAAAGTCTACCTGTCATGGGTACCAGCACACAAGGGAATTGGAGG
961 L I G K D K V Y L S W V P A H K G I G G

2941 AAATGAACAAGTAGATAAACTGGTCAGTTCTGGAATCAGGAAGGTGCTATTTCTAGATGG
981 N E Q V D K L V S S G I R K V L F L D G

3001 GATAGATAAGGCTCAAGAAGAACATGAAAGATATCACAGCAACTGGAGAGCTATGGCTAG
1001 I D K A Q E E H E R Y H S N W R A M A S

3061 TGATTTTAATCTGCCACCTATAGTAGCAAAGGAGATAGTAGCCAGCTGTGATAAATGCCA
1021 D F N L P P I V A K E I V A S C D K C Q

3121 GCTAAAAGGGGAAGCCATGCATGGACAAGTAGACTGCAGTCCAGGAATATGGCAATTAGA
1041 L K G E A M H G Q V D C S P G I W Q L D

3181 CTGCACACATCTAGAAGGAAAAGTAATTCTGGTAGCAGTCCATGTAGCCAGTGGCTATAT
1061 C T H L E G K V I L V A V H V A S G Y I

```
3241    AGAAGCAGAAGTTATCCCAGCAGAAACAGGACAAGAGACAGCATACTTTCTATTAAAATT
1081     E  A  E  V  I  P  A  E  T  G  Q  E  T  A  Y  F  L  L  K  L

3301    AGCAGGAAGATGGCCAGTAAAAACAGTACACACAGACAATGGCAGCAATTTCACCAGTGC
1101     A  G  R  W  P  V  K  T  V  H  T  D  N  G  S  N  F  T  S  A

3361    TGCAGTTAAAGCAGCCTGTTGGTGGGCAGGTATCAAACAGGAATTTGGAATTCCCTACAA
1121     A  V  K  A  A  C  W  W  A  G  I  K  Q  E  F  G  I  P  Y  N

3421    TCCCCAAAGTCAAGGAGTAGTGGAATCTATGAATAAGGAATTAAAGAAAATCATAGGGCA
1141     P  Q  S  Q  G  V  V  E  S  M  N  K  E  L  K  K  I  I  G  Q

3481    GGTAAGAGAGCAAGCTGAACACCTTAAGACAGCAGTACAAATGGCAGTATTCATTCACAA
1161     V  R  E  Q  A  E  H  L  K  T  A  V  Q  M  A  V  F  I  H  N

3541    TTTTAAAAGAAAAGGGGGGATTGGGGGGTACAGTGCAGGGGAAAGAATAATAGACATAAT
1181     F  K  R  K  G  G  I  G  G  Y  S  A  G  E  R  I  I  D  I  I

3601    AGCAACAGACATACAAACTAAAGAATTACAAAAACAAATTACAAAAATTCAAAAATTTCG
1201     A  T  D  I  Q  T  K  E  L  Q  K  Q  I  T  K  I  Q  K  F  R

3661    GGTTTATTACAGGGACAGCAGAGATCCAATTTGGAAAGGACCAGCAAAACTACTCTGGAA
1221     V  Y  Y  R  D  S  R  D  P  I  W  K  G  P  A  K  L  L  W  K

3721    AGGTGAAGGGGCAGTGGTAATACAGGACAATAGTGATATAAAGGTAGTACCAAGAAGAAA
1241     G  E  G  A  V  V  I  Q  D  N  S  D  I  K  V  V  P  R  R  K

3781    AGCAAAGATCCTTAAGGATTATGGAAAACAGATGGCAGGTGATGATTGTGTGGCAGGTAG
1261     A  K  I  L  K  D  Y  G  K  Q  M  A  G  D  D  C  V  A  G  R
1261                         M  E  N  R  W  Q  V  M  I  V  W  Q  V  D

3841    ACAGGATGAGGATTAGAACATGGCACAGTTTAGTAAAACATCATATGTATGTCTCAAGGA
1281     Q  D  E  D
1281      R  M  R  I  R  T  W  H  S  L  V  K  H  H  M  Y  V  S  R  K

3901    AAACTAAAGATTGGTCTTATAGACATCACTATGAAAGCAGACATCCAAGAGTAAGTTCAG
1301      T  K  D  W  S  Y  R  H  H  Y  E  S  R  H  P  R  V  S  S  E

3961    AAGTACACATCCCACTAGGGGACGCTAGAATAATAGTAAAAACATATTGGGGTCTGCATA
1321      V  H  I  P  L  G  D  A  R  I  I  V  K  T  Y  W  G  L  H  T

4021    CAGGAGAAAAAGACTGGCAATTGGGTCATGGGGTCTCCATAGAATGGAGGCTGAAAAGCT
1341      G  E  K  D  W  Q  L  G  H  G  V  S  I  E  W  R  L  K  S  Y
```

```
4081    ATAACACACAAATAGACCCTGACCTGGCAGACCAACTAATTCATCTGCATTATTTTGAAT
1361      N  T  Q  I  D  P  D  L  A  D  Q  L  I  H  L  H  Y  F  E  C

4141    GTTTTTCAGATTCTGCCATAAGGAAAGCCATATTAGGGCGAGTAGTTAACCCTAGGTGTG
1381      F  S  D  S  A  I  R  K  A  I  L  G  R  V  V  N  P  R  C  E

4201    AATATCAAACAGGAAATAAAAAGGTAGGATCTCTACAATATTTAGCACTAAAAGCATTAG
1401      Y  Q  T  G  N  K  K  V  G  S  L  Q  Y  L  A  L  K  A  L  V

4261    TAGGACCAAAAAAGACAAAGCCACCTTTGCCTAGTGTTAGTAAACTAACAGAGGATAGAT
1421                                                              M
1421      G  P  K  K  T  K  P  P  L  P  S  V  S  K  L  T  E  D  R  W

4321    GGAACAAGCCCCAGAAGACCAGGGGCCCCAGAGAGAGCCATACAATGAATGGATGTTAGA
1441    E  Q  A  P  E  D  Q  G  P  Q  R  E  P  Y  N  E  W  M  L  E
1441      N  K  P  Q  K  T  R  G  P  R  E  S  H  T  M  N  G  C

4381    GCTGTTAGAAGAACTTAAGCATGAAGCTGTTAGACATTTCCCTAGACCATGGCTCCAGGG
1461    L  L  E  E  L  K  H  E  A  V  R  H  F  P  R  P  W  L  Q  G

4441    ACTAGGACAATATATCTACAACACCCATGGGGATACTTGGGAAGGAGTTGAAGCTATTAT
1481    L  G  Q  Y  I  Y  N  T  H  G  D  T  W  E  G  V  E  A  I  I

4501    AAGAATTTTGCAGCAACTACTGTTTGTTCATTTCAGGATTGGGTGCCAACACAGCAGAAT
1501    R  I  L  Q  Q  L  L  F  V  H  F  R  I  G  C  Q  H  S  R  I

4561    AGGCATTATTCGAGGGAGAAGAGTCAGAAATGGATCCAGTAGATCCTAACCTAGAGCCCT
1521    G  I  I  R  G  R  R  V  R  N  G  S  S  R  S
1521                                M  D  P  V  D  P  N  L  E  P  W

4621    GGAACCATCCAGGAAGTCAGCCTACAACTCCTTGTAGCAAGTGTTACTGTAAAGCGTGTT
1541      N  H  P  G  S  Q  P  T  T  P  C  S  K  C  Y  C  K  A  C  C

4681    GCTACCATTGCTTAGTTTGCTTTCAGACCAAAGGCTTAGGCATCTCCTATGGCAGGAAGA
1561                                                M  A  G  R
1561      Y  H  C  L  V  C  F  Q  T  K  G  L  G  I  S  Y  G  R  K  K

4741    AGCGGAGACAGCGACGAGGCACTCCTCACAGCCGTACGGATCATCAAAATCCTGTATCAA
1581    S  G  D  S  D  E  A  L  L  T  A  V  R  I  I  K  I  L  Y  Q
1581      R  R  Q  R  R  G  T  P  H  S  R  T  D  H  Q  N  P  V  S  K

4801    AGCAGTAAGTGTTTATATATGTAATGACCCCTTTAGAAATTAGTGCAATAATAGGATTGA
1601    S
1601      Q                 M  T  P  L  E  I  S  A  I  I  G  L  I
```

```
4861    TAGTAGCGCTAATCTTAGCAATAGTTGTATGGACTATAGTAGGTATAGAATATAAGAAAA
1621      V  A  L  I  L  A  I  V  V  W  T  I  V  G  I  E  Y  K  K  I

4921    TAAGAAGGCAAAGAAAAATAGACAGGTTACTTGAGAGAATAAGAGAAAGAGCAGAAGACA
1641      R  R  Q  R  K  I  D  R  L  L  E  R  I  R  E  R  A  E  D  S

4981    GTGGCAATGAGAGTGAGGGGGATACAGATGACTTGGCAGCACTTATTGGGATGGGGAATT
1661         M  R  V  R  G  I  Q  M  T  W  Q  H  L  L  G  W  G  I
1661      G  N  E  S  E  G  D  T  D  D  L  A  A  L  I  G  M  G  N  Y

5041    ATGATCTTGGGGATGATTATAATGTGTAGTACTGCAGACAACTTGTGGGTTACTGTTTAC
1681     M  I  L  G  M  I  I  M  C  S  T  A  D  N  L  W  V  T  V  Y
1681      D  L  G  D  D  Y  N  V

5101    TATGGGGTACCTGTGTGGAAAGATGCAGAGACCACCCTATTTTGTGCATCAGATGCTAAA
1701     Y  G  V  P  V  W  K  D  A  E  T  T  L  F  C  A  S  D  A  K

5161    GCATATGAGAAAGAAGTGCATAATGTCTGGGCTACACATGCCTGTGTACCCACAGACCCC
1721     A  Y  E  K  E  V  H  N  V  W  A  T  H  A  C  V  P  T  D  P

5221    AACCCACAAGAAATACATTTGGTAAATGTGACAGAAAATTTTAATATGTGGAAAAATAAA
1741     N  P  Q  E  I  H  L  V  N  V  T  E  N  F  N  M  W  K  N  K

5281    ATGGTAGAGCAGATGCATGCAGATATAATCAGTCTATGGGACCAAAGCCTAAAGCCATGT
1761     M  V  E  Q  M  H  A  D  I  I  S  L  W  D  Q  S  L  K  P  C

5341    GTAAAGCTAACCCCTCTCTGTGTAACTTTAAATTGTACCAATGCCAATATCACCTATGTC
1781     V  K  L  T  P  L  C  V  T  L  N  C  T  N  A  N  I  T  Y  V

5401    AGTACCAACAGCACGAAGGCCTATGTCACTGTCAACGGCACAACGGAAGAAATAAAAAAC
1801     S  T  N  S  T  K  A  Y  V  T  V  N  G  T  T  E  E  I  K  N

5461    TGCTCTTATAATATGACCACAGAACTAAGGGATAAGAAACAGAAAGTATATTCACTTTTT
1821     C  S  Y  N  M  T  T  E  L  R  D  K  K  Q  K  V  Y  S  L  F

5521    TATAGACTTGATGTAGTACAGATTAATAAAAATAATAATAGTAGAGATAATGATAGTGGT
1841     Y  R  L  D  V  V  Q  I  N  K  N  N  N  S  R  D  N  D  S  G

5581    GAGTATAGATTAATAAATTGTAATACCTCAGCCATTACACAAGCTTGTCCAAAGGTCTCC
1861     E  Y  R  L  I  N  C  N  T  S  A  I  T  Q  A  C  P  K  V  S

5641    TTTGAGCCAATTCCCATACATTATTGTGCTCCAGCTGGTTTTGCGATCCTAAAATGTAAT
1881     F  E  P  I  P  I  H  Y  C  A  P  A  G  F  A  I  L  K  C  N
```

```
5701    GAGGAGGAGTTCAACGGAACAGGGCCATGCAAGAATGTCAGCTCAGTACAATGCACACAT
1901     E  E  E  F  N  G  T  G  P  C  K  N  V  S  S  V  Q  C  T  H

5761    GGAATCAGGCCAGTAGTATCAACTCAACTGCTGTTAAATGGCAGTCTAGCCCAAGGAGAG
1921     G  I  R  P  V  V  S  T  Q  L  L  L  N  G  S  L  A  Q  G  E

5821    GTAAAAATTAGATCTGAAAATATCTCAGACAATGCTAAAACCATAATAGTACAATTTAAC
1941     V  K  I  R  S  E  N  I  S  D  N  A  K  T  I  I  V  Q  F  N

5881    CAGTCTGTAATAATTAATTGTACCAGACCTAGCAACAATACAAGGAGAAGTGTACGTATA
1961     Q  S  V  I  I  N  C  T  R  P  S  N  N  T  R  R  S  V  R  I

5941    GGACCAGGACAAGCATTCTATGCAACAGGTGAGATAATAGGGGACATAAGGAAAGCACAT
1981     G  P  G  Q  A  F  Y  A  T  G  E  I  I  G  D  I  R  K  A  H

6001    TGTAATGTCAGTGAATCAGAATGGAATAAAGCTTTACAACAGGTAGCTACACAATTAGGA
2001     C  N  V  S  E  S  E  W  N  K  A  L  Q  Q  V  A  T  Q  L  G

6061    AGATACTGGAGTAACAAAACAATAATTTTTAATAGCTCCTCAGGAGGGGATTTAGAAATT
2021     R  Y  W  S  N  K  T  I  I  F  N  S  S  S  G  G  D  L  E  I

6121    ACAACACATAGTTTTAATTGTGGAGGAGTATTTTTCTATTGTAATACATCAGGTCTGTTT
2041     T  T  H  S  F  N  C  G  G  V  F  F  Y  C  N  T  S  G  L  F

6181    AGTAGCAGGTGGTTCACTAATGGCACTAACAGCACGGAGTCAAATGGCACAGGCAATATA
2061     S  S  R  W  F  T  N  G  T  N  S  T  E  S  N  G  T  G  N  I

6241    ACTCTCCAATGCAGGATAAAGCAAATTATAAATATGTGGCAGAGAGTAGGACAAGCAACG
2081     T  L  Q  C  R  I  K  Q  I  I  N  M  W  Q  R  V  G  Q  A  T

6301    TACACCCCTCCCATCCAAGGAGAAATAAGGTGTAGATCAAACATTACAGGACTACTATTA
2101     Y  T  P  P  I  Q  G  E  I  R  C  R  S  N  I  T  G  L  L  L

6361    ACAAGAGATGGTGGGATTAACACAACAGAGGAAATCTTCAGACCTGGAGGGGGAAATATG
2121     T  R  D  G  G  I  N  T  T  E  E  I  F  R  P  G  G  G  N  M

6421    AAGGACAATTGGAGAAGTGAATTATATAAGTATAAAGTAGTAAAAATTGAACCACTAGGA
2141     K  D  N  W  R  S  E  L  Y  K  Y  K  V  V  K  I  E  P  L  G

6481    GTAGCACCATCCAAGGCAAAGAGAAGAGTGGTGGGAAGAGAAAAAAGAGCAGTTGGACTG
2161     V  A  P  S  K  A  K  R  R  V  V  G  R  E  K  R  A  V  G  L

6541    GGAGCTGTATTCATTGGGTTCTTGGGAGCAGCAGGAAGCACTATGGGCGCGGCGTCAGTG
2181     G  A  V  F  I  G  F  L  G  A  A  G  S  T  M  G  A  A  S  V
```

```
6601    ACGCTGACGGTACAGGCCAGACAATTATTGTCTGGCATAGTGCAGCAGCAAAGCAATTTG
2201     T  L  T  V  Q  A  R  Q  L  L  S  G  I  V  Q  Q  Q  S  N  L

6661    CTGAGGGCTATAGAGGCTCAACAGCATCTGTTGAAACTCACAGTCTGGGGCATTAAACAG
2221     L  R  A  I  E  A  Q  Q  H  L  L  K  L  T  V  W  G  I  K  Q

6721    CTCCAGGCAAGAGTCCTGGCTCTAGAGAGATACCTAAGGGATCAACAGCTCCTAGGAATT
2241     L  Q  A  R  V  L  A  L  E  R  Y  L  R  D  Q  Q  L  L  G  I

6781    TGGGGCTGCTCTGGAAAACTCATTTGCGCCACTAATGTGCCTTGGAACTCTAGTTGGAGT
2261     W  G  C  S  G  K  L  I  C  A  T  N  V  P  W  N  S  S  W  S

6841    AATAAATCTTATAATGAAATATGGGATAACATGACCTGGCTGCAGTGGGATAAAGAAATT
2281     N  K  S  Y  N  E  I  W  D  N  M  T  W  L  Q  W  D  K  E  I

6901    GACAATTACACAGAAACAATATATAGGCTAATTGAAGAATCGCAAAACCAGCAGGAAAGG
2301     D  N  Y  T  E  T  I  Y  R  L  I  E  E  S  Q  N  Q  Q  E  R

6961    AATGAACAAGACTTATTGGCATTGGACAAGTGGACAAATCTGTGGAGTTGGTTTGACATA
2321     N  E  Q  D  L  L  A  L  D  K  W  T  N  L  W  S  W  F  D  I

7021    TCGAACTGGCTGTGGTATATAAAAATATTTATAATGATAGTAGGAGGCTTAATAGGATTA
2341     S  N  W  L  W  Y  I  K  I  F  I  M  I  V  G  G  L  I  G  L

7081    AGAATAGTTTTTGCTGTGCTTTCTATAATAAATAGAGTTAGGCAGGGATACTCACCTTTG
2361     R  I  V  F  A  V  L  S  I  I  N  R  V  R  Q  G  Y  S  P  L

7141    TCATTTCAGACCCATACCCCAAACCCAGGGGGACTCGACAGGCCCGAAAGAACAGAAGAA
2381     S  F  Q  T  H  T  P  N  P  G  G  L  D  R  P  E  R  T  E  E
2381              P  I  P  Q  T  Q  G  D  S  T  G  P  K  E  Q  K  K
2381               P  Y  P  K  P  R  G  T  R  Q  A  R  K  N  R  R  R

7201    GAAGGTGGAGTGCAAGGCAGAGACAGATCGATTCGATTAGTCAGCGGATTCTTAGCTCTT
2401     E  G  G  V  Q  G  R  D  R  S  I  R  L  V  S  G  F  L  A  L
2401      K  V  E  C  K  A  E  T  D  R  F  D
2401        R  W  S  A  R  Q  R  Q  I  D  S  I  S  Q  R  I  L  S  S  C

7261    GCCTGGGACGATCTGAGGAGCCTGTGCCTTTTCAGCTACCACCGCTTGAGAGACTTCATA
2421     A  W  D  D  L  R  S  L  C  L  F  S  Y  H  R  L  R  D  F  I
2421        L  G  R  S  E  E  P  V  P  F  Q  L  P  P  L  E  R  L  H  I

7321    TTGATTGCAGCGAGGACTGTGGAACTTCTGGGACACAGCAGTCTCAAGGGGCTGAGACTG
2441     L  I  A  A  R  T  V  E  L  L  G  H  S  S  L  K  G  L  R  L
2441        D  C  S  E  D  C  G  T  S  G  T  Q  Q  S  Q  G  A  E  T  G
```

```
7381    GGGTGGGAAGGAATCAAGTATCTGGGGAATCTCCTGTTGTATTGGATTCGGGAACTAAAG
2461     G  W  E  G  I  K  Y  L  G  N  L  L  L  Y  W  I  R  E  L  K
2461       V  G  R  N  Q  V  S  G  E  S  P  V  V  L  D  S  G  T  K  E

7441    AATAGTGCTATTAATTTGCTTGATACCATAGCAATAGCAGTAGCTGGCTGGACAGATAGG
2481     N  S  A  I  N  L  L  D  T  I  A  I  A  V  A  G  W  T  D  R

7501    GTTATAGAAGTAGGACAAAGATTTGGTAGAGCTATTCTCCACATACCTAGAAGGATCAGA
2501     V  I  E  V  G  Q  R  F  G  R  A  I  L  H  I  P  R  R  I  R

7561    CAAGGACTTGAAAGAGCTTTACTATAACATGGGTAGCAAGTGGTCAAAAAGCAGCATAGT
2521     Q  G  L  E  R  A  L  L
2521                                M  G  S  K  W  S  K  S  S  I  V

7621    GGGATGGCCCGAGGTTAGGGAAAGAATGAGACAAGCTCAAGCTCCTCCAGCAGCAAAGGG
2541       G  W  P  E  V  R  E  R  M  R  Q  A  Q  A  P  P  A  A  K  G

7681    AGTAGGAGCAGTATCTCAAGATCTAGAAAAACATGGAGCAATCACAAGCAGCAACATGAA
2561       V  G  A  V  S  Q  D  L  E  K  H  G  A  I  T  S  S  N  M  N

7741    TCATCCTAGTTGTGTCTGGCTGGAAGCACAAGAAGAAGAGGAGGTAGGCTTTCCAGTCAG
2581       H  P  S  C  V  W  L  E  A  Q  E  E  E  E  V  G  F  P  V  R

7801    GCCACAAGTACCTTTAAGACCAATGACTTATAAGGGAGCTCTGGATCTCAGCCACTTTTT
2601       P  Q  V  P  L  R  P  M  T  Y  K  G  A  L  D  L  S  H  F  L

7861    AAAAGAAAAGGGGGGACTGGATGGGTTAATTTACTCCAAAAAGAGACAAGACATCCTTGA
2621       K  E  K  G  G  L  D  G  L  I  Y  S  K  K  R  Q  D  I  L  D

7921    TCTGTGGGTCTACAACACACAAGGCTATTTCCCTGATTGGCAGAATTACACACCAGGGCC
2641       L  W  V  Y  N  T  Q  G  Y  F  P  D  W  Q  N  Y  T  P  G  P

7981    AGGGATTAGATACCCACTAACATTTGGCAGGTGCTTTAAGCTAGTACCAGTGGATCCAGA
2661       G  I  R  Y  P  L  T  F  G  R  C  F  K  L  V  P  V  D  P  E

8041    GGAAGTAGAGAAGGCCAACGAGGGAGAGAACAACAGCCTATTACACCCGGTATGCCAACA
2681       E  V  E  K  A  N  E  G  E  N  N  S  L  L  H  P  V  C  Q  H

8101    TGGAATGGATGATGAGGACAGAGAAGTATTAAAGTGGAGCTTTGACAGTCGCCTGGCACT
2701       G  M  D  D  E  D  R  E  V  L  K  W  S  F  D  S  R  L  A  L

8161    AAAACACAGAGCACAAGAGCTGCATCCGGAGTTCTACAAAGACTGCTGACACAGGAATTG
2721       K  H  R  A  Q  E  L  H  P  E  F  Y  K  D  C
```

Sample TV412 from 1246 – 8254 (coordinates relative to HXB2)
Nucleotide and amino acid composition

gag – Red
pol - Blue
vif - Orange
vpr - Green
tat - Pink
rev – Dark Red
vpu – Sky Blue
env – Sea Green
nef – Dark Teal

```
1       TAGTTAGGACTTTGAATGCATGGGTAAAAGTAATAGAAGAAAAGGCTTTCAGCCCAGAAG
1            R  T  L  N  A  W  V  K  V  I  E  E  K  A  F  S  P  E  V

61      TAATACCCATGTTCTCAGCATTATCAGAAGGAGCCACCCCACAAGATTTAAATATGATGC
21         I  P  M  F  S  A  L  S  E  G  A  T  P  Q  D  L  N  M  M  L

121     TGAACATAGTGGGGGGACACCAGGCAGCCATGCAAATGTTAAAAGATACCATTAATGAGG
41         N  I  V  G  G  H  Q  A  A  M  Q  M  L  K  D  T  I  N  E  E

181     AAGCTGCAGAATGGGACAGAGTACATCCAGTACATGCAGGGCCTATTCCACCAGGCCAGA
61         A  A  E  W  D  R  V  H  P  V  H  A  G  P  I  P  P  G  Q  M

241     TGAGAGAACCAAGGGGAAGTGACATAGCAGGAACTACTAGTACCATTCAAGAACAAATAG
81         R  E  P  R  G  S  D  I  A  G  T  T  S  T  I  Q  E  Q  I  G

301     GATGGATGACAAGCAACCCACCTGTCCCAGTGGGAGAAATCTATAAAAGATGGATAATCC
101        W  M  T  S  N  P  P  V  P  V  G  E  I  Y  K  R  W  I  I  L

361     TGGGATTAAATAAAATAGTAAGAATGTATAGCCCTGTTAGCATTTTGGATATAAAACAAG
121        G  L  N  K  I  V  R  M  Y  S  P  V  S  I  L  D  I  K  Q  G

421     GGCCAAAAGAACCCTTCAGAGATTATGTAGATAGGTTCTTTAAAACTCTCAGAGCAGAGC
141        P  K  E  P  F  R  D  Y  V  D  R  F  F  K  T  L  R  A  E  Q

481     AAGCTACCCAGGAGGTAAAAGGTTGGATGACTGAAACATTACTGGTCCAAAATGCAAATC
161        A  T  Q  E  V  K  G  W  M  T  E  T  L  L  V  Q  N  A  N  P

541     CAGATTGTAAGTCCATTTTAAGAGCATTAGGACCAGGGGCTACATTAGAAGAAATGATGA
181        D  C  K  S  I  L  R  A  L  G  P  G  A  T  L  E  E  M  M  T

601     CAGCATGCCAGGGAGTGGGAGGACCCAGTCATAAAGCAAGAGTCTTGGCTGAGGCAATGA
201        A  C  Q  G  V  G  G  P  S  H  K  A  R  V  L  A  E  A  M  S
```

```
661     GCCAAGCAACAAGTGCAAATGCTGCTATAATGATGCAGAGAGGCAATTTTAAGGGTCCAA
221       Q  A  T  S  A  N  A  A  I  M  M  Q  R  G  N  F  K  G  P  R

721     GGAAAAGCATTAAGTGTTTCAACTGTGGCAAAGAAGGGCACCTAGCAAGAAACTGCAGGG
241       K  S  I  K  C  F  N  C  G  K  E  G  H  L  A  R  N  C  R  A

781     CTCCTAGGAAAAAGGGTTGTTGGAAATGTGGAAGGGAAGGACACCAAATGAGAGATTGCA
261       P  R  K  K  G  C  W  K  C  G  R  E  G  H  Q  M  R  D  C  T

841     CTGAAAGACAGGCTAATTTTTTAGGGAGAATTTGGCCTCTCAACAAGGGGAGGCCAGGAA
281                  F  F  R  E  N  L  A  S  Q  Q  G  E  A  R  K
281       E  R  Q  A  N  F  L  G  R  I  W  P  L  N  K  G  R  P  G  N

901     ATTTTCCTCAGAACAGACTGGAACCAACAGCTCCACCAATGGAGACCTTTGGGATGGGGG
301     F  S  S  E  Q  T  G  T  N  S  S  T  N  G  D  L  W  D  G  G
301       F  P  Q  N  R  L  E  P  T  A  P  P  M  E  T  F  G  M  G  E

961     AAGAGACAGCCTCCCCTCAGAAGCAGGAACAGAAAGGCAGGGAACAGTCCCAACCCTTAA
321     R  D  S  L  P  S  E  A  G  T  E  R  Q  G  T  V  P  T  L  N
321       E  T  A  S  P  Q  K  Q  E  Q  K  G  R  E  Q  S  Q  P  L  I

1021    TTTCCCTCAAATCACTCTTTGGCAACGACCCCTCGTCACAGTAAAGGTAGGGGGGCAGCT
341     F  P  Q  I  T  L  W  Q  R  P  L  V  T  V  K  V  G  G  Q  L
341       S  L  K  S  L  F  G  N  D  P  S  S  Q

1081    AAAAGAAGCTCTATTAGATACAGGAGCAGATGATACAGTATTAGAAGACATAAATTTGCC
361     K  E  A  L  L  D  T  G  A  D  D  T  V  L  E  D  I  N  L  P

1141    AGGAAAATGGAAACCAAAAATGATAGGGGGAATTGGAGGTTTCATTAAAGTAAAACAGTA
381     G  K  W  K  P  K  M  I  G  G  I  G  G  F  I  K  V  K  Q  Y

1201    TGATCAGATACTTATAGAAATTTGTGGAAAAAAGGCTATAGGTACAGTCTTAGTAGGACC
401     D  Q  I  L  I  E  I  C  G  K  K  A  I  G  T  V  L  V  G  P

1261    CACACCTGTCAACATAATTGGAAGAAACATGTTGACCCAGATTGGTTGTACTCTAAATTT
421     T  P  V  N  I  I  G  R  N  M  L  T  Q  I  G  C  T  L  N  F

1321    CCCAATTAGTCCTATTGAGACTGTACCAGTAAAATTAAAGCCAGGAATGGATGGCCCAAG
441     P  I  S  P  I  E  T  V  P  V  K  L  K  P  G  M  D  G  P  R

1381    GGTTAAACAATGGCCATTAACAGAAGAAAAAATAAAAGCATTGACAGAAATTTGTACAGA
461     V  K  Q  W  P  L  T  E  E  K  I  K  A  L  T  E  I  C  T  E
```

```
1441    GATGGAAAAGGAAGGAAAAATTTCAAAAATTGGGCCTGAAAATCCATACAATACTCCAAT
481      M  E  K  E  G  K  I  S  K  I  G  P  E  N  P  Y  N  T  P  I

1501    ATTTGCAATAAAGAAAAAAGATAGCACTAAATGGAGAAAATTAGTAGATTTCAGAGAGCT
501      F  A  I  K  K  K  D  S  T  K  W  R  K  L  V  D  F  R  E  L

1561    CAATAAAAGAACACAAGACTTTTGGGAAGTTCAATTGGGAATACCGCATCCAGCGGGCCT
521      N  K  R  T  Q  D  F  W  E  V  Q  L  G  I  P  H  P  A  G  L

1621    AAAAAAGAAAAAATCAGTAACAGTACTAGATGTGGGGGATGCATATTTTTCAGTTCCTTT
541      K  K  K  K  S  V  T  V  L  D  V  G  D  A  Y  F  S  V  P  L

1681    AGATGTAAACTTTAGAAAGTATACTGCATTCACCATACCTAGTAGAAACAATGAGACACC
561      D  V  N  F  R  K  Y  T  A  F  T  I  P  S  R  N  N  E  T  P

1741    AGGAATCAGGTATCAGTACAATGTGCTTCCACAGGGATGGAAAGGATCACCGGCAATATT
581      G  I  R  Y  Q  Y  N  V  L  P  Q  G  W  K  G  S  P  A  I  F

1801    CCAGAGTAGCATGACAAAAATCTTAGAGCCCTTTAGAACAAAAAATCCAGAACTAATTAT
601      Q  S  S  M  T  K  I  L  E  P  F  R  T  K  N  P  E  L  I  I

1861    CTATCAATACATGGATGACTTGTATGTAGGATCTGATTTAGAAATAGGACAGCATAGAAC
621      Y  Q  Y  M  D  D  L  Y  V  G  S  D  L  E  I  G  Q  H  R  T

1921    AAAAATAGAAGAGTTGAGAGCTCATCTATTGAGCTGGGGATTTACCACACCAGACAAAAA
641      K  I  E  E  L  R  A  H  L  L  S  W  G  F  T  T  P  D  K  K

1981    GCATCAGAAAGAACCTCCATTCCTTTGGATGGGATATGAGCTCCATCCTGACAAGTGGAC
661      H  Q  K  E  P  P  F  L  W  M  G  Y  E  L  H  P  D  K  W  T

2041    AGTCCAGCCTGTAAAGCTGCCAGAAAAAGAGCACTGGACTGTCAATGATATACAGAAATT
681      V  Q  P  V  K  L  P  E  K  E  H  W  T  V  N  D  I  Q  K  L

2101    AGTAGGGAAACTAAATTGGGCAAGTCAAATTTATGCAGGGATTAAAGTAAAGCAATTGTG
701      V  G  K  L  N  W  A  S  Q  I  Y  A  G  I  K  V  K  Q  L  C

2161    CAAGCTCCTCAGGGGAGCCAAAGCATTAACAGACATAGTAACATTGACTGAGGAAGCAGA
721      K  L  L  R  G  A  K  A  L  T  D  I  V  T  L  T  E  E  A  E

2221    ATTAGAATTGGCAGAAAACAGGGAGATTCTAAAAGACCCTGTGCATGGAGTATACTATGA
741      L  E  L  A  E  N  R  E  I  L  K  D  P  V  H  G  V  Y  Y  D

2281    CCCATCAAAAGACTTAATAGCAGAAATACAGAAACAGGGGCAAGACCAATGGACATATCA
761      P  S  K  D  L  I  A  E  I  Q  K  Q  G  Q  D  Q  W  T  Y  Q
```

```
2341    AATTTATCAAGAGCCATTTAAGAATCTGAAAACAGGGAAATATGCAAGAAAAAGATCAGC
781      I  Y  Q  E  P  F  K  N  L  K  T  G  K  Y  A  R  K  R  S  A

2401    ACACACTAATGATGTAAAACAATTAACAGAAGTGGTGCAAAAAGTGGTCATGGAAAGCAT
801      H  T  N  D  V  K  Q  L  T  E  V  V  Q  K  V  V  M  E  S  I

2461    AGTAATATGGGGAAAGACTCCTAAATTTAAACTACCCATACAAAAAGAAACATGGGAAAC
821      V  I  W  G  K  T  P  K  F  K  L  P  I  Q  K  E  T  W  E  T

2521    ATGGTGGATGGACTATTGGCAGGCTACCTGGATTCCTGAATGGGAATTTGTCAATACCCC
841      W  W  M  D  Y  W  Q  A  T  W  I  P  E  W  E  F  V  N  T  P

2581    TCCTCTAGTAAAATTGTGGTACCAATTAGAGAAAGACCCCATAATGGGAGCAGAGACTTT
861      P  L  V  K  L  W  Y  Q  L  E  K  D  P  I  M  G  A  E  T  F

2641    CTATGTAGATGGGGCAGCCAATAGGGAGACTAAGCTAGGAAAAGCAGGGTATGTCACTGA
881      Y  V  D  G  A  A  N  R  E  T  K  L  G  K  A  G  Y  V  T  D

2701    TAGGGGAAGACAAAAGGTTGTCTCCCTAACAGAGACAACAAATCAAAAAACTGAACTACA
901      R  G  R  Q  K  V  V  S  L  T  E  T  T  N  Q  K  T  E  L  H

2761    TGCAATCTATCTAGCCTTGCAGGATTCAGGATCAGAAGTAAACATAGTAACAGACTCACA
921      A  I  Y  L  A  L  Q  D  S  G  S  E  V  N  I  V  T  D  S  Q

2821    GTATGCATTAGGAATCATTCAGGCACAACCAGACAGGAGTGAATCAGAGTTAGTCAATCA
941      Y  A  L  G  I  I  Q  A  Q  P  D  R  S  E  S  E  L  V  N  Q

2881    AATAATAGAGAAGTTAATAGAAAAGGACAAAGTCTATCTGTCATGGGTACCAGCACACAA
961      I  I  E  K  L  I  E  K  D  K  V  Y  L  S  W  V  P  A  H  K

2941    AGGAATTGGAGGAAATGAACAAGTAGATAAATTAGTCAGTAATGGAATCAGGAAGATACT
981      G  I  G  G  N  E  Q  V  D  K  L  V  S  N  G  I  R  K  I  L

3001    ATTTTTAGATGGGATAGATAAGGCTCAAGAAGAACATGAAAGATATCATAGCAATTGGAG
1001     F  L  D  G  I  D  K  A  Q  E  E  H  E  R  Y  H  S  N  W  R

3061    AGCAATGGCTAATGATTTTAACCTGCCACCTGTGGTAGCAAAGGAAATAGTAGCCAGCTG
1021     A  M  A  N  D  F  N  L  P  P  V  V  A  K  E  I  V  A  S  C

3121    TGATAAATGTCAGCTAAAAGGGGAAGCCATGCATGGACAGGTAGACTGTAGTCCAGGAAT
1041     D  K  C  Q  L  K  G  E  A  M  H  G  Q  V  D  C  S  P  G  I

3181    ATGGCAATTAGATTGCACACATCTAGAAGGAAAAGTAATTCTGGTAGCAGTTCATGTAGC
1061     W  Q  L  D  C  T  H  L  E  G  K  V  I  L  V  A  V  H  V  A
```

```
3241    CAGTGGCTATATAGAAGCAGAAGTTATCCCAGCAGAAACAGGACAGGAGACAGCATACTT
1081     S  G  Y  I  E  A  E  V  I  P  A  E  T  G  Q  E  T  A  Y  F

3301    TCTGCTAAAATTAGCAGGAAGGTGGCCAGTAAAAGTAGTTCACACAGACAATGGCAGCAA
1101     L  L  K  L  A  G  R  W  P  V  K  V  V  H  T  D  N  G  S  N

3361    TTTCACCAGTGCTGCAGTTAAAGCAGCCTGTTGGTGGGCAAATATCCAGCAGGAATTTGG
1121     F  T  S  A  A  V  K  A  A  C  W  W  A  N  I  Q  Q  E  F  G

3421    GATTCCCTACAATCCCCAAAGTCAAGGAGTAGTAGAATCTATGAATAAAGAATTAAAGAA
1141     I  P  Y  N  P  Q  S  Q  G  V  V  E  S  M  N  K  E  L  K  K

3481    AATTATAGGACAGGTAAGAGATCAAGCTGAACATCTTAAGACAGCAGTACAAATGGCAGT
1161     I  I  G  Q  V  R  D  Q  A  E  H  L  K  T  A  V  Q  M  A  V

3541    ATTCATTCACAATTTTAAAAGAAAAGGGGGGATTGGGGGGTACAGTGCAGGGGAAAGAAT
1181     F  I  H  N  F  K  R  K  G  G  I  G  G  Y  S  A  G  E  R  I

3601    AATAGACATAATAGCAACAGACATACAAACTAAAGAACTACAAAAACAAATTACAAAAAT
1201     I  D  I  I  A  T  D  I  Q  T  K  E  L  Q  K  Q  I  T  K  I

3661    TCAAAATTTTCGGGTTTATTACAGGGACAGCAGAGATCCAGTTTGGAAAGGACCAGCAAA
1221     Q  N  F  R  V  Y  Y  R  D  S  R  D  P  V  W  K  G  P  A  K

3721    GCTTCTCTGGAAAGGTGAAGGGGCAGTAGTAATACAAGACAATAGTGAAATAAAGGTAGT
1241     L  L  W  K  G  E  G  A  V  V  I  Q  D  N  S  E  I  K  V  V

3781    ACCAAGAAGAAAAGCAAAGATCATTAGGGATTATGGAAAACAGATGGCAGGTGATGATTG
1261     P  R  R  K  A  K  I  I  R  D  Y  G  K  Q  M  A  G  D  D  C
1261                                   M  E  N  R  W  Q  V  M  I  V

3841    TGTGGCAGGTAGACAGGATGAGGATTAAAACATGGAACAGTTTAGTAAAGCATCATATGT
1281     V  A  G  R  Q  D  E  D
1281      W  Q  V  D  R  M  R  I  K  T  W  N  S  L  V  K  H  H  M  Y

3901    ATGTCTCAAAGAAAGCTAAAGATTGGTTCTATAGACATCATTATGAAAGCAGGCATCCAA
1301      V  S  K  K  A  K  D  W  F  Y  R  H  H  Y  E  S  R  H  P  K

3961    AAGTAAGTTCAGAAGTACACATCCCACTCGGAGAAGCTAGACTGGTAGTAAGAACATATT
1321      V  S  S  E  V  H  I  P  L  G  E  A  R  L  V  V  R  T  Y  W

4021    GGGGTCTGCATACAGGAGAGAGAGAATGGCATCTGGGTCAGGGAGTCTCCATAGAATGGA
1341      G  L  H  T  G  E  R  E  W  H  L  G  Q  G  V  S  I  E  W  R
```

```
4081    GGAAAAGGAGATATAGCACACAAATAGACCCTGGCCTGGCAGACCAACTAATTCATATAC
1361      K  R  R  Y  S  T  Q  I  D  P  G  L  A  D  Q  L  I  H  I  H

4141    ATTATTTTGATTGTTTTGCAGAATCTGCTATAAGAAAAGCCATATTAGGACATATAGTTA
1381      Y  F  D  C  F  A  E  S  A  I  R  K  A  I  L  G  H  I  V  T

4201    CTCCTAGGTGTAATTATCAAGCAGGACATAACAAGGTAGGATCTTTACAATATTTGGCAT
1401      P  R  C  N  Y  Q  A  G  H  N  K  V  G  S  L  Q  Y  L  A  L

4261    TAACAGCATTAATAGCACCAAAAAAGATAAAACCACCTCTGCCTAGCGTGAGGAAGCTGA
1421      T  A  L  I  A  P  K  K  I  K  P  P  L  P  S  V  R  K  L  T

4321    CAGAAGATAGATGGAACGAACCCCAGAGGACCAAGGACCACAGAGGGAGCCATGCAATGA
1441              M  E  R  T  P  E  D  Q  G  P  Q  R  E  P  C  N  E
1441      E  D  R  W  N  E  P  Q  R  T  K  D  H  R  G  S  H  A  M  N

4381    ATGGACATTAGAGCTTTTAGAGGAGCTCAAGAGTGAAGCTGTTAGACACTTTCCTAGGCC
1461     W  T  L  E  L  L  E  E  L  K  S  E  A  V  R  H  F  P  R  P
1461      G  H

4441    ATGGCTTCACAGCCTAGGACAATATATCTATGAAACTTATGGGGATACCTGGACAGGAGT
1481     W  L  H  S  L  G  Q  Y  I  Y  E  T  Y  G  D  T  W  T  G  V

4501    TGAAACTATAATAAGAATTCTTCAACAACTACTGTTTATCCATTTCAGAATTGGGTGTCA
1501     E  T  I  I  R  I  L  Q  Q  L  L  F  I  H  F  R  I  G  C  Q

4561    ACATAGCAGAATAGGCATTACTCGACAGAGGAGATCAAGAAATGGACCCAGTAGATTTTA
1521     H  S  R  I  G  I  T  R  Q  R  R  S  R  N  G  P  S  R  F
                                                M  D  P  V  D  F  N

4621    ACCTAGAGCCCTGGAACCATCCAGGAAGTCAGCCTAGGACTCCTTGTAACAGGTGTTATT
1541      L  E  P  W  N  H  P  G  S  Q  P  R  T  P  C  N  R  C  Y  C

4681    GTAAAAAGTGCTGCTATCATTGTCAAGTGTGCTTCGTAACGAAAGGCTTAGGCATCTCCT
1561      K  K  C  C  Y  H  C  Q  V  C  F  V  T  K  G  L  G  I  S  Y

4741    ATGGCAGGAAGAAGCGGAAACAGCGACGAAGACCTCCTGAAGGCGGTCAGGCTCATCAAG
1581     M  A  G  R  S  G  N  S  D  E  D  L  L  K  A  V  R  L  I  K
1581      G  R  K  K  R  K  Q  R  R  R  P  P  E  G  G  Q  A  H  Q  D

4801    ATCCTATACCAAAGCAGTAAGTAGTACATGTAATGTTACCTTTAGTGATATTAGCAATAG
1601     I  L  Y  Q  S
1601      P  I  P  K  Q                M  L  P  L  V  I  L  A  I  V
```

```
4861    TAGCCTTAGTGGTAGCACTAATACTAGCAATAGTTGTGTGGACTATAGTAGCTATAGAGT
1621      A  L  V  V  A  L  I  L  A  I  V  V  W  T  I  V  A  I  E  C

4921    GTATAAGATTAAGAAAGCAAAGAAAAATAGACAGGTTAATTGAAAGAATAAGGGAAAGAG
1641      I  R  L  R  K  Q  R  K  I  D  R  L  I  E  R  I  R  E  R  A

4981    CAGAAGACAGTGGCAATGAGAGTGATGGGGACACAGATGAATTGGCAAAACTGGTGGAGA
1661     Q  K  T  V  A  M  R  V  M  G  T  Q  M  N  W  Q  N  W  W  R
1661      E  D  S  G  N  E  S  D  G  D  T  D  E  L  A  K  L  V  E  M

5041    TGGGGAACCATGATTATGGGGATATTAATAATTTGTAGTACTGCAGAAGAAACGTGGGTT
1681     W  G  T  M  I  M  G  I  L  I  I  C  S  T  A  E  E  T  W  V
1681      G  N  H  D  Y  G  D  I  N  N  L

5101    ACTGTCTACTATGGGGTACCTGTGTGGAGAGACGCAGAGACCACCTTATTTTGTGCATCA
1701     T  V  Y  Y  G  V  P  V  W  R  D  A  E  T  T  L  F  C  A  S

5161    GATGCTAAGGCATATGAGACAGAAAAGCATAATGTCTGGGCTACACATGCCTGTGTACCC
1721     D  A  K  A  Y  E  T  E  K  H  N  V  W  A  T  H  A  C  V  P

5221    ACAGACCCCAGCCCACAAGAAATATATTTGGAAAATGTGACAGAACAGTTTAACATGTGG
1741     T  D  P  S  P  Q  E  I  Y  L  E  N  V  T  E  Q  F  N  M  W

5281    AAAAATAACATGGTAGAGCAAATGCATGCAGATATAATCAGTCTATGGGACCAAAGCTTA
1761     K  N  N  M  V  E  Q  M  H  A  D  I  I  S  L  W  D  Q  S  L

5341    AAGCCATGTGTACGGTTAACCCCTCTCTGTGTTACTTTAGAGTGTAGTGACGTCATTAAC
1781     K  P  C  V  R  L  T  P  L  C  V  T  L  E  C  S  D  V  I  N

5401    AAAACCAGAGGTACCATCAACCAAACCATAGAACAAAGAATGGAAGGAGAAATAAAAAAC
1801     K  T  R  G  T  I  N  Q  T  I  E  Q  R  M  E  G  E  I  K  N

5461    TGCTCTTACAATATGACCACAGAACTAAGGGATAAGAGACAAAAAGTACAGTCATTATTT
1821     C  S  Y  N  M  T  T  E  L  R  D  K  R  Q  K  V  Q  S  L  F

5521    TATAGACTTGATGTAGTAAAAATTAATAAAAATAGTAATAACACAAATACCAGTGAATAT
1841     Y  R  L  D  V  V  K  I  N  K  N  S  N  N  T  N  T  S  E  Y

5581    AGATTAATAAATTGTAATACCTCAGCCATTACACAAGCATGCCCAAAGGTAACCTTTGAG
1861     R  L  I  N  C  N  T  S  A  I  T  Q  A  C  P  K  V  T  F  E

5641    ACAATTCCCATACATTATTGTGCCCCAGCTGGTTTTGCGATTCTAAAATGTAATGACAAA
1881     T  I  P  I  H  Y  C  A  P  A  G  F  A  I  L  K  C  N  D  K
```

5701 GAGTTCAATGGAATAGGGGAATGCAAAAATGTCAGCACAGTCCAATGCACACATGGAATC
1901 E F N G I G E C K N V S T V Q C T H G I

5761 AGGCCAGTAGTAACAACTCAACTGCTGTTAAATGGCAGTCTACCAGACGGAAAGGTAATG
1921 R P V V T T Q L L L N G S L P D G K V M

5821 ATTAGATCTGAAAATATCACAAACAATGCCAAAAATATAATAGTACAATTTAACGAGACT
1941 I R S E N I T N N A K N I I V Q F N E T

5881 GCACAAATTAATTGTACCAGACCTAACAACAATACAAGAAAAAGTGTACGTATAGGACCA
1961 A Q I N C T R P N N N T R K S V R I G P

5941 GGACAAGCATACTATGCAGCAGGTGACATAATAGGGGATATAAGACAAGCATATTGTAAT
1981 G Q A Y Y A A G D I I G D I R Q A Y C N

6001 GTCAGTAAAAAACAATGGGATGGAATGTTGCAAAAAGTAGCCGACCAATTAAGAACACAT
2001 V S K K Q W D G M L Q K V A D Q L R T H

6061 TTTGGGGAAAACAAAACAATAATCTTTGCTAACTCCTCAGGAGGGGACGTACAAATTACA
2021 F G E N K T I I F A N S S G G D V Q I T

6121 ACACATAGTTTTAATTGTGGAGGAGAATTTTTCTATTGTGGTACATCAGACCTGTTTAAT
2041 T H S F N C G G E F F Y C G T S D L F N

6181 AGCATTTGGGATCTCAATAATGCCACAAATGGCTCAGAGTCAACTGACACTATAATAAAA
2061 S I W D L N N A T N G S E S T D T I I K

6241 CTCCCATGCAGAATAAAGCTAATCATAAATATGTGGCAGAGAACAGGACAAGCAATGTAT
2081 L P C R I K L I I N M W Q R T G Q A M Y

6301 CCCCCTCCCCTCCGAGGAGTAATAAGATGTGATTCAAACATTACAGGACTAATATTAACA
2101 P P P L R G V I R C D S N I T G L I L T

6361 AGAGATGGTGGGAATGGGAACAGTAGTACAAATGAAACCTTTAGACCTGGAGGAGGAAAT
2121 R D G G N G N S S T N E T F R P G G G N

6421 ATGAGGGACAATTGGAGAAGTGAATTATATAAGTATAAAGTAGTAAAAATTGAACCACTA
2141 M R D N W R S E L Y K Y K V V K I E P L

6481 GGAGTAGCACCCACCAGGGCAAAGAGAAGAGTGGTGGAGAGAGAAAAAAGAGCAGTTGGA
2161 G V A P T R A K R R V V E R E K R A V G

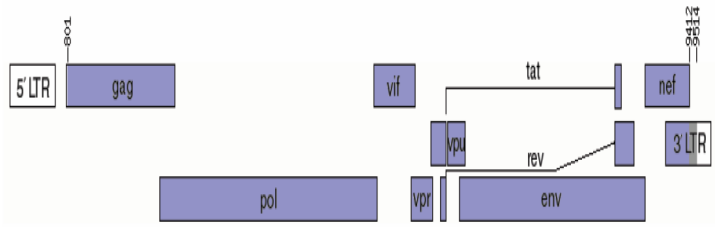
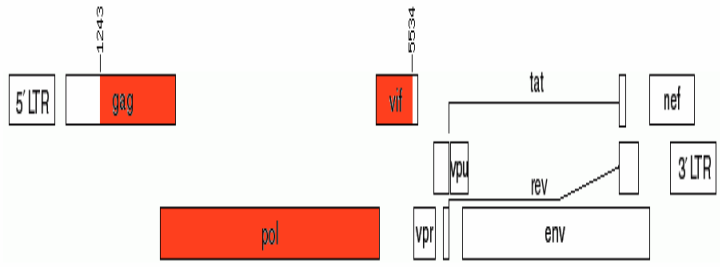
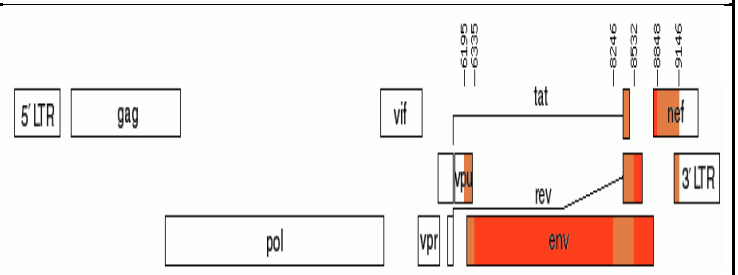
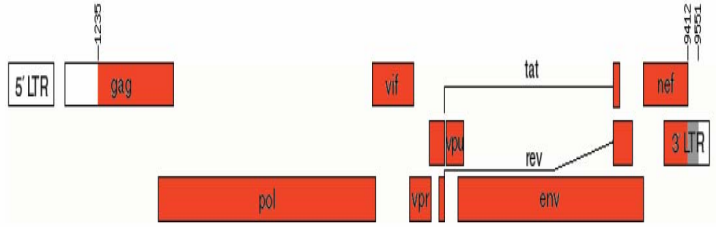
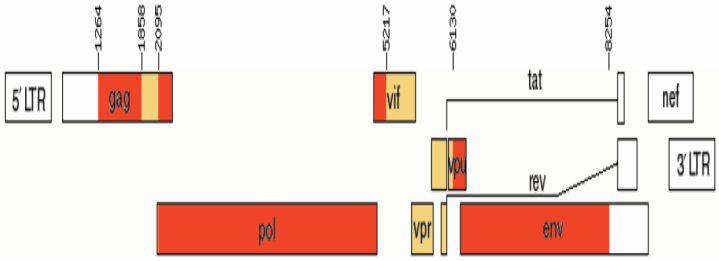
6541 ATAGGAGCTGTTTTCATTGGGTTCTTAGGAGCAGCAGGAAGCACTATGGGCGCGGCGTCA
2181 I G A V F I G F L G A A G S T M G A A S

```
6601    ATAACGCTGACGGTACAGGCCAGACAATTGTTGTCTGGCATAGTGCAGCAGCAAAGCAAT
2201     I  T  L  T  V  Q  A  R  Q  L  L  S  G  I  V  Q  Q  Q  S  N

6661    TTGCTGAGGGCTATAGAGGCTCAACAACATCTGTTGAAACTCACGGTCTGGGGCATTAAA
2221     L  L  R  A  I  E  A  Q  Q  H  L  L  K  L  T  V  W  G  I  K

6721    CAGCTCCAGGCAAGAGTCCTGGCTGTGGAAAGATACCTAAAGGATCAACAGCTCCTAGGA
2241     Q  L  Q  A  R  V  L  A  V  E  R  Y  L  K  D  Q  Q  L  L  G

6781    ATTTGGGGCTGCTCTGGAAAACTCATCTGCACCACTAATGTGCCCTGGAATTCTAGTTGG
2261     I  W  G  C  S  G  K  L  I  C  T  T  N  V  P  W  N  S  S  W

6841    AGTAATAAATCTCAGGATGAGATATGGAATAACATGACCTGGCTGCAGTGGGATAAAGAA
2281     S  N  K  S  Q  D  E  I  W  N  N  M  T  W  L  Q  W  D  K  E

6901    ATTAGCAATTACACAGAAACAATATATAGGCTAATTGAAGAATCGCAAAACCAGCAGGAA
2301     I  S  N  Y  T  E  T  I  Y  R  L  I  E  E  S  Q  N  Q  Q  E

6961    AAGAATGAACAAGACTTATTGGCATTGGACAAGTGGACAAATCTGTGGAATTGGTTTGAC
2321     K  N  E  Q  D  L  L  A  L  D  K  W  T  N  L  W  N  W  F  D

7021    ATATCGAACTGGCTGTGGTATATAAAAATATTTATAATGATAGTAGGAGGCTTAATAGGA
2341     I  S  N  W  L  W  Y  I  K  I  F  I  M  I  V  G  G  L  I  G

7081    TTAAGAATAGTTGCTGTGCTTTCTATAATAAATAGAGTTAGGCAGGGATACTCACCTTTG
2361     L  R  I  V  A  V  L  S  I  I  N  R  V  R  Q  G  Y  S  P  L

7141    TCATTTCAGACCCATACCCCAAACCCAGGGGGACTCGACAGGCCCGAAAGAACAGAAGAA
2381     S  F  Q  T  H  T  P  N  P  G  G  L  D  R  P  E  R  T  E  E
2381              P  I  P  Q  T  Q  G  D  S  T  G  P  K  E  Q  K  K
2381               P  Y  P  K  P  R  G  T  R  Q  A  R  K  N  R  R  R

7201    GAAGGTGGAGTGCAAGGCAGAGACAGATCGATTCGATTAGTCAGCGGATTCTTAGCTCTT
2401     E  G  G  V  Q  G  R  D  R  S  I  R  L  V  S  G  F  L  A  L
2401      K  V  E  C  K  A  E  T  D  R  F  D
2401        R  W  S  A  R  Q  R  Q  I  D  S  I  S  Q  R  I  L  S  S  C

7261    GCCTGGGACGATCTGAGGAGCCTGTGCCTTTTCAGCTGCCGCCGCTTGAGAGACTTCATG
2421     A  W  D  D  L  R  S  L  C  L  F  S  C  R  R  L  R  D  F  M
2421        L  G  R  S  E  E  P  V  P  F  Q  L  P  P  L  E  R  L  H  V

7321    TTGATTGCAGCGAGGACTGTGGAACTTCTGGGACACAGCAGTCTCAAGGGGCTGAGACTG
2441     L  I  A  A  R  T  V  E  L  L  G  H  S  S  L  K  G  L  R  L
2441        D  C  S  E  D  C  G  T  S  G  T  Q  Q  S  Q  G  A  E  T  G
```

```
7381      GGGTGGGAAGGAATCAAGTATCTGGGGAATCTCCTGTTGT
2461       G  W  E  G  I  K  Y  L  G  N  L  L  L
2461         V  G  R  N  Q  V  S  G  E  S  P  V
```

Appendix E

Table 6.8: jpHMM subtyping results of NFLG's fragments.

Sample	Subtype	Results
R84	B	
TV239 *gag-pol*	A1	
TV239 *env-nef*	A1 / C	
TV314	A1	
TV412	A1 / D	

Name	Length	Report	Assignment	Support	Genome
R84	8710bp	Report	HIV-1 Subtype B	100.0	
TV239_gag-pol	4287bp	Report	HIV-1 Subtype A (A1)	100.0	
TV239_env-nef	2966bp	Report	Check the bootscan	NA	
TV314	8344bp	Report	HIV-1 Subtype A (A1)	100.0	
TV412	7420bp	Report	Check the bootscan	NA	

Figure 6.4: REGA subtyping results of NFLG's fragments.