

**Investigation of the molecular epidemiology of HIV-1 in Khayelitsha,  
Cape Town, using serotyping and genotyping techniques**

**Graeme Brendon Jacobs**

Thesis presented in partial fulfillment of the requirements of the degree of Masters of Sciences in Medical Sciences (Medical Virology) at the Faculty of Health Sciences, University of Stellenbosch.



**Promoter**

Professor Susan Engelbrecht

**Co-promoter**

Dr Corena de Beer

**December 2005**

## DECLARATION

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

Signature: \_\_\_\_\_

Date: \_\_\_\_\_



## SUMMARY

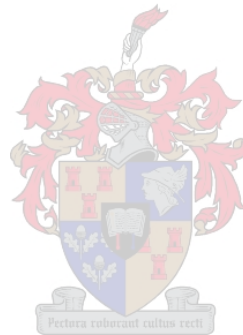
There are currently an estimated 5.3 million people infected with human immunodeficiency virus / acquired immunodeficiency syndrome (HIV/AIDS) in South Africa. HIV-1 group M Subtype C is currently responsible for the majority of HIV infections in sub-Saharan Africa (56% worldwide). The Khayelitsha informal settlement, located 30 km outside Cape Town, has one of the highest HIV prevalence rates in the Western Cape. The objective of this study was to investigate the molecular epidemiology of HIV-1 in Khayelitsha using serotyping and genotyping techniques.

Patient samples were received from the Matthew Goniwe general health clinic located at site C in Khayelitsha. Serotyping was performed through a competitive enzyme-linked immunosorbent assay (cPEIA). RNA was isolated from patient plasma and a two step RT-PCR amplification of the *gag* p24, *env* gp41 IDR, *env* gp120 V3 and *pol* genome regions performed. Sequences obtained were used for detailed sequence and phylogenetic analysis. Neighbour-joining and maximum likelihood phylogenetic trees were drawn to assess the relationship between the Khayelitsha sequences obtained and a set of reference sequences obtained from the Los Alamos National Library (LANL) HIV database (<http://www.hiv.lanl.gov/>).

Through serotyping and genotyping the majority of HIV strains were characterised as HIV-1 group M subtype C. One sample (1154) was characterised as a possible C / D recombinant strain. In 9 other samples HIV-1 recombination cannot be excluded, as only one of the gene regions investigated could be amplified and characterised in these samples. The *gag* p24 genome region was found to be more conserved than the *env* gp41 IDR, with the *env* gp41 IDR more conserved than the *env* gp120 V3. The variability of the *env* gp120 V3 region indicates that patients might be dually infected with variant HIV-1 subtype C strains or quasispecies. Conserved regions identified in the Khayelitsha sequences can induce CD4+ T-cell responses and are important antibody recognition target sites. These conserved regions can play a key role in the development of an effective HIV-1 immunogen reactive against all HIV-1 subtypes. The majority of subtype C viruses were predicted to use CCR5 as their major

chemokine co-receptor. The *pol* sequences analysed indicate that mutations associated with minor resistance to Protease Inhibitors (PIs) might be present in the Khayelitsha community. The identification of resistant mutations is vital for people receiving antiretroviral treatment (ART). It can influence the success of their treatment and delay the onset of AIDS.

Serotyping is a quick characterisation method, but not always accurate. With genotyping detailed molecular analysis can be performed. However, with genotyping the success of amplification often depends on viral load. In Southern Africa a subtype C candidate vaccine appears to be the best option for future vaccine considerations. The sporadic detection of non-subtype C and recombinant subtype C viruses remains a concern and will thus have to be closely monitored. Phylogenetic analysis can help to classify and monitor the spread and evolution of these viruses.



## OPSOMMING

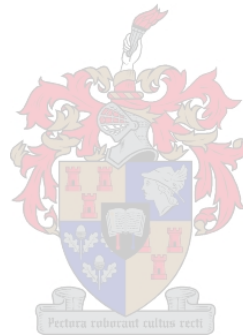
Daar is huidiglik 'n beraamde 5.3 miljoen mense in Suid-Afrika besmet met menslike immuuniteitsgebrek-virus / verworwe immuuniteitsgebreksindroom (MIV/VIGS). MIV-1 groep M sub tipe C is verantwoordelik vir die oorweldigende meerderheid van MIV infeksies in die sub-Sahara gebied van Afrika (56% oor die hele wêreld). Die informele nedersetting van Khayelitsha (omtrent 30 km buite Kaapstad geleë) het een van die hoogste MIV/VIGS syfers in die Wes-Kaap. Die doel van hierdie studie was om die molekulêre epidemiologie van MIV-1 in Khayelitsha na te vors. Dit is gedoen deur gebruik te maak van sero- en genotiperingsmetodes.

Pasiëntmonsters is verkry vanaf die Matthew Goniwe algemene gesondheidskliniek, geleë by terrein C in Khayelitsha. Serotipering is deur 'n kompeterende ensiem-gekoppelde immunologiese toets uitgevoer. RNS is vanaf pasiëntmonsters geïsoleer en 'n tweevoudige amplifikasie van die *gag* p24, *env* gp41 IDR, *env* gp120 V3 en *pol* gene is uitgevoer. Nukleïensuurvolgordes wat bekom is, is in filogenetiese analises gebruik. Beide naaste-verbindinge en maksimum waarskynlikheid filogenetiese bome is geteken om die verhouding tussen die Khayelitsha nukleïensuurvolgordes, en dié wat alreeds op die Los Alamos Nasionale biblioteek (LANL) MIV databasis (<http://www.hiv.lanl.gov>) beskikbaar is te ondersoek.

Deur sero- en genotiperingsmetodes is die meerderheid van MIVs wat ondersoek is as MIV-1 groep M sub tipe C gekarakteriseer. Een monster (1154) is as 'n moontlike sub tipe C / D rekombinante vorm gekarakteriseer. In 9 ander gevalle kan moontlike rekombinasie nie uitgeskakel word nie, aangesien nie genoeg nukleïensuurvolgorde informasie beskikbaar was nie. Die *gag* p24 is meer behoudend as die *env* gp41 IDR, terwyl die *env* gp41 IDR meer behoudend as die *env* gp120 V3 is. Die nukleïensuurvolgordes afwykings in die *env* gp120 V3 geen dui daarop dat moontlike diverse MIV-spesies in die pasiënt teenwoordig is. Gebiede wat behoue bly, kan 'n betekenisvolle rol speel om CD4<sup>+</sup> T-sel reaksies uit te lok en is ook belangrike herkenningsleutels vir teenliggaampies. Hierdie sleutels kan 'n beduidende rol speel in die ontwikkeling van 'n immunogeniese entstof teen MIV wat teen alle MIV-1 subtypes reaktief is. Die meerderheid van die sub tipe C virusse is voorspel om by

voorkeur die CCR5 chemokien koreseptore te gebruik. Die *pol* DNS volgordes dui aan dat weerstand teen Protease Inhibeerders (PIs) dalk in die Khayelitsha gemeenskap teenwoordig kan wees. Om weerstandige mutasies te identifiseer, kan lewensbelangrik wees vir mense wat antivirale behandeling ontvang. Dit kan 'n invloed hê op die doeltreffendheid van die behandeling en die ontwikkeling van VIGS vertraag.

Die studie toon dat serotipering 'n vlugtige karakteriseringsmetode is, maar nie altyd akkuraat is nie. Meer verfynde molekulêre analyses kan met genotipering uitgevoer word. Waarskynlik is 'n MIV-1 sub tipe C entstof vir Suider-Afrika die beste uitweg vorentoe. Die sporadiese identifisering van nie-sub tipe C virusse behoort egter fyn gemonitor te word. Filogenetiese analyses is nuttig om die verspreiding en evolusie van MIV en sy rekombinante te bestudeer en te klassifiseer.



## ACKNOWLEDGEMENTS

I wish to extend my sincere thanks to:

Professor Susan Engelbrecht, my promoter, for all her advice, assistance and guidance throughout my M.Sc project.

Corena de Beer, my co-promoter, for her assistance and insights during the compilation of my thesis.

Dr. John Fincham and his colleagues at the South African Medical Research Council (MRC) for providing the patient samples used during the study.

André Loxton for assisting with the plasma and PBMC isolations.

Annette Laten for aiding with the sequencing reactions.

Annette Laten and Fabian Fiff for performing the viral load assays.

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

My parents, Lawton and Mildred Jacobs for their love and support throughout my life.

My family and friends for their support and encouragement throughout the study period.

My colleagues at the Department of Medical Virology, University of Stellenbosch.

# CONTENTS

	<b>PAGE</b>
Summary	iii
Opsomming	v
Acknowledgements	vii
List of abbreviations	xi
Figures	xvi
Tables	xviii

## **Chapter 1:**

<b>1. Introduction and Literature review</b>	<b>1</b>
1.1 Introduction	3
1.2 Literature review	4
1.2.1 History of HIV-1 infection	4
1.2.2 Origin of HIV	5
1.2.3 The HIV genome, proteins and viral life cycle	6
1.2.4 Envelope protein glycosylation patterns	14
1.2.5 Consensus sequences and conserved genomic regions	15
1.2.6 HIV diversity: A global pandemic	15
1.2.7 The HIV-1 epidemic in South Africa	19
1.2.8 HIV-1 characterisation techniques	23
1.3 Aim of this study	33

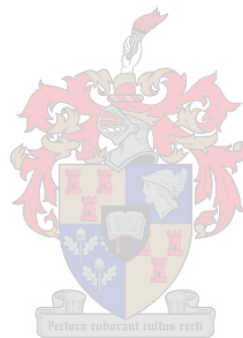
## **Chapter 2:**

<b>2. Materials and Methods</b>	<b>34</b>
2.1 Introduction	35
2.2 Materials	35
2.2.1 Cohort samples	37
2.3 Methods	39
2.3.1 Sample preparation	39
2.3.2 HIV-1 viral load assay	40
2.3.3 <i>env</i> gp120 V3 serotyping assay	40
2.3.4 The polymerase chain reaction	41



	<b>PAGE</b>
2.3.5 Genotyping reactions	41
2.3.6 Agarose gel electrophoresis	42
2.3.7 Purification of PCR products	43
2.3.8 DNA concentration determination	44
2.3.9 DNA cycle sequencing reactions	44
2.3.10 Sequence and phylogenetic analysis	45
2.3.11 Cloning experiments	52
2.3.12 Full length genome analysis	54
<b>Chapter 3:</b>	
<b>3. Results</b>	<b>58</b>
3.1 Introduction	59
3.2 HIV-1 viral load assays	59
3.3 Serotyping with an <i>env</i> gp120 V3 cPEIA	61
3.4 PCR data	62
3.5 Sequencing data	66
3.6 DNA cloning	67
3.7 Sequence and phylogenetic analysis	70
3.8 Near full-length characterisation of possible HIV-1 recombinant strains	108
<b>Chapter 4:</b>	
<b>4. Discussion and Conclusion</b>	<b>109</b>
4.1 Discussion	110
4.1.1 Introduction	110
4.1.2 HIV-1 in Khayelitsha	110
4.1.3 HIV-1 serotyping compared to HIV-1 genotyping	111
4.1.4 HIV-1 nucleotide substitution rates	112
4.1.5 Phylogenetic analysis	112
4.1.6 The role of variable and conserved genome regions of HIV-1	114
4.1.7 The <i>env</i> gp120 V3 loop	115

	<b>PAGE</b>
4.1.8 HIV-1 ART and drug resistance testing	116
4.1.9 Implications for vaccine design	117
4.2 Conclusion	118
 <b>Chapter 5:</b>	
5. References	120
 <b>Appendix</b>	
Appendix A: Ethical approval	158
Appendix B: Complete sequence alignments	160



## LIST OF ABBREVIATIONS

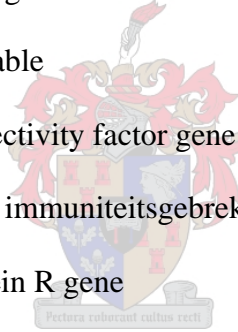
°C	Degree Celsius
©	Copyright
®	Registered
µg	Microgram
µl	Microlitre
A	Absorbance
AIDS	Acquired immunodeficiency syndrome
ART	Antiretroviral treatment
ARV	AIDS-associated retrovirus
AZT	Zidovudine
bDNA	Branched DNA
BLAST	Basic Local Alignment Search Tool
bp	Base pairs
C1 to C5	Constant regions 1 to 5
CA	Capsid protein
cDNA	Complementary DNA
cPEIA	Competitive enzyme linked immunosorbent serotyping assay
CR	Cross-reactive
CRF	Circulating recombinant form
CTL	Cytotoxic T lymphocytes
DBS	Dried blood spots
ddNTPs	Dideoxyribo-nucleoside triphosphates
DNA	Deoxyribonucleic acid
dNTPs	Deoxyribonucleoside triphosphates

dNTPs A, G, C, T	Adenine, Guanine, Cytosine, Thymidine
DRC	Democratic Republic of Congo
EDTA	Ethylene diamine tetra-acetic acid
EIAs	Enzyme immunoassays
<i>env</i>	Envelope gene
Env	Envelope protein
<i>Exo1</i>	Exonuclease 1
FDA	Food and Drug Administration
g	Gram
<i>gag</i>	Group antigen gene
gp	Glycoprotein
GTR	General time reversal model of evolution
HIV	Human immunodeficiency virus
HIV-1	Human immunodeficiency virus type 1
HIV-2	Human immunodeficiency virus type 2
HKY	Hasegawa-Kishino-Yano model of evolution
HTLV	Human T-lymphotropic virus
I + G	Invariant sites and gamma-distribution
IAVI	International AIDS Vaccine Initiative
IDR	Immunodominant region
IDU	Intravenous drug user
IN	Integrase protein
indels	Insertions and deletions
IPTG	Isopropyl- $\beta$ -D-thiogalactopyranosid
kb	Kilo – base pairs

l	Litre
LANL	Los Alamos National Library
LAS	Lymphadenopathy syndrome
LAV	Lymphadenopathy virus
LB	Luria-Bertani
LDL	Lower than the detection limit
LTR	Long terminal region
M	Molar
mAbs	Monoclonal antibodies
MDA	Multiple Displacement Amplification
mg	Milligram
MHC	Major histocompatibility
MIV	Menslike immuniteitsgebrek-virus
ml	Millilitre
mM	Millimolar
MRC	Medical Research Council
MSF	Médecins Sans Frontières, “Doctors without Borders”
MTCT	Mother-to-child transmission
NASBA	Nucleic acid based amplification assay
NC	Nucleocapsid protein
<i>nef</i>	Negative factor gene
ng	Nanogram
NNI	Nearest-neighbour interchange
NNRTIs	Non-nucleoside RT inhibitors
NR	Non-reactive

NRTIs	Nucleoside / nucleotide RT inhibitors
NSI	Non-syncytium inducing
PCR	Polymerase chain reaction
PI	Protease inhibitor
pmol	Picomole
<i>pol</i>	Polymerase gene
PR	Protease enzyme
<i>rev</i>	Regulator of viral expression gene
RNA	Ribonucleic acid
RT	Reverse Transcriptase enzyme
SAAVI	South African AIDS Vaccine Initiative
SAP	Shrimp alkaline phosphatase
SI	Syncytium inducing
SIV	Simian immunodeficiency virus
SPR	Subtree pruning and regrafting
STDs	Sexually transmitted diseases
SU	Surface glycoproteins
SYM	Symmetrical model of evolution
TAC	Treatment Action Campaign
TN	Tamura-Nei (TN93) model of evolution
<i>Taq</i>	<i>Thermus aquaticus</i>
<i>tat</i>	Transcriptional transactivator gene
TB	Tuberculosis
TBR	Tree bisection and reconstruction
<i>Tfl</i>	<i>Thermus flavus</i>

<i>Tgo</i>	<i>Thermococcus gorgonarius</i>
TIM	Transition model of evolution
TM	Transmembrane protein
™	Trademark
TSR	Template suppression reagent
TVM	Transversion model of evolution
U3	Unique 3` region
U5	Unique 5` region
USA	United States of America
UPGMA	Unweighted pair group method with arithmetic mean
V1 to V5	Variable regions 1 to 5
V3	Third variable
<i>vif</i>	Virion infectivity factor gene
VIGS	Verworwe immuniteitsgebreksindroom
<i>vpr</i>	Viral protein R gene
<i>vpu</i>	Viral protein U gene
WHO	World Health Organisation
X-Gal	X-Galactosidase



## FIGURES

	PAGE
<b>Figure 1.1:</b> Estimated number of people living with HIV/AIDS at the end of 2004	4
<b>Figure 1.2:</b> The HIV-1 genome	7
<b>Figure 1.3:</b> A schematic diagram of the HIV virion	7
<b>Figure 1.4:</b> The HIV-1 replication cycle	9
<b>Figure 1.5:</b> The <i>env</i> gp120 core	12
<b>Figure 1.6:</b> Global distribution of HIV-1 group M subtypes and recombinants	17
<b>Figure 1.7:</b> Prevalence of HIV-1 among antenatal care attendees in South Africa, 1990-2003	19
<b>Figure 1.8:</b> Khayelitsha, Western Cape	21
<b>Figure 1.9:</b> An example of an unrooted and rooted phylogenetic tree	30
<b>Figure 1.10:</b> DNA substitution mutations	32
<b>Figure 3.1:</b> Serotype graph	61
<b>Figure 3.2:</b> Example of a 0.8% agarose gel with the <i>env</i> gp120 V3, <i>gag</i> p24 and <i>env</i> gp41 IDR	62
<b>Figure 3.3:</b> <i>pol</i> PCR amplification on a 0.8% agarose gel	63
<b>Figure 3.4:</b> <i>gag</i> p24 PCR fragments used for cloning	67
<b>Figure 3.5:</b> A <i>gag</i> p24 PCR from cultures grown overnight	68
<b>Figure 3.6:</b> Partial restriction enzyme digestion of cloned <i>gag</i> p24 cultures	70
<b>Figure 3.7:</b> A <i>gag</i> p24 neighbour-joining phylogenetic tree with the 3 cloned fragments 1039, 1151 and 1154	77
<b>Figure 3.8:</b> The <i>gag</i> p24 neighbour-joining phylogenetic tree	78
<b>Figure 3.9:</b> The <i>env</i> gp41 neighbour-joining phylogenetic tree	79
<b>Figure 3.10:</b> The <i>env</i> gp120 V3 neighbour-joining phylogenetic tree	80
<b>Figure 3.11:</b> Analysis of possible hypermutant HIV-1 <i>env</i> gp120 V3 sequences	81
<b>Figure 3.12:</b> A subtype D <i>gag</i> p24 neighbour-joining phylogenetic tree	82



	<b>PAGE</b>
<b>Figure 3.13:</b> A <i>pol</i> neighbour-joining phylogenetic tree with samples 1039 and 1151 drawn with the reference sequences	83
<b>Figure 3.14:</b> A <i>pol</i> neighbour-joining phylogenetic tree with the sequences from samples 1039 and 1151	84
<b>Figure 3.15:</b> A <i>gag</i> p24 maximum likelihood phylogenetic tree with the sequences from the 3 cloned samples (1039, 1151 and 1154)	87
<b>Figure 3.16:</b> The <i>gag</i> p24 maximum likelihood phylogenetic tree	88
<b>Figure 3.17:</b> The <i>env</i> gp41 IDR maximum likelihood phylogenetic tree	89
<b>Figure 3.18:</b> A subtype D <i>gag</i> p24 maximum likelihood phylogenetic tree	90
<b>Figure 3.19:</b> A <i>pol</i> maximum likelihood phylogenetic tree with the sequences from samples 1039 and 1151	91
<b>Figure 3.20:</b> A <i>pol</i> maximum likelihood phylogenetic tree with the sequences from samples 1039 and 1151	92
<b>Figure 3.21:</b> The <i>gag</i> p24 similarity plot of the sequence from sample 1154	93
<b>Figure 3.22:</b> The 1.2 kb <i>pol</i> similarity plot of the sequence from sample 1039	94
<b>Figure 3.23:</b> The 1.2 kb <i>pol</i> similarity plot of the sequence from sample 1151	94
<b>Figure 3.24:</b> Percentage variation between the Khayelitsha consensus sequence and other consensus sequences	95
<b>Figure 3.25:</b> The Khayelitsha <i>env</i> gp120 V3 consensus sequence compared to the subtype C consensus and ancestral sequences	96
<b>Figure 3.26:</b> A <i>gag</i> p24 amino acid entropy graph	97
<b>Figure 3.27:</b> An <i>env</i> gp41 IDR amino acid entropy graph	98
<b>Figure 3.28:</b> An <i>env</i> gp120 V3 amino acid entropy graph	98
<b>Figure 3.29:</b> Genotype resistance interpretation results for the sequence of sample 1039	107
<b>Figure 3.30:</b> Genotype resistance interpretation results for the sequence of sample 1151	107
<b>Figure 3.31:</b> Amplification of genomic DNA from sample 1154	108

## TABLES

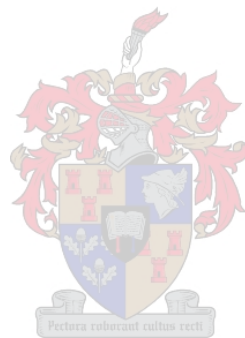
	PAGE
<b>Table 2.1:</b> Equipment used to perform sample assays and analysis	35
<b>Table 2.2:</b> List of commercial products and assays used	36
<b>Table 2.3:</b> Additional chemicals needed for analysis	37
<b>Table 2.4:</b> HIV-1 genotyping primers	43
<b>Table 2.5:</b> HIV- 1 subtype reference sequences used for phylogenetic analysis	49
<b>Table 2.6:</b> Subtype D gag p24 sequences used in phylogenetic analysis	50
<b>Table 2.7:</b> Subtype C pol sequences used in phylogenetic analysis	51
<b>Table 2.8</b> HIV-1 full-length amplification primers	56
<b>Table 2.9:</b> HIV-1 primers used to amplify overlapping genomic regions	57
<b>Table 3.1:</b> RNA viral load results using the Abbott LCx <sup>®</sup> HIV RNA Quantitative assay	
	59
<b>Table 3.2:</b> Summary of serotyping and PCR results	63
<b>Table 3.3:</b> Position of amplified fragments compared to HXB2	66
<b>Table 3.4:</b> Initial BLAST results and sequence analysis of unusual HIV strains	67
<b>Table 3.5:</b> Average number of colonies per plate observed	68
<b>Table 3.6:</b> DNA concentrations of cloned samples	69
<b>Table 3.7:</b> Possible hypermutations in the <i>env</i> gp120 V3 sequences	73
<b>Table 3.8:</b> Sample 1154 <i>gag</i> p24 subtype D sequence similarity	85
<b>Table 3.9:</b> Sample 1039 and 1151 <i>pol</i> subtype C sequence similarity	86
<b>Table 3.10:</b> Conserved amino acid regions	99
<b>Table 3.11:</b> <i>env</i> gp120 V3 co-receptor prediction	101
<b>Table 3.12:</b> Envelope N-Glycosylation numbers	104

# CHAPTER ONE

## 1. Introduction and Literature review

	<b>PAGE</b>
1.1 Introduction	3
1.2 Literature review	4
1.2.1 History of HIV-1 infection	4
1.2.2 Origin of HIV	5
1.2.3 The HIV genome, proteins and viral life cycle	6
1.2.3.1 The virus and genome structure	6
1.2.3.2 The HIV-1 replication cycle	8
1.2.3.3 HIV-1 genome regions relevant to this study	10
1.2.3.3.1 <i>gag</i> p24	10
1.2.3.3.2 <i>env</i> gp41 Immunodominant region	11
1.2.3.3.3 <i>env</i> gp120 V3	12
1.2.3.3.4 The <i>pol</i> gene	13
1.2.4 Envelope protein glycosylation patterns	14
1.2.5 Consensus sequences and conserved genomic regions	15
1.2.6 HIV diversity: A global pandemic	15
1.2.6.1 Subtype C and its recombinants	18
1.2.7 The HIV-1 epidemic in South Africa	19
1.2.7.1 HIV-1 diversity in South Africa	19
1.2.7.2 HIV-1 and its social and economic impact on South Africa	20
1.2.7.3 Addressing the problem in Khayelitsha	21
1.2.8 HIV-1 characterisation techniques	23
1.2.8.1 HIV-1 viral load assays	23
1.2.8.2 The <i>env</i> gp120 V3 serotyping assay	23
1.2.8.3 HIV-1 genotyping	24
1.2.8.4 Nucleic acid extraction	25
1.2.8.5 The polymerase chain reaction	26
1.2.8.6 DNA cloning	26
1.2.8.7 DNA sequencing	27

	<b>PAGE</b>
1.2.8.8 Phylogenetic analysis	28
1.2.8.8.1 What is phylogenetic analysis?	28
1.2.8.8.2 Multiple alignments and phylogenetic trees	29
1.2.8.8.3 Models of evolution	31
1.3 Aim of this study	33



# CHAPTER ONE

## 1. Introduction and Literature Review

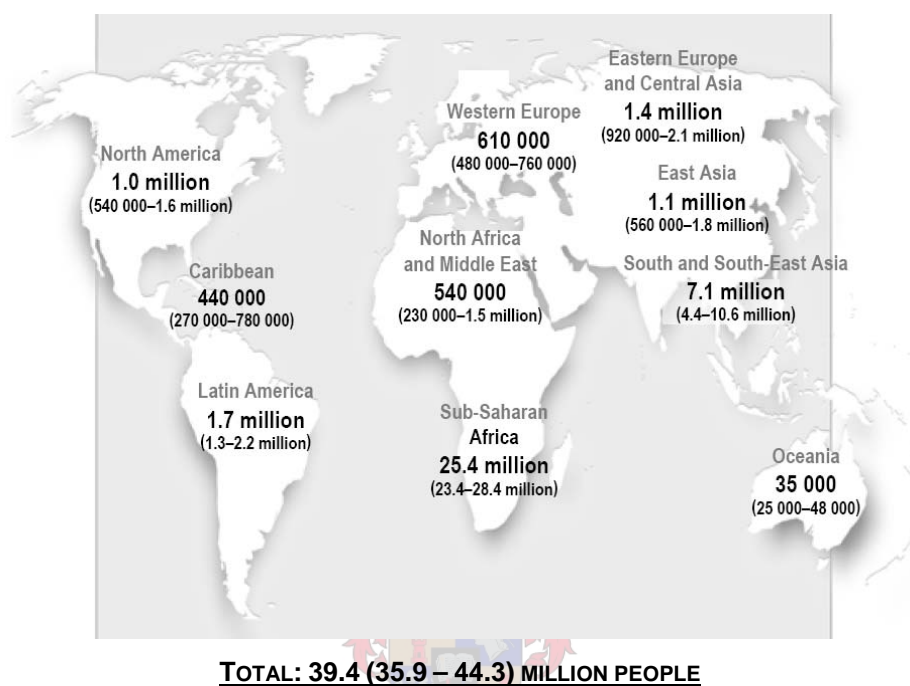
### 1.1 Introduction

Since the discovery of human immunodeficiency virus / acquired immunodeficiency syndrome (HIV/AIDS) in 1981 an estimated 60 million people have become infected with the virus. There seems to be no stopping the current trend of the pandemic, as approximately 4.9 million new infections occurred in 2004, while 3.1 million lives have been lost due to HIV/AIDS related causes (UNAIDS, 2004). This disease clearly has had devastating effects on mankind and the world need to stand together as one community if we are going to combat the virus successfully.

There are currently two known types of HIV: HIV-1 and HIV-2. HIV-1 is responsible for the pandemic we are facing today. It has been divided into three groups: Group M (main), N (Non-M, Non-O/New) and O (Outlier) (Robertson *et al*, 2000, Spira *et al*, 2003). HIV-1 group M subtype C is the major viral subtype found in South Africa, with sporadic reports of other HIV-1 group M subtypes (HIV-1 subtypes) and recombinants (Esparza and Bhamarapavati, 2000; Osmanov *et al*, 2002). This is a major concern, especially considering the development of an effective vaccine against the predominant HIV-1 subtype in a specific geographical area. Sub-Saharan Africa consists mostly of third world and developing countries and yet it is the region that needs the most support in fighting HIV/AIDS. It is estimated that between 23.4 and 28.4 million people in this region are currently living with the disease (Figure 1.1), which account for 64% of all current HIV-1 infections (UNAIDS, 2004).

The Khayelitsha Township has one of the highest HIV-1 prevalence rates (27.2%) within the Western Cape Province of South Africa (Department of Health, 2004). This is a major health and social issue for the township's residents. Khayelitsha, with an estimated population of 400 000 people, is a poor community with restricted resources and without help they will not be able to cope with the existing HIV/AIDS

problem. A recent report in Uganda showed that community-based education and awareness campaigns have drastically reduced the HIV-1 prevalence in that country by 70% over the past few years (Stoneburner and Low-Ber, 2004). Ideally, such an approach should be taken in countries such as South Africa, starting with communities in crisis, such as Khayelitsha.



**Figure 1.1: Estimated number of people living with HIV/AIDS at the end of 2004.** Sub-Saharan Africa and South and South East Asia are the geographical areas with the highest number of HIV-1 infections. The lowest number of infections are found in Oceania (Australasia) and in the Caribbean Islands between North and South America (UNAIDS, 2004).

## 1.2 Literature Review

An overview of the current literature on HIV (HIV-1 and HIV-2) history, the virus structure, replication cycle and HIV diversity is presented. Focus is also placed on HIV-1 in South Africa and Khayelitsha. A brief review on the techniques used during the study is also presented.

### 1.2.1 History of HIV-1 infection

The first reports of HIV/AIDS were described in 1981 in the United States of America (USA) amongst homosexual men who had a rare disease, *Pneumocystis carinii*

pneumonia (Gottlieb *et al*, 1981a; Gottlieb *et al*, 1981b). A few of these patients also developed Kaposi's sarcoma (Friedman-Kien *et al*, 1981). It was believed that this new emerging disease was nothing more than punishment for the lifestyle and high-risk behaviour of individuals such as homosexuals and intravenous drug users (IDUs) (Sepkowitz, 2001; Shilts, 1987). Not long after these initial cases, signs and symptoms often preceding AIDS were also reported in other population groups, such as infants (Oleske *et al*, 1983), female sexual partners of men (Masur *et al*, 1982), haemophiliacs (Bloom, 1984), blood transfusion recipients (Curran *et al*, 1984), as well as the heterosexual population of Zaire (previously Zaire, currently the Democratic Republic of Congo, DRC) in Africa (Piot *et al*, 1984; Sepkowitz, 2001). The search had begun to find the etiological agent causing this new emerging immunodeficiency. In 1983, Barré-Sinoussi and co-workers isolated a retrovirus from a homosexual man who consistently presented with lymphadenopathy syndrome (LAS) (Barré-Sinoussi *et al*, 1983). This was the first time HIV, first called lymphadenopathy virus (LAV), was isolated. The same virus was also identified by Levy and co-workers who called it the AIDS-associated retrovirus (ARV) (Levy *et al*, 1984). Robert Gallo and his colleagues hypothesised that a variant of the human T-lymphotropic virus (HTLV) might be the causative agent of AIDS (Gallo *et al*, 1984). It was independently confirmed that this new retrovirus was indeed the cause of AIDS (Ratner *et al*, 1985a; Ratner *et al*, 1985b). By 1986 the same retrovirus had three designations: LAV, ARV and HTLV-III. This was confusing and the International Committee on the Taxonomy of viruses decided to rename the AIDS virus HIV (Coffin *et al*, 1986a; Coffin *et al*, 1986b). Today, heterosexual transmission is responsible for the majority of new HIV-1 infections (Esparza and Bhamarapavati, 2000; Osmanov *et al*, 2002) and even though Zidovudine (AZT), the first Food and Drug Administration (FDA) (USA bureau) approved drug against HIV/AIDS, was introduced in 1987 (Fischl *et al*, 1987), no known cure has been found to date.

### **1.2.2 Origin of HIV**

HIV has probably been around for many years and reports of possible AIDS cases predating 1981 have retrospectively been identified (Hummer *et al*, 1987). The earliest known report of HIV infection is derived from a HIV sequence from a seropositive patient in Kinshasa, DRC from 1959 (Zhu *et al*, 1998). Most researchers

believe phylogenetic analysis has clarified the argument that HIV is derived from related simian immunodeficiency viruses (SIVs) found in primates. These viruses do not usually result in similar AIDS defining illnesses in our non-human primate counterparts (Hahn *et al*, 2000; Silvestri *et al*, 2003). HIV-1 was probably transmitted from the common chimpanzee, *Pan troglodytes troglodytes* (Gao *et al*, 1999; Hahn *et al*, 2000; Santiago *et al*; 2002), while HIV-2 originates from the sooty mangabey, *Cerocebus atys* (Gao *et al*, 1992; Hahn *et al*, 2000; Hirsch *et al*, 1989). Humans are not the natural host of these viruses and precisely when and how HIV crossed from ape to human will never truly be known. Phylogenetic analysis shows that HIV was introduced into the human population during the 1930s with a  $\pm 20$  year confidence gap (Hahn *et al*, 2000; Korber *et al*, 2000). Several reports speculate that HIV originated in central Africa (Apetrei *et al*, 2004; Nahmias *et al*, 1986) where these transmissions most likely first occurred and probably still do. Some of these primates are slaughtered as a food source or kept as household pets and cross-species transmissions are probably common (Weiss and Wrangham, 1999).

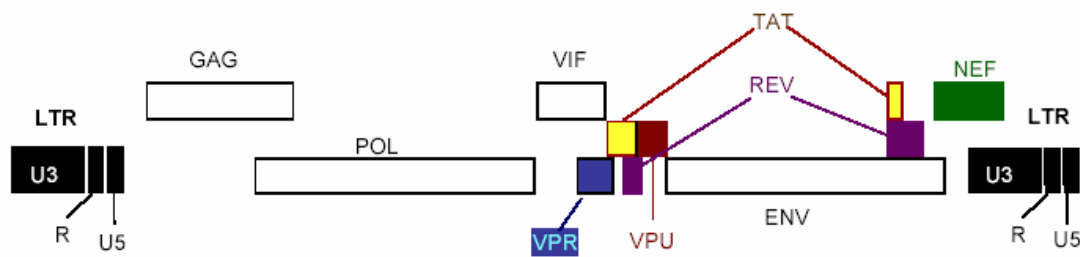
### **1.2.3 The HIV genome, proteins and life cycle**

Detailed reviews of the HIV genome, proteins and life cycle (HIV Biology) are given by the following publications: Briggs *et al*, 2003; Freed, 1998; Freed, 2001; Goto *et al*, 1998; Joshi and Joshi, 1996; Nisole and Saïb, 2004; Turner and Summers, 1999.

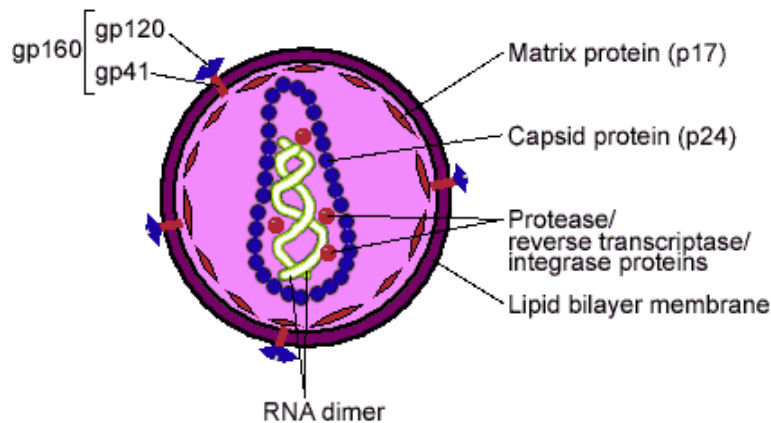
#### **1.2.3.1 The virus and genome structure**

A schematic diagram of the HIV-1 genome is presented in Figure 1.2, while the HIV virion is depicted in Figure 1.3. HIV is a retrovirus and contains two copies of unspliced genomic viral RNA. The virus is enveloped by a lipid membrane derived from the membrane of the host cell. The virus surface contains distinct 72 knob shaped trimers or tetramers of the Envelope (Env) glycoproteins (Gottlinger, 2001; Levy, 1998). These are the exposed surface glycoproteins (SU), which are anchored to the virus via interactions with the transmembrane proteins (TM). These proteins are derived from the *env* gp160 precursor. The *env* gp160 precursor is cleaved into the gp120 derived SU and gp41 derived TM.





**Figure 1.2: The HIV-1 genome.** The different HIV-1 genes, as well as the U3 (unique 3' region), R (terminal redundancy region) and U5 (unique 5' region) Long Terminal Repeat (LTR) regions are indicated. The *env* (envelope), *gag* (group antigen) and *pol* (polymerase) genes encode for the structural proteins. The regulatory and accessory genes *nef* (negative factor gene), *tat* (transcriptional transactivator), *rev* (regulator of viral expression), *vif* (virion infectivity factor), *vpr* (viral protein R) and *vpu* (viral protein U) are involved in viral replication, infectivity and maturation (Gatignol and Jeang, 2000).



**Figure 1.3: A schematic diagram of the HIV virion.** The diagram displays the Envelope (gp160), Gag (p17 and p24) and Pol (Protease, Reverse Transcriptase and Integrase) proteins, as well as the RNA dimer (<http://www.mclcd.co.uk/hiv/>). HIV is an enveloped virus. It consists of two copies of unspliced genomic RNA surrounded by a conically shaped capsid core. The Matrix (MA) proteins cover the inner surfaces of the virus particle.

The lipid bilayer also contains several cellular membrane proteins, including major histocompatibility (MHC) antigens derived from the host cell (Arthur *et al*, 1992). The inner surface of the viral membrane is lined with a matrix shell consisting of p17 derived MA proteins. In the center of the virus particle the conical shaped capsid core consisting of the Capsid protein (CA, derived from the *gag* p24) encapsidates the RNA genome. The virally encoded enzymes Protease (PR), Reverse Transcriptase (RT) and Integrase (IN), as well as the Nucleocapsid (NC) proteins, are closely associated with the ribonucleoprotein complex stabilised RNA in the capsid core.

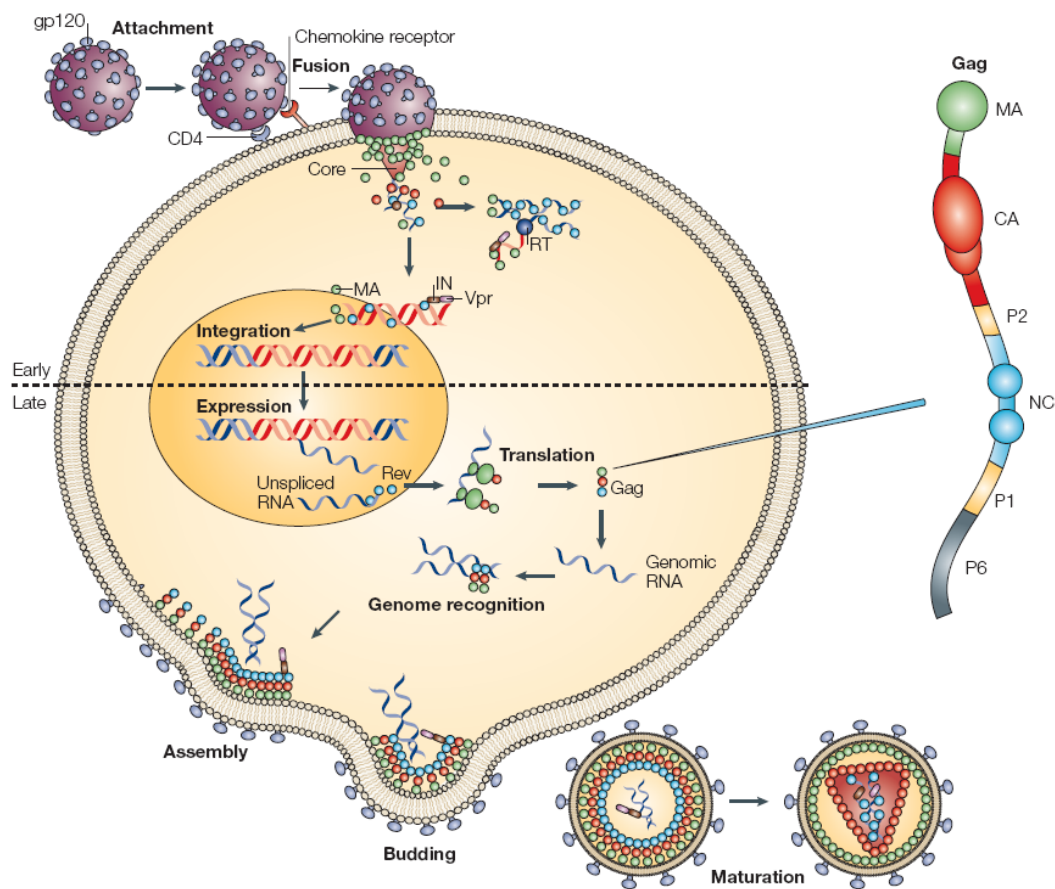
Some accessory proteins (Nef, Vif and Vpr) are also packaged by virus particles in the core.

Together *env*, *gag* *pol* comprise the structural genes. The *env* gene encodes for the gp160 precursor, *gag* for the MA, CA as well as NC, and *pol* for the enzymes PR, RT and IN. The HIV genome also encodes for several regulatory and accessory genes. The regulatory proteins Tat and Rev are encoded by the *tat* and *rev* genes. These are both needed for viral replication *in vitro*. The Tat protein is a viral transcriptional transactivator, while Rev is a regulator of viral protein expression. The Rev protein is also involved in RNA transport (Emerman and Malim, 1998). The accessory genes include *nef*, *vif*, *vpr*, *vpu* and *vpx* (*viral protein X*). They are not necessary for viral replication *in vitro*, but play a variety of roles during the life cycle of HIV. Vif is essential for viral infectivity, as well as virion maturation, while Nef plays a role in CD4 and MHC class I down regulation. The Vpu and Vpx proteins promote virion production, as well as virus release from the host cells, while Vpr enhances viral expression and promotes the extra cellular release of viral particles (Bour and Strebel, 2003). The *vpu* gene is only found in HIV-1 and the *vpx* gene in HIV-2. The 9.2 kb HIV genome is flanked by two LTR regions that do not encode for any proteins. The LTRs do however contain important transcription factors and binding sites for the regulation of viral gene expression (Briggs *et al*, 2003).

Within the HIV-1 genome the highest diversity is seen in the *env* gene (Gordon and Delwart, 2000; Wain-Hobson, 1995) and the lowest in the *pol* gene (Cornelissen *et al*, 1997, Servais *et al*, 2004). Virus diversity between individuals may reach 20% if they are infected with the same subtype (Delwart *et al*, 2002; Karlsson *et al*, 1998), 30% between group M subtypes (Vidal *et al*, 2000a) and up to 50% between the various HIV groups (Simon *et al*, 1998). The viral Env protein continuously has to evade the host immune response, resulting in multiple mutations within the *env* gene. The *pol* gene encodes for vital viral enzymes and mutations may lead to impaired protein function.

### 1.2.3.2 The HIV-1 replication cycle

A schematic representation of the virus replication cycle can be viewed in Figure 1.4.

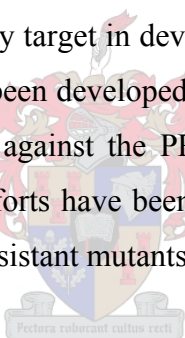


**Figure 1.4: The HIV-1 replication cycle.** The life cycle is divided into two distinct phases: the early phase (upper portion of diagram) up to integration of the proviral DNA and the late phase, which includes all events from transcription to virus budding and maturation (D'Souza and Summers, 2005).

HIV infects cells of the immune system such, as CD4<sup>+</sup> T-cells, Cytotoxic T-lymphocytes (CTLs), CD4<sup>+</sup> monocytes, macrophages and CD4<sup>+</sup> dendritic cells (Stebbing *et al*, 2004). The virus can be found in blood plasma, peripheral blood mononuclear cells (PBMCs), lymph nodes, the central nervous system and various other body fluids and cells after infection (Pierson *et al*, 2000, Stebbing *et al*, 2004). The life cycle can be divided into two distinct phases. The early replication phase extends from virus attachment to integration of the viral genome into the host cell. HIV uses its host CD4 molecule along with a chemokine receptor, either CCR5 or CXCR4, to bind and enter the host cell (Regoes and Bonhoeffer, 2005). Env gp120 binds to CD4 and forces a conformational change that allows the host and viral membranes to fuse. The viral RNA is released into the host cytoplasm and uses the RT enzyme to synthesise a double stranded DNA copy from its RNA. With the help

of the viral IN enzyme the newly synthesised DNA is incorporated into the host genome. The integrated provirus often establishes latency in the infected cell. New viral RNA is synthesised from the provirus. Gene expression is regulated by both cellular and viral factors. Late stage HIV replication includes expression of viral proteins followed by viral budding and maturation. This starts when spliced and unspliced mRNA transcripts are transported out of the nucleus for translation. After genome replication the newly formed virus exits the cell by budding and is free to infect neighbouring cells. The PR enzyme usually cleaves the Gag polyprotein after the newly formed virus has left the host cell (Freed, 1998; Freed 2001). The host cell dies from the effects of continuous immune activation that occurs in HIV-1 infected patients (Badley *et al*, 2003; Badley 2005). Host cell death, or apoptosis, causes severe depletion of CD4+ T-cells and paralyses the host immune system (Roshal *et al*, 2001).

The HIV-1 life cycle has been a key target in developing efficient antiretroviral drugs against HIV-1. Many drugs have been developed to stop viral entry, or interfere with viral protein functions, especially against the PR and RT enzymes as described in section 1.2.3.3.4. However, all efforts have been unsuccessful in eliminating HIV-1 infection thus far and many drug resistant mutants have been identified (Miller, 2001).



### **1.2.3.3 HIV-1 genome regions relevant to this study**

A brief literature review is presented on the *gag* p24, *env* gp41 Immunodominant region (IDR), *env* gp120 V3 and the *pol* gene, as they are important target areas for the purposes of this study. They also play a key role in certain important diagnostic tests, as described in section 1.2.8.3 (Parekh and McDougal, 2005; Swanson *et al*, 2003).

#### **1.2.3.3.1 *gag* p24**

The *gag* p24 encodes for the CA protein, as described in section 1.2.3.1. Capsid assembly is important for viral infectivity, therefore genome and structural studies involving CA are essential (Forshey *et al*, 2002). CA of mature HIV-1 is conically shaped and surrounds the viral RNA nucleoprotein complex (Figure 1.3 and 1.4)

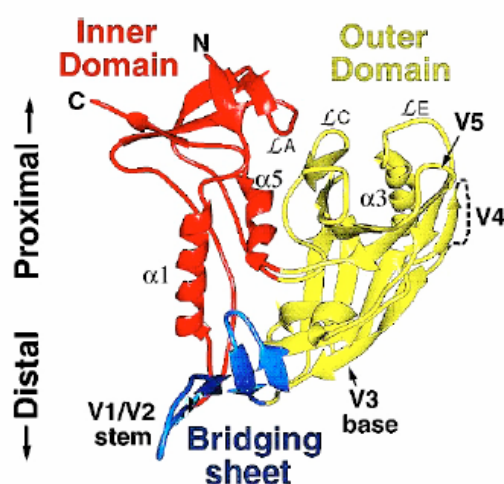
(Freed, 1998; Freed 2001). The high-resolution structure of the Gag proteins have been reviewed in detail by other authors (Turner and Summers, 1999). Processed HIV-1 p24 consists of two  $\alpha$  helical domains. The N-terminal has seven  $\alpha$ -helices and the C-terminal four. They are connected to each other via a flexible linker (Berthet-Colominas *et al*, 1999). The N-terminal domain contributes to viral core formation (Yoo *et al*, 1997; Kaplan, 2002), while the C-terminal domain is involved in the oligomerisation of Gag and Gag-Pol precursors necessary for virion budding (Borsetti *et al*, 1998; Chiu *et al*, 2002; Kaplan, 2002). Mutations in the *gag* p24 have an effect on the viral assembly process and viruses with impaired p24 function are generally non-infectious of nature (Dorfman *et al*, 1994). Amino acid substitutions in the C-terminal and N-terminal regions have also been implicated in affecting viral assembly and release (Abdurahman *et al*, 2004; Scholz *et al*, 2005).

#### **1.2.3.3.2 *env* gp41 Immunodominant region**

The *env* gene encodes two heavily glycosylated proteins, namely the gp120 outer membrane and the carboxy-terminal transmembrane gp41 (Hunter, 1997; Leitner, 1996a). The Env gp41 protein is a multifunctional protein and is important for HIV entry and viral pathogenesis (Hunter, 1997). The IDR is one of many gp41 functional domains. It consists of a cluster I with the CTL epitope and cysteine loop and cluster II with the ectodomain region. Since the viral Env proteins are exposed to its host immune defenses, more than 99% of HIV-1 infected individuals produce antibodies against the *env* gp41 IDR domains (Cano *et al*, 2004; Horal *et al*, 1991; Hunter, 1997). Antibodies recognising these clusters do not normally neutralise HIV-1 infection (Cano *et al*, 2004; Hunter, 1997). Previous neutralisation studies have however identified mutation variations in the ectodomain of gp120 and gp41 that resulted in altered antibody binding, due to changes in conformation or glycosylation patterns (Kalia *et al*, 2005; Kwong *et al*, 2002; Lue *et al*, 2002). Diversity seen within IDRs of structural proteins, such as Env, directly influences antibody detection methods (Dorn *et al*, 2000). The *env* sequences are highly variable. All genetically diverse HIV groups and subtypes have been characterised thoroughly based on sequences from the *env* gene. Thus, *env* is a principal target region for epidemiologically linked subtype studies as it can provide information regarding all circulating subtypes in a certain geographical area (Pieniazek *et al*, 1998).

### 1.2.3.3.3 *env* gp120 V3

HIV-1 *env* gp120 has been divided into five constant (C1 to C5) and five variable regions (V1 to V5) (Starcich *et al*, 1986). The variable regions are mostly found within regions encoding disulfide-constrained loops exposed to the surface and to the host immune system (Leonard *et al*, 1990). The structure of gp120 has been published on extensively (Poignard *et al*, 2001; Wyatt *et al*, 1998). The gp120 core can be viewed in Figure 1.5. Even though other gp120 regions play a role in predicting viral phenotype (Carrillo and Ratner, 1996; Cho *et al*, 1998; Koito *et al*, 1994), the V3 loop has been the focus of most researchers (Bickel *et al*, 1996; Hartley *et al*, 2005; Korber *et al*, 1993; Hoffman *et al*, 2002).



**Figure 1.5: The *env* gp120 core; inner domain (red), outer domain (yellow), bridging sheet (blue).** The inner domain is believed to interact with the gp41 Env glycoprotein, while the outer domain, which is quite variable (V1 to V5 are indicated) and heavily glycosylated, is believed to be exposed on the assembled envelope glycoprotein trimer (Wyatt *et al*, 1998).

Early biological studies have found that HIV-1 either produces non-syncytium inducing (NSI) or syncytium inducing (SI) viruses *in vitro* (Fenyo *et al*, 1988). These phenotypes were associated with differences in growth properties and cytopathicity on PBMCs. A SI cell phenotype is a mass of multinucleated cytoplasm with no internal cell boundaries visible. This phenotype is absent in NSI viruses (De Jong *et al*, 1992; Fenyo *et al*, 1988; Fenyo *et al*, 1997). NSI viruses often use CCR5 as their major chemokine co-receptor along with CD4+ T-cells, whereas SI viruses use the CXCR4 chemokine co-receptor. CXCR4 / SI viruses have also been associated with rapid

progression to AIDS disease (Maas *et al*, 2000; Regoes and Bonhoeffer, 2005). The V3 region has been recognised as a crucial target area for vaccine development (Javaherian *et al*, 1990; Moore and Nara, 1991; Binley *et al*, 2004). The role of V3 tropism and its impact on the development of a HIV-1 vaccine are reviewed in detail by Hartley *et al*, 2005. Briefly, a successful vaccine must be able to generate antibodies against the surface exposed V3 region. Antibodies that recognise certain V3 motifs, such as the Glycine – Proline – Glycine – Arginine (GPGR) motif at the crown of the V3 loop (Gaschen *et al*, 1999), despite subtype diversity, have been identified. The GPGR motif is associated with HIV-1 subtype B, while the Glycine – Proline – Glycine – Glutamine (GPGQ) motif is associated with either HIV-1 subtype A or C. Monoclonal antibodies (mAbs) with possible neutralising capabilities, such as mAb 447 (Zolla-Pazner *et al*, 2004), that recognise both the GPGR and GPGQ motifs, might play a crucial role in developing a vaccine reactive against all HIV-1 groups and subtypes (Gaschen *et al*, 1999; Gorny *et al*, 2004; Zolla-Pazner *et al*, 2004).

#### 1.2.3.3.4 The *pol* gene

The *pol* region of the HIV-1 genome is highly conserved amongst HIV-1 groups and subtypes. This gene encodes for the enzymes IN, RT and PR. The functions of these enzymes in the viral life cycle are explained in section 1.2.3.2. Excessive mutations in these regions would hamper the ability of the virus to replicate in its host cell. This is why many HIV-1 antiretroviral drugs have been aimed at inhibiting the function of these viral enzymes (Cornelissen *et al*, 1997; Lindström and Albert, 2003). *Pol* mutations occur as a result of selection pressure caused by certain inhibiting PR and RT drugs. These drugs include nucleoside / nucleotide RT inhibitors (NRTIs), non-nucleoside RT inhibitors (NNRTIs) and PR inhibitors (PIs) (Johnson *et al*, 2003). NRTIs are analogues of the body's own nucleoside or nucleotide molecules and act as alternative substrates for DNA polymerases. NNRTIs are a set of drugs which binds and physically interacts with the RT enzyme of HIV-1. Most of the current antiretroviral treatment (ART) drugs attempt to stop viral replication by inhibiting the RT gene, stop virus maturation by inhibiting the PR gene or attempt to stop the virus from entry into the host cell. ART has led to the reduction of opportunistic infections, an increased life span and an improved quality of life in many HIV-1 infected

individuals. Mutations can often lead to the failure of ART in patients infected with HIV-1 (Carr and Cooper, 1996; Cornelissen *et al*, 1997; Lindström and Albert, 2003).

#### **1.2.4 Envelope protein glycosylation patterns**

Post-translational modifications, such as acetylation, glycosylation and phosphorylation of HIV-1 RNA transcripts play an important role in viral transport and maturation (Ratner, 1992). Gag proteins and HIV-1 accessory proteins are known to undergo acetylation and phosphorylation (Henderson *et al*, 1992). Acetylation is the addition of an acetyl group to an organic compound, phosphorylation the addition of a phosphate group. Viral Env proteins undergo glycosylation, the addition of saccharides to proteins and lipids, as described below.

The most common and best studied glycosylation pattern is N-linked glycosylation, where oligosaccharides are uniquely added to asparagine (N) in the pattern of N-X-[S or T]. X is any amino acid followed by serine (S) or threonine (T) (Marshall, 1974). Another type of glycosylation is O-linked glycosylation. This pattern involves either simple oligosaccharide chains or glycosaminoglycan chains, where a carbohydrate is covalently linked to a hydroxyl group of S or T. O-linked glycosylation signals are more difficult to predict in protein sequences than N-linked sites (Blom *et al*, 1999; Chackerian *et al*, 1997; Hansen *et al*, 1998).

Glycosylation patterns influence protein folding (Hebert *et al*, 1997; Land and Braakman, 2001; Slater-Handshy *et al*, 2004), as well as protein confirmation (Meunier *et al*, 1999). The HIV Env gp120 protein is amongst the most heavily glycosylated proteins in nature (Myers and Lenroot, 1992). The number of glycosylation sites in the HIV Env protein does not necessarily increase over time, but varies extensively in both HIV and SIV infected individuals facilitating immune escape (Ye *et al*, 2000; Zhang *et al*, 2004).

Glycosylation pattern changes in the Env gp41 transmembrane protein also induce conformational changes in the Env gp120 surface protein. This dramatically diminishes the binding capacity of many gp120-specific antibodies (Si *et al*, 2001). The conformational changes have a huge influence on receptor binding and the



phenotypic properties of viruses (Ogert *et al*, 2001; Pollakis *et al*, 2001). Glycosylation of variable loops, such as the V3 loop, often restricts access to conserved host receptor binding sites. This limits their exposure to the host immune system and HIV Env has been described as having a glycan shield protecting it (Wyatt and Sodroski, 1998; Wei *et al*, 2003). The range of glycosylation patterns observed in different HIV-1 subtypes is very broad with patterns often overlapping between conserved sites in different subtypes (Gao *et al*, 1996; Zhang *et al*, 2004).

### **1.2.5 Consensus sequences and conserved genomic regions**

The human immune response generates antibodies against the exposed Env proteins of HIV (Cano *et al*, 2004; Horal *et al*, 1991; Hunter, 1997). Different mAbs have been identified that neutralise HIV-1 isolates from different genetic subtypes (Burton and Montefiori, 1997; Burton *et al*, 1994; Trkola *et al*, 1998, Moore *et al*, 2001). Many of these molecules with conserved antigenic features are poorly immunogenic (Moore *et al*, 2001). Conserved features amongst different HIV-1 subtypes in the Env glycoprotein can play a crucial role in the development of an effective HIV-1 vaccine (Burton *et al*, 2004). Antibodies recognise protein structures, not DNA sequences, and similar structural features amongst different genetic subtypes can be used to develop an efficient HIV-1 immunogen. Sequences that are similar usually translate into protein products that share common features. It has been suggested that consensus sequences and conserved genomic regions be used in vaccine development to overcome the high genetic diversity of HIV-1. Conserved structures might be expressed in possible vaccine antigens aimed at inducing broadly reactive immune responses (McKinney *et al*, 2004; Moore *et al*, 2001).

### **1.2.6 HIV diversity: A global pandemic**

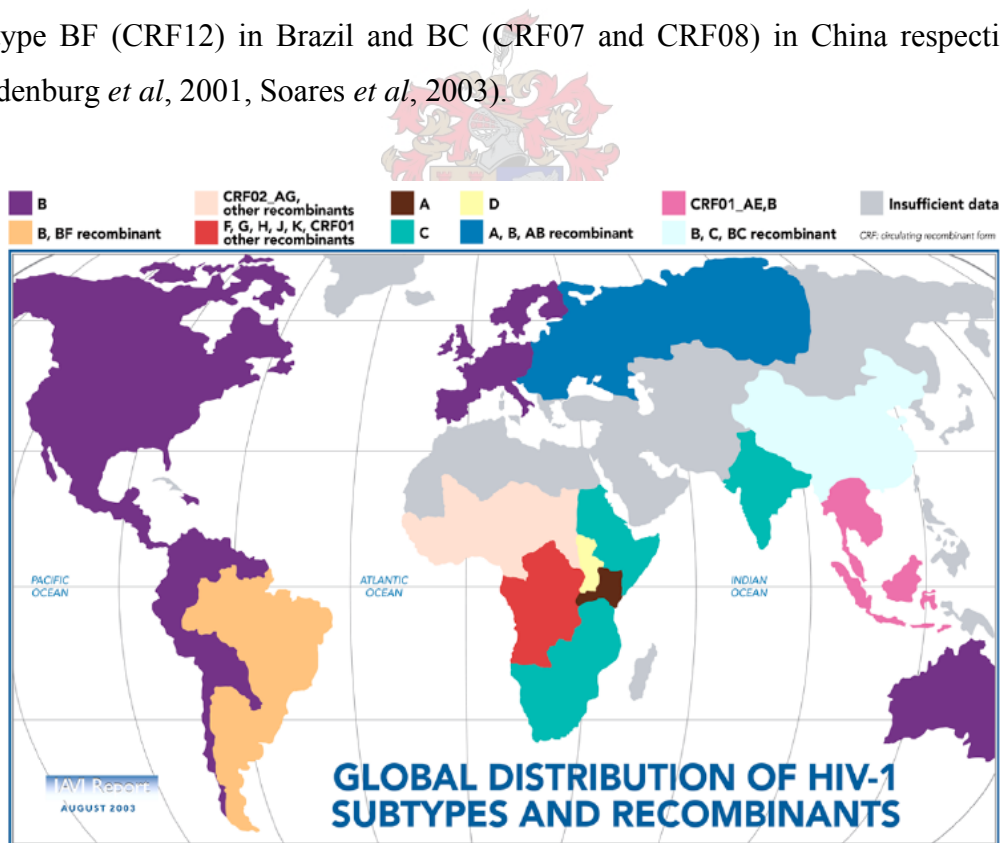
HIV forms part of the *Retroviridae* family (genera *Lentivirus*) based on genomic sequences and phylogenetic comparisons (Sonigo *et al*, 1985). The high degree of diversity of HIV is mainly caused by the high error rate and lack of proofreading ability of RT. The error-prone RT enzyme is also responsible for a phenomenon called hypermutations. Hypermutations result when an excessive number of

substitutions in DNA bp, usually from G → A, occur. They are often induced by host cellular defense mechanisms to produce replication-incompetent viruses (Fitzgibbon *et al*, 1993; Mangeat *et al*, 2003; Rose and Korber, 2000) and are not restricted to HIV (Wain-Hobson *et al*, 1995; Ngui *et al*, 1999). Diversity is further increased by the short replication time of HIV, which results in the fast turnover of new viruses within its human host (Coffin, 1995; Spira *et al*, 2003). Recombination events also contribute to the diversity of these viruses (Robertson *et al*, 1995). Dual infections by genetically diverse viruses have been linked to higher levels of viral replication and faster depletion of CD4+ T-cells. Infections with multiple HIV-1 variants have been associated with faster disease progression (Sagar *et al*, 2004).

The current proposed HIV nomenclature can be found on the Los Alamos National Laboratory (LANL) website (<http://www.hiv.lanl.gov/content/hiv-db/HelpDocs/subtypes-more.html>). Although the genomic organisation of HIV-1 and HIV-2 is similar, they only share about 40 percent nucleotide similarity, with HIV-2 closer related to SIVs (Hirsch *et al*, 1989; Bock and Markovitz, 2001). HIV-2 is predominantly found in West Africa and is much less pathogenic in humans than HIV-1. HIV-1 groups N and O are rare and the degree of their diversity have not yet been differentiated through phylogenetic analysis. However, group N seems to be phylogenetically equidistant from groups M and O (Robertson *et al*, 2000, Spira *et al*, 2003). Group M, responsible for the majority of HIV-1 infections worldwide, has been divided into nine different subtypes (A-D, F-H, J,K) and at least sixteen circulating recombinant forms (CRFs), with new unique recombinants continuously being identified. The formerly designated subtypes E and I have now also been classified as CRFs (Osmanov *et al*, 2002). Within HIV-1 group M subtypes A and F, closely related subclusters have also been identified. They are designated subtypes A1, A2, F1 and F2 respectively (Thomson *et al*, 2002). Subtypes B and D can also be considered subclusters of each other. However, due to historical reasons and previously published work, their original designations have been retained (Thomson *et al*, 2002). In Figure 1.6 the current global distribution trend of HIV-1 group M can be seen. HIV-1 group M subtype C is currently the most prevalent, while subtype B

is widely spread over the continents (Esparza and Bhamarapavati, 2000; Osmanov *et al*, 2002). HIV-2 has also been subdivided into eight subtypes (A-H) based on phylogenetic analysis (Robertson *et al*, 2000; Damond *et al*, 2004).

Africa seems to be the epicenter of HIV diversity, as all subtypes circulate on this continent. HIV diversity in Africa is not behaviourally linked, as many subtypes occur within different risk groups (Neilson *et al*, 1999). However, globally subtype C is mainly spread via heterosexual exposure, especially in southern Africa and India. Although subtype B is not the most prevalent subtype, it is the most widespread, especially in Europe and North America. In some countries certain HIV subtypes are more commonly found in high risk groups. For example, in Thailand subtype AE (CRF01) is commonly found in IDUs, while subtype B occurs more often in the heterosexual population (Nguyen *et al*, 2002; Mastro *et al*, 1997; Tovanabutra *et al*, 2003). In other countries, such as Brazil and China, CRFs are common, for example subtype BF (CRF12) in Brazil and BC (CRF07 and CRF08) in China respectively (Rodenburg *et al*, 2001, Soares *et al*, 2003).



**Figure 1.6: Global distribution of HIV-1 group M subtypes and recombinants.** The diagram shows the distribution and not the prevalence of HIV-1 subtype variants. The majority of the subtypes and recombinant forms are prevalent in Africa, subtype B in the Americas, Europe and Australia and subtype C in sub-Saharan Africa, Ethiopia and India (IAVI, 2003).

### 1.2.6.1 Subtype C and its recombinants

There has been much focus on the development of a subtype C candidate vaccine for Southern Africa, as this is the major subtype found in this geographical area (Novitsky *et al*, 2002; Van Harmelen *et al*, 2003; Williamson *et al*, 2003). The most common ancestor of HIV-1 subtype C dates back to the late 1960s. This is consistent with the theory that HIV-1 group M originated in the 1930s (Travers *et al*, 2004). Subtype C was first discovered in North East Africa in the early 1980s (Salminen *et al*, 1996).

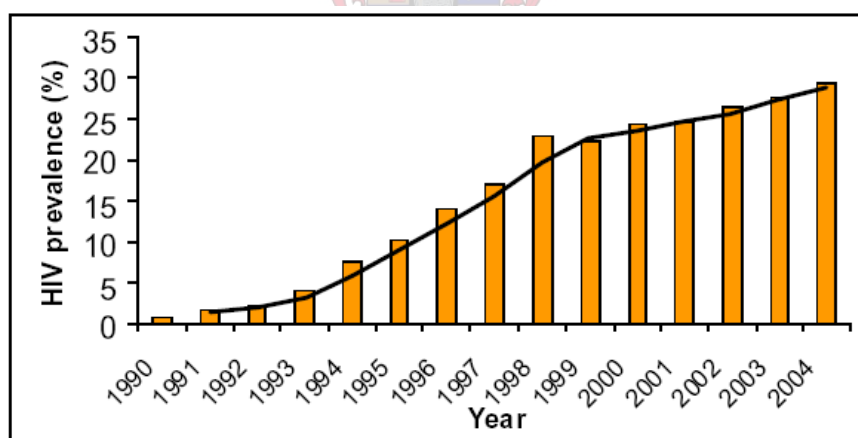
The first documented case of subtype C comes from a sample taken from a Malawian patient in 1983 (McCormack *et al*, 2002). The epidemic gradually spread to Sub-Saharan Africa from North East Africa (Gordon *et al*, 2003; Novitsky *et al*, 1999; Van Harmelen *et al*, 1999a; Van Harmelen *et al*, 1999b). The virus has also become the most dominant HIV-1 subtype in East and Central Africa (Neilson *et al*, 1999; Renjifo *et al*, 1998; Vidal *et al*, 2000b). There have been reports of subtype C in numerous countries, such as Russia (Bobkov *et al*, 1997; India (Shankarappa *et al*, 2001), China (Yu *et al*, 1998) and Brazil (Soares *et al*, 2003). In Ethiopia and India subtype C variants with intersubtype recombination have been characterised (Lole *et al*, 1999; Pollakis *et al*, 2003). In China BC recombinant strains CRF07 and CRF08 are predominant (Piyasirisilp *et al*, 2000; Rodenburg *et al*, 2001; Su *et al*, 2000). Three CD recombinant strains from Tanzania have also been classified as CRFs (CFR10) (Koulinski *et al*, 2001). It is likely that with the rapid expanding subtype C epidemic more subtype C recombinant strains, as well as other complex recombinant forms, will be identified in the near future.

HIV-1 subtype C has very unique genetic characteristics which distinguishes it from other HIV-1 subtypes. These include the presence of extra NF- $\kappa$ B enhancer copies in the LTR, Tat and Rev prematurely truncated proteins and a 15 bp insertion of the 5' end of the *vpu* reading frame (Huang *et al*, 2003; Peeters and Sharp, 2000). Subtype C also has a relatively conserved *env* gp120 V3 loop, with the virus showing preference to using CCR5 as its major co-receptor despite the stage of disease progression (Peeters and Sharp, 2000; Shankarappa *et al*, 2001). Some hypothesise that differences seen in the LTR promoter may be responsible for this rapid expansion

of subtype C (Jeeninga *et al*, 2000; Montano *et al*, 1997; Zacharova *et al*, 1997). The efficiency by which subtype C is transmitted from one person to another has been suggested as a contributing factor to subtype C predominance (Ariën *et al*, 2005; Ball *et al*, 2003; Chen *et al*, 2000). However, subtype C does not have a higher fitness level, defined as an organism's replicative capacity or adaptive ability in a given environment (Domingo and Holland, 1997), compared to the other HIV-1 subtypes (Ariën *et al*, 2005).

### 1.2.7 The HIV-1 epidemic in South Africa

The National HIV-1 prevalence trends in South Africa are presented in Figure 1.7. The survey, recorded since 1990, is conducted amongst pregnant women attending antenatal clinics across South Africa. There are currently an estimated 5.6 million (29.5%) South Africans infected with HIV-1. The prevalence rate has risen from 26.5% in 2002 to 27.9% in 2003 to its current 29.5%. The highest rates are seen in KwaZulu-Natal (40.7%) and the lowest in the Western Cape (15.4%) (Department of Health, 2005).



**Figure 1.7: Prevalence of HIV-1 among antenatal care attendees in South Africa, 1990-2003.** The prevalence rate has risen from 0.7% in 1990 to 29.5% in 2004 (Department of Health, 2005).

#### 1.2.7.1 HIV-1 diversity in South Africa

The first HIV-1 cases in South Africa were reported in 1982 (Ras *et al*, 1983) and the virus was isolated for the first time in the country in 1984 (Becker *et al*, 1985). In the

1980s the HIV-1 epidemic in South Africa was dominated by HIV-1 subtypes B and D, associated with the homosexual population (Engelbrecht *et al*, 1995; Sher, 1989). This has been replaced by the fast spreading subtype C epidemic more commonly found in the heterosexual population (Van Harmelen *et al*, 1997; Van Harmelen *et al*, 1999a). In the Western Cape, depending on the patient group sampled, non-subtype C strains still account for 1-10% of documented cases (Engelbrecht *et al*, unpublished). Non-subtype C and recombinant HIV-1 strains have now also been identified in South Africa (Papathanasopoulos *et al*, 2003), as well as recorded by our own observations (Department of Medical Virology, Tygerberg Campus, University of Stellenbosch). Previous studies have shown that the majority of the subtype C viruses in South Africa use CCR5 as their major chemokine co-receptor and it was believed that CXCR4 strains were non-existent (Bjorndal *et al*, 1999; Treurnicht *et al*, 2002). More recent studies have found that certain subtype C viruses switch to SI variants as in other HIV-1 subtypes (Cilliers *et al*, 2003). CXCR4 using strains have now also been identified in South Africa (Janse van Rensburg *et al*, 2002).

#### **1.2.7.2 HIV-1 and its social and economic impact on South Africa**

HIV/AIDS is the leading cause of death in South Africa (Dorrington, 2001; Hosegood *et al*, 2004). Life expectancy has dropped and adult deaths are rising due to HIV-related diseases (Kapp, 2004). South Africa is still a developing country and HIV/AIDS is currently the most significant threat to slow down the economic development process (Allen *et al*, 2000; Rosen *et al*, 2004). One major contributing factor might be the lower levels of HIV/AIDS awareness in South Africa, compared to other developed and developing countries (Morris and Williamson, 2001). Misinformed youth often still engage in unsafe sexual practices, despite the risk of contracting HIV/AIDS (Eaton *et al*, 2003; Myer *et al*, 2002). Women are also often victims of gender-based violence, such as rape, increasing their risk of contracting the disease (Dunkle *et al*, 2004; Pettifor *et al*, 2004; Wood and Jewkes, 1997). In sub-Saharan Africa, 58% of HIV-1 infected adults are female. This means that women residing in Africa are the most severely affected by HIV/AIDS (UNAIDS, 2004). Other risk factors include promiscuous sex with multiple partners, poverty, the migrant labour system, the practice of commercial sex, lack of formal education, the traditional status of women in their communities, stigmatisation and discrimination

(Abt. Associates Inc., 2000; Allen *et al*, 2000). People need to change their social attitude towards HIV/AIDS and its victims if any impact on the high levels of prevalence rates is to be made in the near future.

### 1.2.7.3 Addressing the problem in Khayelitsha

The Khayelitsha Township (Figure 1.8) was established in 1983 (De Tolly and Nash, 1984). It is a huge, mostly informal settlement located approximately 30 km outside the Cape Town city center. The Xhosa word ‘Khayelitsha’ itself means ‘new home’ or ‘new beginning’. The current population is estimated at 400 000 people; however, this figure varies from 350 000 to 900 000 (Dorrington, 2002). It is an overcrowded township characterised by poverty and often associated with violence (Wood and Jewkes, 1997). Unemployment rates are high and those who are employed are either casual or domestic workers only employed on a temporary basis, earning a basic salary (Muzondo *et al*, 2004).



**Figure 1.8: Khayelitsha, Western Cape** ([www.aerialeye.co.za/khay01.jpg](http://www.aerialeye.co.za/khay01.jpg)). The majority of the township houses are small, self-made shacks. They are often overcrowded, without adequate sanitation and no clean running water.

Khayelitsha currently has the second highest HIV-1 prevalence rate (27.2%) amongst women attending antenatal clinics in the Western Cape, surpassed only by Gugulethu / Nyanga (28.1%) (Department of Health, 2004). Khayelitsha also has the highest Tuberculosis (TB) incidence in the province (20%) (Department of Health, 2002). TB is the second most common opportunistic infection after oral candidiasis and the most common cause of death in HIV-1 infected patients (Department of Health, 2002).

This has led to attempts to integrate HIV-1 and TB services at important clinical sites, such as Khayelitsha, to reduce the health risk and improve the quality of service and treatment received (WHO, 2004). The settlement has long been regarded as a high health risk area with poor nutritional status (Bohm, 1996; Le Roux and Le Roux, 1991), aggravated by prevailing poor sanitation (Fincham, 2004). The measles vaccination campaign was the first effort of a mass vaccination project in this area to try and improve the general health of the community (Berry *et al*, 1991; Coetzee *et al*, 1990).

Despite the launch of AIDS prevention campaigns in informal sector shops as early as 1991 (Marks and Downes, 1991), the HIV-1 prevalence in Khayelitsha still continued to rise in the 1990s. In 1999 a programme was launched to try and prevent mother-to-child transmission (MTCT) at 2 midwife obstetric units with limited resources (Abdullah *et al*, 2001; Chopra *et al*, 2002). In developing countries trials of short course ART have demonstrated drastic reductions in MTCT (Guay *et al*, 1999; Preble and Piwoz, 2001). In association with the local provincial government and *Médecins Sans Frontières* (MSF, “Doctors without Borders”) dedicated services to adults and children living with HIV-1 were established in 2000. This programme was extended in 2001 to offering free ART in the community to those who qualified and could not afford treatment on their own (MSF, 2003; WHO, 2004). Although it is still too early to say if the campaign was a success, positive milestones have been reached. At this stage 95% of all pregnant women are being tested for HIV-1 and receive counseling. By April 2004 more than 1000 people were registered on ART (MSF, 2003).

The Treatment Action Campaign (TAC) of South Africa and MSF are also running a joint treatment literacy programme entitled Project Ulwazi (Knowledge). The programme looks at increasing literacy and knowledge to raise HIV/AIDS awareness in townships, such as Khayelitsha. These programmes attempt to influence the behaviour of the entire community towards HIV/AIDS (TAC, 2005; WHO, 2004). With the joint effort of the government and community success stories might become a reality, not only in Khayelitsha, but throughout South Africa.



## 1.2.8 HIV-1 characterisation techniques

A brief literature review on the principles of the methods used during the study is presented here. The precise methods and assays used are presented in chapter two.

### 1.2.8.1 HIV-1 viral load assays

The HIV-1 RNA level in HIV-1 positive patients is clinically important for evaluating the efficacy of ART and monitoring disease progression (Mellors *et al*, 1997; Swanson *et al*, 2005). The viral load can be measured either through RT – Polymerase Chain Reaction (RT-PCR), the isothermal nucleic acid based amplification assay (NASBA) or by the branched DNA (bDNA) signal amplification assay. All these techniques are dependant on the amplification of HIV-1 with sequence specific primers and / or probes (Swanson *et al*, 2005). They are incorporated into viral load assays, such as the VERSANT<sup>®</sup> HIV-1 RNA 3.0 (bDNA) assay (Bayer Diagnostics, Tarrytown, New York, USA), Amplicor HIV-1 Monitor<sup>®</sup> v1.5 RT-PCR test (Roche Diagnostics, Mannheim, Germany), the NASBA NucliSens<sup>®</sup> HIV-1 QT assay (Biomérieux, Inc., Durham, North Carolina, USA) and the LCx<sup>®</sup> HIV RNA Quantitative assay (Abbott Laboratories, Illinois, USA). In our laboratory (Department of Medical Virology, Tygerberg Campus, University of Stellenbosch) the LCx<sup>®</sup> HIV RNA Quantitative assay based on RT-PCR amplification is used, as this method is sensitive, highly accurate and repeatable (Johanson *et al*, 2001; Zanchetta *et al*, 2000). The assay has been shown to perform better with genetically diverse HIV-1 strains than other available viral load assays (Swanson *et al*, 2005). In areas with poor resources easier, more economical and practical methods, such as dried blood spots (DBS) might be used in the future to determine the concentrations of HIV-1 RNA samples. The technique allows DBS or plasma to be saturated and absorbed onto filter paper for long-term storage. (Alvarez-Munoz *et al*, 2005; Cassol *et al*, 1991; Cassol *et al*, 1997; Mwaba *et al*, 2003).

### 1.2.8.2 The *env* gp120 V3 serotyping assay

The competitive enzyme linked immunosorbent serotyping assay (cPEIA) is based on the *env* gp120 V3 amino acid sequences and uses the antigenic rather than genetic

properties of HIV-1 by detecting type-specific antibodies against HIV-1. An ideal V3 cPEIA should be able to distinguish between all HIV groups and subtypes. Serotyping analysis is not always accurate and can be misleading (Apetrei *et al*, 1998; Barin *et al*, 1996; Plantier *et al*, 1998). HIV-1 peptides can be cross-reactive to the different HIV serotypes, making analysis difficult (Barin *et al*, 1996; Plantier *et al*, 1998). Peptides from subtype A and C, and B and D have been shown to be cross-reactive with each other. Subtype E, which has now been designated as an AE recombinant strain CRF01\_AE (Carr *et al*, 1996; Gao *et al*, 1996; Nguyen *et al*, 2002), also has cross-reactive capabilities with subtype A (Barin *et al*, 1996; Plantier *et al*, 1998).

### 1.2.8.3 HIV-1 genotyping

Characterising HIV-1 strains through genotyping has important implications for HIV-1 vaccine development. Genotyping can identify and keep track of new emerging HIV-1 variants. These new variants might have either increased, or reduced virulence and should thus be closely monitored. Through HIV-1 genotyping conserved as well as unique features can be identified in various HIV-1 strains (Moore *et al*, 2001). HIV-1 genotyping, as with any other sequences from other organisms, should ideally be based on full-length genomic sequences or at least complete gene areas to be absolutely reliable (Salminen *et al*, 1995). However, complete full-length genome amplification of all study samples are not always possible and are far more difficult to perform. Therefore, certain smaller genomic regions or genes are usually targeted for analysis (Carr *et al*, 1998). During this study the *gag* p24, *env* gp41 IDR, *env* gp120 V3 and a part of the *pol* gene region were targeted for HIV-1 genotyping analysis.

These regions were chosen as they form part of diagnostically important HIV-1 antigen and antibody screening assays, as well as several viral load assays (Swanson *et al*, 2003). These include the *gag* p24 antigen detection assays supplied by Roche Diagnostics (Mannheim, Germany) and certain enzyme immunoassays (EIAs), such as the Less-sensitive EIA and the Vironostika<sup>®</sup> HIV-1 EIA (Biomérieux, Inc., Durham, North Carolina, USA; Parekh and McDougal, 2005). Mutations in these regions can alter the sensitivity of the assays in use. The detection of HIV-1 RNA or *gag* p24 antigen prior to the development of antibodies usually indicates a very recent

HIV-1 infection or pre-seroconversion. Antibodies to Gag (p24 and p17) and Env (gp120 and gp41) proteins are usually detected early and get stronger over time compared to the *pol* gene products. Theoretically assays that include these regions should be able to detect all HIV-1 positive individuals despite their time of seroconversion. Within the Env proteins, the *env* gp41 IDR antibodies are elicited early, while antibodies to the *env* gp120 V3 only develop later (Parekh and McDougal, 2005). The *gag* p24 and *env* gp41 IDR regions were also chosen to increase the chances of finding possible HIV-1 recombinant strains (Swanson *et al*, 2003). The *env* gp120 V3 region was used to compare molecular serotyping and genotyping methods and served as an extra gene fragment on which molecular analysis could be performed. The *pol* gene fragment was used to help characterise more complicated HIV-1 strains. Mutations in the *pol* RT and PR genes are important to monitor, as they can lead to HIV-1 drug resistance and ART failure (Hirsch *et al*, 2000; Lindström and Albert, 2003). Molecular genotyping methods used to characterise the Khayelitsha cohort include PCR, DNA cloning, DNA sequencing and phylogenetic analysis. These are described in detail in the sections below.

#### 1.2.8.4 Nucleic acid extraction

Nucleic acids (RNA and DNA) are extracted by releasing them from the cells in which they are found, cell lyses, and deproteinising them. The most common method used is the phenol / chloroform extraction method. Phenol and chloroform denature proteins and solubilise the nucleic acids to obtain maximum yields (Ausubel *et al*, 2003; Kirby, 1957; Palmiter, 1974; Pennman, 1966; Sambrook *et al*, 1989). The nucleic acids can be purified with ethanol to remove the excess chloroform and phenol.

Ethanol causes a structural transition in nucleic acids, which stabilises the DNA (Eickbush and Moudrianakus, 1978). Newer methods of nucleic acid extraction are based on silica membrane spin protocols (Vogelstein and Gillespie, 1979). In the presence of a chaotropic (chaos-forming) salt, DNA binds to a silica membrane present inside a spin column. Chaotropic salts have the ability to disrupt hydrogen bond structures in water. They denature proteins by interfering with their hydrophobic interactions (Hamaguchi and Geiduschek, 1962). After purification,

dissociation of the DNA from the membrane can be achieved with water or a low salt buffer, such as TE buffer [10 mM tris (hydroxymethyl) methylamine-Chloride (Tris-Cl); 1 mM ethylene diamine tetra-acetic acid (EDTA)]. DNA is stable at 4°C and can be stored for prolonged periods at this temperature. RNA purification requires additional enzymatic steps, such as treatment with deoxyribonuclease, to remove DNA. RNA is easily degraded by thermostable RNase enzymes, which are present on fingertips and in dust (Chomczynski, 1992), and have to be frozen in order to keep its stability (Ausubel *et al*, 2003; Sambrook *et al*, 1989).

#### **1.2.8.5 The polymerase chain reaction**

The PCR method was developed in 1985 by Kary B. Mullis (Mullis and Faloona, 1987). A PCR amplification involves concurrent steps of DNA heat denaturation, primer annealing and DNA extension. These steps are repeated several times during PCR cycling (Saiki *et al*, 1988). A specific DNA region is targeted for amplification with oligonucleotide primers that are complementary to sequences that flank the segment of interest. The first protocols for PCR used the Klenow fragment of *E.coli* DNA polymerase 1 for the extension and amplification of targeted DNA (Mullis *et al*, 1986; Mullis and Faloona, 1987; Saiki *et al*, 1988). The polymerase was inactivated during heat denaturations, which lead to the failure of many attempted PCRs. This has now been replaced by more heat stable DNA polymerase enzymes, such as *Thermus aquaticus* (*Taq*) DNA polymerase (Chiën *et al*, 1976). This greatly reduces mispriming events at elevated temperatures (Ausubel *et al*, 2003; Sambrook *et al*, 1989).

#### **1.2.8.6 DNA cloning**

Molecular cloning is the principle by which foreign DNA, such as a PCR product, can be inserted into a specific DNA vector. These vectors are usually circular double-stranded DNA plasmids found in many bacterial species (Ausubel *et al*, 2003; Sambrook *et al*, 1989). Plasmids behave as accessory genetic units that replicate independently of the bacterial chromosome. They mostly contain genes that are advantageous to the host bacteria and confer many different phenotypes. These include the production of antibiotics, degradation of complex organic compounds and

expression of restriction enzymes. (Sambrook *et al*, 1989). The method was first shown to be potentially useful when Cohen and his colleagues demonstrated that biologically functional foreign DNA can be inserted into *E.coli* vectors (Cohen *et al*, 1973). The highest cloning efficiency is achieved with directional cloning. This method produces non-complementary overhangs, also known as sticky ends, which can be cleaved by two different restriction enzymes. Cloning with blunt end DNA products is more difficult, as DNA ends are compatible and can religate with each other. The choice of plasmid / vector is very important, as certain bacterial strains can inhibit or interfere with the reproduction of foreign DNA (Bertani and Weigle, 1953; Murray *et al*, 2001).

#### **1.2.8.7 DNA sequencing**

In modern times DNA sequencing has probably become the most powerful tool for characterising the genomes of different organisms. The characterisation of the human genome (Human Genome Project) has led to many arguments and ethical question-marks. The information obtained from this and other genome projects has the potential to answer many health-related questions, such as the possibility of treating genetically related diseases transmitted from parent to child. This is no different to HIV and DNA sequencing. Sequence analysis can help clarify many questions of modern day HIV biology, such as origin, epidemiology, subtyping, as well as cross-species transmission of HIV (Rodrigo and Learn, 2001).

DNA sequencing is a PCR-based method by which the exact base pair (bp) sequence of a certain DNA fragment being investigated can be revealed. These sequencing reactions are based on the earlier enzymatic method of Sanger *et al* (1977) and the chemical degradation method of Maxam and Gilbert (Maxam and Gilbert, 1992) both resulting in chain termination of the oligonucleotide fragments. Sequencing reactions incorporate both deoxyribonucleoside triphosphates (dNTPs) and dideoxyribonucleoside triphosphates (ddNTPs). A ddNTP incorporation into the DNA fragment results in the DNA chain being terminated. This results in various lengths of DNA strands that can be distinguished from each other. Each of the four dNTPs [Adenine (A), Guanine (G), Cytosine (C), Thymidine (T)] are labeled with different fluorescent dyes for easy recognition.

Today, most sequencing reactions are carried out using automated machines and computers, such as the ABI Prism<sup>®</sup> Genetic Analyzer (Applied Biosystems, Foster City, California, USA). These machines use a polymer in an electrophoresis capillary column in which DNA fragments are separated according to size. A laser detects the different dNTP dyes as they pass through a capillary and generates a visible computer electrophenogram converting the termination signals into peaks that can be easily analysed. The ddNTP connected to a particular length strand, correlates to the dNTP of a particular position in the sequence (Swerdlow and Gesteland, 1990). The discovery of DNA sequencing formed the basis for detailed gene and genome analysis. Having the sequence of a particular DNA strand is only the start and is the foundation on which phylogenetic analysis can be based.

#### **1.2.8.8 Phylogenetic analysis**

Detailed reviews on phylogenetic analysis are presented by the following authors: Page and Holmes, 2002; Nei and Kumar, 2000; Salemi and Vandamme, 2003; Rodrigo and Learn, 2001. A brief summary is presented here.

##### **1.2.8.8.1 What is phylogenetic analysis?**

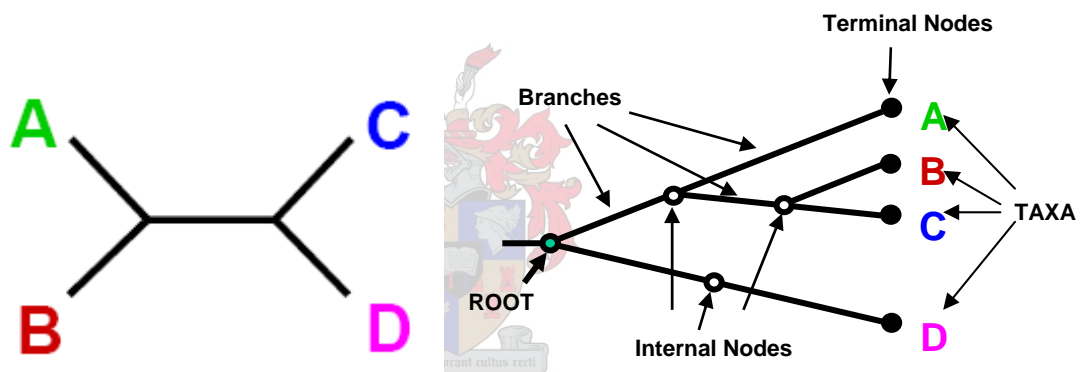
Organisms are classically identified and characterised based on their phenotypic properties. With the expansion of molecular techniques, phylogenetic analysis has become useful in studying organisms at molecular level based on their genome. It allows us to determine the relationship between certain organisms based on assumptions of evolution with accuracy and confidence. A phylogeny can be described as a set of relationships amongst groups of genes or organisms that reflect their evolutionary history based on their DNA and / or protein sequences. Phylogenetic analysis consists of the generation of a multiple alignment, testing the alignment with different evolutionary models and creating a phylogenetic tree, which best describes the data under investigation.

#### 1.2.8.8.2 Multiple alignments and phylogenetic trees

The basis of phylogenetic analysis is to compare similar sequences with each other. This is done by creating multiple alignments with the sequences in question. The sequences in an alignment are computationally compared with each other. The evolutionary relationship based on the model chosen is then illustrated by means of a phylogenetic tree. The tree indicates which group of sequences compared, being that of a gene or organism, has the closest relationship with each other. The generation of alignments is one of the most common tasks in computational sequence analysis. This is because alignments can be used in analysis, such as structure prediction or simply to demonstrate sequence similarity within a family of sequences (Salemi and Vandamme, 2003). The percentage similarity is calculated by simply counting the amount of identical nucleotides or amino acids relative to the length of the sequences (Salemi and Vandamme, 2003). Sequences have different lengths with different coding regions and gaps have to be inserted or shifted in some positions to achieve the optimal alignment (Goldman and Yang, 1994; Muse and Gaut, 1994). Similarity plots based on sequence alignment similarities can be useful in determining breakpoints in recombinant viruses (Lole *et al.*, 1999). The amino acid sequence similarity or degree of variability can be expressed as entropy values (Korber *et al.*, 1994). This is defined as the measure of variability at each amino acid position through a column in an alignment. The entropy value takes into consideration both the variety and frequency of observed amino acids at each aligned position.

A phylogenetic tree consists of nodes and branches (Figure 1.9). The nodes represent the taxonomic units and the branches the relationships between these units. More distantly related taxonomic units have bigger branch lengths. External nodes are the taxa or sequences which are being compared, while internal nodes represent a common ancestor between two or more taxa. The root of a tree is the common ancestor of all the taxa being analysed. An unrooted phylogenetic tree positions the individual taxa relative to each other without indicating the direction of the evolutionary process. If the direction of evolution or common ancestor is known, the tree can be rooted with these sequences.

Tree making methods include distance-matrix methods and discrete data methods. Each method uses different mathematical equations to best describe the sequences being analysed. In distance matrix methods, such as the unweighted pair group method with arithmetic mean (UPGMA, Sneath and Sokal, 1973), neighbour-joining (Saitou and Nei, 1987) and Fitch-Margoliash (Fitch and Margoliash, 1967) methods, the aligned sequences are converted into pairwise distance matrixes. The distances are expressed or calculated as the fraction of sites that differ between two sequences in a multiple alignment. Sequences with the closest distances are grouped closely together on the representative tree. The UPGMA searches for the smallest value in the pairwise distance matrix to construct a phylogenetic tree. The neighbour-joining method sequentially finds its closest neighbours based on the internal branch lengths, while the Fitch-Margoliash method evaluates all possible trees for the shortest overall branch length.



(A) Unrooted phylogenetic tree

(B) Rooted phylogenetic tree

**Figure 1.9: An example of an unrooted (A) and rooted (B) phylogenetic tree:** The branches, nodes, roots and taxa are indicated. In an unrooted phylogenetic tree only the relationship among the taxa is given and the direction of evolution is unknown. In a rooted phylogenetic tree the root represents the most common ancestor between different taxa. Branch lengths indicate how closely related these taxa are to each other (Adapted from Nei and Kumar, 2000).

Discrete data methods, such as maximum parsimony (Eck and Dayhoff, 1966), maximum likelihood (Felsenstein, 1981) and Bayesian methods (Rannala and Yang, 1996; Mau *et al*, 1999), consider each nucleotide site of the alignment directly in order to construct a phylogenetic tree that best accommodate all the sequence data. Maximum parsimony finds the tree topology that can be explained with the smallest



number of character changes. Maximum likelihood calculates the probability of expecting each possible nucleotide or amino acid in the ancestral nodes of the trees.

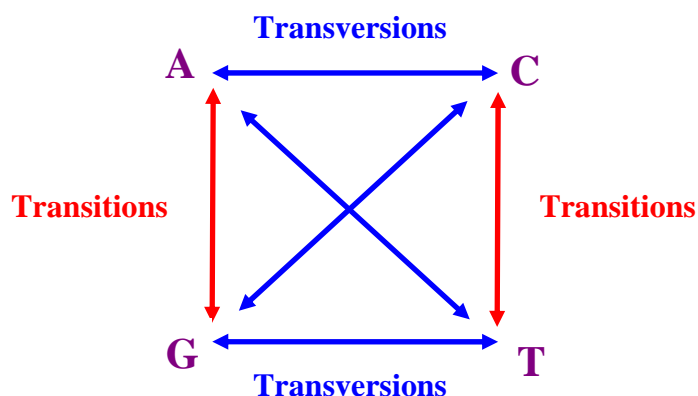
Maximum parsimony and maximum likelihood use the principles of stepwise addition and branch swapping to search for the best phylogenetic tree. Through stepwise addition branches are added in succession of each other at different levels on a phylogenetic tree. Each level is evaluated and the best tree chosen before the addition of the next branch continues. Branch swapping techniques allow for the pre-defined rearrangement of the phylogenetic tree branches. The most common branch swapping methods are tree bisection and reconstruction (TBR), subtree pruning and regrafting (SPR) and nearest-neighbour interchange (NNI). The TBR method is computer intensive. This method cuts the original phylogenetic tree created, usually through stepwise addition, and rejoins each branch to create a completely different topology from the original tree. The tree with the highest score is chosen as the phylogenetic output. SPR only clips subtrees from the original tree, whereas NNI only swaps one branch of a subtree with another (Page and Holmes, 2002; Salemi and Vandamme, 2003).

Bayesian analysis has been developed recently (Rannala and Yang, 1996; Mau *et al*, 1999) and is similar to that of maximum likelihood. Instead of seeking the tree that maximises the likelihood of observing data, Bayesian analysis seeks those trees with the greatest likelihood of the given data. Bayesian models searches for the best tree that is consistent with both the evolutionary model of choice and the data in an alignment.

#### **1.2.8.8.3 Models of evolution**

Evolution can be seen as the mutational changes of genes and gene function over time. Mutations are caused by events such as nucleotide substitutions, insertions / deletions (indels) and recombination. These mutant genes or homologue copies eventually spread through the population by genetic drift and / or natural selection (Hartl and Clark, 1997). Phylogenetic analysis makes assumptions about the process and rate of DNA substitutions or amino acid replacements in the model of evolution they employ (Felsenstein, 1973; Penny *et al*, 1992). Changes take place either

through transitions, when a purine DNA base (A, G) replaces another purine DNA base or a pyrimidine DNA base (C, T) replaces another pyrimidine DNA base or transversions, when a purine bp replaces a pyrimidine bp or *vice versa* (Figure 1.10).



**Figure 1.10: DNA substitution mutations.** Transitions, in red, occur when interchanges of purines (A ↔ G) or of pyrimidines (C ↔ T) occur. Transversions, in blue, occur when a purine bp replaces a pyrimidine bp or *vice versa*. There are twice as many possible transversions (8) compared to transitions (4), but due to the chemical composition of DNA bp transition mutations are more common than transversions (Adapted from Posada and Crandall, 2001).

The simplest mathematical and computer model of nucleotide substitution is the Jukes and Cantor model (Jukes and Cantor, 1969). This model assumes that each nucleotide has an equal chance to be replaced by any other during evolution. DNA bp are chemically different from each other and thus have different binding properties. Each organism also has a unique DNA composition. Therefore a substitution model for DNA evolution must be chosen carefully. Phylogenetic methods and analysis may become less accurate and inconsistent if the wrong model of evolution is assumed during data analysis (Bruno and Halpern, 1999; Huelsenbeck and Hillis, 1993; Posada and Buckley, 2004).

There are currently fifty-six described models of evolution (Posada and Buckley, 2004; Posada and Crandall, 2001). The most commonly used model for HIV DNA sequences is the Kimura 2-parameter model (Kimura, 1980). This model assumes that transitions generally occur more frequently than transversions do, as in the case of HIV (Hillis *et al*, 1994). The general time reversal model (GTR, Tavaré, 1986), which assumes six nucleotide transformation rates and the existence of nonequal base frequencies, is also a popular model commonly used in HIV phylogenetic

analysis (Salemi and Vandamme, 2003). Other models include the F81 model (Felsenstein, 1981), F84 model (Felsenstein, 1993), Hasegawa-Kishino-Yano (HKY) models (Hasegawa *et al*, 1985), the Kimura 3-parameter model (Kimura, 1981), the Symmetrical model (SYM; Zharkikh, 1994), the Tamura-Nei (TN93) model (Tamura and Nei, 1993), the Transition model (TIM; Posada and Buckley, 2004) and the Transversion model (TVM; Posada and Buckley, 2004). The remainders of the fifty-six models are variations of the models of evolution mentioned above. The models of evolution also employ the parameters of gamma-distribution (G), or rate variation, and invariant sites (I). These parameters consider that the substitution rates differ due to the functional constraints of amino acid residues. Invariant models take into account that certain sites never or rarely vary, or rather stay constant over time, such as conserved sites. The gamma-distribution rate calculates at which rate, how fast or how slow, a certain site is evolving from one sequence or amino acid to another (Gu and Zhang, 1997; Gu *et al*, 1995). The simplest model, which best explains the given data, should be used during analysis (Hillis *et al*, 1996).

Amino acids are degenerate and this means that not all nucleotide substitutions result in a change in amino acid. Non-synonymous substitutions correspond with amino acid replacements, while synonymous substitutions are silent with no amino acid change observed. Under neutral evolution, synonymous substitutions occur at a higher rate than nonsynonymous substitutions. This also needs to be considered when working with the different substitution models.

Many software packages are freely available on the internet to perform phylogenetic analysis with (example <http://hiv-web.lanl.gov/>). It is important to carefully choose the correct analysis tool to justify the findings of the specific study.

### **1.3 Aim of this study**

The first objective of this study was to investigate the molecular epidemiology of HIV-1 in Khayelitsha. This was done by identifying HIV-1 subtypes found in the study area. A second objective was to further characterise any unusual HIV-1 strains identified. This study will thus help elucidate whether the HIV-1 strains identified resemble those circulating in the rest of the Western Cape and South Africa.

## CHAPTER TWO

### 2. Materials and Methods

	<b>PAGE</b>
2.1 Introduction	35
2.2 Materials	35
2.2.1 Cohort samples	37
2.3 Methods	39
2.3.1 Sample preparation	39
2.3.2 HIV-1 viral load assay	40
2.3.3 <i>env</i> gp120 V3 serotyping assay	40
2.3.4 The polymerase chain reaction	41
2.3.5 Genotyping reactions	41
2.3.6 Agarose gel electrophoresis	42
2.3.7 Purification of PCR products	43
2.3.8 DNA concentration determination	44
2.3.9 DNA cycle sequencing reactions	44
2.3.10 Sequence and phylogenetic analysis	45
2.3.10.1 DNA sequence assembly	45
2.3.10.2 Sequence alignments and phylogenetic trees	46
2.3.10.3 Analysis of hypermutation variants	47
2.3.10.4 Consensus sequences and the identification of conserved genomic regions	47
2.3.10.5 Similarity plots	48
2.3.10.6 Viral phenotype prediction	48
2.3.10.7 Envelope N-linked glycosylation sites	52
2.3.10.8 Screening for drug resistant mutations	52
2.3.11 Cloning experiments	52
2.3.11.1 Preparation of samples for cloning	52
2.3.11.2 Cloning reactions	53
2.3.11.3 Selection of positive clones	54
2.3.12 Full-length genome analysis	54

## CHAPTER TWO

### 2. Materials and Methods

#### 2.1 Introduction

A total of 127 patient samples were serotyped and genotyped during this study. The techniques used to analyse these samples are described in this chapter. Serotyping was done by performing a cPEIA. The *gag* p24, *env* gp120 V3 and *env* gp41 IDR were amplified by nested PCR. The PCR products were purified and directly sequenced. Possible HIV-1 recombinant strains were cloned and further characterised based on a 1.2 kb *pol* gene fragment. Detailed sequence and phylogenetic analysis were performed on all sequences obtained.

#### 2.2 Materials

The materials used to characterise the Khayelitsha cohort samples are presented in Tables 2.1 to 2.3. In Table 2.1 the equipment needed to perform the necessary assays and analysis are presented. In Table 2.2 the commercial packages used are presented, while additional chemicals not provided by the commercial packages are presented in Table 2.3. The suppliers and catalogue numbers are provided where applicable. The ® and ™ designations indicate that the products are either registered or trademark properties of the suppliers they are provided from.

**Table 2.1:** Equipment used to perform sample assays and analysis

<b>Equipment</b>	<b>Supplier</b>	<b>Location</b>
ABI prism® 310 Genetic Analyzer	Applied Biosystems	California, USA
Abbott LCx® Probe system	Abbott Laboratories	Illinois, USA
Anthos Winread (Windows™ based)	Anthos Labtec.	Salzburg, Austria
AxSYM® Diagnostic system	Abbott Laboratories	Illinois, USA
Beckman Coulter Allegra™ 6R centrifuge	Beckman Inc.	Fullerton, California, USA
geneAMP® 9700 PCR system	Applied Biosystems	California, USA
Labcon® FSIE shaking incubator	Labcon Ltd	Krugersdorp, South Africa
Nanodrop™ ND 1000	Nanodrop Technologies Inc.	Delaware, USA
Speedvac™ SVC100	Savant Instruments Inc.	New York, USA
Syngene™ GeneGenius Computer system	Synoptics Ltd.	Cambridge, United Kingdom
Techne® Gene E Thermal Cycler	Techne Ltd.	Cambridge, United Kingdom

**Table 2.2:** List of commercial products and assays used

<b>Product</b>	<b>Company</b>	<b>Location</b>	<b>Catalogue number</b>
1 kb DNA ladder	Promega	Madison, Wisconsin, USA	G5711
Abbott LCx <sup>®</sup> HIV RNA Quantitative assay; Blank cells	Abbott Laboratories	Illinois, USA	8392-01
Abbott LCx <sup>®</sup> HIV RNA Quantitative assay; Reaction cells	Abbott Laboratories	Illinois, USA	09A48-01
Access RT PCR system *	Promega	Madison, Wisconsin, USA	A1250
Axsym <sup>®</sup> HIV Antigen and Antibody Combo	Abbott Laboratories	Illinois, USA	2G8320
BigDye <sup>™</sup> Terminator Cycle Sequencing Ready Reaction Kit	Applied Biosystems	Foster City, California, USA	0 211 005
Dye Ex <sup>™</sup> 2.05 Spin Kit	Applied Biosystems	Foster City, California, USA	63206
Enzymum-Test <sup>®</sup> Substrat (Enzyme-Substrate Test)	Roche Diagnostics	Mannheim, Germany	1 295 250
Expand <sup>™</sup> High Fidelity PCR system	Roche Diagnostics	Mannheim, Germany	3 300 226
Expand <sup>™</sup> Long Template PCR system	Roche Diagnostics	Mannheim, Germany	1 681 842
Improm II <sup>™</sup> Reverse Transcriptase	Promega	Madison, Wisconsin, USA	A3800
pGem <sup>®</sup> T-Easy Vector system	Promega	Madison, Wisconsin, USA	A1360
QIAamp <sup>®</sup> Blood and Body Fluid Spin Kit	Qiagen GmbH	Hilden, Germany	51104
QIAamp <sup>®</sup> Ultrasens virus Kit	Qiagen GmbH	Hilden, Germany	53706
Qiagen MinElute <sup>®</sup> Gel Extraction Kit	Qiagen GmbH	Hilden, Germany	28604
QIAprep <sup>®</sup> Spin Miniprep Kit	Qiagen GmbH	Hilden, Germany	27104
QIAquick <sup>®</sup> PCR purification Kit	Qiagen GmbH	Hilden, Germany	28106
Repli-g <sup>®</sup> system	Qiagen GmbH	Hilden, Germany	59043
Vironostika <sup>®</sup> HIV Uni-Form II Ag-Ab bench ELISA test	Biomerieux	Durham, North Carolina, USA	84138

\*The Promega PCR systems, such as the Access RT-PCR system, are covered by patent laws.

**Table 2.3:** Additional chemicals needed for analysis

Product	Supplier	Location	Catalogue number
Ampicillin	Gibco™, Invitrogen Corporation	Paisley, United Kingdom	Q10016
dNTPs	Roche Diagnostics	Mannheim, Germany	11 636 103 001
Ethidium bromide	Promega	Madison, Wisconsin, USA	H5041
Exonuclease ( <i>Exo1</i> )	Amersham Pharmacia Biotechnology	New Jersey, USA	E70073Z
Histopaque®-1077	Sigma-Aldrich	Steinheim, Germany	A7054
Isopropyl-β-D-thiogalactopyranosid (IPTG)	Promega	Madison, Wisconsin, USA	V3951
Luria-Bertani (LB) Broth	Fluka Biochemika	Buchs, Switzerland	61748
Sabax® water	Sabax	Johannesburg, South Africa	H/34/4
Seakem LE® agarose	FMC BioProducts	Rockland, Maine, USA	50004
Shrimp alkaline phosphatase (SAP)	Amersham Pharmacia Biotechnology	New Jersey, USA	E70092Y
Template suppression reagent (TSR)	Applied Biosystems	California, USA	401956
X-Galactosidase (X-Gal)	Promega	Madison, Wisconsin, USA	V3941

### 2.2.1 Cohort samples

This project was approved by the Ethics Committee of the University of Stellenbosch (Research Committee C) on 01 September 2004. The project number is N04/06/100 (Appendix A). The patient cohort forms part of a larger study headed by Dr. J. Fincham at the South African Medical Research Council (MRC) entitled: “**Tracking potential for exposure to helminthic antigens to impair SAAVI vaccine trials and efficacy of HIV vaccine**” where the influence of deworming on HIV vaccination is being investigated. The study looks at mass deworming campaigns and the influence the programmes will have on HIV/AIDS and TB (Fincham *et al*, 2002; Fincham *et al*, 2003; Markus and Fincham, 2000). In summary, it is hypothesised that a vaccine against HIV-1 might be ineffective in people who have helminthic infections. Chronic helminthic infections continuously activate the host cellular immune system by the activation of T helper type 2 lymphocytes and their cytokines. Vaccine

development against HIV-1 is partially aimed at inducing and activating the cellular immune responses of an infected individual. With both infections targeting the cellular immune response, cross-regulation, or suppression, of the immune system by helminthic infections may impair an individual to produce a successful immune response targeted against HIV-1. HIV-1 infected individuals already have an immunocompromised status and the onset of AIDS in these individuals might be advanced due to other occurring opportunistic infections.

Approximately 20 ml EDTA blood samples were received from the Matthew Goniwe general health clinic located at site C in Khayelitsha during the period July 2002 to November 2003. EDTA was used as an anticoagulant and prevents the clotting of blood samples (Koepke *et al*, 1989). The patient number corresponds to the number given to the samples by Dr. J. Fincham and his colleagues at the MRC. The cohort of 127 samples consists of 110 females (86.6%) and 17 males (13.4%) with an average age of 31.7 ( $\pm$  7.8) years. Eighteen of these patients (14.2%) were also diagnosed with TB, while 9 patients (7.1%) had other sexually transmitted diseases (STDs). HIV-1 positive patients attending the clinic are part of a support group which continuously receives counselling on HIV/AIDS (Fincham, 2004). The samples were tested positive for HIV-1 in the Department of Medical Virology, Tygerberg Campus, University of Stellenbosch on the Axsym<sup>®</sup> diagnostic system (Abbott Laboratories, Illinois, USA). Their HIV-1 status was confirmed with the Vironostika<sup>®</sup> HIV Uni-Form II Ag-Ab bench ELISA test (Biomerieux, Durham, North Carolina, USA). These tests are based on EIAs and detect specific antigens and antibodies against HIV. The Axsym<sup>®</sup> HIV Ag-Ab assay is an automated immunoassay. The test uses HIV recombinant antigens, representative of HIV-1 group M gp41, HIV-1 group O gp41 and HIV-2 gp36, as well as HIV-1 p24-specific monoclonal antibodies to capture the HIV antigens present in a test sample (Brust *et al*, 2000; Ly *et al*, 2001; Ly *et al*, 2004). The Vironostika test uses HIV p24 antigens to detect HIV-1 group M, HIV-1 group O and HIV-2 antigens (Ly *et al*, 2004; Van Binsbergen *et al*, 1998; Van Binsbergen *et al*, 1999).



## 2.3 Methods

### 2.3.1 Sample preparation

Patient plasma and PBMCs were used during this study for the characterisation of HIV-1 samples. Approximately 3 ml plasma was stored from the 20 ml EDTA blood samples by separating it from the red blood cells, white blood cells and blood platelets. Plasma consists of approximately 90% water, as well as blood proteins, salts and minerals. The plasma was collected after centrifugation (Beckman Coulter Allegra™ 6R, Beckman Inc., Fullerton, California, USA) at 2000 rpm (revolutions per minute) for 10 minutes at 4°C. The samples were stored at -80°C for subsequent analysis. The plasma was used for serotyping (60 µl), RNA extractions (1 ml) and viral load determination (200 µl). PBMCs were extracted by density gradient centrifugation with Histopaque®-1077 (Sigma-Aldrich®, Steinheim, Germany) to separate them from red blood cells and most granulocytes (Janeway *et al*, 2001). The cells were frozen in liquid nitrogen (-120°C to -150°C) for long-term storage. RNA was extracted from one ml plasma using the QIAamp® Ultrasens Virus protocol (Qiagen GmbH, Hilden, Germany). The kit protocol was followed and RNA eluted in 50 µl of low salt buffer AVE containing 0.04% sodium azide. Stored PBMCs were used to extract genomic DNA of 3 possible HIV-1 recombinant samples with low or undetectable viral loads. The DNA was extracted from 500 µl sample with the QIAamp® Blood and Body Fluid Spin Protocol (Qiagen GmbH, Hilden, Germany) following the manufacturer's instructions. The Qiagen RNA and DNA extraction protocols are based on the silica-gel membrane principles of Vogelstein and Gillespie (1979). Briefly, the nucleic acids bind to a silica membrane in a QIAamp® Spin Column during centrifugation or a vacuum step. Salt and pH conditions ensure that proteins and other possible contaminants are not retained on the membrane. Wash steps ensure that residual contaminants are removed and improve the purity of the samples. Elution with water, or a low salt buffer, releases the nucleic acids from the Qiagen silica membrane.

### 2.3.2 HIV-1 viral load assay

The Abbott LCx<sup>®</sup> HIV RNA Quantitative assay (Abbott Laboratories, Illinois, USA) was used according to the manufacturer's instructions. Briefly, the assay is based on RT-PCR technology and targets the *pol* IN gene to determine the viral load concentration. Labelled probes, specific for HIV-1, detect the DNA products in the sample after RT-PCR. The amplified product-probe hybrids are quantitated with a microparticle enzyme immunoassay (MEIA) on an automated Abbott LCx<sup>®</sup> Analyzer (Abbott Laboratories, Illinois, USA). A volume of 200 µl was used to determine the number of HIV-1 RNA copies per ml in each plasma sample. The assay has a range of 50 to 1 million copies per ml if 1 ml sample volume is used. If a 200 µl sample is used, the detection limit ranges from 178 to 5 million copies per ml. The assay can detect and quantitate HIV-1 group M subtypes A-G, as well as HIV-1 group O (Johanson *et al*, 2001; Zanchetta *et al*, 2000).

### 2.3.3 *env* gp120 V3 serotyping assay

The serotype of the HIV-1 strains was determined using a V3 cPEIA (Roche Diagnostics, Mannheim, Germany) as previously described (Barin *et al*, 1996; Engelbrecht *et al*, 1999; Plantier *et al*, 1998). The enzyme-substrate test (Enzymum-Test<sup>®</sup> Substrat) peptides are prepared by Roche Diagnostics (Mannheim, Germany). Briefly, anti-HIV-1 antibodies of the patient samples react with five biotinylated *env* gp120 peptides of HIV-1 subtypes A to E, binding the antibodies to the solid phase. In a second incubation anti-human Fcγ (Fragment crystallisable γ) antibodies from sheep bind to the fixed human IgG antibodies on the solid phase. An indicator reaction with substrate-chromogen results in a colour change which is measured at 405 nm. The subtype mixture giving the lowest signal recovery indicates the serotype of the sample. Signal reduction of less than 30% is usually indicative of the sample subtype. Cross-reactivity occurs when two or more subtypes react to the same extent (less than 10% difference) with other peptides. The absorbance was spectrophotometrically measured with the Windows<sup>™</sup> based Anthos Winread version 2.3 system (Anthos Labtec., Salzburg, Austria). The measurement filter was set at 405 nm, while the reference filter was set at 492 nm. A subtype B positive control

and a negative human serum control, provided by the kit, were included on the EIA plates.

### **2.3.4 The polymerase chain reaction**

The PCR method was used for the HIV-1 genotyping and cycle sequencing reactions. The reactions were carried out on the geneAMP<sup>®</sup> 9700 PCR system (Applied Biosystems, Foster City, California, USA) using the standard 9700 ramp speed (average heating rate - 1.5°C per second; average cooling rate - 1.0°C per second; Applied Biosystems, Foster City, California, USA). The ramp speed is calculated as the speed at which the PCR machine (thermocycler) switches between heating steps. The PCR mixtures were prepared in 0.2 ml thin wall PCR<sup>®</sup> tubes (QSP, Porex BioProducts Inc., California, USA)

### **2.3.5 HIV-1 genotyping reactions**

HIV-1 genotyping reactions were performed based on the *gag* p24, *env* gp41 IDR, *env* gp120 V3 and *pol* gene regions of the HIV-1 genome. The PCR primers and additional sequencing primers used for each gene fragment are listed in Table 2.1. The primers were designed to amplify all HIV-1 group M subtypes (Swanson *et al*, 2003; Lindström and Albert, 2003). The primers also indicate the position of the genomic regions, relative to HXB2 (Ratner *et al*, 1985b), targeted for amplification. HXB2 is the prototype virus of choice, as it is the most commonly used reference strain for different functional studies (Korber *et al*, 1998). The Access RT-PCR system (Promega, Madison, Wisconsin, USA) was used in the first round of PCR amplification and the Expand<sup>™</sup> High Fidelity PCR system (Roche Diagnostic, Mannheim, Germany) in the second round. The Access RT-PCR system uses a *Thermus flavus* (*Tfl*) thermostable DNA polymerase with AMV RT (Avian Myeloblastosis Virus Reverse Transcriptase) to create and amplify complementary DNA (cDNA) from the viral RNA products (Abramovici, 2001). The *Taq* and *Tgo* (*Thermococcus gorgonarius*) DNA polymerases of the Expand<sup>™</sup> High Fidelity PCR system ensure the generation of high yield, high fidelity and high specificity DNA products (Barnes, 1994). Two rounds of replication were sufficient to yield DNA concentrations high enough for visualisation on an agarose gel and consequent

sequencing reactions. Five  $\mu\text{l}$  of HIV-1 RNA was added in the first round of replication cycles. Subsequently, 3  $\mu\text{l}$  of the first round reaction was then used in the second round of amplification.

The following standard method of first round PCRs was used: One cycle of reverse transcription at  $48^{\circ}\text{C}$  for 45 minutes; one cycle of denaturation at  $94^{\circ}\text{C}$  for 2 minutes and 40 cycles of heat denaturation at  $94^{\circ}\text{C}$  for 15 seconds, primer annealing for 30 seconds (according to primer pair annealing temperature) and elongation at  $72^{\circ}\text{C}$  for 1 minute. A final elongation step of  $72^{\circ}\text{C}$  for 10 minutes was included, after which the samples were cooled down and stored at  $4^{\circ}\text{C}$  until used.

The second round of amplification consisted of one cycle of denaturation at  $94^{\circ}\text{C}$  for 2 minutes, followed by 40 cycles of denaturing at  $94^{\circ}\text{C}$  for 30 seconds, primer annealing for 30 seconds (according to primer pair annealing temperature) and elongation at  $72^{\circ}\text{C}$  for 1 minute. A final elongation step of  $72^{\circ}\text{C}$  for 10 minutes was performed, after which the samples were cooled and stored at  $4^{\circ}\text{C}$  until used.

### 2.3.6 Agarose gel electrophoresis

The products of all the PCR amplification were viewed by agarose gel electrophoresis on a 0.8% agarose (Seakem LE<sup>®</sup> agarose; FMC BioProducts, Rockland, Maine, USA) gel in 1 x TAE buffer (0.04 M Tris acetate, 0.001 M EDTA). Five  $\mu\text{l}$  ethidium bromide (0.5  $\mu\text{g}$  per ml; Promega, Madison, Wisconsin, USA) was added to the gel to stain the DNA. Ethidium bromide is a powerful mutagen, is moderately toxic and can become carcinogenic and thus care should be taken when handling it (Ausubel *et al*, 2003; Sambrook *et al*, 1989; Sharp *et al*, 1973). A 1 kb DNA molecular weight marker (Promega, Madison, Wisconsin, USA) was used as an estimation of DNA band size. The DNA bands were visualised under an ultra violet light at a wavelength of 302 nm and photographed with the Syngene<sup>™</sup> GeneGenius computer system (Synoptics Ltd., Cambridge, United Kingdom).

**Table 2.4:** HIV-1 genotyping primers

	Amplification Step	Orientation	PCR Primers (5'-3')	Bases	*T <sub>m</sub> (50mM)	#Position (HXB2)	Reference
<b><i>gag p24</i></b>							
P24-7	cDNA 1st PCR	Reverse	CCCTGRCATGCTGTCATCA	19	57.0	1826	Swanson <i>et al</i> , 2003
P24-1	cDNA 1st PCR	Forward	AGYCAAAATTAYCCYATAGT	20	48.4	1193	Swanson <i>et al</i> , 2003
M/O p24-2	2nd PCR	Forward	AGRACYTTRAAYGCATGGGT	20	55.8	1256	Swanson <i>et al</i> , 2003
M/O p24-6	2nd PCR	Reverse	TGTGWAGCTTGYTCRGCTC	19	56.1	1703	Swanson <i>et al</i> , 2003
<b><i>env gp41</i></b>							
JH38	cDNA 1st PCR	Reverse	GGTGARTATCCCTKCCTAAC	20	52.9	8346	Swanson <i>et al</i> , 2003
JH41	cDNA 1st PCR	Forward	CAGCAGGWAGCACKATGGG	20	57.4	7816	Swanson <i>et al</i> , 2003
Env 27F	2nd PCR	Forward	CTGGYATAGTGCARCARCA	19	54.9	7879	Swanson <i>et al</i> , 2003
Menv 19R	2nd PCR	Reverse	AARCCTCTACTATCATTATRA	22	49.4	8278	Swanson <i>et al</i> , 2003
<b><i>env gp120 V3</i></b>							
ED5	1st PCR	Forward	ATGGGATCAAAGCCTAAAGCCATGTG	26	71.8	6582	Bachmann <i>et al</i> , 1994
ED12	1st PCR	Reverse	AGTGCTTCCTGCTGCTCCCAAGAACCCAAG	30	79.2	7782	Bachmann <i>et al</i> , 1994
Es7x	2nd PCR	Forward	CTGTAAATGGTAGTCTAGC	20	51.9	7021	Bachmann <i>et al</i> , 1994
Es8x	1st PCR	Reverse	CACTTCTCCAATTGTCCCTCA	21	55.7	7648	Bachmann <i>et al</i> , 1994
E125	2nd PCR	Reverse	CAATTTCTGGGTCCCCTCTGAG	23	60.6	7316	Sanders-Buell <i>et al</i> , 1995
<b><i>pol</i></b>							
JA203	1st PCR	Forward	GGAAGAYTGYACTGAGAGACAGGCTAAT	28	64.6	2085	Lindström and Albert, 2003
JA206	1st PCR	Reverse	TTAATCCCTGCRTAAATCTGACTTG	25	59.7	3349	Lindström and Albert, 2003
JA204	2nd PCR	Forward	TTCAGAGCAGACCAGAGCCAACAGC	25	67.9	2159	Lindström and Albert, 2003
JA205	2nd PCR	Reverse	TTTTCCCACTAACTTCTGTATGTCATTG	28	61.7	3311	Lindström and Albert, 2003
JA217	Sequencing	Forward	CTTTTATTTTTCTTCTGTCATG	25	56.4	2622	Lindström and Albert, 2003
Pol1D	Sequencing	Forward	TCCCTCAAATCACTCTTTGGC	21	56.3	2271	Loxton, 2004
Pol2D	Sequencing	Forward	CTATTGAACTGTACC	16	40.1	2575	Loxton, 2004

\*T<sub>m</sub> – primer melting temperature

#The position of the primer relative to the start position of HXB2 is indicated

### 2.3.7 Purification of PCR products

The above-mentioned PCR products were purified using the enzymes Exonuclease 1 (*Exo1*) and Shrimp alkaline phosphatase (SAP) (Amersham Pharmacia Biotech., New Jersey, USA) according to the manufacturer's instructions. These enzymes are used to remove any excess primers and dNTPs that might interfere with the sequencing reactions (Werle *et al*, 1994). Ten µl from the PCR mix was used with 0.5 µl of *Exo1* (1 unit per µl) and 0.5 µl of SAP (1 unit per µl). An incubation of 15 minutes at 37°C

followed by an enzyme inactivation step of 15 minutes at 80°C was carried out on a Techne® Gene E Thermal Cycler (Techne Ltd., Cambridge, United Kingdom). The purified aliquot was stored at -20°C until the sequencing reactions were carried out.

### 2.3.8 DNA concentration determination

The DNA concentration was determined by measuring the absorbance (A) of the sample at 260 nm, while the purity was measured by calculating the absorbance at 260 nm divided by the absorbance at 280 nm (Ausubel *et al*, 2003; Sambrook *et al*, 1989). The following equations apply:

$$\text{DNA concentration} = \frac{A_{260}}{20} \times \text{dilution factor} \qquad \text{DNA purity} = \frac{A_{260}}{A_{280}}$$

All DNA concentrations were determined using the Nanodrop™ ND-1000 system (Nanodrop Technologies Inc.; Delaware, USA), which only requires 1 - 2 µl of sample input. The method is fast, accurate, reproducible and does not require any concentration dilutions, as with conventional spectrophotometers (Ambion, 2004). Pure DNA, without protein or other contaminants, should give a purity reading between 1.7 and 1.9, while a reading from 1.5 to 1.7 is also considered acceptable (Ausubel *et al*, 2003; Sambrook *et al*, 1989).

### 2.3.9 DNA cycle sequencing reactions

Approximately 50 ng of the purified products were used in the sequencing reactions. The BigDye™ Terminator Cycle Sequencing Ready Reaction Kit (Applied Biosystems, Foster City, California, USA) was used for the PCR based sequencing reactions. The manufacturer's instructions were followed. Briefly, 4 µl of the Big Dye terminator enzyme mix, 5 pmol of primer, ~ 50 ng of sample DNA and water were added together to give a final volume of 10 µl. Sabax® water (Sabax, Johannesburg, South Africa) free from RNase, DNase or any other DNA contaminants that might interfere with the biological reactions, was used. The second round PCR primers were used for sequencing. The following cycle sequencing reaction was performed: Denaturation at 96°C for 10 seconds, primer annealing for 5

seconds (annealing temperature according to sequencing primers, Table 2.1) and an elongation step at 60°C for 4 minutes. The cycles were repeated 25 times after which the samples were cooled down to 4°C. The sequencing reactions were purified using the Dye ex™ 2.05 spin kit (Applied Biosystems, Foster City, California, USA). The samples were allowed to dry in a vacuum centrifuge (Speedvac™ SVC100, Savant Instruments, Inc., Farmingdale, New York, USA) and resuspended in 25 µl of template suppression reagent (TSR; Applied Biosystems, Foster City, California, USA). TSR prevents the clogging of the sequencing electrophoresis capillary with the DNA template (Bradley, 1996). The resuspended samples were denatured at 95°C for 5 minutes and loaded onto the automated ABI prism® 310 Genetic Analyzer (Applied Biosystems, Foster City, California, USA). The principles of automated DNA sequencing are described in section 1.2.8.7.

### **2.3.10 Sequence and phylogenetic analysis**

The sequencing data from the ABI prism® 310 Genetic Analyzer was converted into an electropherogram by using the DNA Sequencing Analysis software™, version 3.3 (Applied Biosystems, Foster City, California, USA). The colour coded electropherogram peaks generated correspond to the different dye labelled dNTPs of the DNA being analysed. The software creates a string of DNA sequences that can be exported and analysed by different sequence analysis software packages, as described below in the rest of section 2.3.10. Many of the software packages are copyright (©) protected.

#### **2.3.10.1 DNA sequence assembly**

The electropherogram was analysed and edited in Chromas version 1.45 (Technelysium Ltd. ©, Queensland, Australia). The positive and negative strands of DNA sequences are usually sequenced by two different primers. This is done to ensure that the DNA string of sequences being analysed is correct (Murphy *et al*, 2005). These overlapping fragments of the DNA sequences were assembled in Sequencer™ version 4.1.4 (Gene Codes Corporation, Ann. Arbor., Michigan, USA). A BLAST (Basic Local Alignment Search Tool; Altschul *et al*, 1997; Karlin and Altschul, 1990; Karlin and Altschul, 1993) search was performed on the HIV LANL

(<http://hiv-web.lanl.gov/>) with the DNA sequences to establish similarity between the Khayelitsha and other HIV-1 sequences.

### **2.3.10.2 Sequence alignments and phylogenetic trees**

DNA and protein alignments were created with Clustal X version 1.7 (Thompson © *et al*, 1997). The program uses an algorithm to construct an alignment consensus that best matches all the sequences over the entire length of the sequence (Needleman and Wunsch, 1970). The alignments were manually checked for amino acid codon regions (Goldman and Yang, 1994; Muse and Gaut, 1994) with BioEdit version 5.09 (Hall ©, 2001). All alignments were trimmed so that each set of sequences analysed were the same length. This ensures that any sequence bias can be excluded by the wrongful insertion of gaps in a specific alignment (Thompson *et al*, 1997). DNA sequences were translated into proteins with Genedoc version 2.6.002 (Nicholas © and Nicholas, 1997). Initial neighbour-joining phylogenetic trees (Saitou and Nei, 1987) were drawn with TreeconW for Windows, version 1.3b (Van de Peer © and De Wachter, 1994) using the Kimura 2-parameter (Kimura, 1980). Neighbour-joining phylogenetic trees (Saitou and Nei, 1987) were drawn with TreeconW, as an accurate single phylogeny tree output can be reached with fairly high computational speed (Page and Holmes, 2002). Sequence identity matrices were created from the sequence alignments using BioEdit version 5.09 (Hall ©, 2001).

Modeltest version 3.7 (Posada © and Crandall, 1998) was used to select the best-fit model of nucleotide substitution to use in constructing maximum likelihood (Felsenstein, 1981) phylogenetic trees. Maximum likelihood trees were drawn with PAUP (Phylogenetic Analysis Using Parsimony \*and other methods) version 4.0b10 (Sinauer Associates Inc. ©, Massachusetts, USA; Maddison *et al*, 1997). The PAUP software reconstructs and analysis trees using maximum likelihood, parsimony, or distance methods. Maximum likelihood trees were drawn as they are more precise than neighbour-joining trees and use a more accurate model of evolution for the evaluation of a data set. They are also more computer intensive with the analysis being far more thorough (Salemi and Vandamme, 2003). Maximum likelihood trees were drawn by using stepwise addition and TBR branch swapping algorithms in PAUP.



The confidence level of the neighbour-joining and maximum likelihood trees was tested with 1000 bootstrap values (Felsenstein, 1985). Bootstrap values greater than 95% are generally, but not always considered statistically significant (Felsenstein, 1985; Efron et al, 1996). Bootstrapping for the maximum likelihood trees was done in Tree-puzzle version 5.2 (Schmidt © *et al*, 2002). Bootstrapping for the neighbour-joining trees forms part of the TreeconW package. The maximum likelihood trees were viewed and rooted with an outgroup in Treeview version 1.6.6 (Page ©, 1996). Reference sequences used in the phylogenetic trees were obtained from the LANL. In Table 2.5 the set of subtype reference sequences used in phylogenetic analysis is presented. The sequences represent all the major HIV-1 group M subtypes. In Table 2.6 the subtype D *gag* p24 sequences used in phylogenetic analysis and in Table 2.7 the subtype C *pol* sequences are presented. The subtype D *gag* p24 and subtype C *pol* sequences were chosen randomly to use for more detailed analysis of possible HIV-1 recombinant strains identified.

#### **2.3.10.3 Analysis of hypermutation variants**

Query sequences that were considered to be possible hypervariable mutants were analysed with the LANL Hypermut program (Rose and Korber, 2000). The query sequences are compared against a user defined reference sequence. G → A transitions are associated with hypermutant variants and are highlighted by the program.

#### **2.3.10.4 Consensus sequences and the identification of conserved genomic regions**

The BioEdit software package was used to create consensus sequences, calculate amino acid entropy values and search for conserved genomic regions amongst the Khayelitsha amino acid sequences. Consensus sequences were created by using the default 60 percent threshold setting and treating gaps in the sequences as an extra residue. This means that if a sequence occurs in 60%, or more, of the sequences analysed, that sequence will be allocated as the general consensus sequences at that position. The entropy value was used to measure the degree of variability between the Khayelitsha amino acid sequences. The higher an entropy value, the more

variable, or dislike, the sequences are at a particular position in an alignment. Conserved amino acid epitopes recognised were aligned with other epitopes by using the EPILIGN program available on the LANL.

### 2.3.10.5 Similarity plots

Similarity plots were drawn with Simplot version 2.5 (Lole *et al*, 1999) to identify breakpoints in possible HIV-1 recombinant strains (Lole *et al*, 1999; Salminen *et al*, 1995). The program uses a sliding window moving across an alignment in small increment steps to generate a similarity plot. The query sequence is compared to a set of reference sequences in a specified alignment. It is based on the Kimura 2-parameter substitution model. Analysis was done with a window size of either 100 bp for the shorter *gag* p24 DNA sequences, or 200 bp for the longer *pol* DNA sequences. Increment steps of 10 bp were used.

### 2.3.10.6 Viral phenotype prediction

The viral phenotype (co-receptor usage) was predicted by calculating the net *env* gp120 V3 loop charge. This is done by subtracting the number of negatively charged amino acids [aspartic acid (D) or glutamic acid (E)] from the number of positively charged amino acids [lysine (K), arginine (R) or histidine (H)]. SI variants have been shown to have a higher net V3 charges ( $\geq 5$ ) compared to their NSI counterparts (Briggs *et al*, 2000; Kwa *et al*, 2003; Pollakis *et al*, 2004). A basic substitution (amino acids K, R and H) at position 11 or 25 in the V3 loop has also been shown to predict either CXCR4 or combined CCR5 and CXCR4 co-receptor usage (Fouchier *et al*, 1992; Fouchier *et al*, 1995; Hoffman *et al*, 2002).

The co-receptor predictions were made on the following website using the Sinsi matrix: <http://ubik.microbiol.washington.edu/computing/pssm/> (Jensen *et al*, 2003). The matrix employs position-specific scoring matrices (PSSMs) based on *env* gp120 V3 genomic sequences. A PSSM uses reference sequences of viruses with known phenotypes and predicts the probability of a query sequence having that same phenotype. Virtual phenotype predictions based on PSSMs are much quicker and less labour intensive (Jensen *et al*, 2003). Two scoring matrices are available based on

two different available functional assays. The X4R5 matrix uses indicator cell lines producing exogenous CD4 (Vodicka *et al*, 1997) and the Sinsi matrix uses assays based on MT2, Human HTLV producing T-cells, cell lines (Koot *et al*, 1992). The Sinsi matrix was preferred, as this setting gives a phenotype prediction that correlates with ART and HIV/AIDS disease progression (Brumme *et al*, 2004; Jensen *et al*, 2003).

**Table 2.5:** HIV- 1 subtype reference sequences used for phylogenetic analysis

Name	Accession number	Origin of sample	Year of sampling	Subtype	Reference
Q23_17	AF004885	Kenya	1994	A1	Poss <i>et al</i> , 1998
92UGO37.1	U51190	Uganda	1992	A1	Gao <i>et al</i> , 1996
CDKS10	AF286241	DRC	1997	A2	Gao <i>et al</i> , 2001
CDKFE4	AF286240	DRC	1997	A2	Gao <i>et al</i> , 2001
94CYO17-41	AF286242	DRC	1997	A2	Gao <i>et al</i> , 2001
CDKTB48	AF286238	DRC	1997	A2	Gao <i>et al</i> , 2001
JRFL	U63632	USA	1986	B	O'Brien <i>et al</i> , 1990
HXB2	K03455	France	1983	B	Ratner <i>et al</i> , 1985b
WEAU160	U21135	USA	1990	B	Clark <i>et al</i> , 1991
RF	M17451	USA	1983	B	Starcich <i>et al</i> , 1986
ETH2220	U46016	Ethiopia	1986	C	Salminen <i>et al</i> , 1996
92BRO25.8	U52953	Brazil	1992	C	Gao <i>et al</i> , 1996
IN21068	AF067155	India	1995	C	Lole <i>et al</i> , 1999
96BW05.02	AF110967	Botswana	1996	C	Novitsky <i>et al</i> , 1999
94UG114.1	U88824	Uganda	1994	D	Gao <i>et al</i> , 1998
84ZR085	U88822	DRC	1984	D	Gao <i>et al</i> , 1998
NDK	M27323	DRC	1983	D	Spire <i>et al</i> , 1989
ELI	K03454	DRC	1983	D	Alizon <i>et al</i> , 1986
MP411	AJ249238	France	1996	F1	Triques <i>et al</i> , 1999
V1850	AF077336	Belgium (DRC)	1993	F1	Carr <i>et al</i> , 1998
MP257	AJ249237	Cameroon	1995	F2	Triques <i>et al</i> , 1999
MP255	AJ249236	Cameroon	1995	F2	Triques <i>et al</i> , 1999
DRCBL	AF084936	Belgium	1996	G	Oelrichs <i>et al</i> , 1999
SE6165	AF061642	Sweden (DRC)	1993	G	Carr <i>et al</i> , 1998
92NG083	U88826	Nigeria	1992	G	Gao <i>et al</i> , 1998
VI991	AF19012	Belgium	1993	H	Janssens <i>et al</i> , 2000
90CF056	AF005496	Central African Republic	1990	H	Murphy <i>et al</i> , 1993
VI997	AF190128	Belgium	1993	H	Janssens <i>et al</i> , 2000
SE7022	AF082394	Sweden	1994	J	Laukkanen <i>et al</i> , 2000
SE7887	AF082395	Sweden	1993	J	Laukkanen <i>et al</i> , 2000
EQTB11C	AJ249235	DRC	1997	K	Triques <i>et al</i> , 1999
MP535	AJ249239	Cameroon	1996	K	Triques <i>et al</i> , 1999
Ant70	AF147884	Cameroon	1987	Group O	Janssens <i>et al</i> , 1999

**Table 2.6:** Subtype D *gag* p24 sequences used in phylogenetic analysis

Name	Accession number	Origin of sample	Year of sampling	Reference
MP259	AJ286372	Cameroon	1995	Montavon <i>et al</i> , 2000
MP571	AJ286394	Cameroon	1997	Montavon <i>et al</i> , 2000
0175BA	AY371156	Cameroon	2001	Kijak <i>et al</i> , 2004
MN011	AJ488926	Chad	1999	Vidal <i>et al</i> , 2003
MN012	AJ488927	Chad	1999	Vidal <i>et al</i> , 2003
MN018	AJ491005	Chad	1999	Vidal <i>et al</i> , 2003
MN019	AJ491006	Chad	1999	Vidal <i>et al</i> , 2003
84ZR085	84ZR085	DRC	1984	Gao <i>et al</i> , 1998
ELI	K03454	DRC	1983	Alizon <i>et al</i> , 1986
KFE339	AJ404250	DRC	1997	Vidal <i>et al</i> , 2000b
KS26	AJ404267	DRC	1997	Vidal <i>et al</i> , 2000b
KTB23	AJ404278	DRC	1997	Vidal <i>et al</i> , 2000b
MBFE183	AJ404285	DRC	1997	Vidal <i>et al</i> , 2000b
NDK	M27323	DRC	1983	Spire <i>et al</i> , 1989
VI203	L11784	DRC	1993	Louwagie <i>et al</i> , 1993
VI205	L11785	DRC	1993	Louwagie <i>et al</i> , 1993
Z2Z6_Z2	M22639	DRC	1985	Srinivasan <i>et al</i> , 1987
MP613	AJ286535	France	1997	Montavon <i>et al</i> , 2000
97GA_972	AJ286426	Gabon	1997	Montavon <i>et al</i> , 2000
97GA_G27	AJ286443	Gabon	1997	Montavon <i>et al</i> , 2000
97GA_GB144	AJ286453	Gabon	1997	Montavon <i>et al</i> , 2000
97GA_TB125	AJ286479	Gabon	1997	Montavon <i>et al</i> , 2000
97GA_URG7	AJ286499	Gabon	1997	Montavon <i>et al</i> , 2000
M115gag	AY772952	Kenya	*N/A	Steain <i>et al</i> , 2004
M150gag	AY772956	Kenya	N/A	Steain <i>et al</i> , 2004
MB2059	AF133821	Kenya	1993	Neilson <i>et al</i> , 1999
ML415_2_1997	AY322189	Kenya	1997	Fang <i>et al</i> , 2004
NKU3006	AF457090	Kenya	2001	Dowling <i>et al</i> , 2002
112HPD	AJ274561	Senegal	1998	Toure-Kane <i>et al</i> , 2000
96SE_1029	AJ274543	Senegal	1996	Toure-Kane <i>et al</i> , 2000
96SE_1116	AJ274541	Senegal	1996	Toure-Kane <i>et al</i> , 2000
SN365	L11797	Senegal	1990	Louwagie <i>et al</i> , 1993
D_ZA_R2	AY773338	South Africa	1984	Loxton <i>et al</i> , 2005
D_ZA_R214	AY773339	South Africa	1985	Loxton <i>et al</i> , 2005
D_ZA_R286	AY773340	South Africa	1985	Loxton <i>et al</i> , 2005
D_ZA_R482	AY773341	South Africa	1986	Loxton <i>et al</i> , 2005
94UG114	U88824	Uganda	1994	Gao <i>et al</i> , 1998
99UGB32394	AF484483	Uganda	1999	Harris <i>et al</i> , 2002
99UGD23550	AF484485	Uganda	1999	Harris <i>et al</i> , 2002
99UGD26830	AF484486	Uganda	1999	Harris <i>et al</i> , 2002
99UGE08364	AF484487	Uganda	1999	Harris <i>et al</i> , 2002
99UGE23438	AF484489	Uganda	1999	Harris <i>et al</i> , 2002
AMK10	U08192	USA	1994	Gao <i>et al</i> , 1994

\*N/A - Data not available

**Table 2.7:** Subtype *C pol* sequences used in phylogenetic analysis

<b>Name</b>	<b>Accession number</b>	<b>Origin of sample</b>	<b>Year of sampling</b>	<b>Reference</b>
p99_249	AF338992	Belgium	1999	Snoeck <i>et al</i> , 2002
96BW01B03	AF110959	Botswana	1996	Novitsky <i>et al</i> , 1999
96BW11B01	AF110971	Botswana	1996	Novitsky <i>et al</i> , 1999
96BW15C05	AF110975	Botswana	1996	Novitsky <i>et al</i> , 1999
96BW16D14	AF110977	Botswana	1996	Novitsky <i>et al</i> , 1999
00BW18802	AF443100	Botswana	2000	Novitsky <i>et al</i> , 2002
92BR025	U52953	Brazil	1992	Gao <i>et al</i> , 1996
261	AJ419476	Denmark	2002	Jorgensen <i>et al</i> , 2003
DJ259	AF447839	Djibouti	1991	Huang <i>et al</i> , 2003
ETH2220	U46016	Ethiopia	1986	Salminen <i>et al</i> , 1996
DD88379GP	AY242591	Ethiopia	1988	Pollakis <i>et al</i> , 2003
DE88404GP	AY242594	Ethiopia	1988	Pollakis <i>et al</i> , 2003
JM96111GP	AY242588	Ethiopia	1996	Pollakis <i>et al</i> , 2003
s03_928	AY371693	Ethiopia	2003	Mono <i>et al</i> , unpublished
93IN101	AB023804	India	1993	Mochizuki <i>et al</i> , 1999
93IN905	AF067158	India	1993	Lole <i>et al</i> , 1999
93IN9999	AF067154	India	1993	Lole <i>et al</i> , 1999
98IN012	AF286231	India	1998	Rodenburg <i>et al</i> , 2001
98IN022	AF286232	India	1998	Rodenburg <i>et al</i> , 2001
mIDU101_3	AB097871	Myanmar	1999	Takebe <i>et al</i> , 2003
SE364	AF447842	Senegal	1990	Huang <i>et al</i> , 2003
99SE_17178	AY165213	Senegal	1999	Maljkovic <i>et al</i> , 2003
00SE_18474	AY165224	Senegal	2000	Maljkovic <i>et al</i> , 2003
SE_15076	AY165196	Senegal	2003	Maljkovic <i>et al</i> , 2003
SM145	AF447850	Somalia	1989	Huang <i>et al</i> , 2003
TV001	AY162223	South Africa	1998	Zur Megede <i>et al</i> , 2002
03ZASK011B2	AY901965	South Africa	2003	Kiepiela <i>et al</i> , 2004
SK065B1	AY772694	South Africa	2003	Kiepiela <i>et al</i> , 2004
SK134B1	AY703909	South Africa	2004	Kiepiela <i>et al</i> , 2004
97TZ05	AF361875	Tanzania	1997	Hoelscher <i>et al</i> , 2001
98TZ013	AF286234	Tanzania	1998	Rodenburg <i>et al</i> , 2001
A125	AY253304	Tanzania	2001	Arroyo <i>et al</i> , 2004
A246	AY253308	Tanzania	2001	Arroyo <i>et al</i> , 2004
121108V	AF410207	Uganda	2001	Eshleman <i>et al</i> , 2002
613	AF388102	Uganda	2001	Eshleman <i>et al</i> , 2002
823	AF388161	Uganda	2001	Eshleman <i>et al</i> , 2002
NC5625POL	AY032091	USA	1999	Gonzales <i>et al</i> , 2001
96ZM751	AF286225	Zimbabwe	1996	Rodenburg <i>et al</i> , 2001

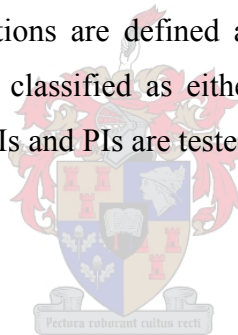
### 2.3.10.7 Envelope N-linked glycosylation sites

An N-linked glycosylation pattern means that asparagine (N) requires the context of N-X-[S or T], where X is any amino acid followed by serine (S) or threonine (T) (Marshall, 1974). The N-linked glycosylation sites were determined at the LANL with the GLYCOSITE program (Kasturi *et al*, 1997; Marshall, 1974; Mellquist *et al*, 1998). Both the *env* gp41 IDR and the *env* gp120 V3 regions contain Env glycosylation sites.

### 2.3.10.8 Screening for drug resistant mutations

The *pol* sequences of samples 1039 and 1151 were screened for pre-existing drug resistant mutations on the HIV drug resistance database at Stanford University (<http://hivdb.stanford.edu/>). Query sequences are compared to the consensus subtype B PR an RT sequences. Mutations are defined as differences from the wild-type consensus B sequences and are classified as either major or minor. The level of resistance against NRTIs, NNRTIs and PIs are tested (Kantor *et al*, 2005).

### 2.3.11 Cloning experiments



Possible HIV-1 recombinant samples, which could not be clearly subtyped, were cloned to identify potential sets of quasispecies amongst them. If divergent quasispecies are present in a viral population, direct sequencing can lead to the presence of mixed bases at a single point in a particular DNA sequence (Kapoor *et al*, 2004; Paolucci *et al*, 2001). Some of these quasispecies can be identified by sequencing multiple subclones derived from a PCR product (Liu *et al*, 1996; Kapoor *et al*, 2004).

#### 2.3.11.1 Preparation of samples for cloning

PCR products targeted for cloning were gel-purified with the Qiagen MinElute<sup>®</sup> Gel Extraction Kit Protocol (Qiagen GmbH, Hilden, Germany) according to the manufacturer's instructions with DNA aliquots eluted in Sabax<sup>®</sup> water (Sabax, Johannesburg, South Africa). Approximately 400 µl of PCR product (multiple PCRs

performed on same sample) were pooled for gel-extraction purposes to obtain better DNA concentrations and yields.

### 2.3.11.2 Cloning reactions

Cloning reactions were done with the pGEM<sup>®</sup> T-Easy Vector System (Promega, Madison, Wisconsin, USA) according to the manufacturer's technical manual. It is an efficient system to ligate PCR products smaller than 2 kb into a DNA plasmid and allows the release of the cloned fragment from the plasmid through restriction enzyme digestion (pGEM<sup>®</sup> T- and pGEM<sup>®</sup> T-Easy Vector System Technical Manual; Promega, Madison, Wisconsin, USA). Briefly, approximately 25 ng of gel-purified DNA was ligated with 1 µl (50 ng) of the pGEM<sup>®</sup> T-Easy Vector. The vector was transformed into *E.coli* JM109 High Efficiency Competent Cells (Promega, Madison, Wisconsin, USA) and cultured onto petri dishes (QSP, Porex BioProducts Inc., California, USA) with Luria-Bertani (LB) media [(10g per l bacto-tryptone, 5g per l bacto-yeast-extract, 10g per l NaCl, 15g per l bacto-agar) (Fluka Biochemika, Buchs, Switzerland) (Ausubel *et al*, 2003; Sambrook *et al*, 1989)].

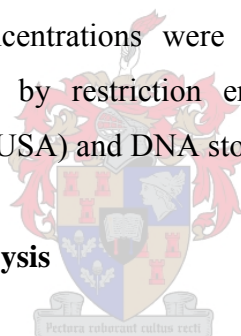
The petri dishes were prepared with 100 µg per ml of the antibiotic ampicillin (Gibco<sup>™</sup>, Invitrogen Corporation, Paisley, United Kingdom) to inhibit the growth of possible bacteria (Ausubel *et al*, 2003; Sambrook *et al*, 1989). The transformed JM109 cells are resistant to ampicillin and grow efficiently in the LB media. Twenty µl of 50 mg per ml X-Gal (Promega, Madison, Wisconsin) per plate and 100 µl of 100 mM IPTG (Promega, Madison, Wisconsin, USA) per plate were also added to assist with blue and white colony selection. IPTG induces the expression of the *LacZ* gene in the pGem<sup>®</sup> T-Easy Vector synthesis, while X-Gal is an indicator chemical responsible for producing blue colonies when cloning is unsuccessful (pGEM<sup>®</sup> T- and pGEM<sup>®</sup> T-Easy Vector System Technical Manual; Promega, Madison, Wisconsin, USA; Ausubel *et al*, 2003; Sambrook *et al*, 1989). The colonies were grown overnight at 37°C in a temperature control room incubator.

### 2.3.11.3 Selection of positive clones

Positive clones are usually white, although blue colonies may also sometimes contain the correct DNA insert if cloned in-frame with the *lacZ* gene (pGEM<sup>®</sup> T- and pGEM<sup>®</sup> T-Easy Vector System Technical Manual; Promega, Madison, Wisconsin, USA). Positive colonies were picked and a colony PCR performed by adding picked cultured cells in a PCR mixture containing the original PCR primers, as in section 2.3.4 and 2.3.5. Positive cloned products were sequenced and analysed, as described in section 2.3.9 and 2.3.10.

Positive colonies were also grown overnight at 37°C in 3 ml LB media with ampicillin in a 37°C Labcon FSIE shaking incubator (Labcon Ltd, Krugersdorp, South Africa; (Ausubel *et al*, 2003; Sambrook *et al*, 1989). Plasmid DNA was extracted from grown cultures with the QIAprep<sup>®</sup> Spin Miniprep Kit Protocol (Qiagen GmbH, Hilden, Germany). DNA concentrations were determined as in section 2.3.8. Positive clones were analysed by restriction enzyme digestion using *Eco* R1 (Promega, Madison, Wisconsin, USA) and DNA stored for possible future use.

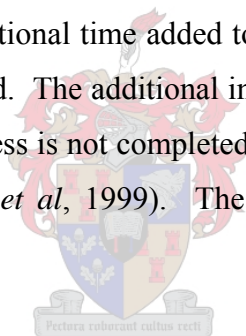
### 2.3.12 Full-length genome analysis



To amplify a possible HIV-1 recombinant sample (1154) the amplification of the near full-length 9 kb HIV-1 genome was attempted. Full-length genomic amplification is required to identify possible breakpoints in a recombinant genome (Salminen *et al*, 1995; Lole *et al*, 1999). Methods published previously by other authors (Dittmar *et al*, 1997; Fang *et al*, 1996; Gao *et al*, 1998; Loxton, 2004; Papathanasopoulos *et al*, 2003; Van Harmelen *et al*, 2001; Yamaguchi *et al*, 2003; Zur Megede *et al*, 2002) were used. Briefly, the Expand<sup>™</sup> Long Template PCR system (Roche Diagnostics, Mannheim, Germany) with two rounds of amplification was used to amplify genomic fragments longer than 2 kb. The system contains a unique combination of the thermostable *Taq* and *Tgo* DNA polymerases, with proofreading activity, to allow for the amplification of large DNA fragments (Frey and Suppmann, 1995; Hopfner *et al*, 1999). The methods attempt to amplify the full-length 9 kb genome with primers complementary to the 5' and 3' LTR regions and primers spanning the complete HIV-1 genome. The primers used for full-length amplification are presented in Table 2.8.



A combination of different forward and reverse primer pairs (such as MSF12/MSR5, MSF12/FGR94, FGF61/MSR5 and FGF61/FGR94) was used in an attempt to amplify the complete HIV-1 genome. A method was also used to amplify four smaller regions of the HIV-1 genome: the LTR-*gag*, *gag-pol*, *pol-env* and *env-LTR*. By this method a complete genome sequence can be obtained, which comprises overlapping genomic sequences (Yamaguchi *et al*, 2003). The primers used for this method were obtained from Dr. J. Hackett (Abbott Laboratories, Illinois, USA) and are presented in Table 2.9. Their position relative to the HXB2 (Ratner *et al*, 1985b) reference strain is indicated. The thermal cycling method was as follows: One cycle of denaturation at 94°C for 2 minutes followed by 15 cycles of denaturing at 94°C for 10 seconds, primer annealing (annealing temperature according to sequencing primers, Table 2.8) for 30 seconds and elongation at 68°C for 6 minutes. An additional 25 cycles of denaturing at 94°C for 10 seconds, primer annealing (annealing temperature according to sequencing primers, Table 2.8) for 30 seconds and elongation at 68°C for 10 minutes with 15-second increments (additional time added to the elongation step) per cycle at the elongation step were included. The additional increment steps ensure the addition of dNTPs in the elongation process is not completed before the next cycle starts (Frey and Suppmann, 1995; Hopfner *et al*, 1999). The samples were cooled down and stored at 4°C until used.



Attempts were also made to increase the PBMC DNA concentration from sample 1154 with the Repli-g<sup>®</sup> system (Qiagen GmbH, Hilden, Germany). The manufacturer's instructions were followed. The system is based on Multiple Displacement Amplification (MDA) technology allowing the genome amplification of up to 100 kb. The DNA polymerase utilised, phage  $\phi$ 29 DNA polymerase (Blanco *et al*, 1989), is stable and does not dissociate from the genomic DNA during replication. It has a 3'→5' exonuclease proofreading activity to maintain the high fidelity during replication. (Dean *et al*, 2002; Hosono *et al*, 2003; Yan *et al*, 2004). Briefly, the sample material is lysed releasing the genomic DNA. An alkaline lysis buffer denatures the DNA, which enables amplification to proceed across the whole genome with minimum sequence bias. After neutralisation, a master mix with the DNA polymerase is added to allow for the heat stable amplification of DNA overnight at 30°C.

**Table 2.8** HIV-1 full-length amplification primers

	Orientation	PCR Primers (5'-3')	Bases	*T <sub>m</sub> (50mM)	#Position (HXB2)	Reference
MSF12	Forward	AAATCTCTAGCAGTGGCGCCCGAACAG	28	69.0	650	Salminen <i>et al</i> , 1995
MSR5	Reverse	GCACTCAAGGCAAGCTTTATTGAGGCT	27	62.9	9606	Salminen <i>et al</i> , 1995
UP1A	Forward	AGTGGCGCCCGAACAGG	17	70.0	650	Gao <i>et al</i> , 1998
S2Full	Reverse	ATAAGAATGCGGCCGCTGCTAGAGA-TTTCCACACTACCA	40	70.3	9697	Zur Megede <i>et al</i> , 2002
Low2	Reverse	TGAGGCTTAAGCAGTGGGTTTC	22	66.3	9591	Gao <i>et al</i> , 1998
FGF61	Forward	TAGTCAGTGTGAAAATCTCTAGCAGT	27	66.3	636	Fang <i>et al</i> , 1996
FGR94	Reverse	CTCGATGTCAGCAGTTCTTGAAGTACTC	28	67.9	9397	Fang <i>et al</i> , 1996
FGF60	Forward	CAGACCCTTTTAGTCAGTGTGAAAATC	28	68.1	627	Fang <i>et al</i> , 1996
FGR53	Reverse	TACTTGTGTGCTATATCTCTTTTCTCC	29	67.2	5306	Fang <i>et al</i> , 1996
FGF46	Forward	GCATTCCTACAATCCCCAAAG	22	57.3	4669	Fang <i>et al</i> , 1996
FGR95	Reverse	GGTCTAACCCAGAGAGACCCAGTACAG	26	61.1	9532	Fang <i>et al</i> , 1996
OMFR1	Reverse	TGAGATCTCTAGTTACCAGAGTC	23	52.6	9662	Van Harmelen <i>et al</i> , 2001
F2NST	Forward	GCGAGGCTAGAAGAGAGAGATG	22	55.5	791	Van Harmelen <i>et al</i> , 2001
OFM19	Reverse	GCACTCAAGGCAAGCTTTATTGAGGCTTA	29	60.9	9604	Van Harmelen <i>et al</i> , 2001
626(+)	Forward	AGGGGGCCAAGTCGGCCTCTCTAGCAGT-GCGCCCGAACAGGG	43	69.5	626	Dittmar <i>et al</i> , 1997
9690(-)	Reverse	AGTCGCGGCCGCGGTCTGAGGGATCTCTAGT-TACCAGAGTC	41	70.3	9690	Dittmar <i>et al</i> , 1997
4955(+)	Forward	TAGTAGACGTCTGAAAAGGTGAAGGGCAGTAGTA	35	62.7	4955	Dittmar <i>et al</i> , 1997
9624(-)	Reverse	TAAGGCGGCCGCGCAAGCTTTATTGAGGCTTAG	34	65.2	9624	Dittmar <i>et al</i> , 1997
5048(+)	Forward	TGTGTGACGTCACAGATGGCAGGTGATGATTGTGT	35	62.7	5048	Dittmar <i>et al</i> , 1997

\*T<sub>m</sub> – primer melting temperature

#The position of the primer relative to the start position of HXB2 is indicated

**Table 2.9:** HIV-1 primers used to amplify overlapping genomic regions

	Amplification Step	Orientation	PCR Primers (5'-3')	Bases	*T <sub>m</sub> (50mM)	#Position (HXB2)
<b>env-LTR</b>						
7496F	RT-PCR	Forward	CCTKGCYCTGGAAAGATACCA	22	57.2	7984
LTR9131R-G	RT-PCR	Reverse	CTCYCAGGCTCARATCTGGTCT	22	59.4	9635
7542F	2nd PCR	Forward	TGGGGCTGCTCTGGAAACT	20	60.4	8029
9110R	2nd PCR	Reverse	GCAAGAGAGACCCAGTACAG	20	55.4	9532
<b>pol-env</b>						
JH38	cDNA	Reverse	GGTGARTATCCCTKCCTAAC	20	52.9	8346
5541912R	cDNA & 1st PCR	Reverse	CTTTCGGGCTGTCTGGGTTCC	21	63.7	8399
pol4274F	1st PCR	Forward	ACAGCAGTACAGATGGCAGTATTCATTC	28	60.6	4776
LP7728R	1st PCR	Reverse	CCACTTGTCCAATGCCAATAAGTCTTGT	28	61.5	8195
poli2	1st PCR	Forward	TAAARACARYAGTACWAATGGCA	23	52.9	4766
pol4277F	2nd PCR	Forward	GCAGTACAGATGGCAGTATTCAT	23	56.4	4774
LP7725R	2nd PCR	Reverse	GTCCAATGCCAATAAGTCTTGTTTC	24	56.1	8193
<b>gag-pol</b>						
poli8R	cDNA	Reverse	TAGTGGGATGTGTACTTCTGAAC	22	52.6	5195
LPgag 3F	1st PCR	Forward	TTTCAGCCCAGAAGTAATACCCATGTTT	28	60.6	1305
LPpoli 2R	1st PCR	Reverse	ATCACCTGCCATCTGTTTTCCATAATCC	28	61.1	5036
LPgag 6F	2nd PCR	Forward	CAGCCCAGAAGTAATACCCATGTT	24	58.5	1304
LPpoli 4R	2nd PCR	Reverse	ACCTGCCATCTGTTTTCCATAATC	24	57.1	5037
<b>LTR-gag</b>						
R9737F	1st PCR	Forward	CTCTCTTGCTAGACCAGAT	19	51.2	476
M/O p24-6	1st PCR	Reverse	TGTGWAGCTTGTCRGCTC	19	56.1	1703

Primers provided by Dr. J. Hackett (Abbott Laboratories, Illinois, USA)

\*T<sub>m</sub> – primer melting temperature

#The position of the primer relative to the start position of HXB2 is indicated

## CHAPTER THREE

### 3. Results

	<b>PAGE</b>
3.1 Introduction	59
3.2 HIV-1 viral load assays	59
3.3 Serotyping with an <i>env</i> gp120 V3 cPEIA	61
3.4 PCR data	62
3.5 Sequencing data	66
3.6 DNA Cloning	67
3.7 Sequence and phylogenetic analysis	70
3.7.1 Sequence alignments	70
3.7.2 Phylogenetic tree analysis	71
3.7.2.1 Neighbour-joining phylogenetic trees	71
3.7.2.2 Sequence identity matrices	73
3.7.2.3 Models of evolution	74
3.7.2.4 Maximum likelihood phylogenetic trees	74
3.7.3 Similarity plots	93
3.7.4 Consensus sequences	95
3.7.5 Entropy values and conserved genomic regions	96
3.7.6 Co-receptor predictions	100
3.7.7 Envelope N-Glycosylation sites	103
3.7.8 HIV-1 drug resistant mutations	106
3.8 Near full-length characterisation of possible HIV-1 recombinant strains	108

## CHAPTER THREE

### 3. Results

#### 3.1 Introduction

A total of 127 samples were serotyped and genotyped for the purpose of this study. Serotyping was based on an *env* gp120 V3 cPEIA, while genotyping was performed on the *gag* p24, *env* gp41 IDR, *env* gp120 V3 and a 1.2 kb *pol* fragment of the HIV-1 genome. Detailed sequence and phylogenetic analysis were performed. Based on the preliminary sequencing results three *gag* p24 samples (2.4%) (1039, 1151 and 1154) were cloned before analysing them. The results are presented in this chapter.

#### 3.2 HIV-1 viral load assays

The viral load results determined for each patient are presented in Table 3.1. The RNA copy number ranges from 217 (sample 1144) to 3 637 244 (sample 1094) copies per ml. Seven samples (5.5%) (1029, 1033, 1039, 1063, 1139, 1150 and 1154) had viral RNA copy numbers that were lower than the assays detection limit (LDL).

**Table 3.1:** RNA viral load results using the Abbott LCx<sup>®</sup> HIV RNA Quantitative assay

Sample	Viral load (RNA copies / ml)	Sample date	Sample	Viral load (RNA copies / ml)	Sample date
1001	3 295	2002/09/09	1024	143 471	2002/08/26
1002	66 779	2002/08/19	1025	66 940	2002/08/27
1003	262 962	2002/08/19	1026	107 270	2002/08/27
1005	4 290	2002/08/19	1027	4 269	2002/08/26
1006	111 907	2002/08/19	1029	*LDL	2002/08/27
1008	1 110	2003/11/21	1031	12 569	2002/08/27
1009	20 036	2002/08/20	1033	LDL	2002/08/28
1010	602 560	2002/08/27	1034	165 421	2002/09/03
1011	92 575	2003/11/20	1037	2 953	2002/08/26
1012	780	2003/05/06	1038	828 120	2002/09/09
1013	106 388	2002/08/19	1039	LDL	2002/09/09
1015	806	2002/08/26	1040	41 401	2002/08/28
1016	4 884	2002/08/27	1041	13 691	2002/08/19
1017	70 615	2002/08/26	1042	17 015	2002/08/19
1018	9 762	2002/08/20	1043	16 609	2002/08/19
1019	1 818	2002/08/26	1044	199 859	2002/08/26
1021	1 886	2002/08/20	1045	47 114	2002/08/19
1023	64 757	2002/08/27	1047	412 373	2002/08/19

\*LDL: Viral loads that were lower than the detection limit

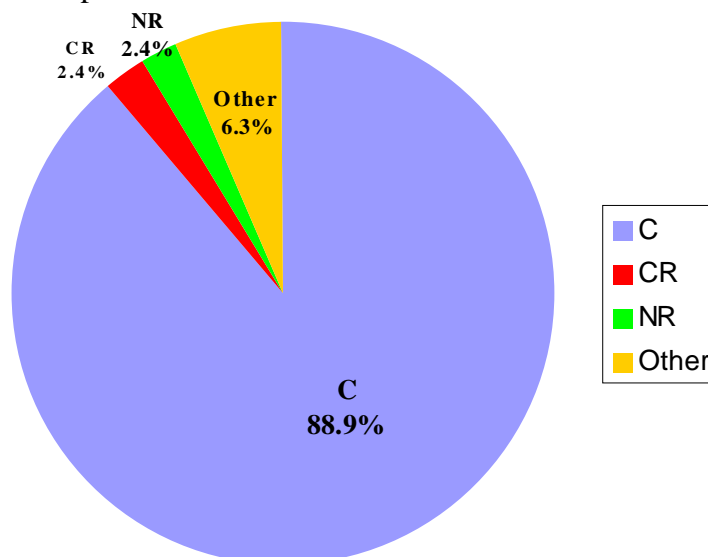
**Table 3.1 continue:** RNA viral load results using the Abbott LCx<sup>®</sup> HIV RNA Quantitative assay

Sample	Viral load (RNA copies / ml)	Sample date	Sample	Viral load (RNA copies / ml)	Sample date
1048	118 909	2002/08/19	1115	91 201	2003/11/20
1049	2 548	2002/08/19	1116	141 254	2003/05/06
1050	130 268	2002/08/19	1118	452	2003/05/05
1052	150 063	2002/08/27	1119	1 660	2003/05/05
1054	1 013 265	2002/08/27	1120	531 898	2003/05/06
1055	70 515	2002/08/27	1121	928 635	2003/05/05
1056	1 301	2002/08/27	1123	109 438	2003/05/05
1057	12 731	2002/08/19	1125	411 053	2003/11/20
1058	405 187	2002/08/20	1127	84 056	2003/05/07
1059	963 468	2002/08/27	1129	4 651	2003/05/12
1060	509	2002/09/03	1131	45 713	2003/05/06
1061	92 133	2002/08/26	1132	1 051	2003/05/20
1062	26 045	2002/08/26	1133	2 827	2003/11/20
1063	*LDL	2002/08/28	1134	37 927	2003/05/06
1064	65 069	2002/08/26	1135	437 938	2003/05/07
1067	11 597	2002/08/26	1136	28 706	2003/05/06
1068	67 448	2002/08/27	1137	59 107	2003/05/06
1069	2 497	2002/08/26	1138	154 762	2003/05/07
1072	44 639	2002/08/20	1139	LDL	2003/05/07
1073	4 929	2002/08/27	1140	87 813	2003/05/20
1075	88 309	2002/08/27	1141	378	2003/05/06
1076	2 740	2002/08/26	1142	152 703	2003/05/12
1077	44 099	2002/08/19	1143	89 393	2003/05/20
1078	74 554	2002/08/20	1144	217	2003/05/07
1079	4 817	2002/08/27	1146	1 574 180	2003/05/13
1083	3 286	2002/08/27	1147	440	2003/05/13
1084	1 817 719	2002/08/27	1148	24 547	2003/11/20
1088	1 352	2002/08/28	1149	228 079	2003/05/12
1089	97 158	2002/08/26	1150	LDL	2003/05/13
1090	92 260	2002/08/26	1151	438	2003/05/13
1094	3 637 244	2002/08/27	1152	25 442	2003/05/12
1096	39 284	2002/08/19	1153	27 289	2003/05/06
1097	307 768	2002/08/19	1154	LDL	2003/05/12
1098	135 870	2002/08/20	1155	1 711 249	2003/05/07
1099	448 447	2002/08/26	1156	27 809	2003/05/06
1100	120 288	2002/08/26	1157	4 178	2003/05/06
1101	24 302	2002/08/28	1160	38 875	2003/05/05
1102	41 449	2003/05/07	1162	14 150	2003/11/21
1104	LDL	2003/05/05	1163	5 813	2003/05/07
1106	81 283	2003/11/20	1165	205 265	2003/11/18
1108	376 429	2003/11/18	1169	24 118	2003/05/07
1109	29 512	2003/11/18	1172	265	2003/11/21
1110	213 796	2003/11/18	1173	7 563	2003/05/12
1112	100 807	2003/11/18	1174	1 090 299	2003/05/19
1113	41 497	2003/11/21	1175	1 328	2003/11/18
1114	269 091	2003/11/20			

\*LDL: Viral loads that were lower than the detection limit

### 3.3 Serotyping with an *env* gp120 V3 cPEIA

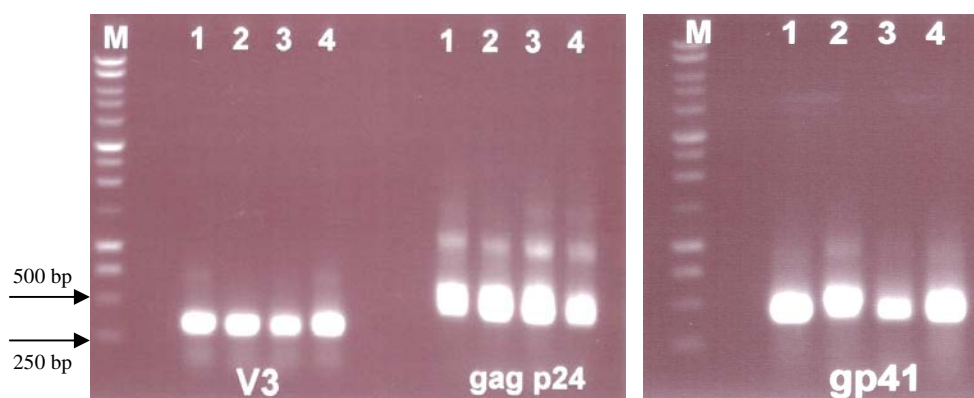
Serotyping was performed by comparing each plasma sample to a reference set of HIV-1 peptide strains representing subtypes A to E. A pie chart summary of the V3 cPEIA results is presented in Figure 3.1. The majority of the samples (89.9%) (113) reacted strongly with serotype C peptides. Three samples (2.4%) (1010, 1060 and 1083) were cross-reactive (CR) with all five serotype peptides (A, B, C, D and E), while 8 samples (6.3%) (1068, 1089, 1096, 1109, 1142, 1148, 1151, 1156) were CR with multiple peptide combinations, such as with peptides A and C or peptides A, C and E. They are indicated as “other” CR samples in Figure 3.1. A strong secondary serotype A reaction was noted in 38 (29.9%) serotype C samples, highlighting the potential cross-reactivity between serotypes A and C (data not shown). Cross-reactivity was thus responsible for 8.7% of the cohort samples that could not be accurately serotyped. In 3 samples (2.4) (1138, 1143 and 1162) no signal reduction activity was seen and they were characterised as non-reactive (NR). The 3 samples had relatively high viral loads (91 180, 148 088 and 14 150 copies per ml respectively) and the lack of activity seen was most probably due to peptide variation in the V3 region of the *env* gene and not their HIV-1 RNA copy number present in the plasma. Together, the CR and NR responses accounted for 11.1 % of samples that could not be accurately serotyped and had to be followed-up with genotyping methods, as described in chapter one.



**Figure 3.1: Serotype graph.** The Khayelitsha samples were serotyped with a V3 cPEIA. Anti-HIV-1 antibodies, present in the plasma sample of each HIV-1 positive patient, were tested against peptides of HIV-1 representing subtypes A to E. The majority of samples (88.9%) reacted with subtype C, indicated in blue. CR samples are shown in red and NR in green. Other samples that could not be serotyped with certainty are indicated in yellow.

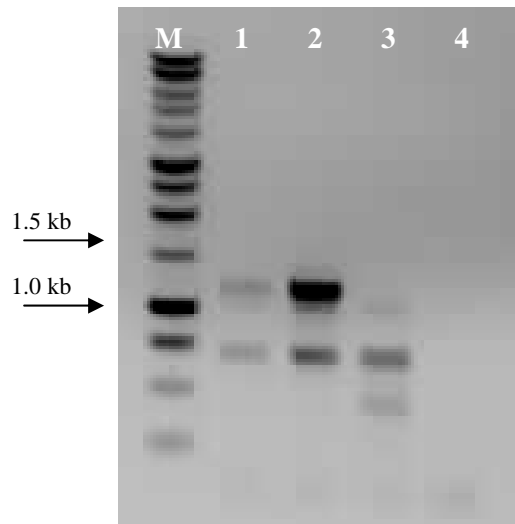
### 3.4 PCR data

PCRs were performed on the *gag* p24, *env* p41 IDR, *env* gp120 V3 and a 1.2 kb *pol* gene fragment of the HIV-1 genome. The *gag* p24, *env* gp41 IDR and *env* gp120 V3 amplification were performed on all 127 Khayelitsha samples. Figure 3.2 shows an example of positive amplification with samples TV167, TV671, TV48 and R3714, as viewed by agarose gel electrophoresis. These HIV-1 subtype C samples from the Department of Medical Virology, University of Stellenbosch, were previously identified as HIV-1 positive and used as controls in the PCR experiments. The amplification results and serotype for all these samples are summarised in Table 3.2. One-hundred-and-eleven samples (87.4%) were amplifiable in *gag* p24, 99 (78.0%) in *env* gp41 IDR and 107 (84.3%) in *env* gp120 V3. A 1.2 kb PCR, based on the *pol* gene of HIV-1, was also carried out on 3 possible HIV-1 recombinant samples [(1039, 1151 and 1154) (Figure 3.3; sequence data shown in section 3.6)]. This was done in an attempt to further characterise other areas of the possible recombinant genomes. A positive control was not included, as this method was not fully optimised at the time the PCR was performed. Although Figures 3.2 and 3.3 are both 0.8% agarose gels, they appear to be different, one with a black and one with a grey background. This is because the gel photographs that gave the best resolution, with the best clarity, are presented where possible.



**Figure 3.2: Example of a 0.8% agarose gel with the *env* gp120 V3, *gag* p24 and *env* gp41 IDR.** Lanes: Lane M – 1 kb DNA Marker; Lane 1- TV167; Lane 2 – TV671; Lane 3 – TV48; Lane 4 – R3714. These HIV-1 subtype C samples were tested positive previously and were used as positive controls during the study.





**Figure 3.3: *pol* PCR amplification on a 0.8% agarose gel.** Lanes: Lane M – 1 kb DNA Marker; Lane 1- 1039; Lane 2 – 1151; Lane 3 – 1154; Lane 4 – Negative control. The best amplification was seen with sample 1151. Two amplified bands are present. The *pol* band is 1.2 kb in length, while another unknown amplified fragment is present at  $\pm 750$  bp.

**Table 3.2:** Summary of serotyping and PCR results

Sample	Serotype	<i>gag p24</i>	<i>env gp41</i> IDR	<i>env gp120</i> V3
1001	C	+	+	-
1002	C	+	+	+
1003	C	+	+	+
1005	C	+	+	+
1006	C	+	+	+
1008	C	+	+	+
1009	C	+	+	+
1010	CR	+	-	+
1011	C	+	+	+
1012	C	+	+	+
1013	C	+	+	+
1015	C	+	+	-
1016	C	+	+	-
1017	C	+	+	+
1018	C	+	+	+
1019	C	-	-	-
1021	C	+	+	+
1023	C	+	+	+
1024	C	+	+	+
1025	C	-	+	+
1026	C	+	+	+
1027	C	-	+	+
1029	C	+	+	-
1031	C	+	+	+
1033	C	+	-	+
1034	C	+	+	+
1037	C	+	+	+
1038	C	+	-	+
1039	C	+	-	-
1040	C	+	+	+
1041	C	-	-	+
1042	C	+	+	+

**Table 3.2 continue:** Summary of serotyping and PCR results

Sample	Serotype	<i>gag</i> p24	<i>env gp41</i> IDR	<i>env gp120</i> V3
1043	C	+	+	+
1044	C	+	+	+
1045	C	+	+	+
1047	C	+	+	+
1048	C	+	+	+
1049	C	+	+	+
1050	C	+	+	+
1052	C	+	+	+
1054	C	+	+	-
1055	C	+	-	+
1056	C	+	-	-
1057	C	+	+	+
1058	C	+	+	+
1059	C	+	+	+
1060	CR	+	+	-
1061	C	+	+	-
1062	C	+	-	+
1063	C	+	-	-
1064	C	+	+	+
1067	C	+	+	-
1068	A/C	+	+	+
1069	C	+	+	-
1072	C	+	-	-
1073	C	+	-	+
1075	C	+	+	+
1076	C	+	+	+
1077	C	+	+	+
1078	C	-	-	+
1079	C	+	+	+
1083	CR	+	+	+
1084	C	+	-	+
1088	C	+	+	-
1089	C/E	+	+	+
1090	C	+	+	+
1094	C	+	+	+
1096	A/B/C/E	+	+	+
1097	C	+	+	+
1098	C	+	+	+
1099	C	+	+	+
1100	C	-	+	+
1101	C	+	-	+
1102	C	+	+	+
1104	C	+	+	+
1106	C	+	+	+
1108	C	+	+	+
1109	A/C/D/E	-	+	-
1110	C	+	+	+
1112	C	+	+	+
1113	C	+	+	+
1114	C	+	+	+
1115	C	+	-	+
1116	not done	+	+	+
1118	C	+	+	+
1119	C	+	+	+

**Table 3.2 continue:** Summary of serotyping and PCR results

Sample	Serotype	<i>gag</i> p24	<i>env gp41</i> IDR	<i>env gp120</i> V3
1120	C	+	+	+
1121	C	+	+	+
1123	C	+	+	+
1125	C	+	+	+
1127	C	+	-	+
1129	C	+	-	+
1131	C	+	+	+
1132	C	+	-	+
1133	C	+	-	-
1134	C	-	+	+
1135	C	+	+	+
1136	C	+	+	-
1137	C	+	+	+
1138	NR	+	+	+
1139	C	-	-	-
1140	C	+	-	+
1141	C	+	+	+
1142	B/C	+	+	+
1143	NR	+	+	+
1144	C	-	-	+
1146	C	+	+	+
1147	A/C	-	+	+
1148	C	+	-	+
1149	C	-	+	-
1150	C	-	-	-
1151	A/C	+	+	+
1152	C	+	+	+
1153	C	+	+	+
1154	C	+	-	+
1155	C	+	+	+
1156	C/D	+	+	+
1157	C	+	-	+
1160	C	+	+	+
1162	NR	+	+	+
1163	C	-	+	+
1165	C	-	+	+
1169	C	+	+	+
1172	C	-	+	+
1173	C	+	+	+
1174	C	+	+	+
1175	C	-	-	+
		<b>111</b>	<b>99</b>	<b>107</b>
		<b>87.4</b>	<b>78.0</b>	<b>84.3</b>

### 3.5 Sequencing data

All positively amplified PCR products were sequenced. RNA was used for PCR amplification, except for the *gag* p24 samples of the 3 possible HIV-1 recombinant strains (1039, 1151 and 1154). There were only 2 samples (1.6%) (1139 and 1150) that had no sequence data. For the *gag* p24, *env* gp41 IDR and *env* gp120 V3 analysis 82 samples (64.6%) were amplifiable in all three fragments, in 36 samples (28.3%) only two and in 9 samples (7.1%) only one fragment. In these 9 samples recombination cannot be ruled out, as analysis is only based on a small section of the HIV-1 genome. A further ten samples (7.9%) (1001, 1019, 1039, 1054, 1060, 1109, 1133, 1139, 1149, 1150) were amplifiable in the *env* gp120 V3, but could not generate positive sequences. Double peaks present in the sequences suggest that these patients might be dually infected with two or more diverse HIV-1 quasispecies. Table 3.3 gives an indication of the genomic position of each of the fragments compared to the reference strain HXB2. With the *pol* PCR the sequences from samples 1039 and 1151 resulted in a 1050 DNA sequence product (position 2289 to 3339 relative to HXB2). Only 227 bp from sample 1154 could be sequenced.

**Table 3.3:** Position of amplified fragments compared to HXB2

	<i>gag</i> p24	<i>env</i> gp41	<i>env</i> gp120 V3	<i>pol</i>
<b>Base Pairs</b>	444	441	261	1050
<b>Amino Acids</b>	148	147	87	350
<b>Base Position</b>	1258-1701	7860-8301	7002-7267	2289-3339

All 113 individuals who were serotyped as subtype C were also genotyped as subtype C, showing that serotyping is an efficient screening tool, but not sufficient enough to do detailed molecular analysis with. The remaining 11.1% of samples had to be genotyped to confirm their HIV-1 subtype. The best success rate of PCR amplification was seen with the *gag* p24 (87.4%). This is expected, as the *gag* p24 region is more conserved than the *env* gp41 IDR and *env* gp120 V3 regions. Even though the *env* gp41 IDR is more conserved than the *env* gp120 V3 region, more samples were amplifiable in the *env* gp120 V3 region (84.3%). A possible reason might be that the *env* gp120 V3 primers were better optimised, or more specific, for amplifying HIV-1 subtype C than the *env* gp41 IDR primers used.

Less variation in a population of viruses means that the chances of complementary primers in a PCR recognising more of these strains are much higher. The success of amplification is often dependent on the starting concentration, or viral load, of a sample. The more RNA copies present, the better the success of amplification. Optimal genotyping analysis, as well as to rule out any possible recombination, requires that more than one genome region should be targeted for amplification.

### 3.6 DNA Cloning

Based on the initial *gag* p24 BLAST results, 3 samples (1039, 1151 and 1154) were identified as possible HIV-1 recombinant strains. The data from these 3 samples is summarised in Table 3.4. PBMCs were used to extract genomic DNA and amplify the *gag* p24 region of these patient viruses. The PCR DNA was gel-purified (Figure 3.4) and cloned to identify possible quasispecies amongst them.

**Table 3.4:** Initial BLAST results and sequence analysis of unusual HIV strains

Samples	Genotype (BLAST results)			V3 Serotype	Viral load (RNA copies / ml)
	<i>gag</i> p24	<i>env</i> gp41 IDR	<i>env</i> gp120 V3		
1039	*A/C (474 bp)	†N/A	N/A	C	††LDL
1151	#Complex (470 bp)	C	C	**A/C	438
1154	Complex (472 bp)	N/A	C	C	LDL

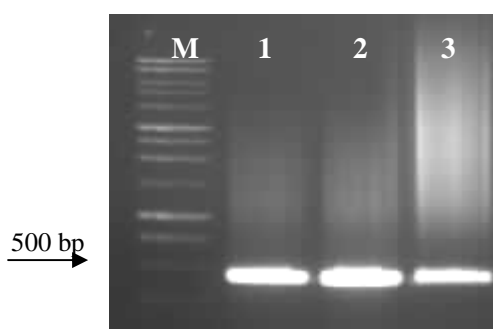
†N/A – Data not available; regions could not be amplified

††LDL – Viral loads that were lower than the detection limit

\*A/C – Samples are possible subtype A and subtype C recombinant strains

\*\*A/C – The serotype of this sample was characterised as an A/C serotype

#Complex – Samples are possible complex recombinant forms from multiple HIV-1 subtypes

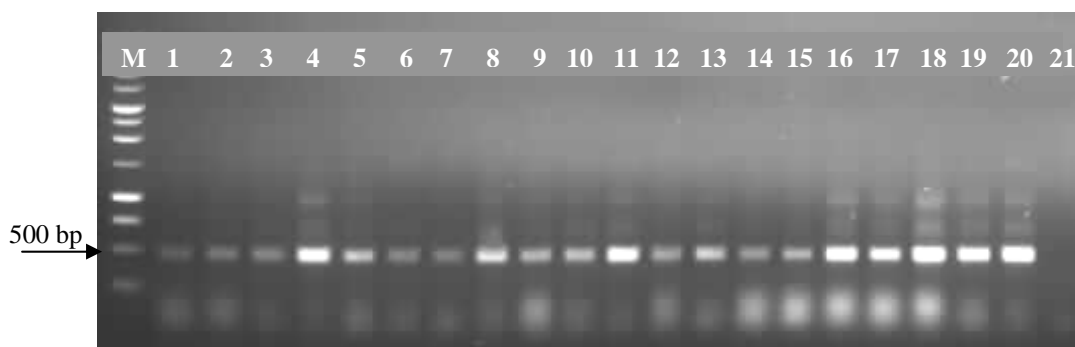


**Figure 3.4: *gag* p24 PCR fragments used for cloning.** The fragments were gel-purified and cloned into the pGem T-easy Vector system. Lanes: Lane M – 1 kb DNA Marker; - Lane 1 - 1039; Lane 2 – 1151; Lane 3 – 1154.

The HIV-1 DNA was inserted into an *E.Coli* JM109 High Efficiency competent bacterial vector. The *E. Coli* cells, containing the vector, were plated onto a LB media petri dish and incubated overnight at 37°C, as described in chapter two. The average number of colonies per plate observed is summarised in Table 3.5. The positive control had an average of 192 colonies per plate, of which 177 were white and 15 blue. This means that at least 92.4% of the observed colonies were presumed positive. The background control only had blue colonies as expected, as the vector contained no DNA insert. Samples 1039, 1151 and 1154 had an average of 21, 26 and 7 colonies per plate respectively. The observed cloning efficiency for 1039, 1151 and 1154 was 4.8%, 53.8% and 14.3% respectively, if assuming that all the white colonies were positive and the blue colonies negative. The high success rate of sample 1151 was probably due to better quality DNA and a higher concentration of HIV-1 DNA products present in the sample. The viral load of sample 1151 was 438 RNA copies per ml, while the viral loads of samples 1039 and 1154 were undetectable. DNA from the positive clones from the colony PCRs (Figure 3.5) were sequenced and sequence analysis were performed with the rest of the Khayelitsha sequences obtained.

**Table 3.5:** Average number of colonies per plate observed

Sample	Blue	White	Total
Positive Control	15	177	192
Background Control	11	0	11
1039	20	1	21
1151	12	14	26
1154	6	1	7



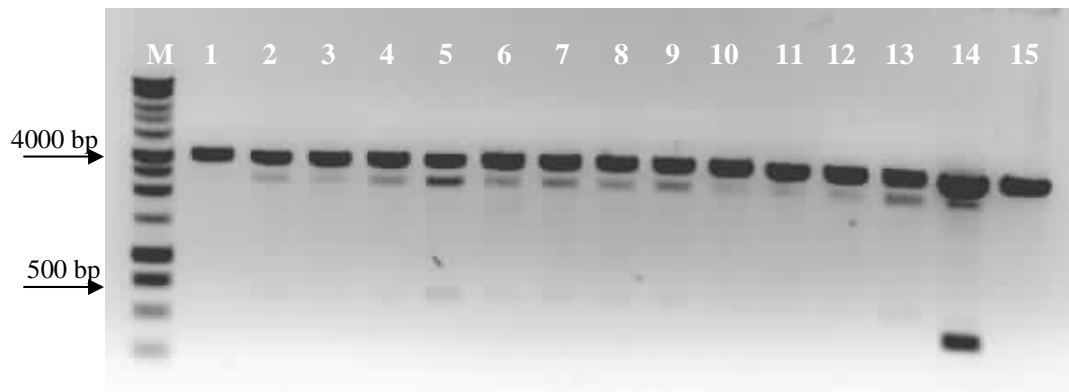
**Figure 3.5: A gag p24 PCR from cultures grown overnight.** Lanes: Lane M – 1 kb DNA Marker; Lane 1 to 20 – the cloned gag p24 fragments from sample 1154, 20 different colonies were picked and amplified; Lane 21 – negative control.

Twenty colonies from each sample were also picked and grown overnight at 37°C, as described in chapter two. DNA was extracted from the cultures and DNA concentrations determined with the Nanodrop system. The concentration for each respective colony is presented in Table 3.6. A partial restriction enzyme digestion with *Eco* R1 (Figure 3.6) was also performed to identify the cloned vector with, as well as without the insert. The top DNA fragment visible at 4000 bp is the vector containing the DNA insert, followed by the vector without the insert and the cut *gag* p24 cloned fragment at 500 bp. In Figure 3.6 the sample in lane one had an unsuccessful enzyme digestion, as this colony probably did not contain any inserted HIV-1 DNA fragments. In lane fifteen no digestion took place, as this was a background control sample. The *gag* p24 bands are only partially visible, as the digestion was incomplete. The bands are much clearer on the Syngene™ GeneGenius computer system with which the gel photos were taken with. Although the DNA quality was poor, with purity values ranging from 1.02 to 2.07, agarose gel electrophoresis and DNA sequence analysis proved that the cloning reactions were successful (Figures 3.5 to 3.6).

**Table 3.6:** DNA concentrations of cloned samples

n	1039			1151			1154		
	ng/μl	A260	Purity	ng/μl	A260	Purity	ng/μl	A260	Purity
1	65.11	1.302	1.05	201.07	4.021	1.91	76.42	1.53	1.95
2	60.58	1.212	1.02	100.25	2.005	1.91	169.05	3.38	1.42
3	69.73	1.395	1.05	39.84	0.797	1.89	102.80	2.06	1.95
4	76.78	1.536	1.13	18.35	0.367	1.91	80.63	1.61	1.96
5	84.07	1.681	1.10	12.66	0.253	1.32	50.08	1.00	1.99
6	49.15	0.983	1.20	80.85	1.617	1.83	45.52	0.91	2.02
7	45.62	0.912	1.36	97.64	1.953	1.84	66.65	1.33	1.98
8	46.34	0.927	1.26	96.3	1.926	1.83	66.21	1.32	2.01
9	51.61	1.032	1.21	122.65	2.453	1.85	88.53	1.77	1.99
10	51.50	1.030	1.34	97.22	1.944	1.80	63.66	1.27	1.94
11	60.98	1.220	1.25	66.53	1.331	1.97	79.82	1.60	1.96
12	53.71	1.074	1.32	47.99	0.960	2.04	83.94	1.68	1.91
13	41.68	0.834	1.32	61.65	1.233	1.97	75.92	1.52	1.99
14	50.53	1.011	1.39	80.87	1.617	1.96	89.45	1.79	1.97
15	41.44	0.829	1.19	85.41	1.708	1.93	105.02	2.10	1.95
16	45.76	0.915	1.10	30.22	0.604	2.07	75.56	1.51	1.93
17	35.12	0.702	1.43	87.79	1.756	2.04	90.65	1.81	1.94
18	36.96	0.739	1.25	83.72	1.674	1.95	52.56	1.05	1.94
19	32.37	0.647	1.04	66.89	1.338	1.95	63.04	1.26	1.93
20	50.66	1.013	1.33	67.96	1.359	2.03	70.58	1.41	1.94
*+	80.60	0.124	1.60						
*B	17.55	0.027	4.50						

\*A positive (+) and background (B) control was included

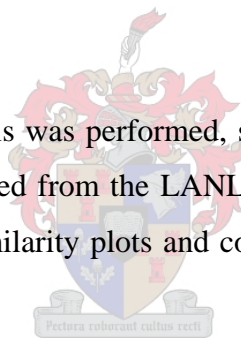


**Figure 3.6 Partial restriction enzyme digestion of cloned *gag* p24 cultures.** Lanes: Lane M – 1 kb DNA marker; Lanes 1 to 13 – sample 1154 restriction enzyme digestion, representing 13 different clones; Lane 14 – positive control; Lane 15 – uncut vector used as a negative control. The *gag* p24 bands are faintly visible at  $\pm$  500 bp and can be more easily seen on the Syngene™ GeneGenius computer system.

### 3.7 Sequence and phylogenetic analysis

#### 3.7.1 Sequence alignments

Before detailed sequence analysis was performed, sequences were aligned with each other and with sequences obtained from the LANL. These alignments were used to construct phylogenetic trees, similarity plots and consensus sequences as used in the analysis.



With the *gag* p24 alignment no amino acid insertions or deletions (indels) were observed. In the *env* gp41 only one sequence, from sample 1043, had a two amino acid deletion at position 8073 to 8078 relative to the HXB2 reference strain. Compared to the other Khayelitsha samples the first amino acid deletion resulted in either an Arginine or Lysine deletion, while the second amino acid deletion is most likely Serine. As the result of to the highly variable nature of the V3 region, many deletions in different samples were observed. The sequence from sample 1123 had a Metionine insertion at position 7228 to 7230 relative to HXB2. This was the only V3 insertion observed. In the V3 only one amino acid sequence, from sample 1098, contained the GPGR tetramer motif, while the remaining sequences contained the GPGQ motif. The complete amino acid sequence alignments are given in Appendix B. No indels were detected in the *pol* alignment for the sequences from samples 1039 and 1151. The 227 *pol* bp region of sample 1154, most likely a recombinant HIV-1



strain, could not be aligned to any of the reference strains present in the alignment. The BLAST results indicate that this sequence has the closest similarity to a 980 bp Chinese HIV-1 subtype C sequence, Binyang\_610, accession number AY635754 (data not shown).

### 3.7.2 Phylogenetic tree analysis

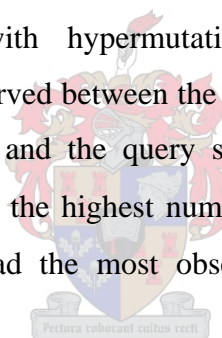
Detailed sequence analysis was performed by drawing neighbour-joining and maximum likelihood phylogenetic trees. The *gag* p24, *env* gp41 IDR, *env* gp120 V3 and *pol* neighbour-joining phylogenetic trees were drawn with reference strains obtained from the LANL. Bootstrap values are indicated where possible, with a distance scale drawn on each phylogenetic tree. The trees were rooted with HIV-1 sequences distantly related, usually the most common ancestor of the phylogenetic group, from the group of sequences being analysed. The HIV-1 group M reference trees were rooted with a HIV-1 group O sequence, Ant70. The HIV-1 subtype D phylogenetic trees were rooted with a subtype C reference sequence, ETH2220, while the HIV-1 subtype B phylogenetic trees were rooted with a subtype B reference sequence, HXB2.

#### 3.7.2.1 Neighbour-joining phylogenetic trees

The neighbour-joining phylogenetic trees are presented in Figures 3.7 to 3.10 and Figures 3.12 to 3.14. A clear phylogenetic distinction, supported by high bootstrap values, can be made between the HIV-1 subtypes (A-D, F-H, J, K). In Figure 3.7 the 3 cloned *gag* p24 samples (1039, 1151 and 1154), each with ten representative sequences, are phylogenetically compared to the set of reference sequences used. The sequences from samples 1039 and 1151 group with the subtype C reference sequences, while the sequences from sample 1154 clearly group with the subtype D reference sequences. No quasispecies were detected amongst these cloned variants. The sequences from each of the cloned samples are grouped together with a 100% bootstrap value. This indicates that the major HIV-1 variant, present in the virus population of each patient, is being resampled multiple times.

In Figure 3.8 the *gag* p24 phylogenetic tree, with all the Khayelitsha *gag* p24 sequences, is presented. The representative sequences of the cloned samples (1039, 1151 and 1154) are shown in blue. All the Khayelitsha sequences, with the exception of the sequence from sample 1154, group with the HIV-1 subtype C sequences. Interestingly, sample pairs 1010 and 1018, 1113 and 1115, as well as 1142 and 1143 group closely together with 100% bootstrap values. These sequences might be from samples originating from patients that are epidemiologically linked.

In Figure 3.9 and Figure 3.10 the *env* gp41 IDR and *env* gp120 V3 phylogenetic trees are presented. All the Khayelitsha sequences from these genome regions cluster with HIV-1 subtype C. In Figure 3.10 unusual long branch lengths are noted for the sequences from samples 1024, 1040, 1041, 1090, 1096, 1108 and 1154. The short, highly variable V3 bp sequences and possible hypermutations (Figure 3.11) might account for this observation. Table 3.7 highlights the amount of observed mutations (G↔A) usually associated with hypermutations. A considerable nucleotide composition difference was observed between the consensus subtype C sequence, the subtype C reference sequences and the query sequences from Khayelitsha. The sequence from sample 1041 had the highest number of G→A transitions, while the sequence from sample 1108 had the most observed differences from the query sequences.



In Figure 3.12 the subtype D *gag* p24 sequence of sample 1154 was compared to a group of randomly selected subtype D sequences from the LANL. The sequences all originated from African patients (Cameroon, Chad, DRC, Gabon, Kenya, Senegal, South Africa) where subtype D is commonly found, with the exception of MP613 (France) and AMK10 (USA). The sequence from 1154 forms a phylogenetic group with the South African subtype D sequences.

The *pol* sequences obtained from the possible HIV-1 recombinant samples 1039 and 1151 were also phylogenetically compared with the group of *pol* reference sequences (Figure 3.13). The 227 bp *pol* sequence from sample 1154 could not be aligned to any set of reference sequences and was subsequently not used in phylogenetic analysis. The sequences from samples 1039 and 1151 clearly group with the HIV-1

subtype C reference sequences. In Figure 3.14 these sequences form a phylogenetic relationship with two other South African sequences, TV001 and 03ZASK011B2.

**Table 3.7:** Possible hypermutations in the *env* gp120 V3 sequences

Sequence names	*N	#Percentage Gs	A→G	G→A	†Dinucleotide context			
					GG	GA	GC	GT
C.BW.96.96BW0502	30	13.0	8	6	0	5	0	1
C.IN.95.95IN21068	15	2.2	5	1	0	0	0	1
C.ET.86.ETH2220	23	22.2	4	1	0	0	0	1
C.BR.92.92BR025	29	10.9	8	5	0	4	0	1
1024	42	8.7	7	4	0	3	1	0
1040	44	6.5	10	3	0	3	0	0
1041	47	6.5	16	3	0	1	1	1
1090	44	8.7	11	4	0	1	1	2
1096	55	6.5	8	3	1	2	0	0
1108	58	13.0	10	6	1	3	0	2
1154	49	10.9	4	5	1	1	0	3

\*The total number of positions in which the given sequence differs from the reference sequence.

#The percentage of Gs in the reference sequence that have undergone G↔A transitions.

†A tally of the dinucleotide contexts of the G↔A transitions. This represents two contiguous bases in the reference strain and summarises the context of changes.

### 3.7.2.2 Sequence identity matrices

The *gag* p24 subtype D nucleotide similarity of sample 1154, compared to the randomly selected subtype D sequences used in phylogenetic analysis, is presented in Table 3.8. The sequence has the highest similarity with D\_ZA\_84\_R2, a South African subtype D sequence originating from a 1984 patient sample. The lowest similarity is noted with a sequence from the DRC, KS26. This relationship is also evident in Figure 3.12 with the South African sequences clustering closely together, with sequence KS26 very distantly related to the South African sequences. The *pol* subtype C nucleotide similarities of the sequences from samples 1039 and 1151 are noted in Table 3.9. The sequence from sample 1039 has the closest nucleotide similarity with a Somalian sequence, SM145. The sequence from sample 1151 has the closest similarity with SM145, as well as 98TZ013, a Tanzanian sequence.

Interestingly, the *pol* sequences do not have the closest nucleotide similarity to the South African sequences, as the phylogenetic relationship (Figure 3.14) indicates.

### 3.7.2.3 Models of evolution

The neighbour-joining phylogenetic trees were all drawn with the Kimura 2-parameter model of evolution. This method is computationally fast and reliable and has been used extensively in HIV-1 sequence analysis. However, not all data can be treated the same as rates of DNA evolution differ, as described in chapter one. Modeltest was thus used to test for the most reliable model of evolution to apply in order to construct more accurate maximum likelihood phylogenetic trees. Each alignment was tested against the 56 mathematical models of evolution currently available. None of the alignment scores predicted the Kimura 2-parameter model of evolution to be the correct model to use in constructing phylogenetic trees. The relationship with the *gag* p24 cloned fragments should be drawn with the TrN + I + G model. The *gag* p24 reference tree, with all the Khayelitsha sequences, should be drawn with the TIM + I (Invariant sites) + G (gamma-distribution) model, and the *gag* p24 subtype D sequence tree with the TrN + I + G model. The GTR model with I + G should be used for the *env* gp41 IDR, *env* gp120 V3 and the *pol* phylogenetic trees.



### 3.7.2.4 Maximum likelihood phylogenetic trees.

The maximum likelihood phylogenetic trees were drawn with PAUP, as described in section 2.3.10.2. The trees were based on the evolution models as obtained from the Modeltest software. In Figure 3.15 the *gag* p24 sequences from the cloned samples 1039, 1151 and 1154 were re-evaluated. The sequences from samples 1039 and 1151 form a phylogenetic group with HIV-1 subtype C, while the sequences from sample 1154 form a phylogenetic group with HIV-1 subtype D.

The ten representative sequences from each clone do not however form a monophyletic cluster amongst themselves, as in Figure 3.7. The sequences from the clones of sample 1039 seem to form three distinct phylogenetic clusters. The sequences from clones 1, 3, 4 and 8 seems to be closer related to each other than the sequences from clones 5, 6, 7, 9, 10 with the sequence from clone 2 separating the

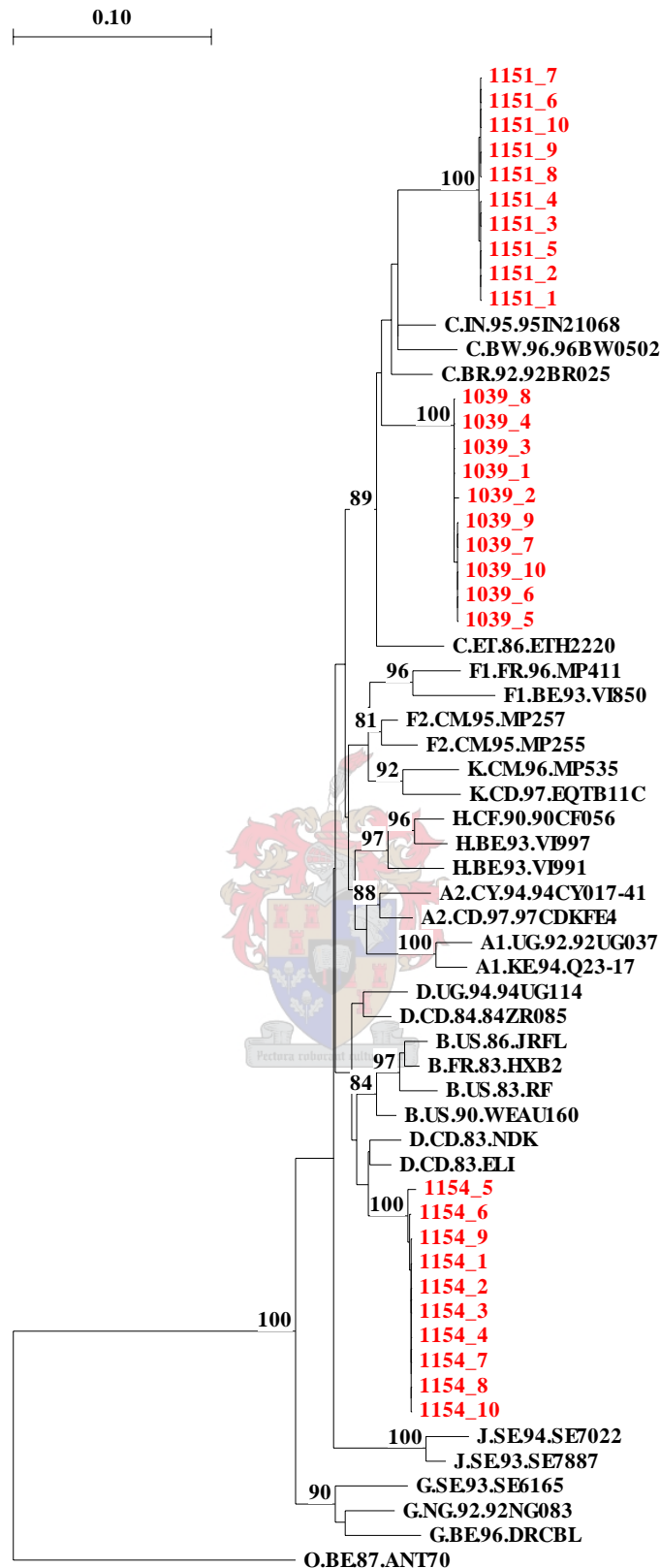
two clusters. These phylogenetic relations are the result of single bp differences noted in the *gag* p24 alignment containing the sequences from the cloned samples. The sequences from clones 5, 6, 7, 9 and 10 contain a G bp instead of an A at position 100 in the *gag* p24 gene. Clone 2 differs with a G bp at position 407 in its sequence, compared to an A bp present in the other sequences. The sequences from clones 1, 2, 3, 4, 5 and clones 6, 7, 8, 9, 10 from sample 1151 also seem to be phylogenetically different. Sequence differences for the sequences from sample 1151 are located at position 186, with the sequences either containing an A or a G bp. This is also noted for the sequences from sample 1154, with clones 5 and 6 forming a separate phylogenetic cluster. Sequence differences are located at positions 21 (T instead of C), 91 (T instead of C) and 199 (G instead of A) of the sequence alignment.

In Figure 3.16 and Figure 3.17 the *gag* p24 and *env* gp41 IDR maximum likelihood phylogenetic trees, drawn with the reference sequences obtained from the LANL, are presented. Bootstrap values are not indicated on these trees, as it is too computer intensive and the PAUP and Treepuzzle software are unable to handle bootstrapping the large datasets. Closely related sequence pairs are indicated in purple on the phylogenetic trees. The *gag* p24 pairs, 1010 and 1018, 1113 and 1115, as well as 1142 and 1143, correlate with the observation seen in the neighbour-joining phylogenetic tree (Figure 3.8). However, in the *env* gp41 maximum likelihood phylogenetic tree five closely related sequence pairs (1027 and 1112, 1033 and 1038, 1057 and 1059, 1064 and 1076, as well as 1113 and 1115) are identified. The samples these sequences were obtained from might be epidemiologically linked, having infected each other. This might especially be true for samples 1113 and 1115, who have closely related sequences in both the *gag* p24 and *env* gp41 IDR. The *env* gp120 V3 maximum likelihood tree is not presented. The short branch lengths of the *env* gp41 IDR maximum likelihood tree indicate that this region of the *env* genome is conserved amongst the Khayelitsha sequences. The short sequence length of the V3 alignment (261 bp), together with the large data set, results in a phylogenetic tree unable to distinguish between HIV-1 group M subtypes (data not shown).

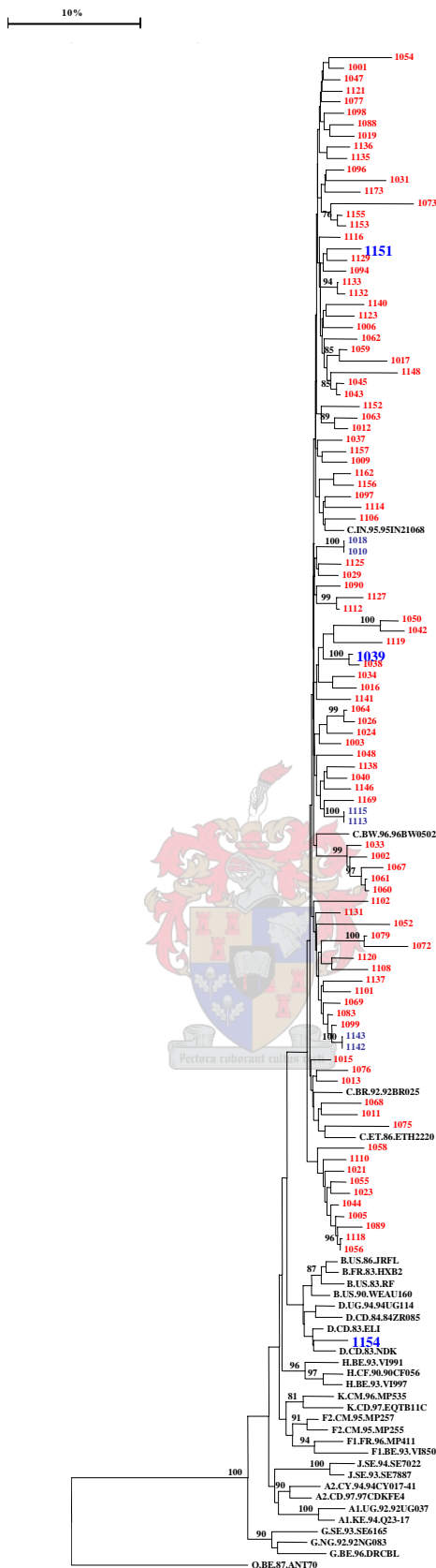
The *gag* p24 subtype D maximum likelihood phylogenetic tree (Figure 3.18) confirms the close relationship the sequence from sample 1154 has with the South African subtype D sequences, also noted in the neighbour-joining phylogenetic tree (Figure

3.12). Bootstrap values are not indicated as the subtype D *gag* p24 sequences are very similar and bootstrapping the tree would collapse the branches. This would lead to a phylogenetic tree with very short branch lengths, and the relationship between the different subtype D sequences would not be comparable with each other. Although the virus from sample 1154 has only been characterised during this study, the other South African subtype D sequences originate from patients sampled as early as 1984 (D\_ZA\_84\_R2). The phylogenetic relationship between the South African subtype D sequences indicates that this subtype has been present, and still is, since the beginning of the epidemic in South Africa.

The *pol* maximum likelihood phylogenetic tree (Figure 3.19) confirms that the *pol* sequences from samples 1039 and 1151 belong to HIV-1 group M subtype C. However, in (Figure 3.20) the sequences from 1039 and 1151 do not form a phylogenetic group with each other. The sequence from sample 1039 form a cluster with the sequences SM145, 97TZ05 and SK065B1. These sequences originate from Somalia, Tanzania and South Africa respectively. The sequence from sample 1151 forms a phylogenetic cluster with sequences 261 and 98TZ013. These sequences originate from Denmark and Tanzania respectively. The subtype C phylogenetic trees indicate that the subtype C variants present in South Africa are not unique to this country. The subtype C virus clade has probably been introduced into the South African population from neighbouring countries such as Botswana, with multiple introductions taking place. The phylogenetic results and observations are discussed in more detail in chapter four.

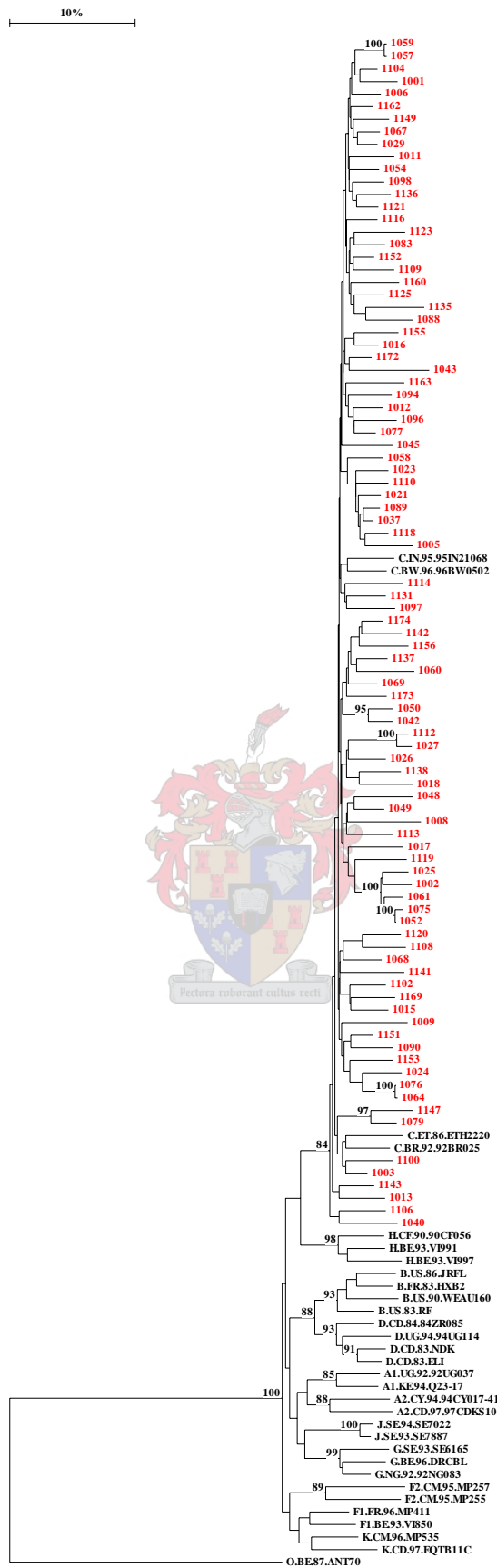


**Figure 3.7: A *gag* p24 neighbour-joining phylogenetic tree with the 3 cloned fragments 1039, 1151 and 1154.** Bootstrap values are indicated. The sequences from the cloned samples are indicated in bold red. Ten different clones of each sample were sequenced and analysed. The *gag* p24 sequence alignment was 444 bp in length. Branch lengths are drawn to scale. Sequences from samples 1039 and 1151 group with the subtype C reference sequences, while sequences from sample 1154 group with the subtype D reference sequences.

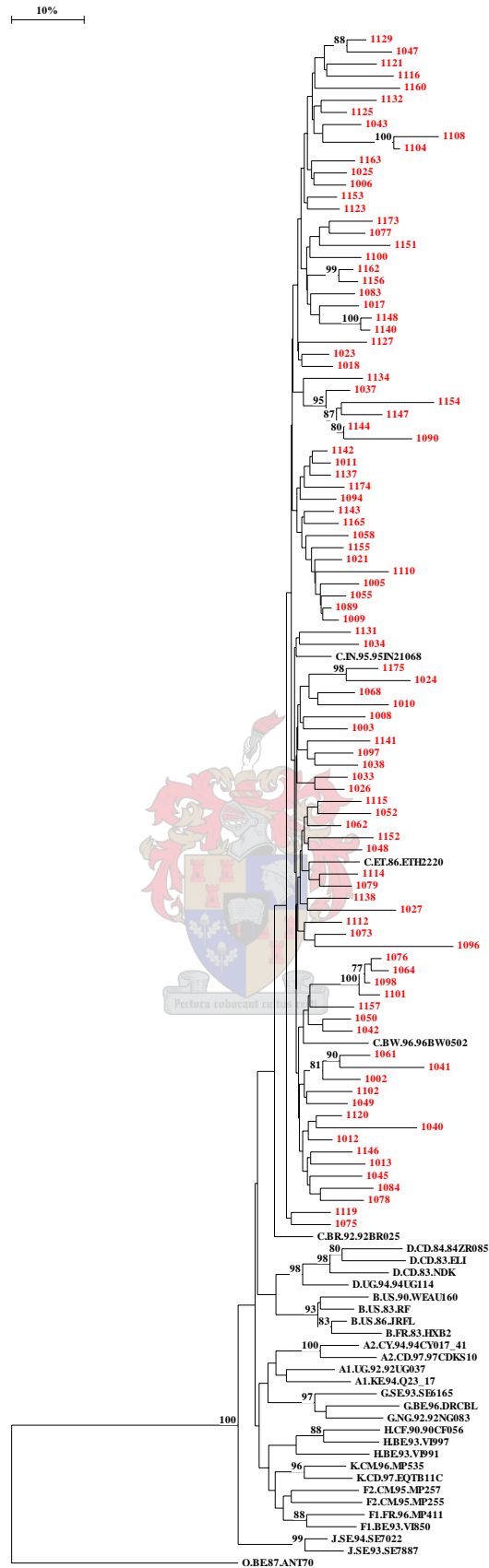


**Figure 3.8: The *gag* p24 neighbour-joining phylogenetic tree.** Bootstrap values are indicated. The sequences from the cloned samples (1039, 1151 and 1154) are indicated in blue, while sample pairs 1010 and 1018, 1113 and 1115, as well as 1142 and 1143 are indicated in purple. The rest of the Khayelitsha sequences are indicated in bold red. The *gag* p24 sequence alignment was 444 bp in length. Branch lengths are drawn to scale. All the Khayelitsha sequences, with the exception of 1154, group with the HIV-1 subtype C reference sequences.

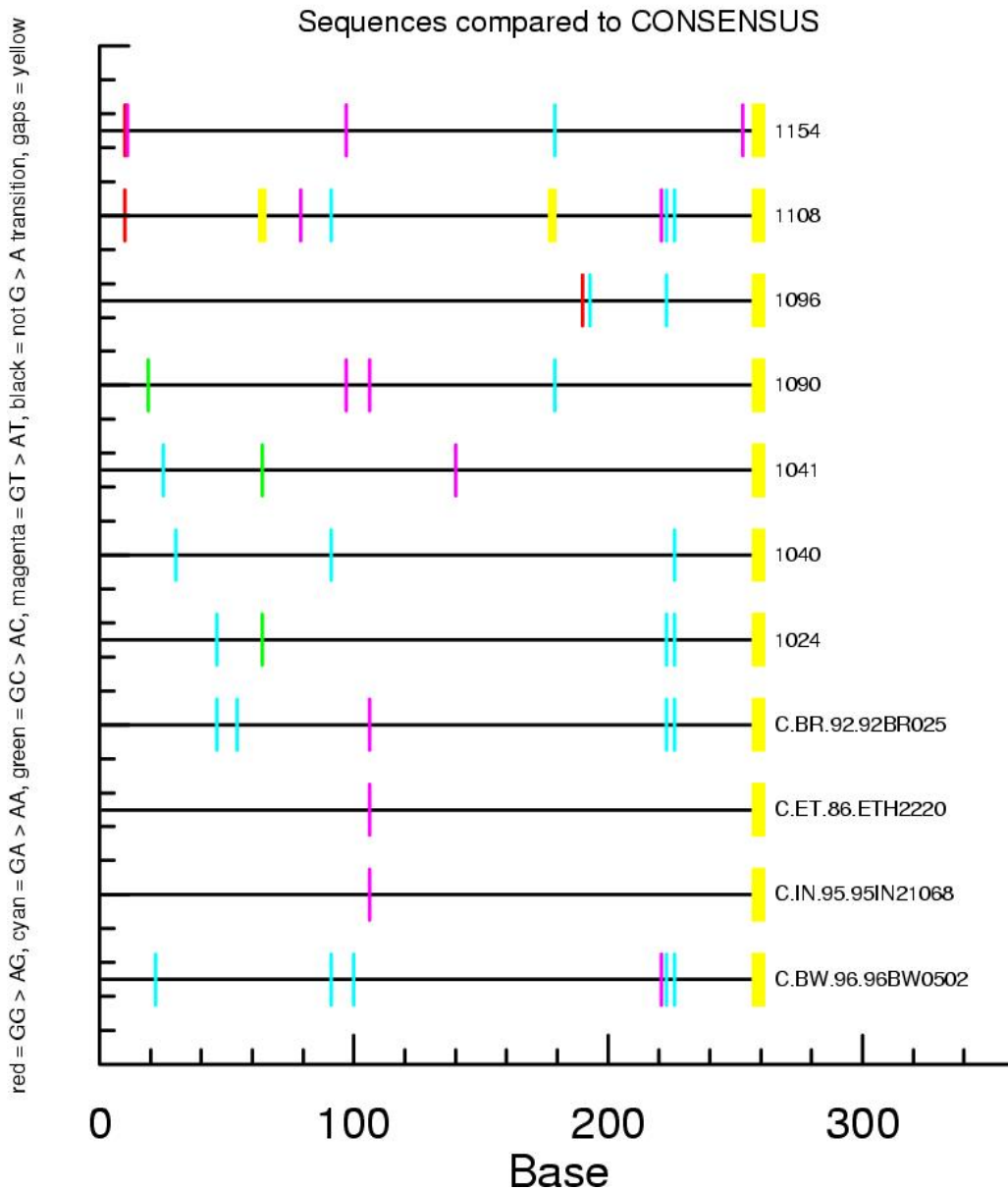




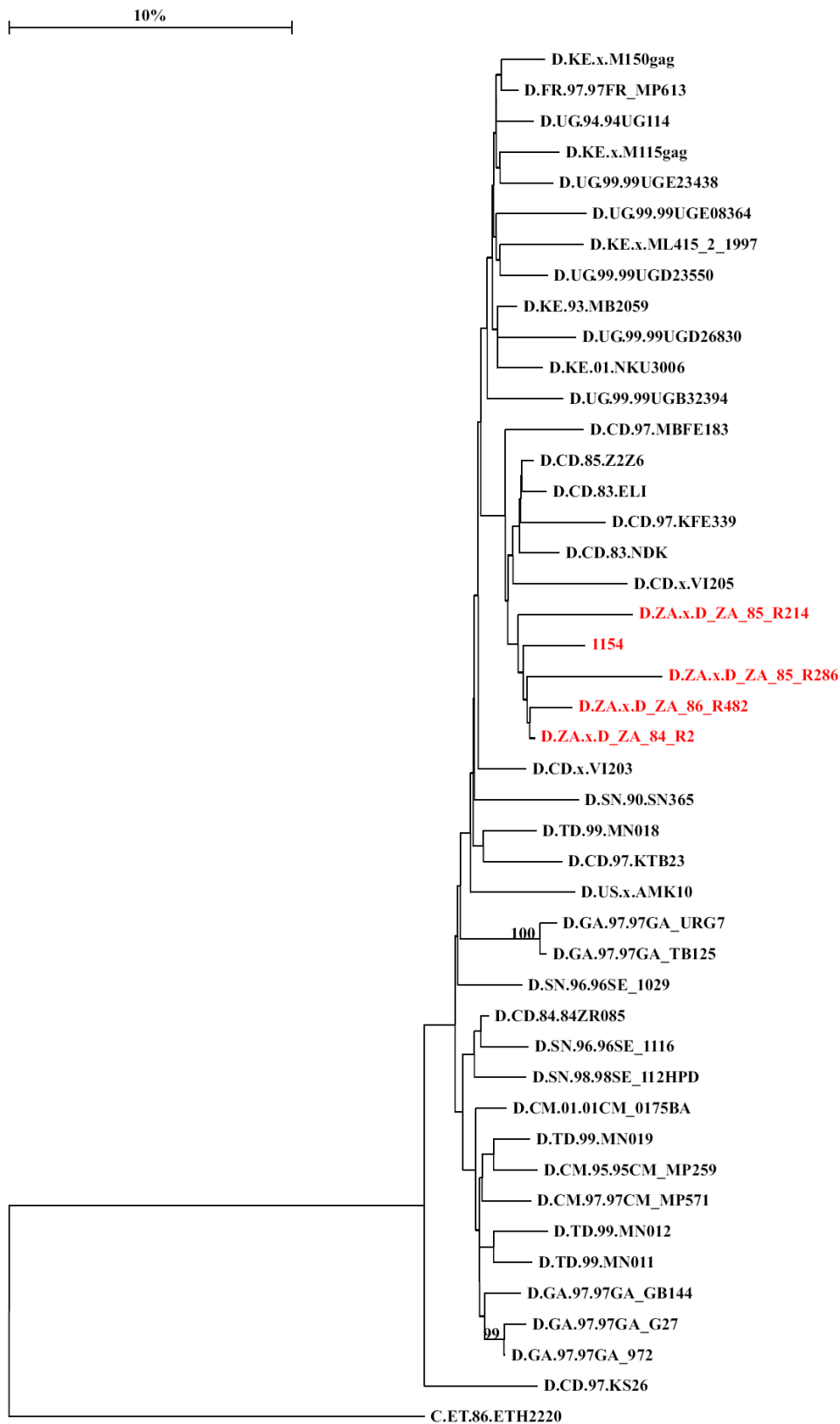
**Figure 3.9: The *env* gp41 neighbour-joining phylogenetic tree.** The Khayelitsha *env* gp41 sequences are indicated in bold red. The sequence alignment was 441 bp in length. Bootstrap values are indicated, with the branch lengths drawn to scale. All the *env* gp41 Khayelitsha sequences cluster with the HIV-1 subtype C reference sequences.



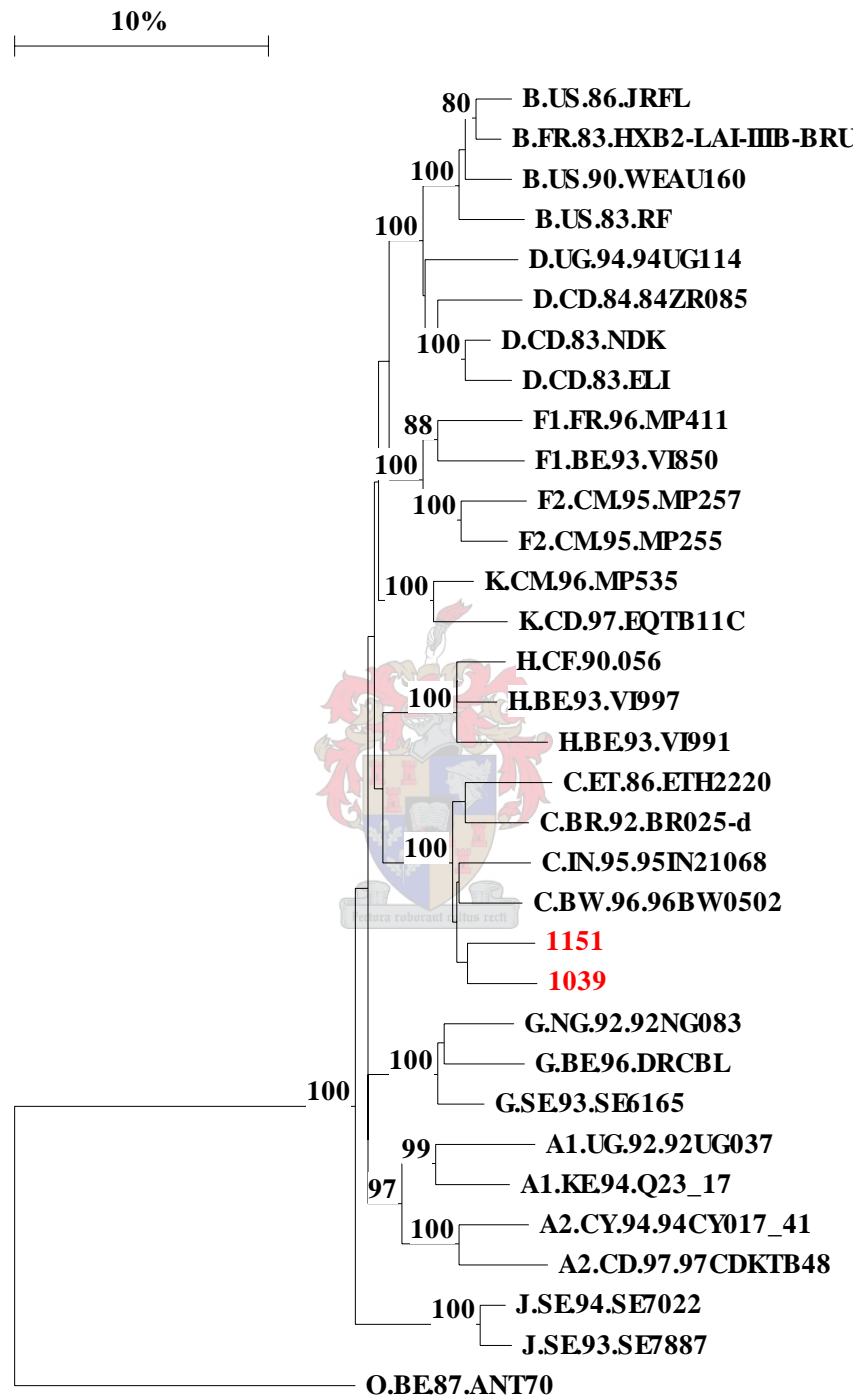
**Figure 3.10: The *env* gp120 V3 neighbour-joining phylogenetic tree.** The *env* gp120 V3 sequence alignment was 261 bp in length. The Khayelitsha sequences are indicated in bold red. Bootstrap values are also indicated, with branch lengths drawn to scale. All the *env* gp120 V3 sequences group with the HIV-1 subtype C reference sequences. The long branch lengths are indicative of the variability of the V3 region.



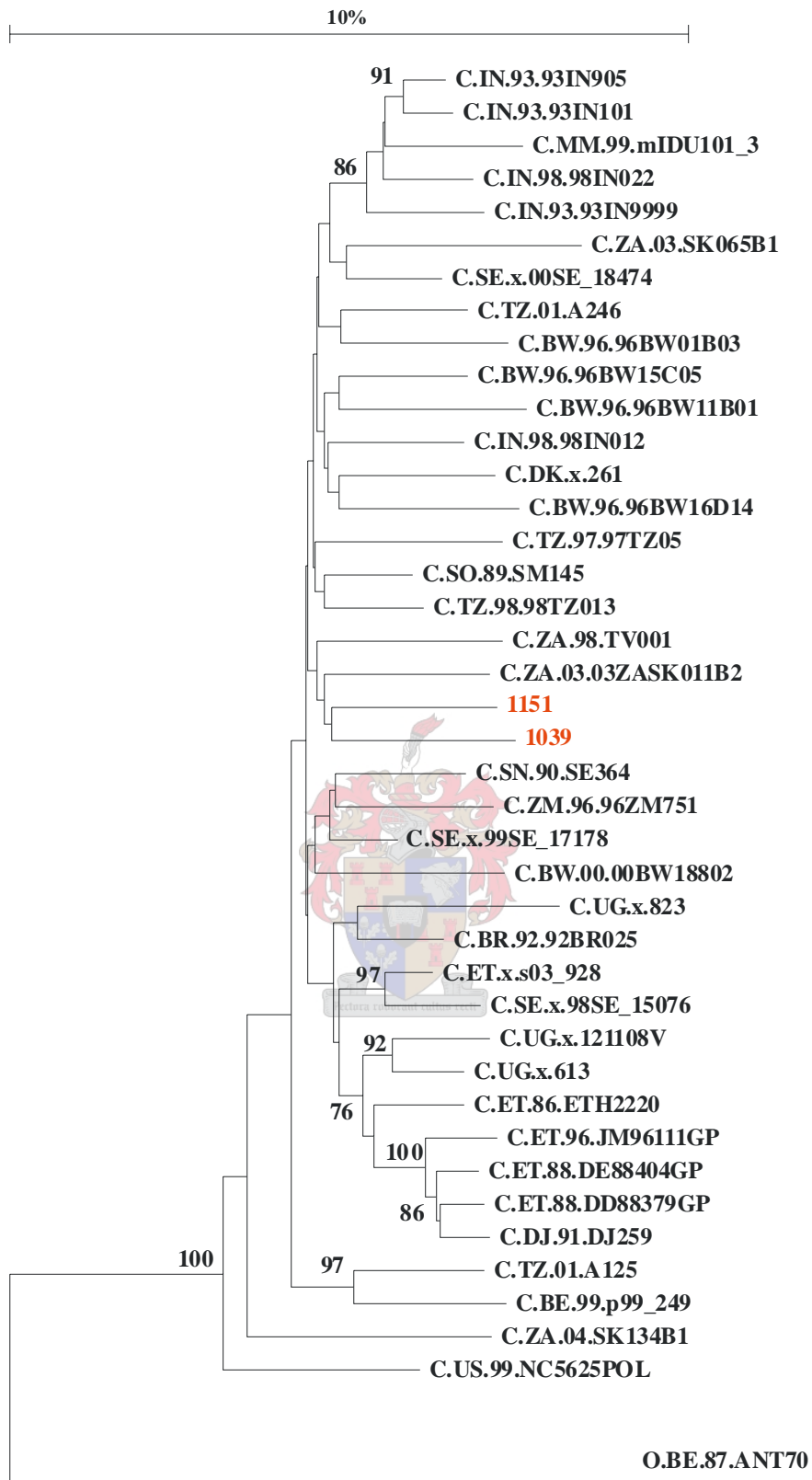
**Figure 3.11: Analysis of possible hypermutant HIV-1 *env* gp120 V3 sequences.** The x-axis represents the sequence length, while the hash marks indicate differences in the sequences compared to the subtype C *env* gp120 V3 consensus sequences. Hypermutant variants are often responsible for long branch lengths on phylogenetic trees, as observed in Figure 3.10. They are caused by the error-prone RT enzyme and are induced by host defence mechanisms to produce non-viable sequence.



**Figure 3.12: A subtype D gag p24 neighbour-joining phylogenetic tree.** The South African subtype D sequences, including the sequence from sample 1154 are indicated in bold red. The sequence from sample 1154 clearly forms a phylogenetic group with the rest of the South African sequences. The gag p24 sequence alignment was 444 bp in length. Bootstrap values are indicated and branch lengths drawn to scale.



**Figure 3.13: A *pol* neighbour-joining phylogenetic tree with samples 1039 and 1151 drawn with the reference sequences.** Bootstrap values are indicated and branch lengths drawn to scale. The Khayelitsha samples are indicated in bold red. The *pol* sequence alignment was 1050 bp in length. The *pol* sequences group with the HIV-1 subtype C reference sequences.



**Figure 3.14: A *pol* neighbour-joining phylogenetic tree with the sequences from samples 1039 and 1151.** Randomly selected subtype C *pol* sequences were included in the phylogenetic tree. Bootstrap values are indicated with branch lengths drawn to scale. The Khayelitsha samples are indicated in bold red. The *pol* sequence alignment was 1050 bp in length. The sequences from 1039 and 1151 form a phylogenetic cluster with the South African sequences TV001 and 03ZASK011B2.

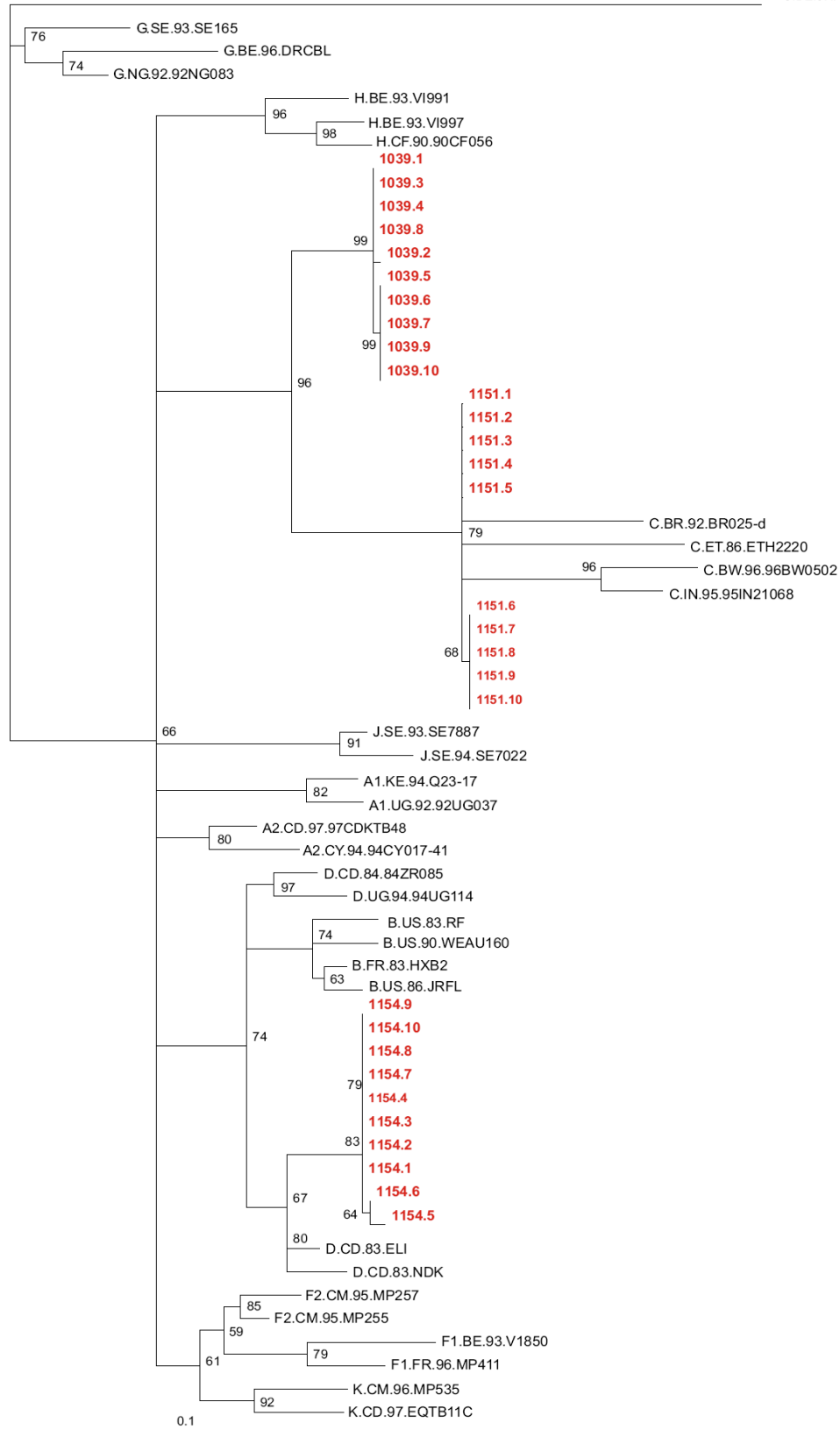
**Table 3.8:** Sample 1154 gag p24 subtype D sequence similarity

<b>Name of sequence</b>	<b>Accession number</b>	<b>Origin of sample</b>	<b>Percentage similarity</b>
MP259	AJ286372	Cameroon	92.7
MP571	AJ286394	Cameroon	93.4
0175BA	AY371156	Cameroon	94.4
MN019	AJ491006	Chad	93.0
MN012	AJ488927	Chad	93.6
MN011	AJ488926	Chad	93.8
MN018	AJ491005	Chad	94.0
KS26	AJ404267	DRC	91.5
VI205	L11785	DRC	93.6
KTB23	AJ404278	DRC	93.8
84ZR085	84ZR085	DRC	94.6
KFE339	AJ404250	DRC	94.6
VI203	L11784	DRC	94.8
MBFE183	AJ404285	DRC	95.2
NDK	M27323	DRC	95.8
ELI	K03454	DRC	96.2
Z2Z6_Z2	M22639	DRC	97.1
MP613	AJ286535	France	95.6
97GA_URG7	AJ286499	Gabon	92.5
97GA_TB125	AJ286479	Gabon	93.0
97GA_G27	AJ286443	Gabon	93.4
97GA_972	AJ286426	Gabon	93.8
97GA_GB144	AJ286453	Gabon	94.0
ML415_2_1997	AY322189	Kenya	93.0
NKU3006	AF457090	Kenya	94.0
M115gag	AY772952	Kenya	94.2
MB2059	AF133821	Kenya	94.6
M150gag	AY772956	Kenya	94.8
SN365	L11797	Senegal	92.7
96SE_1116	AJ274541	Senegal	93.2
112HPD	AJ274561	Senegal	93.6
96SE_1029	AJ274543	Senegal	93.6
D_ZA_R286	AY773340	South Africa	93.4
D_ZA_R214	AY773339	South Africa	94.0
D_ZA_R482	AY773341	South Africa	96.2
D_ZA_R2	AY773338	South Africa	97.5
99UGE08364	AF484487	Uganda	93.2
99UGD26830	AF484486	Uganda	93.4
99UGD23550	AF484485	Uganda	93.8
99UGB32394	AF484483	Uganda	94.4
94UG114	U88824	Uganda	94.6
99UGE23438	AF484489	Uganda	94.6
AMK10	U08192	USA	92.5

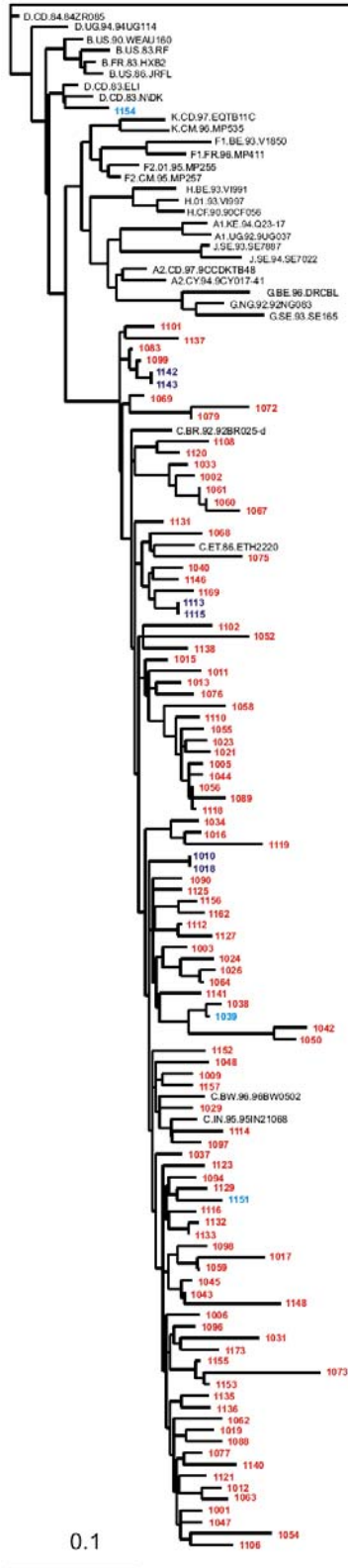
**Table 3.9:** Sample 1039 and 1151 *pol* subtype C sequence similarity

Name of sequence	Accession number	Origin of sample	Percentage similarity	
			1039	1151
p99_249	AF338992	Belgium	92.8	93.5
96BW11B01	AF110971	Botswana	93.6	94.0
00BW18802	AF443100	Botswana	94.0	94.2
96BW16D14	AF110977	Botswana	94.2	94.6
96BW01B03	AF110959	Botswana	94.4	94.8
96BW15C05	AF110975	Botswana	95.2	95.3
92BR025	U52953	Brazil	94.7	95.6
261	AJ419476	Denmark	94.3	94.9
DJ259	AF447839	Djibouti	94.0	94.2
ETH2220	U46016	Ethiopia	93.9	94.7
DE88404GP	AY242594	Ethiopia	94.0	94.5
JM96111GP	AY242588	Ethiopia	94.0	94.4
s03_928	AY371693	Ethiopia	94.0	94.2
DD88379GP	AY242591	Ethiopia	94.1	94.7
98IN022	AF286232	India	94.7	94.9
93IN9999	AF067154	India	94.8	94.5
93IN101	AB023804	India	94.9	95.3
98IN012	AF286231	India	94.9	95.2
93IN905	AF067158	India	95.2	95.3
mIDU101_3	AB097871	Myanmar	93.9	94.3
99SE_17178	AY165213	Senegal	92.8	93.5
SE_15076	AY165196	Senegal	93.5	93.7
SE364	AF447842	Senegal	93.5	93.7
00SE_18474	AY165224	Senegal	95.3	95.2
SM145	AF447850	Somalia	95.9	96.0
TV001	AY162223	South Africa	94.4	94.9
SK065B1	AY772694	South Africa	92.9	93.8
SK134B1	AY703909	South Africa	93.0	93.0
03ZASK011B2	AY901965	South Africa	95.1	95.2
A125	AY253304	Tanzania	93.9	94.0
97TZ05	AF361875	Tanzania	94.4	93.9
98TZ013	AF286234	Tanzania	95.3	96.0
A246	AY253308	Tanzania	95.3	95.0
823	AF388161	Uganda	93.4	93.4
121108V	AF410207	Uganda	93.8	94.5
613	AF388102	Uganda	94.2	95.2
NC5625POL	AY032091	USA	91.6	92.1
96ZM751	AF286225	Zimbabwe	94.0	94.1
1039		<i>Khayelitsha</i>	100	95.1
1151		<i>Khayelitsha</i>	95.1	100

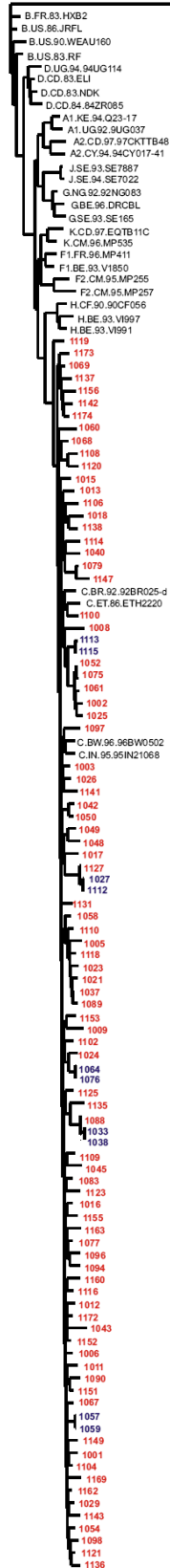




**Figure 3.15: A *gag* p24 maximum likelihood phylogenetic tree with the sequences from the 3 cloned samples (1039, 1151 and 1154).** Bootstrap values are indicated with the sequences from the cloned samples indicated in bold red. Ten different clones of each sample were sequenced and analysed. The *gag* p24 sequence alignment was 444 bp in length.

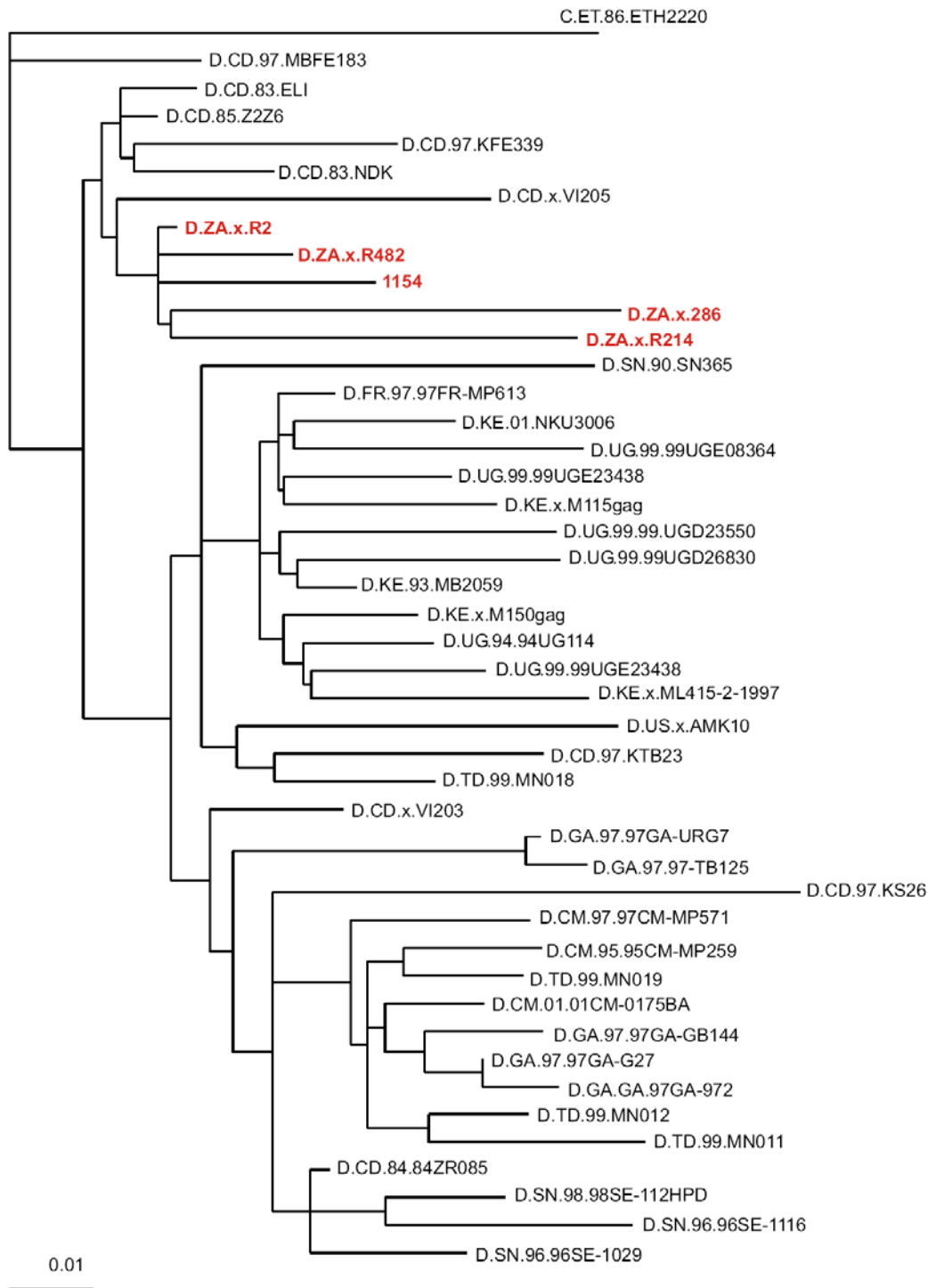


**Figure 3.16: The *gag* p24 maximum likelihood phylogenetic tree.** The sequences from the cloned samples (1039, 1151 and 1154) are indicated in blue, while sample pairs 1010 and 1018, 1113 and 1115, as well as 1142 and 1143 are indicated in purple. The *gag* p24 sequence alignment was 444 bp in length. Branch lengths are drawn to scale. All the Khayelitsha sequences, with the exception of 1154, group with the HIV-1 subtype C reference sequences.

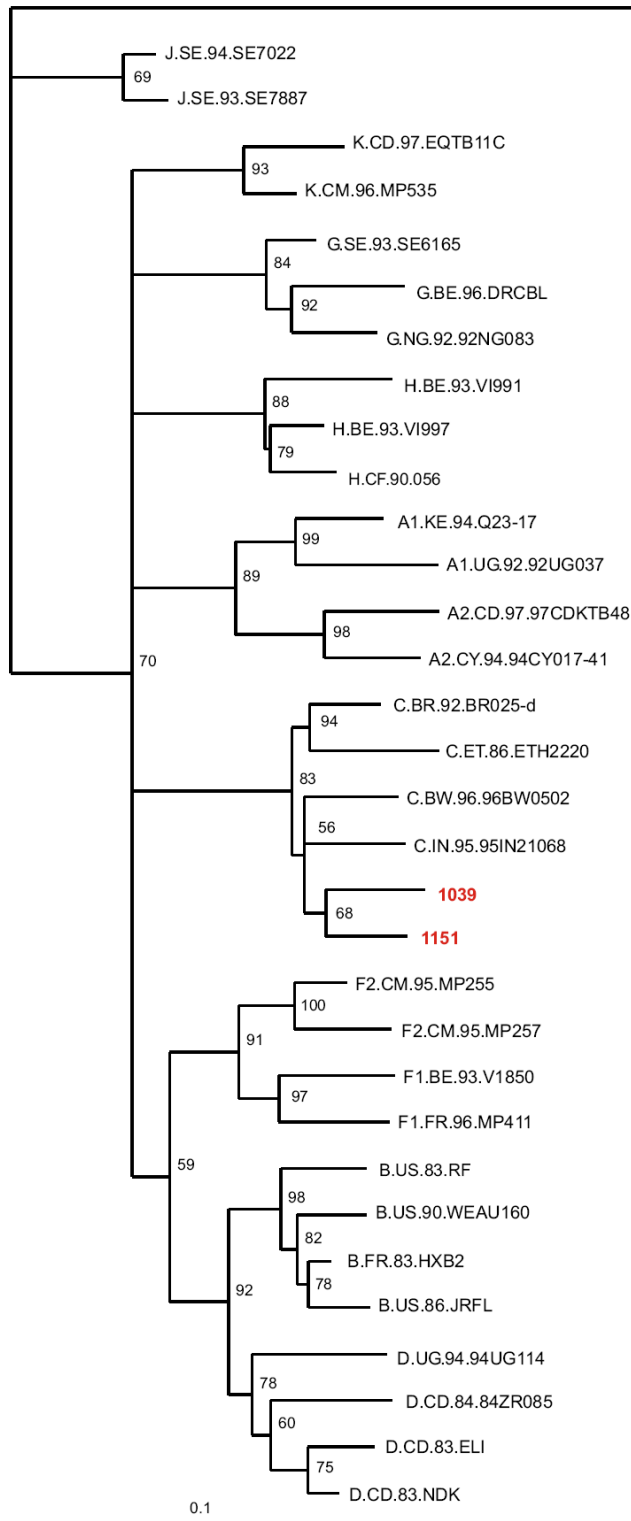


0.1

**Figure 3.17: The *env* gp41 IDR maximum likelihood phylogenetic tree.** Closely related sequences are indicated in purple, while the rest of the Khayelitsha sequences are indicated in red. The *env* gp41 IDR sequence alignment was 441 bp in length. Branch lengths are drawn to scale. All the Khayelitsha sequences group with the HIV-1 subtype C reference sequences.



**Figure 3.18: A subtype D gag p24 maximum likelihood phylogenetic tree.** The South African subtype D sequences, including the sequence from sample 1154 are indicated in bold red. Branch lengths are drawn to scale. The gag p24 sequence alignment was 444 bp in length. The South African subtype D gag p24 sequences form a unique phylogenetic group.

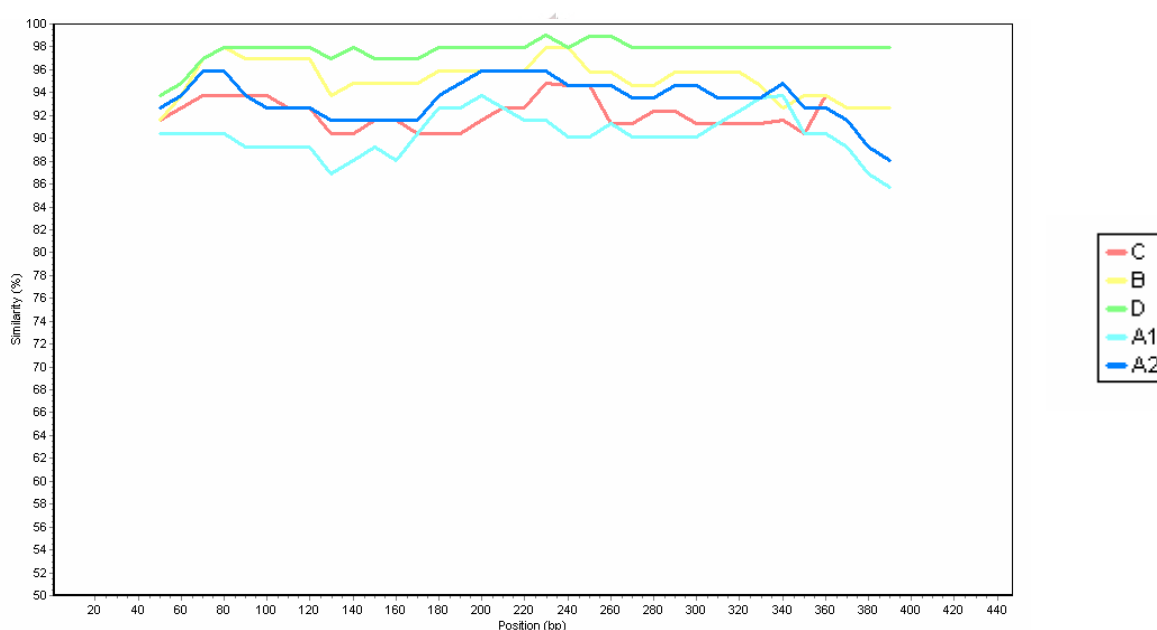


**Figure 3.19: A *pol* maximum likelihood phylogenetic tree with the sequences from samples 1039 and 1151.** The tree was drawn with the reference sequences obtained from the LANL database. Bootstrap values are indicated with branch lengths drawn to scale. The *pol* sequence alignment was 1050 bp in length. The South African sequences, indicated in bold red, form a phylogenetic cluster with the subtype C reference sequences.

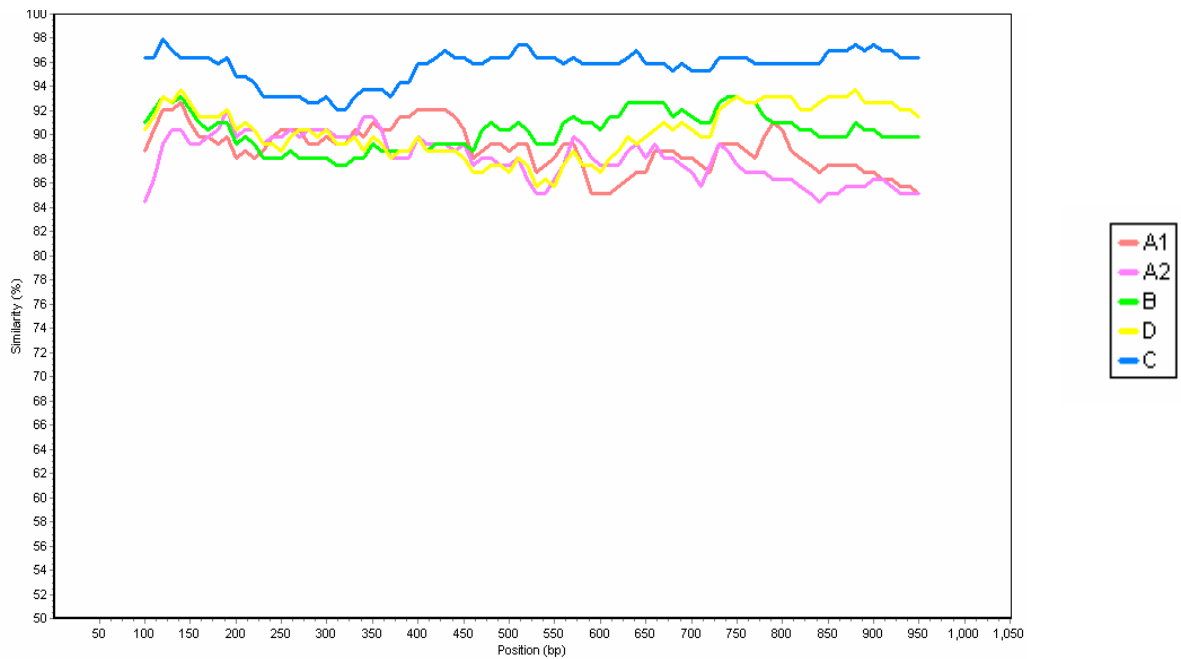


### 3.7.3 Similarity plots

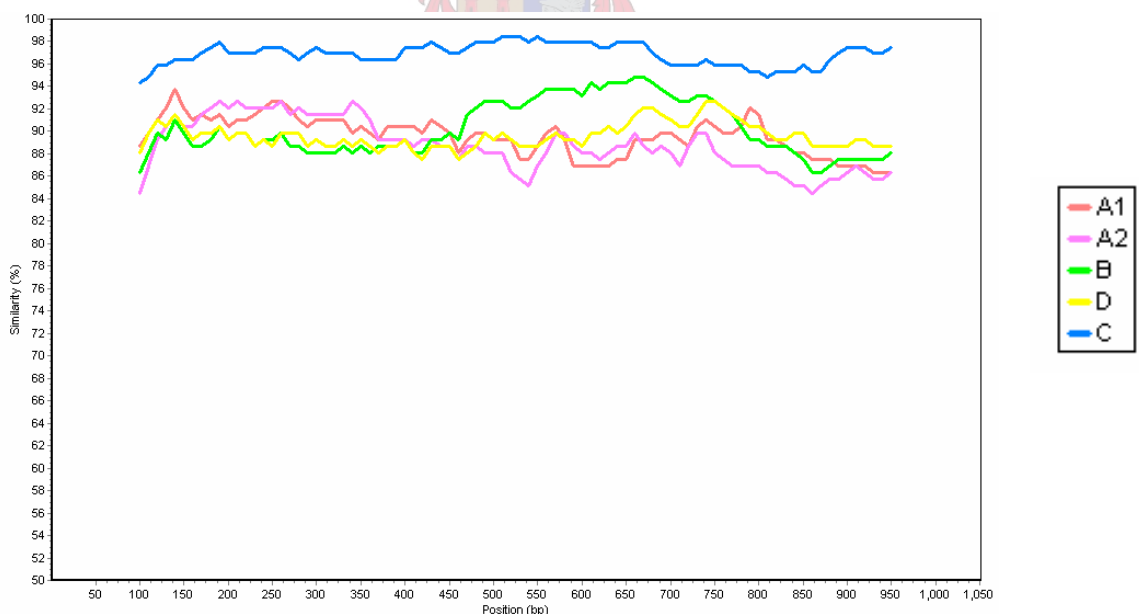
Similarity plots were used to help genotype possible HIV-1 recombinant strains that proved difficult to characterise. The Khayelitsha sequences were screened against reference sequences from HIV-1 subtypes A1, A2, B, C and D. Not all the reference sequences used in phylogenetic analysis were used in Simplot, as the program can only incorporate a limited number of sequences into its software analysis package. The query sequences were only screened against the major circulating subtypes, the most commonly found in South Africa. The similarity plots clearly indicate that the *gag* p24 sequence from sample 1154 has the closest similarity to the subtype D reference sequences (Figure 3.21). The *pol* sequences from samples 1039 (Figure 3.22) and 1151 (Figure 3.23) are the most comparable to the subtype C reference sequences.



**Figure 3.21: The *gag* p24 similarity plot of the sequence from sample 1154.** The top part of the graph shows the colour code for the reference sequence with the highest similarity score. The group M subtype colour codes (subtypes A1, A2, B, C and D) are indicated on the right hand side. The sequence positions are indicated on the x-axis, while the percentage similarity is indicated on the y-axis. A window size of 100 bp with 10 increment steps was used to construct the similarity plot. The *gag* p24 sequence has the highest similarity score with the subtype D reference sequences.



**Figure 3.22: The 1.2 kb *pol* similarity plot of the sequence from sample 1039.** The similarity plot was drawn with a window size of 200 bp and 10 increments per step. The sequence from sample 1039 shows the highest similarity score to the subtype C reference sequences, as indicated in blue at the top of the graph. The group M subtype colour codes (subtypes A1, A2, B, C and D) are indicated on the right hand side. The sequence positions are indicated on the x-axis, while the percentage similarity is indicated on the y-axis.

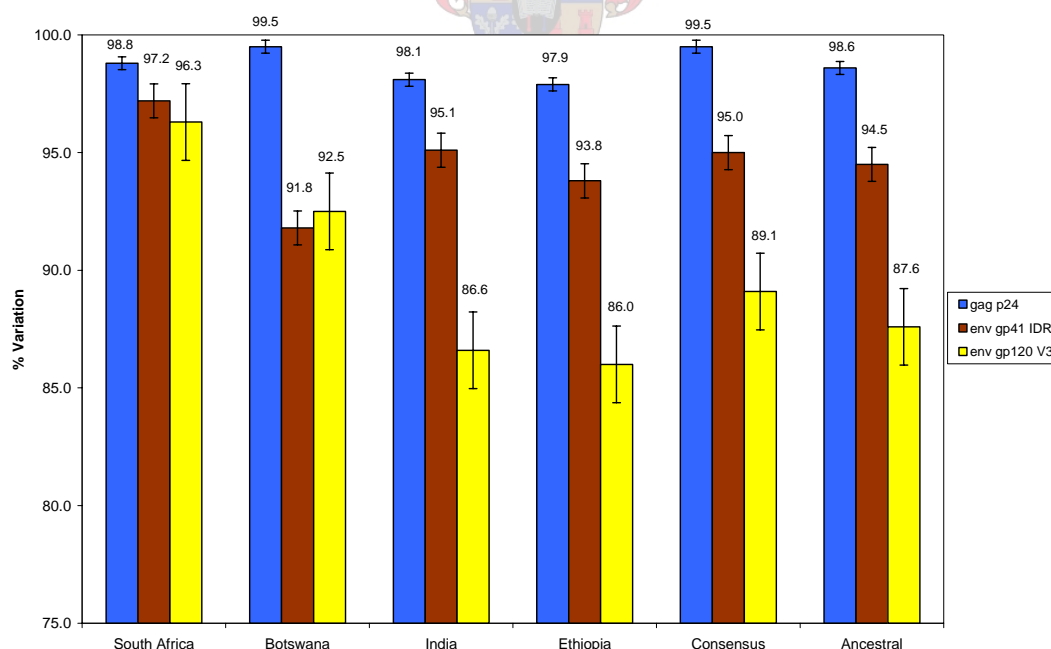


**Figure 3.23: The 1.2 kb *pol* similarity plot of the sequence from sample 1151.** The similarity plot was drawn with a window size of 200 bp and 10 increments per step. The sequence from sample 1151 shows the highest similarity score to the subtype C reference sequences, as indicated in blue at the top of the graph. The group M subtype colour codes (subtypes A1, A2, B, C and D) are indicated on the right hand side. The sequence positions are indicated on the x-axis, while the percentage similarity is indicated on the y-axis.



### 3.7.5 Consensus sequences

Khayelitsha consensus sequences were created from the nucleotide sequence alignments from the *gag* p24, *env* gp41 IDR and *env* gp120 V3 genome regions. The consensus sequences was compared against consensus sequences from Botswana, Ethiopia, India, South Africa, as well as an overall consensus subtype C sequence obtained from the LANL. From each alignment a sequence identity matrix was constructed. The matrix shows the percentage identity each consensus sequence has relative to one another. The *gag* p24 percentage similarity varies between 97.2% and 99.5%. In the *env* gp41 IDR the variation is between 90.9% and 99.5% and in the *env* gp120 V3 the variation is from 83.2% to 96.3%. Variation between the *gag* p24 sequences is very small, with only a 2.3% difference between the most similar and least similar consensus sequences (data not shown). The matrix scores were compared to the Khayelitsha scores and summarised in Figure 3.24. The Khayelitsha sequences show the closest similarity to the South African sequences in both the *env* gp41 IDR (97.2%) and *env* gp120 V3 (96.3%). In the *gag* p24 the Khayelitsha sequences are the most comparable to the Botswana and consensus sequences (99.5%).



**Figure 3.24: Percentage variation between the Khayelitsha consensus sequence and other consensus sequences.** The *gag* p24 sequences show the least variation, followed by the *env* gp41 IDR. The *env* gp120 V3 region is the most variable. In the *env* gp41 IDR and *env* gp120 V3 the Khayelitsha sequences are the most similar to the South African sequences, 97.2% and 96.3% respectively. With the *gag* p24 the highest similarity is seen with the Botswana and Consensus sequence, 99.5%

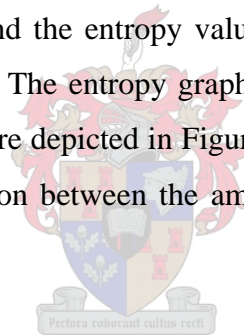
In Figure 3.25 the 35 amino acid V3 loop is compared to the subtype C consensus and ancestral V3 loop sequences. In four regions in the Khayelitsha sequences no consensus could be reached due to the variability amongst the Khayelitsha sequences.

	*	20	*					
Consensus	:	CTRPNNTRKSI	RIGPGQTFYAT	GDIIGDIRQAHC	:	35		
Ancestral	:	CTRPNNTRKSI	RIGPGQTFYAT	GDIIGDIRQAHC	:	35		
Khayelitsha	:	CTRPNN	-TRKS-	RIGPGQTFYAT	--	IIGDIRQAHC	:	31
		CTRPNNnTRKSiRIGPGQTFYATgDIIGDIRQAHC						

**Figure 3.25: The Khayelitsha env gp120 V3 consensus sequence compared to the subtype C consensus and ancestral sequences.** The gaps in the Khayelitsha sequence represent areas where no consensus could be reached due to the *env* gp120 variability of the V3 Khayelitsha sequences.

### 3.7.6 Entropy values and conserved genomic regions

The *gag* p24, *env* gp41 IDR and *env* gp120 V3 Khayelitsha amino acid sequences were aligned with each other and the entropy value (degree of variability) for each amino acid position calculated. The entropy graphs, showing the variation over the three genome regions targeted, are depicted in Figures 3.26, 3.27 and 3.28. Where an entropy value is zero no variation between the amino acids in a specific alignment was observed.

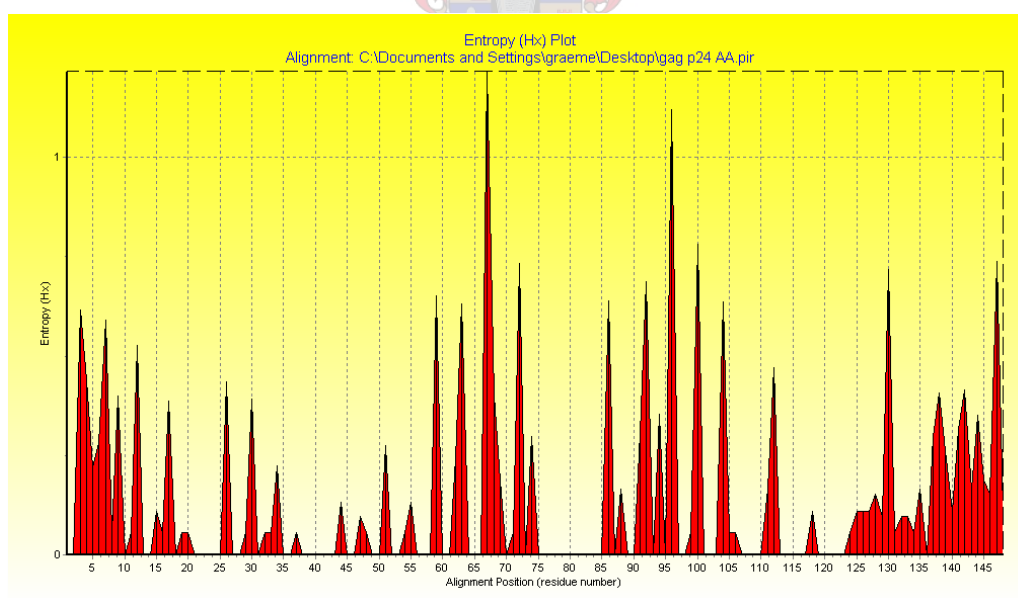


The *gag* p24 entropy value ranges from 0 to 1.215 with a mean value of 0.143 ( $\pm$  0.2). The *env* gp41 IDR entropy value ranges from 0 to 1.941 with a mean value of 0.244 ( $\pm$  0.4), while the *env* gp120 V3 entropy value ranges from 0 to 2.322 with a mean value of 0.548 ( $\pm$  0.6). The most variable regions of the *gag* p24 are at sequence positions 67 and 96. In the *env* gp41 IDR they are at positions 74 and 75 in the amino acid sequence and in the *env* gp120 V3 they are at positions 77 and 78.

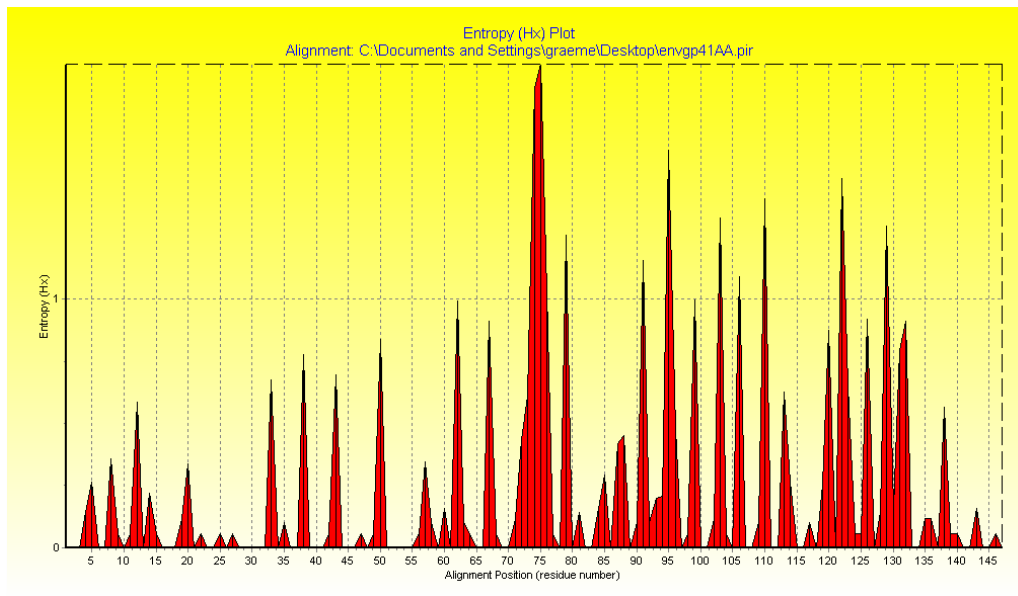
The *gag* gene recorded the lowest entropy values. The gene is highly conserved, encoding for structurally important protein peptides. Different epitopes of the *gag* gene are also important for CD4+ T-cell responses in HIV-1 infected individuals, as discussed in chapter four. In the *env* gp41 IDR the most variable amino acid epitopes are SNKSLEQ at position 71 to 75 of the *env* gp41 IDR protein sequence and LDKWAS at position 118 to 123. The SNKSLEQ motif has been recognised as an important CD4+ T cell epitope, while the LDKWAS motif is an important antibody

recognition site for mAbs. In the *env* gp120 V3 (Figure 3.28) amino acid residues 52 to 55 represent the highly conserved crown regions of the V3 loop, usually containing the GPGR or GPGQ motif. It is notable that the HIV-1 subtype C sequences from Khayelitsha seem to be more variable towards the end of the V3 loop.

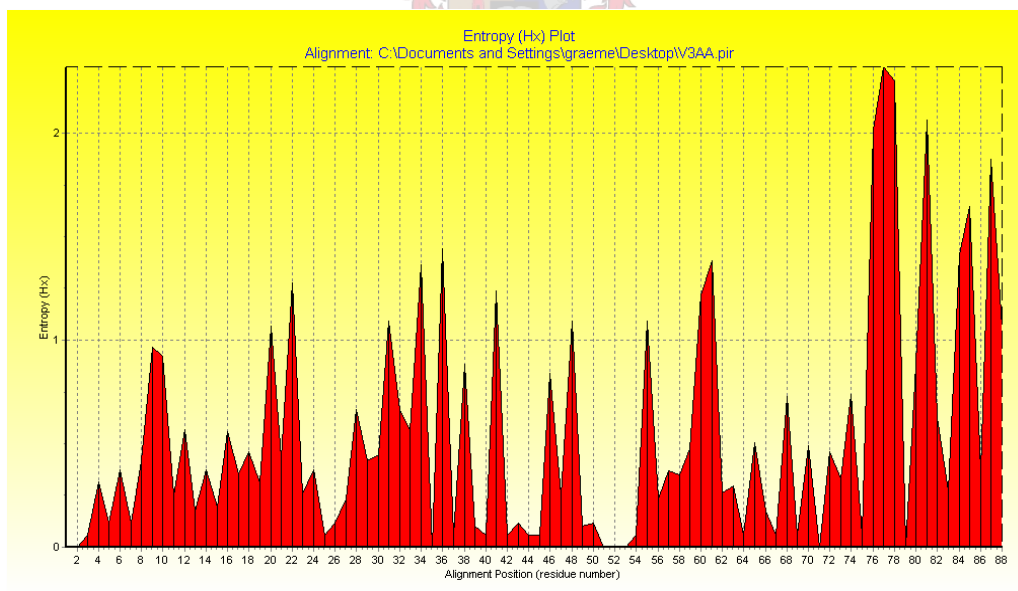
A minimum segment length of four amino acid sequences was used in BioEdit to search for conserved genomic regions amongst the Khayelitsha sequences. A length of four amino acids was used to search for all possible conserved epitopes. Amino acid residues of less than four will pick up redundant regions that might occur frequently, but not actually be conserved. Nine conserved regions were found in the *gag* p24, 14 in the *env* gp41 IDR and only two in the *env* gp120 V3. However, the longest conserved peptides (length of amino acid residues) were found in the *gag* p24. An Epilign of these conserved amino acid epitopes indicates how often they occur in HIV-1 subtype C. The data is presented in Table 3.10. It is important to keep track of the different epitope variants, as a single amino acid mutation might lead to the change in protein function. The implications of HIV-1 variability and conserved genomic regions are discussed in more detail in chapter four.



**Figure 3.26: A *gag* p24 amino acid entropy graph.** The graph was created with BioEdit version 5.09 using a protein alignment from the Khayelitsha sequences. The y-axis indicates the entropy log scores and the x-axis the amino acid position of the sequences in the alignment. Where the entropy value is zero no variation between the amino acids in the specific alignment was observed. The higher the entropy values, red bars, the more variable the amino acid sequences at that particular position. In the *gag* p24 the most variable sites are at position 67 and 96 in the sequence.



**Figure 3.27: An *env* gp41 IDR amino acid entropy graph.** The graph was created with BioEdit version 5.09 with the *env* gp41 IDR protein alignment. The entropy values are indicated by the red bars. The most variable sites are at position 74 and 75 in the sequence. The variable region containing the NKSLE motif (position 71 to 75) is an important CD4+ T cell epitope recognising region, while the LDKWAS motif (position 118 to 123) is targeted by various mAbs.



**Figure 3.28: An *env* gp120 V3 amino acid entropy graph.** The graph was created with BioEdit version 5.09 with the *env* gp120 V3 protein alignment. The entropy values are indicated at the red bars. In the *env* gp120 V3 the most variable sites are at position 77 and 78 in the sequence. The V3 regions entropy scores are much higher than the *gag* p24 and the *env* gp41 IDR entropy scores. The conserved crown region of the V3 loop is located at position 52 to 55 of the V3 amino acid sequence. The Khayelitsha sequences are highly variable towards the end of the *env* gp120 V3 sequence.

**Table 3.10:** Conserved amino acid regions

<b><i>gag p24</i></b> (*n = 9)				
<b>Position</b>	<b>#HXB2</b>	<b>Sequence</b>	<b>Length</b>	<b>†Epilign (%)</b>
13 - 16	169 - 172	IPMF	4	98.3
18 - 25	174 - 181	ALSEGATP	8	94.1
35 - 50	191 - 206	VGGHQAAMQMLKDTIN	16	95.8
52 - 58	208 - 214	EAAEWDR	7	92.4
75 - 85	231 - 241	PRGSDIAGTTS	11	93.2
87 - 90	243 - 246	LQEQ	4	97.5
105 - 111	261 - 267	IYKRWII	7	94.1
113 - 129	269 - 285	GLNKIVRMYSVILDI	17	93.2
131 - 136	287 - 292	QGPKEP	6	96.6
<b><i>env gp41 IDR</i></b> (n = 14)				
<b>Position</b>	<b>HXB2</b>	<b>Sequence</b>	<b>Length</b>	<b>Epilign (%)</b>
1 - 4	546 - 549	SGIV	4	95.9
15 - 19	560 - 564	EAQQH	5	94.3
21 - 32	566 - 577	LQLTVWGIKQLQ	12	93.4
34 - 37	579 - 582	RVLA	4	95.1
39 - 42	584 - 587	ERYL	4	95.9
44 - 49	589 - 594	DQQLLG	6	95.1
51 - 56	596 - 601	WGCSGK	6	95.1
58 - 61	603 - 606	ICTT	4	93.4
63 - 66	608 - 611	VPWN	4	94.3
68 - 71	613 - 616	SWSN	4	91.8
80 - 84	625 - 629	NMTWM	5	82.8
115 - 118	660 - 663	LLAL	4	93.4
133 - 137	678 - 682	WLWYI	5	95.1
139 - 147	684 - 692	IFIMIVGGL	9	74.6
<b><i>env gp120 V3</i></b> (n = 2)				
<b>Position</b>	<b>HXB2</b>	<b>Sequence</b>	<b>Length</b>	<b>Epilign (%)</b>
42 - 45	301 - 304	NNTR	4	94.3
49 - 54	310 - 315	RIGPGQ	6	93.4

\*Number of conserved regions in each gene fragment

# Amino acid position relative to protein start in HXB2

†The Epilign indicates the percentage a specific epitope sequence occurs in other HIV-1 subtype C sequences

### 3.7.7 Co-receptor prediction

A PSSM scoring matrix was used to predict the HIV-1 co-receptor usage of the Khayelitsha cohort samples. Most co-receptor prediction matrices are optimised for HIV-1 subtype B and results were interpreted carefully. From the 96 sequences used in sequence analysis 88 (91.7%) viruses were predicted to use CCR5 as co-receptor, while 8 viruses (8.3%) (1047, 1076, 1096, 1101, 1110, 1127, 1129 and 1154) were predicted to use CXCR4. Serine (S) is the most common amino acid at position 11 and is found in 92 of the V3 sequences (95.8%). Amino acids arginine (R), asparagine (N) and glycine (G) are also present. Aspartic acid (D) is the most common amino acid at position 25 and is present in 49 (51.0%) of the V3 sequences. Amino acids glutamic acid (E), isoleucine (I), lysine (K), A, G and N also occur at position 25. In sample 1019 an amino acid deletion at position 25 is present. Basic amino acid substitutions (K, R or H) at positions 11 and / or 25 are usually associated with a shift in chemokine co-receptor usage from CCR5 to CXCR4. However, basic amino acid substitutions are only present in two of the CXCR4 using viruses. These are K at position 25 of the sequence from sample 1047 and R at position 11 of the sequence from sample 1096. The true viral phenotype of a virus can thus only be predicted with functional assays and caution should be taken when using sequence prediction methods. The Khayelitsha *env* gp120 V3 sequences have a mean positive charge of 5.9 ( $\pm$  0.7) and a mean net charge of 4.2 ( $\pm$  1.0). The CCR5 strains alone have a mean charge of 4.1 ( $\pm$  0.9), while the CXCR4 strains have a mean net V3 charge of 5.4 ( $\pm$  1.2). CXCR4 strains are usually associated with higher net V3 charges and a SI phenotype, with a more advanced stage of disease progression. The reference subtype C strains were also tested against the criteria, as well as the HXB2 reference strain. The subtype C strains all use CCR5 as a co-receptor. HXB2 uses CXCR4 as a co-receptor, has a high net V3 charge and basic amino acids, (R) and (K), at positions 11 and 25 of its V3 region. The data is presented in Table 3.11.

**Table 3.11:** *env* gp120 V3 co-receptor prediction

Sample	Amino acid position		Positive charge	Net charge	Co-receptor predictions
	11	25			
1001	*Not available				
1002	S	D	5	3	CCR5
1003	S	D	6	4	CCR5
1005	S	D	6	4	CCR5
1006	S	D	5	3	CCR5
1008	S	E	6	5	CCR5
1009	S	E	6	5	CCR5
1010	S	D	6	4	CCR5
1011	S	D	6	4	CCR5
1012	S	D	6	4	CCR5
1013	S	D	6	4	CCR5
1015	Not available				
1016	Not available				
1017	S	G	6	4	CCR5
1018	S	-	5	5	CCR5
1019	Not available				
1021	S	E	5	3	CCR5
1023	S	D	6	4	CCR5
1024	S	D	5	3	CCR5
1025	S	G	6	5	CCR5
1026	S	G	6	5	CCR5
1027	S	G	5	5	CCR5
1029	Not available				
1031	Not available				
1033	S	G	6	6	CCR5
1034	S	E	5	4	CCR5
1037	S	D	6	4	CCR5
1038	S	D	5	2	CCR5
1039	Not available				
1040	S	D	5	3	CCR5
1041	N	D	5	3	CCR5
1042	S	E	6	4	CCR5
1043	S	A	6	4	CCR5
1044	Not available				
1045	S	E	6	4	CCR5
1047	G	K	8	7	CXCR4
1048	S	D	6	4	CCR5
1049	S	E	6	4	CCR5
1050	S	E	7	5	CCR5
1052	S	D	4	2	CCR5
1054	Not available				
1055	S	G	6	5	CCR5
1056	Not available				
1057	Not available				
1058	S	E	5	3	CCR5
1059	Not available				
1060	Not available				
1061	S	D	6	5	CCR5
1062	S	D	6	4	CCR5
1063	Not available				

\*N/A – sequences not available, PCR negative samples

**Table 3.11 continue:** *env* gp120 V3 co-receptor prediction

Sample	Amino acid position		Positive charge	Net charge	Co-receptor prediction
	11	25			
1064	S	E	7	5	CCR5
1067	*Not available				
1068	S	D	6	4	CCR5
1069	Not available				
1072	Not available				
1073	S	D	6	4	CCR5
1075	G	D	5	2	CCR5
1076	S	E	7	5	CXCR4
1077	S	G	6	5	CCR5
1078	S	D	5	3	CCR5
1079	S	D	6	4	CCR5
1083	S	D	6	4	CCR5
1084	S	D	5	2	CCR5
1088	Not available				
1089	S	D	6	5	CCR5
1090	S	D	6	3	CCR5
1094	S	E	6	4	CCR5
1096	R	D	8	7	CXCR4
1097	S	D	7	6	CCR5
1098	S	E	7	5	CCR5
1099	Not available				
1100	S	G	7	6	CCR5
1101	S	E	7	5	CXCR4
1102	S	D	6	4	CCR5
1104	S	A	6	4	CCR5
1106	Not available				
1108	S	D	5	2	CCR5
1109	Not available				
1110	S	I	7	6	CXCR4
1112	S	D	6	4	CCR5
1113	Not available				
1114	S	E	6	4	CCR5
1115	S	E	6	5	CCR5
1116	S	E	7	5	CCR5
1118	Not available				
1119	S	D	5	3	CCR5
1120	S	D	6	5	CCR5
1121	S	D	6	3	CCR5
1123	S	A	6	4	CCR5
1125	S	D	6	4	CCR5
1127	S	N	5	4	CXCR4
1129	S	S	6	4	CXCR4
1131	S	D	6	4	CCR5
1132	S	D	6	4	CCR5
1133	Not available				
1134	S	D	5	3	CCR5
1135	Not available				
1136	Not available				
1137	S	D	5	3	CCR5
1138	S	E	6	4	CCR5

\*N/A – sequences not available, PCR negative samples



**Table 3.11 continue:** *env* gp120 V3 co-receptor prediction

Sample	Amino acid position		Positive charge	Net charge	Co-receptor prediction
	11	25			
1139	*Not available				
1140	S	A	6	4	CCR5
1141	S	S	6	6	CCR5
1142	S	G	7	6	CCR5
1143	S	E	6	5	CCR5
1144	S	D	6	3	CCR5
1146	S	D	7	5	CCR5
1147	S	G	5	4	CCR5
1148	S	A	7	5	CCR5
1149	Not available				
1150	Not available				
1151	S	D	6	4	CCR5
1152	S	D	6	4	CCR5
1153	S	I	5	3	CCR5
1154	S	A	8	5	CXCR4
1155	S	D	5	4	CCR5
1156	S	A	6	5	CCR5
1157	S	D	6	4	CCR5
1160	S	D	6	4	CCR5
1162	S	D	7	5	CCR5
1163	S	D	6	4	CCR5
1165	S	A	6	5	CCR5
1169	Not available				
1172	Not available				
1173	S	D	6	4	CCR5
1174	S	D	6	4	CCR5
1175	S	D	5	4	CCR5
<b>Average</b>			5.9	4.2	
<b>HXB2</b>	R	K	10	10	CXCR4
<b>96BW0502</b>	S	E	5	3	CCR5
<b>95IN21068</b>	S	D	6	3	CCR5
<b>ETH2220</b>	S	D	5	2	CCR5
<b>92BR025</b>	S	E	6	4	CCR5
<b>Consensus_C</b>	S	D	6	4	CCR5
<b>Ancestral_C</b>	S	D	6	4	CCR5

\*N/A – sequences not available, PCR negative samples

### 3.7.8 Envelope N-Glycosylation sites

The *env* gp41 IDR N-glycosylation number ranges from 3 to 5, with a mean value of 4.1 ( $\pm$  0.4). The *env* gp120 V3 N-glycosylation number ranges from 2 to 7, with a mean value of 5.6 ( $\pm$  0.9). The data is presented in Table 3.12. The four subtype C reference strains, the consensus and ancestral subtype C sequences, as well as the HXB2 reference strain glycosylation numbers, are also listed.

**Table 3.12:** Envelope N-Glycosylation numbers

<b>Sample</b>	<b><i>env</i> gp41 IDR</b>	<b><i>env</i> gp120 V3</b>
1001	4	*N/A
1002	4	5
1003	4	6
1005	5	6
1006	4	6
1008	4	6
1009	4	6
1010	N/A	4
1011	4	6
1012	4	6
1013	N/A	6
1015	4	N/A
1016	4	N/A
1017	4	6
1018	4	6
1019	N/A	N/A
1021	4	6
1023	4	6
1024	4	4
1025	4	6
1026	4	4
1027	4	5
1029	5	N/A
1031	N/A	N/A
1033	N/A	6
1034	N/A	6
1037	4	5
1038	N/A	7
1039	N/A	N/A
1040	4	3
1041		6
1042	5	5
1043	3	5
1044	N/A	N/A
1045	4	6
1047	N/A	6
1048	4	6
1049	4	5
1050	4	6
1052	4	5
1054	4	N/A
1055	N/A	6
1056	N/A	N/A
1057	4	N/A
1058	5	6
1059	4	N/A
1060	4	N/A
1061	4	5
1062	N/A	7
1063	N/A	N/A

\*N/A – sequences not available, PCR negative samples

**Table 3.12 continue:** Envelope N-Glycosylation numbers

<b>Sample</b>	<b><i>env</i> gp41 IDR</b>	<b><i>env</i> gp120 V3</b>
1064	4	5
1067	5	*N/A
1068	4	6
1069	4	N/A
1072	N/A	N/A
1073	N/A	5
1075	4	6
1076	4	6
1077	4	7
1078	N/A	5
1079	4	6
1083	4	6
1084	N/A	4
1088	4	N/A
1089	4	6
1090	4	4
1094	5	5
1096	5	2
1097	3	7
1098	4	6
1099	N/A	N/A
1100	4	6
1101	N/A	5
1102	4	6
1104	4	6
1106	4	N/A
1108	4	6
1109	4	N/A
1110	5	4
1112	4	6
1113	5	N/A
1114	4	6
1115	N/A	6
1116	4	5
1118	4	N/A
1119	4	6
1120	4	6
1121	4	6
1123	4	7
1125	4	4
1127	N/A	5
1129	N/A	7
1131	4	6
1132	N/A	7
1133	N/A	N/A
1134	N/A	6
1135	4	N/A
1136	4	N/A
1137	4	6
1138	5	6

\*N/A – sequences not available, PCR negative samples

**Table 3.12 continue:** Envelope N-Glycosylation numbers

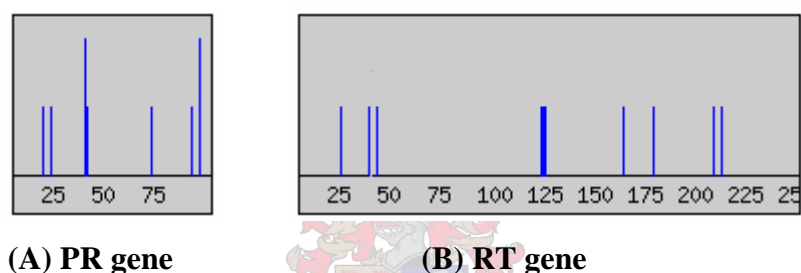
<b>Sample</b>	<b><i>env</i> gp41 IDR</b>	<b><i>env</i> gp120 V3</b>
1139	*N/A	N/A
1140	N/A	5
1141	3	5
1142	4	6
1143	4	6
1144	N/A	5
1146	N/A	5
1147	4	5
1148	N/A	5
1149	4	N/A
1150	N/A	N/A
1151	5	6
1152	5	5
1153	4	6
1154	N/A	3
1155	4	6
1156	4	6
1157	N/A	5
1160	5	6
1162	4	7
1163	4	6
1165	N/A	6
1169	4	N/A
1172	4	N/A
1173	5	5
1174	4	5
1175	N/A	5
<b>Average</b>	<b>4.1</b>	<b>5.6</b>
<b>HXB2</b>	6	8
<b>96BW0502</b>	5	6
<b>95IN21068</b>	5	7
<b>ETH2220</b>	6	7
<b>92BR025</b>	5	8
<b>Consensus_C</b>	5	6
<b>Ancestral_C</b>	5	6

\*N/A – sequences not available, PCR negative samples

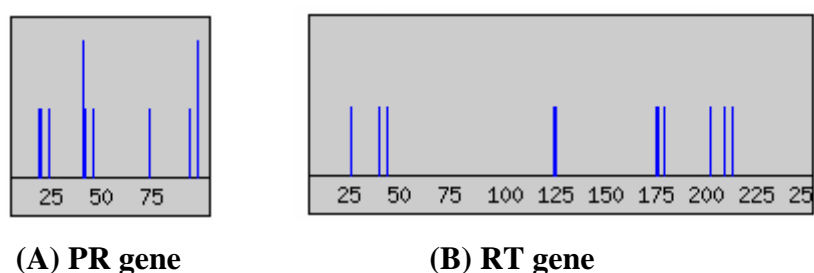
#### 3.7.4 HIV-1 drug resistant mutations

The *pol* sequences from samples 1039 and 1151 include positions 13 to 99 (relative to HXB2) of the PR gene and the complete RT gene. Both sequences predict that the patients will be susceptible to NRTIs and NNRTIs. These sequences were also predicted to have minor PI resistant mutations at positions 36 and 93 in the PR gene. These were M36I and I93L for the sequence from 1039, and M36L and I93L for the sequence from sample 1151. The letter in front of the gene position indicates the

amino acid residue present in the wild-type sequence, whereas the letter after the position indicator is the amino acid residue present in the query sequence. In other words, for sample 1039 the amino acid at position 36 in the PR gene is isoleucine (I), compared to methionine (M) in the wild-type. The sequences usually cause ART drug resistance in conjunction with other mutations. Differences between the sequences from samples 1039 and 1151 and the wild-type subtype B sequences are displayed in Figures 3.29 and 3.30. These differences are attributed to subtype difference between HIV-1 sequences. Both sequences from samples 1039 and 1151 belong to HIV-1 subtype C, while the wild-type sequence on the resistance database is a consensus HIV-1 subtype B sequence.



**Figure 3.29: Genotype resistance interpretation results for the sequence of sample 1039.** The sequences were screened for resistance in the PR (A) and RT (B) genes. The blue lines indicate differences from the consensus B sequences. The tall blue lines are areas associated with drug resistant mutations. Two PI minor resistant mutations were identified in the *pol* sequences of sample 1039. These were M36I and I93L.



**Figure 3.30: Genotype resistance interpretation results for the sequence of sample 1151.** The sequences were screened for resistance in the PR (A) and RT (B) genes. Two PI minor resistant mutations were identified in the *pol* sequences of sample 1151. These were M36L and I93L. The blue lines indicate differences from the consensus B sequences, with the tall blue lines indicating the positions of possible drug resistant mutations.

### 3.8 Near full-length characterisation of possible HIV-1 recombinant strains

Attempts were made to characterise the complete genomes of the 3 possible recombinant samples, especially sample 1154, a possible C/D recombinant strain. None of the full-length amplification procedures (methods described in chapter two) was successful. Due to the low viral load of sample 1154 (LDL), attempts were made to amplify the genomic DNA of this sample with the Repli-g system (Figure 3.31). The DNA concentration was successfully increased from 41.7 ng per  $\mu\text{l}$  to 2581.3 ng per  $\mu\text{l}$  with a purity value of 1.82. The positive control DNA concentration was increased from 10 ng per  $\mu\text{l}$  to 981.5 ng per  $\mu\text{l}$  with a purity value of 1.73. However, even after genomic DNA amplification no full-length PCR products could be generated. Full-length sequences are needed to identify breakpoints of recombination events and without the complete genome sequences sample 1154 remains difficult to characterise.

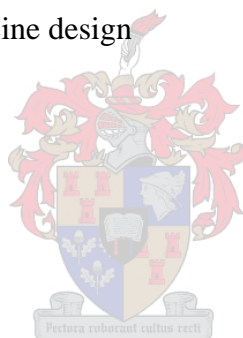


**Figure 3.31: Amplification of genomic DNA from sample 1154.** Lanes: Lane 1 – control DNA; Lane 2- 1154 DNA before amplification; Lane 3 – 1154 DNA after amplification with Repli-g kit; Lane M – 1 kb DNA Marker.

## CHAPTER FOUR

### 4. Discussion and Conclusion

	<b>PAGE</b>
4.1 Discussion	110
4.1.1 Introduction	110
4.1.2 HIV-1 in Khayelitsha	110
4.1.3 HIV-1 serotyping compared to HIV-1 genotyping	111
4.1.4 HIV-1 nucleotide substitution rates	112
4.1.5 Phylogenetic analysis	112
4.1.6 The role of variable and conserved genome regions of HIV-1	114
4.1.7 The <i>env</i> gp120 V3 loop	115
4.1.8 HIV-1 ART and drug resistance testing	116
4.1.9 Implications for vaccine design	117
4.2 Conclusion	118



## **4.1 Discussion**

### **4.1.1 Introduction**

The objective of this study was to identify HIV-1 subtypes circulating in the Khayelitsha community. This was done by characterising HIV-1 positive samples received from the Matthew Goniwe clinic in Khayelitsha. Molecular techniques included serotyping and genotyping methods. Detailed phylogenetic analysis was done on the HIV-1 sequences obtained. The major findings of the study are discussed in this chapter.

### **4.1.2 HIV-1 in Khayelitsha**

Khayelitsha is a huge, mostly poverty-stricken, informal settlement within the Western Cape Province of South Africa. The Khayelitsha and Gugulethu / Nyanga districts have the highest HIV-1 prevalence rates in the Western Cape, 27.2% and 28.1% respectively. These are also the poorest areas in the province. The lack of education and unsafe sexual practices play a key role in contributing towards the growing HIV/AIDS prevalence rates seen in Khayelitsha and Gugulethu / Nyanga. The biggest HIV/AIDS risk group in Khayelitsha, and the rest of the Western Cape, are women between the ages of 15 and 24 (Department of Health, 2004). Khayelitsha is one of the first sites in which the South African National Government ART rollout programme was implemented. There are currently more than 1000 people on ART drug regimens in the district that are being followed-up at regularly intervals (MSF, 2003; WHO, 2004).

The patient samples of this study were collected before the ART programme was implemented and patients had not received treatment at the time of sampling. Samples were collected anonymously and patient follow-up was a difficult task. A previously established study population, headed by Dr. J. Fincham at the MRC, made it easier to gain access to HIV-1 positive patients residing in Khayelitsha. For an epidemiological survey to be more representative of the complete study population, samples ideally should be taken from as many different sources as possible. If the ART programme grows and proves to be successful, such studies might be possible in the future. This



could lead to the expansion of community based HIV-1 research opportunities in South Africa and improved awareness and treatment campaigns.

#### **4.1.3 HIV-1 serotyping compared to HIV-1 genotyping**

Serotyping is an easy and rapid method to assess the HIV-1 epidemiology in a specific geographical area and has been found to correlate well with genotyping (Apetrei *et al*, 1998). The method has been used extensively before in England, Romania and South Africa (Apetrei *et al*, 1998; Engelbrecht *et al*, 1999; Smith *et al*, 2005). The antigen-antibody cross-reactivity between HIV-1 subtypes is one of the major disadvantages of this technique, as observed during this study. This can lead to the incorrect characterisation of HIV-1 subtypes (Apetrei *et al*, 1998). The assay often cannot differentiate between HIV-1 subtypes A and C (Murphy *et al*, 1999). Although the nucleotide sequences of HIV-1 subtype A and C seem to be very different, the *env* gp120 V3 region encode for structurally and thus functionally similar peptides. Thus, antigens recognising HIV-1 subtype A peptides will often recognise HIV-1 subtype C peptides, and *vice versa* (Murphy *et al*, 1999). This is also true for subtype B and D peptides, which have been shown to be closely related and possibly share a common ancestor (Thomson *et al*, 2002). This is an important factor, as subtype C is the dominant HIV-1 subtype in South Africa (Gordon *et al*, 2003; Novitsky *et al*, 1999; Van Harmelen *et al*, 1999a; Van Harmelen *et al*, 1999b) with subtype A much less prevalent (Bredell *et al*, 2000; Hunt *et al*, 2001; Moodley *et al*, 2003; Papathanasopoulos *et al*, 2002; Van Harmelen *et al*, 1999a). Another negative aspect of the cPEIA is the miss-characterisation of false-negative samples, such as the NR Khayelitsha samples. The result is often caused by HIV divergence, leaving human antibodies to HIV unable to recognise synthetic antigen peptides (McAlpine *et al*, 1995; Preiser *et al*, 2000).

Since the cPEIA is based on a small antigenic HIV-1 Env domain, recombination in other areas of the genome cannot be detected with serotyping. Any organism's genotype can be investigated and characterised by determining the relevant nucleotide base sequence of that organism. Genotyping is based on larger, or different genomic domain areas, can detect HIV-1 recombination and is much more specific than serotyping (Kandathil *et al*, 2005). The DNA sequence obtained does not reveal any

structural information about the gene or protein being investigated. Assumptions of a protein structure can be made by comparing sequence similarities of proteins or genes with similar features or properties. If the complete genome cannot be characterised, smaller genes or gene fragments can be used for more accurate genotyping. To characterise a complete genome of any organism can be labour-intensive, time-consuming and expensive (Kandathil *et al*, 2005). It is not always possible to genotype all HIV-1 positive samples in a study population due to the high variability of the HIV-1 genome (Coffin, 1995; Spira *et al*, 2003). Serotyping can thus be employed as a fast screening method with genotyping being reserved to further investigate possible HIV-1 recombinant strains, or any other interesting strains observed.

#### **4.1.4 HIV-1 nucleotide substitution rates**

By testing aligned datasets against different nucleotide substitution models of evolution, it is evident that not all sequence data can be treated the same way. The nucleotide substitution rate across the HIV-1 genome is not uniform. This is especially true for the *env* gp120 region where five highly variable regions occur (V1 to V5) (Starcich *et al*, 1986), the V3 region being important for this study. The *gag*, *env* and *pol* genes have very different nucleotide compositions and functions. The rate at which each of these genes evolves, or change its nucleotide composition, over time is clearly different (Leitner *et al*, 1997). Different models of evolution should thus be applied for different data sets, as well as for different genomic regions. Some genomic areas, such as the *env* gp120 variable loops, rapidly change their nucleotide composition. This often happens under selection pressure, such as to evade the host immune response (Kato *et al*, 1999; Yang *et al*, 1994). In other functionally important genomic areas, such as the *pol* gene that encodes for biologically important enzymes (Caumont *et al*, 2001; Cornelissen *et al*, 1997), nucleotide substitutions occur less frequently.

#### **4.1.5 Phylogenetic analysis**

As the result of the rapid rate of HIV evolution, phylogenetic analysis has become a useful tool in studying HIV. The method not only determines the subtype of a query sequence, but also examines the relationship between a set of aligned sequences. Phylogenetic analysis has supported the hypothesis that HIVs are derived from related

SIVs commonly found in non-human primates (Hahn *et al*, 2000). It has also demonstrated that HIV-1 group M probably first evolved during the 1930s ( $\pm 20$  years) (Hahn *et al*, 2000; Korber *et al*, 2000) and has identified distinct HIV groups and subtypes. Phylogenetic relationships between the different HIV strains from different epidemiological areas provide valuable information on the origin and propagation of viruses in a population (Salemi and Vandamme, 2003; Thomson *et al*, 2002). Another useful phylogenetic application is the identification of epidemiologically linked groups of sequences. By identifying such sequences the possible transmission history of HIV-1 strains can be investigated (Leitner *et al*, 1996b).

Both neighbour-joining and maximum likelihood methods were used to characterise the sequences obtained from the Khayelitsha samples. Although maximum likelihood trees are assumed to be more correct based on the models of evolution employed, both forms of trees generated clearly distinguish between the different HIV-1 phylogenetic subtypes. The neighbour-joining method is usually preferred for a quick analysis of HIV-1 query samples, while more intensive analysis, such as origin of subtypes, is preferred with the maximum likelihood method (Salemi and Vandamme, 2003). Neighbour-joining trees only show the relationship between the sequences based on sequence similarity and holds no information on the history or direction of HIV-1 evolution and transmission. Maximum likelihood intensively compares the sequences with each other based on algorithms that hypothesis a possible history of transmission, based on possible ancestral sequences. Therefore, maximum likelihood analysis is computer intensive, time-consuming and not always necessary. For HIV-1 subtype characterisation and epidemiological studies, based on large data sets, neighbour-joining analysis seems to be sufficient (Nei and Kumar, 2000; Page and Holmes, 2002; Salemi and Vandamme, 2003).

Phylogenetic analysis of the Khayelitsha sequences reveals that HIV-1 subtype C most probably predominates in this community. HIV-1 subtype C is responsible for at least 46% of all HIV-1 infections in Africa (Esparza and Bhamarapavati, 2000; Neilson *et al*, 1999; Osmanov *et al*, 2002; Renjifo *et al*, 1998; Vidal *et al*, 2000b). A wide range of phylogenetic subclusters of subtype C is observed in the *gag* p24, *env* gp41 IDR and *env* gp120 V3 genome regions. This is indicative that the subtype C population in Khayelitsha does not originate from a single common ancestor, but is the result of

multiple introductions of the virus from different geographical sources (Abebe *et al*, 2001; Novitsky *et al*, 1999; Travers *et al*, 2004). Closely related subtype C sequence pairs, from samples from patients that might possibly be related, were also identified and highlighted. A single subtype D *gag* p24 sequence (sample 1154; *gag* p24 D, *env* gp120 V3 C) was detected amongst the Khayelitsha sequences analysed. This subtype, along with subtype B, was responsible for the initial HIV-1 epidemic in South Africa (Engelbrecht *et al*, 1995; Sher, 1989). The phylogenetic relationship of this sequence reveals that HIV-1 subtype D is still present and circulating in the community, and probably the rest of South Africa, although to a much lesser extent than subtype C (Loxton *et al*, 2005).

#### **4.1.6 The role of variable and conserved genome regions of HIV-1**

Within the HIV-1 genome the *env* gene is highly variable, while the *pol* gene is conserved as it contains important coding regions for biologically important viral enzymes. The *gag* gene, encoding for the structural proteins, is also highly conserved (Caumont *et al*, 2001; Swanson *et al*, 2003). In this study sequence analysis, based on consensus sequences, conserved genomic regions and entropy values indicate that the *gag* p24 region is the most conserved of the HIV-1 genome regions investigated. The Khayelitsha *gag* p24 sequences had the lowest entropy values.

The *gag* gene has been identified as an important antigenic region. The host immune response of an infected individual has been shown to induce specific CD4+ T-cell responses to conserved *gag* and *gag* like peptides (Venturini *et al*, 2002). The amino acid region 201 to 300, relative to HXB2, can be considered the IDR of *gag*. CD4+ T-cell responses also occur after vaccination of HIV-1 negative individuals with *gag* p24 like particles (Norris *et al*, 2001). CD4+ T-cell responses are important for controlling the host immune response to HIV-1 and other pathogens. CD4+ T-cells stimulate cells of the immune system (CTLs, B lymphocytes) to divide and grow and enhances the ability of macrophages to destroy microbes (Janeway *et al*, 2001). It is thus important to identify key conserved *gag* p24 epitopes, such as the PRGSDIAGTTS (position 231 to 241) identified in the Khayelitsha sequences. These epitopes and their importance during an immune response can subsequently be evaluated.

Within the *env* gene, the *env* gp120 V3 genome region is more variable than the *env* gp41 IDR region. Conserved epitopes within the *env* gene are important antibody recognition sites and many mAbs have been identified that target these conserved areas. These include mAb 2F5 that recognise the LDKWAS conserved epitope in the *env* gp41 IDR (Conley *et al*, 1994; McGaughey *et al*, 2003) and mAb 447 that targets the GPGQ motif in the *env* gp120 V3 region (Zolla-Pazner *et al*, 2004). These antibodies recognise conserved viral epitopes despite the designated subtype of the HIV-1 strains. These antibodies and their HIV-1 neutralising properties can play a crucial role in vaccine development (Gaschen *et al*, 1999; Gorny *et al*, 2004). Other conserved regions within the *env* gene, such as the SNKSLEQ conserved epitope identified in the Khayelitsha sequences, are also important antigenic sites that are recognised by CD4+ T-cells.

#### **4.1.7 The *env* gp120 V3 loop**

The 35 amino acid V3 loop is important for inducing an NSI / SI phenotype and predicting chemokine co-receptor usage (Fenyo *et al*, 1997; Hartley *et al*, 2005; Regoes and Bonhoeffer, 2005). It is also an important immunogenic region targeted by various neutralising antibodies, as described in section 4.1.6. Attempts by HIV-1 to escape the neutralisation effect of antibodies can lead to hypervariable mutations in the HIV-1 genome, especially in the V3 region (Inouye *et al*, 1998). Hypermutations are induced by the host defence systems to create incompetent viruses, while HIV-1 uses this mechanism to facilitate immune escape (Fitzgibbon *et al*, 1993; Mangeat *et al*, 2003; Rose and Korber, 2000). The GPGQ tetramer present at the tip with the V3 loop is found within 90% of subtype A and 98% of subtype C viruses. The tetramer is usually highly conserved amongst subtypes, with the GPGR motif associated with subtype B (Gaschen *et al*, 1999; Gorny *et al*, 2004; Zolla-Pazner *et al*, 2004). A single subtype C variant (sequence from sample 1098) containing the GPGR motif has been identified amongst the Khayelitsha sequences. HIV-1 subtype C variants from Khayelitsha predicted to use CXCR4 as their major chemokine co-receptor have also been identified. This prediction has not been phenotypically confirmed. Although not common, this feature has recently been identified in other subtype C characterisation studies as well (Cilliers *et al*, 2003; Janse van Rensburg *et al*, 2002).

The region downstream of the V3 loop, usually after the second cysteine residue, has an even higher degree of variation. This feature is common amongst all HIV-1 group M subtypes and helps facilitate host immune evasion. This observation was also noted for the Khayelitsha amino acid sequences, having higher entropy values towards the end of the region investigated (Figure 3.27). It would be interesting to compare the Khayelitsha subtype C *env* gp120 V3 loop against other subtypes in South Africa. Previous studies have shown that the V3 loop of subtype C is fairly conserved, compared to other HIV-1 group M subtypes (Peeters and Sharp, 2000; Shankarappa *et al*, 2001). *Env* gp120 V3 sequences presenting with double peaks, dual bp sequences at the same position in the sequence, indicate that diverse HIV-1 subtype C strains might be present in some of the Khayelitsha patients. The presence of multiple HIV-1 variants is associated with an increased viral load set point and can lead to a more progressive onset of AIDS (Gottlieb *et al*, 2004; Grobler *et al*, 2004; Sagar *et al*, 2004). It would be worthy to further investigate dual infections in the future to assess the implications they have on aspects of viral diversity and evolution.

#### **4.1.8 HIV-1 ART and drug resistance testing**

It is important to monitor drug resistant HIV-1 quasispecies in an infected population as drug resistant mutations can lead to an individual failing ART. Genotypic resistance testing, based on the PR and RT genes of HIV-1, may prevent patients failing ART by identifying the drug resistant mutations present in the virus population of an individual beforehand (Kantor *et al*, 2004; Lindström and Albert, 2003). This can lead to a reduced risk of HIV-1 transmission rates in the larger population. By decreasing the HIV-1 RNA plasma levels in an individual through ART, the risk of transmitting the virus to another individual in the population is reduced (Daar and Richman, 2005). Patients on ART receive a combination of NRTIs, NNRTIs, PIs and host cell entry inhibitors. By screening for drug resistant mutations, a choice of effective drug regimens can be chosen to treat HIV-1 infected patients (Johnson *et al*, 2003). Drug resistant mutations enable the virus to have a replicative advantage *in vivo*.

Two of the Khayelitsha samples (1039 and 1151) were screened for possible mutations in the PR and RT genes. They have minor PI resistance mutations at positions 36 and 93 of the *pol* gene. These mutations only infer resistance in conjunction with mutations

in other areas of the *pol* genes. Khayelitsha patients receiving ART should be tested for drug resistant mutations in the future. This will be done by characterising the RT and PR genes of HIV-1 from these patients.

#### **4.1.9 Implications for vaccine design**

An effective vaccine against HIV-1 should be able to induce both cell-mediated CTL responses and neutralising antibody immune responses (Burton *et al* 1997; Burton *et al*, 2004; Moore *et al*, 2001). There is still a lack of knowledge about the influence of genetically diverse HIV-1 subtypes on vaccine development (Moore *et al*, 2001; McKinney *et al*, 2004). A vaccine immunogen should ideally be effective against all HIV-1 subtypes. However, realistically this might not be possible due to the high diversity of HIV-1 (McKinney *et al*, 2004). Two vaccine strategies are currently being employed. The first focuses on the development of a subtype specific vaccine, based on ancestral or consensus sequences within geographical areas (Korber *et al*, 2001; Novitsky *et al*, 2002). The second regards the use of highly conserved features common amongst all group M subtypes (Wilson *et al*, 2003). Considerable efforts have been made to test a subtype C candidate vaccine for Southern Africa (Novitsky *et al*, 2002; Van Harmelen *et al*, 2003; Williamson *et al*, 2003). This is the major circulating subtype found in this region, as confirmed with the Khayelitsha sequences as well. Conserved subtype C epitopes, such as the Khayelitsha sequence epitopes identified, can thus also play a crucial role in future vaccine development strategies. Features commonly identified in HIV-1 subtype C, present in other HIV-1 group M subtypes can thus be used to help develop an HIV-1 immunogen.

The detection of a possible HIV-1 recombinant virus in Khayelitsha, and the differences observed in the subtype C strains characterised, indicate that the diversity of HIV-1 in South Africa is rapidly changing. HIV-1 CRFs, such as subtype C and D recombinant forms might become more prevalent in South Africa in the near future. Characterisation studies will help keep track of the evolution, and the extent of the diversity, of HIV-1 in this country.

## 4.2 Conclusion

The majority of HIV samples analysed belongs to HIV-1 group M subtype C. One possible HIV-1 recombinant virus (*gag* D, *env* C) was detected in Khayelitsha. However, due to a lack of sequence information recombination cannot be ruled out in 9 other samples, with a further 10 patients possibly dually infected with two or more diverse HIV-1 quasispecies. The genotyping method was found to be superior to serotyping techniques. Most of the Khayelitsha subtype C sequences resemble the typical subtype C reference sequences previously analysed and available on the LANL database with a few exceptions being detected. The *env* gp120 V3 region was found to be the most variable of the regions studied, while the *gag* p24 showed the least variation. Sequence and phylogenetic analysis indicate that the sequences from Khayelitsha are comparable to the sequences circulating in the rest of South Africa. The *gag* p24 gene is essential for capsid assembly, while the *env* gp41 IDR and *env* gp120 V3 are important coding regions of the Env proteins forming the viral envelope. Regions in the *env* gp41 IDR and *env* gp120 V3 are also important recognition sites for mAbs and play important roles in CD4<sup>+</sup> T-cell responses. The *pol* sequences analysed are known to undergo mutations that support immune evasion and can cause ART failure. PI minor resistant mutations are present in the two patients analysed for drug resistance. All the HIV-1 positive samples on ART might be tested for drug resistant mutations in the near future.

The Khayelitsha study cohort consisting of 127 patients represents only a fraction of the 5.6 million South Africans currently infected with HIV/AIDS. This is the first published epidemiological study of HIV-1 in the Western Cape from a singular point, community. In the future such studies might also include large areas from other parts of the Western Cape. It is also important to continue with epidemiological surveys to keep trend of the HIV-1 diversity in South Africa. This can help form a global picture on the dynamics of the HIV virus population. The sequences of the viruses characterised from the Khayelitsha samples will contribute to the growing HIV database. Unique features identified amongst these sequences might help the scientific community understand the HIV-1 epidemiological transmission patterns in South Africa. One such problem is the fast spreading pace of HIV-1 subtype C. Information from the Khayelitsha sequences,



together with other possible HIV-1 subtyping or epidemiological studies in South Africa might elucidate this problem in the near future.

The sporadic detection of non-subtype C and recombinant subtype C viruses in South Africa remains a concern, especially considering developing an effective vaccine against all HIV subtypes. The spread of HIV-1 and its recombinant strains will have to be closely monitored in the future. It is thus important to do full-length genome analysis on possible HIV-1 recombinant samples or unique strains identified. There is still a huge amount of work to be done before we can say with confidence that we understand the dynamics of HIV epidemiology and evolution, not only in this country, but worldwide. Ultimately, the sequence information available should be used to help develop an effective HIV-1 vaccine that induces both cellular and humoral immune responses reactive against all HIV-1 groups and subtypes.



## CHAPTER FIVE

### 5. References

Abdullah MF, Young T, Bitalo L, Coetzee N and Myers JE. Public health lessons from a pilot programme to reduce mother-to-child transmission of HIV-1 in Khayelitsha. *S. Afr. Med. J.* 2001: 91; 579-583.

Abdurahman S, Høglund S, Goobar-Larsson L and Vahlne A. Selected amino acid substitutions in the C-terminal region of human immunodeficiency virus type 1 capsid protein affect virus assembly and release. *J. Gen. Virol.* 2004: 85; 2903-2913.

Abebe A, Demissie D, Goudsmit J, Brouwer M, Kuiken CL, Pollakis G, Schuitemaker H, Fontanet AL and Rinke de Wit TF. HIV-1 subtype C syncytium- and non-syncytium-inducing phenotypes and coreceptor usage among Ethiopian patients with AIDS. *AIDS.* 1999: 13; 1305-1311.

Abramovici H. Promega RT-PCR systems explained. *Promega Notes.* 2001: 78; 21-22.

Abt. Associates Inc. and the AIDS Research Unit Metropolitan Life Ltd. Demographic impacts of HIV/AIDS in South Africa. Abt Associates Inc. Sandton, Johannesburg, South Africa. 2000.

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

Allen D, Simelela A and Makubalo L. Epidemiology of HIV/AIDS in South Africa. *Southern African Journal of HIV Medicine.* July 2000.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ. Gapped Blast and PSI-Blast: a new generation of protein database search programs. *Nucleic Acid Research.* 1997: 25; 3389-3402.

Alvarez-Munoz MT, Zaragoza-Rodriguez S, Rojas-Montes O, Palacios-Saucedo G, Vazquez-Rosales G, Gomez-Delgado A, Torres J and Munoz O. High correlation of human immunodeficiency virus type-1 viral load measured in dried-blood spot samples and in plasma under different storage conditions. *Arch. Med. Res.* 2005: 36; 382-386.

Ambion Inc. Finally, an Easier Way to Measure Nucleic Acid Concentration. Ambion and others recommend the NanoDrop® ND-1000. A Spectrophotometer for highly reproducible and accurate measurements of nucleic acid concentrations using just 1-2 µl of sample. TechNotes. Ambion Inc. Austin, Texas, USA. March 2004.

Apetrei C, Robertson DL and Marx PA. The history of SIVS and AIDS: epidemiology, phylogeny and biology of isolates from naturally SIV infected non-human primates (NHP) in Africa. *Front. Biosci.* 2004; 9; 225-254.

Apetrei C, Necula A, Holm-Hansen C, Loussert-Ajaka I, Pandrea I, Cozmei C, Streinu-Cercel A, Pascu FR, Negut E, Molnar G, Duca M, Pecec M, Brun-Vezinet F and Simon F. HIV-1 diversity in Romania. *AIDS.* 1998; 12; 1079-1085.

Ariën KK, Abraha A, Quinones-Mateu ME, Kestens L, Vanham G and Arts EJ. The replicative fitness of primary human immunodeficiency virus type 1 (HIV-1) group M, HIV-1 group O, and HIV-2 isolates. *J. Virol.* 2005; 79; 8979-8990.

Arroyo MA, Hoelscher M, Sanders-Buell E, Herbinger KH, Samky E, Maboko L, Hoffmann O, Robb MR, Birx DL and McCutchan FE. HIV type 1 subtypes among blood donors in the Mbeya region of southwest Tanzania. *AIDS Res. Hum. Retroviruses.* 2004; 20; 895-901.

Arthur LO, Bess JW Jr, Sowder RC 2nd, Benveniste RE, Mann DL, Chermann JC and Henderson LE. Cellular proteins bound to immunodeficiency viruses: implications for pathogenesis and vaccines. *Science.* 1992; 18; 1935-1938.

Ausubel FM, Brent R, Kingston RE, Moore DD, Seidman JG, Smith JA and Struhl K. *Current protocols in Molecular Biology.* John Wiley and Sons. New York, USA. 2003.

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

Badley AD. In vitro and in vivo effects of HIV protease inhibitors on apoptosis. *Cell Death Differ.* 2005; 12; S1: 924-931.

Badley AD, Roumier T, Lum JJ and Kroemer G. Mitochondrion-mediated apoptosis in HIV-1 infection. *Trends Pharmacol. Sci.* 2003; 24: 298-305.

Ball SC, Abraha A, Collins KR, Marozsan AJ, Baird H, Quinones-Mateu ME, Penn-Nicholson A, Murray M, Richard N, Lobritz M, Zimmerman PA, Kawamura T, Blauvelt A and Arts EJ. Comparing the ex vivo fitness of CCR5-tropic human immunodeficiency virus type 1 isolates of subtypes B and C. *J. Virol.* 2003; 77; 1021-1038.

Barin F, Lahbabi Y, Buzelay L, Lejeune B, Baillou-Beaufils A, Denis F, Mathiot C, M'Boup S, Vithayasai V, Dietrich U and Goudeau A. Diversity of antibody binding to V3 peptides representing consensus sequences of HIV type 1 genotypes A to E: an approach for HIV type 1 serological subtyping. *AIDS Res. Hum. Retroviruses.* 1996; 12; 1279-1289.

Barnes WM. PCR amplification of up to 35-kb DNA with high fidelity and high yield from lambda bacteriophage templates. *Proc. Natl. Acad. Sci. USA.* 1994; 91; 2216-2220.

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

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

Berry DJ, Yach D and Hennink MH. An evaluation of the national measles vaccination campaign in the new shanty areas of Khayelitsha. *S. Afr. Med. J.* 1991; 79; :433-436.

Bertani G and Weigle JJ. Host controlled variation in bacterial viruses. *J. Bacteriol.* 1953; 65; 113-121.

Berthet-Colominas C, Monaco S, Novelli A, Sibai G, Mallet F and Cusack S. Head-to-tail dimers and interdomain flexibility revealed by the crystal structure of HIV-1 capsid protein (p24) complexed with a monoclonal antibody Fab. *EMBO. J.* 1999; 18; 1124-1136.

Bickel PJ, Cosman PC, Olshen RA, Spector PC, Rodrigo AG and Mullins JI. Covariability of V3 loop amino acids. *AIDS Res. Hum. Retroviruses.* 1996; 12; 1401-1411.

Binley JM, Wrinn T, Korber B, Zwick MB, Wang M, Chappey C, Stiegler G, Kunert R, Zolla-Pazner S, Katinger H, Petropoulos CJ and Burton DR. Comprehensive cross-clade neutralization analysis of a panel of anti-human antibodies. *J. Virol.* 2004; 78; 13232-13252.

Bjorndal A, Sonnerborg A, Tscherning C, Albert J and Fenyo EM. Phenotypic characteristics of human immunodeficiency virus type 1 subtype C isolates of Ethiopian AIDS patients. *AIDS Res. Hum. Retroviruses.* 1999; 15; 647-653.

Blanco L, Bernad A, Lazaro JM, Martin G, Garmendia C and Salas M. Highly efficient DNA synthesis by the phage phi 29 DNA polymerase. Symmetrical mode of DNA replication. *J. Biol. Chem.* 1989; 264; 8935-8940.

Bloom AL. Acquired immunodeficiency syndrome and other possible immunological disorders in European haemophiliacs. *Lancet.* 1984; 1; 1452-1455.

Blom N, Gammeltoft S and Brunak S. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol.* 1999; 294; 1351-1362.

Bobkov AF, Pokrovskii VV, Selimova LM, Kazennova EV, Karaseva NG, Ladnaia NN, Kravchenko AV, Cheingsong-Popov R and Veber D. Genotyping and phylogenetic analysis of HIV-1 isolates circulating in Russia. *Vopr. Virusol.* 1997; 42; 13-16.

Bock PJ and Markovitz DM. Infection with HIV-2. *AIDS.* 2001; 15; S35-S45.

Bohm L. Primary health care in Khayelitsha. *S. Afr. Med. J.* 1996; 86; 847-848.

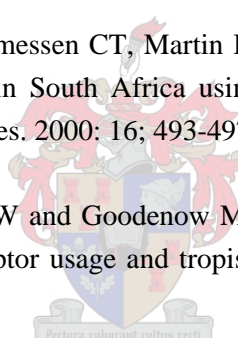
Borsetti A, Ohagen A and Gottlinger HG. The C-terminal half of the human immunodeficiency virus type 1 Gag precursor is sufficient for efficient particle assembly. *J. Virol.* 1998; 72; 9313-9317.

Bour S and Strebel K. The HIV-1 Vpu protein: a multifunctional enhancer of viral particle release. *Microbes. Infect.* 2003; 5; 1029-1039.

Bradley AF. Recent developments in automatic DNasequencing. *Pure and App. Chern.* 1996; 68; 1907-1912.

Bredell H, Hunt G, Morgan B, Tiemessen CT, Martin DJ and Morris L. Identification of HIV type 1 intersubtype recombinants in South Africa using env and gag heteroduplex mobility assays. *AIDS Res. Hum. Retroviruses.* 2000; 16; 493-497.

Briggs DR, Tuttle DL, Sleasman JW and Goodenow MM. Envelope V3 amino acid sequence predicts HIV-1 phenotype (co-receptor usage and tropism for macrophages). *AIDS.* 2000; 14; 2937-2939.



Briggs JAG, Wilk T, Welker R, Kräusslich H, Stephen D and Fuller SD. Structural organization of authentic, mature HIV-1 virions and cores. *EMBO. J.* 2003; 22; 1707-1715.

Brumme ZL, Dong WW, Yip B, Wynhoven B, Hoffman NG, Swanstrom R, Jensen MA, Mullins JI, Hogg RS, Montaner JS and Harrigan PR. Clinical and immunological impact of HIV envelope V3 sequence variation after starting initial triple antiretroviral therapy. *AIDS.* 2004; 18; F1-F9.

Bruno WJ and Halpern AL. Topological bias and inconsistency of maximum likelihood using wrong models. *Mol. Biol. Evol.* 1999; 16; 564-566.

Brust S, Duttman H, Feldner J, Gurtler L, Thorstensson R and Simon F. Shortening of the diagnostic window with a new combined HIV p24 antigen and anti-HIV-1/2/O screening test. *J. Virol. Methods.* 2000; 90; 153-165.

Burton DR and Montefiori DC. The antibody response in HIV-1 infection. *AIDS.* 1997; 11; S87-S98.

Burton DR, Desrosiers RC, Doms RW, Koff WC, Kwong PD, Moore JP, Nabel GJ, Sodroski J, Wilson IA and Wyatt RT. HIV vaccine design and the neutralizing antibody problem. *Nat. Immunol.* 2004; 5; 233–236.

Burton DR, Pyati J, Koduri R, Sharp SJ, Thornton GB, Parren PWHL., Sawyer LS, Hendry RM, Dunlop N and Nara PL. Efficient neutralization of primary isolates of HIV-1 by a recombinant human monoclonal antibody. *Science.* 1994; 266; 1024–1027.

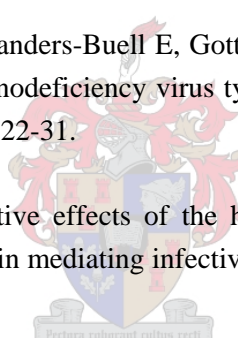
Cano A, Viveros M, Acero G, Govezensky T, Munguia ME, Gonzalez E, Soto L, Gevorkian G and Manoutcharian K. Antigenic properties of phage displayed peptides comprising disulfide-bonded loop of the immunodominant region of HIV-1 gp41. *Immunol. Lett.* 2004; 95: 207-212.

Carr A and Cooper DA. HIV protease inhibitors. *AIDS.* 1996; 10; S151-157.

Carr JK, Salminen MO, Koch C, Gotte D, Artenstein AW, Hegerich PA, St Louis D, Burke DS and McCutchan FE. Full-length sequence and mosaic structure of a human immunodeficiency virus type 1 isolate from Thailand. *J. Virol.* 1996; 70; 5935-5943.

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

Carrillo A and Ratner L. Cooperative effects of the human immunodeficiency virus type 1 envelope variable loops V1 and V3 in mediating infectivity for T cells. *J. Virol.* 1996; 70; 1310-1316.



Cassol S, Gill MJ, Pilon R, Cormier M, Voigt RF, Willoughby B and Forbes J. Quantification of human immunodeficiency virus type 1 RNA from dried plasma spots collected on filter paper. *J. Clin. Microbiol.* 1997; 35; 2795-2801.

Cassol S, Salas T, Arella M, Neumann P, Schechter MT and O'Shaughnessy M. Use of dried blood spot specimens in the detection of human immunodeficiency virus type 1 by the polymerase chain reaction. *J. Clin. Microbiol.* 1991; 29; 667-671.

Caumont A, Lan NT, Uyen NT, Hung PV, Schvoerer E, Urriza MS, Roques P, Schrive MH, Lien TT, Lafon ME, Dormont D, Barre-Sinoussi F and Fleury HJ. Sequence analysis of env C2/V3, gag p17/p24, and pol protease regions of 25 HIV type 1 sequences from Ho Chi Minh City, Vietnam. *AIDS Res. Hum. Retroviruses.* 2001; 17; 1285-1291.

Chackerian B, Rudensey LM and Overbaugh J. Specific N-linked and O-linked glycosylation modifications in the envelope V1 domain of simian immunodeficiency virus variants that evolve in the host alter recognition by neutralizing antibodies. *J. Virol.* 1997; 71; 7719-7727.

Chen Z, Huang Y, Zhao X, Skulsky E, Lin D, Ip J, Gettie A and Ho DD. Enhanced infectivity of an R5-tropic simian/human immunodeficiency virus carrying human immunodeficiency virus type 1 subtype C envelope after serial passages in pig-tailed macaques (*Macaca nemestrina*). *J. Virol.* 2000; 74; 6501-6510.

Chièn A, Edgar DB and Trela JM. Deoxyribonucleic acid polymerase from the extreme thermophile *Thermus aquaticus*. *J. Bacteriol.* 1976; 127; 1550-1557.

Chiu HC, Yao SY and Wang CT. Coding sequences upstream of the human immunodeficiency virus type 1 reverse transcriptase domain in Gag-Pol are not essential for incorporation of the Pr160(gag-pol) into virus particles. *J. Virol.* 2002; 76; 3221-3231.

Cho MW, Lee MK, Carney MC, Berson JF, Doms RW and Martin MA. Identification of determinants on a dualtropic human immunodeficiency virus type 1 envelope glycoprotein that confer usage of CXCR4. *J. Virol.* 1998; 72; 2509-2515.

Chomczynski P. Solubilization in formamide protects RNA from degradation. *Nucleic Acids Res.* 1992; 20; 3791-3792.

Chopra M, Piwoz E, Sengwana J, Schaay N, Dunnett L and Sadlers D. Effect of a mother-to-child HIV prevention programme on infant feeding and caring practices in South Africa. *S. Afr. Med. J.* 2002; 92; 298-302.

Cilliers T, Nhlapo J, Coetzer M, Orlovic D, Ketas T, Olson WC, Moore JP, Trkola A and Morris L. The CCR5 and CXCR4 coreceptors are both used by human immunodeficiency virus type 1 primary isolates from subtype C. *J. Virol.* 2003; 77; 4449-4456.

Clark SJ, Saag MS, Decker WD, Campbell-Hill S, Roberson JL, Veldkamp PJ, Kappes JC, Hahn BH and Shaw GM. High titers of cytopathic virus in plasma of patients with symptomatic primary HIV-1 infection. *N. Engl. J. Med.* 1991; 324; 954-960.

Coetzee N, Yach D, Blignaut R and Fisher SA. Measles vaccination coverage and its determinants in a rapidly growing peri-urban area. *S. Afr. Med. J.* 1990; 78; 733-737.

Coffin JM. HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy. *Science.* 1995; 267; 483-489.

Coffin JM, Haase A and Levy JA. What to call the AIDS virus? *Nature.* 1986b; 321; 10.

Coffin JM, Haase A, Levy JA, Montagnier L, Oroszlan S, Teich N, Temin H, Toyoshima K, Varmus H, Vogt P and Weiss R. Human immunodeficiency viruses. *Science.* 1986a; 232; 697.

Cohen SN, Chang AC, Boyer HW and Helling RB. Construction of biologically functional bacterial plasmids in vitro. *Proc. Natl. Acad. Sci. USA.* 1973: 70; 3240-3244.

Cornelissen M, van den Burg R, Zorgdrager F, Lukashov V and Goudsmit J. *pol* Gene diversity of five human immunodeficiency virus type 1 subtypes: Evidence for naturally occurring mutations that contribute to drug resistance, limited recombination patterns and common ancestry for subtypes B and D. *J. Virol.* 1997: 71; 6348-6358.

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

D'Souza V and Summers MF. How Retroviruses select their genomes. *Nature Reviews Microbiology.* 2005: 3; 643-655.

Daar ES and Richman DD. Confronting the emergence of drug-resistant HIV type 1: impact of antiretroviral therapy on individual and population resistance. *AIDS Res. Hum. Retroviruses.* 2005: 21; 343-357.

Damond F, Worobey M, Campa P, Farfara I, Colin G, Matheron S, Brun-Vezinet F, Robertson DL and Simon F. Identification of a highly divergent HIV type 2 and proposal for a change in HIV type 2 classification. *AIDS Res. Hum. Retroviruses.* 2004: 20; 666-672.

De Jong JJ, De Ronde A, Keulen W, Tersmette M, Goudsmit J. Minimal requirements for the human immunodeficiency virus type 1 V3 domain to support the syncytium-inducing phenotype: analysis by single amino acid substitution. *J. Virol.* 1992: 66; 6777-6780.

De Tolley J and Nash M. The Road to Khayelitsha and Beyond, Black Sash National Conference, Cape Town, South Africa. 1984.

Dean FB, Hosono S, Fang L, Wu X, Faruqi AF, Bray-Ward P, Sun Z, Zong Q, Du Y, Du J, Driscoll M, Song W, Kingsmore SF, Egholm M and Lasken RS. Comprehensive human genome amplification using multiple displacement amplification. *Proc. Natl. Acad. Sci. USA.* 2002: 99; 5261-5266.

Delwart E, Magierowska M, Royz M, Foley B, Peddada L, Smith R, Heldebrant C, Conrad A and Busch M. Homogeneous quasispecies in 16 out of 17 individuals during very early HIV-1 primary infection. *AIDS.* 2002: 16: 189-195.

Department of Health. City of Cape Town Department of Health. Tuberculosis Service Review, Khayelitsha Health District. Cape Town, South Africa: 2002.



Department of Health. City of Cape Town Department of Health. Demographic characteristics of Cape Town. Cape Town, South Africa: 2004.

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

Dittmar MT, Simmons G, Donaldson Y, Simmonds P, Clapham PR, Schulz TF and Weiss RA. Biological characterization of human immunodeficiency virus type 1 clones derived from different organs of an AIDS patient by long-range PCR. *J. Virol.* 1997; 71; 5140-5147.

Domingo E, Menendez-Arias L and Holland JJ. RNA virus fitness. *Rev. Med. Virol.* 1997; 7; 87-96.

Dorfman T, Bukovsky A, Ohagen A, Hoglund S and Gottlinger HG. Functional domains of the capsid protein of human immunodeficiency virus type 1. *J. Virol.* 1994; 68; 8180-8187.

Dorn J, Masciotra S, Yang C, Downing R, Biryahwaho B, Mastro TD, Nkengasong J, Pieniazek D, Rayfield MA, Hu DJ and Lal RB. Analysis of genetic variability within the immunodominant epitopes of envelope gp41 from human immunodeficiency virus type 1 (HIV-1) group M and its impact on HIV-1 antibody detection. *J. Clin. Microbiol.* 2000; 38; 773-780.

Dorrington R. Population Projections for the Western Cape to 2025. Cape Town, South Africa: Centre for Actuarial Research, University of Cape Town; 2002.

Dorrington R, Bourne D, Bradshaw D, Laubscher R and Timaeus IM. The impact of HIV/AIDS on adult mortality in South Africa. Technical report. Tygerberg: Burden of Disease Research Unit of the Medical Research Council of South Africa. 2001: 6.

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

Dunkle KL, Jewkes RK, Brown HC, Gray GE, McIntyre JA and Harlow SD. Gender-based violence, relationship power, and risk of HIV infection in women attending antenatal clinics in South Africa. *Lancet.* 2004; 363; 1415-1421.

Eaton L, Flisher AJ and Aaro LE. Unsafe sexual behaviour in South African youth. *Soc. Sci. Med.* 2003; 56; 149-165.

Eck RV and Dayhoff MO. Atlas of Protein Sequence and Structure. National Biomedical Research Foundation, Silver Springs, Maryland, USA. 1966.

Eickbush TH and Moudrianakis EN. The compaction of DNA helices into either continuous supercoils or folded-fiber rods and toroids. *Cell*. 1978; 13; 295–306.

Efron B, Halloran E and Holmes S. Bootstrap confidence levels for phylogenetic trees. *Proc. Natl. Acad. Sci. USA*. 1996; 12; 13429-13434.

Emerman M and Malim MH. HIV-1 regulatory / accessory genes: keys to unraveling viral and host cell biology. *Science*. 1998; 280; 1880-1884.

Engelbrecht S, Laten J, Smith TL and J. van Rensburg E. Identification of *env* subtypes in fourteen HIV type 1 isolates from South Africa. *AIDS Res. Hum. Retroviruses*. 1995; 11; 1269-1271.

Engelbrecht S, Smith TL, Kasper P, Faatz E, Zeier M, Moodley D, Clay CG and van Rensburg EJ. HIV type 1 V3 Domain serotyping and genotyping in Gauteng, Mpumalanga, Kwazulu-Natal and Western Cape provinces of South Africa. *AIDS Res. Hum. Retroviruses*. 1999; 15; 325-328.

Eshleman SH, Guay LA, Fleming T, Mwatha A, Mracna M, Becker-Pergola G, Musoke P, Mmiro F and Jackson JB. Survival of Ugandan infants with subtype A and D HIV-1 infection (HIVNET 012). *J. Acquir. Immune. Defic. Syndr*. 2002; 31; :327-330.

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

Fang G, Kuiken C, Weiser B, Rowland-Jones S, Plummer F, Chen CH, Kaul R, Anzala AO, Bwayo J, Kimani J, Philpott SM, Kitchen C, Sinsheimer JS, Gaschen B, Lang D, Shi B, Kemal KS, Rostron T, Brunner C, Beddows S, Sattenau Q, Paxinos E, Oyugi J and Burger H. Long-term survivors in Nairobi: complete HIV-1 RNA sequences and immunogenetic associations. *J. Infect. Dis*. 2004; 190; 697-701.

Fang G, Weiser B, Visosky AA, Townsend L and Burger H. Molecular cloning of full-length HIV-1 genomes directly from plasma viral RNA. *J. Acquir. Imm. Def. Syndrome and Hum. Retrovirology*. 1996; 12; 352-357.

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

Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol*. 1981; 17; 368-376.

Felsenstein J. Maximum likelihood and minimum-steps methods for estimating evolutionary trees. *Syst. Zool*. 1973; 22; 240-249.

Felsenstein J. PHYLIP (Phylogeny Inference Package). 3.5c. Department of Genetics, University of Washington, Seattle, USA. 1993.

Fenyo EM, Morfeldt-Manson L, Chiodi F, Lind B, von Gegerfelt A, Albert J, Olausson E and Asjo B. Distinct replicative and cytopathic characteristics of human immunodeficiency virus isolates. *J. Virol.* 1988; 62; 4414-4419.

Fenyo EM, Schuitemaker H, Asjö B, McKeatin J, Sattentau Q and the EC Concerted Action HIV Variability. The History of HIV-1 Biological Phenotypes. Past, Present and Future. In: Human Retroviruses and AIDS 1997: Korber B, Hahn BH, Foley B, Mellors J, Leitner T, Myers G, McCutchan F and Kuiken C. A compilation and analysis of nucleic acid and amino acid sequences.. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, New Mexico, USA. 1997.

Fincham JE. Tracking potential for exposure to helminthic antigens to impair SAAVI vaccine trials and efficacy of HIV vaccine. HIV vaccination and deworming. Protocol. 2004.

Fincham JE, Markus MB and Brombacher F. Vaccination against helminths: influence on HIV/AIDS and TB. *Trends Parasitol.* 2002; 18; 385-386.

Fincham JE, Markus MB and Adams VJ. Could control of soil-transmitted helminthic infection influence the HIV/AIDS pandemic. *Acta. Trop.* 2003; 86; 315-333.

Fischl MA, Richman DD, Grieco MH, Gottlieb MS, Volberding PA, Laskin OL, Leedom JM, Groopman JE, Mildvan D and Schooley RT. The efficacy of azidothymidine (AZT) in the treatment of patients with AIDS and AIDS-related complex. A double-blind, placebo-controlled trial. *N. Engl. J. Med.* 1987; 317; 185-191.

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

Fitzgibbon JE, Mazar S and Dubin DT. A new type of G-->A hypermutation affecting human immunodeficiency virus. *AIDS Res. Hum. Retroviruses.* 1993; 9; 833-838.

Forshey BM, von Schwedler U, Sundquist WI and Aiken C. Formation of a human immunodeficiency virus type 1 core of optimal stability is crucial for viral replication. *J. Virol.* 2002; 76; 5667-5677.

Fouchier RA, Broersen SM, Brouwer M, Tersmette M, Van't Wout AB, Groenink M and Schuitemaker H. Temporal relationship between elongation of the HIV type 1 glycoprotein 120 V2 domain and the conversion toward a syncytium-inducing phenotype. *AIDS Res. Hum. Retroviruses.* 1995; 11; 1473-1478.

Fouchier RA, Groenink M, Kootstra NA, Tersmette M, Huisman HG, Miedema F and Schuitemaker H. Phenotype-associated sequence variation in the third variable domain of the human immunodeficiency virus type 1 gp120 molecule. *J. Virol.* 1992; 66; 3183-3187.

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

Freed EO. HIV-1 replication. *Somat. Cell. Mol. Genet.* 2001; 26; 13-33.

Frey B, Suppmann B. Demonstration of the Expand PCR System's Greater Fidelity and Higher Yields with a *lacI*-based PCR Fidelity Assay. *Biochemica.* 1995; 2; 8-9.

Friedman-Kien AE, Laubenstein L, Marmor M, Hymes K, Green J, Ragaz A, Gottlieb J, Muggia F, Demopoulos R, Weintraub M, Williams D, Oliveri R, Marmer J, Wallace J, Halperin I, Gillooley JF, Prose N, Klein E, Vogel J, Safai B, Myskowski P, Urmacher C, Koziner B, Nisce L, Kris M, Armstrong D, Gold J, Mildran D, Tapper M, Weissman JB, Rothenberg R, Friedman SM, Siegal FP, Groundwater J, Gilmore J, Coleman D, Follansbee S, Gullett J, Stegman SJ, Wofsy C, Bush D, Drew L, Braff E, Dritz S, Klein M, Preiksaitis JK, Gottlieb MS, Jung R, Chin J and Goedert J. Kaposi's sarcoma and Pneumocystis pneumonia among homosexual men - New York City and California. *MMWR.* 1981; 30:305-308.

Gallo RC, Salahuddin SZ, Popovic M, Shearer GM, Kaplan M, Haynes BF, Palker TJ, Redfield R, Oleske J and Safai B. Frequent detection and isolation of cytopathic retroviruses (HTLV-III) from patients with AIDS and at risk for AIDS. *Science.* 1984; 224; 500-503.

Gao F, Bailes E, Robertson DL, Chen Y, Rodenburg CM, Michael SF, Cummins LB, Arthur LO, Peeters M, Shaw GM, Sharp PM and Hahn BH. Origin of HIV-1 in the chimpanzee *Pan troglodytes*. *Nature.* 1999; 397; 4360-4441.

Gao F, Morrison SG, Robertson DL, Thornton CL, Craig S, Karlsson G, Sodroski J, Morgado M, Galvao-Castro B, von Briesen H, Beddows S, Weber J, Sharp PM, Shaw GM and Hahn BH. Molecular cloning and analysis of functional envelope genes from human immunodeficiency virus type 1 sequence subtypes A through G. *J. Virol.* 1996; 70; 1651-1667.

Gao F, Robertson DL, Carruthers CD, Morrison SG, Jian B, Chen Y, Barré-Sinoussi F, Girard M, Srinivasan A, Abimiku AG, Shaw GM, Sharp PM and Hahn BH. A comprehensive panel of near-full-length clones and reference sequences for non-subtype B sequences of Human Immunodeficiency Virus type 1. *J. Virol.* 1998; 72; 5680-5698.

Gao F, Vidal N, Li Y, Trask SA, Chen Y, Kostrikis LG, Ho DD, Kim J, Oh MD, Choe K, Salminen M, Robertson DL, Shaw GM, Hahn BH and Peeters M. Evidence of two distinct subsubtypes within the HIV-1 subtype A radiation. *AIDS Res. Hum. Retroviruses.* 2001; 17; 675-688.

Gao F, Yue L and White AT. Human infection by genetically diverse SIVSM-related HIV-2 in west Africa. *Nature*. 1992: 358; 495-499.

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

Gaschen B, Korber BT, and Foley DT: Global variation in the HIV-1 V3 region. In: *Human Retroviruses and AIDS 1999*: Kuiken C, Foley B, Hahn B, Marx P, McCutchan F, Mellors JW, Mullins J, Wolinsky S and Korber B. A compilation and analysis of nucleic acid and amino acid sequences. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, New Mexico, USA. 1999.

Gatignol A and Jeang KT. Tat as a transcriptional activator and a potential therapeutic target for HIV-1. *Adv. Pharmacol.* 2000: 48: 209-227.

Goldman N and Yang Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* 1994: 11; 725-736.

Gonzales MJ, Machekano RN and Shafer RW. Human immunodeficiency virus type 1 reverse-transcriptase and protease subtypes: classification, amino acid mutation patterns, and prevalence in a northern California clinic-based population. *J. Infect. Dis.* 2001: 184; 998-1006.

Gordon M, De Oliveira T, Bishop K, Coovadia HM, Madurai L, Engelbrecht S, Janse van Rensburg E, Mosam A, Smith A and Cassol S. Molecular characteristics of human immunodeficiency virus type 1 subtype C viruses from KwaZulu-Natal, South Africa: implications for vaccine and antiretroviral control strategies. *J. Virol.* 2003: 77; 2587-2599.

Gordon CJ and Delwart EL. Genetic diversity of primary HIV-1 isolates and their sensitivity to antibody-mediated neutralization. *Virology*. 2000. 272: 326-330.

Gorny MK, Revesz K, Williams C, Volsky B, Louder MK, Anyangwe CA, Krachmarov C, Kayman SC, Pinter A, Nadas A, Nyambi PN, Mascola JR and Zolla-Pazner S. The V3 loop is accessible on the surface of most human immunodeficiency virus type 1 primary isolates and serves as a neutralization epitope. *J. Virol.* 2004: 78; 2394-2404.

Goto T, Nakai M and Ikuta K. The life-cycle of human immunodeficiency virus type 1. *Micron*. 1998: 2-3; 123-138.

Gottlieb GS, Nickle DC, Jensen MA, Wong KG, Grobler J, Li F, Liu SL, Rademeyer C, Learn GH, Karim SS, Williamson C, Corey L, Margolick JB and Mullins JI. Dual HIV-1 infection associated with rapid disease progression. *Lancet*. 2004: 363; 619-622.

Gottlieb MS, Schanker HM, Fan PT, Saxon A and Weisman JD. Pneumocystis Pneumonia - Los Angeles. MMWR. 1981a: 30; 250-252.

Gottlieb MS, Schroff R, Schanker HM, Weisman JD, Fan PT, Wolf RA and Saxon A. Pneumocystis Carinii pneumonia and mucosal candidiasis in previously healthy homosexual men. N. Eng. J. Med. 1981b: 305; 1425-1431.

Gottlinger H. The HIV-1 assembly machine. AIDS. 2001: 15: S13 - S20.

Grobler J, Gray CM, Rademeyer C, Seoghe C, Ramjee G, Karim SA, Morris L and Williamson C. Incidence of HIV-1 dual infection and its association with increased viral load set point in a cohort of HIV-1 subtype C-infected female sex workers. J. Infect. Dis. 2004: 190; 1355-1359.

Gu X and Zhang J. A simple method for estimating the parameter of substitution rates variation among sites. Molecular Biology and Evolution. 1997: 14; 1106-1113.

Gu X, Fu YX and Li WH. Maximum-likelihood estimation of the heterogeneity of substitution rates among nucleotide sites. Molecular Biology and Evolution. 1995: 12; 546-557.

Guay LA, Musoke P, Fleming T, Bagenda D, Allen M, Nakabiito C, Sherman J, Bakaki P, Ducar C, Deseyve M, Emel L, Mirochnick M, Fowler MG, Mofenson L, Miotti P, Dransfield K, Bray D, Mmiro F and Jackson JB. Intrapartum and neonatal single-dose nevirapine compared with zidovudine for prevention of mother-to-child transmission of HIV-1 in Kampala, Uganda: HIVNET 012 randomised trial. Lancet. 1999: 354; 795-802.

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

Hall T. BioEdit, Biological sequence alignment for Windows 95/98/NT. (<http://www.mbio.ncsu.edu/BioEdit/bioedit.html>). 2001.

Hamaguchi K and Geiduschek EP. The effect of electrolytes on the stability of deoxyribonucleate helix. J. Am. Chem. Soc. 1962: 84; 1329-1337.

Hansen JE, Lund O, Nilsson J, Rapacki K and Brunak S. O-GLYCBASE Version 3.0: a revised database of O-glycosylated proteins. Nucleic. Acids. Res. 1998: 1; 387-389.

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

Hartl DL and Clark AG. Principals of population genetics. Sinauer Associates, Sunderland, Massachusetts, USA. 1997.

Hartley O, Klasse PJ, Sattentau QJ and Moore JP. V3: HIV's switch-hitter. *AIDS Res. Hum. Retroviruses*. 2005; 21; 171-189.

Hasegawa M, Kishino H and Yano T. Dating of the human-ape splitting by a mitochondrial clock of mitochondrial DNA. *J. Mol. Evol.* 1985; 12; 152-162.

Hebert DN, Zhang JX, Chen W, Foellmer B and Helenius A. The number and location of glycans on influenza hemagglutinin determine folding and association with calnexin and calreticulin. *J. Cell. Biol.* 1997; 139; 613-623.

Henderson LE, Bowers MA, Sowder RC 2nd, Serabyn SA, Johnson DG, Bess JW Jr, Arthur LO, Bryant DK and Fenselau C. Gag proteins of the highly replicative MN strain of human immunodeficiency virus type 1: posttranslational modifications, proteolytic processings, and complete amino acid sequences. *J. Virol.* 1992; 66; 1856-1865.

Hillis DM, Huelsenbeck JP and Cunningham CW. Application and accuracy of molecular phylogenies. *Science*. 1994; 264; 671-677.

Hillis DM, Moritz C and Mable BK. *Molecular Systematics*, Second Edition. Sinauer Associates, Inc., Sunderland, Massachusetts, USA. 1996.

Hirsch MS, Brun-Vezinet F, D'Aquila RT, Hammer SM, Johnson VA, Kuritzkes DR, Loveday C, Mellors JW, Clotet B, Conway B, Demeter LM, Vella S, Jacobsen DM, Richman DD. Antiretroviral drug resistance testing in adult HIV-1 infection: recommendations of an International AIDS Society-USA Panel. *JAMA*. 2000; 283; 2417-2426.

Hirsch V, Olmsted R, Murphy-Corb M, Pырcell and Johnson P. An African primate lentivirus (SIVsm) closely related to HIV-2. *Nature*. 1989; 339; 389-392.

Hoelscher M, Kim B, Maboko L, Mhalu F, von Sonnenburg F, Birx DL, McCutchan FE; UNAIDS Network for HIV Isolation and Characterization. High proportion of unrelated HIV-1 intersubtype recombinants in the Mbeya region of southwest Tanzania. *AIDS*. 2001; 15; 1461-1470.

Hoffman NG, Seillier-Moiseiwitsch F, Ahn J, Walker JM and Swanstrom R. Variability in the human immunodeficiency virus type 1 gp120 Env protein linked to phenotype-associated changes in the V3 loop. *J. Virol.* 2002; 76; 3852-3864.

Hopfner KP, Eichinger A, Engh RA, Laue F, Ankenbauer W, Huber R and Angerer B. Crystal structure of a thermostable type B DNA polymerase from *Thermococcus gorgonarius*. *Proc. Natl. Acad. Sci. USA*. 1999; 96; 3600-3605.

Horal P, Svennerholm B, Jeansson S, Rymo L, Hall WW and Vahlne A. Continuous epitopes of the human immunodeficiency virus type 1 (HIV-1) transmembrane glycoprotein and reactivity of human sera to synthetic peptides representing various HIV-1 isolates. *J. Virol.* 1991; 65; 2718-2723.

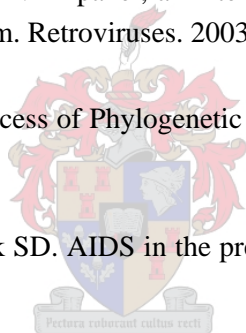
Hosegood V, Vanneste AM and Timaeus IM. Levels and causes of adult mortality in rural South Africa: the impact of AIDS. *AIDS*. 2004; 18; 663-671.

Hosono S, Faruqi AF, Dean FB, Du Y, Sun Z, Wu X, Du J, Kingsmore SF, Egholm M and Lasken RS. Unbiased whole-genome amplification directly from clinical samples. *Genome Res.* 2003; 13; 954-964.

Huang DD, Giesler TA and Bremer JW. Sequence characterization of the protease and partial reverse transcriptase proteins of the NED panel, an international HIV type 1 subtype reference and standards panel. *AIDS Res. Hum. Retroviruses*. 2003; 19; 321-328.

Huelsenbeck JP and Hillis DM. Success of Phylogenetic Methods in the Four-Taxon Case. *Sys. Biol.* 1993; 42; 247-264.

Hummer D, Rosenfeld JB and Pitlik SD. AIDS in the pre-AIDS era. *Rev. Infect. Dis.* 1987; 9; 1102-1108.



Hunt GM, Johnson D and Tiemesse CT. Characterisation of the long terminal repeat regions of South African human immunodeficiency virus type 1 isolates. *Virus Genes*. 2001; 23; 27-34.

Hunter E. gp41, A Multifunctional Protein Involved in HIV Entry and Pathogenesis. In: *Human Retroviruses and AIDS 1997*: Korber B, Hahn BH, Foley B, Mellors J, Leitner T, Myers G, McCutchan F and Kuiken C. A compilation and analysis of nucleic acid and amino acid sequences.. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, New Mexico, USA. 1997.

IAVI. International AIDS vaccine Initiative. Global distribution of HIV-1 subtypes and recombinants. 2003. ([www.iavi.org](http://www.iavi.org)).

Inouye P, Cherry E, Hsu M, Zolla-Pazner S and Wainberg MA. Neutralizing antibodies directed against the V3 loop select for different escape variants in a virus with mutated reverse transcriptase (M184V) than in wild-type human immunodeficiency virus type 1. *AIDS Res. Hum. Retroviruses*. 1998; 14; 735-740.



Janeway CA, Travers P, Walport M and Shlomchik M. Immunobiology: The immune system in health and disease. Fifth Edition. Garland Publishing, New York, USA. 2001.

Janse van Rensburg E, Smith TL, Zeier M, Robson B, Sampson C, Treurnicht F and Engelbrecht S. Change in co-receptor usage of current South African HIV-1 subtype C primary sequences. *AIDS*. 2002; 16; 2479-2480.

Janssens W, Laukkanen T, Salminen MO, et Carr JK, Van der Auwera G, Heyndrickx L, van der Groen G and McCutchan FE. HIV-1 subtype H near-full length genome reference strains and analysis of subtype-H-containing inter-subtype recombinants. *AIDS*. 2000; 14; 1533-1543.

Janssens W, Nkengasong J, Heyndrickx L, Van der Auwera G, Vereecken K, Coppens S, Willems B, Beirnaert E, Franssen K, Montavon C and Van der Groen G. Inpatient variability of HIV type 1 group O AN70 during a 10-year follow-up. *AIDS Res. Hum. Retroviruses*. 1999; 15; 1325-1332.

Javaherian K, Langlois AJ, LaRosa GJ, Protty AT, Bolognesi DP, Herlihy WC, Putney SD and Matthews TJ. Broadly neutralizing antibodies elicited by the hypervariable neutralizing determinant of HIV-1. *Science*. 1990; 250; 1590-1593.

Jeeninga RE, Hoogenkamp M, Armand-Ugon M, de Baar M, Verhoef K and Berkhout B. Functional differences between the long terminal repeat transcriptional promoters of human immunodeficiency virus type 1 subtypes A through G. *J. Virol*. 2000; 74; 3740-3751.

Jensen MA, Li FS, van 't Wout AB, Nickle DC, Shriner D, He HX, McLaughlin S, Shankarappa R, Margolick JB and Mullins JI. Improved coreceptor usage prediction and genotypic monitoring of R5-to-X4 transition by motif analysis of human immunodeficiency virus type 1 env V3 loop sequences. *J. Virol*. 2003; 77; 13376-13388.

Johanson J, Abravaya K, Caminiti W, Erickson D, Flanders R, Leckie G, Marshall E, Mullen C, Ohhashi Y, Perry R, Ricci J, Salituro J, Smith A, Tang N, Vi M and Robinson J. A new ultrasensitive assay for quantitation of HIV-1 RNA in plasma. *J. Virol. Methods*. 2001; 95; 81-92.

Johnson VA, Brun-Vezinet F, Clotet B, Conway B, D'Aquila RT, Demeter LM, Kuritzkes DR, Pillay D, Schapiro JM, Telenti A, Richman DD and the International AIDS Society-USA Drug Resistance Mutations Group. Drug resistance mutations in HIV-1. *Top. HIV Med*. 2003; 11; 215-221.

Jorgensen LB, Christensen MB, Gerstoft J, Mathiesen LR, Obel N, Pedersen C, Nielsen H and Nielsen C. Prevalence of drug resistance mutations and non-B subtypes in newly diagnosed HIV-1 patients in Denmark. *Scand. J. Infect. Dis*. 2003; 35; 800-807.

Joshi S and Joshi RL. Molecular biology of human immunodeficiency virus type-1. *Transfus. Sci.* 1996; 17; 351-378.

Jukes TH and Cantor CR. Evolution of protein molecules. In: *Mammalian Protein Metabolism* 1969: Munro HM (Ed.). Academic Press. New York, USA. 1969.

Kalia V, Sarkar S, Gupta P and Montelaro RC. Antibody neutralization escape mediated by point mutations in the intracytoplasmic tail of human immunodeficiency virus type 1 gp41. *J. Virol.* 2005; 79; 2097-2107.

Kandathil AJ, Ramalingam S, Kannangai R, David S and Sridharan G. Molecular epidemiology of HIV. *Indian J. Med. Res.* 2005; 121; 333-344.

Kantor R, Katzenstein DA, Efron B, Carvalho AP, Wynhoven B, Cane P, Clarke J, Sirivichayakul S, Soares MA, Snoeck J, Pillay C, Rudich H, Rodrigues R, Holguin A, Ariyoshi K, Bouzas MB, Cahn P, Sugiura W, Soriano V, Brigido LF, Grossman Z, Morris L, Vandamme AM, Tanuri A, Phanuphak P, Weber JN, Pillay D, Harrigan PR, Camacho R, Schapiro JM and Shafer RW. Impact of HIV-1 subtype and antiretroviral therapy on protease and reverse transcriptase genotype: results of a global collaboration. *PLoS Med.* 2005; 2; e112.

Kantor R, Shafer RW, Follansbee S, Taylor J, Shilane D, Hurley L, Nguyen DP, Katzenstein D and Fessel WJ. Evolution of resistance to drugs in HIV-1-infected patients failing antiretroviral therapy. *AIDS.* 2004; 18; :1503-1511.

Kaplan AH. Assembly of the HIV-1 core particle. *AIDS Rev.* 2002; 4; 104-111.

Kapoor A, Jones M, Shafer RW, Rhee SY, Kazanjian P, Delwart EL. Sequencing-based detection of low-frequency human immunodeficiency virus type 1 drug-resistant mutants by an RNA/DNA heteroduplex generator-tracking assay. *J. Virol.* 2004; 78; 7112-7123.

Kapp C. HIV overshadows South African health advances. Life expectancy is on the decline, and adult deaths are soaring-mainly due to HIV-related diseases. *Lancet.* 2004; 363; 1202-1203.

Karlin S and Altschul SF. Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc. Natl. Acad. Sci. USA.* 1993; 90; 5873-5877.

Karlin S and Altschul SF. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. USA.* 1990; 87; 2264-2268.

Karlsson AC, Lindback S, Gaines H and Sonnerborg A. Characterization of the viral population during primary HIV-1 infection. *AIDS.* 1998; 12; 839-847.

Kasturi L, Chen H and Shakin-Eshleman SH. Regulation of N-linked core glycosylation: use of a site-directed mutagenesis approach to identify Asn-Xaa-Ser/Thr sequons that are poor oligosaccharide acceptors. *Biochem. J.* 1997; 323; 415-419.

Kato K, Sato H and Takebe Y. Role of naturally occurring basic amino acid substitutions in the human immunodeficiency type 1 subtype E envelope V3 loop on viral coreceptor usage and cell tropism. *J. Virol.* 1999; 73; 5520-5526.

Kiepiela P, Leslie AJ, Honeyborne I, Ramduth D, Thobakgale C, Chetty S, Rathnavalu P, Moore C, Pfafferott KJ, Hilton L, Zimbwa P, Moore S, Allen T, Brander C, Addo MM, Altfeld M, James I, Mallal S, Bunce M, Barber LD, Szinger J, Day C, Klenerman P, Mullins J, Korber B, Coovadia HM, Walker BD and Goulder PJ. Dominant influence of HLA-B in mediating the potential co-evolution of HIV and HLA. *Nature.* 2004; 432; 769-775.

Kijak GH, Sanders-Buell E, Wolfe ND, Mpoudi-Ngole E, Kim B, Brown B, Robb ML, Birx DL, Burke DS, Carr JK and McCutchan FE. Development and application of a high-throughput HIV type 1 genotyping assay to identify CRF02\_AG in West/West Central Africa. *AIDS Res. Hum. Retroviruses.* 2004; 20; 521-530.

Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 1980; 16; 111-120.

Kimura M. Estimation of evolutionary distances between homologous nucleotide sequences. *Proc. Natl. Acad. Sci. USA.* 1981; 78; 454-458.

Kirby KS. A new method for the isolation of deoxyribonucleic acid: Evidence on the nature of bonds between deoxyribonucleic acid and protein. *Biochem. J.* 1957; 66; 495-504.

Koepke JA, van Assendelft OW, Bull BS and Richardson-Jones A. Standardization of EDTA anticoagulation for blood counting procedures. *Labmedica.* 1988 / 1989; 5; 15-17.

Koito A, Harrowe G, Levy JA and Cheng-Mayer C. Functional role of the V1/V2 region of human immunodeficiency virus type 1 envelope glycoprotein gp120 in infection of primary macrophages and soluble CD4 neutralization. *J. Virol.* 1994; 68; 2253-2259.

Koot M, Vos AH, Keet RP, de Goede RE, Dercksen MW, Terpstra FG, Coutinho RA, Miedema F and Tersmette M. HIV-1 biological phenotype in long-term infected individuals evaluated with an MT-2 cocultivation assay. *AIDS.* 1992; 6; 49-54.

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

Korber BT, Farber RM, Wolpert DH and Lapedes AS. Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. *Proc. Natl. Acad. Sci. USA.* 1993; 90; 7176-7180.

Korber BT, Foley BT, Kuiken CL, Pillai S and Sodroski JG. Numbering positions in HIV relative to HXB2CG. In: *Human Retroviruses and AIDS 1998: A compilation and analysis of nucleic acid and amino acid sequences.* Korber BT, Kuiken CL, Foley BT, Hahn B, McCutchan F, Mellors J and Sodroski JG. Theoretical Biology and Biophysics Group. Los Alamos National Laboratory, Los Alamos, New Mexico, USA. 1998.

Korber BT, Kunstman KJ, Patterson BK, Furtado M, McEvilly MM, Levy R and Wolinsky SM. Genetic differences between blood- and brain-derived viral sequences from human immunodeficiency virus type 1-infected patients: evidence of conserved elements in the V3 region of the envelope protein of brain-derived sequences. *J. Virol.* 1994; 68: 7467-7481.

Koulinska IN, Ndung'u T, Mwakagile D, Msamanga G, Kagoma C, Fawzi W, Essex M and Renjifo B. A new human immunodeficiency virus type 1 circulating recombinant form from Tanzania. *AIDS Res. Hum. Retroviruses* 2001; 17; 423-431.

Kwa D, Vingerhoed J, Boeser B and Schuitemaker H. Increased in vitro cytopathicity of CC chemokine receptor 5-restricted human immunodeficiency virus type 1 primary isolates correlates with a progressive clinical course of infection. *J. Infect. Dis.* 2003; 187; 1397-1403.

Kwong PD, Doyle ML, Casper DJ, Cicala C, Leavitt SA, Majeed S, Steenbeke TD, Venturi M, Chaiken I, Fung M, Katinger H, Parren PW, Robinson J, Van Ryk D, Wang L, Burton DR, Freire E, Wyatt R, Sodroski J, Hendrickson WA and Arthos J. HIV-1 evades antibody-mediated neutralization through conformational masking of receptor-binding sites. *Nature.* 2002; 420; 678-682.

Land A and Braakman I. Folding of the human immunodeficiency virus type 1 envelope glycoprotein in the endoplasmic reticulum. *Biochimie.* 2001; 83; 783-790.

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

Le Roux IM and Le Roux PJ. Survey of the health and nutrition status of a squatter community in Khayelitsha. *S. Afr. Med. J.* 1991; 79; 500-503.

Leitner T. Genetic subtypes of HIV-1. In: *Human Retroviruses and AIDS 1996:* Myers GB, Korber B, Foley B, Jeang KT, Mellors JW and Wain-Hobson S. A compilation and analysis of

nucleic acid and amino acid sequences. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, New Mexico, USA. 1996a.

Leitner T, Escanilla D, Franzen C, Uhlen M, Albert J. Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. *Proc. Natl. Acad. Sci. USA.* 1996b: 93; 10864-10869.

Leitner T, Kumar S and Albert J. Tempo and mode of nucleotide substitutions in gag and env gene fragments in human immunodeficiency virus type 1 populations with a known transmission history. *J. Virol.* 1997: 71; 4761-4770.

Leonard CK, Spellman MW, Riddle L, Harris RJ, Thomas JN and Gregory TJ. Assignment of intrachain disulfide bonds and characterization of potential glycosylation sites of the type 1 recombinant human immunodeficiency virus envelope glycoprotein (gp120) expressed in Chinese hamster ovary cells. *J. Biol. Chem.* 1990: 265; 10373-10382.

Levy JA. HIV and the Pathogenesis of AIDS. Second edition. ASM Press. Washington DC, USA. 1998.

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

Lindström A and Albert J. A simple and sensitive 'in-house' method for determining genotypic drug resistance in HIV-1. *J. Virol. Methods.* 2003: 107; 45-51.

Liu SL, Rodrigo AG, Shankarappa R, Learn GH, Hsu L, Davidov O, Zhao LP and Mullins JI. HIV quasispecies and resampling. *Science.* 1996: 73; 415-416.

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

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

Loxton AG. Characterisation of new full-length subtype D viruses from South Africa. M.Sc (Medical Sciences). University of Stellenbosch, South Africa. 2004.

Loxton AG, Treurnicht F, Laten A, van Rensburg EJ and Engelbrecht S. Sequence analysis of near full-length HIV type 1 subtype D primary strains isolated in Cape Town, South Africa, from 1984 to 1986. *AIDS Res. Hum. Retroviruses*. 2005: 21; 410-413.

Lue J, Hsu M, Yang D, Marx P, Chen Z and Cheng-Mayer C. Addition of a single gp120 glycan confers increased binding to dendritic cell-specific ICAM-3-grabbing nonintegrin and neutralization escape to human immunodeficiency virus type 1. *J. Virol*. 2002: 76; 10299-10306.

Ly TD, Laperche S and Courouce AM. Early detection of human immunodeficiency virus infection using third- and fourth-generation screening assays. *Eur. J. Clin. Microbiol. Infect. Dis*. 2001: 20; 104-110.

Ly TD, Laperche S, Brennan C, Vallari A, Ebel A, Hunt J, Martin L, Daghfal D, Schochetman G and Devare S. Evaluation of the sensitivity and specificity of six HIV combined p24 antigen and antibody assays. *J. Virol. Methods*. 2004: 122; 185-194.

Maas JJ, Gange SJ, Schuitemaker H, Coutinho RA, van Leeuwen R and Margolick JB. Strong association between failure of T cell homeostasis and the syncytium-inducing phenotype among HIV-1-infected men in the Amsterdam Cohort Study. *AIDS*. 2000: 14; 1155-1161.

Maddison DR, Swofford DL and Maddison WP. NEXUS: An extensible file format for systematic information. *Systematic Biology*. 1997: 46; 590-621.

Maljkovic I, Wilbe K, Solver E, Alaeus A and Leitner T. Limited transmission of drug-resistant HIV type 1 in 100 Swedish newly detected and drug-naive patients infected with subtypes A, B, C, D, G, U, and CRF01\_AE. *AIDS Res. Hum. Retroviruses*. 2003: 19; 989-997.

Mangeat B, Turelli P, Caron G, Friedli M, Perrin L and Trono D. Broad antiretroviral defence by human APOBEC3G through lethal editing of nascent reverse transcripts. *Nature*. 2003: 424; 99-103.

Marks AS and Downes GM. Informal sector shops and AIDS prevention. An exploratory social marketing investigation. *S. Afr. Med. J*. 1991: 79; 496-499.

Markus MB and Fincham JE. Helminths, HIV/AIDS and tuberculosis. *S. Afr. Med. J*. 2000: 90; 834- 836.

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

Mastro T, Kuanusont C, Dondero T and Wasi C. Why do HIV-1 subtypes segregate among persons with different risk behaviors in South Africa and Thailand? *AIDS*. 1997: 11; 113-116.

Masur H, Michelis MA, Wormser GP, Lewin S, Gold J, Tapper ML, Giron J, Lerner CW, Armstrong D, Setia U, Sender JA, Siebken RS, Nicholas P, Arlen Z, Maayan S, Ernst JA, Siegal FP and Cunningham-Rundles S. Opportunistic infection in previously healthy women. *Ann. Intern. Med.* 1982; 97; 533-539.

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

Maxam AM and Gilbert W. A new method for sequencing DNA. *Biotechnology.* 1992; 24; 99-103.

McAlpine L, Parry JV, Shanson D and Mortimer PP. False negative results in enzyme linked immunosorbent assays using synthetic HIV antigens. *J. Clin. Pathol.* 1995; 48; 490-493.

McCormack GP, Glynn JR, Crampin AC, Sibande F, Mulawa D, Bliss L, Broadbent P, Abarca K, Ponnighaus JM, Fine PE and Clewley JP. Early evolution of the human immunodeficiency virus type 1 subtype C epidemic in rural Malawi. *J. Virol.* 2002; 76; 12890-12899.

McKinney DM, Skvoretz R, Livingston BD, Wilson CC, Anders M, Chesnut RW, Sette A, Essex M, Novitsky V and Newman MJ. Recognition of variant HIV-1 epitopes from diverse viral subtypes by vaccine-induced CTL. *J. Immunol.* 2004; 173; 1941-1950.

Mellors JW, Munoz A, Giorgi JV, Margolick JB, Tassoni CJ, Gupta P, Kingsley LA, Todd JA, Saah AJ, Detels R, Phair JP and Rinaldo CR Jr. Plasma viral load and CD4+ lymphocytes as prognostic markers of HIV-1 infection. *Ann. Intern. Med.* 1997; 126; 946-954.

Mellquist JL, Kasturi L, Spitalnik SL and Shakin-Eshleman SH. The amino acid following an asn-X-Ser/Thr sequon is an important determinant of N-linked core glycosylation efficiency. *Biochemistry.* 1998; 37; 6833-6837.

Meunier JC, Fournillier A, Choukhi A, Cahour A, Cocquerel L, Dubuisson J, Wychowski C. Analysis of the glycosylation sites of hepatitis C virus (HCV) glycoprotein E1 and the influence of E1 glycans on the formation of the HCV glycoprotein complex. *J. Gen. Virol.* 1999; 80; 887-896.

Miller V. HIV drug resistance: overview of clinical data. *J. HIV Ther.* 2001; 6; 68-71.

Mochizuki N, Otsuka N, Matsuo K, Shiino T, Kojima A, Kurata T, Sakai K, Yamamoto N, Isomura S, Dhole TN, Takebe Y, Matsuda M and Tatsumi M. An infectious DNA clone of HIV type 1 subtype C. *AIDS Res. Hum. Retroviruses.* 1999; 15; 1321-1324.

Monno L, Punzi G, Scarabaggio T, La Gioia A, Brindicci G, Angarano G, Di Bari C. Sequence analysis of HIV-1 from patients living in Apulia. Unpublished. GenBank accession number AY371693.

Montano MA, Novitsky VA, Blackard JT, Cho NL, Katzenstein DA and Essex M. Divergent transcriptional regulation among expanding human immunodeficiency virus type 1 subtypes. *J. Virol.* 1997; 71; 8657-8665.

Montavon C, Toure-Kane C, Liegeois F, Mpoudi E, Bourgeois A, Vergne L, Perret JL, Boumah A, Saman E, Mboup S, Delaporte E and Peeters M. Most env and gag subtype A HIV-1 viruses circulating in West and West Central Africa are similar to the prototype AG recombinant virus IBNG. *J Acquir. Immune. Defic. Syndr.* 2000; 23; 363-374.

Moodley D, Smith TL, Van Rensburg EJ, Moodley J and Engelbrecht S. HIV type 1 V3 region subtyping in KwaZulu-Natal, a high-seroprevalence South African region. *AIDS Res. Hum. Retroviruses.* 1998; 14; 1015-1018.

Moore JP and Nara PL. The role of the V3 loop of gp120 in HIV infection. *AIDS.* 1991; 5; S21-S33.

Moore JP, Parren PW, Burton DR. Genetic subtypes, humoral immunity, and human immunodeficiency virus type 1 vaccine development. *J. Virol.* 2001; 75; 5721-5729.

Morris L and Williamson C. Host and Viral factors that impact on HIV-1 transmission and disease progression in South Africa. *S. Afr. Med. J.* 2001; 91; 212-215.

MSF "Médecins Sans Frontières" South Africa. Providing HIV services including antiretroviral therapy at primary health care clinics in resource-poor settings: The experience from Khayelitsha. Activity report. Infectious Disease Epidemiology Unit, School of Public Health and Family Medicine, University of Cape Town. 2003. ([www.msf.org](http://www.msf.org)).

Mullis KB and Faloona FA. Specific synthesis of DNA *in vitro* via a polymerase catalyzed chain reaction. *Methods Enzymol.* 1987; 155; 335-350.

Mullis KB, Faloona FA, Scharf S, Saiki R, Horn G and Erlich H. Specific enzymatic amplification of DNA *in vitro*: the polymerase chain reaction. *Cold Spring Harb. Symp. Quant. Biol.* 1986; 51; 263-273.

Murphy KM, Berg KD and Eshleman JR. Sequencing of Genomic DNA by Combined Amplification and Cycle Sequencing Reaction. *Clinical Chemistry.* 2005; 51; 35-39.



Murphy G, Belda FJ, Pau CP, Clewley JP and Parry JV. Discrimination of subtype B and non-subtype B strains of human immunodeficiency virus type 1 by serotyping: correlation with genotyping. *J. Clin. Microbiol.* 1999; 37; 1356-1360.

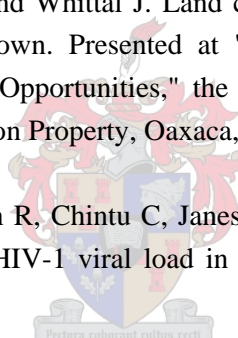
Murphy E, Korber B, Georges-Courbot MC, You B, Pinter A, Cook D, Kieny MP, Georges A, Mathiot C, Barre-Sinoussi F and Girard M. Diversity of V3 region sequences of human immunodeficiency viruses type 1 from the central African Republic. *AIDS Res. Hum. Retroviruses.* 1993; 9; 997-1006.

Murray AE, Lies D, Li G, Neelson K, Zhou J and Tiedje JM. DNA/DNA hybridization to microarrays reveals gene-specific differences between closely related microbial genomes. *Proc. Natl. Acad. Sci. USA.* 2001; 98; 9853-9858.

Muse SV and Gaut BS. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.* 1994; 11; 715-724.

Muzondo IF, Barry M, Dewar D and Whittal J. Land conflict resolution: A case study of the Khayelitsha settlement in Cape Town. Presented at "The Commons in an Age of Global Transition: Challenges, Risks and Opportunities," the Tenth Conference of the International Association for the Study of Common Property, Oaxaca, Mexico, August 9-13. 2004.

Mwaba P, Cassol S, Nunn A, Pilon R, Chintu C, Janes M and Zumla A. Whole blood versus plasma spots for measurement of HIV-1 viral load in HIV-infected African patients. *Lancet.* 2003; 362; 2067-2068.



Myer L, Mathews C and Little F. Improving the accessibility of condoms in South Africa: the role of informal distribution. *AIDS Care.* 2002; 14; 773-778.

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

Nahmias AJ, Weiss J Yao X, Lee F, Kodosi R, Schanfield M, Matthews T, Bolognesi D, Durack D, Motulsky A, Kanki P and Essex M. Evidence for human infection with an HTLV III/ LAV like virus in central Africa, 1959. *Lancet.* 1986; 1; 1279-1280.

Needleman SB and Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 1970; 48; 443-453.

Nei M and Kumar S. *Molecular evolution and phylogenetics.* Oxford University Press, Inc. New York, USA. 2000.

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

Ngui SL, Hallet R and Teo CG. Natural and iatrogenic variation in hepatitis B virus. *Rev. Med. Virol.* 1999; 9; 183-209.

Nguyen L, Hu DJ, Choopanya K, Vanichseni S, Kitayaporn D, van Griensven F, Mock PA, Kittikraisak W, Young NL, Mastro TD and Subbarao S. Genetic analysis of incident HIV-1 strains among injection drug users in Bangkok: evidence for multiple transmission clusters during a period of high incidence. *J. Acquir. Immune. Defic. Syndr.* 2002; 30; 248-256.

Nicholas KB and Nicholas JR. Genedoc, a tool for editing and annotating multiple sequence alignments. Distributed by the authors. 1997.

Nisole S and Saïb A. Early steps of retrovirus replicative cycle. *Retrovirology.* 2004; 1; 1-20.

Norris PJ, Sumaroka M, Brander C, Moffett HF, Boswell SL, Nguyen T, Sykulev Y, Walker BD and Rosenberg ES. Multiple effector functions mediated by human immunodeficiency virus-specific CD4(+) T-cell clones. *J. Virol.* 2001; 75; 9771-9779.

Novitsky V, Smith UR, Gilbert P, McLane MF, Chigwedere P, Williamson C, Ndung'u T, Klein I, Chang SY, Peter T, Thior I, Foley BT, Gaolekwe S, Rybak N, Gaseitsiwe S, Vannberg F, Marlink R, Lee TH and Essex M. Human immunodeficiency virus type 1 subtype C molecular phylogeny: consensus sequence for an AIDS vaccine design? *J. Virol.* 2002; 76; 5435-5451.

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

O'Brien WA, Koyanagi Y, Namazie A, Zhao JQ, Diagne A, Idler K, Zack Jaand Chen SY. HIV-1 tropism for mononuclear phagocytes can be determined by regions of gp120 outside the CD4-binding domain. *Nature.* 1990; 348; 69-73.

Oelrichs RB, Vandamme AM, Van Laethem K, Debyser Z, McCutchan FE and Deacon NJ. Full-length genomic sequence of an HIV type 1 subtype G from Kinshasa. *AIDS Res. Hum. Retroviruses* 1999; 15; 585-589.

Ogert RA, Lee MK, Ross W, Buckler-White A, Martin MA and Cho MW. N-linked glycosylation sites adjacent to and within the V1/V2 and the V3 loops of dualtropic human

immunodeficiency virus type 1 sequence DH12 gp120 affect coreceptor usage and cellular tropism. *J. Virol.* 2001; 75; 5998-6006.

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

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

Page RDM. TREEVIEW: An application to display phylogenetic trees on personal computers. *Computer Applications in the Biosciences.* 1996; 12; 357-358.

Page RDM and Holmes EC. *Molecular evolution: A phylogenetic approach.* Blackwell Science Ltd. Oxford, United Kingdom. 2002.

Palmiter RD. Magnesium precipitation of ribonucleoprotein complexes. Expedient techniques for the isolation of undergraded polysomes and messenger ribonucleic acid. *Biochemistry.* 1974; 13; 3606-3615.

Papathanasopoulos MA, Cilliers T, Morris L, Mokili JL, Dowling W, Birx DL and McCutchan FE. Full-length genome analysis of HIV-1 subtype C utilizing CXCR4 and intersubtype recombinants isolated in South Africa. *AIDS Res. Hum. Retroviruses.* 2002; 18; 879-886.

Papathanasopoulos MA, Patience T, Meyers TM, Morris L and McCutchan F. Full-length genome characterisation of HIV type 1 subtype C sequences from two slow-progressing perinatally infected siblings in South Africa. *AIDS Res. Hum. Retroviruses* 2003; 19; 1033-1037.

Parekh BS and McDougal JS. Application of laboratory methods for estimation of HIV-1 incidence. *Indian J. Med. Res.* 2005; 121; 510-518.

Paolucci S, Baldanti F, Campanini G, Zavattoni M, Cattaneo E, Dossena L, Gerna G. Analysis of HIV drug-resistant quasispecies in plasma, peripheral blood mononuclear cells and viral isolates from treatment-naive and HAART patients. *J. Med. Virol.* 2001; 65; 207-217.

Peeters M and Sharp PM. Genetic diversity of HIV-1: the moving target. *AIDS.* 2000; 14; S129-S140.

Penman S. RNA metabolism in the HeLa cell nucleus. *J. Mol. Biol.* 1966; 17; 117-130.

Penny D, Hendy MD and Steel MA. Progress with methods for constructing evolutionary trees. *Trends. Ecol. Evol.* 1992; 7; 73-79.

Pettifor AE, Measham DM, Rees HV and Padian NS. Sexual power and HIV risk, South Africa. *Emerg. Infect. Dis.* 2004; 10; 1996-2004.

Pieniasek D, Baggs J, Hu DJ, Matar GM, Abdelnoor AM, Mokhbat JE, Uwaydah M, Bizri AR, Ramos A, Janini LM, Tanuri A, Fridlund C, Schable C, Heyndrickx L, Rayfield MA and Heneine W. Introduction of HIV-2 and multiple HIV-1 subtypes to Lebanon. *Emerg. Infect. Dis.* 1998; 4; 649-656.

Pierson T, McArthur J and Siliciano RF. Reservoirs for HIV-1: Mechanisms for Viral Persistence in the Presence of Antiviral Immune Responses and Antiretroviral Therapy. *Annu. Rev. Immunol.* 2000; 18; 665-708.

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

Piyasirisilp S, McCutchan FE, Carr JK, Sanders-Buell E, Liu W, Chen J, Wagner R, Wolf H, Shao Y, Lai S, Beyrer C and Yu XF. A recent outbreak of human immunodeficiency virus type 1 infection in southern China was initiated by two highly homogeneous, geographically separated strains, circulating recombinant form AE and a novel BC recombinant. *J. Virol.* 2000; 74; 11286-11295.

Plantier JC, Le Pogam S, Poisson F, Buzelay L, Lejeune B and Barin F. Extent of antigenic diversity in the V3 region of the surface glycoprotein, gp120, of human immunodeficiency virus type 1 group M and consequences for serotyping. *J. Virol.* 1998; 72; 677-683.

Poignard P, Saphire EO, Parren PW and Burton DR. gp120: Biologic aspects of structural features. *Annu. Rev. Immunol.* 2001; 19; 253-274.

Pollakis G, Abebe A, Kliphuis A, Chalaby MI, Bakker M, Mengistu Y, Brouwer M, Goudsmit J, Schuitemaker H and Paxton WA. Phenotypic and genotypic comparisons of CCR5- and CXCR4-tropic human immunodeficiency virus type 1 biological clones isolated from subtype C-infected individuals. *J. Virol.* 2004; 78; 2841-2852.

Pollakis G, Abebe A, Kliphuis A, De Wit TF, Fisseha B, Tegbaru B, Tesfaye G, Negassa H, Mengistu Y, Fontanet AL, Cornelissen M and Goudsmit J. Recombination of HIV type 1C (C'/C'') in Ethiopia: possible link of EthHIV-1C' to subtype C sequences from the high-prevalence epidemics in India and Southern Africa. *AIDS Res. Hum. Retroviruses.* 2003; 19; 999-1008.

Pollakis G, Kang S, Kliphuis A, Chalaby MI, Goudsmit J and Paxton WA. 2001 N-linked glycosylation of the HIV type-1 gp120 envelope glycoprotein as a major determinant of CCR5 and CXCR4 coreceptor utilization. *J. Biol. Chem.* 2001; 276; 13433-13441.

Posada D and Buckley TR. Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Syst. Biol.* 2004; 53; 793-808.

Posada D and Crandall KA. Modeltest: testing the model of DNA substitution. *Bioinformatics.* 1998; 14; 817-818.

Posada D and Crandall KA. Selecting the best-fit model of nucleotide substitution. *Syst. Biol.* 2001; 50; 580-601.

Poss M, Rodrigo AG, Gosink JJ, Learn GH, de Vange Panteleeff D, Martin HL Jr, Bwayo J, Kreiss JK and Overbaugh J. Evolution of envelope sequences from the genital tract and peripheral blood of women infected with clade A human immunodeficiency virus type 1. *J. Virol.* 1998; 72; 8240-8251.

Preble E and Piwoz E. Prevention of mother-to-child transmission of HIV in Africa: practical guidance for programs. SARA Project, Academy for Educational Development. Washington DC, USA. 2001.

Preiser W, Brink NS, Hayman A, Waite J, Balfe P and Tedder RS. False-negative HIV antibody test results. *J. Med. Virol.* 2000; 60; 43-47.

Rannala B and Yang Z. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *J. Mol. Evol.* 1996; 43; 304-311.

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

Ratner L. Glucosidase inhibitors for treatment of HIV-1 infection. *AIDS. Res. Hum. Retroviruses.* 1992; 8; 165-173.

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

Ratner L, Haseltine W, Patarca R, Livak KJ, Starcich B, Josephs SF, Doran ER, Rafalski A, Whitehorn EA, Baumeister K, Ivanoff L, Petteway SR, Pearson ML, Lauenberger JA, Papas TS, Ghayeb J, Chang NT, Callo RC and Wong-Staal F. Complete nucleotide sequence of the AIDS virus, HTLV-III. *Nature.* 1985b; 313; 277-284.

Regoes RR and Bonhoeffer S. The HIV coreceptor switch: a population dynamical perspective. *Trends. Microbiol.* 2005; 13; 269-277.

Renjifo B, Chaplin B, Mwakagile D, Shah P, Vannberg F, Msamana G, Hunter D, Fawzi W and Essex M. Epidemic expansion of HIV type 1 subtype C and recombinant genotype in Tanzania. *AIDS Res. Hum. Retroviruses*. 1998; 4; 635-638.

Robertson DL, Anderson JP, Bradac JA, Carr JK, Foley B, Funkhouser RK, Gao F, Hahn BH, Kalish ML, Kuiken C, Learn GH, Leitner T, McCutchan F, Osmanov S, Peeters M, Pieniazek D, Salminen M, Sharp PM, Wolinsky S and Korber B. HIV-1 nomenclature proposal. *Science*. 2000; 288; 55-57.

Robertson DL, Sharp PM, McCutchan FE and Hahn BH. Recombination in HIV-1. *Nature*. 1995; 374; 124-126.

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

Rodrigo AG and Learn GH. Computational and evolutionary analysis of HIV molecular sequences. Kluwer Academic Publishers, Boston, USA. 2001.

Rose PP and Korber BT. Detecting hypermutations in viral sequences with an emphasis on G --> A hypermutation. *Bioinformatics*. 2000; 16; 400-401.

Rosen S, Vincent JR, MacLeod W, Fox M, Thea DM and Simon JL. The cost of HIV/AIDS to businesses in southern Africa. *AIDS*. 2004; 23; 317-324.

Roshal M, Zhu Y and Planelles V. Apoptosis in AIDS. *Apoptosis*. 2001; 6; 103-116.

Sagar M, Kirkegaard E, Long EM, Celum C, Buchbinder S, Daar ES and Overbaugh J. Human immunodeficiency virus type 1 (HIV-1) diversity at time of infection is not restricted to certain risk groups or specific HIV-1 subtypes. *J. Virol*. 2004; 78; 7279-7283.

Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, Mullis KB and Erlich HA. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*. 1988; 239; 487-491.

Saitou N and Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol*. 1987; 4; 406-425.

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

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

Salminen MO, Johansson B, Sonnerborg A, Ayehunie S, Gotte D, Leinikki P, Burke DS and McCutchan FE. Full-length sequence of an Ethiopian human immunodeficiency virus type 1 (HIV-1) isolate of genetic subtype C. *AIDS Res. Hum. Retroviruses*. 1996; 12; 1329-1339.

Sambrook J, Fritsch EF and Maniatis T. *Molecular cloning: A laboratory Manual* second edition. Cold Spring Harbor Laboratory Press. New York, USA. 1989.

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

Sanger F, Nicklen S and Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA*. 1977; 74; 5463-5467.

Santiago ML, Rodenburg CM, Kamenya S, Bibollet-Ruche F, Gao F, Bailes E, Meleth S, Soong S, Kilby JM, Moldoveanu Z, Fahey B, Muller MN, Ayouba A, Nerrienet E, McClure HM, Heeney JL, Pusey AE, Collins DA, Boesch C, Wrangham RW, Goodall J, Sharp PM, Shaw GM and Hahn BH. SIVcpz in wild chimpanzees. *Science*. 2002; 295; 465.

Schmidt HA, Strimmer K, Vingron M and von Haeseler A. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*. 2002; 18; 502-504.

Scholz I, Arvidson B, Huseby D and Barklis E. Virus particle core defects caused by mutations in the human immunodeficiency virus capsid N-terminal domain. *J. Virol*. 2005; 79; 1470-1479.  
Sepkowitz KA. AIDS - the first 20 years. *N. Engl. J. Med*. 2001; 344; 1764-1772.

Servais J, Lambert C, Karita E, Vanhove D, Fischer A, Baurith T, Schmit JC, Schneider F, Hemmer R and Arendt V. HIV type 1 pol gene diversity and archived nevirapine resistance mutation in pregnant women in Rwanda. *AIDS. Res. Hum. Retroviruses*. 2004; 20: 279-283.

Shankarappa R, Chatterjee R, Learn GH, Neogi D, Ding M, Roy P, Ghosh A, Kingsley L, Harrison L, Mullins JI and Gupta P. Human immunodeficiency virus type 1 env sequences from Calcutta in eastern India: identification of features that distinguish subtype C sequences in India from other subtype C sequences. *J. Virol*. 2001; 75; 10479-10487.

Sharp PA, Sugden B and Sambrook J. Detection of two restriction endonuclease activities in *Haemophilus parainfluenzae* using analytical agarose - ethidium bromide electrophoresis. *Biochemistry*. 1973; 12; 3055-3063.

Sher R. HIV infection in South Africa, 1982-1988 - a review. *S. Afr. Med. J.* 1989. 76; 314-318.

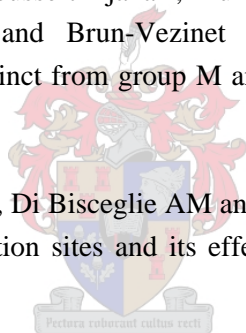
Shilts RM. *And the band played on: Politics, people and the AIDS epidemic*. St. Martin's Press. New York, USA. 1987.

Si Z, Cayabyab M and Sodroski J. Envelope glycoprotein determinants of neutralization resistance in a simian-human immunodeficiency virus (SHIV-HXBc2P 3.2) derived by passage in monkeys. *J. Virol.* 2001; 75; 4208-4218.

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

Simon F, Mauclore P, Roques P, Loussert-Ajaka I, Muller-Trutwin MC, Saragosti S, Georges-Courbot MC, Barre-Sinoussi F and Brun-Vezinet F. Identification of a new human immunodeficiency virus type 1 distinct from group M and group O. *Nat. Med.* 1998; 4: 1032-1037.

Slater-Handshy T, Droll DA, Fan X, Di Bisceglie AM and Chambers TJ. HCV E2 glycoprotein: mutagenesis of N-linked glycosylation sites and its effects on E2 expression and processing. *Virology*. 2004; 319; 36-48.



Smith M, Geretti AM, Osner N, Easterbrook P and Zuckerman M. High levels of discordance between sequencing and serological subtyping in a predominantly non-B subtype HIV-1 infected cohort. *J. Clin. Virol.* 2005; 33; 312-318.

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

Snoeck J, Van Dooren S, Van Laethem K, Derdelinckx I, Van Wijng E, De Clercq E, Vandamme AM. Prevalence and origin of HIV-1 group M subtypes among patients attending a Belgian hospital in 1999. *Virus Res.* 2002; 85; 95-107.

Soares MA, de Oliveira T, Brindeiro RM, Diaz RS, Sabino EC, Brigido L, Pires IL, Morgado MC, Dantas MC, Barreira D, Teixeira PR, Cassol S and Tanuri A. A specific subtype C of human immunodeficiency virus type 1 circulates in Brazil. *AIDS*. 2003; 17; 11-21.

Sonigo P, Alizon M, Staskus K, Klatzmann D, Cole S, Danos O, Retzel E, Tiollais P, Haase A and Wain-Hobson S. Nucleotide sequence of the Visna Lentivirus: relationship to the AIDS virus. *Cell*. 1985; 42; 369-382.



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

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

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

Starcich BR, Hahn BH and Shaw GM. Identification and characterization of conserved and variable regions in the envelope gene of HTLV-III/LAV, the retrovirus of AIDS. *Cell*. 1986; 45; 637-648.

Stein MS, Wang B, Dwyer DE, Saksena NK. HIV-1 coinfection, superinfection and recombination. *Sexual Health*. 2004; 1; 239-250.

Stebbing J, Gazzard B and Douek DC. Where does HIV live? *N. Engl. J. Med*. 2004; 350; 1872-1880.

Stoneburner RL and Low-Beer D. Population-level HIV declines and behavioral risk avoidance in Uganda. *Science*. 2004; 304; 714-718.

Su L, Graf M, Zhang Y, von Briesen H, Xing H, Kostler J, Melzl H, Wolf H, Shao Y and Wagner R. Characterization of a virtually full-length human immunodeficiency virus type 1 genome of a prevalent intersubtype (C/B') recombinant strain in China. *J. Virol*. 2000; 74; 11367-11376.

Swanson P, Devare SG and Hackett JR. Molecular characterization of 39 HIV-1 isolates representing group M (subtypes A-G) and group O: Sequence analysis of gag p24, pol Integrase and env gp41. *AIDS Res. Hum. Retroviruses*. 2003; 19; 625-629.

Swanson P, de Mendoza C, Joshi Y, Golden A, Hodinka RL, Soriano V, Devare SG and Hackett J Jr. Impact of human immunodeficiency virus type 1 (HIV-1) genetic diversity on performance of four commercial viral load assays: LCx HIV RNA Quantitative, AMPLICOR HIV-1 MONITOR v1.5, VERSANT HIV-1 RNA 3.0, and NucliSens HIV-1 QT. *J. Clin. Microbiol*. 2005; 43; 3860-3868.

Swerdlow H and Gesteland R. Capillary gel electrophoresis for rapid, high resolution DNA sequencing. *Nucleic Acids Research*. 1990; 18; 1415-1419.

TAC. The treatment action campaign (TAC) of South Africa. 2005. ([www.tac.org.za](http://www.tac.org.za)).

Takebe Y, Motomura K, Tatsumi M, Lwin HH, Zaw M and Kusagawa S. High prevalence of diverse forms of HIV-1 intersubtype recombinants in Central Myanmar: Geographical hot spot of extensive recombination. *AIDS*. 2003; 17; 2077-2087.

Tamura K Nei M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* 1993; 10; 512-526.

Tavaré S. Some probabilistic and statistical problems in the analysis of DNA sequences. In: Miura RM. (Ed.). *Some Mathematical Questions in Biology/DNA Sequence Analysis*. American Mathematical Society, Providence, Rhode Island, USA. 1986.

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

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

Toure-Kane C, Montavon C, Faye MA, Gueye PM, Sow PS, Ndoeye I, Gaye-Diallo A, Delaporte E, Peeters M and Mboup S. Identification of all HIV type 1 group M subtypes in Senegal, a country with low and stable seroprevalence. *AIDS Res. Hum. Retroviruses*. 2000; 16; 603-609.

Tovanabutra S, Watanaveeradej V, Viputtikul K, De Souza M, Razak MH, Suriyanon V, Jittiwutikarn J, Sriplienchan S, Nitayaphan S, Benenson MW, Sirisopana N, Renzullo PO, Brown AE, Robb ML, Beyrer C, Celentano DD, McNeil JG, Birx DL, Carr JK and McCutchan FE. A new circulating recombinant form, CRF15\_01B, reinforces the linkage between IDU and heterosexual epidemics in Thailand. *AIDS Res. Hum. Retroviruses*. 2003; 19; 561-567.

Travers SA, Clewley JP, Glynn JR, Fine PE, Crampin AC, Sibande F, Mulawa D, McInerney JO and McCormack GP. Timing and reconstruction of the most recent common ancestor of the subtype C clade of human immunodeficiency virus type 1. *J. Virol.* 2004; 78; 10501-10506.

Treurnicht FK, Smith T.-L, Engelbrecht S, Claassen M, Robson BA, Zeier M and van Rensburg EJ. Genotypic and phenotypic analyses of the env genes from South African HIV-1 subtype B and C isolates. *J. Med. Virology*. 2002; 68; 141-146.

Triques K, Bourgeois A, Saragosti S, Vidal N, Mpoudi-Ngole E, Nzilambi N, Apetrei C, Ekwalinga M, Bibollet-Ruche F, Peeters M. High diversity of HIV-1 subtype F strains in Central Africa. *Virology*. 1999; 259; 99-109.

Trkola A, Paxton WA, Monard SP, Hoxie JA, Siani MA, Thompson DA, Wu L, Mackay CR, Horuk R, Moore JP. Genetic subtype-independent inhibition of human immunodeficiency virus type 1 replication by CC and CXC chemokines. *J. Virol.* 1998; 72; 396-404.

Turner BG and Summers MF. Structural biology of HIV. *J. Mol. Biol.* 1999; 8; 1-32.

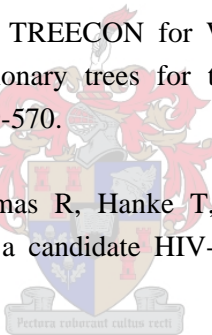
UNAIDS/WHO working group on global HIV/AIDS and STD surveillance: AIDS epidemic update. December 2004 ([www.unaids.org](http://www.unaids.org)).

Van Binsbergen J, Keur W, Siebelink A, van de Graaf M, Jacobs A, de Rijk D, Nijholt L, Toonen J and Gurtler LG. Strongly enhanced sensitivity of a direct anti-HIV-1/-2 assay in seroconversion by incorporation of HIV p24 ag detection: a new generation vironostika HIV Uni-Form II. *J. Virol. Methods.* 1998; 76; 59-71.

Van Binsbergen J, Siebelink A, Jacobs A, Keur W, Bruynis F, van de Graaf M, van der Heijden J, Kambel D and Toonen J. Improved performance of seroconversion with a 4th generation HIV antigen/antibody assay. *J. Virol. Methods.* 1999; 82; 77-84.

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

Van Harmelen J, Shephard E, Thomas R, Hanke T, Williamson AL and Williamson C. Construction and characterisation of a candidate HIV-1 subtype C DNA vaccine for South Africa. *Vaccine.* 2003; 21; 4380-4389.



Van Harmelen J, Van der Ryst E, Loubser AS, York D, Madurai S, Lyons S, Wood R and Williamson C. A Predominantly HIV Type 1 Subtype C-Restricted Epidemic in South African Urban Populations. *AIDS Res. Hum. Retroviruses.* 1999a; 15; 395-398.

Van Harmelen J, Van der Ryst E, Wood R, Lyons SF and Williamson C. Restriction fragment length polymorphism analysis for rapid gag subtype determination of human immunodeficiency virus type 1 in South Africa. *J. Virol. Methods.* 1999b; 78; 51-59.

Van Harmelen J, Williamson C, Kim B, Morris L, Carr J, Karim SS and McCutchan F. Characterisation of full-length HIV type 1 subtype C sequences from South Africa. *AIDS Res. Hum. Retroviruses* 2001; 17; 1527-1531.

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

Venturini S, Mosier DE, Burton DR and Poignard P. Characterization of human immunodeficiency virus type 1 (HIV-1) Gag- and Gag peptide-specific CD4(+) T-cell clones

from an HIV-1-seronegative donor following in vitro immunization. *J Virol.* 2002: 76; 6987-6999.

Vidal N, Koyalta D, Richard V, Lechiche C, Ndinaromtan T, Djimasngar A, Delaporte E and Peeters M. High genetic diversity of HIV-1 strains in Chad, West Central Africa. *J. Acquir. Immune. Defic. Syndr.* 2003: 33; 239-246.

Vidal N, Mulanga-Kabeya C, Nzilambi N, Delaporte E, Peeters M. Identification of a complex env subtype E HIV type 1 virus from the democratic republic of congo, recombinant with A, G, H, J, K, and unknown subtypes. *AIDS. Res. Hum. Retroviruses.* 2000a: 16; 2059-2064.

Vidal N, Peeters M, Mulanga-Kabeya G, Nzilambi N, Robertson D, Ilunga W, Sema H, Tshimanga K, Bongo B and Delaporte E. Unprecedented degree of human immunodeficiency virus type 1 (HIV-1) group M genetic diversity in the Democratic Republic of the Congo suggest that the HIV-1 pandemic originated in Central Africa. *J. Virol.* 2000b: 74; 10498-10507

Vodicka MA, Goh WC, Wu LI, Rogel ME, Bartz SR, Schweickart VL, Raport CJ and Emerman M. Indicator cell lines for detection of primary strains of human and simian immunodeficiency viruses. *Virology.* 1997: 233; 193-198.

Vogelstein B and Gillespie D. Preparative and analytical purification of DNA from agarose. *Proc. Natl. Acad. Sci. USA.* 1979: 76; 615-619.

Wain-Hobson S. Virological mayhem. *Nature.* 1995: 373; 102.

Wain-Hobson S, Sonigo P, Guyader M, Gazit A and Henry M. Erratic G→A hypermutation within a complete caprine arthritis-encephalitis virus (CAEV) provirus. *Virology.* 1995: 209; 297-303.

Wei X, Decker JM, Wang S, Hui H, Kappes JC, Wu X, Salazar-Gonzalez JF, Salazar MG, Kilby JM, Saag MS, Komarova NL, Nowak MA, Hahn BH, Kwong PD and Shaw GM. Antibody neutralization and escape by HIV-1. *Nature.* 2003: 422; 307-312.

Weiss RA and Wrangham RW. From pandemic to pandemic. *Nature.* 1999: 397; 385-386.

Werle E, Schneider C, Renner M, Volker M and Fiehn W. Convenient single-step, one tube purification of PCR products for direct sequencing. *Nucleic Acids Res.* 1994: 11; 4354-4355.

WHO. World Health Organization. Antiretroviral therapy in primary health care: experience of the Khayelitsha programme in South Africa, case study. Geneva. WHO. 2004.

Williamson C, Morris L, Maughan MF, Ping LH, Dryga SA, Thomas R, Reap EA, Cilliers T, van Harmelen J, Pascual A, Ramjee G, Gray G, Johnston R, Karim SA and Swanstrom R.

Characterization and selection of HIV-1 subtype C isolates for use in vaccine development. *AIDS Res. Hum. Retroviruses*. 2003; 19; 133-144.

Wilson CC, McKinney D, Anders M, MaWhinney S, Forster J, Crimi C, Southwood S, Sette A, Chesnut R, Newman MJ and Livingston BD. Development of a DNA vaccine designed to induce cytotoxic T lymphocyte responses to multiple conserved epitopes in HIV-1. *J. Immunol*. 2003; 171; 5611-5623.

Wood K and Jewkes R. Violence, rape, and sexual coercion: everyday love in a South African township. *Gend. Dev*. 1997; 5; 41-46.

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

Wyatt R, Kwong PD, Desjardins E, Sweet RW, Robinson J, Hendrickson WA and Sodroski JG. The antigenic structure of the HIV gp120 envelope glycoprotein. *Nature*. 1998; 393; 705-711.

Yamaguchi J, Bodelle P, Kaptue L, Zekeng L, Gurtler LG, Devare SG, and Brennan CA. Near full-length genomes of 15 HIV type 1 group O isolates. *AIDS Res. Hum. Retroviruses*. 2003; 19: 979-988.

Yan J, Feng J, Hosono S and Sommer SS. Assessment of multiple displacement amplification in molecular epidemiology. *Biotechniques*. 2004; ;37; 136-143.

Yang Z, Goldman N, Friday A. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Mol. Biol. Evol*. 1994; 11; 316-324.

Ye Y, Si ZH, Moore JP and Sodroski J. Association of structural changes in the V2 and V3 loops of the gp120 envelope glycoprotein with acquisition of neutralization resistance in a simian-human immunodeficiency virus passaged in vivo. *J. Virol*. 2000; 74; 11955-11962.

Yoo S, Myszka DG, Yeh C, McMurray M, Hill CP and Sundquist WI. Molecular recognition in the HIV-1 capsid/cyclophilin A complex. *J. Mol. Biol*. 1997; 269; 780-795.

Yu XF, Chen J, Shao Y, Beyrer C and Lai S. Two subtypes of HIV-1 among injection-drug users in southern China. *Lancet*. 1998; 351; 1250.

Zacharova V, Becker ML, Zachar V, Ebbesen P and Goustin AS. DNA sequence analysis of the long terminal repeat of the C subtype of human immunodeficiency virus type 1 from Southern Africa reveals a dichotomy between B subtype and African subtypes on the basis of upstream NF-IL6 motif *AIDS Res. Hum. Retroviruses*. 1997; 13; 719-724.

Zanchetta N, Nardi G, Tocalli L, Drago L, Bossi C, Pulvirenti FR, Galli C and Gismondo MR. Evaluation of the abbot LCx HIV-1 RNA quantitative, a new assay for quantitative determination of human immunodeficiency virus type 1 RNA. *J. Clin. Microbiol.* 2000; 38; 3882-3886.

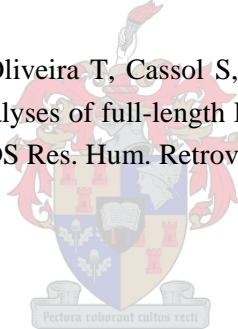
Zhang M, Gaschen B, Blay W, Foley B, Haigwood N, Kuiken C and Korber B. Tracking global patterns of N-linked glycosylation site variation in highly variable viral glycoproteins: HIV, SIV, and HCV envelopes and influenza hemagglutinin. *Glycobiology.* 2004; 14; 1229-1246.

Zharkikh A. Estimation of evolutionary distances between nucleotide sequences. *J. Mol. Evol.* 1994; 39; 315-329.

Zhu T, Korber BT, Nahmias AJ, Hooper E, Sharp PM and Ho DD. An African HIV-1 sequence from 1959 and implications for the origin of the epidemic. *Nature.* 1998; 594-597.

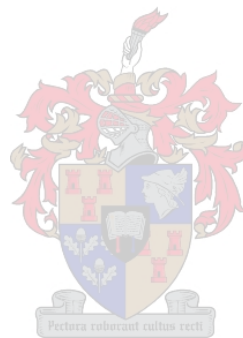
Zolla-Pazner S, Zhong P, Revesz K, Volsky B, Williams C, Nyambi P and Gorny MK. The cross-clade neutralizing activity of a human monoclonal antibody is determined by the GPGR V3 motif of HIV type 1. *AIDS Res. Hum. Retroviruses.* 2004; 20; 1254-1258.

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



## APPENDIX

		<b>PAGE</b>
Appendix A	Ethical approval	158
Appendix B	Amino acid sequence alignments	160



## APPENDIX A

### Ethical approval

#### PAGE

A1: Letter confirming ethical approval

159

This study has been ethically approved by the Committee for Human Research at the University of Stellenbosch. The project number is N04/06/100. A letter (in Afrikaans) confirming the approval has been included.





## A1: Letter confirming ethical approval



UNIVERSITEIT • STELLENBOSCH • UNIVERSITY  
jou kennisvenoot • your knowledge partner

1 September 2004

Prof S Engelbrecht  
Departement Geneeskundige Virologie

Geagte prof Engelbrecht

**NAVORSINGSPROJEK:** "INVESTIGATION OF THE MOLECULAR EPIDEMIOLOGY  
OF HIV-1 IN KHAYELITSHA, CAPE TOWN, USING  
GENOTYPING TECHNIQUES"

**PROJEKNOMMER** : N04/06/100

Dit is vir my aangenaam om u mee te deel dat die Komitee vir Mensnavorsing op sy vergadering van 4 Augustus 2004 bogenoemde projek goedgekeur het, ook wat die etiese aspekte daarvan betref.

Die projek is nou geregistreer en u kan voortgaan met die werk. U moet asseblief in verdere korrespondensie na bogenoemde projeknommer verwys.

Ek vestig graag u aandag daarop dat pasiënte wat deelneem aan 'n navorsingsprojek in Tygerberg Akademiese Hospitaal nie gratis behandeling sal ontvang nie aangesien die PGWK nie navorsing finansiële ondersteun nie.

Die verpleegkorps van die Tygerberg Akademiese Hospitaal kan ook nie omvattende verpleeghulp met navorsingsprojekte lewer nie weens die swaar werkslading waaronder hulle reeds gebuk gaan. Dit kan dus van 'n navorser verwag word om in sulke gevalle privaat verpleegkundiges te verkry.

Met vriendelike groete

**CJ VAN TONDER**  
**NAVORSINGSONTWIKKELING EN -STEUN (TYGERBERG)**

CJVT/ev



C:\DOCUMENTS AND SETTINGS\REVISAGIE\000\MY DOCUMENTS\SKM\PROJEKTE\2004\N04-06-100-001.DOC

Fakulteit Gesondheidswetenskappe • Faculty of Health Sciences



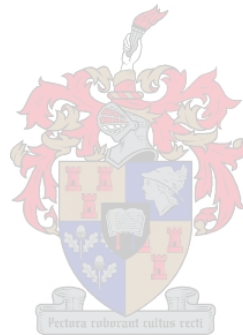
Verbind tot Optimale Gesondheid • Committed to Optimal Health  
Afdeling Navorsingsontwikkeling en -steun • Division of Research Development and Support  
Posbus/PO Box 19063 • Tygerberg 7505 • Suid-Afrika/South Africa  
Tel: +27 21 938 9207 • Faks/Fax: +27 21 933 6330  
E-pos/E-mail: cjvt@sun.ac.za

## APPENDIX B

### Amino acid sequence alignments

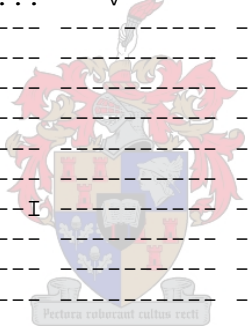
	<b>PAGE</b>
B1 <i>gag</i> p24 amino acid sequence alignment	161
B2 <i>env</i> gp41 IDR amino acid sequence alignment	168
B3 <i>env</i> gp120 V3 amino acid sequence alignment	175
B4 <i>pol</i> amino acid sequence alignment	182

The *gag* p24, *env* gp41 IDR, *env* gp120 V3 and *pol* amino acid sequence alignments compared to HXB2 and the four HIV-1 subtype C reference strains are presented here. A stripe indicates place holders where the amino acids are the same relative to HXB2, while a dot indicates places where no sequence data was available.



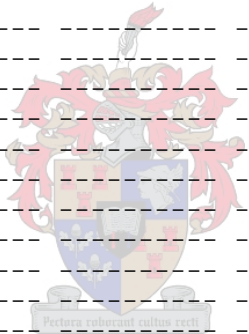
**B1: gag p24 amino acid sequence alignment**

	KVVEEKAFSP	EVIPMFSALS	EGATPQDLNT	MLNTVGGHQA	AMQMLKETIN	EEAAEWDRVH	PVHAGPIAPG	QMREP
B.FR.83.HXB2								
C.BW.96.96BW0502	--I-----	-----T--	-----	-----	-----D--	-----L-	--Q--V--	---D-
C.IN.95.95IN21068	--I-----	-----T--	-----	-----	-----D--	-----L-	--P-----	-L---
C.ET.86.ETH2220	-----	-----T--	-----	-----	-----D--	-----L-	-----V--	---D-
C.BR.92.92BR025	-----	-----T--	-----	-----	-----D--	-----L-	-----V--	-----
1001	--I-----	-----T--	-----	-----	-----D--	-----L-	--Q--A--	-L---
1002	--E-----	-----T--	-----	-----	-----D--	-----L-	-----V--	-I---
1003	--I-----	-A--T--	-----	-----	-----D--	-----L-	--Q--L-H-	-----
1005	-----	-----T--	-----	-----	-----D--	-----L-	--Q-----	-I---
1006	--I-----	-----T--	-----	-----	-----D--	D-----L-	-----	-----
1008	.....	.....	.....	--V-----	-----D--	---D--L-	-----	-I---
1009	--I-----	-----T--	-----	-----	-----D--	-----L-	-----	-----
1010	--I-----	-----T--	-----	-----	-----D--	-----L-	--Q-----	-----
1011	-----	-----I--	-----	-----	I--D--	-----L-	-----V--	-----
1012	--R-----	-----R--	-----	-----	D--D--	-----L-	-----V--	-----
1013	--I-----	-----T--	-----	-----	-----D--	-----L-	-----V--	-----
1015	--I-----	-----T--	-----	-----	-----D--	-----L-	-----V--	-----
1016	---G-N-	-----A--	---S--I--	-----	-----D--	-----L-	-----V--	-----
1017	---D-----	-----T--	-----	-----	-----D--	-----L-	-----Q--	-----
1018	--I-----	-----T--	-----	-----	-----D--	-----L-	--Q-----	-----
1019	--I-----	-----T--	-E-----	-----	-----D--	-----L-	-----N	-I---
1021	--I-----	-----T--	-----	-----	-----D--	-----L-	--Q-----	-I-D-
1023	-----	-I--T--	-----	-----	-----D--	-----L-	-AQ--V--	-I-D-
1024	--ID-E--	-----T--	-----	-----	-----D--	-----L-	-AQ--FPA-	-I---
1026	--I--S--	-I--T--	-----	-----	-----D--	-----L-	--Q--YPA-	-----
1029	--I-----	-----T--	-----	-----	-----D--	-----L-	-----	-----
1031	--NK-----	-----T--	-----	-----	-----D--	-----L-	-----	P----
1033	--I--G--	-----T--	-----	-----	-----D--	---D--L-	-----V--	-----
1034	---G-N-	-----	-----	---D-----	-----D--	D-----L-	-----V--	-----
1037	--I-----	-----T--	-----	-----	-----D--	-----L-	-----	-----
1038	--I-----	-----T--	---S--G--	---I-----	-----D--	-----L-	--Q-----	-I---
1039	--I-----	-----T--	---S--G--	---I-----	-----D--	-----M-	-----	-----
1040	--I-----	-----T--	-----	-----	-----D--	-----L-	--Q--V--	-I-D-
1042	--I--G-N-	-----T--	---S--C--	---I-----	-----D--	---D-----	-----V--	-I---



**B1 continue:** gag p24 amino acid sequence alignment

B.FR.83.HXB2	KVVEEKAFSP	EVIPMFSALS	EGATPQDLNT	MLNTVGGHQA	AMQMLKETIN	EEAAEWDVRVH	PVHAGPIAPG	QMREP
1043	--I-----	-----T--	-----	-----	-----D--	-----L-	-----NP--	-----
1044	--IG-----	-----T--	-----	-----	-----D--	-----L-	--Q-----	-----
1045	--I-----	-----T--	-----	-----	-----D--	D-----L-	-----NP--	-----
1047	--I-----	-----T--	-----	-----	-----D--	-----L-	-----V--	-----
1048	--I-----	-----T--	-----	-----	-----D--	-----L-	-----	-I----
1049	.....	.....	.....-	-----D--	--L--D--	-----L-	-----	-----
1050	--I--G-N-	-----T--	-----S--S	--I-----	-----D--	---D--L-	-----V--	-I----
1052	--IGG----	-A---T--	-----	-----	-----	---I---L-	-----	-----
1054	--IGD-G--	-A---T--	-----	-----	-----D--	-----L-	--Q--V--	-I----
1055	--I-----	-----T--	-----	-----	-----D--	-----L-	--Q-----	-I----
1056	--I-----	-----T--	-----	-----	-----D--	-----L-	--Q-----	-I----
1057	.....	.....	..--K--	-----	-----D--	-----L-	-----V--	-----
1058	--I-----	-----	-----	-----	-----D--	-----L-	-----	-----
1059	--I-----	-----T--	-----	-----	-----D--	-----L-	-----Q--	-----
1060	--I-----	-----T--	-----	-----	-----D--	-----L-	--Q--V--	-I-D-
1061	--I-----	-----T--	-----	-----	-----D--	-----L-	--Q--V--	-I-D-
1062	--I-----	-----T--	-----	-----	-----D--	-----L-	-----NV--	-----
1063	--I-----	-----T--	-----	-----	-----D--	D-----L-	-----V--	-----
1064	--I-----	-I---T--	-----	-----	-----D--	-----L-	--Q--YPA-	-----
1067	--I-----	-----T--	-----	-----	-----D--	-----L-	--Q--V--	-I----
1068	--I-----	-----T--	-----	-----	--I--D--	-----L-	-----V--	-I----
1069	--I-----	-----T--	-----	-----	-----D--	-----L-	-----V--	-----
1072	--IG---P-	-----T--	-----	-----	-----D--	-----L-	-----	---D-
1073	---R-RV--	R---L--	-----	-F-----	-----D--	-----F-	-----V--	-----
1075	---G-----	---IMT-C-	-----L--Y-	--T-----	-----D--	-----L-	-----A--	-----
1076	-----G-	-----T--	-----	-----	-----D--	-----	-----V--	-----
1077	--I-----	-----T--	-----	-----	-----D--	D-----L-	--P--V--	---D-
1079	--I-----	-----T--	-----	-----	-----D--	-----L-	-----	---D-
1083	--I-----	-----T--	-----	-----	-----D--	-----L-	-----	-----
1084	-SNRGE---	-L-A-T--	-----	-----	-----D--	-----L-	-AQ---YPA-	-I----
1088	--I-----T-	-----T--	-----	-----	-----D--	-----L-	--Q-----	-I----
1089	--G-E---	-----A--	-----	-----	-----D--	-----L-	--Q-----	-I----
1090	--I-----	-----T--	-----	-----	-----D--	-----	-----	-I----



**B1 continue:** gag p24 amino acid sequence alignment

B.FR.83.HXB2	KVVEEKAFSP	EVIPMFSALS	EGATPQDLNT	MLNTVGGHQA	AMQMLKETIN	EEAAEWDVRH	PVHAGPIAPG	QMREP
1094	--I-----	-----T--	-----	-----	-----D--	-----L-	-I----V--	-----
1096	--I-----	-----T--	-----	-----	-----D--	-----L-	-----V--	-----
1097	-----	-----T--	-----T--	-----	-----D--	-----L-	-----	-L----
1098	--I-----	-I----T--	-----	-----	-----D--	-----L-	--Q----P--	-----
1099	--I-----	-----T--	-----	-----	-----D--	-----L-	-----	-----
1101	--I--E----	-----T--	-----	-----	-----D--	-----L-	-----	-----
1102	--I--R---	-----T--	-----	-----	-----D--	-----L-	-----V--	-V----
1104	-GI-----	-I----T--	-----	-----	-----L-D-T	-----L-	-----V	-----
1106	-----G-N-	-----T--	-----	-----	-----D--	-----L-	-----	-----
1108	--I-----	-I----T--	-----	-----	-----D--	-----M-	-----	-----
1110	--I-----	-----T--	-----	-----	-----D--	-----I-	-QQ-----	-I----
1112	--I-----	-I----T--	-----	-----	-----D--	-----	-----N--	-----
1113	--I-----	-----T--	-----	-----	-----D--	-----L-	-----	-----
1114	--I-----	-----T--	-----	-----	-----D--	-----L-	-----	-L----
1115	--I-----	-----T--	-----	-----	-----D--	-----L-	-----	-----
1116	--I-----	-----T--	-----T--	-----	-----D--	-----L-	-----	-----
1118	--I-----	-----T--	-----	-----	-----D--	-----L-	--Q-----	-I----
1119	--I--G-N-	-----	-----S--M	-----I	-----D--	-----L-	-----V	-----
1120	--I-----	-I----T--	-----	-----	-----D--	-----L-	-----V	-----
1121	--I--D---	-I----T--	-----	-----	-----D--	-----L-	-----V--	---D--
1123	--I--MS-T-	-L----T--	-----	-----	-----D--	D-----L-	-----	-----
1125	--I-----	-----T--	-----	-----	-----I-D-	-----M-	-----	-----
1127	--I-----	-I----T--	-----	-----	-----D--	-----	-----N--	-----
1129	--I-----	-----T--	-----S--	-----	-----D--	-----L-	-----V--	-----
1131	--I-----	-----T--	-----	-----	-----D--	-----L-	-----V--	-----
1132	-----	-----T--	-----	-----	-----D--	-----L-	-----	-I----
1133	--I-----	-----T--	-----	-----	-----D--	-----L-	-----	-I----
1135	--I-----	-----T--	-----	-----	-----D--	-----I-	-----	-----
1136	--I-----	-----T--	-----	-----	-----D--	-----M-	-----N--	-----
1137	--I--R---	-----T--	-----S--	-----I	-----D--	-----L-	-----	-----
1138	--I-----	-I----T--	-----	-----	-----D--	-----L-	-AQ---V--	-I----
1140	--I-----	-----T--	-----	-----	-----D--	-----L-	-----	-----
1141	--I-V--L--	-----	-----	-----	-----D--	-----	-----	-----



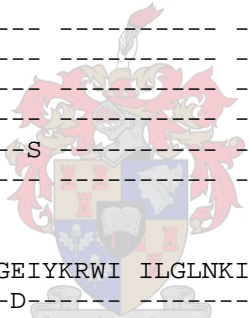
**B1 continue: gag p24 amino acid sequence alignment**

B.FR.83.HXB2	KVVEEKAFSP	EVIPMFSALS	EGATPQDLNT	MLNTVGGHQA	AMQMLKETIN	EEAAEWDVRH	PVHAGPIAPG	QMREP
1142	--I-----	-----T--	-----	-----	-----D--	-----L-	-----V--	-----
1143	--I-----	-----T--	-----	-----	-----D--	-----L-	-----V--	-----
1146	--I--S---	-I---T---	-----	-----	-----D--	-----L-	--Q--V--	-----
1148	---KK---	-----T--	----T---	-----	-----D--	-----L-	-----NP--	-----
1151	--I-----	-----T--	-----	-----	-----D--	-----L-	-----P--	-----
1152	--I-----	-----T--	-----	-----	-----D--	D-----L-	-----A--	-----
1153	--I-----	-----T--	-----	-----	-----D--	-----L-	-----	-----
1154	--I-K---	-----	-----	-----	-----	-----L-	-----	-----
1155	--I-----	-----T--	-----	-----	-----D--	-----L-	-----	-----
1156	--I---N-	-----T--	-----	-----	-----D--	-----	-----	-----
1157	--I-----	-----T--	-----	-----	-----D--	-----L-	-----	-----
1160	.....	...--F-	D---G-	-----	-----D--	-----L-	-----	-----
1162	--I-----	-----T--	-----	-----	-----D--	-----	-----	-----
1169	--I-----	-----T--	-----	-----	-----D--	-----L-	-----V--	-----
1173	--I-----	-----T--	-----S-	-----	-----DA--	-----L-	--Q-----	-----
1174	.....-	-A---T--	-----	-----	-----D--	-----L-	-P---V--	-I---

75

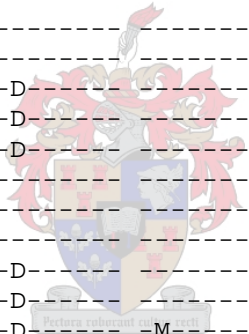
B.FR.83.HXB2	RGSDI	AGTTSTLQEQ	IGWMTNPPPI	PVGEIYKRWI	ILGLNKIVRM	YSPTSILDIR	QGPKEPFRDY	VDRFYKTL
C.BW.96.96BW0502	-----	--A-----	-A--S--V	--D-----	-----	---V-----	-----	----F--
C.IN.95.95IN21068	-----	-----	-A-----V	--D-----	-----	---V-----	-----	----F--
C.ET.86.ETH2220	-----	-----	-A--G--V	--D-----	-----	---V---K	-----	----F--
C.BR.92.92BR025	-----	-----	-T-----V	--D-----	-----	---V---K	-----	----F--
1001	-----	-----	-A-----V	-----	-----	---V-----	-----	----F--
1002	-----	-----	-A--S--	--D-----	-----	---V---K	-----	----F--
1003	-----	-----	-A-----	-----	-----	---V---K	-----	----F--
1005	-----	-----	-A--G--	--D-----	-I-----	---V---K	-----	----F--
1006	-----	-----	-A--S--	-----	-----	---V--VH--	-----	----F--
1008	-----	-----	-A--AY--	-----	----I-----	.....	.....	.....
1009	-----	---N---	-A--S--	--D-----	-----	---V-----	-----	----F--
1010	-----	-----	-A--G--V	--D-----	-----	---V---K	-----	----F--
1011	-----	---N---	-A-I-G--V	--D-----	V-----	---V---K	-----	----F--F
1012	-----	-----	-A--S--V	-----	-----	---V---E	-R-----	----F--

148



**B1 continue:** gag p24 amino acid sequence alignment

B.FR.83.HXB2	RGSDI	AGTTSTLQEQ	IGWMTNNPPI	PVGEIYKRWI	ILGLNKIVRM	YSPTSILDIR	QGPKEPFRDY	VDRFYKTL
1013	-----	-----	-A--G----	---D-----	-----	---V----K	-----	----F---
1015	-----	-----	-T--S----	---D-----	-----	---V----K	-----	----F---
1016	-----	-----	-A-T-G----	---D-----	-----	---V----K	-----	----F---
1017	-----	-----	-A--G--V	-----	-----M--	---V--SH-	----Q-----	GVL-F-S-
1018	-----	-----	-A--G--V	---D-----	-----	---V----K	-----	----F---
1019	-----	-----	-T--S--V	-----	-----	---V-----	-----	----F---
1021	-----	-----	-A--G----	---D-----	-M-----	---VT---K	-----	EA--F---
1023	-----	-----	-A--S----	---D-----	-I-----	---V----K	-----	----F---
1024	-----	-----	-T-----	-----	-----	---V----K	-----	----F-A-
1026	-----	-----	-A--S--M	-----	-----	---V----K	-----	-A--F---
1029	-----	-----	-A--G--V	-----	-----	---V----K	-----	----F-V-
1031	-----	-----	VA-I-G----	-----	-----	---VR---K	K--R--V--	----F---
1033	-----	-----	-A--S----	---D-----	-----	---V----K	-----	----F---
1034	-----	-----	-A--S----	---D-----	-----	---V----K	-----	----L-V-
1037	-----	-----S----	-A--S----	---D-----	-----	---V-----	-----	----LF---
1038	-----	-----	-A--G----	-----	-----	---V----K	-----	----F---
1039	-----	-----	-A--S----	-----	-----	---V----K	-----	----F---
1040	-----	-----	-T-----V	-----	-----	---V----K	-----	----F---
1042	-----	-----	VT--S----	---D-----	-----	---V----K	-----	----F---
1043	-----	-----	-A--G----	---D-----	-----	---V----K	-----	----F---
1044	-----	-----	-A--S----	---D-----	-----M-----	---V----K	-----	----F---
1045	-----	-----	-A--G----	---D-----	-----	---V----K	-----	----F---
1047	-----	-----	-A-----V	---D-----	-----	---V-----	-----	----F---
1048	-----	-----	-T-I-----	-----	-----	---V-----	-----	----F---
1049	-----	-----	-A--S----	---D-----	-----	---V-.....	.....	.....
1050	-----	-----N----	VT-----	-----	-----	---V----K	-----	----F---
1052	-----	-----	-A--S----	-----	-----	---V----K	-----V-E-	AV--FR--
1054	-----	-----N----	-A--H--V	-----	V-----	---V-----	-----E-V-	-VG-F---
1055	-----	-----N----	-A-----	---D-----	-I-----	---V----K	-----	----F-I-
1056	-----	-----N----	-A--S----	---D-----	-I-----	---V----K	-----	----F---
1057	-----	-----	-A--S--M	---D-----	-----	---V--..	.....	.....
1058	-----	-----	-A-----	---D-----	-M-----	---V----K	-R-----	-N--F---
1059	-----	-----	-A--G--V	-----	-----	---V-----	-----	----F-V-



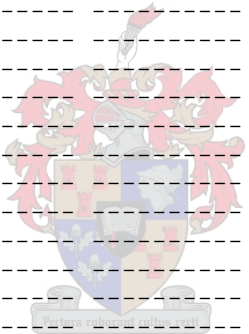
**B1 continue: gag p24 amino acid sequence alignment**

B.FR.83.HXB2	RGSDI	AGTTSTLQEQ	IGWMTNNPPI	PVGEIYKRWI	ILGLNKIVRM	YSPTSILDIR	QGPKEPFRDY	VDRFYKTL
1060	-----	-----	-T---G----	---D-----	-----	---V----K	-----	----F-V-
1061	-----	-----	-A---G----	---D-----	-----	---V----K	-----	----F-V-
1062	-----	-----	-A--S--V	---D-----	-----	---V-----	-----I--	----F-F-
1063	-----	-----	-A--S--V	---D-----	-----	---V-N---K	-----EK-	----F---
1064	-----	-----	-A--S--M	---D-----	-----	---V----K	-----	----F---
1067	-----	-----	-T---G----	---D-----	-----	---V----K	-----G	I--VF-S-
1068	-----	-----	-T---S--V	---D-----	-----	---V----K	-----	----F---
1069	-----	---N----	-A--S----	---D-----	-M-----	---V----K	-----	----F---
1072	-----	---N----	-A--A----	---D-----	-M-----	---V----K	-----V-V-	-A--L-G-
1073	-----	-----	-A-I-G----	-----	V-----	---V-----	---Q-----	---VL---
1075	-----	-----	-A--S--V	---D-----	-----	---V----K	-----	----F-F-
1076	-----	-----	-A-----	---D-----	-----	---V--V-K	-----	----F---
1077	-----	-----	-A--S--V	-----	-----	---V-----	-----	----F---
1079	-----	---N----	-A--A----	---D-----	-M-----	---V----K	-----	----F---
1083	-----	-----	-A-----	---D-----	-----	---V----K	-----	----F---
1084	-----	-----	-A--G----	-----	R-----	-----	-----	-----
1088	-----	---E----	VR--S--V	-----	-----	---V-----	-----	----F---
1089	-----	---N----	-T---S----	---D-----	I-----	---V----K	-----	----F---
1090	-----	-----	-A--S----	---D-----	-----	---V--VK	-----	----F-I-
1094	-----	-----	-A--S----	---D-----	-----	---V----K	-----	----F-C-
1096	-----	-----	-A-I-G----	-----	-----	---V----G	-----	----F---
1097	-----	-----	-A-----V	---D-----	-----	---V-----	-----	----F---
1098	-----	---N----	-A--S--V	---D-----	-----	---V-----	-----	----F---
1099	-----	---S----	-A--S----	---D-----	-----	---V----K	-----	----F---
1101	-----	-----	VA--S----	---D-----	-----	---V----K	-----I--	---VF---
1102	-----	---N----	-N--G----	-----	-----	---V----K	-----I--	---F---
1104	-----	-----	-A--S--V	---D-----	-----	-----	-----	-----
1106	-----	---S----	-A-----V	-----	V-----	---V-----	-----	----F---
1108	-----	-----	-N--S----	---D-----	-----	---V----K	-----L--	G--F---
1110	-----	-----	-T---S----	---D-----	-----	---V----K	-----	---FR---
1112	-----	-----	-A--S----	---D-----	-----	---V----K	-----	----F---
1113	-----	-----	-A--G--V	---D-----	-----	---V----K	-----	----F---
1114	-----	-----	-A-----V	---D-----	-----	---V-----	-----	----F---



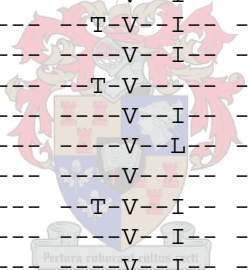
**B1 continue: gag p24 amino acid sequence alignment**

B.FR.83.HXB2	RGSDI	AGTTSTLQEQ	IGWMTNPNPI	PVGEIYKRWI	ILGLNKIVRM	YSPTSILDIR	QGPKEPFRDY	VDRFYKTL
1115	----	-----	-A--G---V	---D-----	-----	---V----K	-----	----F---
1116	----	-----	-A--S---	---D-----	-----	---V----K	-----	----F---
1118	----	---N----	-A--S---	---D-----	-I-----	---V----K	-----	----F---
1119	----	---N----	-A--G---V	---D-----	-----	---V----K	-----	----F---
1120	----	---N-A--	-A--S---V	---D-----	-----	---V----K	-----	----F---
1121	----	-----	-A--S---V	---D-----	-----	---V-----	-----	---PF---
1123	----	-----	-A--S---	---D-----	-----	---V-----	-----	----F---
1125	----	-----	-A--S---	---D-----	-----	---V----K	---K---	----F---
1127	----	-----	-A--S---	---D-----	-----	---V----K	---D-.-	LYP-F-L-
1129	----	-----	-A-----V	---D-----	-----	---V----K	-----	----F---
1131	----	---N----	-A--S---	---D-----	-----	---V----K	-----	----F---
1132	----	-----	-T--S---M	---D-----	-----	---V----K	-----	----F---
1133	----	-----	-T--S---M	---D-----	-----	---V----K	-----	----F---
1135	----	-----	-Q--S---V	---D-----	-----	---V-----	-----	----F---
1136	----	-----	-A--S---V	---D-----	-----	---V-----	-----	----F---
1137	----	-----	VA--S---	---D-----	-----	---V----K	---V---	----F---
1138	----	-----	-T--S---	---L-----	-----	---V----K	-----	----F---
1140	----	-----	-A-I-S--S-	-----	-----	---V-----	-----	---N--F--
1141	----	-----	-A--S---	-----	-----	---V----K	-----	----F---
1142	----	---S----	-T--S---	---D-----	-----	---V----K	-----	----F---
1143	----	---S----	-T--S---	---D-----	-----	---V----K	-----	----F---
1146	----	---N----	VA--S---V	---D-----	-----	---V----K	-----	----F---
1148	----	-----	-A--S---	---D-----	-----	---V-L---	---LQ--	---AF---
1151	----	-----	-A--A---V	---D-----	-----	---V----K	-----	----F---
1152	----	-----	---S---	---D-----	-----	---V-----	-----	----F---
1153	----	-----	-A-I-G---	-----	-----	---V-----	---Q---	----F---
1154	----	-----	-A-----	---D-----	-----	---V-----	-----	----F---
1155	----	-----	-A-I-G---	-----	-----	---V-----	-----	----F---
1156	----	---L----	-A--G---	---D-N---	-----	---V----K	-----	----F---
1157	----	-----	-A--S---	---D-----	-M-----	---V-----	-----	----F---
1160	----	-----	-A--S---	---D-----	TM-.....	.....	.....	.....
1162	----	-----	-A--S---V	---D-----	-----	---V-----	-----	----F---
1169	----	-----	MA-I-G--V	-----	---E---	---V----K	-----	---F-A-
1173	----	-----	-A--G---	-----	-----	---I--V-	-----	----F---
1174	----	---P----	-A-----	---D-----	-----	---V----K	---Q-.....	.....



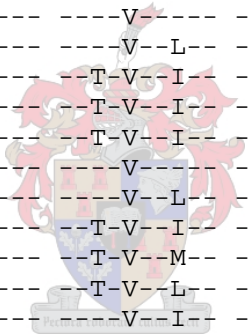
**B2: env gp41 IDR amino acid sequence alignment**

B.FR.83.HXB2	SGIVQQQNNL	LRAIEAQQHL	LQLTVWGIKQ	LQARILAVER	YLKDQQLLGI	WGCSGKLICT	TAVPWNASWS	NKSLE
C.BW.96.96BW0502	-----S--	-----M	-----I--	T-----	-----L	-----S--	-----R--	HD
C.IN.95.95IN21068	-----S--	-K-----M	-----I--	T-V-I--	H-R-----	-----S--	-----Q-	RTQK
C.ET.86.ETH2220	-----S--	-----M	-----I--	T-V-I--	--R-----	-----S--	-----R-Q-	
C.BR.92.92BR025	---GP-----	-----M	-----I--	V-----	-----L	-----N---	-----S--	QT
1001	-----S--	-----M	-----I--	T-V-L--	--R-----	-----T---	-----Y---	RTEK
1002	---P-S-----	-----M	-----I--	T-V-I--	-----	-----S--	-----R--	RTQD
1003	-----S--	Q-----M	-----I--	-----	-----	-----S--	-----S--	Q-
1005	-----S--	-----M	-----I--	V-----	-----L	-----D---	-----S--	EA
1006	-----S--	-----M	-----I--	T-V-I--	-----	-----D-L--	-----S--	K-
1008	---E-S-----	-----M	-----I--	T-V-L--	-----	-----C---	-----	KS
1009	-----S--	-----M	-----I--	V-----	-----L	-----N---	-----S--	QQ
1011	-----S--	-----M	-----I--	T-V-I--	-----L	-----P-----	-----S--	D--QK
1012	-----S--	-M-----M	-----I--	V-----	--R-----	-----S--	-----	H-
1013	-----S--	-----M	-----I--	T-V-I--	--R-----	-----S--	-----	QH
1015	-----S--	-----M	-----I--	V-----	-Q-----L	-----S--	-----	QT
1016	-----S--	-----M	-----I--	V-L--	-T-----M	-----E---	-----	Q-
1017	-----S--	-----M	-----I--	V-----	--Q-----	-----N---	-----	FH
1018	-----S--	-----M	-----I--	T-V-I--	-----	-----S--	-----	TQ-
1021	-----S--	-K-----M	-----I--	V-----	-----	-----S--	-----	Q-
1023	-----S--	-----M	-----I--	V-----	-----	-----S--	-----	VT
1024	-----S--	-----NM	-----I--	V-----	-I-----	-----S--	-----	RTV-
1025	-----S--	-----M	-----I--	V-----	--Q-----	-----N---	-----	R-Q-
1026	-----S--	-----M	-----I--	V-----	-----L	-----N---	-----	YT
1027	-----S--	-----M	-----I--	T-V-I--	-----L	-----N---	-----	QG
1029	-----S--	-----M	-----I--	V-----	-----L	-----N---	-----	Q-
1031	.....	.....	.....	V-----	-----L	-----N---	-----	NQT
1034	.....	-----M	-----I--	V-----	-----L	-----V---	-----S--	QT
1037	-----S--	-----M	-----I--	V-----	-----	-----S--	-----	Q-
1040	-----S--	---V---M	---A---	T-V-I--	-----	-----N---	-----	QD
1042	-----S--	-M-----M	-----I--	V-----	-----L	-----S--	-----	QT
1043	-----SH-	---M---	---D---	V-----	-----L	-----S--	-----	R..T
1044	.....	.....-P-M	-----I--	V-----	-----	-----S--	-----	YD
1045	-----S--	-----M	-----I--	T-V-I--	-----L	-----S--	-----	YG



**B2 continue: env gp41 IDR amino acid sequence alignment**

B.FR.83.HXB2	SGIVQQQNNL	LRAIEAQQHL	LQLTVWGIKQ	LQARILAVER	YLKDQQLLGI	WGCSGKLICT	TAVPWNASWS	NKSLE
1047	.....	....-----M	-----	---V-I--	-----L	-----	-N---S---	---KT
1048	---P-S--	-K-----M	-----	---V-I--	-----M	-----	--L---N---	-RTQK
1049	---P-S--	-----M	-----	--T-V-I--	-----	-----	--L---T---	---Q-
1050	-----S-	-M-----M	-----	---V-I--	-----L	-----S	-----S---	----T
1052	-----S-	-----M	-----	---V---	--R-----	-----	-----S---	--TEK
1054	-----S-	-----M	-----	--T-V-I--	-----L	-----	-N---S---	---QA
1057	-----S-	-----M	-----	---V-L--	-----L	-----	-T---S---	---Q-
1058	-----S-	-----M	-----	---V-I--	-----	-----	-----S---	---Q-
1059	-----S-	-----M	-----	---V-L--	-----L	-----	-T---S---	---Q-
1060	-----S-	-----M	-----	---V-I--	-----	-----	-N---S---	----D
1061	-----S-	-K-----M	-----	---V---	--R-----	-----	-----S---	--TE-
1064	-----S-	-----M	-----	---V-L--	--R-----	-----V--	-----	---EA
1067	---T-S--	-----M	-----	--T-V-I--	-----L	-----	-N---S---	---QT
1068	---T-S--	-----M	-----	--T-V-I--	-----	-----	-----S---	---KD
1069	---T-S--	-----M	-----	--T-V-I--	-----	-----	-----S---	---E-
1075	---T-S--	-----M	-----	---V---	--R-----	-----	-----S---	--TEK
1076	---T-S--	-----M	-----	---V-L--	--R-----	-----V--	-----	---EA
1077	---T-S--	-----M	-----	--T-V-I--	-----L	-----	-----S---	---QT
1079	---T-S--	-----M	-----	--T-V-M--	-----	-----	-----N---	---QA
1083	---T-S--	-----M	-----	--T-V-L--	-----L	-----	-----S---	---QA
1088	---T-S--	-----M	-----	---V-I--	-----L	-----V--	-----S---	---YA
1089	---T-S--	-----M	-----	---V-I--	-----	-----	-----S---	---H-
1090	---T-S--	-----M	-----	--T-V-I--	--R-----	-----	-N---S---	---A
1094	---T-S--	-----M	-----	--T-V-I--	-----L	-----	-----S---	---K-
1096	---T-S--	-K-----M	-----	--T-V-I--	-----L	-----I---	-----S---	---DT
1097	---T-S--	-K-----	-----	--T-V-I--	-----L	-----	-----K---	S--E-
1098	---T-S--	-----M	-----	--T-V-I--	-----L	-----	-N---S---	-R-ED
1099	---T-S--	---L-----M	-----	---V-I--	--Q-----	-----	-----T---	-R-E-
1100	---T-S--	-----M	-----	--T-V-L--	-----	-----	-T---S---	--THD
1102	---T-S--	-----M	-----	--T-V-I--	-----	-----	-D---S---	---QT
1104	---T-S--	-----M	-----	---V-I--	-----L	-----	-N---S---	-R-QT
1106	---T-S--	-----M	-----	---V-I--	-----	-----	-----S---	---HN
1108	---T-S--	-----	-----	--T-V-I--	--Q-----	-----	-----S---	-R-KD
1109	---T-S--	-----M	-----	---V---	-----L	-----	-----S---	---HK



**B2 continue: env gp41 IDR amino acid sequence alignment**

B.FR.83.HXB2	SGIVQQNNL	LRAIEAQQHL	LQLTVWGIKQ	LQARILAVER	YLKDQQLLGI	WGCSGKLICT	TAVPWNASWS	NKSLE
1110	-----S--	-----M	-----	---V--I--	-----F--	-----	-----S--	--TQ-
1112	-----	-----M	-----	---V--I--	-----L	-----	-----N--	---YT
1113	-----S--	-----M	-----	---V--L--	--Q-----M	-----	-----S--	--TYN
1114	-----S--	-K--Q----	-----	---V--I--	-----	-----	-----NT--	---HD
1116	-----S--	---V-----M	-----	--T-V--I--	-----L	-----	-----S--	---K-
1118	-----S--	-----M	-----	---V--I--	--R-----L	-----	-N---S---	---Q-
1119	-----S--	-----	-----	--T-V-----	--R-----	-----	-D---S---	---D
1120	-----S--	-----	-----	--T-V--I--	-----	-----	-N---S---	---YN
1121	-----S--	-----M	-----	---V--I--	-----L	-----	-N---S---	---QA
1123	-----S--	-----M	-----	---V--I--	-----L	-----	-T---S---	---EA
1125	-----S--	-K-----M	-----	---V--I--	-----L	-----	-----S--	---K-
1131	-----S--	-K-----	-----	---V--I--	-----	-----P	-----S--	---Q-
1134	.....	.....-M	-----	--T-V-SI--	-----	-----	-N---S---	---H
1135	-----S--	-----M	-K-----	-----	-----L	-----I--	-----S--	-R-H-
1136	-----S--	-----M	-----	--T-V--I--	-----L	-----	-N---S---	---KT
1137	-----S--	-----M	-----	---V--I--	-----	-----	-T---S---	---E-
1138	-----	-----M	-----	---V--I--	-----	-----I--	-----N--	--TYN
1141	-----	-----	-----	---V--I--	--Q-----	-----	-N---S---	-R-Q-
1142	-----S--	---L---M	-----	---V--I--	-----	-----	-----	---EK
1143	-----S--	-----M	-----	--T-V--I--	-----L	-----	-N---H--	---G
1146	.....	.....	-----	---V--L--	--R-----	-----N	-N---S---	-I--N
1147	-----S--	-----M	-----	--T-V--I--	-----	-----I--	-----S--	--T-G
1149	-----S--	-----	-----	--T-V--I--	-----L	-----I--	-N---S---	---KA
1151	-----S--	-----M	-----	---V--I--	-----L	-----	-N---S---	---TA
1152	-----S--	-----M	-----	--T-V-----	-----L	-----	-----S--	---QA
1153	-----S--	-----M	-----	--T-V--I--	-----	-----	-----S--	--N-T
1155	-----S--	-----M	-----	---V-----	--Q-----L	-----	-----S--	--N-T
1156	-----	-M-----M	-----	---V--I--	-----	-----	-D---S---	---EK
1160	--G--S--	-----M	-----	---V--I--	-----SL	-----	-T---S---	-RTKD
1162	-----S--	-----M	-----	--T-V--I--	-----L	-----	-N---S---	---K-
1163	-----S--	-K-----	-----	--T-V--I--	-----L	-----	-----L--	--TQ-
1165	.....	.....-M	-----	---V--I--	-----	-----	-----S--	---HD
1169	-----S--	-----RM	-----	--T-V-----	-----	-----	-N---S---	-R-ET
1172	-----S--	-----M	-----	---V--I--	-----	-----	-----T--	---QP

**B2 continue: env gp41 IDR amino acid sequence alignment**

B.FR.83.HXB2 SGIVQQNNL LRAIEAQQHL LQLTVWGIKQ LQARILAVER YLKDQQLLGI WGCSGKLICT TAVPWNASWS NKSLE  
 1173 -----S-- -K-----M -----V--I-- -----R-----S--- ---EA  
 1174 -----S-- -K-----M -----V--I-- -----S--- ---EK

75

B.FR.83.HXB2 QIWNH TTWMEWDREI NNYTSLIHSL IEESQNQQEK NEQELLELDK WASLWNWFNI TNWLWYIKLF IMIVGGL  
 C.BW.96.96BW0502 E--DN M---Q----- ----DT-YR- L----- --KD--A--S -QN-----S- -----I- -----  
 C.IN.95.95IN21068 E--DN M---Q----- ----NT-YR- L-----E --KD--A--S -KN-----D- -K-----I- -I-----  
 C.ET.86.ETH2220 E--DN M---Q----- S---DI-YN- L-V-----D- --KD--A--- -EN-----I- -----V  
 C.BR.92.92BR025 D--N M---Q----- S---NT-YR- L-D----- --D--A--- -QN--T--G- -----I- -K-----  
 1001 D--DN M---Q----- R---DT-YL- L---S--- --KD--A--S -KN--S--D- S-----RI- -----  
 1002 D--DN M---Q-E--- D---DT-YR- L---S--- --D--A--R -ND-----G- -R-----I- --M-----  
 1003 D--N M---Q----- ----NT-YR- L-V-----E --KD--A--NN -QN-----D- -----I- -----  
 1005 E--SN M---Q--K-- ----NT-YR- L---S--- --KD--A--S -NNSG--S- A---S--I- -----  
 1006 D--EN M---Q----- S---NI-YG- L-Y----- --KD--A--S -KN-----D- -----I- -----  
 1008 E--EN M---Q----- S---DT-YR- L---S---R -N--A--S -N-----S- -----I- --V-----  
 1009 E--DN M---Q----- S---AT-YK- L-D---I---Q --KD--A--S -Q-----S- -R-----I- -----  
 1011 D--DN M---Q----- ----NT-YR- L---T---N --KD--A--S -NN-----DL -K-----I- -----  
 1012 D--DN M---Q----- S---GT-YR- L----- --RD--A--S -NN--D---- -K-----I- -----  
 1013 D--DN M---Q----- S---G--YK- L-D---I---N --KD--A--S -NN-----D- -----I- -----  
 1015 D--DN L---Q----- S---DT-YR- L-D-----R --KD--A--S -NT--S--D- S-----I- -----  
 1016 D--DN M---Q----- S--SNT-YR- L----- --KD--A--S -NN--S--S- -----I- -----  
 1017 D--DN M---Q----- S---NT-YR- L-N----- --KD----- -SN-----S- -K-----I- -----  
 1018 D--DN M---Q----- S--SNT-YK- L---I---Q --KD--A--- -Q-----S- A-----I- -----  
 1021 E--GN M---Q----- S---NT-YR- L-----N --KD--A--S -EN-----D- -K-----I- -----  
 1023 E--GN M---Q----- S---NT-YR- L-D-----E --KN--A--S -N-----D- SK-----RI- -----  
 1024 E--DN M-----E--- D---ET-YR- L-I-----Q --KD--A--- -QN--S--D- S-----RI- -----  
 1025 D--DN M---Q-E--- D---NT-YR- L----- --D--A--R -NN-----G- -R--S--I- --M-----  
 1026 D--DN M---Q----- S---NT-YR- L-D----- --KD--A--SS -QN--S--S- -----I- -----  
 1027 D--DN M---Q----- S---KT-YR- L-D--S---E --KD--A--S -KN--S--S- SK-----I- -----  
 1029 D--DN M---Q----- S---NT-YR- L-D-----E --KD--A--S -KN-----S- -----I- -----  
 1031 E--DK M---Q----- D---I--YG- L-D----- --KD--A--S -NN-----D- -K-----I- -----  
 1034 E--DN M---Q----- S--SYT-YR- L----- --KD--A--HS -EN--S--S- -----SL-I- ---E--  
 1037 D--GN M---Q----- S---NT-YR- L-D----- --KD--A--S -E-----S- -K-----I- -----  
 1040 Y--GN M---Q--K-- ----DT-YR- LG-A----- --K-----R -GN-----D- -K-----I- -----

147

**B2 continue: env gp41 IDR amino acid sequence alignment**

B.FR.83.HXB2	QIWNH	TTWMEWDREI	NNYTSLIHSL	IEESQNQQEK	NEQELLELDK	WASLWNWFNI	TNWLWYIKLF	IMIVGGL
1042	D--DN	M---Q-----	S---HT-YK-	L-D--S-----	--KD--A--S	-NN-----	-K-----I-	-----
1043	D--DN	M---Q--K--	S---DI-YG-	L-D-----	--KD--A--S	-N---S-LTL	S----L--I-	-----
1044	E--GN	M---Q-----	---NT-YR-	L-----	--KD--A--S	-EN-----S-	-K-----I-	-----
1045	S--DN	M---Q-----	-K--NT-Y--	L-----Q	--KD--A--S	-Q-----	-K-----I-	-----
1047	D--DN	M---Q-----	S---DT-YM-	L-D--I-----	--KD--A--S	-NN-----	S-----I-	-----
1048	D--DN	M-----	S---DT-YR-	L-D-----R	--KD--A--S	-NN-----S-	-----I-	-----
1049	E--EN	M---Q-----	---ET-YR-	L-D-----R	--D--A--S	-NN-----S-	-K-----I-	-----
1050	D--EN	M---Q-----	---NT-YR-	L---S-----	--KD--A--GS	-NN-----D-	-----I-	-----
1052	D--DN	M---Q-E---	D---DT-YR-	L-----D-	--D--A--R	-NN-----G-	-K-----I-	-----
1054	D--EN	M---Q-----	S-F-NT-YR-	L---S--N	--KD--A--S	-NN-----D-	S-----I-	-----
1057	D--DN	M---Q-----	S---NT-YR-	L-D-----R	--KD--A--S	-MN--S--D-	S-----I-	-----
1058	E--DN	M--IQ-----	S---DT-YR-	L-D-----	--KD--A--S	-KN--S---	-----I-	-----S-
1059	D--DN	M---Q-----	S---NT-YR-	L-D-----R	--KD--A--S	-MN--S--D-	S-----I-	-----
1060	YV--N	M--LQ-E---	D---D--YK-	L-----D-	--D--A--R	-QN--S--S-	-K-----I-	-----
1061	D--EN	M---Q-E---	D---DT-YR-	L-K-----	--D--A--R	-NN-----D-	-R-----I-	-----
1064	E---N	M---Q-----	---DT-YR-	L-V--T---Q	--KD--A---	-QN--S--D-	-----I-	-----
1067	D--DN	M---Q--K--	---DT-YR-	L-A-----E	--KD--A--S	-KN-----	S-----I-	-----
1068	S--DN	M---Q-----	S---DI-YR-	L---I---Q	--D--A--R	-KN-----A-	-----RI-	-----
1069	E---N	M---Q-----	S---NT-YR-	L-D-----	--D--A--R	-KN--T--D-	-----RI-	-----
1075	D--DN	M---Q-E---	D---DT-YR-	L-----D-	--D--A--R	-NN-----GM	-K-----IL	-----
1076	E---N	M---Q-----	---DT-YR-	L-V--T---Q	--KD--A---	-QN--S--D-	-----I-	-----
1077	D--DN	M---Q--K--	---GI-YR-	L-D-----	--KD--A--S	-KD--T--S-	-K-----I-	-----
1079	E--EN	M-----	---DT-YK-	L-I-----E	--KD--A--N	-KN-----D-	-----I-	-----
1083	E---N	M---Q-----	S---GT-YR-	L-D-----	--KD--A--S	-KN--S--D-	-----I-	-----
1088	E--GN	M---Q--E--	---DT-YR-	L-V--T---N	--KD--A--S	-KN--S--D-	S-----I-	-----
1089	D--DN	M---Q-----	---NT-YR-	L-D--S-----	--KD--A--S	-EN-----S-	-K-----I-	-----
1090	E---N	M---Q-EK--	S---GT-YR-	L-D--T---Q	--KD--A--S	-KN-----D-	-----I-	-----
1094	D--DN	M---Q-----	S---GI-YQ-	L-D--S-----	--KD--A--S	-QN-----	S-----RI-	-----
1096	D--DN	M---Q-----	S---GI-YR-	L-D-----	--D--A--NS	-NN-----	SQ--R--QI-	-----
1097	D--DN	M---Q-----	A---NT-YQ-	L-----	--KD--A--S	-N-----D-	-K-----I-	-----
1098	Y--DN	M---Q-----	---DT-YR-	L-----Q	--KD--A--S	-KN--S--T-	S-----I-	-----
1099	E--DN	M---Q--S--	R---GI-Y--	L-D-----	--KD--A--S	-KN-----D-	-----	-----
1100	E---N	M---Q-----	---NT-YN-	L-V-----Q	--KD--A---	--N-----D-	-----I-	-----
1102	D--DN	M---Q-----	S---DT-YR-	L-I-----Q	--KD--A--S	-KN--S--D-	S-----I-	-----

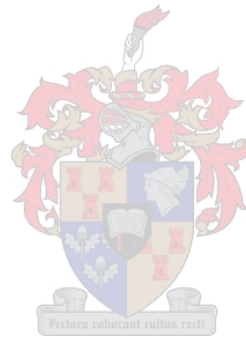
**B2 continue: env gp41 IDR amino acid sequence alignment**

B.FR.83.HXB2	QIWNH	TTWMEWDREI	NNYTSLIHSL	IEESQNQQEK	NEQELLELDK	WASLWNWFNI	TNWLWYIKLF	IMIVGGL
1104	D--DN	M---Q-----	S---ET-YW-	L-D-----	--KD--A--S	-NN-----D-	SK-----I-	-----
1106	D--DN	M---Q-----	S--SET-YR-	L-----	--KD--A--S	-TN-----S-	S-----RI-	-----
1108	E---N	M---Q-----	----DT-YR-	L-K--T---Q	-----A--	-KN--T--D-	SS-----I-	-----
1109	I--EN	M---Q-E---	S---GT-YR-	L-D-----	E--KD--A--S	-KN--S--D-	-----I-	-----
1110	E--DN	M---Q-----	S---NT-YK-	L-D--I----	---D--A--S	-NN--S----	-----RI-	-----
1112	D--DN	M---Q-----	S----T-YR-	L-D--S---E	--KD--A--S	-KN--S--S-	S-----RI-	-----
1113	E--DN	M---Q-E---	G---DT-YR-	L-D-----	R--KD--A--R	-NN--S----	-----I-	-----
1114	E--GN	M---Q-----	S---YT-Y--	L----T---Q	--KD--A--S	-QN-----D-	-K-----I-	-----
1116	D--EN	M---Q-----	S---YT-YR-	L-D-----	E--KD--A--S	-NN-----D-	-K-----I-	-----
1118	E--GN	M--IQ-----	----NT-YR-	L-D-----	--KD--A--S	-KN-----S-	-----RI-	-----
1119	Y--N	M--D---V	---ET-YR-	L-D-----	---D--A--R	-NN-----G-	-K-----I-	-----
1120	D--DN	M-----	----DT-YR-	L-K--T---Q	--KD--A--S	-KN--S-LS-	S-----RI-	-----
1121	D---N	M---Q-----	S---NT-YK-	L-----	--KD--A--S	-KN-----D-	S-----I-	-----
1123	D--DN	M---Q-----	----GT-YQ-	L---T---	--RD--A--	-N--S--D-	-----TI-	-----
1125	D---N	M---Q-E---	S---GT-YR-	L-D--T---	--KD--A--S	-KN--T--D-	S-----I-	-----
1131	D---N	M-----	S---TT-YR-	L-D--S---	--KD--A--S	-KF-----D-	-R-----I-	-----
1134	D--DN	M---Q-----	----NI-Y--	L---I---Q	--KD--A--	-QN-----S-	-H-----I-	-----
1135	E---N	L---Q-EK--	D---DT-YR-	L-D--T---T	--K--A--S	-NT--S-LS-	-----RI-	-----
1136	D--DN	M---Q-----	S---NT-YR-	L-----	--KD--A--S	-KN--T--D-	SK-----I-	-----
1137	E---N	M---Q-----	G---NI-YR-	L-----	--KD--A--S	-QN--S--S-	-----RI-	-----
1138	D--DN	M---Q-----	----DT-YR-	L-D-----	--KD--A--S	-N-----	-K-----I-	-----
1141	D--KN	M---Q--K-V	SKH-NT-YR-	L-D--I---Q	--KD--A-NS	-D-----D-	-----R-	-----
1142	E--DN	M---Q-----	R--SGI-Y--	L-----	---D--A--R	-EN--S--S-	S-----I-	-----
1143	D--DN	M---Q-----	----YT-YR-	L-D--I---N	--KD--A--S	-KN-----D-	-----I-	-----
1146	H--QN	M-----	---NI-YT-	L-----	---A--S	--N--S---	S-----I-	-----
1147	E--QN	M-----	S---NT-YK-	L-I--I---E	--KD--A--S	-KN-----D-	-----I-	-----
1149	E--DN	M---Q-----	S---NT-YQ-	LAD--S---	--K--A--S	-NN-----S-	S-----I-	-----
1151	E--DN	M---Q-----	----GT-YR-	L-D--T---Q	--KD--A--S	-KN--T----	-----I-	-----
1152	D--DN	M---Q-----	S---DT-YR-	L-D-----	--KD--A--S	-NN-----	-----I-	-----
1153	D---N	M---Q--K--	S--SNT-YK-	L-D--I---Q	--KD--A--	-QN--S--D-	-----I-	-----
1155	T--DN	M---Q-----	S---HT-YR-	L-D-----	--KD--A--S	-KN--S--S-	S-----RI-	-----
1156	E--EN	M---Q-----	D-H-DI-YR-	L-K-----	---D--A--S	-KN--S--D-	SK-----I-	-----
1160	D--EN	M---Q-E---	S---DT-YR-	L-V-----	--KD--A--S	-NN--S----	-----I-	-----
1162	S--DN	M---Q-----	S---DT-YR-	L-D-----	--RD--A--S	-EN-----D-	S-----I-	-----

**B2 continue: env gp41 IDR amino acid sequence alignment**

B.FR.83.HXB2	QIWNH	TTWMEWDREI	NNYTSLIHSL	IEESQNQQEK	NEQELLELDK	WASLWNWFNI	TNWLWYIKLF	IMIVGGL
1163	D---N	M---Q-----	----D--YA-	L-----	--KD--A--S	-KN--S--S-	-----RI-	-----
1165	E--DN	M---Q-----	-----I-YR-	L-----N	--KD--A--S	-E-----S-	-K-----I-	-----
1169	D---N	L---Q-----	S---DT-YR-	L-D-----R	--KD--A--N	-KN-----D-	SR-----I-	-----
1172	D--DN	M---Q-----	S-H-DT-YR-	L-D-----Q	--KD--A--S	-NN-----P	-----I-	-----
1173	D--DN	M---Q-----	----ET-FR-	L-D-----	--KD--A---	-NN-----	-K-----I-	-----
1174	E--DN	M---Q-----	S---GI-Y--	L-D-----	--KD--A---	-NT--S--D-	-----I-	-----

147





**B3: env gp120 V3 amino acid sequence alignment**

```

B.FR.83.HXB2      LLNGSLAEEE VVIRSVNFTD NAKTIIVQLN TSVEINCTRP NNNTRKRIRI QRGPGRAFVT
C.BW.96.96BW0502  -----V-KG- II---E-L-N ---I----- KP-K-V-V-- -----SV-- ..---QT-YA
C.IN.95.95IN21068 -----GG II---E-L-N -V-----H-- QP---M---- D-----S--- ..---QT-YA
C.ET.86.ETH2220   -----I--G- TI--FE-L-N ---I----- E----T---- S----ES--- ..---QT-YA
C.BR.92.92BR025   -----II---K-L-- -V-----H-- E-----S--- ..---Q--YA
1002              -----G II---K-L-- -G-----H-- KAAK-V---- Q-----SV-- ..---QTIYA
1003              -----G II---E-L-T -I-----E-- E--K-V---- -----S--- ..---Q--YG
1005              -----II---K-LSN --Y---HF- E-L--V-S-- S-----SV-- ..---QT-YA
1006              -----G II---E-L-- -T-----H-- E--Q-V---- -----SV-- ..---QT-YA
1008              -----D II---E-L-N -V-----H-- E-I--E--- S-----SM-- ..---QT-YA
1009              -----II---E-L-N -----HF- E---V--- G-----S--- ..---QT-YA
1010              -----GD II---E-L-N -----H-- E-I--V---A S---QS--- ..---Q--FA
1011              -----II---E-LAN -----H-- E-----S--- ..---Q--YA
1012              -----II---E-L-N -----H-- K---VW--- -----SM-- ..---QT-YA
1013              -----II---E-L-N -V-----H-- Q---K-I-- G---RSM-- ..---QT-YA
1017              -----G II---E-L-N -----HF- E-ID-V--- -----S--- ..---QT-YA
1018              -----G II-K-E-L-N -----H-- E---V--- -----S--- ..---QT-YA
1021              -----II---K-L-N --NI--H-- D---V--- -----TS--- ..---QT-YA
1023              -----G II---E-L-N -----H-- E--G-E--- -----S--- ..---Q--YA
1024              --F--R-A-D -I-S-IYL-A DTI---H-- Q---V--- A---QS--- ..---Q--YA
1025              -----G II---E-M-N -----AH-- E--A-V--- -----S--- ..---QT-YA
1026              -----II-K-E-LG- -T-----Q---V--- -----S--- ..---QT-YA
1027              -----KG- II---E-L-N ---I--H-- E--N-E-Y-- -----SV-- ..---QT-YA
1031              -----K- II---E-P-- -V-----L-- E--NVV--S- -----S--- ..---QTLA-
1033              -----G II---E-L-- -F-----H-- Q---V--- -----S--- ..---QT-YA
1034              -----G II-S-E-L-A -V-----E---E-V-- T-----S--- ..---QT-YA
1037              -----GG II---E-L-N -----H-- E-I--A-VS- -----SV-- ..---QV-YA
1038              -----II---E-L-N -----ET-L----- G---QS--- ..---QT-YA
1040              ---A--P--- IIVS-EILPT R--P--H-- K---V--- A--Q--SM-- ..---QT-YA
1041              -----AK- -I---S-M-A -S-----H-H ED--V-A-- G--T-NV-- ..---QPSFA
1042              -----II---E-L-N T-----KP---V---- -----QSV-- ..---QT-YA
1043              -----GD II---E-L-- -V-----H-- E---V--- G-----S--- ..---Q--YA
1044              ..... -H-- E---V-A-- G-----S--- ..---QM-FA
1045              -----K- II---E-L-- -V-----H-- Q-I--V--- S-----SV-- ..---QT-YA

```

**B3 continue: env 120 V3 amino acid sequence alignment**

```

B.FR.83.HXB2      LLNGSLAEEE VVIRSVNFTD NAKTIIVQLN TSVEINCTRP NNNTRKRIRI QRGPGRAFVT
1047      -----G- II-K-EYI-N -V-----L-- K--A----- G-----G--- ..---QVVYA
1048      -----II---E-L-- ---I---H-- E----V---- G----QS--- ..---QT-YA
1049      -----D II---E-L-N -V-----H-- E--K-V-V-- -----SV-- ..---QT-YA
1050      -----II---E-L-N -V-----F- K----E----- SV-- ..---QT-YA
1052      -----II--FE-LA- -T-----H-- E-I----- G----TSM-- ..---QT-YA
1055      -----II---E-L-N -----HF- E----V-S-- -----SV-- ..---QT-YA
1057      -----II---K-LA- -----H-- E-I--V-A-- R----TSV-- ..---QT-YA
1058      -----II---E-L-N -----P-- E--A-V---- G----SV-- ..---QT-YA
1059      ..... -V-----H-- E----V-I-- G-----G--- ..---QT-YA
1061      -----K- II---K-M-- -S-----H-- EK-K-V---- G----SV-- ..---QT-YA
1062      -----G- II---E-L-N -----H-- D-----S-----SV-- ..---QT-YA
1064      ---D---A-- II---E-L-N -----F- KP--V---- G---RSV-- ..---Q--Y-
1068      -----K- II---E-L-N -----H-- E-I--K-I-- G----QS--- ..---QT-FA
1073      -----P--GG II---E-L-N D-----H-- E----V---- S----SM-- ..---QT-YA
1075      -----I---K-L-- -T-----H-- ET--T-I-- D-----G--- ..---QT-YA
1076      ---N---AK- II---E-L-N -----F- K----V---- G---RSV-- ..---QP-Y-
1077      -----II---E-L-- -V----- Q--D----- S----- ..---Q--YA
1078      -----GG I---E-L-- -----Q NP--V-A-- -----TSV-- ..---QT-YA
1079      -----G- II---E-L-N ---I---H-- E--K-T---- G----SV-- ..---QT-YA
1083      -----D II-K-E-L-N -----H-- N-I--V---- G-----S--- ..---Q--YA
1084      ---R----- I---E-L-N -V---G--K EP-K-G---- -----ESV-- ..---QT-YA
1089      -----II---E-L-N -----HF- E----V---- -----S--- ..---QT-YA
1090      ----VT-G- IG--FGYV-I TV-A-T-H-- E-I--T-V-- -----SV-- ..---QV-YA
1094      -----II---E-L-N -----H-K EP--V---- -----SV-- ..---QTWYA
1096      ----AA--AV II--F-TV-- T-I--LLH-- E---A----- TTK-----GT ..---QT-SA
1097      -----G- II---E-L-- -----H-- E--Q----- G-----S--- ..---T-YA
1098      -----II---E-L-N -----F- R---V---- G---RSV-- ..---Q--Y-
1099      ..... N T-----H-- E----V---- -----S--- ..---QT-YA
1100      -----K II---E-L-- ---I---H-- E----V---- -----S--- ..---QT-YA
1101      -----K- II---E-L-N -----F- EP--V-I-- G---RSV-- ..---Q--Y-
1102      -----K- I---E-L-- -----H-- K----L-V-- G-----S--- ..---QT-Y-
1104      ---S-V-G- IIS--D-T-- -T-----H-- E--D-V---- G-T--S--- ..---QT-YA
1106      -----II---E-I-N ---I----- EP--T---- S-----SV-- ..---QT-YA
1108      ---S-A---- IIR--D-TA- -.N---IL-- K--A-V---- G-----S-G- ..---QT-YA

```

**B3 continue: env 120 V3 amino acid sequence alignment**

```

B.FR.83.HXB2      LLNGSLAEEE VVIRSVNFTD NAKTIIVQLN TSVEINCTRP NNNTRKRIRI QRGPGRAFVT
1110      ----- II---E-L-N -----HFD EA---A----- -----RS--- ..---QS-SA
1112      -----GG II---E-L-N S---T--H-- E---V----- -----SV-- ..---Q--YA
1113      ..... .. ...---H-- E----- GH----SM-- ..---QT-YA
1114      -----KG- II---E-L-N ---I---H-- E--Q-T-I-- G---SV-- ..---QT-YA
1115      ----- II---E-L-A -T-----H-- E---S--- G---SM-- ..---QP-YA
1116      -----G- II---EDM-N -V-----H-T E--T----- G---S--- ..---QT-YA
1118      -----K- I---E-L-- -----H-- K---L-V-- G---S--- ..---QT-Y-
1119      ----- IR---E-L-- -V-----H-- E---T-I-- G---RS--- ..---Q--YA
1120      -----D I---E-L-N -V-----H-- E--K-V--- -----SM-- ..---QT-FA
1121      -----G- II--CE-L-N -V-----P-- E--A----- D---S--- ..---QV-YA
1123      -----G- I---E-L-N -----H-- K---V----- S--- ..---QT-YA
1125      -----G- IM---E-L-- -V-----H-- E---K-Q-- G---S--- ..---QT-YA
1127      -----A-- IM---E-L-- -V-----H-- E--A-V-V-- -----QG-- ..---QT-Y-
1129      -----G- II---E-M-- -V-----H-- E--A----- D---SM-- ..---QV-YA
1131      -----K- II---E-L-- -V-----H-- E-I--T--- -----S--- ..---QT-YA
1132      -----G II---K-L-- SV-----H-- E--V----- G---S--- ..---QT-YA
1134      -----G II---E-L-- -----H-- E-I--T--- -----S--- ..---QV-YA
1135      ..... .. --Q---H-- E---E-I-- G---SV-- ..---QT-FA
1137      -----K- II---E-L-N -----H-- E---V----- -----SV-- ..---Q--YA
1138      -----G- LI---E-L-N -----H-- E---V----- -----RSM-- ..---QV-YA
1140      ----- II---E-L-- -----H-- K-IS-V----- S--- ..---QT-YA
1141      ----- II---E-L-N -V-----HF- ST-T-E---- S-----S--- ..---QS-YA
1142      ----- II---E-L-- -V-----H-- E---E----- -----S--- ..---Q--YA
1143      ----- II---E-L-N -V-----H-- E---V----- -----S--- ..---Q--YA
1144      -----GG II---E-L-N -V-----H-- E-I--T-V-- -----SV-- ..---QV-YA
1146      -----K IT---E-L-N -----H-K DP--V----- G---SM-- ..---QT-YA
1147      -----GA II---E-L-N -V-----H-- E-I--T-V-- -----SV-- ..---QL-YA
1148      ----- II---E-L-- -----H-- K-IS-V----- -----SK-- ..---QT-YA
1151      ----- II---E-L-E -S-I---H-- E--P-V-V-- -----S--- ..---Q--YA
1152      -----G- II---E-L-N -V-----H-- E---E----- G---SV-- ..---QT-YA
1153      -----G- II---E-L-N -V-----H-- E---V--S--- -----RS--- ..---QT-SA
1154      ---N-Q-GGG II---E-L-H -G-----H-D ECI--L-V-- -----RSA-R ..---QV-DA
1155      -----D II-K-E-L-N -V-----HF- E--K-V-V-- -----RSV-- ..---QT-YA
1156      ----- -I---E-L-A -----H-- E---T----- -----S--- ..---Q--YA

```

**B3 continue: env 120 V3 amino acid sequence alignment**

```

B.FR.83.HXB2      LLNGSLAEEE VVIRSVNFTD NAKTIIIVQLN TSVEINCTRP NNNTRKRIRI QRGPGRAFVT
1157             ----- II---E-L-N -----FD A----V----- G-----S--- ..---Q--YA
1160             -----GV II---E-L-- -V-----R-- E--A----- -----RSR-- ..---QT-YS
1162             ----- II---E-L-- -----H-- E-----S--- ..---Q--YA
1163             .--- II---E-L-- -----H-- K--P-V----- -----S--- ..---QT-YA
1165             ----- II---E-L-N -----H-- E-----T-----S--- ..---QT-YA
1169             .....N ---I----- Q---V----- -----SM-- ..---QT-YA
1172             -----GGR -I--F--L-N -----H-- E-G--L----- C---S-S-S- ..---Q--YA
1173             -----G II---E-L-N -I-----H-- Q---V----- -----S--- ..---QT-YA
1174             -----G- IM---EDL-N SV-----H-- K---V----- -----SV-- ..---Q--YA
1175             -----D I--S-E-L-- -T-----H-- Q--T-V----- A----QS--- ..---Q--YA

```

60

```

B.FR.83.HXB2      IG.KIGNMRQ AHCNIS.RAK WNNTLKQIAS K
C.BW.96.96BW0502 T-EI--DI-- -Y-I-N-KTE --S--QGVSK .
C.IN.95.95IN21068 T-DI--DI-- -----ED- --E--QNVSK .
C.ET.86.ETH2220  T-DI--DI-- -----EE- --K--QKVKE .
C.BR.92.92BR025  T-EI--DI-- -----TA  --K--QEVGK .
1002             T-DI--DI-- -Y-----G.. --E--EGVKK .
1003             .MDI--DI-- -----ASN -TK--QRVSE .
1005             TNDI--DI-- -----G-- --R--Q-VGK .
1006             .NDI--DI-- -Y-----TQE --K--ERVKK .
1008             T-EI---I-- -----GN -STMMQRVSE .
1009             TNEI---I-- -----ED- --K--Q-VGK .
1010             TKDV--DI-- ---H---GD --EA--RVSR .
1011             TNDI--DI-- -----AD- --K--Q-VGK .
1012             T-DI--DI-- -----ERP --D---GVSE .
1013             T-DI--DI-- -----SGT -----EVVK .
1017             TNGI--DI-E -----T--Q- -----E-VKE .
1018             TN-I---I-- -Y-----KGN --K--ERVKE .
1021             T-EI--DI-- -----ED- --K--Q-VGK .
1023             TNDI--DI-- -----TE- -----RVRE .
1024             T-DI--DI-- -----T-KRN -TE--QKVNK .
1025             TNGI--DI-- -----KDE --K--ERVKT .
1026             T-GI--DI-- -----..G --D--EKVRE .

```




91

**B3 continue: env 120 V3 amino acid sequence alignment**

```

B.FR.83.HXB2      IG.KIGNMRQ AHCNIS.RAK WNNTLKQIAS K
1027              T-GI---I-- -Y-E-R-EKT --Q--DKVKK .
1031              N.NIK-DI-- -.....-... .....
```



```

1033              T-GIV--I-- -----ES- --G--QRVRE .
1034              T-EI---I-L -Y-----EE- --K--HRVSK .
1037              T-DI--DI-- -----AD- -IT--QREEE .
1038              T-DI--DI-E -----TD- --T--ERVKK .
1040              T-DI--DI-- -P-T---EKD --E--NNVRK .
1041              T-DI--DI-- -----EGR --K-YNGVKK .
1042              T-EI--DI-K -----N-KTL -----QEVEK .
1043              TDAI--DI-- -----V--K-R --QM-ERVKE .
1044              N.DI--DI-- -Y-FVN-GT- --K-FQ-VGK .
1045              T-EI--DI-- -----ES- -----Q-V-K .
1047              TNKV--DI-K -----KTQ --E----VRE .
1048              R-DI--DI-R -Y-A-N-ES- --I--QRVSE .
1049              T-EI--DI-- -----KGG --KA-EGVRE .
1050              T-EI--DI-R -----N-IT- -----EKVKK .
1052              T-DI--DI-- -Y-----TS- -HT--ERVKK .
1055              TDGI---I-- -----AE- --E--Q-VGK .
1057              T-DI--DI-- -----.....
```

**B3 continue: env 120 V3 amino acid sequence alignment**

```

B.FR.83.HXB2      IG.KIGNMRQ AHCNIS.RAK WNNTLKQIAS K
1090              TDDI--DI-- -----AD- --TN-QRVSK .
1094              T-EI--DI-- -----AKN --K--Q-VGE .
1096              T-DI-R-I-- -----NW- -TLP--GVSK .
1097              T-DI---I-- -----G-  --V--QRVKG .
1098              T-EV--DT-K -----N-ASL --E--QGVGK .
1099              TNVIV-DIK- -R-I-G-GRR C-T--QLVEK .
1100              TNGI--DI-K -----KSN -TR----V-E .
1101              A-EV--DT-R -----N-ASL --E--QGVRK .
1102              T-DI--DI-- -----KSE --K--ESVKK .
1104              T.AI--DI-E ----VN--SL --A--E-VKK .
1106              T-DI--DI-- .....-... ..
1108              T.DI--DI-E ----VN--SL --A--E-VKK .
1110              N.KI--DI-- P----N-ANT --T--Q-V-K .
1112              T-DI--DI-- -----T--SN --E--QVRK .
1113              T-EI--GI-- -----ES- -----SRVSE .
1114              T-EI--DI-- -----EKN --D--QRVSE .
1115              T-EI--GI-- -----AS- -----LRVSK .
1116              H.EV--DI-- -----N-GTQ --Q--A-VKE .
1118              T-DI--DI-- -----KSE --K..... .
1119              T-DI--DI-- -Y--L--ES- --Q--QKVGK .
1120              T-DI---I-- -----EKE --E--YRVSK .
1121              NNDI--DI-- ---I---GEQ --R--G-VEE .
1123              T-AI--DI-E -----NGTI --T--EMVKK .
1125              .NDI--DI-- ---I---.G -KT--EKVRK .
1127              N.NI--DI-- -----T-KGQ -K---EKESK .
1129              TNSI--DI-- -----ERQ --D--Q-VRE .
1131              T-DIV-DI-- -----NKN -TTA-QRVSE .
1132              TNDI--DI-- ----L---QA --K-IE--RK .
1134              TNDI--DI-- -Y-----TVS --D--QKVVK .
1135              T-AI--DI-K -----N-QT- --T--ERVKR .
1137              TNDI--DI-- -Y-----EE- --K--Q-VVK .
1138              T-EI--DI-K -Y-----AE- --K--EMVRE .
1140              TNAI--DI-E -----KKE -QT--E-EGR .
1141              T-SI---I-- -----.-SKQ -YR--QRVKE .

```

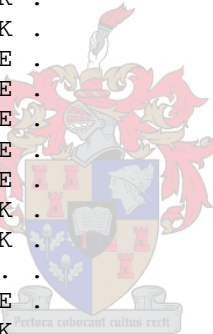


**B3 continue: env 120 V3 amino acid sequence alignment**

```

B.FR.83.HXB2      IG.KIGNMRQ AHCNIS.RAK WNNTLKQIAS K
1142              TKGI--DI-- -----ES- --K--Q-VGK .
1143              T-EI---I-- -----EKD --K--Q-VGK .
1144              TDDI--DI-- -----AD- --TN-QRVSE .
1146              T-DIT-DI-K -----EE- --K--QEVGK .
1147              T-GI--DIT- -----ADN -DDKFQRVSK .
1148              TNAI--DI-E -----KKE -QK--E-VKG .
1151              TNDI--DI-- -----N-KTS --K--T-VKE .
1152              T-DI--DI-- -----V--KEQ -DR--RMVRA .
1153              T.DI--DI-- -----T-- --T--EGVRE .
1154              TDAIR-DI-- ---H---ADR -D-N-QR-SK .
1155              T-DI---I-- -Y-----ED- --K--Q-VTK .
1156              TNAI--DI-- -----ER --K--E-VKE .
1157              T-DI--DI-- -----N-SSQ -----Q-RV-E .
1160              T-DI--DI-- -Y-E-N--E- --G--ALGKE .
1162              TKDI--DI-- -----KH --E--ERVKE .
1163              T-DI--DI-- -----GEE --T-ME--KE .
1165              T-AI--DI-- ---T-N-ETT --K--Q-VGK .
1169              T-GI--DI-- -----EKI --K--NEVGK .
1172              T-DI--DI-- -----.....
1173              .NDI--DI-- -----KTS -STA-RNVTE .
1174              T-DIT-DI-- -----AD- --E--Q-VGK .
1175              T-DI---I-- -----T--R- -TEA-QNVNN .

```



**B4: pol amino acid sequence alignment**

B.FR.83.HXB2-LAI-IIIB-BRU	IKIGGQLKEA	LLDTGADDTV	LEEMSLPGRW	KPKMIGGIGG	FIKVRQYDQI	LIEICGHKAI	
C.BR.92.BR025-d	--V-----	-----	---IK---N-	-----	-----	-----K---	
C.ET.86.ETH2220	-----	-----	---IN---K-	-----	-----	I-----K---	
C.BW.96.96BW0502	--V---I---	-----N--	---IN---K-	-----	-----	V-----K---	
C.IN.95.95IN21068	-RV---I---	-----	---V---K-	R-----	-----EE-	P-----K---	
1039	--V---I---	-----	---ID-----	-----	-----	-----K---	
1151	-RV---I---	--A-----	---L---K-	-----	-----	-----K---	60
B.FR.83.HXB2-LAI-IIIB-BRU	GTVLVGPTPV	NIIGRNLLTQ	IGCTLNFPIS	PIETVPVKLK	PGMDGPKVKQ	WPLTEEKIKA	
C.BR.92.BR025-d	-----	-----M---	L-----	-----	-----	-L-----	
C.ET.86.ETH2220	-----	-----M---	L-R-----	-----	-----	-----	
C.BW.96.96BW0502	-----	-----M---	L-----	-----	-----	-T-----	
C.IN.95.95IN21068	-----	-----M---	L-----	-----	-----	-----	
1039	-----	-----M---	L-----	T-----Q-	-----I-	-----	
1151	-----	-----M---	L-----	-----	-----I-	-----	120
B.FR.83.HXB2-LAI-IIIB-BRU	LVEICTEMEK	EGKISKIGPE	NPYNTPVFAI	KKKDSTKWRK	LVDFRELNKR	TQDFWEVQLG	
C.BR.92.BR025-d	-TA--D---R	---T-----	-----	-----	-----	-*-----	
C.ET.86.ETH2220	-TA--E---Q	---R-----	-----	-----	-----	-----	
C.BW.96.96BW0502	-T--E-----	---T-----	-----	-----	-----	-----	
C.IN.95.95IN21068	-TA--D-----	---T-----	-----I-	-----	-----	-----	
1039	-TA--E-----	-----	-----	-----	-----	-----	
1151	-TA--E-----	-----	-----	-----	-----	-----	180
B.FR.83.HXB2-LAI-IIIB-BRU	IPHPAGLKKK	KSVTVLDVGD	AYFSVPLDED	FRKYTAFTIP	SINNETPGIR	YQYNVLPQGW	
C.BR.92.BR025-d	-----	-----	-----G-	-----	-----	-----	
C.ET.86.ETH2220	-----	-----	-----G-	-----	-T-----	-----	
C.BW.96.96BW0502	-----	-----M--	-----G-	-----	-----	-----	
C.IN.95.95IN21068	-----	-----	-----Y-	-----	-----	-----	
1039	-----	-----	-----Y-G	-----	-----S---	-----	
1151	-----	-----	-----G-	-----	-----	-----	240

**B4: pol amino acid sequence alignment**

B.FR.83.HXB2-LAI-IIIB-BRU	KGSPAIFQSS	MTKILEPFRK	QNPDIIVIQY	MDDLIVGSDL	EIGQHRTKIE	ELRQHLLRWG	
C.BR.92.BR025-d	---S-----	T-----A	---E-I----	-----	-----A---	---E---K--	
C.ET.86.ETH2220	---P-----	-PQ-----	P--E-----	-----	-----AP--	---E---K--	
C.BW.96.96BW0502	-----	-----L	---E-----	-----	-----R-AQ--	---E---K--	



C.IN.95.95IN21068  
1039  
1151

-----N- --R-----A ---E----- -----A--- ---K-----  
-----C- -----A ---E----- -----E---K--  
-----T K--E----- -K---KA--- ---A---K--

300

B.FR.83.HXB2-LAI-IIIB-BRU  
C.BR.92.BR025-d  
C.ET.86.ETH2220  
C.BW.96.96BW0502  
C.IN.95.95IN21068  
1039  
1151

LTTTPDKKHQK EPPFLWMGYE LHPDKWTVQP IVLPEKDSWT VNDIQKLVGK  
F----- -Q-----  
F----- -Q-----  
F----- -Q--D-----  
F----- -Q-----  
F----- -E-- -Q-----  
F----- -Q-----

