Molecular characterisation of HIV-1 recombinants and non-subtype C viruses in South Africa

by
Olivette Varathan

Thesis presented in fulfilment of the requirements for the degree of Master of Science in the Faculty of Medicine and Health Sciences at Stellenbosch University

Supervisor: Dr. Graeme Brendon Jacobs
Co-supervisor: Prof. Susan Engelbrecht

April 2019

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

----------	07/02/2019
Signature	Date

Acknowledgements

I would like to extend my most sincere gratitude to the following individuals and institutions, without whom I would not have been able to complete this degree:

Dr. Graeme Brendon Jacobs, my supervisor, for his continued support and encouragement throughout this project. Thank you for always being so approachable and understanding in everything. I am grateful to you for providing me with the opportunity of this degree.

Prof. Susan Engelbrecht, for her expert advice and impartation of knowledge. Thank you for your positive outlook towards every obstacle encountered in this project. Your patience in tutoring me is appreciated.

My office mates, Mr Emmanuel Obasa, Miss Janca Ferreira and Miss Michaela Venter, for their constant support, encouragement and comedic relief. Our time together will be treasured.

My fellow colleagues at the Division of Medical Virology for their moral support and friendship.

The National Research Foundation (NRF) and Poliomyelitis Research Foundation (PRF) for their financial assistance.

My parents, Emmanuel and Kookie Varathan, and siblings for their unwavering support, encouragement and belief in me.

I am most grateful to God for his favour and grace to complete this degree. Without Him, I am nothing.

"*Trust in the Lord with all your heart and lean not on your own understanding. In all your ways acknowledge Him and he will direct your path*." ~ Proverbs 3:5-6

Opsomming

MIV / VIGS is 'n ernstige gesondheidsbelasting, wat teen die einde van 2017 wêreldwyd 36,9 miljoen mense affekteer. Suid-Afrika het die grootste MIV-1-epidemie ter wêreld, met 7,2 miljoen mense geafekteer teen die einde van 2017. Die MIV-1-epidemie in Suid-Afrika word oorheers deur MIV-1 subtipe C, wat verantwoordelik is vir 'n geskatte 98,2% van die infeksies in die land, gebaseer op die virusse in die Los Alamos Nasionale Biblioteek (LANL) databasis. Tot op datum is 22 referate gepubliseer oor nie-subtype C virusse in Suid-Afrika tot die beste van ons wete. Hierdie studie het ten doel om twee naby vollengte-genoomvolgorde van nie-subtype C virusse in Suid-Afrika te karakteriseer.

Die studiemonsters is verkry deur die diagnostiese dienste van die Nasionale Gesondheids Laboratoriumdienste (NHLS), wat roetine dwelm weerstand toetsing in die Afdeling Mediese Virologie, Universiteit Stellenbosch, uitvoer. Alle monsters wat ontvang is, se genoom volgorde van die gedeeltlike pol streek (~ 1.4kb) was bepaal deur die NHLS. Die moontlike subtipe van die virus is geïdentifiseer deur die volgorde te analiseer met behulp van aanlyn subtipieërings programme. Alle vologordes en monsters wat as moontlike nie-subtype C virusse geïdentifiseer is, is in 'n aparte C-kohort van nie-subtype aangeteken. In 2011 is 'n moontlike C, D rekombinante virus vir die eerste keer geïdentifiseer in hierdie kohort. Teen die einde van 2015 is 30 soort gelyke rekombinante virusse waargeneem in die kohort, wat 'n moontlike opkoms van hierdie rekombinante stam aandui. Twee van die monsters wat as moontlike C, D-rekombinante geïdentifiseer is, is gekies vir naby-vol lengte-genoom (NFLG) karakterisering.

Provirale DNA, van monster EC148, is onttrek uit PBMCs en virale RNA. Plasma is onttrek van uit monster WC416. Die RNA was getru getransskribeer na DNA via cDNA sintese. Beide virus monsters het twee rondes PKR ondergaan. Die eerste ronde het die versterking van die MIV-1 NFLG (8978bp) geteiken en die tweede het die versterking van twee oorvleuelende fragmente (5455bp en 4909bp) geteiken. Positiewe PKR amplikone is gesuiwer en volgorde bepaal. Die gegenereerde volgordes is gelees en ontleed voordat dit gebruik was,deur gebruik te maak van aanlyn subtipeërings programme. Die jumping profile hidden markov model (jpHMM), REGA en rekombinasie-identifikasieprogramme (RIP) is gebruik om voorlopige subtipes vir beide monsters te bepaal. Verder is filogenetiese ontledings gebruik om die aanlyn subtipeërings program se resultate te bevestig of te verwerp.

Die aanlyn subtipeërings programme het monsters EC148 en WC416 as komplekse A, C, D rekombinante geïdentifiseer. Filogenetiese analise het die aanlyn subtipeërings program se resultate vir die volgorde van monster WC416 bevestig as komplekse A, C, D rekombinante. Filogenetiese analise het aangedui dat die volgorde van monster EC148 in ooreenstemming is met die resultate wat vanaf die aanlyn subtipeërings programme waargeneem word. Elke MIV-1-volgorde wat geïdentifiseer is as 'n unieke komplekse rekombinante vorm omskryf, omrede die breekpunte tussen die verskillende subtipes verskil het.

Die opkoms van nuwe en unieke nie-subtipe C rekombinante in Suid-Afrika dui aan dat die epidemie kompleks en ontwikkelend is. Dit is dus belangrik om die verspreiding van verskillende MIV-subtipes wat in Suid-Afrika versprei word, te monitor.

Abstract

HIV/AIDS is a severe health burden, affecting 36.9 million people worldwide by the end of 2017. South Africa has the largest HIV-1 epidemic in the world, estimated at 7.2 million infected individuals by the end of 2017. The HIV-1 epidemic in South Africa is dominated by HIV-1 subtype C, accounting for an estimated 98.2% of the infections in the country, based on the viral sequences in the Los Alamos National Library (LANL) database. To date, to the best of our knowledge, 22 papers have been published on non-subtype C viruses in South Africa. This study aimed to characterise two near full-length genome sequences of non-subtype C viruses in South Africa.

The study samples were obtained through the diagnostic services of the National Health Laboratory Services (NHLS), performing routine drug resistance testing within the Division of Medical Virology, Stellenbosch University. All samples received were sequenced in the partial *pol* region (~1.4kb) of the viral genome by the NHLS. The possible subtype of the virus was identified from the sequences using online subtyping programmes. All sequences and samples that identified as possible non-subtype C viruses were recorded in a separate non-subtype C cohort. In 2011, a possible C, D recombinant virus was first identified in this cohort. By the end of 2015, 30 similar recombinant viruses were observed in the cohort, indicating a possible emergence of this recombinant strain. Two of the samples that identified as possible C, D recombinants were selected for further near full-length genome (NFLG) characterisation.

Proviral DNA, from sample EC148, was extracted from PBMCs and viral RNA, from sample WC416, was extracted from plasma. The RNA was reverse transcribed to DNA via cDNA synthesis. Both sample viruses were amplified by PCR in two rounds. The first round targeted the amplification of the HIV-1 NFLG (8978bp) and the second targeted the amplification of two overlapping fragments (5455bp and 4909bp). Positive PCR amplicons were purified and sequenced. The generated sequences were read and analysed before being used in online subtyping programmes. The jumping profile hidden markov model (jpHMM), REGA and recombination identification programmes (RIP) were used to preliminary assign subtypes to both samples. Phylogenetic analyses was inferred to confirm / reject the online subtyping programme results.

Online subtyping programmes identified the virus sequences of samples EC148 and WC416 as complex A, C, D recombinants. Phylogenetic analysis confirmed the online subtyping programme results for the sequence of sample WC416 in identifying it as a complex A, C, D recombinant. Phylogenetic analysis indicated that the sequence of sample EC148 is consistent with the results observed from the online subtyping programmes. Each HIV-1 sequence identified as a unique complex recombinant form as the breakpoints between the different subtypes differed.

The emergence of new and unique non-subtype C recombinants in South Africa indicates that the epidemic is complex and evolving. It is therefore important to monitor the spread of different HIV subtypes circulating in South Africa.

Table of Contents

List of Figures

List of Tables

List of abbreviations

°C	Degree Celcius
µl	Microlitre
AIDS	Acquired Immunodeficiency Syndrome
ART	Antiretroviral Therapy
BIC	Bayesian Information Criterion
CAF	Central Analytical Facility
CDC	Centre for Disease Control
cDNA	Complimentary Deoxyribose Nucleic Acid
CRF	Circulating Recombinant Form
DNA	Deoxyribose Nucleic Acid
DRC	Democratic Republic of Congo
EDTA	Ethylenediaminetetra acetic acid
FASTA	Fast-All
HAART	Highly Active Antiretroviral Therapy
HIV	Human Immunodeficiency Virus
HIV-1	Human Immunodeficiency Virus Type One
HKY	Hasegawa-Kishino-Yano
HREC	Health Research Ethics Committee
HTVL	Human T-cell Leukaemia Associated Virus
JC	Jukes Cantor
jpHMM	Jumping Profile Hidden Markov Model
KS	Kaposi's Sarcoma
Kb	Kilobases
LANL	Los Alamos National Laboratory
LAV	Lymphadenopathy Associated Virus
LB	Lysogeny Broth
MAFFT	Multiple Alignment using Fast Fourier Transform
MEGA	Molecular Evolutionary Genetics Analysis
ML	Maximum Likelihood
MMWR	Morbidity and Mortality Weekly Report
MSA	Multiple Sequence Alignment
NEB	New England BioLabs

NFLG	Near Full-Length Genome
NGS	Next Generation Sequencing
NHLS	National Health Laboratory Services
NJ	Neighbour Joining
PAUP	Phylogenetic Analysis Using Parsimony
PBMC	Peripheral Blood Mononuclear Cells
PCP	*Pneumocystis carinii* pneumonia
PCR	Polymerase Chain Reaction
RAM	Resistance associated mutation
REV	Reversible
RIP	Recombination Identification Programme
RNA	Ribonucleic Acid
RT	Reverse Transcriptase
SA	South Africa
SIV	Simian Immunodeficiency Virus
SSIV	SuperScript IV
TAE	Tris base, acetic acid and EDTA
UPGMA	Unweighted Pair Group Method with Arithmetic Mean
URF	Unique recombinant form
USA	United States of America
USSR	Union of Soviet Socialist Republics
UV	Ultra Violet

Chapter 1

Introduction

1.1. Introduction

Human immunodeficiency virus (HIV) is the causative agent of acquired immune deficiency syndrome (AIDS) (Barré-Sinoussi *et al.*, 1983). HIV has infected over 76 million individuals since the start of the epidemic. It has become a major health burden since its discovery, affecting 36.7 million people globally by the end of 2017. In 2016 alone, 1 million individuals died from AIDS related illnesses. Eastern and Southern Africa are the most affected geographical region in the world, having 19.6 million individuals living with HIV and was responsible for 380 000 AIDS-related deaths by the end of 2017. South Africa is the country with the largest population of HIV-infected individuals in the world, which was 7.2 million by the end of 2017. In that same year there were110 000 AIDS-related deaths in South Africa.

HIV is a retrovirus with a diploid RNA genome. A unique feature of retroviruses is that they possess reverse transcriptase (RT) that allows the virus to enter a host cell, reverse transcribe the RNA genome to DNA and then integrate it into the host cell genome. Upon entry of a HIV particle into a host cell, the virus is able to use the host cells' machinery to replicate and produce progeny viruses (Krieg and Steinberg, 1990).

It is believed that HIV was transmitted to humans via a zoonotic transmission event from infected African primates and has spread rapidly through the human population since then (Essex and Kanki, 1988). Phylogenetic analyses have revealed that HIV exists in two forms, HIV type 1 (HIV-1) and HIV type 2 (HIV-2). HIV-1 has spread across the world; however, HIV-2 is endemic to West Africa. HIV-1 is divided into groups M (major), N (Non-M), O (Non-N) and P with group M being responsible for the pandemic. Group M is further divided into subtypes and recombinant forms. Recombinant forms can either be a unique recombinant form (URF) or a circulating recombinant form (CRF). A CRF is found in three individuals who are not epidemiologically related whereas a URF is found in only individual and has the potential to become a CRF. (Hahn *et al.*, 2000).

HIV-1 (group M) subtypes A, B and C are responsible for the majority of infections worldwide. The different subtypes are not evenly distributed and are predominant in different regions. Subtype B is predominant in Europe and the United States of America (USA), subtype A is predominant in Asia and subtype C is largely predominant in sub-Saharan Africa (van Harmelen *et al.*, 1999). Looking at South Africa in particular, the epidemic is almost completely dominated by subtype C viruses (Hemelaar, 2012). However, the emergence of non-subtype C viruses have been increasing over recent years. Recombinant viruses are becoming more prevalent across the world. This is seen in South East Asia where the epidemic is led by the recombinant CRF01_AE (Lau *et al.*, 2013). By October 2018, 97 recombinant forms have been identified and documented in the HIV Los Alamos National Library (LANL) database (https://www.hiv.lanl.gov/content/sequence/HIV/CRFs/CRFs.html; Accessed 31/10/18).

Recombinant viruses have been shown to have an impact on viral diagnostics, vaccine development, antiretroviral drug resistance as well as disease progression. Therefore, it is vital that the emergence of these viruses are monitored (Hemelaar, 2013).

1.2. Aims

The aim of the project is to characterize two possible new HIV-1 C, D recombinant viruses in South Africa through recombinant and phylogenetic analyses.

1.3. Objectives

1. DNA/RNA extraction, PCR and conventional Sanger sequencing of HIV-1 NFLG
2. HIV-1 recombinant analyses
3. Phylogenetic analyses

1.4. Rationale

South Africa has the highest incidence of HIV-infected individuals in the world. The epidemic is dominated by HIV-1 subtype C; however, there is an increasing emergence of non-subtype C viruses in the country. This study focuses on non-subtype C viruses in South Africa, particularly C, D recombinants. HIV-1 subtype C has been extensively studied in the South African context with only a limited number of studies being conducted on non-subtype C viruses. Studying the non-subtype C viruses in the country will increase our knowledge and understanding of the entire epidemic in South Africa. It is important that subtype diversity be monitored as studies have shown different subtypes to have faster progression to AIDS and to be more resistant to antiretroviral treatment than other subtypes. Recombinant viruses, being composed of 2 or more different viruses, has the potential to be more resistant to ARV treatment than pure subtypes. Constant monitoring of all circulating viruses will give a more accurate representation of the evolution of HIV-1 in South Africa. In 2011, four possible C, D recombinants were noticed in a non-subtype C cohort obtained from the NHLS. By the end of 2015, this number increased to 30. The sudden emergence and increase of this particular recombinant required that it be fully characterised.

Chapter 2

Literature Review

2.1. History

2.1.1. Discovery

The first official report of acquired immunodeficiency syndrome (AIDS) was issued by the United States of America's (USA) Centers for Disease Control and Prevention (CDC) in 1981. A Morbidity and Mortality Weekly Report (MMWR) was released stating that previously healthy homosexual men presented with a lung infection known as *Pneumocystis carinii pneumonia* (PCP) and some presented with a rare cancer called Kaposi's Sarcoma (KS) (CDC, 1981). A year later the CDC first used the term AIDS. The CDC defined a case of AIDS as: "a disease at least moderately predictive of a defect in cell mediated immunity, occurring in a person with no known cause for diminished resistance for that disease" (CDC, 1982).

Soon after the identification of AIDS and its associated characteristics (CDC, 1982), Dr. Robert Gallo and his team identified a virus that shared characteristics with the causative agent of AIDS in that it was transferred between individuals via mother-to-child transmission, sexual contact and through blood. The virus also resulted in lower CD4+ T-cells in infected individuals (Barré-Sinoussi *et al*., 1983; Gallo *et al.,* 1984). The identified virus was named human T-cell leukaemia associated virus-3 (HTVL-III) (Gallo *et al*., 1984). At the same time, a research group at the Pasteur Institute in France, who focused on retrovirus research set out to discover if a retrovirus was the causative agent for AIDS. Their research led to the discovery of the lymphadenopathy associated virus (LAV), which also displayed characteristics similar to those seen in AIDS patients (Ellrodt *et al.,* 1984; Barré-Sinoussi *et al.*, 1983). Upon further investigation of HTVL-III and LAV, it was discovered that they were the same virus and that it was the causative agent for AIDS. The virus was later named human immunodeficiency virus (HIV) (Coffin *et al.*, 1986).

2.1.2. Origin

In 1984 the primate Lentivirus, Simian Immunodeficiency Virus (SIV), was isolated from primates displaying an AIDS-like disease (Marx *et al.*, 1985; Chakrabarti *et al.*, 1987). Characterization of SIV revealed that this virus was very similar to HIV in that they both target the same CD4 subset of lymphocytes. SIV and HIV were also shown to have the same internal core proteins and their structural and regulatory genes were organized in an almost identical manner (Essex and Kanki, 1988).

In 1985, a virus related to SIV was found in blood samples of prostitutes from Senegal (Clavel *et al.*, 1986). The blood samples contained antigens for both HIV and SIV. The antigens present in the blood samples had higher reactivity with SIV than it did to HIV. The reactivity of the antigens to SIV in these blood samples resembled the antigen to SIV reactivity seen in SIV-infected Asian macaques.

This suggested that this virus was different to the HIV infecting people in the USA and Europe. Upon further characterisation of the virus, it was identified as a close relative to HIV-1 and was named HIV-2 (Essex and Kanki, 1988; Etienne *et al*., 2011). The transmission of SIV from Sooty mangabeys (SIVsm) to humans in West Africa was discovered, through phylogenetic analysis, as the origin of HIV-2 (Gao *et al.*, 1999; Etienne *et al*., 2011).

Most primates harbour a species-specific strain of SIV (example: Sooty mangabeys harbour SIVsm) suggesting that transmission of that specific SIV strain is transferred between members of the same species (Etienne *et al*., 2011). HIV-1 viral sequences were most closely related to *Pan troglodytes.troglodytes*, a sub-species of the *P.troglodytes* lineage and was discovered to be the origin of HIV-1 infections in humans (Gao *et al.*, 1999; Sharp and Hahn, 2011). HIV is therefore most likely a result of a transmission event between humans and African primates. The cross-species transmission between the primates and humans was presumably due to direct exposure to the blood of the primates, which could have been as a result of eating the meat raw, hunting or butchering of the animals (Buonaguro *et al.,* 2007). Molecular clock analyses dates the origin of HIV-1 group M to 1931 (Korber *et al.*, 2000). Figure 2.1 shows the origin of HIV transmission events from African primates to humans.

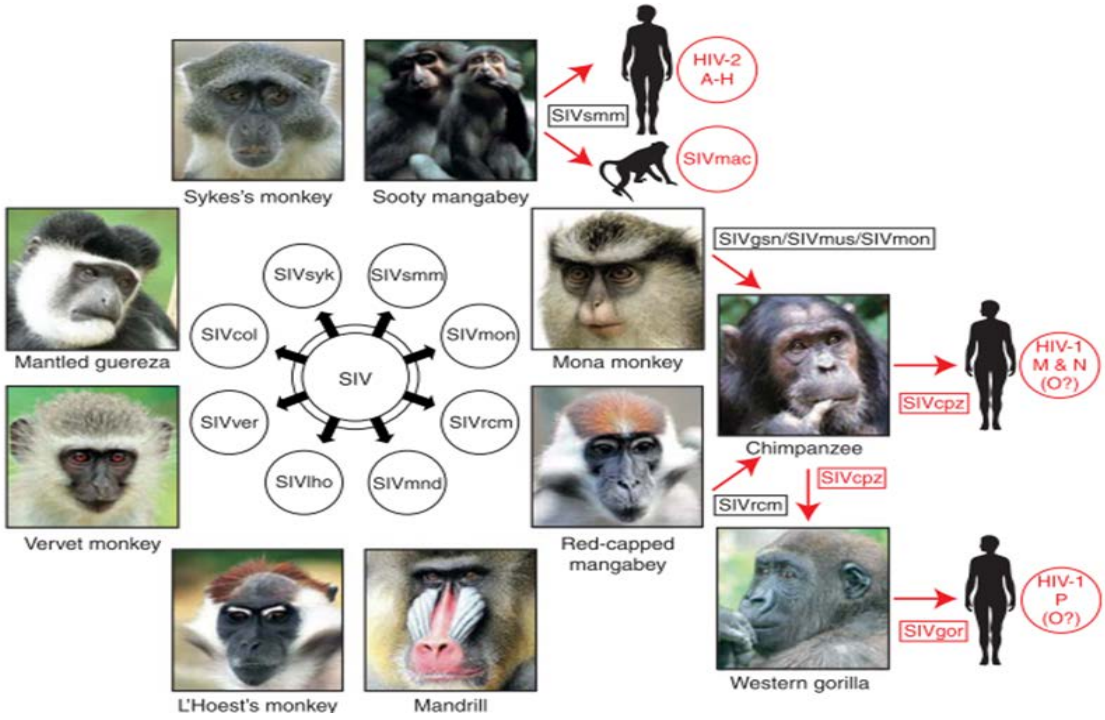

Figure 2.1: Transmission of SIV to humans (Gallo *et al*., 1999)
The transmission of SIVsmm from Sooty mangabeys gave rise to HIV-2 in humans. HIV-1 originated from a cross species transmission event of SIVcpz from the chimpanzee subspecies, *Pan troglodytes troglodytes*. SIV is able to cross the species barrier and transmit to both humans and non-human primates.

2.2. Basic virology of HIV

2.2.1. Classification

Retroviruses can be classified as either having simple genomes (the alpha, beta, gamma and epsilson retroviruses) or complex genomes (the lentiviruses, deltaviruses and spumaviruses) (Weiss, 2006). All retroviruses are composed of the group antigen (*gag)*, polymerase (*pol)* and envelope (*env)* genes as well as long terminal repeats (LTR), which contain promoter and enhancer regions (Krieg and Steinberg, 1990). Figure 2.2 depicts the phylogenetic relatedness of retroviruses.

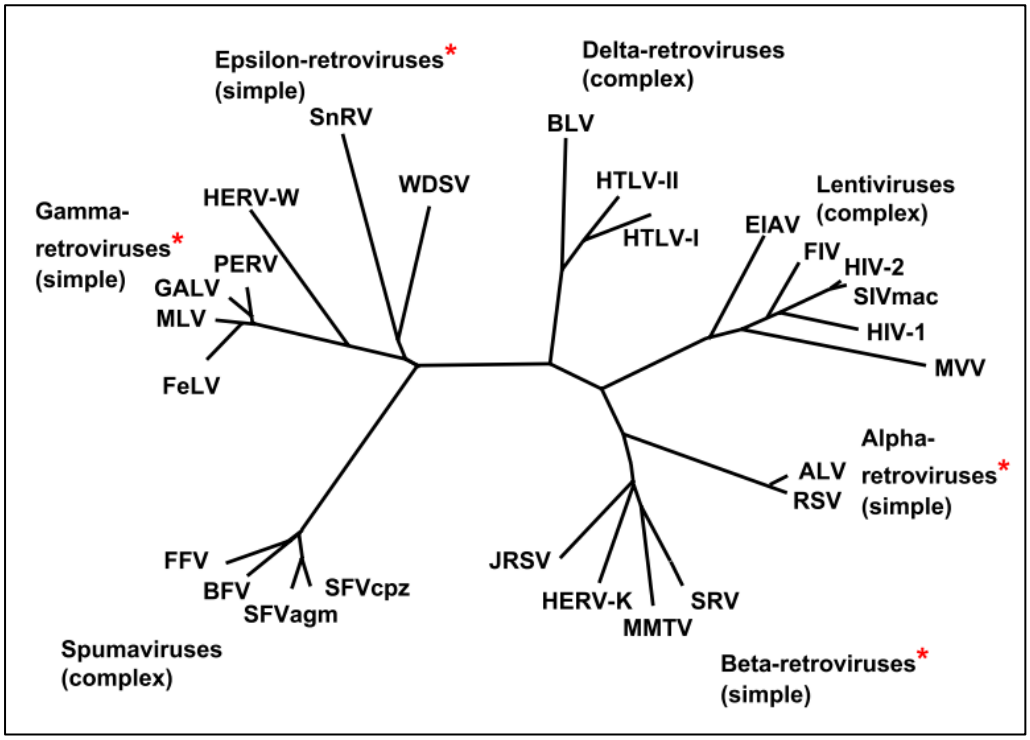

Figure 2.2: Phylogeny of retroviruses (Weiss, 2006)
The phylogenetic tree illustrates the various classifications of retroviruses. HIV-1 and HIV-2 are both classified as a lentivirus, which is a complex retrovirus.

In 1910, Peyton Rous from the Rockefeller Institute for Medicine in New York, identified a sarcoma of the common fowl, which proved to be transplantable to other fowls. The virus responsible for the sarcoma was named the Rous Sarcoma Virus and was the first documented discovery of a retrovirus (Rous, 1910). Since then retroviruses have been identified in tumours of many animals including mice, chicken and cats (Huebner and Todaro, 1969).

The central dogma of molecular biology states that the natural flow of genetic material is from DNA to RNA to protein (Gallo, 2005). However, in retroviruses, the genetic material is transferred from RNA to DNA to RNA and then to protein. The RNA of an infecting virus is first converted to proviral DNA that integrates into the host cell's genome and is then used as a template for the generation of progeny RNA, which is homologous in sequence to the RNA of the infecting virus (Gallo, 1986).

If the proviral DNA integrates into a host germ cell, the virus will be passed down onto the host's progeny as an endogenous retrovirus (Hayward, 2017). Howard Temin from the University of Wisconsin and David Baltimore at Massachusetts Institute of Technology, both independently discovered reverse transcriptase, the enzyme unique to retroviruses that is responsible for the conversion of RNA to proviral DNA (Baltimore, 1970; Temin, 1976). The discovery of reverse transcriptase allowed scientists to fully understand how retroviruses cause disease. The proviral state of the retrovirus allows for it to remain dormant within the host's cells genome and to reactivate and replicate at a later stage (Weiss, 2006). The first isolated human retrovirus was HTLV-1 and was isolated in 1979 from a patient with a T-cell malignancy (Poiesz *et al*., 1980; Gallo, 2005).

2.2.2. Virion and genome structure

The HIV genome is encapsulated by a lipid bilayer derived from the host cell's membrane, known as the viral envelope. Surface glycoproteins are anchored in the envelope and protrude outwards to facilitate the attachment of the virion to a host cells' receptor molecules (Briggs *et al.*, 2003). Within the viral envelope is a core that is composed of a number of proteins and a capsid that contains two molecules of RNA. It is within the core that the necessary genetic information for viral replication is stored. The reverse transcriptase is bound to the RNA molecules and is responsible for the transcription of a DNA molecule by using the RNA as a template (Turner and Summers,

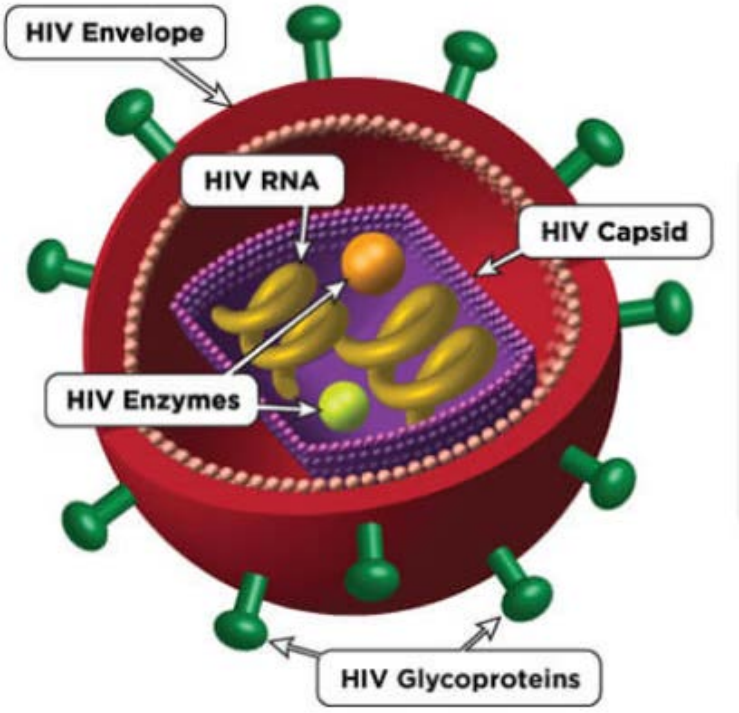

Figure 2.3: Virion structure (https://aidsinfo.nih.gov/understanding-hiv-aids/fact-sheets/19/73/the-hiv-life-cycle; Accessed 22/10/18)
The basic structure of a retrovirus consists of two RNA molecules and enzymes such as reverse transcriptase, enclosed in a viral capsid. The capsid is surrounded by a lipid envelope membrane, which anchors proteins (glycoproteins) that protrude from its surface and facilitate attachment of the virus to host cells.

1999). Figure 2.3 illustrates the components and structure of the HIV virion.

The HIV genome is complex and comprises of structural, regulatory and accessory genes. The following genes make up the HIV genome: the *gag*, p*ol*, *env,* transactivator *(tat*), regulatory factor (*rev*), viral infectivity factor *(vif*), viral protein R (*vpr),* viral protein U (*vpu*) and the negative regulatory factor (*nef*). The *gag*, *pol* and *env* are structural genes; *tat* and *rev* are regulatory genes while *vif, vpr,vpu* and *nef* are accessory genes (Costin, 2007).

The structural proteins are essential components of the viral particle. Figure 2.4 shows the arrangement of genes within the genome. The *gag* gene encodes the capsid proteins that houses the RNA molecules; the *pol* gene produces protease, reverse transcriptase, RNase and integrase; the *env* gene encodes the envelope glycoproteins gp120 and gp41. The *tat* and *rev* regulatory genes code for RNA binding proteins that are responsible for HIV gene expression. The accessory genes code for proteins, which have multiple functions. Vif is the viral infectivity factor that promotes the infectivity of the virus whereas vpu functions to reduce CD4 expression. The nef protein is mostly cytoplasmic and is associated with the plasma membrane (Costin, 2007; http://www.hiv.lanl.gov/; accessed: 30 August 2018).

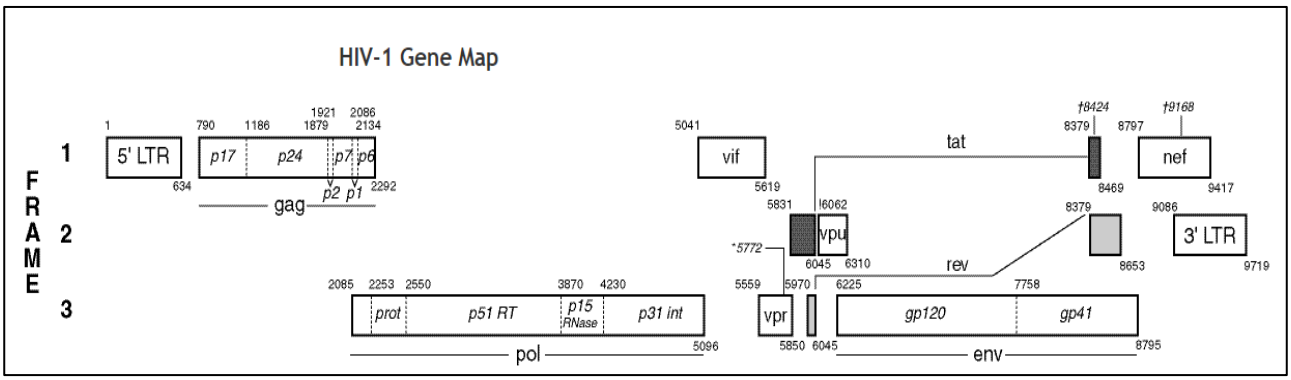

Figure 2.4: HIV-1 gene map (www.hivlanl.gov/ [Accessed on 10 May 2018])
The figure illustrates the positions of the different genes that make up the HIV genome. The numbers at the top denote the start position of the genes and the numbers at the bottom indicate their end positions. The genome begins at the 5'LTR and ends at the 3'LTR.

2.2.3. HIV Life cycle

The HIV life cycle (Figure 2.5) is divided into in eight steps: attachment, fusion, reverse transcription, integration, replication, assembly, budding and maturation. The cycle is separated into an early and late phase. The early phase begins with the attachment of the virus to a host cell and includes all processes up to the integration of DNA into the host genome. The late phase includes all stages after integration until maturation.

A free virus attaches to a host cell by binding of the viral envelope glycoprotein to the host cell receptor. HIV infects vital cells of the immune system such as helper T-cells. HIV-1 requires the presence of a co-receptor, known as chemokine co-receptors, to promote fusion of the virus into the host cell. HIV most commonly binds specifically to the CD4+ cell receptor and the CCR5 or CXCR4 chemokine co-receptor on the T-cell to promote fusion (Lawn, 2004). Fusion involves the transfer of the viral capsid through the host cell membrane and into its cytoplasm. Once in the cytoplasm, a process of uncoating takes place in which the contents of the capsid is released. Reverse transcriptase converts viral RNA to complimentary DNA (cDNA) via reverse transcription. The DNA and the integrase enzyme move from the cytoplasm into the nucleus where the enzyme's activity enables the integration of the DNA into the host cells' genome, as a provirus. At this stage, the early phase is complete (Turner and Summers, 1999). A unique feature of retroviruses is the ability of the provirus to remain dormant in the host cell genome, allowing it to evade antiretroviral therapy (ART), which could stop the virus from replicating (Finzi *et al.*, 1997).

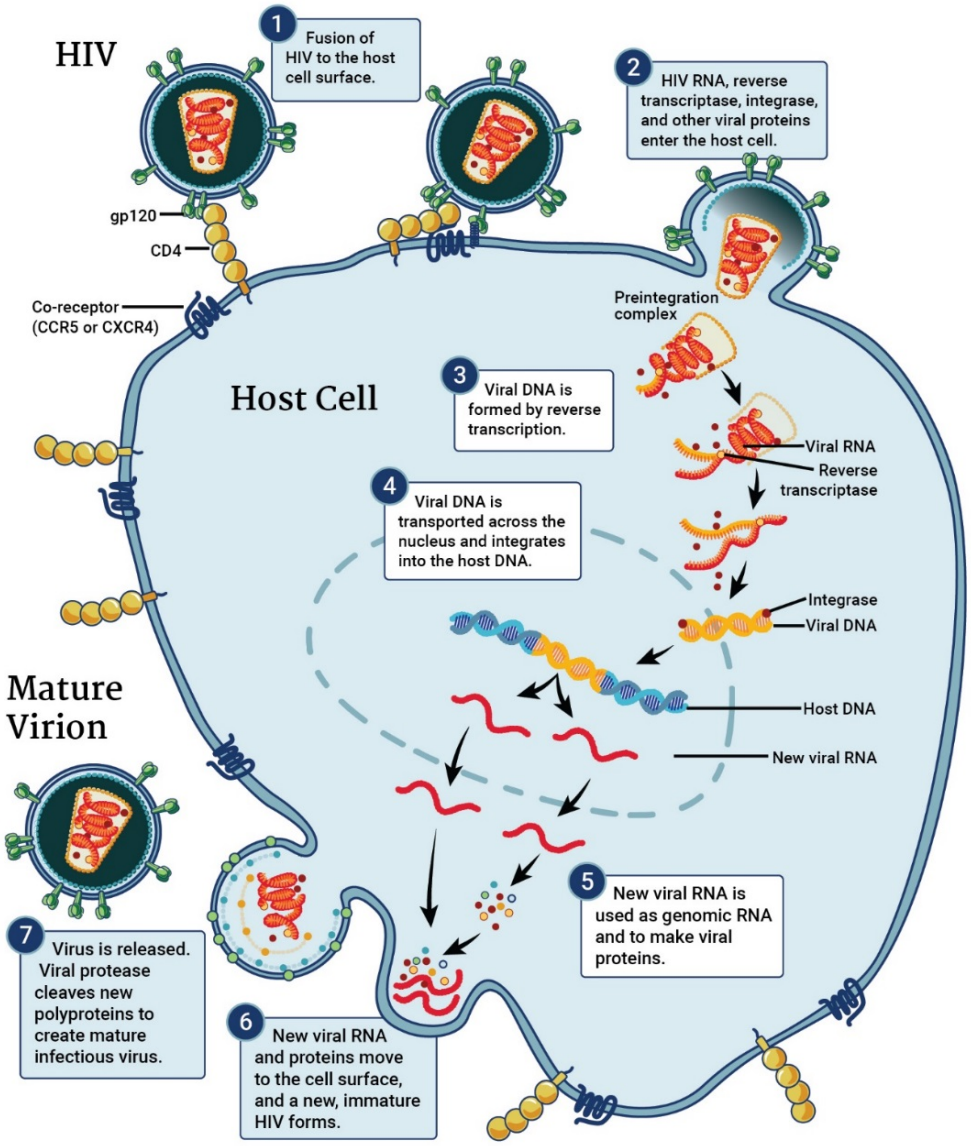

Figure 2.5: HIV life cycle, NIH, 2018 (accessed 22 June 2018)
The figure illustrates the different stages of the HIV life cycle. A free virus attaches itself to a host cell and ejects its contents into the host cell. Viral RNA is converted to DNA which is then integrated into the host cell's DNA. The integrated DNA is used to make copies of the viral RNA and proteins. The newly synthesised components are assembled and released from the cell. Once the new virus has matured, it is able to infect neighbouring host cells.

The late phase begins with the replication step in which the proviral DNA uses the host cells' machinery to transcribe messenger RNA (mRNA). The mRNA is exported from the nucleus to the cytoplasm where it is translated into regulatory and structural proteins. The RNA and proteins are assembled and packaged into a new immature virus particle at the surface of the cell. The virus exits the host cell by pushing itself through the cell membrane in a process called budding. This process allows the virus to use the host's cell membrane as its viral envelope membrane. Upon release, viral protease cleaves the proteins to create a mature infectious virus. The new mature virus is able to infect neighbouring cells and continue the cycle. The HIV life cycle is complex and

involves a number of intricate processes to ensure the success of the virus in the host organism (Sundquist and Krusslich, 2012).

2.3. HIV diversity

2.3.1. HIV-1 and HIV-2

A key characteristic of HIV-1 is the extreme genetic diversity of its viral genome, which is associated with the persistent nature of the virus, drug resistance to ART, as well as the difficulty in designing a successful vaccine (Smyth *et al*., 2012). The two major mechanisms responsible for HIV diversity is mutation and retroviral recombination. The mutations most often arise as a result of an error by the RNA polymerase, reverse transcriptase, as it lacks proof reading activity and therefore has a high error rate (Lau *et al.*, 2013). Retroviral recombination occurs when regions of different viruses recombine to form a new virus. These regions usually contain properties that are beneficial to the virus. This process allows the virus to acquire advantageous mutations and remove mutations that could reduce viral fitness (Korber *et al.*, 2001; Smyth *et al,* 2012).

HIV is divided into two forms, HIV-1 and HIV-2, which can be identified based on the organisation of their genome and their phylogenetic relationship with other primate lentiviruses (Hahn *et al.*, 2000). The two types of HIV can be divided into HIV-1 groups M (major), N (Non-M), O (Non-N) and P and HIV-2 groups A-H, with each group most likely originating as a result of an independent zoonotic transmission from African primates to humans (Duri *et al,* 2013). HIV-1 group M and N originate from SIVcpz found in the chimpanzee subspecies, *Pan troglodytes troglodytes,* endemic to West-central Africa*.* Whereas HIV-1 groups O and P originate from SIVgor found naturally in gorillas located in Cameroon (Hemelaar, 2012). HIV-1 group M is responsible for the HIV pandemic and can be further divided into subtypes, sub-subtypes and recombinant forms. Nine subtypes have been identified and are denoted by the letters A-D, F-H, J and K (Hemelaar, 2013). The intra-subtype genetic variation ranges between 15-20% and the inter-subtype genetic variation ranges between 25-35% (Korber *et al.*, 2001). The co-circulation of the different subtypes have resulted in the emergence of recombinant forms. A recombinants form is a virus composed of segments from two or more different subtypes. Recombinants are classified as either a circulating recombinant form (CRF) or a unique recombinant form (URF). A URF is classified as a recombinant that has been identified in less than three epidemiologically unrelated individuals (Perrin *et al*, 2003).The spread of a URF gives rise to a new CRF (Hemelaar, 2013). A recombinant virus is classified as a CRF if it is fully sequenced and identified in at least three epidemiologically unrelated individuals. Currently there are 97 CRFs globally (https://www.hiv.lanl.gov/content/sequence/HIV/CRFs/CRFs.html; [Accessed October 2018]).

2.3.2. Global diversity

HIV-1 group M is responsible for the HIV/AIDS pandemic. The different subtypes from this group are spread across the world, with subtypes A, B and C being most prevalent as seen in Figure 2.6 (Hemelaar *et al.*, 2011; Hemelaar, 2012). Subtype C accounts for almost 50% of the global infections, making it the dominant subtype of the pandemic.

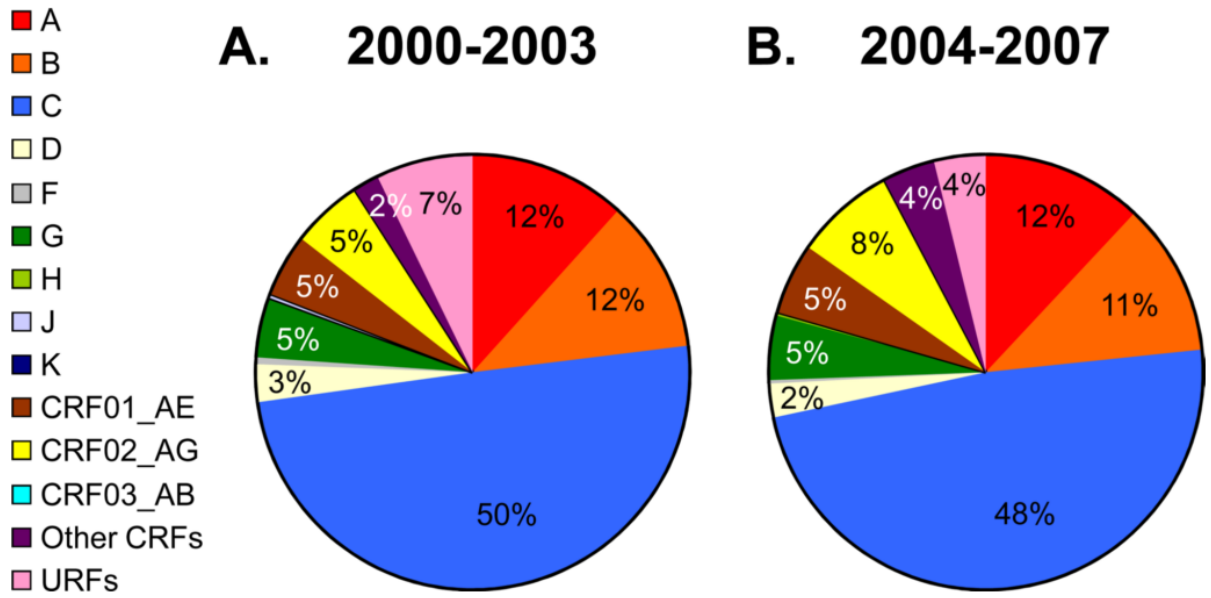

Figure 2.6: Global distribution of HIV- 1 pure subtypes and recombinants (Hemelaar *et al.,* 2012)
There is a slight variation in the results observed in the two time points. In 2004-2007, subtype C was the most prevalent global subtype followed by subtype B and A. CRF02_AG is the most prevalent of the CRFs followed by CRF01_AE.

The unequal global distribution of HIV-1 group M variants is a result of epidemiologic transmissions; the dynamic nature of the HIV pandemic and global variation in the HIV-strains (Lau *et al.*, 2013). Figure 2.7 shows the distribution of the subtypes across the world between 2004-2007. This is the latest representation of HIV-1 global diversity found in the literature. Subtype A is

dominant in East and Central Africa as well as in Eastern Europe. Subtype B is the main variant in Western and central Europe, North and South America, Australia, countries in central Asia, northern Africa and the Middle East. Subtype C is responsible for a large majority of the epidemics in South Africa, Ethiopia and India. The epidemics in these regions are almost exclusively subtype

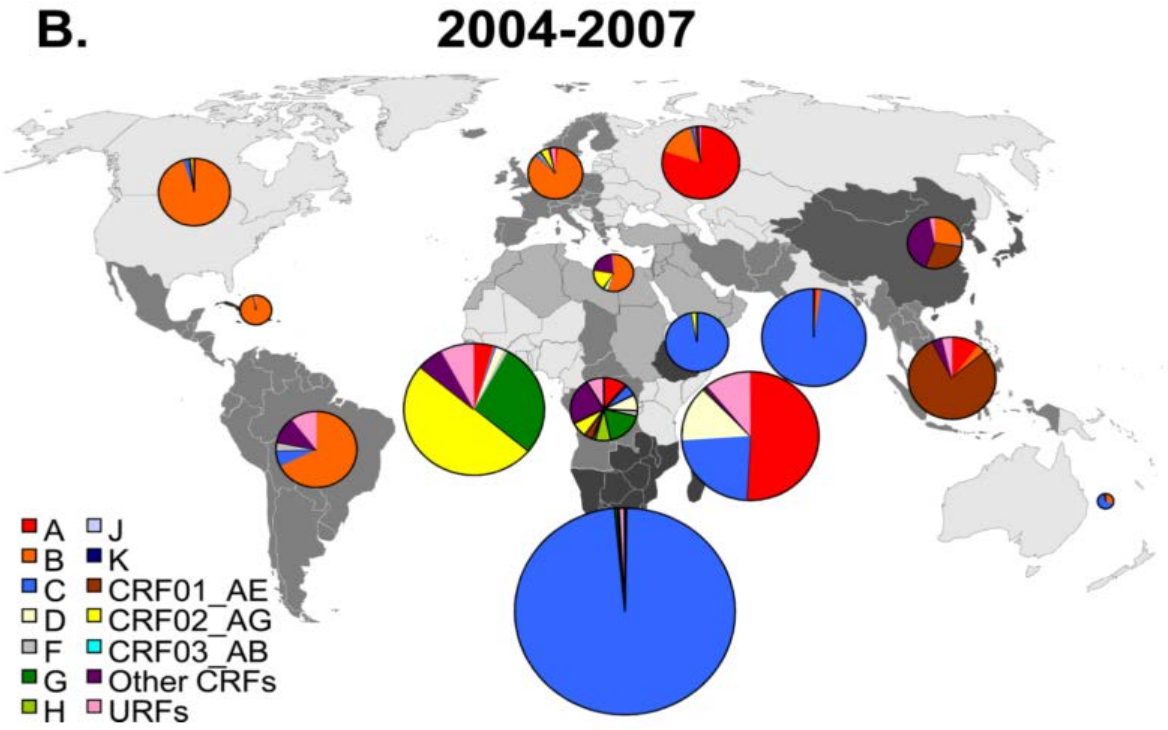

Figure 2.7: Regional distribution of HIV-1 subtypes and recombinants in 2004-2007 (Hemelaar et al., 2012) The map shows subtype B being dominant in the Americas, Europe and Northern Africa; Subtype A dominant in eastern Europe and East Africa whereas subtype C is dominant in South Africa, India and China.

C, with only a small proportion of non-subtype C infections. All subtypes are present in West Africa, with subtype G and CRF02_AG being dominant. In South and East Asia, the epidemic is dominated by CRF01_AE. Subtype D is mostly found in East Africa. The subtype B epidemic is widely and evenly spread across the world. The greatest diversity is observed in central Africa where all subtypes and numerous CRFs are present (Hemelaar *et al.*, 2011).

2.3.3. HIV-1 diversity in South Africa

The initial HIV-1 epidemic in South Africa, in the early 1980s, was led by subtype B and D viruses and was associated with homo- and bisexual transmission (Engelbrecht *et al*., 1995; Becker *et al.,* 1995). Soon after (late 80's, early 1990s), a subtype C epidemic was recognized among heterosexual individuals. Heterosexual and mother-to-child transmission are the main routes of HIV transmission in South Africa (van Harmelen *et al.*, 2001; Jacobs *et al.*, 2009; Wilkinson and Engelbrecht, 2009). It is believed that the HIV-1 subtype C virus originates from HIV-1 group M in central Africa. The ancestral subtype C is believed to have migrated from Kinshasa in the

Democratic Republic of Congo (DRC). Phylodynamics suggests that this migration occurred around the 1930s and that this virus then spread to East and Southern Africa. The HIV-1 subtype C virus is believed to have been introduced in South Africa in 1960 via immigrant mine workers from neighbouring African workers (Wilkinson *et al.*, 2015a).

Of the 36.7 million individuals infected with HIV globally, South Africa had 7.2 million people living with infections by the end of 2017, making it the country with the highest number of HIV infected individuals in the world (UNAIDS, 2018). According to the Los Alamos National Library database, 98% of the infections are subtype C and the remaining 2.2% are non-subtype C viruses. The non-subtype C viruses include subtypes A, B, D and G as well as AC, AG, CD, BC and complex recombinants (https://www.hiv.lanl.gov/; accessed 30 August 2018).

2.3.4. Recombinant strains of HIV-1

CRF01_AE is the first documented and reported HIV-1 recombinant. This CRF was first identified in Thailand in 1992, although, it was discovered to be of African origin (Carr *et al.*, 1996; Gao *et al.*, 1996). Recombination is not limited to inter-subtype events, but is also observed between sequences of different HIV-1 groups. This was seen in a study in which a HIV-1 group M/group O recombinant was identified in a Cameroon women's blood sample (Peeters *et al.*, 1999). The discovery of this rare recombinant illustrates the possibility of new HIV-1 variants emerging.

The prevelance of inter-subtype recombinants in the global pandemic is estimated at between 18-20% (Buonaguro *et al*., 2007; Hemelaar *et al*., 2011). The formation of a viral recombinant requires co-infection or super-infection of two or more viruses into the same target cell (Liu *et al.*, 2002). Co-infection occurs when an individual is infected with two different strains at the same time, whereas super-infection occurs as a result of an already HIV-1 infected individual being infected with a new strain (Yerly *et al.*, 2004).

HIV is a diploid virus, containing two strands of RNA. Homologous recombination can occur in a cell that is co-infected with two different, but related strains (Burke, 1997). Retroviral recombination occurs often during reverse transcription and is a consequence of the virus having two viral genomes packaged into one virion (Hu and Temin, 1990; Moutouh *et al.*, 1996). The viral RNA of both viruses (For example, a subtype A and subtype B virus) enter the host cell and is transcribed to DNA by the reverse transcriptase enzyme. The converted DNA of the subtype A virus and subtype B virus are integrated into the host cell's genome. The proviral DNA strands are converted to RNA by the host cell's RNA polymerase and is transported out of the cell's nucleus and into the cytoplasm where they are packaged into the same virion. Once released from the cell, the new virion goes on to infect a neighbouring host cell in which its subtype A and B RNA strands are reverse transcribed (Taylor *et al.*, 2008). The reverse transcriptase jumps back and forth between the two RNA strands, generating a recombinant of the two parent strands (Burke, 1997). The newly synthesised subtype A/B viral strands are integrated into the host cell's genome. Upon replication,

mosaic subtype A and subtype B RNA strands are packaged into a new virion, creating an A, B recombinant virus (Figure 2.8). Each RNA strands of the recombinant virus will be composed of both parental subtypes (Taylor *et al.*, 2008).

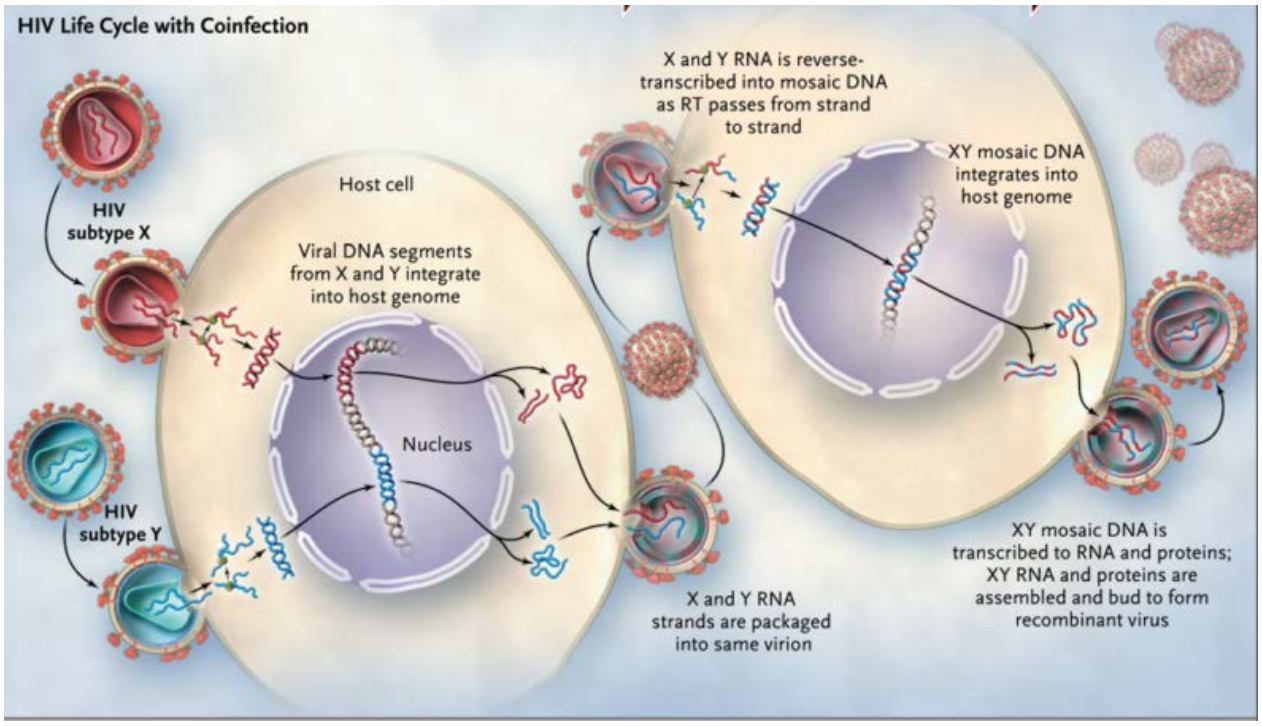

Figure 2.8: Process of the formation of a recombinant HIV virus (Taylor, 2008)
The process shows one target cell being infected with to different HIV subtypes. The viruses are reverse transcribed, integrated in to the host genome, then replicated and packaged into new virion. The new virion infects neighbouring cells and generate recombinant viruses.

Recombinant viruses are most prevalent in regions where many subtypes co-circulate (Lau and Wong, 2013). A recombinant virus is characterised by having distinct regions of their genome match the consensus sequences for different pure subtypes. A region defined as a break point distinguishes between the two different subtype sequences within a recombinant genome. CRFs are named according to the subtypes present in their genome as well as a corresponding number indicating when the virus was identified (Ramirez *et al.*, 2008). For example, CRF01_AE is the first reported and documented CRF and is composed of subtype A and subtype E genomic regions (Lau *et al.*, 2013). Complex recombinants (denoted as CRF_cpx) have genomes that are composed of three or more different subtypes (Ramirez *et al.*, 2008). The emergence of new CRFs are increasing considerably over the years. In 2013 there were 55 documented CRFs in the HIV Los Alamos database (Lau *et al.*, 2013) and by October 2018 there were 97 (www.hiv.lanl.gov ; Accessed 30 October 2018). In a period of five years, 35 new CRFs have been documented.

2.4. Phylogenetic analyses methods

2.4.1. Phylogenetics

Phylogenetics is the study of finding relationships between species or genes using molecular biology and mathematics (Nee, May and Harvey, 1994; Lemey, Salemi and Vandamme, 2009). Molecular phylogenetics uses the information in DNA or genetic sequences to construct a phylogenetic/evolutionary tree based on the similarities and/or differences between the sequences. The more similar two sequences are, the closer they will be positioned on the tree (Wright, 2017). Phylogenetics is an integral component in HIV studies as it is able to determine the relatedness or divergence between different HIV groups/subtypes. Phylogenetic analyses are a key aspect of this project and a brief outline of each major step is described below.

2.4.2. Aligning homologous sequences

A multiple sequence alignment (MSA) is the foundation of any molecular phylogenetic analysis. A MSA is an alignment of sequences in which all nucleotides or amino acids in the same column of the alignment are considered to have a common evolutionary origin (Katoh, 2017). If the same nucleotide is observed at the same position in multiple sequences, the position is believed to be conserved in evolution. However, if the nucleotide in the same position differs between multiple sequences, it is believed that both derived from the same ancestral state, which could be one of the sequences in the multiple alignment or an entirely different sequence. The evolutionary theory states that all organisms have evolved from a common ancestor (Bacon and Anderson, 1986). A number of MSA programmes are available, including ClustalX, Multiple Alignment using Fast Fourier Transform (MAFFT), SAM T-coffee and Muscle. The programmes differ in several ways, mainly in their ability to align large numbers of sequences or complex sequences, and taxa that are only distantly related (Katoh *et al.*, 2017; Wright, 2017).

2.4.3. Selecting a model of evolution

The evolutionary distance between two sequences is estimated based on the number of nucleotide substitutions between them. Nucleotide substitutions are categorised as either transitions or transversions. A transition is the substitution of a purine for a purine (adenosine or guanine) or a pyrimidine for a pyrimidine (thymine or cytosine). All other forms of substitutions are transversions (purine for pyrmidine or vice versa) (Nei and Kumar, 2000). Various mathematical models are used to estimate the evolutionary distance between sequences based on nucleotide substitutions. The most widely used methods are included in the MEGA model test tool. These include the Jukes-Cantor, Tajima-Nei, Kimura 2-Parameter, Tamura 3-Parameter and Tamura-Nei models (Tamura *et al., 2013)*. The three models most commonly used on HIV datasets are the Hasegawa, Kishino, Yano (HKY) (Hasegawa *et al.,* 1985); the General Time-Reversible (GTR) (Yang *et al*., 1994) and Kimura 2-parameter (Kimura, 1980) models. These models are best at estimating the evolutionary

distance between ancestral and query HIV sequences. A model test is therefore carried out on a dataset to determine which model best estimates the real evolutionary distance, as opposed to the observed distance, between the sequences.

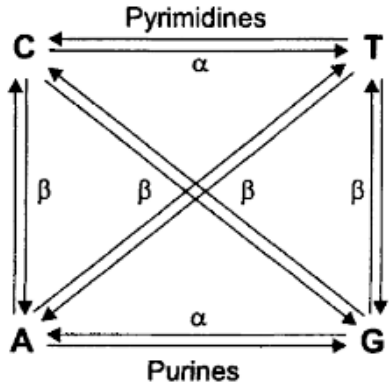

Figure 2.9: Nucleotide substitutions (Nei and Kumar, 2000)
Nucleotide substitutions can be in the form of transitions (purine to purine/ pyrimidine to pyrimidine) or transversions (purine to pyrimidine and vice versa). α and β are the rates of the substitutions.

2.4.4. Phylogenetic tree inference

A phylogenetic tree is a graphical representation of the evolutionary relationships between species based on their genetic closeness (Wright, 2017). A phylogenetic tree is useful in providing information about the origin and spread of disease and to discover evolutionary relationships between groups of organisms (Mahapatro *et al.*, 2012). A phylogenetic tree consists of nodes connected by branches (Figure 2.10) (Hall, 2011). An internal node represents the last common ancestor of the branches diverting from it. External nodes represent the sequences used to derive the tree. The length of the branches correspond to the amount of evolution between the sequences. For example, a longer branch length indicates higher divergence (Baldauf, 2003).

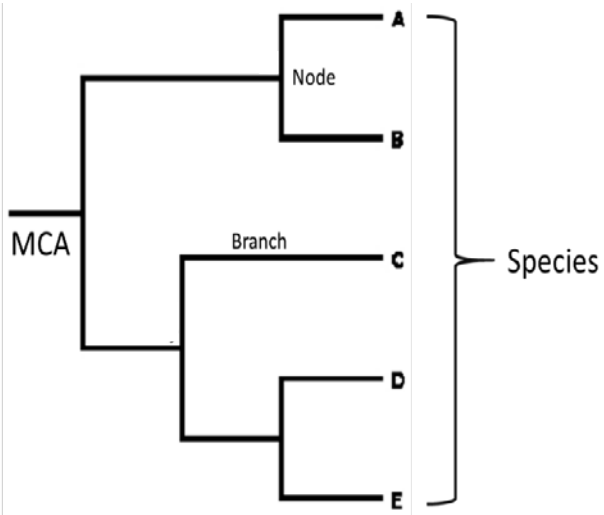

Figure 2.10: A simplistic representation of a phylogenetic tree (adapted from Baum, 2008)
The letters A-E represent the species. The branches of the most closely related species are connected by a node. All branches diverge from the most common ancestor to all the species.
MCA= Most common ancestor

The methods for generating a phylogenetic tree can be divided into two general categories: distance and character based. The distance method is the more simplistic method. In this method, the distance of all pairwise combinations of sequences is calculated and a tree is assembled based on the calculated distances. Common examples of the distance method are the Unweighted pair-Group Method with Arithmetic Mean (UPGMA) and the Neighbour Joining (NJ) algorithmic methods. The character-based methods, including Parsimony, Maximum Likelihood and Bayesian Inference, compares characters within each column of a multiple alignment, providing more information (Hall, 2011). Parsimony identifies the most likely tree to be the one that requires the fewest number of changes to explain the data in the alignment. Maximum likelihood searches for a tree that maximises the chance likelihood of observing the data using a model of evolution. Bayesian inference is a variant of maximum likelihood (Wright, 2017; Felsenstein, 1981).

Bootstrapping is a statistical method used to test phylogenetic accuracy. It involves repeatedly building trees from subsamples of the dataset and calculating the frequency at which the parts of the tree are reproduced in each of the random subsamples. If a group is identified in every subsample tree, the bootstrap value would be 100%. If it is only found in half of the subsample trees, the bootstrap value will be 50% (Wright, 2017; Baldauf, 2003).

Chapter 3

Methodology

3.1. Introduction

The aim of this study was to characterise two possible new HIV-1 C, D near full-length genome recombinant forms in South Africa. The following chapter describes the materials and methods used to achieve the aim and objectives. Figure 3.1 depicts the workflow that was followed to achieve the objectives.

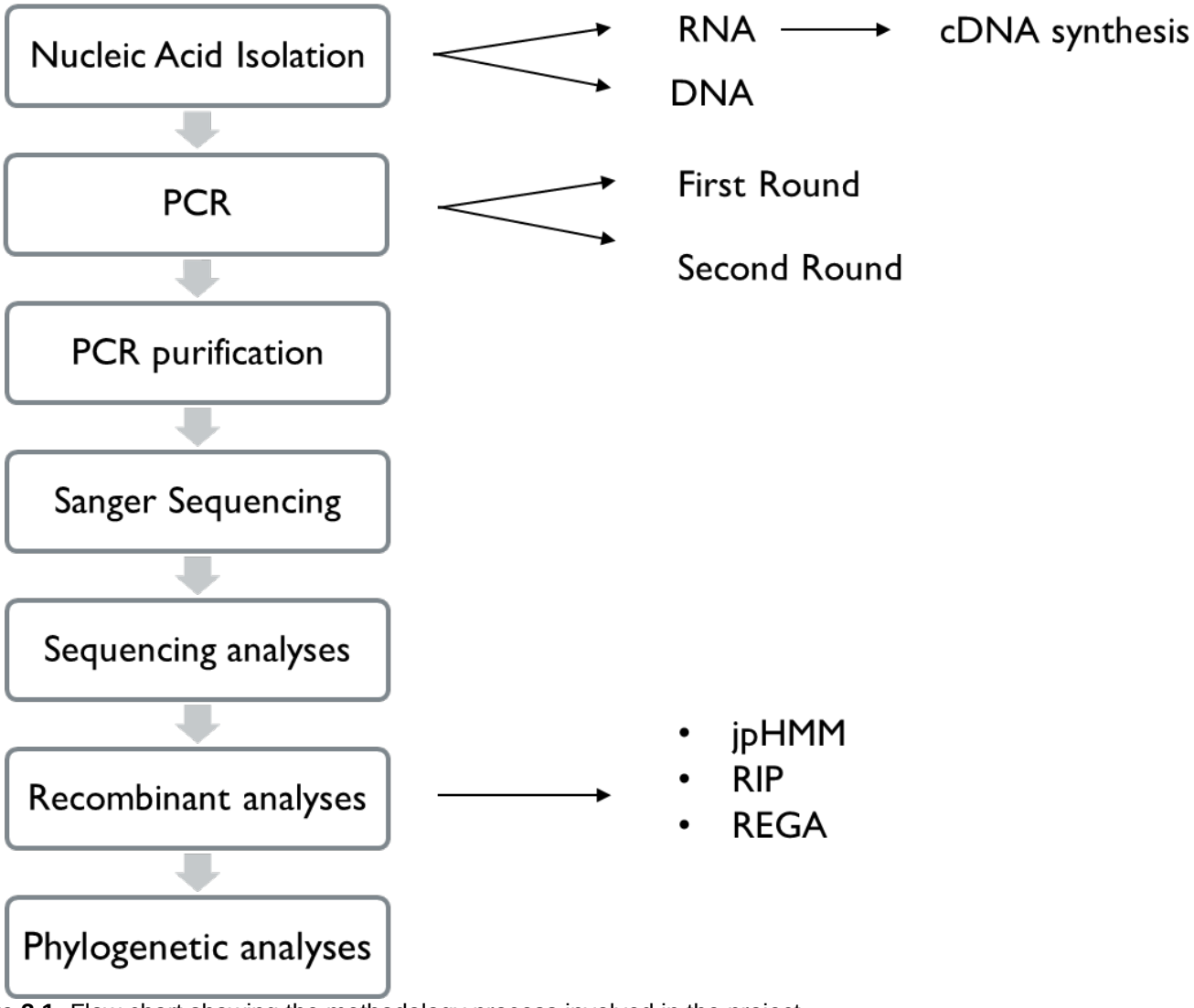

Figure 3.1: Flow chart showing the methodology process involved in the project

3.2. Reagents and equipment

Several different reagents and equipment were used to achieve the objectives set out. The following tables identify all kits (Table 3.1), equipment (Table 3.2) and software (Table 3.3) that contributed to the completion of this study.

Table 3-1: Table of all reagents used

Product/ kit	Application	Supplying company	Catalogue number
Benzonase Nuclease	Pre-treatment step	Merck, Germany	9025-65-4
QIAmp viral RNA mini kit	RNA extraction	Qiagen, Germany	52906
Superscript IV First-strand synthesis system	cDNA synthesis	Invitrogen, USA	18091050
GoTaq Long PCR mastermix	PCR	Promega, USA	PRM0421
KAPA HiFi HotStart ReadyMix PCR Kit	PCR	Kapa Biosystems, USA	KK2601
Nuclease free water	All experiments	Qiagen, Germany	145045078
Seakem LE® agarose	Agarose gels	FMC BioProducts,, USA	50004
GRgreen	Staining dye	Excellgen, USA	EG-1008
QiaQuick PCR purification	Purification of PCR amplicons	Qiagen, Germany	28106
NucleoSpin gel and PCR clean up kit	Gel extraction	Machery-nagel, Germany	740609.50
BigDyeTM Terminator cycle sequence ready Kit	Sequencing PCR	Applied Biosystems, USA	4 337 035
5 x Sequencing buffer	Sequencing PCR	Applied Biosystems, USA	4 305 603
BigDye Xterminator purification kit	Sequencing clean up	Applied Biosystems, USA	4 374 408
NEB PCR cloning kit	Cloning	New England Biolabs, USA	E1202
LB agar	Cloning	Sigma Life Science, USA	L2897
LB broth	Cloning	Fluka BioChemika, USA	61748
GeneJet plasmid miniprep kit	Plasmid purification	ThermoFischer Scientific, USA	K0502

Table 3-2: Table of equipment used

Equipment	Application	Supplying Company
Veriti thermal cycler	PCR and cDNA synthesis	Applied Biosystems, USA
Nanodrop™ ND 1000	Spectrophotometric reading of DNA/RNA	NanoDrop technologies inc.
E-Gel® iBase™ Power System	Gel electrophoresis	Invitrogen, USA
UV-ITEC Prochem Gel Dock System	Gel visualisation	Whitehead Scientific, SA
Eppendorf Centrifuge 5415D	All experiments	Eppendorf, Germany
Degicen 21R centrifuge	Plate centifugation	Ortoalresa, Spain
ABI prism 3130XL automated DNA genetic analyser	Sequencing	Applied Biosystems, USA

Table 3-3: Table of all the software used

Software	Reference	Website
Sequencher version 5	GeneCodes corporation, USA	https://www.genecodes.com/product/sequencher?page=7
REGA version 3	de Oliveira *et al.*, 2005	http://dbpartners.stanford.edu:8080/RegaSubtyping/stanford-hiv/typingtool/
RIP	Siepel A *et al., 1995*	https://www.hiv.lanl.gov/content/sequence/RIP/RIP.html
jpHMM	Zhang M *et al.*, 2002	http://jphmm.gobics.de/
MEGA version 10.0.4	Tamura *et al.*, 2013	https://www.megasoftware.net/
Geneious version 11.1.3	Kearse *et al.*, 2012	https://www.geneious.com/
Genome detective	Vilsker *et al.*, 2017	https://www.genomedetective.com/

3.3. Ethical considerations

Ethics permission for this study was obtained from the Health Research Ethics Committee (HREC) of Stellenbosch University under the protocol number N15/08/071 and was renewed annually (see appendix). The HREC complies with the South African National Health Act No.61 of 2003. This committee abides by the ethical norms and principles for research, established by the Declaration of Helsinki (World Medical Association, 2013), the South African Good Clinical Practices Guidelines as well as the Guidelines for Ethical Research: Principles Structures and Processes 2004 (Department of Health).

3.4. Sample population and sample collection

Between June 2008 and June 2015, the National Health Laboratory Services (NHLS), Virology Division, Tygerberg hospital received 7169 samples for routine HIV drug resistance testing. We received ethics permission for sequencing and phylogenetic analysis of these samples during this period. The majority of the samples were obtained from clinics / hospitals in the Eastern Cape, Western Cape and Gauteng, from both rural communities and large urban cities. From the 7169 samples, ~6800 HIV sequences were obtained, of which 6156 sequences were used for phylogenetic analysis. The NHLS routinely sequences the partial *pol* region of the HIV genome to detect and determine the presence of drug resistance associated mutations (RAMs) against protease inhibitiors and reverse transcriptase inhibitors. From this cohort, 161 non-subtype C viruses were identified based on sequencing and phylogenetic analysis of the partial *pol* region (~1,4kb; HXB2 positions ~2151-3334) of the HIV genome. In 2011 four possible unique C, D recombinants were first noticed within this cohort. These recombinants had similar breakpoints in the *pol* gene. By the end of 2015 there were 30 C, D recombinants within the non-subtype C cohort (S. Engelbrecht, personal communication). The samples were coded to hide true patient identity. The code is generated by using the province abbreviation and an assigned numerical value. For example, EC148 originated from the Eastern Cape and is number 148 in the cohort. Recombinant Identification Program (RIP), Molecular Evolutionary Genetics Analysis (MEGA), jumping profile Hidden Markov Model (jpHMM), Context-based Modelling for Expeditious Typing (COMET) and phylogenetic analysis in MEGA were used to identify the possible subtype of the virus based on sequences obtained for this region of the genome.

We selected two of these possible C, D recombinants from different provinces and at different time points for further characterisation. The selection of these samples were based on sample volume and availability.

3.5. Nucleic acid extraction

3.5.1. DNA extraction

Human genomic DNA from sample EC148 was extracted with the QiaAmp DNA kit (Qiagen, Germany) using the vacuum method. The genomic DNA was extracted from 140µl of PBMCs. The process was performed according to the manufacturer's instruction manual. Briefly, the DNA was collected on the silica membrane of the spin column during the vacuum step. The DNA subsequently underwent two wash steps using buffers AW1 (Wash buffer 1) and AW2 (Wash buffer 2), removing any residual contaminants. The DNA was eluted with an elution buffer (10 mM TrisCl; 0.5 mM EDTA (Ethylenediaminetetraacetic acid); pH 9.0) and the extracted DNA was collected in a 1,5ml microcentrifuge tube. The extracted DNA was stored at -20°C for later use.

3.5.2. RNA extraction

A pre-treatment step using Benzonase was used to improve the quality of viral RNA. Benzonase (Merck, Darmstadt, Germany) is an endonuclease originating from *Serratia marcescens*. It aims to improve cDNA synthesis efficiency by removing circulating nucleic acid in human plasma (Benzonase Brochure, Merck, Germany). The pre-treatment step is performed by adding 10µl of the enzyme solution to 140µl of plasma sample and is incubated at 37^{o}C for 30 minutes. To inactivate the Benzonase nuclease, 20µl of 2mM EDTA buffer is added to each sample and the sample is incubated on ice for 10 minutes. Thereafter, RNA extraction is done as per normal instructions.

Viral RNA was extracted from 140µl of plasma using the QiaAmp RNA kit (Qiagen, Germany). The kit combines the properties of a silica membrane and the speed of a centrifuge to effectively extract purified RNA. All centrifuge steps were performed in an Eppendorf 5415D centrifuge (Merck, USA) at 13000 x g (gravitational force). The spin protocol was followed as per the manufacturer's instructions. Carrier RNA was used in the process to enhance the binding of the nucleic acid to the silica membrane and to reduce the chance of RNA degradation (QiaAmp RNA kit handbook, Qiagen). In short, the sample is lysed under denaturing conditions to inactivate RNases. The salt and pH conditions in the lysate ensures that substances in the sample, other than RNA, are not retained in the membrane. The RNA binds to the silica membrane of the spin column and undergoes two wash steps using buffers AW1 and AW2 to ensure that any residual contaminants are removed. Finally, the RNA is eluted with 30µl of elution buffer and the purified RNA is collected in a 1,5ml microcentrifuge tube (Thermofischer Scientific, USA) and stored at -80^{0}C. The lower the volume of elution buffer used, the more concentrated the purified nucleic acid will be.

3.6. cDNA synthesis

The extracted RNA was reverse transcribed to complimentary DNA (cDNA) using the Superscript IV First Strand Synthesis system (Invitrogen, USA). This kit makes use of a reverse transcriptase that is purified from *Escherichia coli (E.coli*) and is modified to increase its effectivity (Invitrogen Superscript IV handbook). Briefly, a specific primer binds to the 3' end of the template RNA and the reverse transcriptase generates a cDNA strand by adding deoxynucleotide bases resulting in a RNA-DNA hybrid molecule. The RNA strand of the hybrid is degraded and a second, complimentary strand of DNA is synthesised in place of it, resulting in a double stranded cDNA molecule. The LOW2 primer (TGAGGCTTAAGCAGTGGGTTTC , Gao *et al.*, 1998) was used for this reaction. The protocol was followed as per the manufacturer's instructions with one modification (Table 3.4). An extra incubation step at 55^{0}C for ten minutes was included before the inactivation incubation. This was done in an attempt to increase the success of the assay. The thermocycler conditions were carried out in a Veriti thermocycler (Applied Biosystems, USA). The cDNA was stored at -20^{0}C and used for PCR.

Table 3-4:cDNA synthesis reaction assembly

Reagent	**Final Conc.**	**Volume**	**Temperature (°C)**	**Time (minutes)**
Primer (Low 2)	0.1µM	2		
10mM dNTP mix	0.5mM each	1		
RNA template	<5 µg	10		
		13	65	5
5X buffer	1 x	4		
100mM DTT	5mM	1		
Ribonuclease (OUT)	2.0 U/µL	1		
SSIV		1		
		7		
			55	10
			55	10
			80	10
Rnase H		1	37	20
Total Volume		**11**		

3.7. Nucleic acid quantification

The Nanodrop (Thermofischer, USA) technology was used to spectrophotometrically determine the concentration of DNA in each sample. Nucleic acid quantification is measured by calculating the nucleic acid concentration at a wavelength of 260nm. One microliter of the sample was placed on the pedestal to determine the DNA concentration per microliter. The nanodrop reading provides measurements of DNA concentration as well as DNA purity. The concentration of DNA is recorded as ng/µl. Pure DNA falls between the range of 1.7 and 1.9. Values lower than 1.7 indicate protein contamination while values higher than 1.9 indicated the presence of RNA in the sample (Desjardins and Conklin, 2011).

3.8. Polymerase chain reaction (PCR)

PCR is a molecular assay that makes use of an enzyme to make multiple copies of a DNA strand (Saiki *et al.*, 1988; Mullis *et al.*, 1986). One of the enzymes frequently used for PCR is extracted from the organism *Thermus aquaticus* and is commonly known as T*aq* (Saiki *et al.*, 1988). The Promega GoTaq long PCR mastermix (Promega, USA) was used in this study. The mastermix contains all components required to perform long-range PCRs. These components include an optimized buffer, dNTPs, $MgCl_2$ and an optimized hot start enzyme blend for long-range PCR. The forward and reverse primers, nuclease free water and the DNA template are added to a reaction tube containing the mastermix to perform a reaction. Briefly, the DNA is heated to separate the double strands and generate single strands. The primers bind to their corresponding nucleotide bases on the template. Thereafter the enzyme (*Taq* polymerase) binds to the primer and initiates the addition of complimentary nucleotide bases to the template DNA, generating a double stranded DNA molecule. This is repeated to generate multiple copies of the template DNA. The PCR process is performed in a thermocycler using varying temperatures for the different steps. The denaturation step requires a high temperature to separate the double stranded DNA molecule. The temperature is then lowered for an annealing step in which the primer binds to the single stranded DNA template. Finally, an elongation step, at a slightly higher temperature, involves the incorporation of the nucleotide bases. All PCRs were carried out in a Veriti thermocycler (Applied Biosystems, USA).

3.8.1. NFLG PCR amplification strategy

In an attempt to amplify the near full-length genome (NFLG) of HIV-1 (HXB2 location 634 – 9612bp), the amplification was performed in two rounds of PCRs. The first round PCR targeted the amplification of the NFLG (8957kb) and the second round PCRs targeted the amplification of two overlapping fragments: of a *gag-vpu* fragment (5455bp) and a *pol*-3'LTR fragment (4909bp) (Figure 3.2). A combination of an in-house and an unpublished protocol was followed for both rounds of PCR. All fragments were amplified using the Promega GoTaq Long PCR 2X mastermix.

Primers were diluted to 10µM. Each reaction contained 25µl of the reaction mix, 16µl of nuclease free water, 2µl of each primer and 5µl of template, equating to a final volume of 50µl. Negative controls were included with test reactions. Each negative control was comprised of the same components as the test reactions with the exception of 5µl of template. 5µl of nuclease free water was added to the negative control reaction in place of the template.

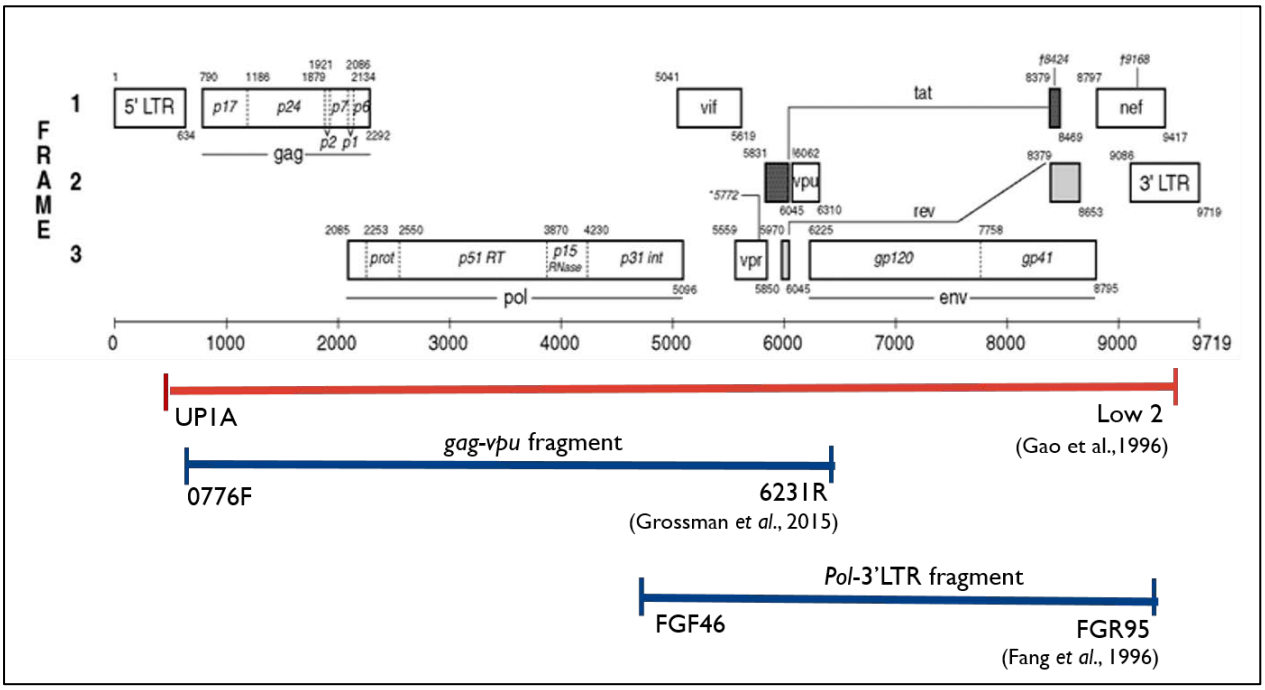

Figure 3.2: PCR amplification strategy used for first and second round PCRs
The red bar represents the first round PCR fragment and the blue bars represent the second round PCR fragments. The blue bar on the left represents the *gag-vpu* fragment and the one to the right represents the *pol*-3'LTR fragment. Adapted from Gao *et al.,* 1996 Grossman *et al.,* 2015 and Fang *et al*., 1996.

The PCRs include a number of heating and cooling steps required for successful amplification of the target fragment/s (Table 3.5). The thermocycling conditions for the first and second round PCRs were the same with the exception of the extension times. The first round PCR required a 9 minute extension time whereas the second round PCRs only required 6 minutes. The extension time is calculated as 1 minute per kb.

Table 3-5: Thermocycler conditions for first and second round PCR
Min=minutes; Sec=seconds

Step	Temperature (^{0}C)	Time	No. of cycles
Initial denaturation	95	3 min.	1
Denaturation	98	20 sec.	30
Annealing	65	30 sec.	
Extension	72	9/6 min.	
Final extension	72	5 min.	1
Soak	4	Indefinite	1

3.8.2. First round PCR

The first round PCR reaction targeted the amplification of a 8957bp NFLG of HIV-1 (HXB2 position 634-9612) using the UP1A and LOW2 primer pair (Table 3.6) described by (Gao *et al.*, 1996). The first round PCR amplification strategy was performed according to the protocol described by (Gao *et al.*, 1998).

Table 3-6: Primers used for first round of the NFLG amplification of HIV-1

Primer orientation	Primer name	Oligonucleotide sequence	HXB2 position
Forward	UP1A	5'AGTGGCGCCCGAACAGG3'	634 → 650
Reverse	LOW2	3'TGAGGCTTAAGCAGTGGGTTTC5'	9612← 9591
Expected size	8957bp		

3.8.3. Second round PCR

The second round PCRs targeted the amplification of the HIV-1 *gag-vpu* fragment (first 5455bp of the NFLG) and the *pol*-3'LTR fragment (last 4909bp of the NFLG).The first round PCR product was used as the template the second round PCRs. The primers used for the amplification of the *gag-vpu* fragment were 0776F and 6231R (Grossmann *et al.*, 2015) (Table 3.7). This region of the genome includes the *gag*, *pol*, *vif* and *vpr* genes of the HIV genome. The primers used for the amplification of the *pol*-3'LTR fragment were FGF46 and FGR95 (Fang *et al.,* 1996) (Table 3.8). This region is from position 9557 to position 2564 relative to the HXB2 genome. The genes covered in this region of the genome are the partial integrase, *vif, vpr, vpu env* and *nef* genes.

Table 3-7: Primers used for amplification of the *pol-vpu* fragment

Primer orientation	Primer name	Oligonucleotide sequence	HXB2 position
Forward	0776F	5'CTAGAAGGAGAGAGAGATGGGTGCGAG3'	776→ 800
Reverse	6231R	3'CTCTCATTGCCACTGTCTTCTGCTC5'	6231←6207
Expected size	5455bp		

Table 3-8: Primers used for amplification of the *pol*-3'LTR fragment

Primer orientation	Primer name	Oligonucleotide sequence	HXB2 position
Forward	FGF46	5'GCATTCCCTACAATCCCCAAAG3'	4648→4669
Reverse	FGR95	3'GGTCTAACCAGAGAGACCCAGTACAG5'	9557← 9532
Expected size	4909bp		

3.8.4. Second round PCRs using Kapa HiFi

The Kapa HiFi HotStart Ready Mix PCR kit (Kapa Biosystems, USA) was used to re-amplify the *gag-vpu* and *pol*-3'LTR fragments of sample WC416 in an attempt to obtain better quality sequences for a region of the genome that was difficult to sequence. The kit makes use of a novel B-family DNA polymerase that is engineered to have increased affinity for DNA (Kapa HiFi handbook). In the ready mix kit, the KAPA HiFi hotstart DNA polymerase is provided in a 2X readymix, containing all reaction components except primers, water and template. The same second round PCR primers (0776F and 6231R; FGR95 and FGF46) were used. The reaction was made up to 50µl and comprised of 25µl of the 2X readymix, 19µl of nuclease free water, 2.5µl of each primer and 1µl of template .The thermocycling conditions for this kit differed to the conditions used for the GoTaq long PCR mastermix kit (Table 3.9).

Table 3-9: Thermocycling conditions for second round PCR using KAPA HiFi

Step	Temperature (^{0}C)	Time	No. of cycles
Initial denaturation	95	3 minutes	1
Denaturation	98	20 seconds	30
Annealing	65	30 seconds	
Extension	72	3 minutes	
Final extension	72	5 minutes	1
Soak	4	Indefinite	1

3.9. Agarose gel electrophoresis

The E-gel (Life technologies, USA) agarose gel electrophoresis system is a bufferless system for gel electrophoresis analysis of DNA samples. The E-gel system contains a pre-cast agarose gel that includes electrodes and is packaged in a disposable, ultraviolet (UV) transparent cassette. These gels are run in a unique design that is a base and power supply in one device.

The E-gel was used to determine the size of the amplified DNA PCR products. A 0.8% gel was used and run at a pre-set voltage and time. Once the run was complete, the gel was visualised under a UV light at a wavelength between 260 and 320 nm and a gel photo was taken with the Syngene™ GeneGenius computer system (Synoptics Ltd.,United Kingdom).

When the E-gel system wasn't available, agarose gels were made manually. A 0.8% agarose gel was made up by combining 0,8 grams of agarose powder (Seakem LE® agarose; USA) with 100ml 1X TAE (Tris-acetate-EDTA) buffer (0.04 M Tris acetate, 0.001 M EDTA). A lower percentage gel results in larger pore size in the gel. This is preferred for large fragments as it allows for easier movement of the large fragment. The gel was cooled before 10µl of GRgreen staining dye (Excellgen, USA) was added to the liquid solution. The solution was poured into a casting tray and allowed to set before samples could be loaded. These gels were run at 80 volts for 1 hour.

For both types of gels, a 1kb DNA ladder (Promega, USA) was used to determine the size of the fragments on the gel. The gels were analysed with the Syngene™ computer system (Synoptics Ltd., United Kingdom).

3.10. PCR product purification

3.10.1. Standard PCR purification

The QiaQuick PCR purification kit (Qiagen, Germany) was used to purify the second round PCR products. This process removes the impurities that may be present in the DNA. The purification

was done according to the manufacturer's instructions. Briefly, the DNA is first bound to the membrane of a spin column by centrifugation at 13 000 x g in an Eppendorf 5415D centrifuge (Merck, USA). Thereafter, the DNA undergoes two wash step with buffer PE (wash buffer) and finally the purified DNA is eluted with 30µl elution buffer. The purified DNA is stored at -20^{0}C until it is to be used again.

3.10.2. Gel extraction

In instances where the second round PCR products displayed non-specific bands on the gel, indicating the amplification of fragments that were not targeted, gel extraction was done to excise the desired fragment only. The NucleoSpin® gel and PCR clean up kit was used for the extraction (Machery-Nagel, Germany). The target fragment was cut out of the gel with the aid of a UV dock system that allowed for visualisation of the fragment in the gel. The excised piece of gel containing the target fragment was mixed with binding buffer NTI and heated at 50˚C for 5-10 minutes until the gel completely dissolved. The DNA was bound to the NucleoSpin column by centrifugation at 11 000 x g for 30 seconds in an Eppendorf 5415D centrifuge (Merck, USA). The DNA underwent two wash steps with buffer NT3 and was eluted with 30µl of elution buffer NE. The collected eluted DNA was stored at -20°C for further use.

3.11. Molecular cloning

3.11.1. Cloning reactions

HIV sequencing results of sample EC148 identified quasispecies as indicated by variability within the *env* region. Quasispecies is a steady-state population of sequences variants derived from a master sequence through mutations (Domingo *et al.,* 2012; Gross *et al.*, 2014). In an attempt to obtain single copy viral sequences, the 3'LTR fragment was cloned. The NEB PCR cloning kit (with NEB-10 beta competent *E.coli* cells) (New England BioLabs, UK) was used. The protocol comprised of three steps: ligation, transformation and plating. Briefly, in the ligation step the insert was mixed with the vector and cloning mix 1 and 2. The ligation reaction was incubated at room temperature for 15 minutes, followed by incubation on ice for 2 minutes. Two microliters of the ligation reaction was added to 50µl of competent *E.coli* cells and then incubated on ice for 20 minutes. Thereafter, the cells were heat shocked at 42°C for 30 seconds. Nine hundred and fifty microliters (950µl) of NEB stable outgrowth medium was added to the cells and was shaken at 37°C for 1 hour. 50µl of the outgrowth was spread on ampicillin selection plates and incubated at 37°C overnight. 100µl and 200µl volumes were also spread on separate plates to determine the optimal volume for bacterial growth.

The LB agar for the plates was made up by mixing 10,5g of lysogeny broth (LB) (Lennox broth with low sodium chloride concentration) agar into 300ml of distilled water. The mixture was autoclaved

and once cooled, 150µl of ampicillin (Sigma, USA) was added. Twenty-five millilitres (25µl) of the LB agar was dispensed into each plate and allowed to cool until it set.

A colony PCR targeting the amplification of the *pol*-3'LTR was done on one colony from each plate. Instead of adding 5µl of template, a single colony was lightly touched with a pipette tip, which was consequently swirled into a PCR reaction mastermix before undergoing the PCR thermocycling conditions as described in Table 3.9.

The cloning PCR products were run on agarose gels. If the target fragment was successfully amplified, it indicated that the cells in that colony successfully took up the insert fragment. In such instances, the same colony used for the PCR was cultured by transferring it into 5ml of LB broth containing ampicillin. The broth containing the colony was shaken at 37°C overnight.

3.11.2. MiniPrep

The GeneJet MiniPrep kit (Thermoscientific, USA) was used to purify the plasmid DNA obtained from the cloning procedure. The cells containing the plasmid were harvested by centrifugation for 2 minutes at 13000 x g using the Eppendorf 5415D centrifuge (Merck, USA). The supernatant was decanted and the remaining pellet was purified as per the manufacturer's instructions. Briefly, the cells were re-suspended, lysed, neutralised and centrifuged. The resulting supernatant was transferred to a spin column and centrifuged to bind the DNA to the spin column membrane. The column was washed twice with 500µl wash solution followed by a dry spin to remove any residual wash buffer. Finally, the DNA was eluted with 30µl of elution buffer and stored at -20°C.

3.12. Sanger Sequencing

3.12.1. Sequencing PCR

The sequencing PCRs were performed using the BigDye™ Terminator Cycle Sequencing Ready Reaction kit, version 5 (Applied Biosystems, USA). The manufacturer's instructions were slightly optimised. Instead of using 1µl of the reaction mix and 6µl of 5X buffer, 0,5µl and 3µl was used, respectively. The volumes were halved as a preferred protocol showed it to yield the same quality of results as the recommended volumes. Firstly, a reaction mix consisted of 4.5µl nuclease free water, 0.5µl reaction mix and 3µl of 5X EDTA buffer. Eight microliters (8µl) of the reaction mix was added to each well. One microliter (1µl) of primer and 1µl of DNA sample was added to each well separately resulting in a total volume of 10µl. The wells were sealed and spun down at 13 000 x g for 2 minutes on an Eppendorf 5415D centrifuge (Merck, USA). The sequencing PCR included one cycle of denaturation at 96^0C for 10 seconds followed by 25 cycles of annealing and elongation at 55^0C for 7 seconds and 60^0C for 4 minutes, respectively. A final hold / soak step was included at 4^0C for an indefinite time.

3.12.2. Sequencing clean up

The BigDye X-Terminator™ Purification Kit (Applied Biosystems, USA) was used to remove impurities that may have accumulated from the sequencing PCR reaction. The manufacturer's protocol was followed. Briefly, a mixture of 49.5µl of Sam's solution and 11µl of Exterminator was prepared per reaction. Fifty-five microliter of the mixture was added to each well. The plate was shaken for 30 minutes before being centrifuged for two minutes at 13 000 x g on a Digicen 21R centrifuge (OrtoAlresa, Spain). The plates were sequenced using the automated ABI3730xl sequencing machine (Applied Biosystems, USA) at the Stellenbosch University Central Analytic Facility (CAF; https://www.sun.ac.za/english/faculty/science/CAF).

3.13. Sequencing analyses

The generated raw ABI files were imported into Sequencher version 5.0 (GeneCodes corporation, USA). The software determined the quality of the sequence and allocated a percentage to each sequence based on the quality. Firstly, the quality of the sequences were assessed and sequences of less than 85% quality were removed. The ends of the sequences were trimmed to improve the read quality. Thereafter, a contig was generated and aligned to the HXB2 reference sequence (GenBank accession number K03455). This gave an indication of the position of the fragment according to the HIV genome. The sequences were read and the contig was exported as FASTA format for further analyses.

3.14. Quality control

The quality of the exported sequences were assessed in the HIV LANL quality control tool (https://www.hiv.lanl.gov/content/sequence/QC/index.html). The tool uses a number of tests to find common problems with the query sequence. The analyses determines the subtype, most similar database sequence, phylogenetic trees for each and all sequences with subtype references, the number of stop codons and frameshift mutations and hypermutations within the query sequence. The quality control tool generates a true indication of the quality of the sequences and it shows areas of the sequences that have problems and this allows for the region to be re-checked.

3.15. HIV subtyping programmes

Online HIV subtyping programmes determine the possible subtype of the viral sequence of the samples. Each tool uses a different algorithm to assign a subtype to the virus. jpHMM (http://jphmm.gobics.de/submission_hiv), RIP (https://www.hiv.lanl.gov/content/sequence/RIP/RIP.html) and REGA (http://dbpartners.stanford.edu:8080/RegaSubtyping/stanford-hiv/typingtool/) were used, as they are the most suitable programmes for determining recombinant subtypes. jpHMM is a probabilistic approach, it doesn't align the input sequence to the alignment of possible subtypes as a whole, but it instead does this in local segments. This is useful for

identifying the subtype of recombinant viruses. REGA constructs a multiple sequence alignment with the query sequence and subtype sequences and trees are constructed from this alignment using the HKY distance method in PAUP* software (Swofford, 2002).

3.16. Reference sequences

The partial genome and full length genome reference sequences needed to generate a multiple sequence alignment with the query sequences were obtained from the HIV LANL database (www.hiv.lanl.gov). The references were selected from the curative alignments page. Subtype reference alignment type was selected and all group M (without recombinants) reference sequences were downloaded and used. This set of reference sequences was used for all phylogenetic inferences.

3.17. Multiple sequence alignment

The reference sequences and sample query sequences were imported into Geneious version 11.1.3 (Kearse *et al.*, 2012) for multiple sequence alignments using the MAFFT plugin (Katoh *et al*., 2017). MAFFT aligned the reference sequences to the sample sequence and identified the regions of homology between the sequences. The alignment was trimmed at the beginning and end, according to the query sequence length, to ensure that all sequences were the same length. The alignment was exported as .meg and .fasta files.

3.18. Model test

A model test was conducted on each dataset to determine which evolutionary model to use in inferring the phylogenetic trees. The model test was done in MEGA version 6 (Tamura *et al.,* 2013). From the table of results generated by the model test, the model with the lowest Bayesian Information Criterion (BIC) value was selected as the best model for the dataset.

3.19. Inferring phylogenetic trees

The multiple sequence alignment generated using the MAFFT plug-in in Geneious version 11.1.3 was used to construct a phylogenetic tree in MEGA. Maximum Likelihood phylogenetic trees were constructed for the datasets by selecting the “Construct/test Maximum Likelihood Tree” option under the phylogeny tab in MEGA. The test of phylogeny option was set to Bootstrap method and the number of replicates were set to 1000 for accuracy. The specific model that correlates to the outcome of the model test was selected for each dataset. All other options were left at default. Phylogenetic trees were inferred for the complete genomes as well as for each region of the genomes that identified as different subtypes.

Chapter 4

Results

4.1. Demographics of study samples

Thirty (30) partial HIV-1 *pol* sequences (HXB2 position 2151-3334) obtained from samples in the non-subtype C cohort identified as possible C, D recombinants (Table 4.1). Two samples were selected for further NFLG characterisation based on sample availability and sample volume. The first is a PBMC sample from a 31 year old female from the Eastern Cape, obtained in 2012 and the second is a plasma sample from a 34 year old female from the Western Cape obtained in 2014. The samples were collected at different time points and at different geographical regions within the country, therefore they are not epidemiologically linked. The samples were also of different nucleic acid types. Genomic proviral DNA was extracted from the PBMC sample and viral RNA from the plasma sample. Multiple samples were used to try obtaining NFLG amplification however, these samples did not amplify and only EC148 and WC416 was successfully amplified and were subsequently used for further characterisation.

Table 4-1: Sample demographics of all 30 possible C, D recombinants in non-subtype C cohort
Highlighted samples were selected for characterisation of NFLGs

Patient Code	Specimen date	Age	Gender	Province
GT434	07 May 2008	25	F	Gauteng
GT764	29 April 2009	36	F	Gauteng
EC013	22 June 2009	33	M	Eastern Cape
GT1112	17 February 2010	56	F	Gauteng
WC123	17 February 2011	36	F	Western Cape
GT1608	25 March 2011	57	F	Gauteng
EC045	23 June 2011	28	F	Eastern Cape
WC183	12 August 2011	38	F	Western Cape
EC134	30 November 2011	32	F	Eastern Cape
EC148	12 January 2012	31	F	Eastern Cape
EC199	19 March 2012	31	F	Eastern Cape
EC212	12 April 2012	9	Unknown	Eastern Cape
EC284	12 June 2012	29	F	Eastern Cape
EC532	12 February 2013	36	F	Eastern Cape
EC746	13 August 2013	51	M	Eastern Cape
EC818	11 October 2013	32	F	Eastern Cape

EC852	05 November 2013	32	M	Eastern Cape
EC1031	26 February 2014	9	M	Eastern Cape
WC416	05 March 2014	34	F	Western Cape
EC1097	31 March 2014	28	F	Eastern Cape
Pr136	01 April 2014	Unknown	Unknown	Unknown
EC1128	14 April 2014	4	F	Eastern Cape
EC1143	23 April 2014	45	F	Eastern Cape
Pr147	03 June 2014	Unknown	Unknown	Unknown
WC500	18 July 2014	36	F	Western Cape
EC1406	08 October 2014	2	M	Eastern Cape
WC539	20 October 2014	17	M	Western Cape
EC1456	13 November 2014	26	F	Eastern Cape
EC1484	24 November 2014	28	F	Eastern Cape
EC1562	29 January 2015	38	Unknown	Eastern Cape

4.2. Nucleic acid quantification of WC416 and EC148

The concentration of DNA was determined after the cDNA synthesis step for sample WC416 and after extraction of sample EC148 (Table 4.2).

Table 4-2: Sample DNA concentrations and purity

Sample name	ng/µl	260/280	260/230
WC 416	95.3	1.81	1.05
EC 148	106.3	1.85	1.35

4.3. NFLG amplification of WC416 and EC148

Agarose gel photographs of the first round PCRs were not included due to the bands usually not being visible on the gels. The following gel photographs are of the second round PCRs. The *gag-vpu* and *pol*-3'LTR fragments were successfully amplified for sample WC416 using Promega GoTaq Long PCR (Figure 4.1) and with KAPA HiFi hotstart ready mix (Figure 4.2). Samples GT1112, EC818, Pr136, EC1143 and 118037 were not successful in amplifying both fragments and were excluded further from the study (Figure 4.1). The *gag-vpu* and *pol*-3'LTR fragments were successfully amplified for EC148 (Figure 4.3). With the overlapping fragments being amplified for WC416 and EC148, the NFLG was obtained for both samples.

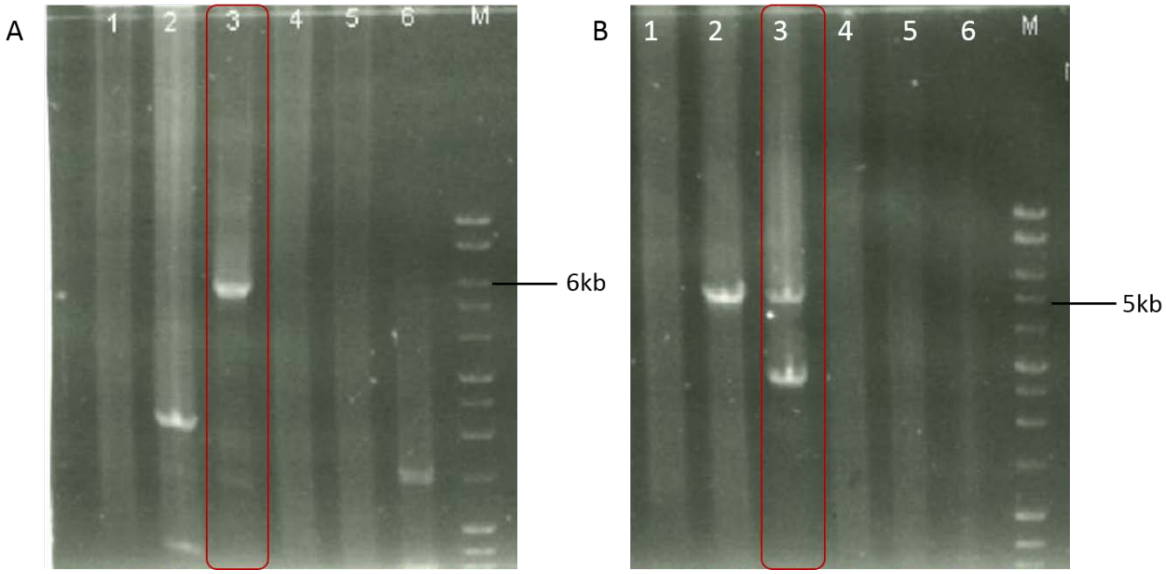

Figure 4.1: Amplification of *gag-vpu* and pol-3'LTR for sample WC416
The bands in the red lane represent the amplification of WC416. Figure A shows the *gag-vpu* amplicon (5455bp) in lane 3. Figure B shows the *pol*-3'LTR amplicon (4909bp) in lane 3. Both fragments correspond with the expected fragment size on the marker. A 0.8% gel was used.
Lane 1 - GT1112, lane 2 - EC818, lane 3 – WC416; lane 4 - Pr136; lane 5 - EC1143, lane 6 – 118037.

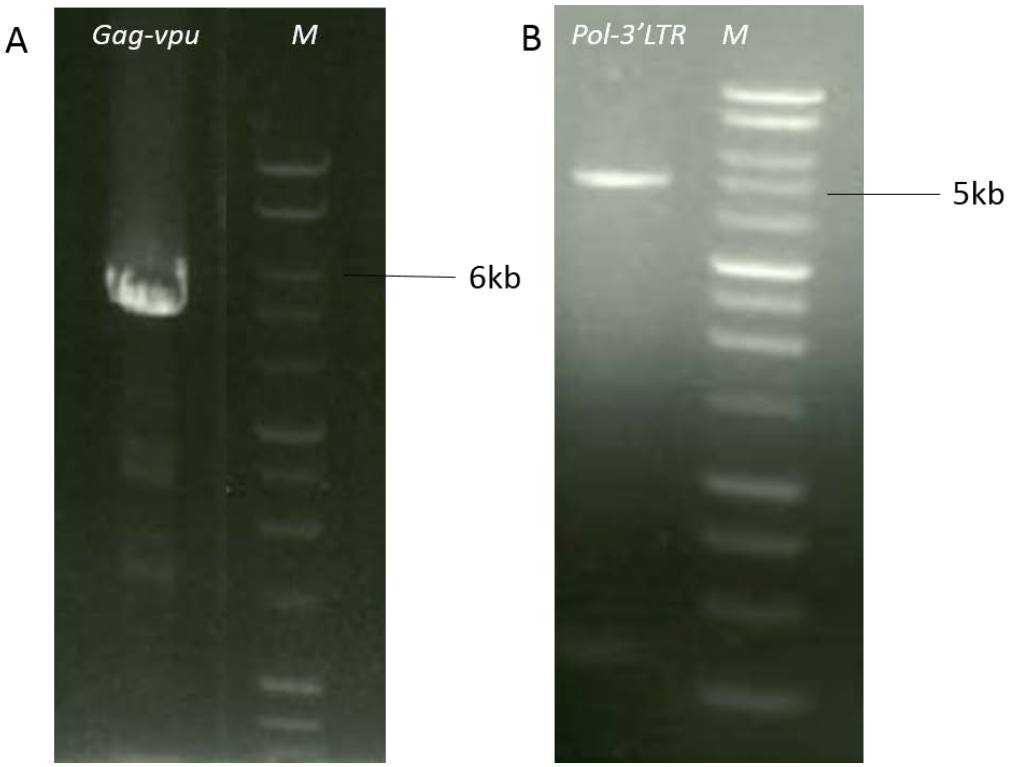

Figure 4.3: Amplification of *gag-vpu* and *pol*-3'LTR for EC148
Figure A shows the *gag-vpu* amplicon (5455bp). Figure B shows the *pol*-3'LTR amplicon (4909bp). Both amplicons align with the expected size on the 1kb DNA marker. A 0.8% gel was used.

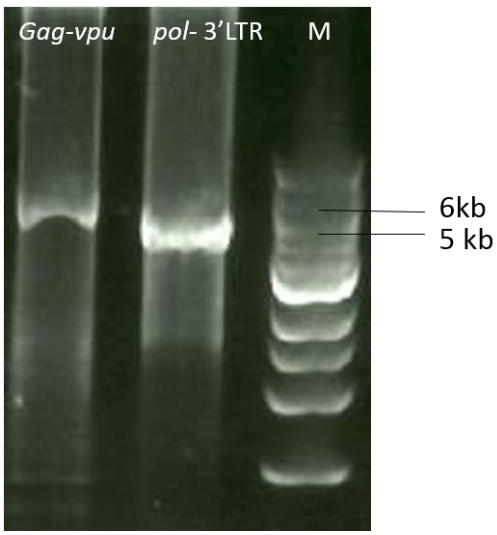

Figure 4.2: Amplification of *gag-vpu* and *pol*-3'LTR for WC416 using KAPA
The *gag-vpu* amplicon (5455bp) is seen in the left lane; the *pol*-3'LTR fragment amplicon (4909bp) in the middle lane and the 1kb DNA ladder in the third lane. Both bands correspond with the desired fragment size. T. A 0.8% gel was used.

4.4. Cloning of *pol*-3'LTR fragment of sample EC148

A cloning colony PCR was done to determine if the fragment was inserted in the plasmids. Amplification of the *pol*-3'LTR fragment was successful, indicating that the fragment was inserted in the plasmid (Figure 4.4). The correct size DNA band fragments (4909bp) were gel extracted and purified for further downstream processing.

The colonies that had positive colony PCR results were cultured and the DNA containing the target fragment was extracted. This DNA was sequenced using primers specific for the insert fragment; however, the sequencing was not successful. In order to obtain the DNA of the insert fragment that was successfully cloned, the bands containing this fragment in Figure 4.4 were gel extracted and purified. The resulting DNA was successfully sequenced.

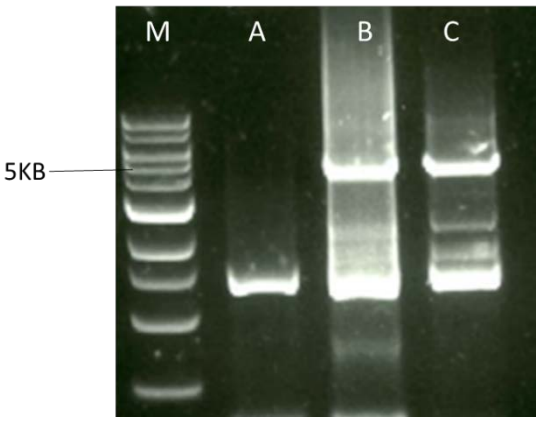

Figure 4.4: Cloning PCR of *pol*-3'LTR of EC148
The *pol*-3'LTR amplicons (4909bp) were successfully amplified in lanes B and C. The amplicons in these lanes corresponded to the expected band size of the marker. The amplification in lane A was not successful as the target fragment was not amplified. A 0.8% gel was used.

4.5. NFLG Sequencing of EC148 and WC416

The HIV-1 NFLG for both samples were successfully sequenced. A 8893bp and a 7720bp NFLG were generated for the viruses of samples WC416 and EC148, respectively. Sequences were imported into the Sequencher software where the quality of each was assessed. The ends and beginning of the sequences were trimmed to increase the overall quality of the sequence. Sequences with quality values below 85% were removed. A contig of all qualifying sequences was generated for each sample.

The sequences were aligned to reference sequence HXB2 using Sequencher version 5 to decipher which portions of the genome they represent. Sample WC416 was successfully sequenced from position 826 – 9719 according to HXB2 numbering, generating a 8893bp NFLG (Figure 4.5). Sample EC148 was sequenced from HXB2 position 779 to 8499, creating a 7720bp NFLG (Figure 4.6).

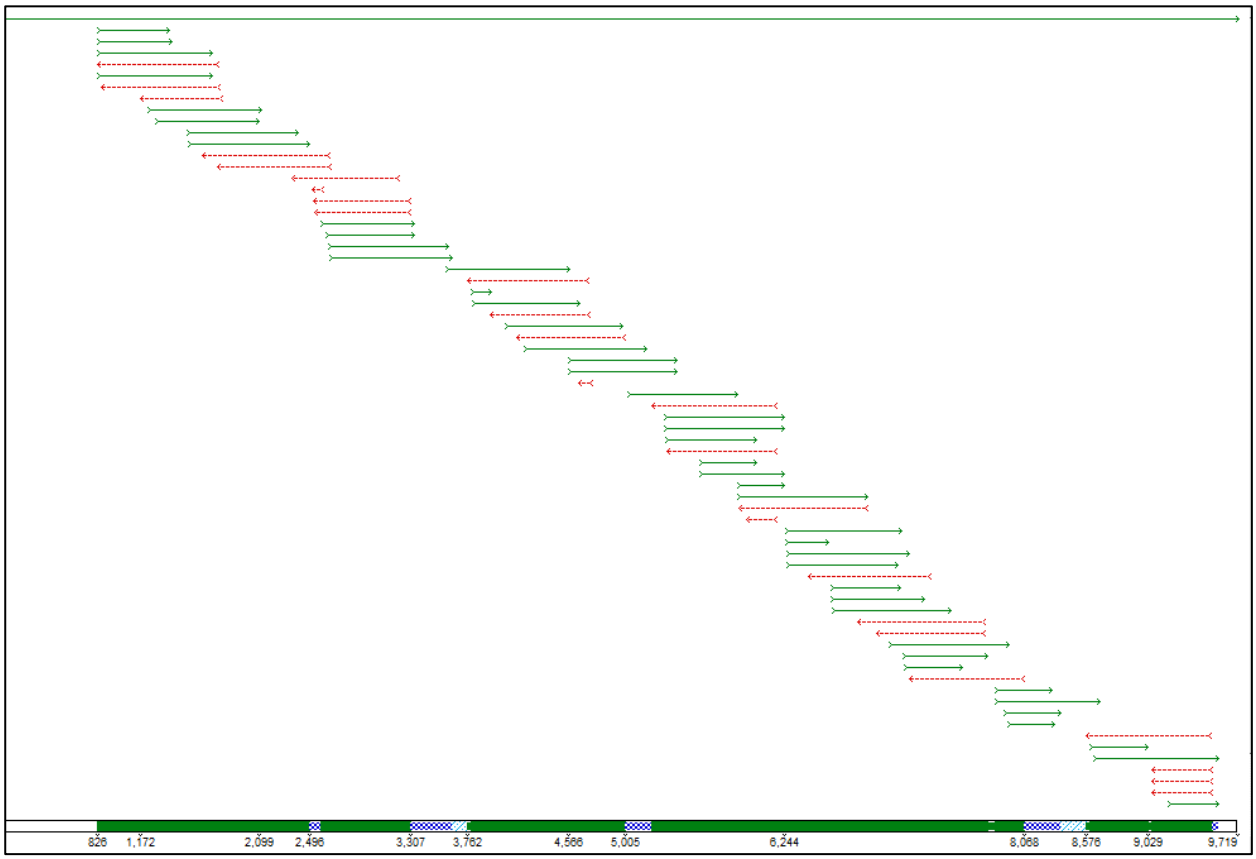

Figure 4.5: NFLG contig of sample WC416
The green and red arrows represent forward and reverse primers, respectively. The green bar at the bottom of the image indicates the region of the genome that was sequenced with reverse and forward primers. Regions of blue represent the portions of the genome that was sequenced with either a reverse or forward primer.

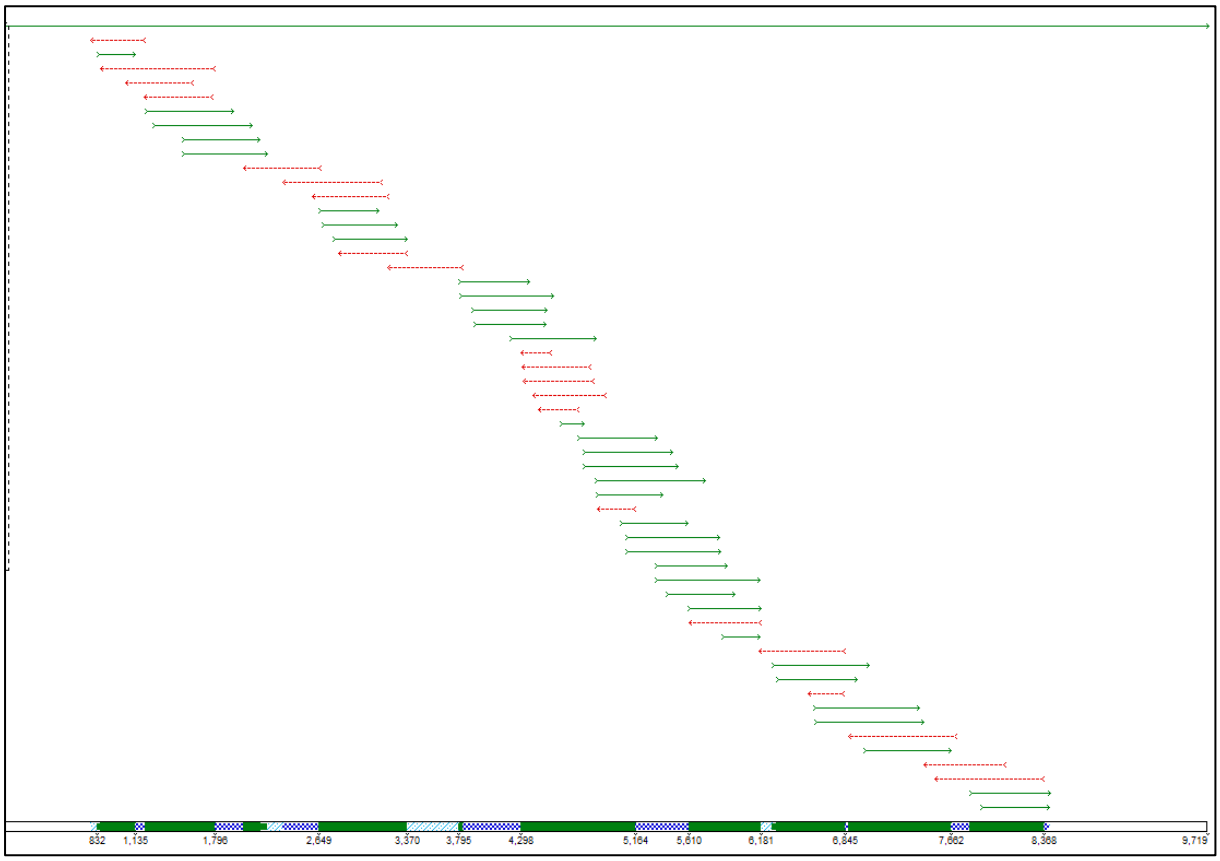

Figure 4.6: NFLG contig of sample EC148
Contig of all sequences for the sample assembled into a contig and aligned to the HXB2 reference genome. Red and green arrows represent reverse and forward primers used. The green bar at the bottom of the image indicates the region of the genome that was sequenced with reverse and forward primers. Regions of blue represent the portions of the genome that was sequenced with either a reverse or forward primer

4.6. Analyses of HIV-1 sequences

After contigs were generated for each sample, the sequences were read to verify any discrepancies that may have resulted from the sequencing reactions. The sequences were scanned for bases that could not be accurately identified by the Sequencher software. This was done by analysing the corresponding chromatograms. The chromatograms were analysed to manually assign a nucleotide to the ambiguous bases, where possible. Once the NFLG sequences were checked, the sequence was uploaded for quality control. Discrepancies identified by the quality control tool were once again checked and verified in Sequencher. This process was repeated until there were no or minimal discrepancies in the quality control results.

4.7. Quality assessment of sequences

The quality control tool was used to determine the overall quality of the sample sequence. The quality control results for sample WC416 detected one stop codon and one frameshift mutation within the sequence. No hypermutations were detected and there was a 0.1% of non-ACGT content. The RIP subtyping used by the quality control tool identified the sample as an A1, C, D recombinant and the BLAST result showed the sample sequence to be most similar to a Tanzanian subtype C sequence from 1992 (Accession number A253310). The quality control results for sample EC148 detected 7 stop codons, 6 frameshifts and possible hypermutations. There was a 0.0% of non-AGCT content within the sample sequence. RIP identified the sample as an A1, C, D recombinant. The sequence most closely related to the sample sequence was detected by BLAST to be a subtype C reference sequence from Botswana in 1992 (Accessioin number KY658708).

4.8. Subtype identification with online subtyping programmes

The jpHMM, REGA and RIP online subtyping programmes identified the possible subtype of each sample. The results from each of these tools are given below.

The results obtained from each subtyping tool were consistent for WC416. Each identified the sample as a complex A, C, D recombinant (Figure 4.7). All three results identified subtype A and C in the *gag*; D and C in the *pol;* C in the *vif, vpr and vpu;* C and A in the *env* and subtype A and C in

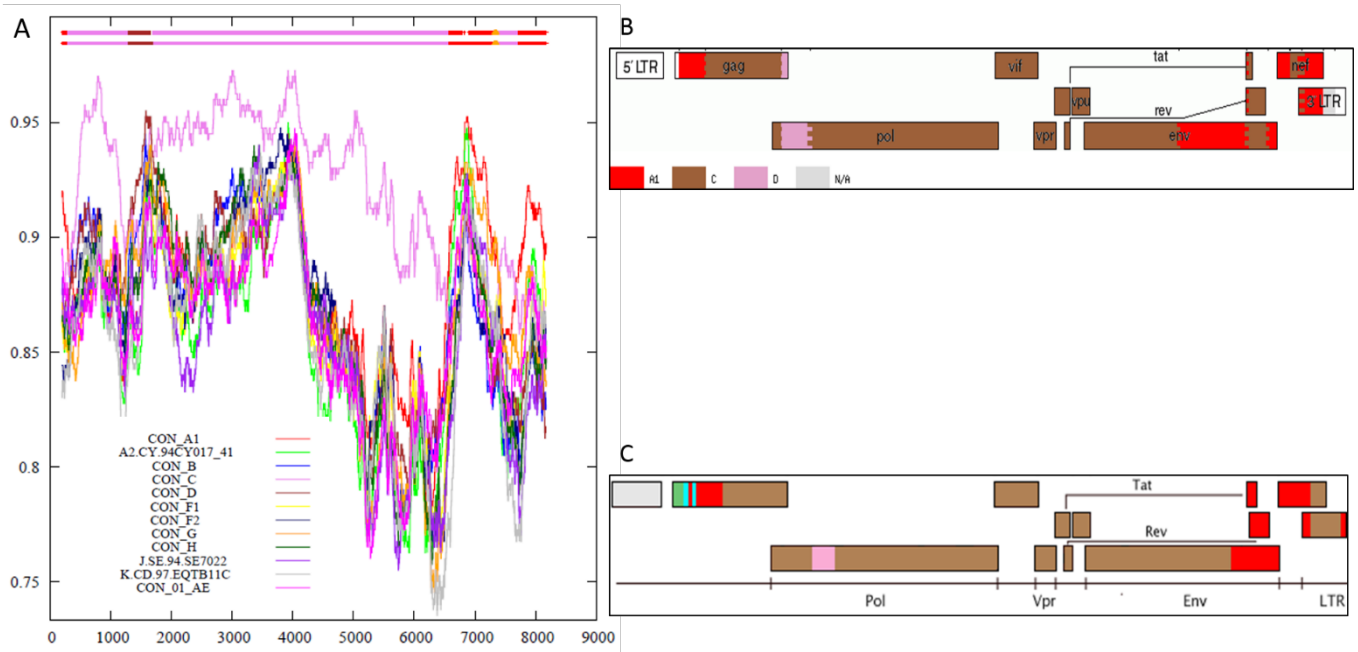

Figure 4.7: Online subtyping tools results for WC416
Figure A represents the RIP similarity plot. The bar at the top of the figure represents the different subtypes identified in the sample; the x axis indicates the distance in nucleotides and the y axis represents the similarity. Pink represents subtype C, brown subtype D and red subtype A. Figure B represents the jpHMM result. The figure shows the distribution of the different subtypes in the viral genome. Subtype A is seen as red, subtype C as brown and subtype D as pink. Figure C represents the REGA results. The figure shows the distribution of the different subtypes within the viral genome. Subtype A is seen as red, subtype C as brown, subtype D as pink and a region of uncertainty is seen as green. All three results identify the sample as a complex A, C and D recombinant.

the *nef* and 3'LTR.

All three online subtyping tools identified EC148 as a complex recombinant of subtypes A, C and D, with jpHMM and REGA identifying regions of additional subtypes (Figure 4.8). RIP identified the sample as a combination of subtypes A, C and D (Figure 4.8A). jpHMM results showed subtypes A, C and D in the *gag,* C, D and K in the *pol;* K and C in the *vif* and subtype C in the *vpr*, *vpu* and *env* (Figure 4.8B). jpHMM did not detect subtype A in the env as RIP and REGA did. To query this result, just the *env* region of the NFLG was analysed in jpHMM. Figure 4.10 shows the jpHMM result for the extracted region and it identifies subtype A in the *env.* The REGA tool assigned subtypes A, C and B in the *gag*; C, B, D and G in the *pol;* C in the *vif, vpr, vpu* and subtype A and C in the *env.* The REGA analysis image (Figure 4.8 C**)** is the approximate recombination pattern without bootstrap confidence. Figure 4.9 illustrates the results of the REGA bootscan analysis and it shows that only subtypes A, C and D are detectable at over 70% support.

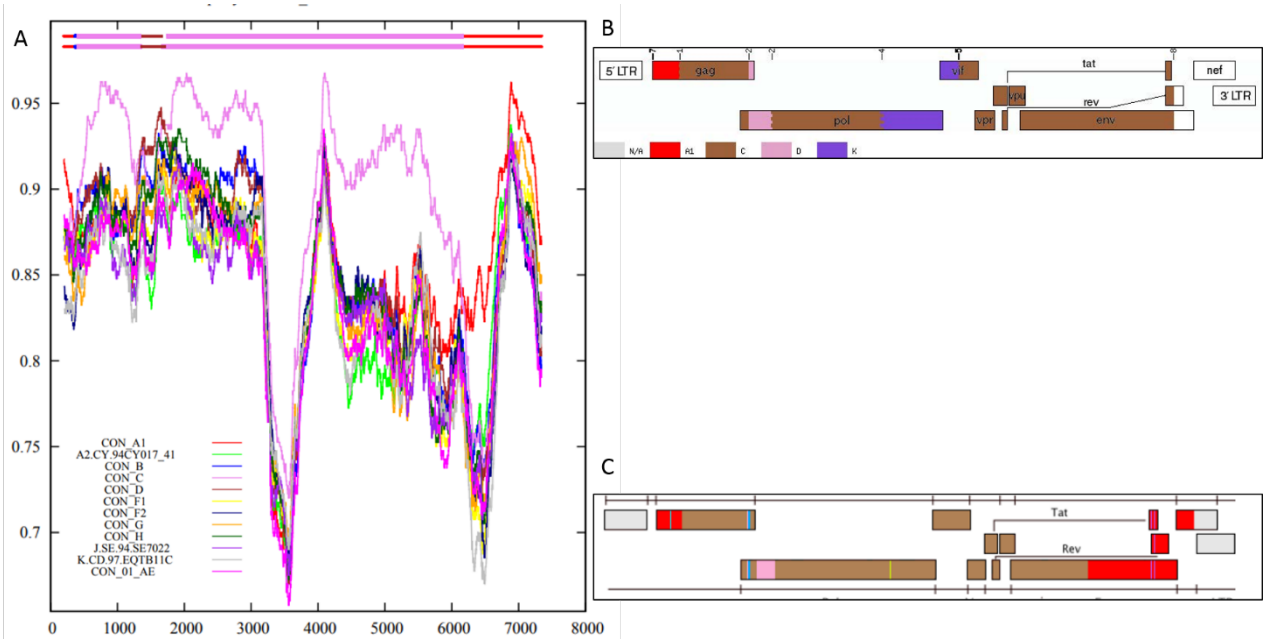

Figure 4.8: Online subtyping tool results for EC148
Figure A represents the RIP result. it identifies subtype A, C and D in the sample. The bar at the top of the figure represents the different subtypes identified in the sample; the x axis indicates the distance in nucleotides and the y axis represents the similarity. Pink represents subtype C, brown subtype D and red subtype. Figure B represents the jpHMM result. The figure shows the distribution of the different subtypes in the viral genome. Subtype A is identified as red, subtype C as brown, subtype D as pink and subtype K as purple. Figure C represents the REGA results. The figure shows the distribution of the different subtypes

Table 4-3: Summary of the outcomes from each online subtyping tool for WC416 and EC148

Genotyping tool	**WC 416**	**EC148**
RIP	A,C,D,C,A,C,A	A,C,D,C,A
jpHMM	A,C,D,C,A,C,A	A,C,D,C,K,C
REGA	A,C,D,C,A,C,A	A,C,D,C,A

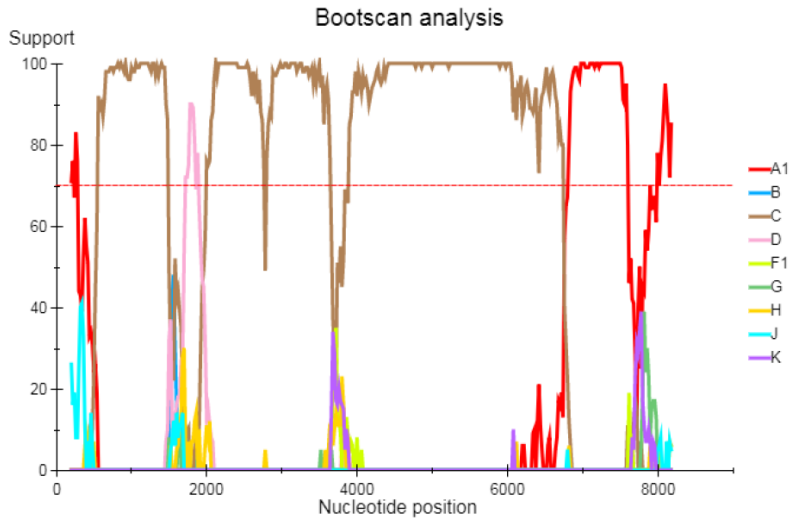

Figure 4.9: REGA bootscan analysis result for EC148
Peaks above the red dotted represents results that are detected at over 70% bootscan support. The y axis represents the support value and the x axis indicates the nucleotide position.

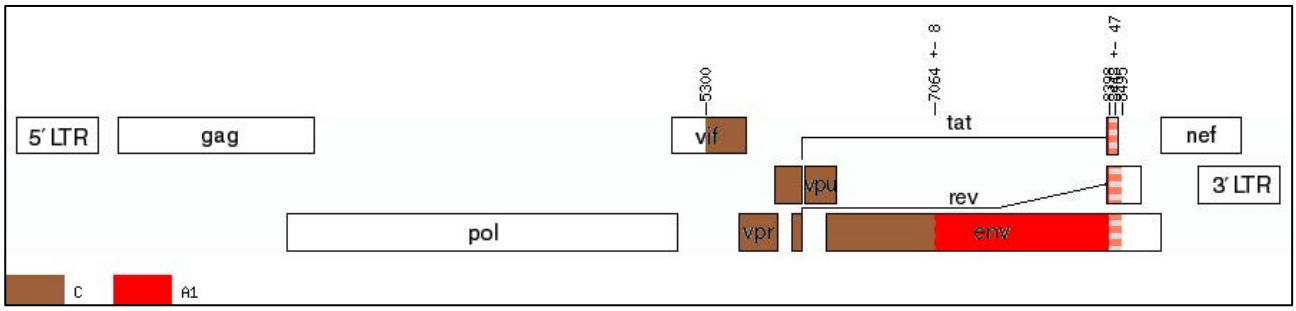

Figure 4.10: jpHMM result for only the *vif, vpr, vpu* and *env*
The *vif, vpr, vpu* and a portion of the *env* identified as subtype C while the remainder of the *env* identified as subtype A.

4.9. Choosing an evolutionary model

A model test was conducted to determine which nucleic acid substitution model best fit each dataset. The test was performed on each dataset that would be used to infer a phylogenetic tree. A phylogenetic tree was inferred for each recombinant segment. For sample WC416, 11 phylogenetic trees were inferred - one for the NFLG and 10 for each portion of the NFLG that identified as a distinct subtype by the jpHMM result. Each portion is described as a fragment. For sample EC148, 8 phylogenetic trees were inferred – 1 for the NFLG and 7 for each subtype. The

model tests used for each of the generated trees is shown in Table 4.3 and Table 4.4. The model with the lowest BIC value was selected as the model most suited for the dataset.

Table 4-4: Model tests used to generate each tree for WC416

Fragment	Size of fragment (bp)	HXB2 position	Selected model
WC416			
NFLG	8736	826-9562	GTR+G+1
1	369	826-1195	HKY+G
2	999	1197-2196	TN93+G+I
3	384	2201-2585	HKY+G
4	4901	2587-7488	GTR+G+I
5	891	7490-8381	GTR+G+I
6	278	8383-8661	K2+G
7	301	8664-8965	K2+G
8	181	8971-9152	K2+G
9	255	9154-9409	K2+G
10	160	9400-9560	K2+G

Table 4-5: Model test used to draw each tree for EC148

Fragment	Size of fragment (bp)	HXB2 position	Selected model
	EC148		
NFLG	7709	790-8499	GTR+G
1	381	790-1171	HKY+G
2	1028	1172-2200	T93+G
3	336	2201-2528	HKY+G
4	1659	2529-4188	GTR+G+I
5	550	4189-4731	T92+G
6	2306	4732-7003	GTR+G+I
7	1343	7005-8327	GTR+G

4.10. Inferring phylogenetic trees

Multiple phylogenetic trees were inferred for each sample. A tree was inferred for the complete NFLG and multiple trees were inferred for each recombinant segment within the sample. All trees were constructed as Maximum Likelihood trees with bootstrap replicates of 1000. Maximum Likelihood trees were inferred as they provide more information and 1000 bootstrap replicates were used for accuracy. The reference sequence set used for the trees were retrieved from HIV LANL and is comprised of 39 group M sequences. For the complete NFLGs trees, an additional group N sequence was included in the reference sequence set, as an outlier.

4.10.1. Sample WC416

In total, eleven trees were inferred for sample WC416. The numbers 1-10 correspond to the position of each fragment relative to the HIV genome (Figure 4.11). Fragments 1, 5, 7 and 9 identified as subtype A by jpHMM. In each tree the fragment sequences were most closely related to the subtype A reference sequences as they formed a distinct cluster together. The trees for fragments 1, 5 and 7 (Figure 4.12, Figure 4.16 and Figure 4.18**,** respectively) had bootstrap values greater than 70% and therefore the results can be accepted with confidence. However, for fragment 9 (Figure 4.20), the phylogenetic tree showed the sequence to be an outlier to the H and J reference sequences. The results displayed on this phylogenetic tree cannot be accepted with confidence due to the lack of bootstrap values over 70% and the short length (255bp) of the fragment sequence.

Fragments 2, 4, 6 and 8 were assigned as subtype C by jpHMM. The phylogenetic tree for each fragment showed them to each cluster with the subtype C reference sequences. Fragment 2, 6 and 8 (Figure 4.13, Figure 4.17, Figure 4.19**,** respectively) sequences formed distinct clusters with the subtype C reference sequences, however, fragment 4 (Figure 4.15) sequence was an outlier to the subtype C reference sequences. Fragment 2, 4 and 6 trees had bootstrap values over 70%. The tree for fragment 8 (Figure 19) had no bootstrap values over 70% and the sequences were short, therefore this result cannot be taken with confidence.

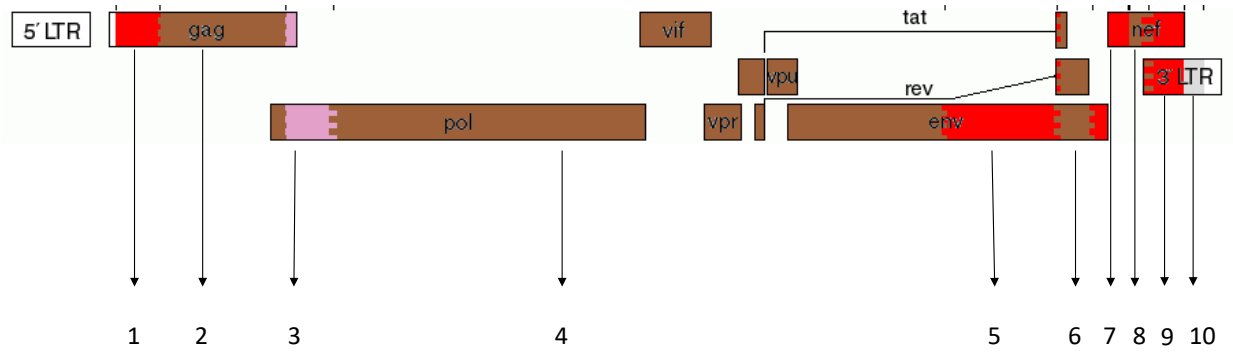

Figure 4.11: Correspondence of fragment number to region on NFLG of WC416.
The NFLG was divided into fragments based on the identification of different subtypes at different regions. The fragments were labelled 1-10 and a phylogenetic tree was inferred for each.

Fragment 3 was identified as subtype D by jpHMM. The phylogenetic tree showed the fragment sequence to be most closely related to subtype D, confirming the jpHMM results (Figure 4.14). The bootstrap values on the tree were over 70% and therefore the results could be accepted with confidence. The jpHMM results identified fragment 10 sequence as being located in the 5'LTR and was unable to assign a subtype to it. The phylogenetic tree inferred showed fragment 10 (Figure 4.21) to be an outlier to subtype J reference sequences; however, this can't be accepted with confidence due to the lack of bootstrap values greater than 70% on the tree and the short length (160bp) of the fragment sequence.

The jpHMM results were confirmed by the phylogenetic tree results in 8 out of the 10 fragments (not in fragments 9 and 10). The sequential order of the different subtypes within the sample is identified as A, C, D, C, A, C, A, C, A. The phylogenetic tree inferred for the complete NFLG (Figure 4.22) is 8736bp and spans from the *gag* to the 3'LTR. The tree shows that the sample sequence is an outlier to the subtype C reference sequences. The tree nodes display bootstrap values of greater than 70% indicating that the results can be accepted with confidence.

Therefore, the online subtyping programme and phylogenetic analyses identified sample WC416 as a complex A, C, D recombinant form.

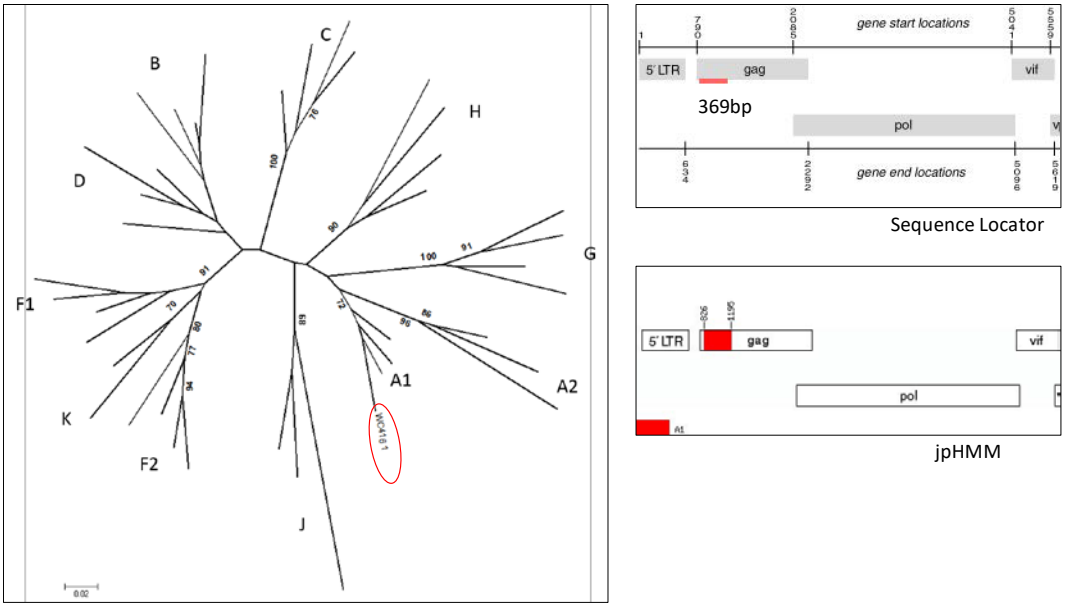

Figure 4.12: Phylogenetic analysis of WC416 fragment 1.
The sample sequence (circled in red) clusters with the A1 reference sequences. Bootstrap values >70% are shown at the nodes; scale of 0.02 used for branch lengths. The sequence locator result shows that the fragment is 369bp and located in the *gag*, HXB2 positions 826-1195. The jpHMM result identified the fragment as subtype A which is seen in red.

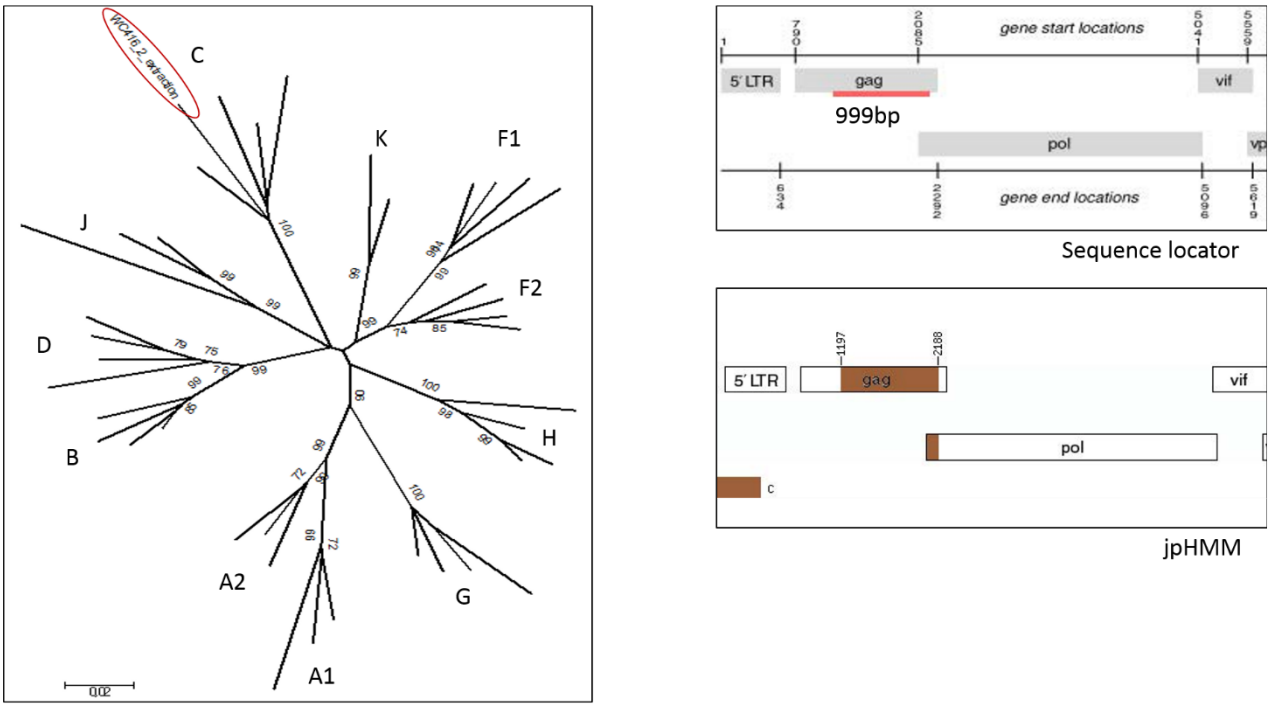

Figure 4.13: Phylogenetic analysis of WC416 fragment 2
The sample sequence is circled in red on the phylogenetic tree and clusters with the subtype C reference sequences. A horizontal scale of 0.01 was used for the branch lengths. The sequence locator result shows that the 999bp fragment is located in the *gag*, HXB2 1197-2196. jpHMM identified the fragment as subtype C which is seen in brown.

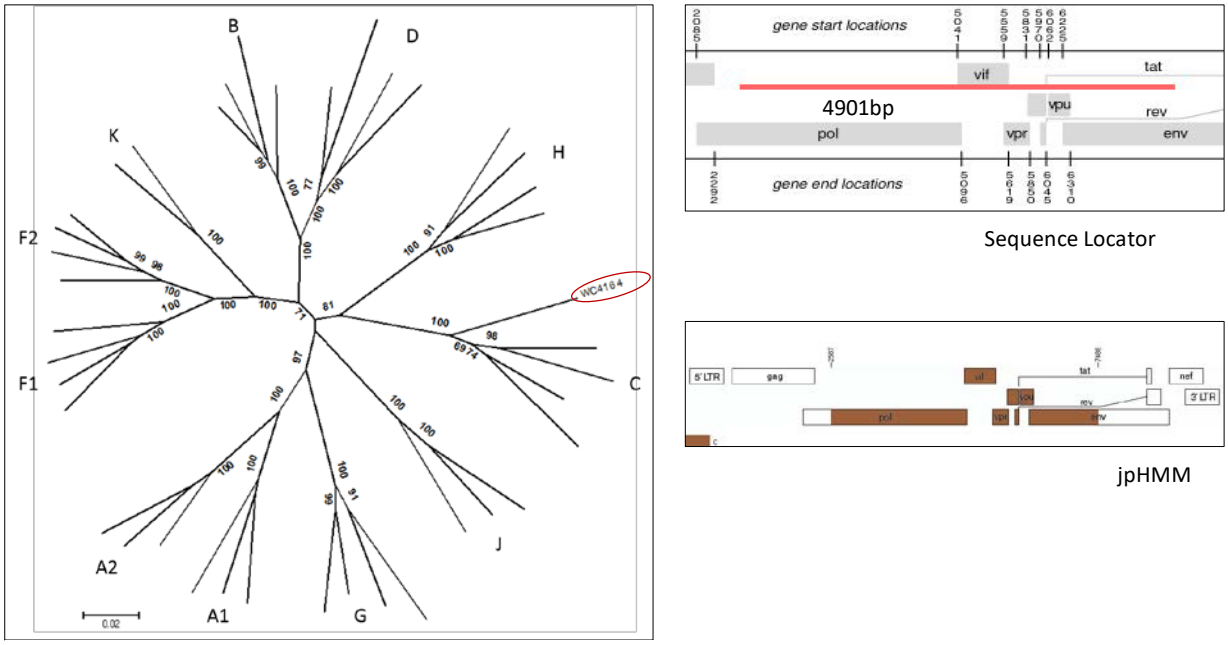

Figure 4.15: Phylogenetic analysis of WC416 fragment 4
The circled fragment sequences clusters with the subtype C reference sequences. A horizontal branch length of 0.02 was used for branch length. The sequence locator result shows the fragment to be 4901bp and located from the *pol* to the *env;* HXB2 2587-7488*.* The jpHMM identified the fragment as subtype C, seen by the brown.

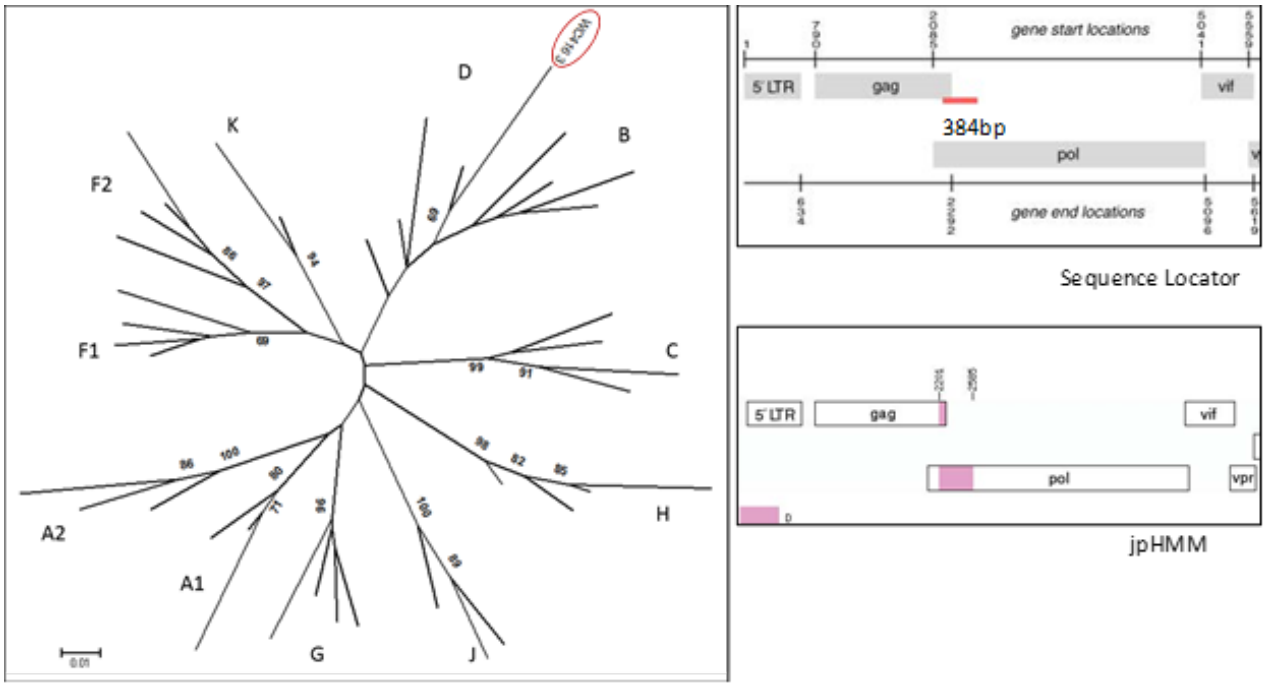

Figure 4.14: Phylogenetic analysis of WC416 fragment 3
The circled fragment sequence clusters most closely with the subtype D reference sequences. A 0.01 horizontal scale was used for branch lengths.The sequence locator result shows that the fragment is 384bp and is in the *pol*, HXB2 2201-2585*.* The jpHMM result identified the fragment as subtype D which is seen as pink.

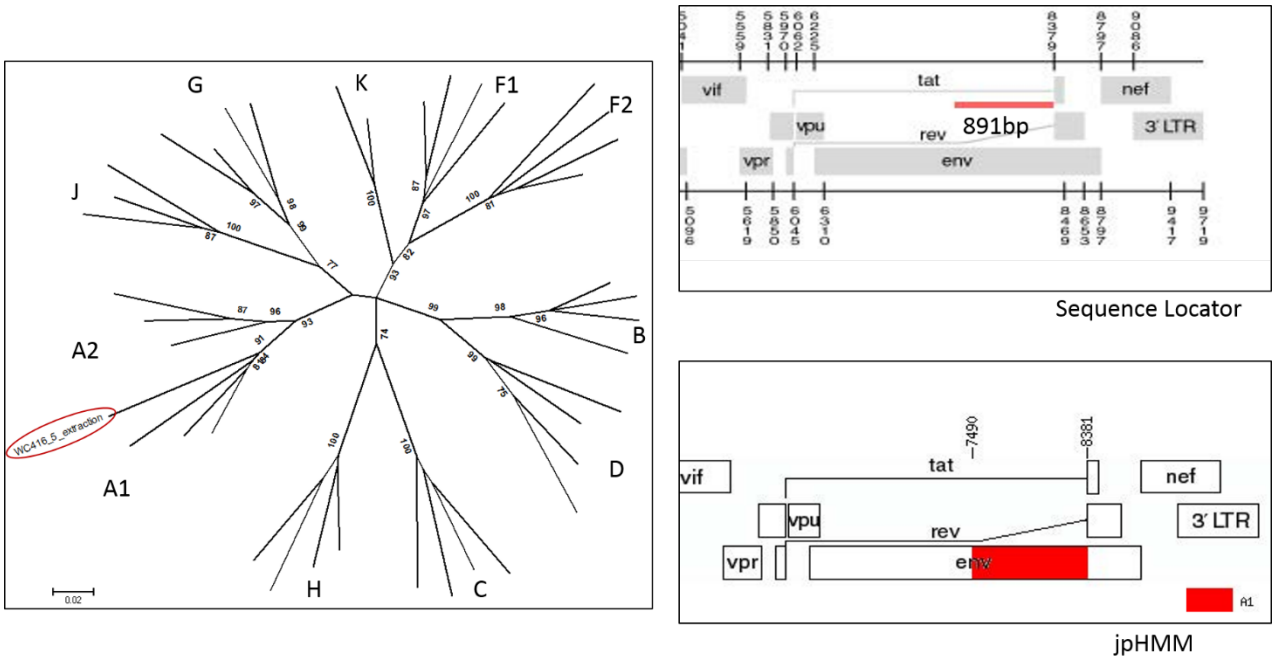

Figure 4.16: Phylogenetic analysis of WC416 fragment 5
The circled fragment sequence clusters with the subtype A1 reference sequences. A 0.02 horizontal scale was used for branch length. The sequence locator result shows the sample to be 891bp and is located in the *env*; HXB2 7490-8381. jpHMM assigned the fragment as subtype A as is seen by the red.

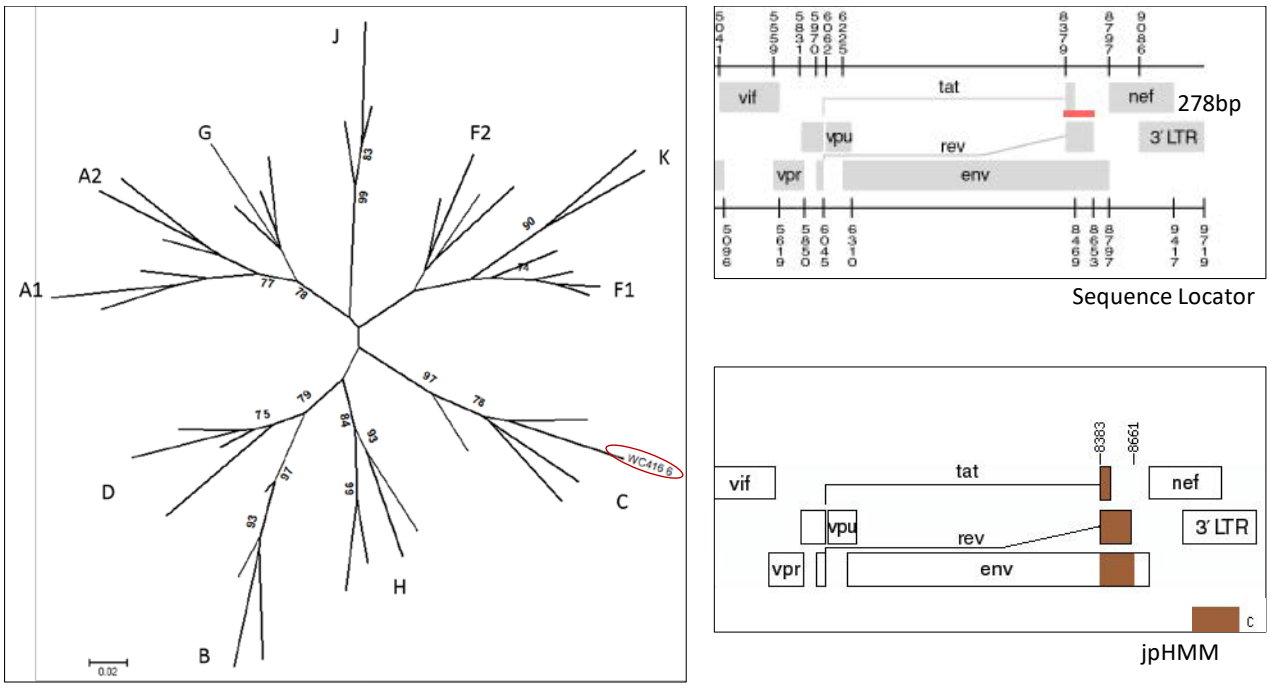

Figure 4.17: Phylogenetic analysis of WC416 fragment 6
The fragment sequence clusters with the subtype C reference sequences. A 0.02 horizontal scale was used for branch lengths. Sequence locator shows the fragment to be 278bp and located in the *env*; HXB2 8383-8661. jpHMM assigned the sample as subtype C shown by the brown.

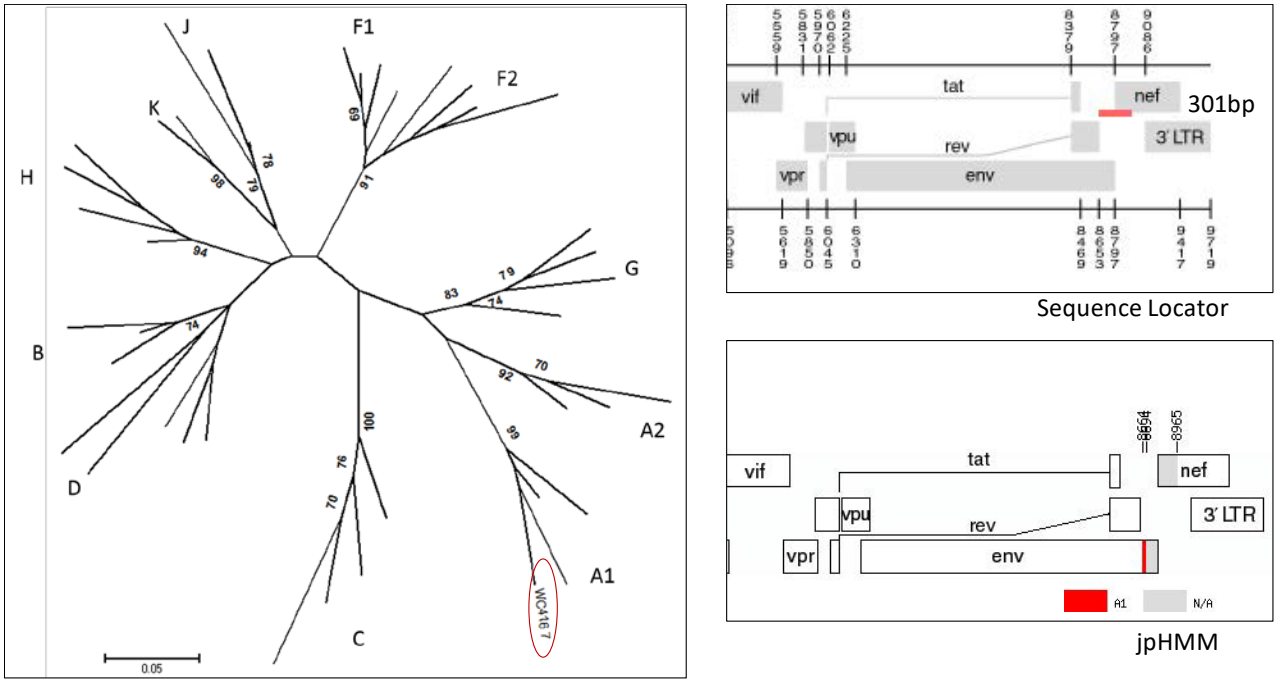

Figure 4.19: Phylogenetic analysis of WC416 fragment 7
The phylogenetic tree shows the fragment sequence clustering with the A1 reference sequences. A 0.05 horizontal scale was used for branch length. The sequence locator result shows that the fragment is 301bp and located in the *env* and *nef*. HXB2 8664-8965. jpHMM assigned this fragment as subtype A, indicated by the red.

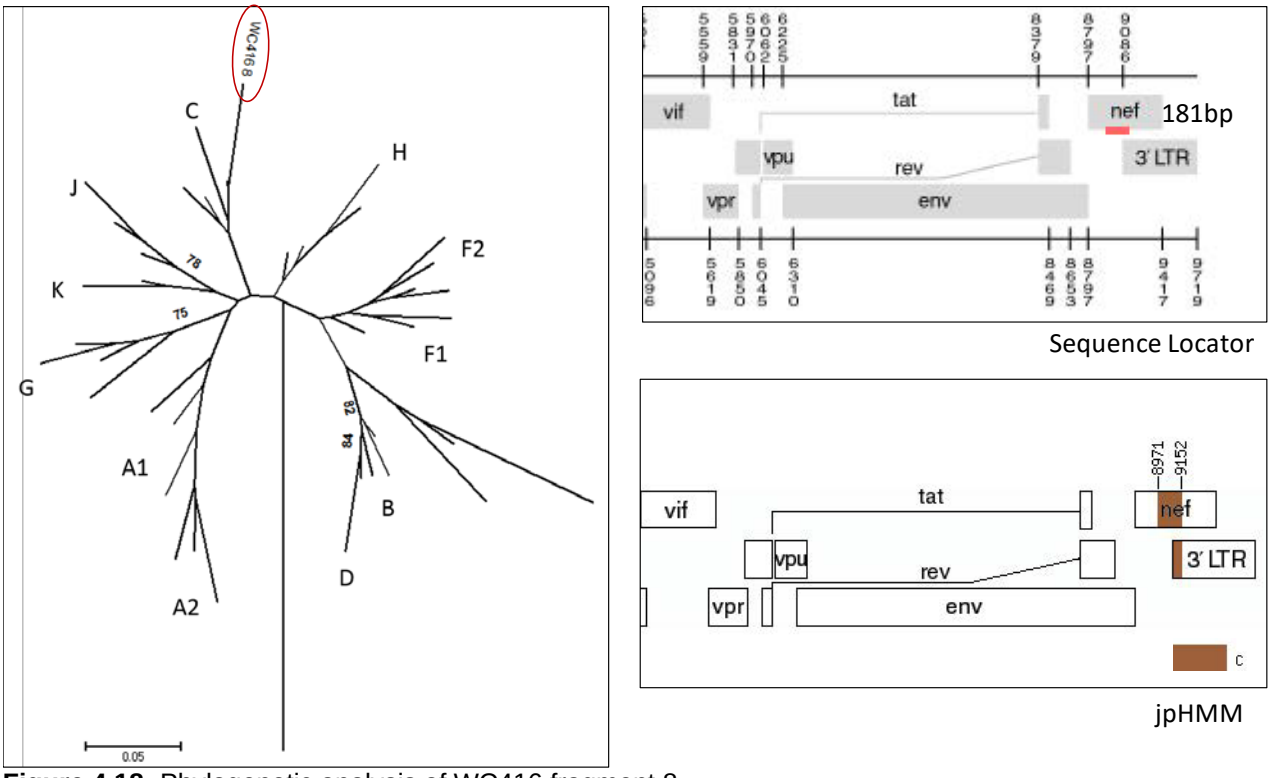

Figure 4.18: Phylogenetic analysis of WC416 fragment 8
The phylogenetic tree shows the fragment sample clusters with the subtype C reference sequences. A horizontal scale of 0.05 was used for branch length. The fragment was identified to be 255bp and located in the *nef* by sequence locator; HXB2 8971-9152. jpHMM was not able to assign a subtype to the sample.

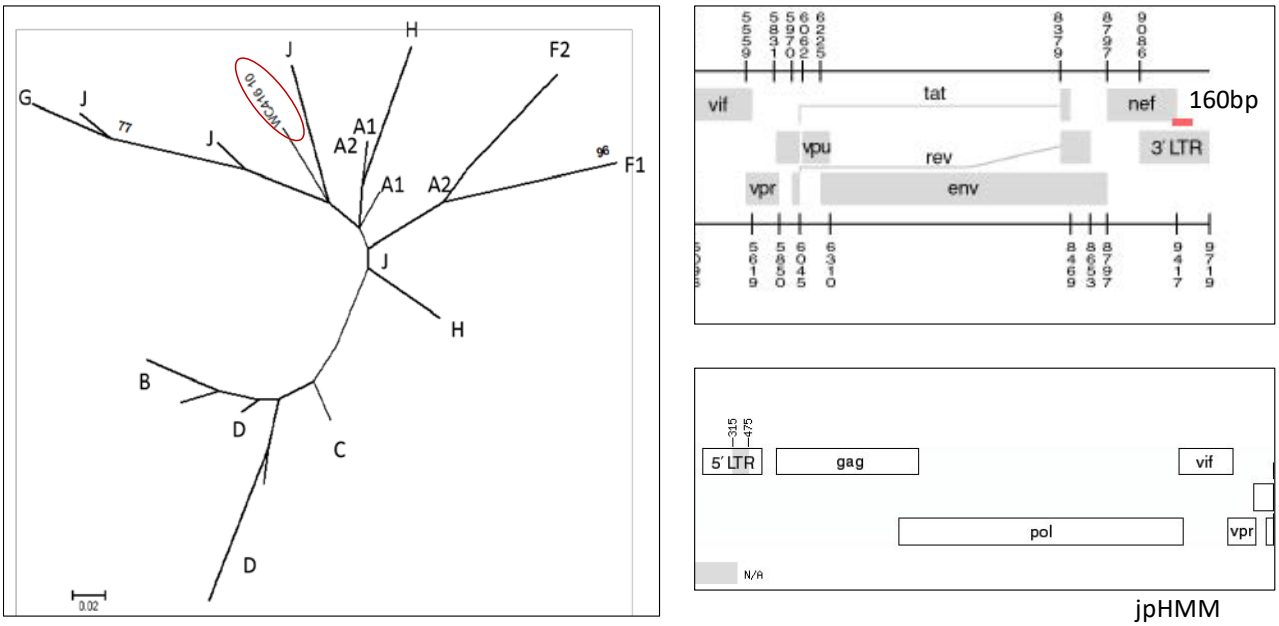

Figure 4.21: Phylogenetic analysis of WC416 fragment 10
The phylogenetic tree shows the fragment clustering with subtype J reference sequences. No bootstrap values are indicated on the tree. A 0.02 scale was used for branch lengths. The sequence locator results show the sample to be 160bp and located in the 3'LTR; HXB2 9400-9560. jpHMM could not assign a subtype to the fragment.

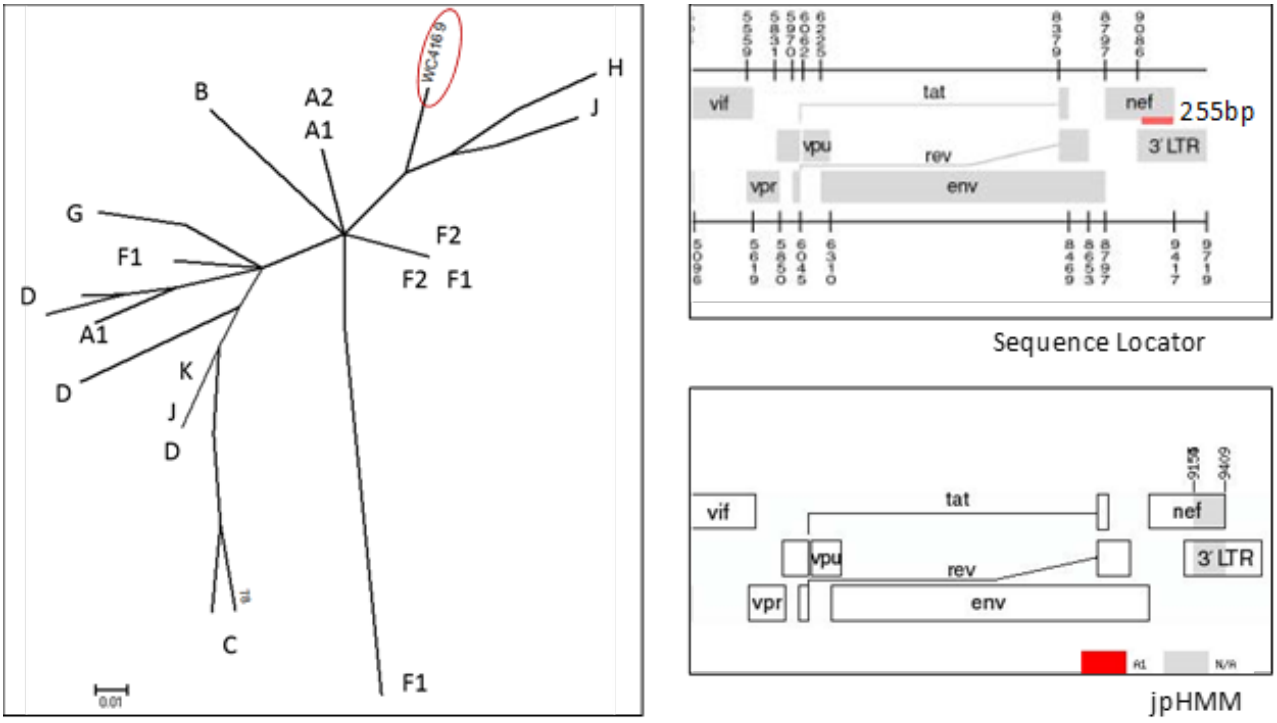

Figure 4.20: Phylogenetic analysis of WC416 fragment 9
The phylogenetic tree shows the fragment to be an outlier to subtype H and J reference sequences however, the tree has no bootstrap values. A 0.01 scale was used for branch lengths. The sequence locator result shows the fragment to be 255bp and located in the *nef*/ 3'LTR; HXB2 9154-9409. jpHMM was not able to assign a subtype to this fragment.

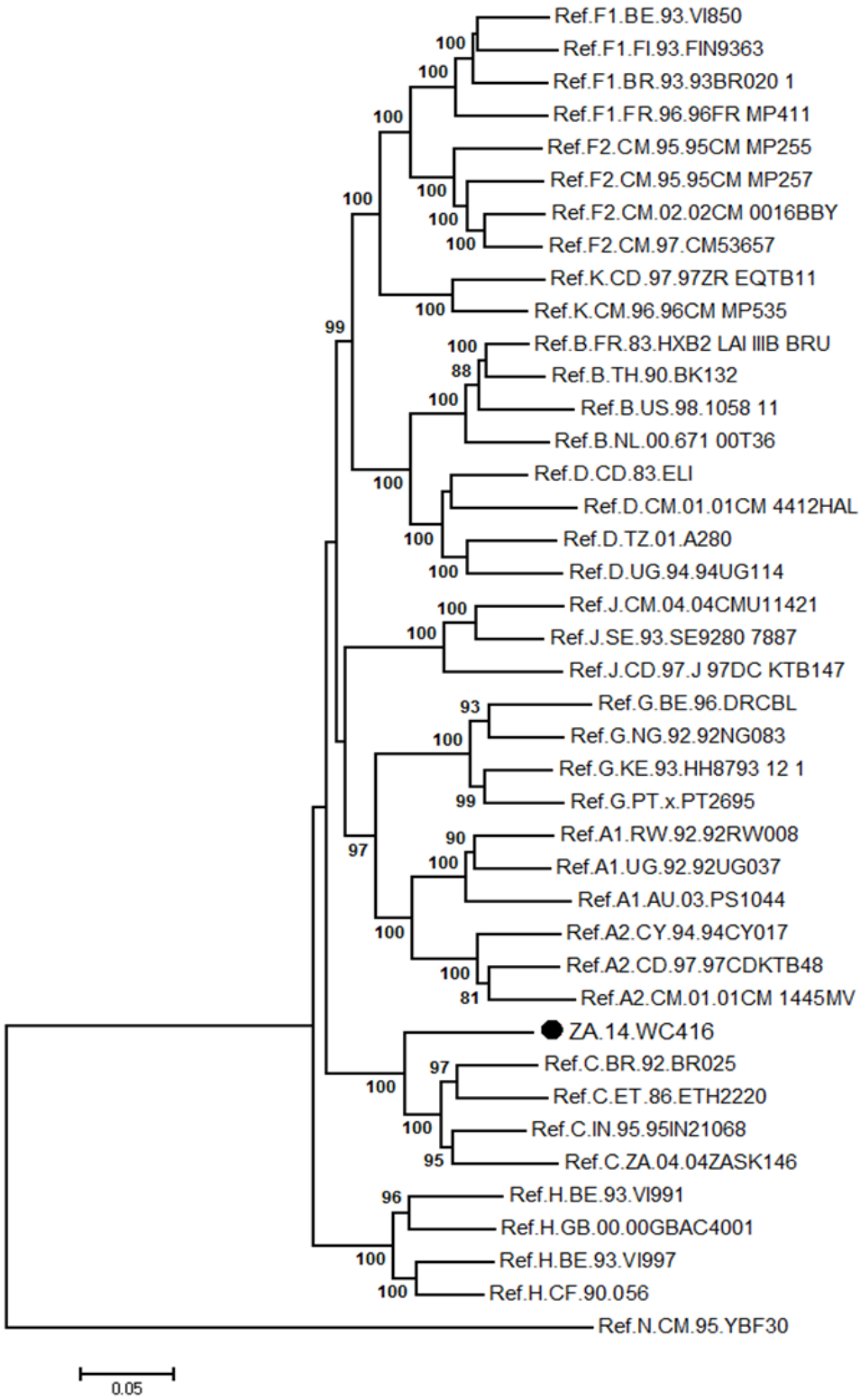

Figure 4.22: Phylogenetic tree of WC416 NFLG
The phylogenetic tree inferred for WC416 complete NFLG showed the sample to be an outlier to the subtype C reference sequences. The sample sequence is denoted as 14.ZA.WC416 on the tree. Bootstrap values greater than 70% are shown at the nodes. A horizontal scale of 0.05 was used for the branch length. HIV-1 group M NFLG reference sequences, obtained from the Los Alamos database, were used to infer the tree.

4.10.2. Sample EC148

In total, 8 trees were inferred for this sample. The jpHMM results for the NFLG (Figure 4.8B) and just the *env* region of the NFLG (Figure 4.10) were combined (Figure 4.23) to give a more accurate representation of all subtypes identified in the sample. Each region of the NFLG that identified as a distinct subtype were separated into individual fragments. Seven fragments were identified in this sample. The fragments numbered 1-7 correspond to their positions relative to the HIV genome (Figure 4.23).

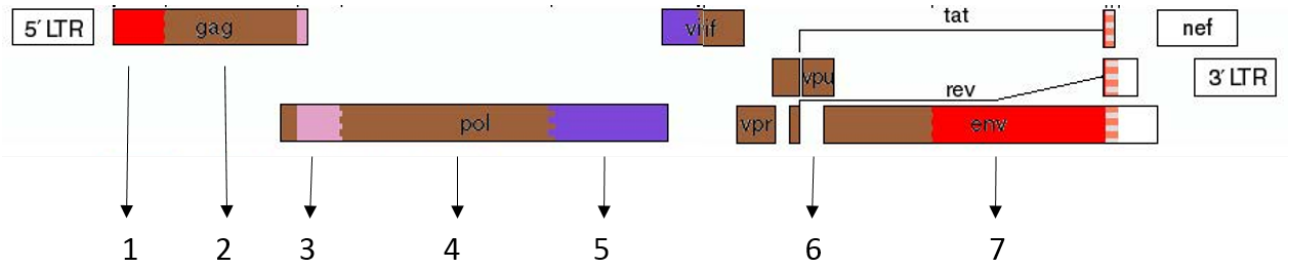

Figure 4.23 Correspondence of fragment number to position on HIV genome for EC148
Fragments 1 and 2 are in the *gag*; 3, 4 and 5 are in the *pol*; 6 covers half of the *vif,* the *vpr, vpu* and half the *env* and 7 covers the remainder of the *env.*

Fragment 1 was identified as subtype A by jpHMM. The sequences clustered most closely to the subtype A1 reference sequences in the phylogenetic tree (Figure 4.24). The bootstrap value for this cluster was over 70% and therefore can be accepted with confidence, thereby confirming the jpHMM result. Fragment 2 was identified as subtype C by jpHMM. The phylogenetic tree for fragment 2 showed that the fragment sequence clustered with the subtype C reference sequences (Figure 4.25) with a bootstrap value of 100% at the node of the cluster. The phylogenetic result confirmed the jpHMM result for this fragment. The phylogenetic tree for fragment 3 illustrated that the fragment sequence clustered closely with the subtype D reference sequences (Figure 4.26). The bootstrap value at the node closest to the fragment sequence is greater than 70% therefore confirming the jpHMM result. Fragment 4 was identified by jpHMM as subtype C. Fragment 4 sequences fell within the cluster of subtype C sequences on the phylogenetic tree, with bootstrap values greater than 70% (Figure 4.27). The jpHMM result and the phylogenetic tree result both identified this fragment as subtype C. Fragment 5 was identified as subtype K by jpHMM. The phylogenetic tree for this fragment showed that the fragment sequence was most closely related to the subtype C reference sequences (Figure 3.28). Although, the long branch length indicates that the query sequence is divergent from the subtype C reference sequences and there is no bootstrap value greater than 70% at that cluster. Fragment 6 was identified as subtype C by jpHMM. The phylogenetic tree for fragment 6 shows the fragment sequences clustering with the subtype C reference sequences, with greater than 70% bootstrap at the closest node (Figure 4.30). The jpHMM and phylogenetic tree results are consistent in assigning subtype C to this fragment.

Fragment 7 was identified by jpHMM as subtype A. The sequences for fragment 7 cluster with the subtype A1 reference sequences in the phylogenetic tree, with bootstrap values greater than 70% (Figure 4.29). Both the jpHMM and phylogenetic tree results identify fragment 7 as subtype A1.

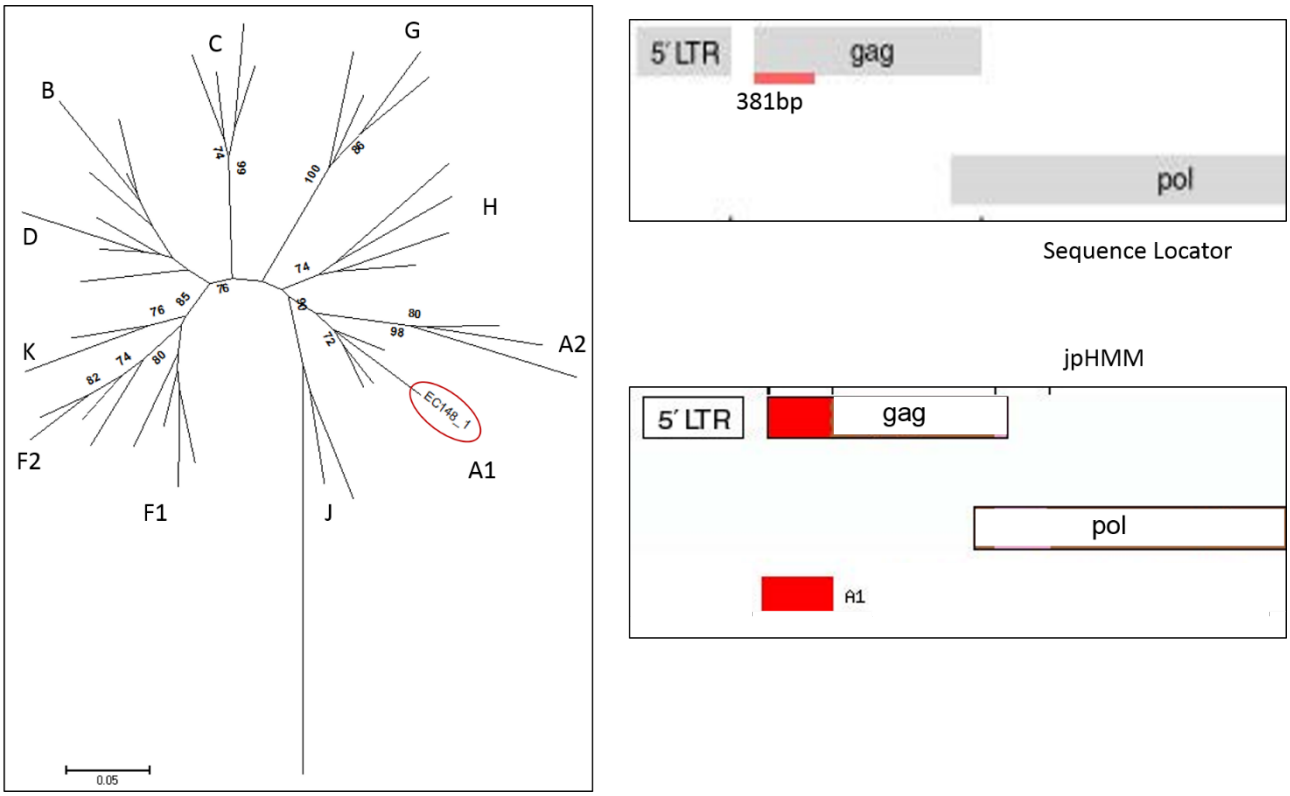

Figure 4.24: Phylogenetic analysis of EC148 fragment 1
The phylogenetic tree shows the fragment sequence to be clustering with the A1 reference sequence. The fragment sequence is circled in red. A 0.05 horizontal branch scale was used. The sequence locator results show that the fragment is 381bp and located in the *gag*; HXB2 790-1171. jpHMM identified the fragment as subtype A which is indicated in red.

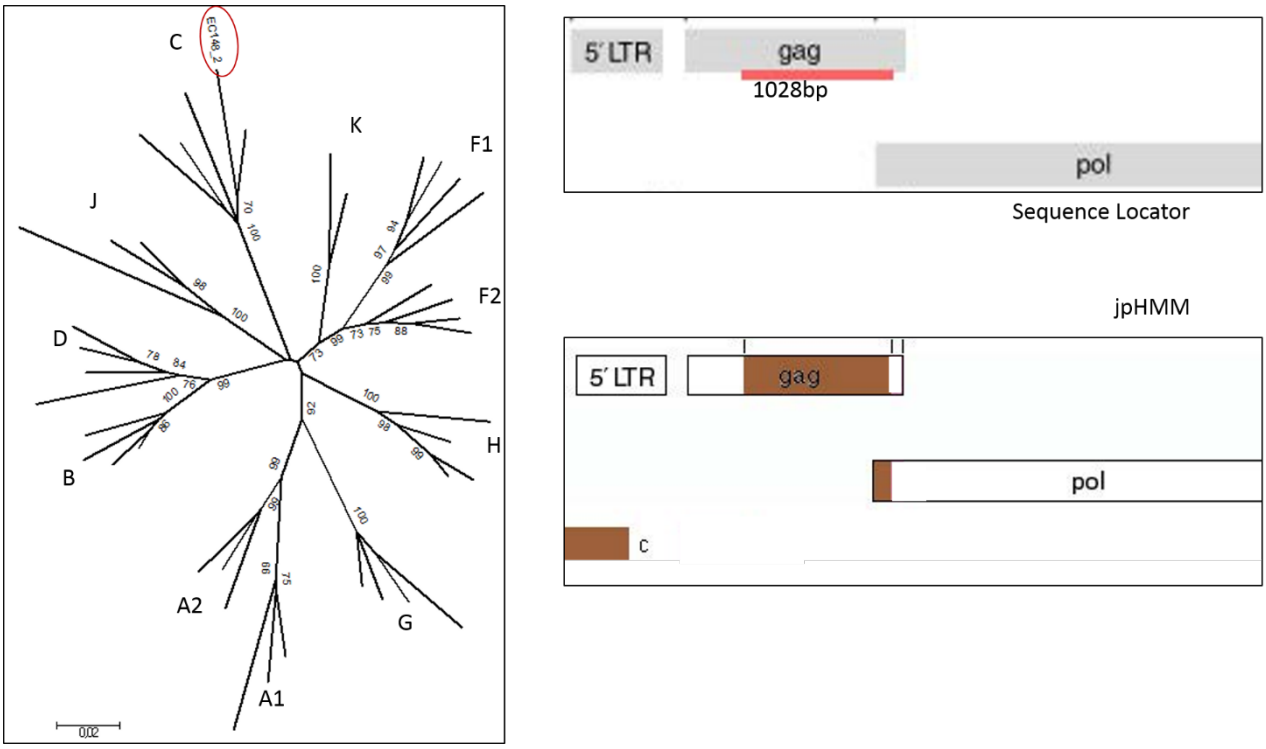

Figure 4.25: Phylogenetic analysis of EC148 fragment 2
The phylogenetic tree illustrates the fragment sequence, circled in red, clustering with the subtype C reference sequences. A 0.02 horizontal scale was used for branch length. The sequence locator results show that the 1028bp fragment is located in the *gag*; HXB2 1172-2200. jpHMM identified the fragment as subtype C.

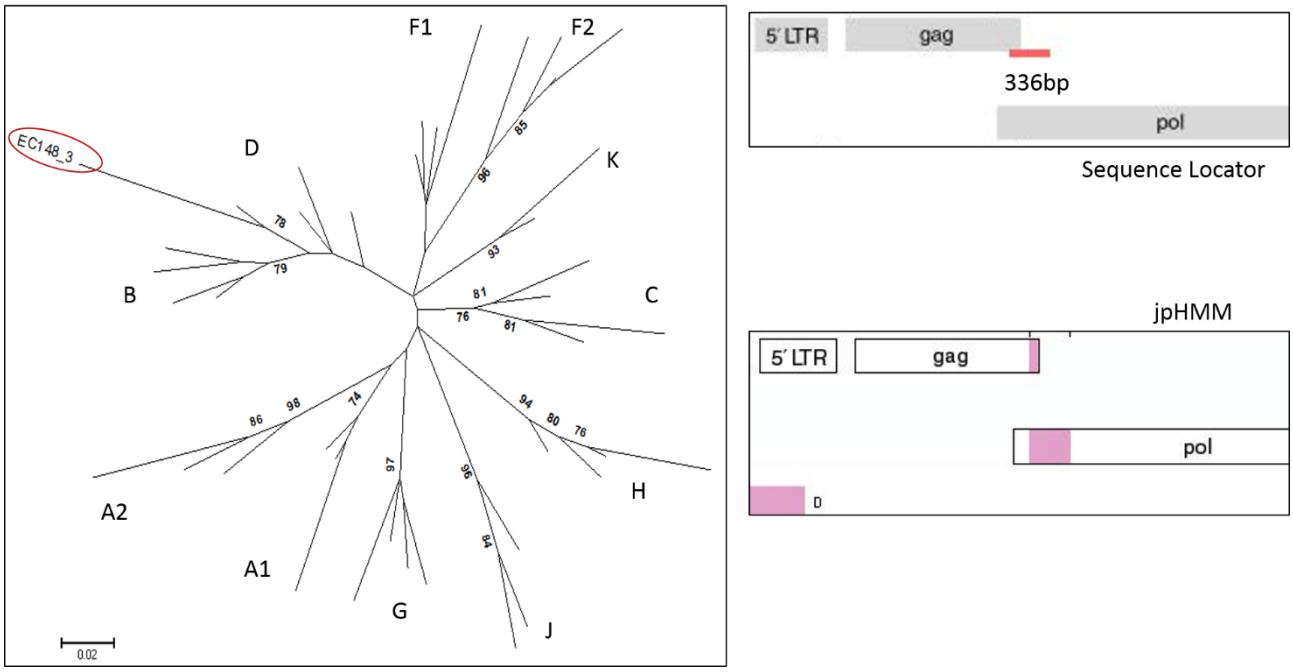

Figure 4.26: Phylogenetic analysis of EC148 fragment 3
The phylogenetic tree shows the fragment sequence (circled in red) to cluster with the subtype D reference sequences. A 0.02 horizontal scale was used for branch length. The sequence locator image shows that the fragment is 336bp and located in the *pol;* HXB2 2201-2528. jpHMM assigned the fragment as subtype D as seen by the pink.

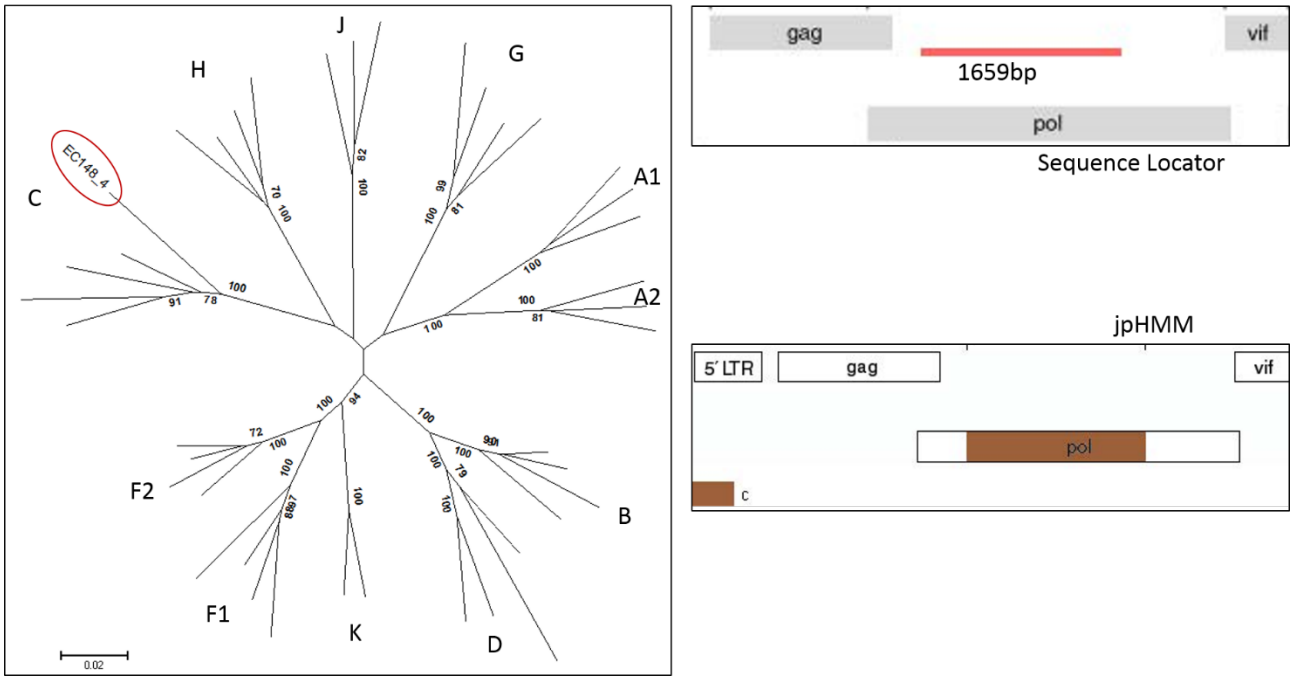

Figure 4.27: Phylogenetic analysis of EC148 fragment 4
The phylogenetic tree shows the fragment sequence, circled in red, to cluster with the subtype C reference sequences. A 0.02 horizontal scale was used for branch length. The sequence locator image shows that the fragment is 1659bp and is located in the *pol*; *HXB2* 2529-4188. jpHMM assigned the fragment as subtype C which is identified as brown.

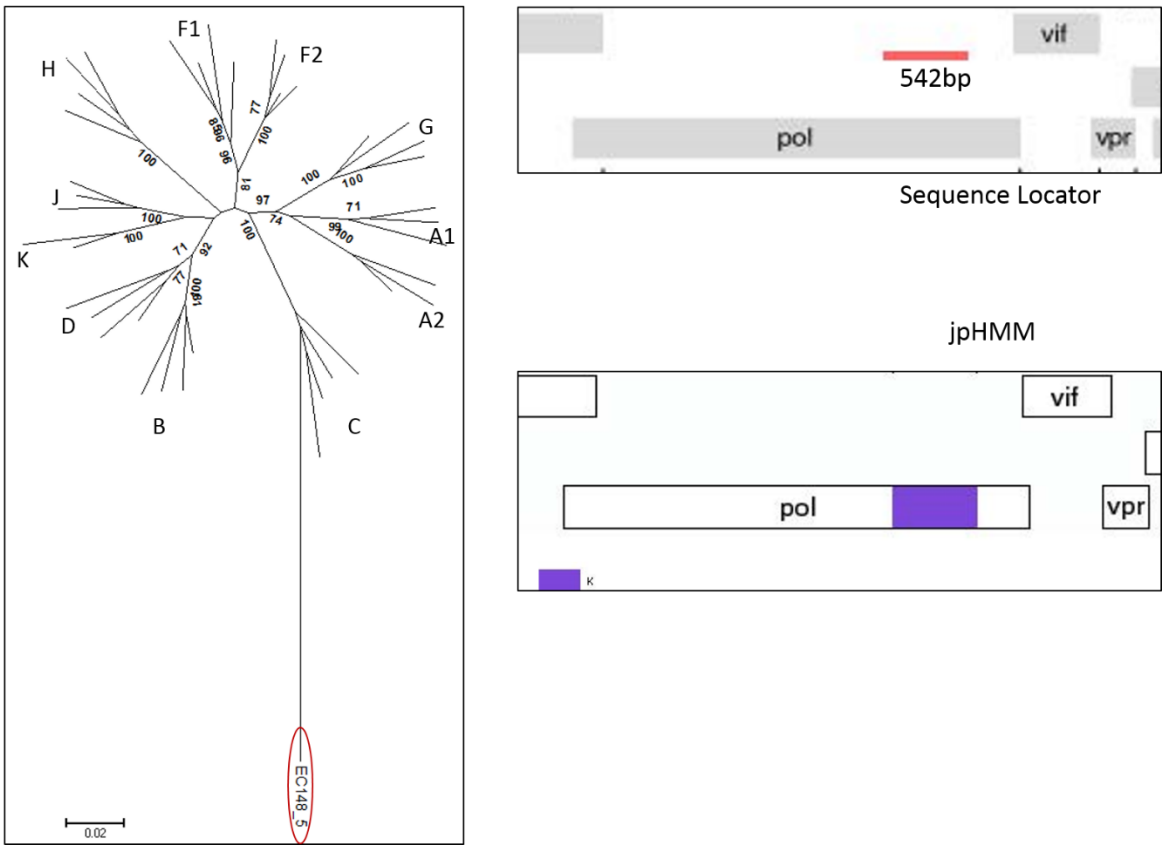

Figure 4.28: Phylogenetic analysis of EC148 fragment 5
The phylogenetic tree shows that the fragment clusters with subtype C reference sequences. The fragment sequence is circled in red. A 0.02 horizontal scale was used for branch length. Sequence locator shows that the 542bp fragment is located in the *pol;* HXB2 4189-4731. jpHMM assigned the fragment as subtype K as seen by the purple.

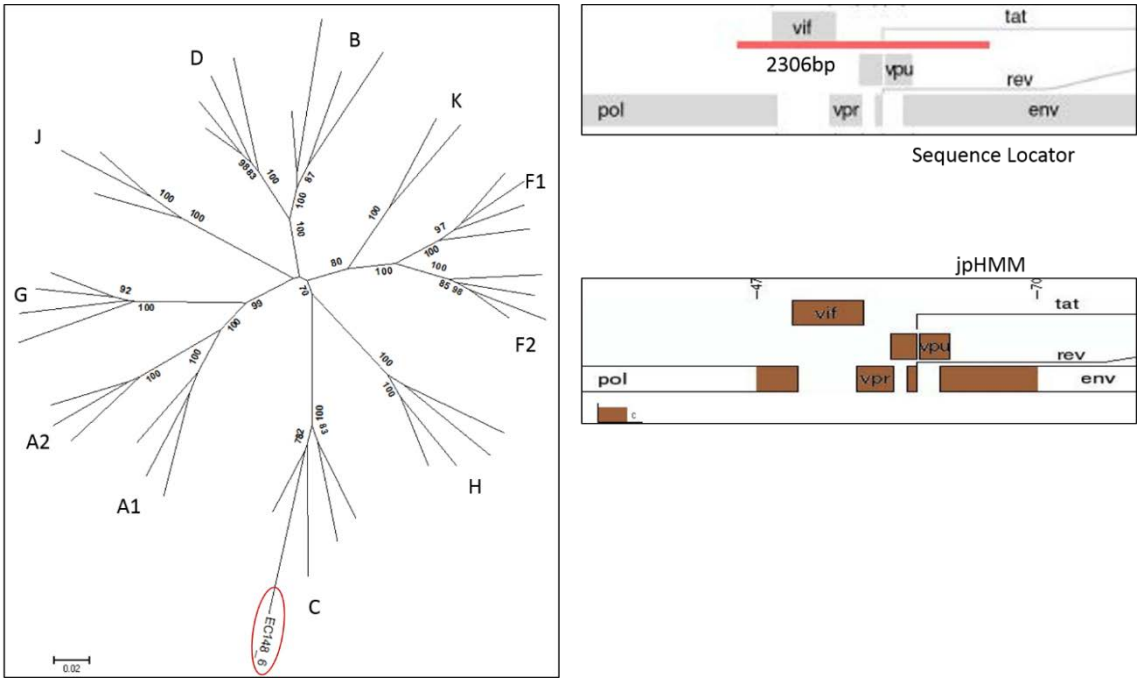

Figure 4.29: Phylogenetic analysis of EC148 fragment 6
The phylogenetic tree shows the fragment to cluster with the subtype C reference sequences. The fragment is circled in red. A 0.02 horizontal scale is used for branch length. The sequence locator result shows the fragment to be 2306bp and located across the *pol, vif, vpr, vpu* and *env;* HXB2 4732-7003. jpHMM identified the fragment as subtype C which is identified by brown.

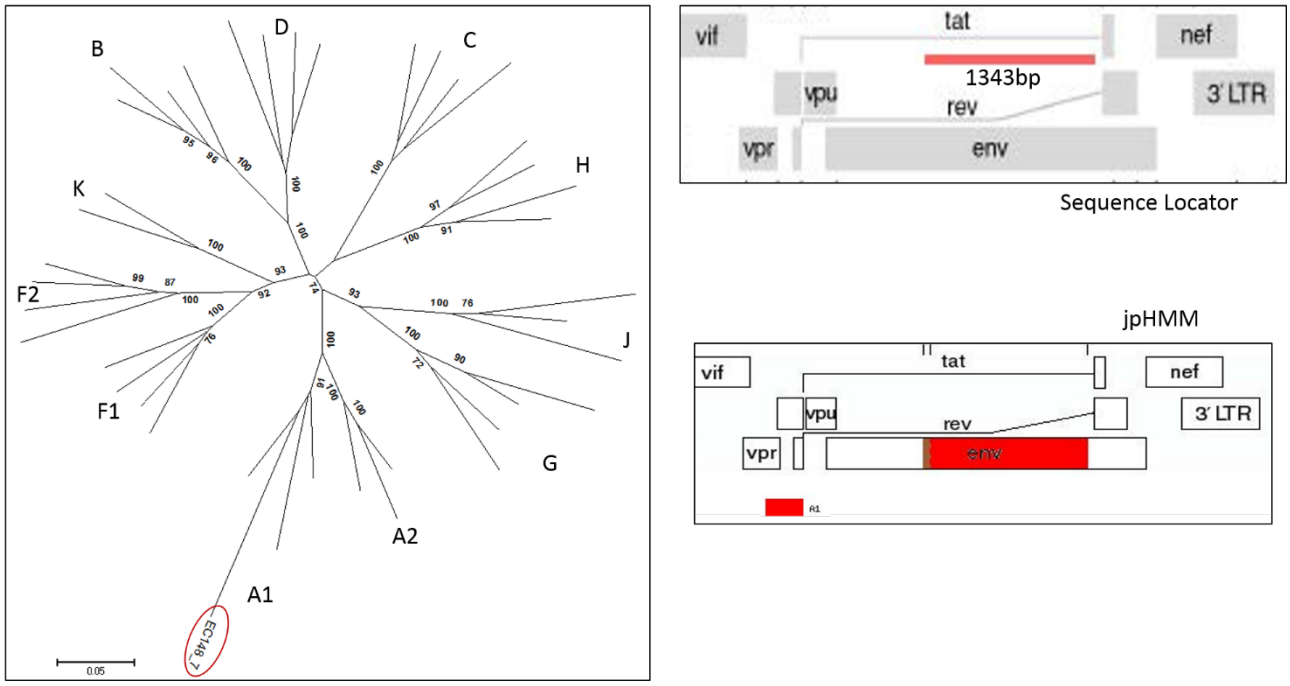

Figure 4.30: Phylogenetic analysis of EC148 fragment 7
The Phylogenetic tree shows that the fragment sequences cluster with the subtype A1 reference sequences. The fragment sequence is circles in red. A 0.05 horizontal scale was used for branch length. Sequence locator identified the 1343bp fragment to be located in the *env; HXB2* 7005-8327. jpHMM assigned the fragment as subtype A as is seen by the red.

The sample NFLG sequence begins in the *gag* and ends in the gp41 region of the *env*. The NFLG phylogenetic tree (Figure 4.31) shows the sample is an outlier to the subtype C reference sequences, with a 100% bootstrap value at the closest node. Bootstrap values greater than 70% are displayed at the nodes on the tree. The high bootstrap values on the tree indicate that the results can be accepted with confidence.

Phylogenetic analysis confirmed the jpHMM results for all fragments expect for fragment 5. jpHMM identified this fragment as subtype K whereas the phylogenetic tree shows it to be most closely related to subtype C reference sequences. The phylogenetic analysis for each fragment and for the NFLG therefore confirms that the sample is a complex A, C, D recombinant.

A phylogenetic tree was inferred with sample EC148 and WC416 NFLGs (Figure 4.32). The tree illustrates that both sample sequences cluster with each other in a separate cluster that is an outlier to the subtype C reference sequences in the tree. There is a 100% bootstrap value at the node of the cluster with the two sample sequences. There is also a 100% bootstrap value at the node joining this cluster to the subtype C reference sequences. Although sample EC148 and WC416 NFLG sequences cluster together, the longer branch length of EC148 indicate that they are different viruses.

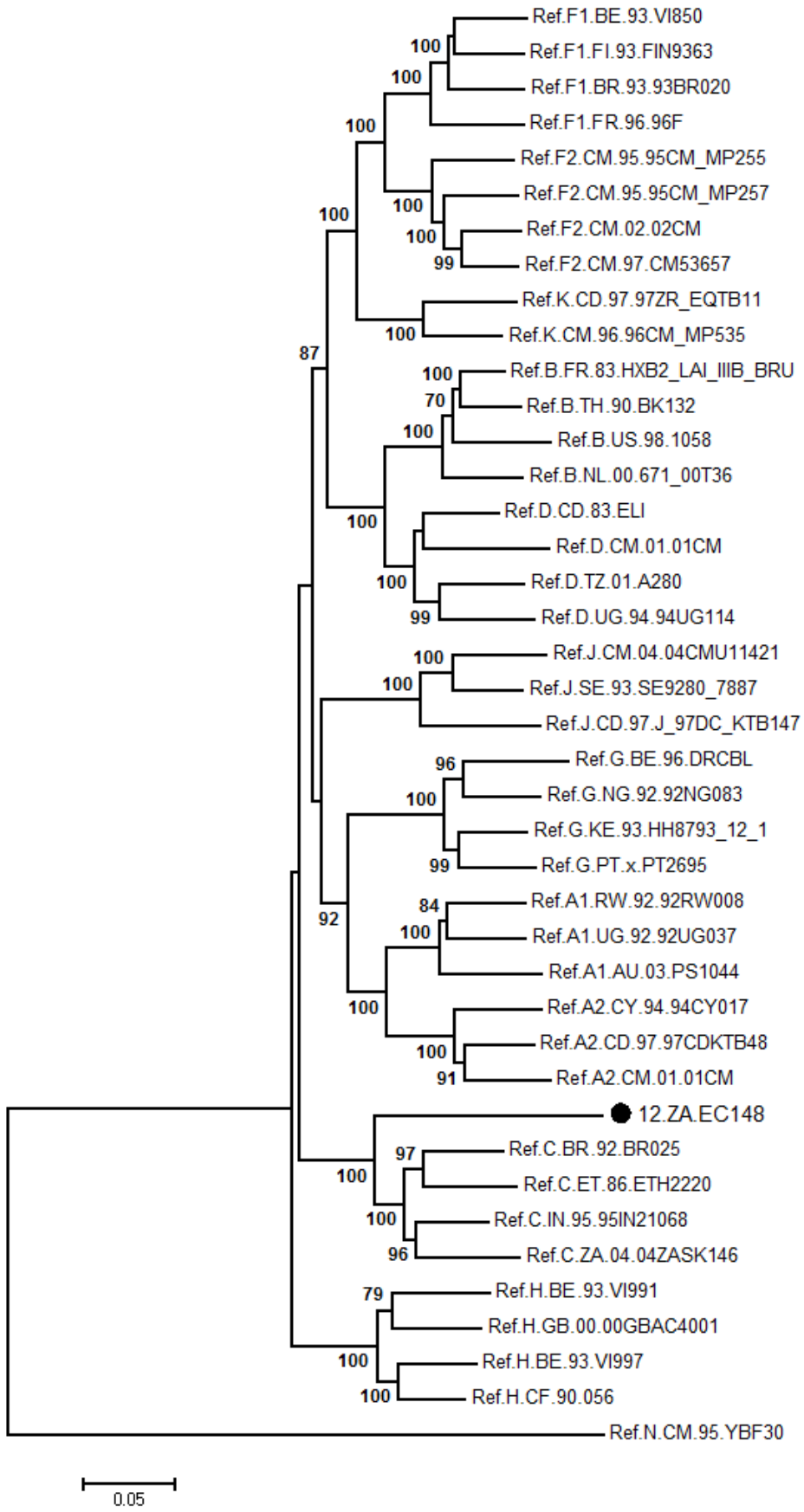

Figure 4.31: Phylogenetic tree of EC148 NFLG
The phylogenetic tree shows the sample sequence to be an outlier to the subtype C reference sequences. The sample sequence is denoted as 12.ZA.EC.148. The numbers on the tree indicate bootstrap values above 70%. A 0.05 horizontal scale was used for branch length.

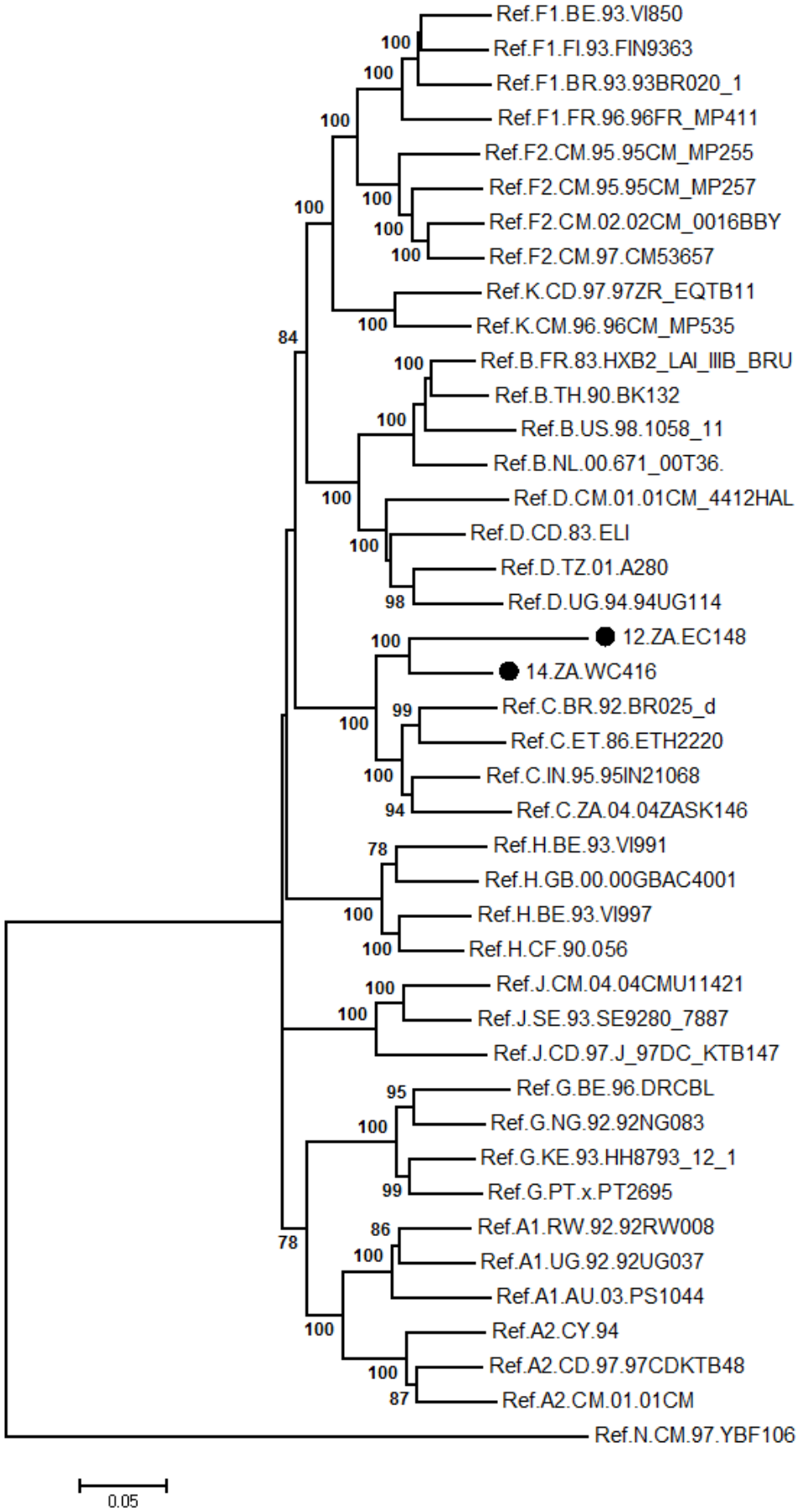

Figure 4.32: Phylogenetic tree of sample EC148 and WC416 NFLGs
Both sample sequences form an independent cluster that is an outlier to the subtype C reference sequences. Branch lengths greater than 70% are indicated on the tree. A branch scale of 0.05 was used. HIV-1 group M NFLG reference sequences, obtained from the Los Alamos database, were used to infer the tree.

Chapter 5

Discussion

5.1. HIV-1 in South Africa

5.1.1. Diversity in South Africa

South Africa has the highest number of individuals infected with HIV worldwide. By the end of 2017 this number was at 7.2 million individuals (UNAIDS, 2017). Based on sequences in the LANL database, subtype C accounts for 97.8% of the epidemic in South Africa, leaving 2.2% of the epidemic as non-subtype C (www.hivlanl.gov; Accessed 27/09/2018). Although 2.2% may appear as slightly insignificant in comparison to 97.8%, 2.2% of the 7.2 million infected individuals equates to 158 400 individuals. This figure of individuals infected by non-subtype C viruses in South Africa is larger than the total population of HIV infected individuals in countries such as Argentina and Mali, where the number of people living with HIV are 120 000 and 130 000, respectively (UNAIDS, 2017). However, these numbers only represent sequences from which the sampling country is known and is not an accurate representation of the true statistics. This study focused on characterisation of non-subtype C viruses circulating in South Africa.

The co-circulation of subtype C and non-subtype C viruses in South Africa indicate the complexity of the epidemic. The characterisation of the two non-subtype C viruses in this study will contribute to understanding the epidemic in South Africa. The characteristics of different HIV-1 strains and their interaction with the human host may play a role in HIV transmission and disease progression. For example, strains that use the CCR5 chemokine receptor are transmitted more frequently than those that use the CXCR4 receptor (Taylor *et al.*, 2008). HIV-1 diversity also has implications in molecular diagnostic assays, ART and vaccine development (Peeters, 2001). A standard diagnostic assay, ART regime or vaccine would not be suitable for an epidemic of high diversity (Lau and Wong, 2013).

5.1.2. Implications of HIV-1 diversity

Our cohort of possible C, D recombinant viruses in South Africa (Table 4.1) indicate that the epidemic is diverse. With over 6000 non-subtype C viral sequences obtained by the NHLS in a period of 7 years (2008-2015), it is evident that large numbers of non-subtype C viruses are circulating in the country (unpublished data; Prof S. Engelbrecht – personal communication). HIV diversity can have an impact on diagnosis and the ability to accurately determine viral load measurements (Bourlet *et al*., 2011; Church *et al*., 2011). Viral diversity may also play a role in the emergence of drug resistance by affecting the response to ART (Tatem *et al*., 2012). Studies have shown that subtypes may differ in their rate of progression to disease as well as in the rate of transmission of the virus (Kiwanuka *et al*., 2008, 2009). An example of this is seen with the K65R mutation that confers drug resistance mutation in HIV-1 subtype C viruses (Doualla-Bell *et al*., 2006). A study by Hemelaar, 2013 identified numerous aspects of HIV infection that are affected by HIV genetic variability. These aspects include vaccine development, transmission routes, disease

progression, drug resistance, response to ART, viral load measurement, diagnosis, pathogenesis and immune response and escape (Figure 5.1) (Hemelaar, 2013). Monitoring the diversity within the country is therefore crucial as it assists in important aspects of HIV infection.

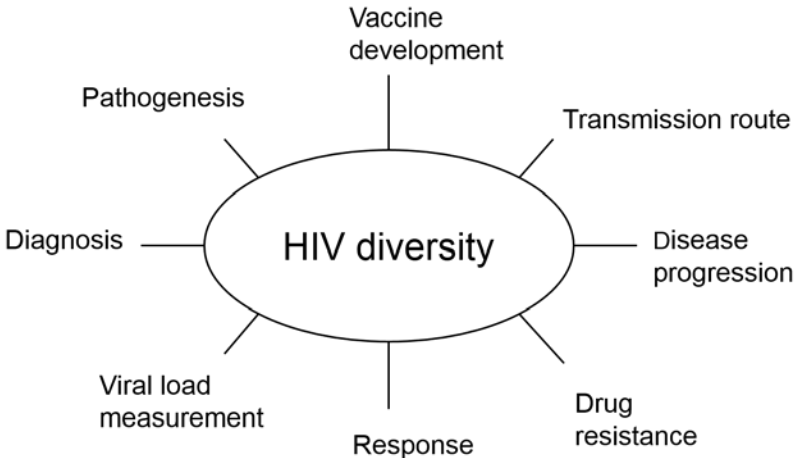

Figure 5.1: Aspects of HIV infection affected by HIV diversity (adapted from Hemelaar, 2013)
Diagram depicting the different aspects found to be associated with HIV infection and HIV diversity. HIV diversity plays a role in multiple different aspects of HIV research.

5.1.3. Characterisation of HIV-1 non-subtype C viruses in South Africa

To date, to the best of our knowledge and based on available sequences in the LANL database, 22 non-subtype C NFLG virus sequences have been characterised in South Africa, 14 are pure subtypes and 8 are recombinant forms. The pure subtypes are comprised of 3 subtype A1 viruses (Rousseau *et al.*, 2006; Wilkinson and Engelbrecht, 2009; Wilkinson *et al.*, 2015b); 5 subtype B viruses (Rousseau *et al.*, 2006; Wilkinson and Engelbrecht, 2009; Wilkinson *et al.,* 2015b); 5 subtype D (Loxton *et al.*, 2005; Jacobs *et al.*, 2007) and 1 subtype G virus (Wilkinson *et al*., 2015b). The recombinant viruses comprise of 5 A, C recombinants (Papathanasopoulos *et al.*, 2002; Rousseau *et al.*, 2006; Iweriebor *et al.*, 2011; Wilkinson *et al.*, 2015b); 2 A, D recombinants (Wilkinson and Engelbrecht, 2009; Wilkinson *et al.*, 2015b) and 1 complex recombinant (Papathanasopoulos *et al.*, 2002).

Characterisation of the two samples from this study will increase the number of non subtype C NFLGs in South Africa from 22 to 24. The two study samples were selected based on the sample availability and sample volume. Numerous samples from the C, D cohort were used to attempt to amplify the NFLG, however they did not yield positive results. only two of the samples, EC148 and WC416, were available, of sufficient volume and yielded positive results from the NFLG PCR amplification. Therefore only these two samples could be used for further characterisation. The two sequences from the viruses characterised in this study each identify as a complex A, C, D URF. Although, subtypes A, C and D are identified in both samples, the breakpoints between the subtypes differ in each. In WC416, BLAST results showed the subtype A segment to be most similar to an A, C recombinant from Malawi in 2009 and the subtype D segment showed highest

similarity to an A, C, D ,G complex recombinant from Tanzania in 2009. The subtype A segment of EC148 is most similar to an A, C recombinant from Tanzania in 1997 and the subtype D segment was most similar to a subtype D virus from Uganda in 2005. These results reveal that EC148 and WC416 viruses are not the same.

Subtype A is dominant in Central African Republic and is responsible for 73.4% of the epidemic. Subtype A is also highly prevalent in the former Union of Soviet Socialist Republics (USSR) region, especially in Kazakstan and Kyrgyzstan, where it dominates the epidemics by more than 90%. Subtype D is most prevalent in Central African countries such as Uganda, Sudan and Chad where subtype D dominates their epidemics by 59.1%, 56.4% and 41.7%, respectively. Subtype C is highly prevalent in South Africa, accounting for 97.8% of South African HIV-1 sequences in the LANL database. These statistics are based only on sequences available in the HIV LANL database (https://www.hiv.lanl.gov; Accessed 20/09/18). Subtype D, along with subtype B, was also once a predominant subtype in South Africa during the first epidemic in the 1980s. With subtype A and D being predominant in Central African counties, it is possible that these subtypes come to South Africa via refugees that are driven out of their native countries due to political unrest (Wilkinson and Engelbrecht, 2009). World travel also influences HIV diversity and could be responsible for the transfer of subtypes into South Africa from other parts of the world (Perrin, 2003). Given that subtype A, C and D have been identified as pure subtypes and as recombinant forms (A, C and A, D) circulating in South Africa, the formation of an A, C, D complex recombinant is indeed plausible.

5.2. Importance of HIV-1 recombinant forms

5.2.1. Significance

No less than 20% of global HIV-1 isolates sequenced are recombinant forms (Hemelaar *et al.*, 2006). By the end of October 2018, 97 recombinant forms have been identified in the LANLdatabase (https://www.hiv.lanl.gov/content/sequence/HIV/CRFs/CRFs.html; Accessed 27/10/2018). The most prevalent recombinant is CRF01_AE, which accounts for 5.9% of all HIV-1 sequences available in the LANL database. CRF01_AE is prevalent in South East Asia and is the predominant subtype in countries including Thailand, Cambodia and Vietnam, having displaced the previous existing HIV-1 subtype. Another common recombinant is CRF02_AG (Haynes *et al*., 1995, Hemelaar, 2012), which is responsible for over 9 million infections globally (Haynes *et al.*, 1995). This particular recombinant virus is the predominant strain in West and West Central Africa, particularly in Libya where 100% (n =148) of the HIV-1 sequences in LANL from that country are CRF02_AG. The emergence of recombinant forms has been on the increase since the discovery of the first CRF in 1992 (Figure 5.2). HIV-1 recombinant viruses are emerging at a high frequency as a results of co/super infection of individuals and due to the co-circulation of multiple subtypes in nearly all geographical regions. Africa, Asia and South America have been identified as the three geographical regions with the highest prevalence of CRFs, globally (Lau and Wong, 2013).

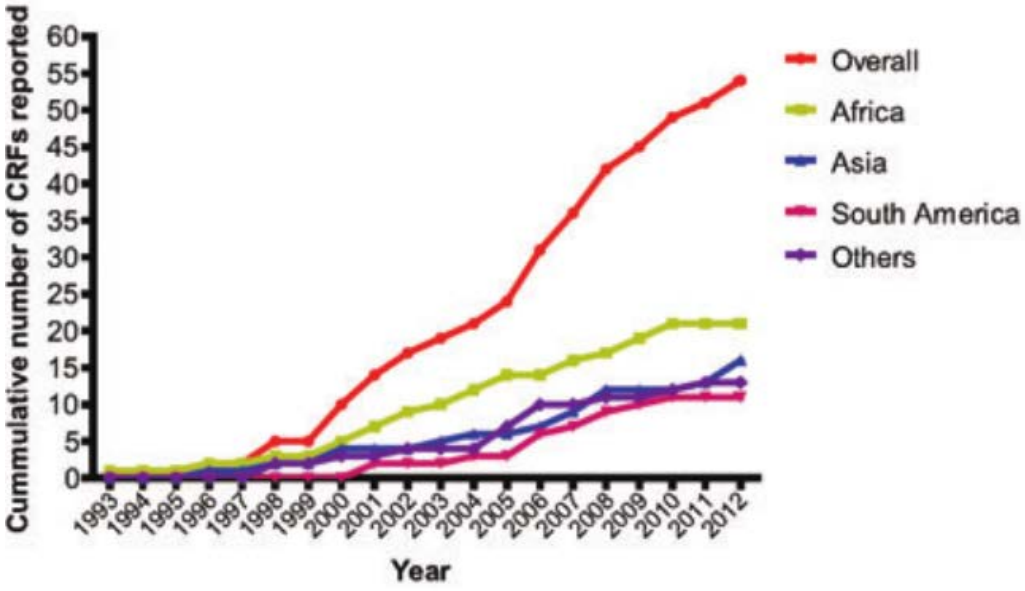

Figure 5.2: Increased emergence of CRFs (Lau and Wong, 2013)
The y axis of the graph represents the number of reported CRFs and the x axis represents the years from 1993-2012. The red line represents the global number, green represents Africa, blue represents Asia, pink represents South America and purple represents other regions.

5.2.2. Implications of HIV-1 recombination

Both samples used in this study identified as complex HIV-1 recombinant forms. Recombination of different HIV-1 subtypes is believed to have an impact on viral diversity and fitness, drug resistance, disease progression and escape from the immune system. Recombination is often observed in the evolution of HIV in infected patients. Such recombination events are able to confer new combinations of phenotypic traits that are frequently advantageous for the virus. These advantageous traits have the potential to contribute to the increase in fitness of the virus (Brown *et al.*, 2011). Recombination of HIV-1 within an infected patient has shown to contribute to the viral diversity of viral quasispecies by generating novel genotypes in HIV-1 genes (Charpentier *et al.*, 2006). Genetic recombination of divergent HIV-1 strains in a co-infected individual can lead to the emergence of drug resistant mutant viruses. When a virus that confers resistance to drug A combines with a virus that confers resistance to drug B, a mutant virus is generated that confers resistance to both drug A and B (Moutouh *et al.*, 1996). Recombination also plays a role in the alteration of HIV-1 to changing selective pressure such as from ART. This can be done by producing viruses with new combinations of resistance mutations, compared to the parent viruses or by increasing diversity in the genomic regions that are not exposed to such selective pressure. For example while diversity decreases in the *pol* due to selective pressure, there is an increase in diversity in the *gag* and *env* (Nora *et al.*, 2007). This highlights the importance of complete genome/NFLG characterisation to detect the true diversity in a sample sequence. Recombination is also associated with immunological escape. In an instance where superinfection occurs in an

infected individual, the superinfecting strain is able to evade the existing immune control possibly by recombination with the initial strain in critical regions under immune selection pressure (Streeck *et al.*, 2008). Recombination between multiple viruses show a greater likelihood of escaping primary CD8+ T-cells as compared to single virus infections (Ritchie *et al.*, 2014). Recombination of divergent HIV-1 strains is sometimes associated with increased viral fitness leading to the ability to evade the human immune cells and therefore possibly accelerate the progression to AIDS (Liu *et al.*, 2002). A study by Kiwanuka *et al.,* 2008 shows that infection with an A, D recombinant lead to a more rapid progression to AIDS than infection with a pure subtype A virus. CRF19_cpx (Subtypes D, A and G) in the Cuban population is also associated with faster progression to AIDS and high pathogenicity (Kouri *et al., 2015)*.

5.3. Near full-length genome characterisation

Characterisation of partial genes or small fragments of the HIV genome to subtype an isolate has been widely used in the past (Jacobs *et al*., 2009; Grossmann *et al*, 2015; Neogi *et al.*, 2012). Such characterisation may only give an indication of the possible subtype of the isolate. Partial genome sequencing is especially not useful in identifying recombinants as a recombinant segment can be at any region within the genome.

The samples in the non-subtype C cohort, from which the study samples were obtained, were all initially only sequenced in the partial *pol* region of the genome (~1.4kb) and identified as possible C, D recombinants. using only the *pol* sequences of a virus to identify recombination will lead to an inaccurate representation of the true recombinant viruses circulating in the country. NFLG sequencing done in this study identified the samples instead as complex A, C, D recombinants, highlighting the importance of NFLG characterisation for recombinant identification. NFLG characterisation allows one to make a more accurate assumption of the subtype and recombination pattern of an isolate (Wilkinson and Engelbrecht, 2009). It allows for the detection of all possible subtypes within a sample even if it is only in a small region. NFLG must be used in order to identify possible recombinants circulating in the country.

5.4. Study strengths and limitations

In this study we were able to characterise two complex A, C, D unique recombinants. The PCR protocols were optimised to work for DNA and viral RNA samples. We were able to obtain NFLG sequences of over 7000bp for each sample using in house primers as well as designed primers for regions that proved difficult to sequence. All protocols displayed in this study can be repeated and can therefore contribute to an increase in the number of NFLG sequences in South Africa.

A limitation to this study was the discrepancies between the online subtyping programs. For sample EC148, jpHMM identified a substantial region of the genome as subtype K whereas RIP and REGA assigned the same region as subtype C. An additional limitation was conferring

phylogenetic trees for small regions that identified as distinct subtypes. The phylogenetic trees results for these regions didn't have bootstrap values >70% as they were less than 500bp in length.

5.5. Conclusion

Two unique HIV-1 NFLG complex A, C, D recombinants have been identified in this study. It is important to monitor the diversity of HIV-1 in the country as the emergence of new recombinants indicates that the epidemic is complex and constantly evolving. NFLG characterisation of HIV-1 allows for a more accurate representation of recombinant forms and increases our knowledge on circulating viruses in the country.

Reference List

Bacon D and Anderson W. (1986). Multiple sequence alignment. *Journal of Molecular Biology*, 191(2), pp.153-161.

Baldauf S. (2003). Phylogeny for the faint of heart: a tutorial. *Trends in Genetics*, 19 (6), pp. 345-351.

Baum D. (2008). Reading a Phylogenetic Tree: The meaning of monophyletic groups. *Nature Education,* 1(1), pp.190.

Baltimore D. (1970) Viral RNA-dependent DNA polymerase: RNA-dependent DNA polymerase in virions of RNA tumour viruses. *Nature*, pp. 1209–1211.

Barré-Sinoussi F, Chermann J, Rey F. (1983). Isolation of a T-lymphotropic Retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science*, 220(4599), pp. 868–871.

Becker M, De Jager G and Becker W. (1995). Analysis of Partial *gag* and *env* gene sequences of HIV type 1 strains from Southern Africa. *AIDS Research and Human Retroviruses*, 11(10), pp.1265-1267.

Bourlet T, Signori-Schmuck A, Roche L *et al.,* (2010). HIV-1 load comparison using four commercial real-time assays. *Journal of Clinical Microbiology*, 49(1), pp.292-297.

Briggs J, Wilk T, Welker R, Krausslich H and Fuller D. (2003). Structural organization of authentic, mature HIV-1 virions and cores. *EMBO Journal*, 22(7), pp. 1707–1715.

Brown R, Peters P, Caron C *et al.,* (2011). Intercompartmental recombination of HIV-1 contributes to *env* intrahost diversity and modulates viral tropism and sensitivity to entry inhibitors. *Journal of Virology*, 85(12), pp.6024-6037.

Buonaguro L, Tornesello M and Buonaguro F. (2007) MINIREVIEW Human Immunodeficiency Virus Type 1 subtype distribution in the worldwide epidemic : pathogenetic and therapeutic implications. *Journal of Virology*, 81(19), pp. 10209–10219.

Burke D. (1997). Recombination in HIV: an important viral evolutionary strategy. *Emerging Infectious Diseases*, 3(3), pp.253-259.

Carr J, Salmien M, Koch C. *et al.* (1996). Full-length sequence and mosaic structure of a human immunodeficiency virus type 1 isolate from Thailand. *Journal of virology*, 70(9), pp. 5935–43.

Centers for Disease Control & Prevention (CDC) (1981) Kaposi 's Sarcoma and Pneumocystis Pneumonia among homosexual men — New York City and California Source : Morbidity and Mortality Weekly Report , 30(25), pp. 305-308.

Centers for Disease Control & Prevetion (CDC) (1982). *Current Trends Update on Acquired Immune Deficiency Syndrome (AIDS)*, 31(37), pp. 513–514.

Chakrabarti L, Guyader M, Alizon M. *et al.,* (1987). Sequence of simian immunodeficiency virus

from macaque and its relationship to other human and simian retroviruses. *Nature*, 328, pp. 543–547.

Charpentier C, Nora T, Tenallion O, Clavel F, Hance A. (2006). Extensive recombination among human immunodeficiency virus type 1 quasispecies makes an important contribution to viral diversity in individual patients. *Journal of Virology*, 80(5), pp. 2472–2482.

Church D, Gregson D, Lloyd T. *et al.* (2011). Comparison of the realtime HIV-1, COBAS TaqMan 48 v1.0, easy Q v1.2, and Versant v3.0 assays for determination of HIV-1 viral loads in a cohort of Canadian patients with diverse HIV subtype infections. *Journal of Clinical Microbiology*, 49(1), pp. 118–124.

Clavel F, Guetard D, Brun-Vezinet F. *et al*., (1986). Isolation of a new human retrovirus from West African patients with AIDS. *Science*, 233(4761), pp.343-346.

Coffin J, Haase A, Levy J. *et al.,* (1986). Human immunodeficiency virus. *Science*, 232, p. 697.

Costin J. (2007). Cytopathic mechanisms of HIV-1. *Virology Journal*, 4, pp. 1–23.

Desjardins P and Conklin D. (2011). Microvolume quantitation of nucleic acids. *Current Protocols in Molecular Biology*, pp. 1–4.

de Oliveira T, Deforche K, Cassol S. *et al.,* (2005). An automated genotyping system for analysis of HIV-1 and other microbial sequences. *Bioinformatics Applications Note,* 21, pp.3797–3800.

Domingo E, Sheldon J and Perales C. (2012). Viral Quasispecies Evolution. *Microbiology and Molecular Biology Reviews*, 76(2), pp. 159–216.

Doualla-Bell F, Avalos A, Brenner B. *et al.,* (2006). High prevalence of the K65R mutation in human immunodeficiency virus type 1 subtype C isolates from infected patients in Botswana treated with didanosine-based regimens. *Antimicrobial Agents and Chemotherapy*, 50(12), pp. 4182–4185.

Duri K, Stray-Pedersen B and Muller F. (2013). HIV diversity and classification, role in transmission. *Advances in infectious diseases*, 3, pp. 146–156.

Ellrodt A, Le Bras P, Palazzo. *et al.,*(1984). Isolation of human t-lymphotropic retrovirus (lav) from Zairian married couple, one with aids, one with prodromes. *The Lancet*, 323(8391), pp.1383-1385.

Engelbrecht S, Laten J, Smith T and van Rensburg E. (1995). Identification of env Subtypes in fourteen HIV Type 1 isolates from South Africa. *AIDS Research and Human Retroviruses*, 11(10), pp.1269-1271.

Essex M. and Kanki P. (1988). The Origins of the AIDS Virus. *Scientific American*, 259(4), pp. 64–71.

Etienne L, Delaporte E and Peeters M. (2011). Origin and Emergence of HIV/AIDS. *Genetics and Evolution of Infectious Diseases*. pp 689-710.

Fang G, Weiser B, Visosky A, Townsend L and Burger H. (1996). Molecular cloning of full-length HIV-1 genomes directly from plasma viral RNA. *Journal of Acquired Immune Deficiency*

Syndromes and Human Retrovirology, 12(4), pp. 352–357.

Felsenstein J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17(6), pp. 368–376.

Finzi D, Hermankova M, Pierson T. *et al.,* (1997). Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy. *Science*, 278(5341), pp. 1295–1300.

Gallo R, Salahuddin S, Popovic M. *et al.* (1984). Frequent detection and isoltaion of cytopathic retroviruses (HTLV-III) from patients with AIDS and risk for AIDS. *Science*, 224(4648), pp. 500–503.

Gallo R. (1986). The First Human Retrovirus. *Scientific American*, 255(6), pp. 88–101.

Gallo R. (2005). The discovery of the first human retrovirus: HTLV-1 and HTLV-2. *Retrovirology*, 2, pp. 1–8.

Gao F, Robertson D, Morrison S. *et al.,* (1996). The heterosexual human immunodeficiency virus type 1 epidemic in Thailand is caused by an intersubtype (A/E) recombinant of African origin., *Journal of virology.* 70(10), pp. 7013–7029.

Gao F, Robertson D, Carruthers C. *et al.,* (1998). A comprehensive panel of near-full-length clones and reference sequences for non-subtype B isolates of Human Immunodeficiency Virus Type 1. *Journal of Virology*, 72(7)pp. 5680–5698.

Gao F, Bailes E, Robertson D. *et al.,* (1999). Origin of HIV-1 in the chimpanzee *Pan troglodytes troglodytes. Nature*, 397(6718), pp. 436–441.

Gross R, Fouxon I, Lancet D, Markovitch O. (2014). Quasispecies in population of compositional assemblies. *BMC Evolutionary Biology*, 14(1), pp. 1–11.

Grossmann S, Nowak P and Neogi U. (2015). Subtype-independent near full-length HIV-1 genome sequencing and assembly to be used in large molecular epidemiological studies and clinical management. *Journal of the International AIDS Society*, 18(1), p.20035

Hahn B. (2000). AIDS as a Zoonosis: Scientific and Public Health Implications. *Science*, 287(5453), pp.607-614.

Hall B. (2011). Phylogenetic trees made easy. Sunderland, Massachusetts: Sinauer.

Haynes B, Moody M, Heinley C, Korber B, Millard W and Scearce R. (1995). Sequence analysis of the glycoprotein 120 coding region of a new HIV type 1 subtype A strain (HIV-1IbNg) from Nigeria. *AIDS Research and Human Retroviruses*, 10(2), pp. 1755–1757.

Hayward A. (2017). Origin of the retroviruses: when, where, and how?. *Current Opinion in Virology*. 25, pp. 23–27.

Hasegawa A, Tsujimoto H, Maki N *et al.,* (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*; 22(2), pp. 160-174

Hemelaar J, Gouws E and Ghys P. (2006). Global and regional distribution of HIV-1 genetic subtypes and recombinants in 2004. *AIDS*, 20, pp. 13–23.

Hemelaar J, Gouws E, Ghys P. and Osmanov S. (2011). Global trends in molecular epidemiology of HIV-1 during 2000–2007. *AIDS*, 25(5), pp.679-689.

Hemelaar J. (2012). The origin and diversity of the HIV-1 pandemic. *Trends in Molecular Medicine*, 18(3), pp. 182–192.

Hemelaar J. (2013). Implications of HIV diversity for the HIV-1 pandemic. *Journal of Infection*. 66(5), pp. 391–400.

Hu W and Temin H. (1990). Genetic consequences of packaging two RNA genomes in one retroviral particle : pseudodiploidy and high rate of genetic recombination. *Microbiology*, 87(4), pp. 1556–1560.

Huebner R and Todaro G. (1969). Oncogenes of RNA tumor viruses as determinants of cancer. *Microbiology*, 64(3), pp. 1087–94.

Iweriebor B, Bessong P, Mavhandu L, Masebe T, Nwobegahay J, Moyo S. and Mphahlele J. (2011). Genetic analysis of the near full-length genome of an HIV Type 1 A1/C unique recombinant form from northern South Africa. *AIDS Research and Human Retroviruses*, 27(8), pp.911-915.

Jacobs G, Loxton A, Laten A, Engelbrecht S. (2007). Complete genome sequencing of a non-syncytium-inducing HIV Type 1 subtype D strain from Cape Town, South Africa. *AIDS Research and Human Retroviruses*, 23(12), pp. 1575–1578.

Jacobs G, Loxton A, Laten A, Robson B, Janse Van Rensburg E and Englebrecht S. (2009) Emergence and diversity of different HIV-1 subtypes in South Africa, 2000-2001. *Journal of Medical Virology*, 81(11), pp. 1852–1859.

Jacobs G, Wilkinson E, Isaacs S. *et al.,* (2014). HIV-1 subtypes B and C unique recombinant forms (URFs) and transmitted drug resistance identified in the Western Cape Province, South Africa. *PLoS ONE*, 9(3), p.e90845.

Katoh K, Rozewicki J and Yamada K. (2017). MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics*, (March), pp. 1–7.

Kearse M, Moir R, Wilson A. *et al.,* (2012). Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28(12), pp. 1647–1649.

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

Kiwanuka N, Laeyendecker O, Robb M. *et al.,* (2008). Effect of human immunodeficiency virus type 1 (HIV-1) subtype on disease progression in persons from Rakai, Uganda, with incident HIV-1 Infection. *The Journal of Infectious Diseases*, 197(5), pp. 707–713.

Kiwanuka N, Laeyendecker O, Quinn T. *et al.,* (2009). HIV-1 subtypes and differences in heterosexual HIV transmission among HIV-discordant couples in Rakai, Uganda. *AIDS*, 23(18), pp.

2479–2484.

Korber A, Muldoon M, Theiler J. *et al.,* (2000). Timing the Ancestor of the HIV-1 Pandemic Strains. *Science*, 288(5472), pp. 1789–1796.

Korber B, Gaschen B, Yusim K, Thakallapally, Kesmir C, Detours V. (2001). Evolutionary and immunological implications of contemporary HIV-1 variation. *British Medical Bulletin*, 58, pp. 19–42.

Kouri V, Khouri R, Alemán Y. *et al.,* (2015). CRF19_cpx is an evolutionary fit HIV-1 variant strongly associated with rapid progression to AIDS in Cuba. *EBioMedicine*, 2(3), pp.244-254.

Krieg A and Steinberg A. (1990). Retroviruses and autoimmunity. *Journal of Autoimmunity*, 3(2), pp. 137–166.

Lau K and Wong J. (2013). Current trends of HIV recombination worldwide. *Infectious Disease Reports*, 5(14), pp. 16-20.

Lawn S. (2004). AIDS in Africa: The impact of coinfections on the pathogenesis of HIV-1 infection. *Journal of Infection*, 48(1), pp. 1–12.

Lemey P, Salemi M and Vandamme A. (2009) *The Phylogenetic Handbook*.

Liu S, Mittler J, Nickle D. *et al.,* (2002). Selection for human immunodeficiency virus Type 1 recombinants in a patient with rapid progression to AIDS. *Journal of virology*, 76(21), pp. 10674–10684.

Loxton A, Treurnicht F, Laten A, Janse Van Rensburg E and Engelnbrecht S. (2005). 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*, 21(5), pp. 410–413.

Mahapatro G, Mishra D, Shaw D, Mishra S, Jena T. (2012). Phylogenetic tree construction for DNA sequences using clustering methods. *Procedia Engineering*, 38, pp. 1362–1366.

Marx P, Bryant M, Osborn K. *et al.* (1985). Isolation of a new serotype of simian acquired immune deficiency syndrome type D retrovirus from Celebes black macaques (Macaca nigra) with immune deficiency and retroperitoneal fibromatosis. *Journal of Virology*, 56(20), pp. 571–578.

Moutouh L, Corbeil J. and Richman D. (1996). Recombination leads to the rapid emergence of HIV-1 dually resistant mutants under selective drug pressure. *Proceedings of the National Academy of Sciences*, 93(12), pp. 6106–6111.

Mullis K, Faloona F, Scharf S, Saiki R, Horn G and Erlich H. (1986). Specific Enzymatic Amplification of DNA In Vitro: The Polymerase Chain Reaction. *Cold Spring Harbor Symposia on Quantitative Biology*, 51(0), pp.263-273.

Nee S, May R and Harvey P. (1994). The Reconstructed Evolutionary Process. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 344(1309), pp. 305–311.

Nei M and Kumar S. (2000). Molecular evolution and phylogenetics. New York: Oxford University Press.

Neogi U, Bontell I, Shet A. *et al.,* (2012). Molecular epidemiology of HIV-1 subtypes in India: origin and evolutionary history of the predominant subtype C. *PLoS ONE*, 7(6), p.e39819.

Nora T, Charpentier C, Tenaillon O, Hoede C, Clavel F and Hance A. (2007). Contribution of recombination to the evolution of human immunodeficiency viruses expressing resistance to antiretroviral treatment. *Journal of Virology*, 81(14), pp.7620-7628.

Papathanasopoulos M, Cilliers T, Morris L. *et al.,* (2002). Full-length genome analysis of HIV-1 subtype C utilizing CXCR4 and intersubtype recombinants isolated in South Africa. *AIDS Research and Human Retroviruses*, 18(12), pp.879-886.

Peeters M, Liegeois F, Torimiro N. *et al.,* (1999). Characterization of a highly replicative intergroup M/O human immunodeficiency virus type 1 recombinant isolated from a Cameroonian patient. *Journal of virology*, 73(9), pp. 7368–75.

Peeters M. (2001). The genetic variability of HIV-1 and its implications. *Transfus Clinical Biology*, pp. 222–225.

Perrin L, Kaiser L and Yerly S. (2003). Travel and the spread of HIV-1 genetic variants. *Lancet Infectious Diseases*, 3(1), pp. 22–27.

Poiesz B, Ruscetti F, Gazdar A, Bunn P, Minna J. and Gallo R. (1980). Detection and isolation of type C retrovirus particles from fresh and cultured lymphocytes of a patient with cutaneous T-cell lymphoma. *Proceedings of the National Academy of Sciences*, 77(12), pp.7415-7419.

Ramirez B, Simon-Loriere E, Galetto R. and Negroni M. (2008). Implications of recombination for HIV diversity. *Virus Research*, 134(1-2), pp.64-73.

Ritchie A, Cai F, Smith N. *et al.,* (2014). Recombination-mediated escape from primary CD8+ T cells in acute HIV-1 infection. *Retrovirology*, 11(1).

Rous P. (1910). A transmissible avian neoplasm (Sarcoma of the common fowl). *The Journal of Experimental Medicine*, 12(5), pp. 696–705.

Rousseau C, Birditt B, McKay A. *et al.,* (2006). Large-scale amplification, cloning and sequencing of near full-length HIV-1 subtype C genomes. *Journal of Virological Methods*, 136(1-2), pp.118-125.

Saiki R, Gelfand D, Stoffel S.*et al.,* (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, 239(4839), pp.487-491.

Sharp P and Hahn B (2011). Origins of HIV and the AIDS Pandemic. *Cold Spring Harbour Perspectives in Medicine*, pp. 1–22.

Siepel A, Halpern A, Macken C. and Korber B. (1995). A Computer Program Designed to Screen Rapidly for HIV Type 1 Intersubtype Recombinant Sequences. *AIDS Research and Human Retroviruses*, 11(11), pp.1413-1416.

Smyth R, Davenport M and Mak J. (2012). The origin of genetic diversity in HIV-1', *Virus Research*. Elsevier B.V., 169(2), pp. 415–429.

Streeck H, Li B, Poon A. *et al.,* (2008). Immune-driven recombination and loss of control after HIV superinfection. *The Journal of Experimental Medicine*, 205(8), pp. 1789–1796.

Sundquist W and Krusslich H. (2012). HIV-1 assembly, budding, and maturation. *Cold Spring Harbor Perspectives in Medicine*, 2(7), pp. 1–24.

Swofford, D. (2002). PAUP* 4.0: Phylogenetic Analysis Using Parsimony (and Other Methods), Version 4.0b2a. Sinauer Associates, Sun- derland, MA

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. *et al.,* (2013). MEGA6: Molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution*, 30(12), pp. 2725–2729.

Tatem A, Hemelaar J, Gray R. and Salemi M. (2012). Spatial accessibility and the spread of HIV-1 subtypes and recombinants. *AIDS*, 26(18), pp.2351-2360.

Taylor B. and Hammer S. (2008). The Challenge of HIV-1 subtype diversity. *New England Journal of Medicine*, 359(18), pp.1965-1966.

Temin H. (1976). The DNA Provirus Hypothesis. *Science*, 192(4244), pp. 1075–1080.

Turner B. and Summers M. (1999). Structural biology of HIV. *Journal of Molecular Biology*, 285(1), pp. 1–32.

UNAIDS (2017). Homepage. [online] Available at: http://www.unaids.org/ [Accessed 24 Oct. 2018].

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

van Harmelen J, Ryst E, Loubser A.(1999). A predominantly HIV type 1 subtype C-restricted epidemic in South African urban populations. *AIDS Research and Human Retroviruses*, 15(4), pp.395-398.

van Harmelen J, Williamson C, Kim B, Morris L, Carr J, Karim S, Cutchan F. (2001). Characterisation of full-length HIV type 1 subtype C sequences from South Africa. *AIDS research and human retroviruses;* 17(16), pp. 1527-1531.

Vilsker M, Moosa Y, Nooij S. *et al.,* (2018). Genome Detective: An Automated System for Virus Identification from High-throughput sequencing data. *Bioinformatics*.

Weiss R. (2006). The discovery of endogenous retroviruses. *Retrovirology*, 11, pp. 1–11.

Wilkinson E and Engelbrecht S. (2009). Molecular characterization of non-subtype C and recombinant HIV-1 viruses from Cape Town, South Africa. *Infection, Genetics and Evolution*, 9(5), pp. 840–846.

Wilkinson E, Engelbrecht S. and de Oliveira T. (2015). History and origin of the HIV-1 subtype C epidemic in South Africa and the greater southern African region. *Scientific Reports*, 5(1), pp. 1-12.

Wilkinson E, Holzmayer V and Jacobs G. (2015). Sequencing and phylogenetic analysis of near full-length HIV-1 subtypes A, B, G and unique recombinant AC and AD viral strains identified in South Africa. *AIDS Research and Human Retroviruses*, 31(4), pp. 412–420.

World Medical Association Declaration of Helsinki. (2013). *JAMA*, 310(20), p.2191.

Wright J. (2017). Phylogenetic-Manual workshop 2017.

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

Yerly S, Jost S, Monnat M. *et al.,* (2004). HIV-1 co/super-infection in intravenous drug users. *AIDS*, 18(10), pp.1413-1421.

Zhang M, Schultz A, Calef C. *et al*., 2006. jpHMM at GOBICS: a web server to detect genomic recombi- nations in HIV-1. *Nucleic Acids Research* 34, 463–465.

Appendix

Table 0-1: Sequencing primers used to sequence sample EC148

Sample EC148				
Primer Name	**Sequence**	**HXB2 Position**	**Reference**	**Orientation**
***Gag-vpu* fragment**				
ABB55-3-790R	AGGCCTTTTCTTCTACTACTTTTA	1279-1256	Dr Hackett JR	Reverse
P24-7	CCCTGRCATGCTGTCATCA	1832-1850	Swanson *et al.*,2003	Forward
P24_2	AGACYTTAAGCATGGGT	1237-1256	Dr Hackett JR	Forward
RTC	CATTTGTCAGGATGGAGTTCATA	3265-3243	Rosseau *et al.*, 2006	Reverse
RTC	CATTTGTCAGGATGGAGTTCATA	3265-3243	Rosseau *et al.*, 2006	Reverse
3885R	CTGCTCCATCTACATAGAA	3885-3876	Brehm *et al.*, 2012	Reverse
F1	GGGAGACATGGTGGACAGACTATTGGCAAGCCACCTGG	3742-3779	Designed	Forward
POLI05	CACACAAAGGRATTGGAGGAAATG	4177-4200	Swanson *et al.*,2003	Forward
6231R	CTCTCATTGCCACTGTCTTCTGCTC	6231-6207	Grossman *et al.*,2015	Reverse
5550F	AGAGAAGATGGAACAAGCCCCAG	5550-5574	Grossman *et al.*,2015	Forward
VIF1F	GGAATTTGGGTCATGGAGTCTCCATA	5276-5301	Dr Hackett JR	Forward
PANHIV_1	CTTWTATGCAGCWTCTGAGGG	432-412	Gall *et al.*, 2012	Reverse
***Pol-3'LTR* fragment**				
SQ10FC	GGAGCCAGTAGATCCTAACCTAGAG	5833-5857	Rosseau *et al.*, 2006	Forward
55-ENV-22R	AATCGCAAAACCAGCTGGAGCAC	6899-6875	Dr Hackett JR	Reverse
FGF46	GCATTCCCTACAATCCCCAAAG	4648-4669	Gao *et al.*,1996	Forward
ED5	5 (ATGGGATCAAAGCCTAAAG CCATGTG	6134 to 6159	Delwart *et al.*,1993	Forward
ENV-1	CACCGGCTTAGGCATCTCCTATGGCAGGAAGAA	5940-5982	In-house	Forward
LP7725R	GTCCAATGCCAATAAGTCTTGTTC	8216-8193	Dr Hackett JR	Reverse
GP40F1	TCTTAGGAGCAGCAGGAAGCACTATGGG	7789-7816	Dr Hackett JR	Forward
MENV24R	AARCCTCCTACTATCATTATRA	8299-8278	Swanson *et al*, 2003	Reverse

Table 0-2: Sequencing primers used to sequence sample WC416

Sample 416				
Primer Name	**Sequence**	**HXB2 Position**	**Reference**	**Orientation**
***Gag-vpu* fragment**				
0776F	CTAGAAGGAGAGAGATGGGTGCGAG	0776-800	Grossman *et al*.,2015	Forward
P24-7	CCCTGRCATGCTGTCATCA	1832-1850	Swanson *et al*.,2003	Forward
P24-2	AGRACYTTRAAYGCATGGGT	1237-1256	Dr Hackett JR	Forward
GAGa	AGAGAACCAAGGGGAAGTGA	1654 - 1673	Kemp *et al*, 1989	Forward
RTC	CATTTGTCAGGATGGAGTTCATA	3265-3243	Rosseau *et al.*, 2006	Reverse
2713R	GGATTTTCAGGCCCAATTTTTG	2713-2692	Grossman *et al*.,1996	Reverse
SQ5F	AAACAATGGCCATTAACAGAAGAGA	2613-2637	Rosseau *et al.*, 2006	Forward
KVL083	GAATACTGCCATTTGTACTGCTG	4772-4750	Van Laethem, A *et al*.,2008	Reverse
VIF1F	GGAATTTGGGTCATGGAGTCTCCATA	5276-5301	Dr Hackett JR	Forward
5550F	AGARGAYAGATGGAACAAGCCCCAG	5550-5574	Grossman *et al*.,2015	Forward
R 1	GGACAGAAAACAGCATACTACATACTAAAATTGG	4509-4542	Designed	Reverse
SQ10FC	GCCATTGTCTGTATGTATTACTTTGACTG	4583-4555	Rosseau *et al.*, 2006	Reverse
Poli5	CACACAAAGGRATTGGAGGAAATG	4177-4200	Swanson *et al*.,2003	Forward
6231R	CTCTCATTGCCACTGTCTTCTGCTC	6231-6207	Grossman *et al*.,2015	Reverse
F1	GGGAGACATGGTGGACAGACTATTGGCAAGCCACCTGG	3742-3779	Designed	Forward
p24-6	TGTGWAGCTTGYTCRGCTC	1721-1703	Swanson *et al*.,2003	Reverse
***Pol*-3'LTR**				
ENVB	AGAAAGAGCAGAAGACAGTGGCAATGA	6202-6228	In-house	Forward
ENV00	TAGAAAGAGCAGAAGACAGTGGCAATG	6201-6227	Sanders-Buell *et al*, 1995	Forward
ED12	AGTGCTTCCTGCTGCTCCCAAGAACCCA	7811-7784	Delwart *et al*, 1993	Reverse
GP40F1	TCTTAGGAGCAGCAGGAAGCACTATGGG	7789-7816	Dr Hackett JR	Forward
NEF416F	GGGACGCAGCAGTCTCCAGGGAC	9287-9307	Designed	Forward
NEF416R	CAAGGCTACTTCCCTGACTGGCAG	9171-9148	Designed	Reverse

ENV-1	CACCGGCTTAGGC ATCTCCTATGGCAG GAAGAA	5940-5982	In-house	Forward
SQ10RC	GCCATTGTCTGTAT GTATTACTTTGACT G	4583-4555	Rosseau *et al*., 2006	Reverse
55-Env-22R	AATCGCAAAACCAG CTGGAGCAC	6899-6875	Dr Hackett JR	Reverse
ED5	ATGGGATCAAAGC CTAAAG CCATGTG	6134 -6159	Delwart *et al*.,1993	Forward
LP-7725R	GTCCAATGCCAATA AGTCTTGTTC	8216-8193	Dr Hackett JR	Reverse
FGF46	GCATTCCCTACAAT CCCCAAAG	4648-4669	Gao *et al*.,1996	Forward
FGR95	GGTCTAACCAGAG AGACCCAGTACAG	9557-9532	Gao *et al*., 1996	Reverse

UNIVERSITEIT•STELLENBOSCH•UNIVERSITY

Ethics Letter

29-Nov-2017

Ethics Reference #: N15/08/071

Title: Tracking the molecular epidemiology and resistance patterns of HIV-1 in South Africa

Dear Dr Graeme Jacobs,

Your request for extension/annual renewal of ethics approval dated 23 October 2017 refers.

The Health Research Ethics Committee reviewed and approved the annual progress report through an expedited review process.

The approval of the research project is extended for a further year.

Approval Date: 29 November 2017

Expiry Date: 28 November 2018

Kindly be reminded to submit progress reports two (2) months before expiry date.

Where to submit any documentation

Kindly submit **ONE HARD COPY** to Elvira Rohland, RDSD, Room 5007, Teaching Building, and **ONE ELECTRONIC COPY** to ethics@sun.ac.za .

Please remember to use your **protocol number (N15/08/071)** on any documents or correspondence with the HREC concerning your research protocol.

Federal Wide Assurance Number: 00001372
Institutional Review Board (IRB) Number: IRB0005240 for HREC1
Institutional Review Board (IRB) Number: IRB0005239 for HREC2

STELLENBOSCH UNIVERSITY
Health Research Ethics Committee
29 NOV 2017
STELLENBOSCH UNIVERSITEIT
Gesondheidsnavorsing Etiekkomitee

Fakulteit Geneeskunde en Gesondheidswetenskappe
Faculty of Medicine and Health Sciences

Afdeling Navorsingsontwikkeling en -Steun • Research Development and Support Division

Posbus/PO Box 241 • Cape Town 8000 • Suid-Afrika/South Africa
Tel: +27 (0) 21 938 9677

Figure 0.1: Study ethics approval letter for 2018