

**Sequence analysis of an HIV-1 subtype C
acutely infected cohort from Durban, South
Africa**

Submitted by

Stanley Carries

In fulfillment of the requirements for the degree of
Master of Medical Science in Virology

In the

Faculty of Health Sciences

School of Laboratory Medicine and Medical Sciences


University of KwaZulu-Natal

2018

Preface

The experimental work described in this dissertation was carried out at the KwaZulu-Natal Research Innovation and Sequencing Platform (KRISP), Nelson R. Mandela School of Medicine, University of KwaZulu-Natal, Durban, from April 2017 to November 2018 under the supervision of Dr Michelle Gordon.

This study represents original work by the author and has not otherwise been submitted in any form for any degree or diploma to any other University. Where use has been made of the work of others, it is duly acknowledged in the text.


Signed:  Date: 14/01/2019
Stanley Carries (Student)

Signed:  Date: 14/01/2019
Dr Michelle Gordon (Supervisor)

Declaration

I, Stanley Carries, declare that:

- i. The research reported in this dissertation, except where otherwise indicated, is my original work.
- ii. This dissertation has not been submitted for any degree or examination at any other university.
- iii. This dissertation does not contain other persons' data, pictures, graphs or other information unless specifically acknowledged as being sourced from other persons.
- iv. This dissertation does not contain other persons' writing unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:
 - a. Their words have been re-written but the general information attributed to them has been referenced;
 - b. Where their exact words have been used, their writing has been placed inside quotation marks and referenced.
- v. This dissertation does not contain text, graphics or tables copied and pasted from the internet, unless specifically acknowledged, and the source being detailed in the dissertation and in the reference sections.

Signed:  Date: 14 / 01 / 2019
Stanley Carries (Student)

Signed:  Date: 14 / 01 / 2019
Dr Michelle Gordon (Supervisor)

Ethical Approval

Ethical approval for this study was obtained from the Biomedical Research Ethics Committee of the Nelson R. Mandela School of Medicine, University of KwaZulu- Natal (BE521/17).

Acknowledgments

I firstly would like to thank the Lord God Almighty for the wonderful opportunity that He has granted me to share a little of His infinite knowledge and understanding through this research. Thank you Father for Your provision and for granting me strength to journey on.

This journey would not have reached its conclusion were it not for the following people, to whom I am greatly indebted:

- My children, who unwittingly bore the brunt of my distraction as I focused on completing this work. I love you guys infinity[∞]
- Phumzile Khumalo, for your constant support and for unselfishly giving of yourself. Your strength and optimism spurred me on through the tempest.
- My supervisor, Michelle Gordon, for your guidance and for believing in me. Through you, I have learnt that clarity comes with diligence and pursuance.
- The students at the KwaZulu-Natal Research and Innovation Sequencing Platform (KRISP), it has been a pleasure working with all of you. It was comforting to know that we were all “licking our wounds” in one form or another.

I would also like to express my gratitude to the College of Health Sciences at the University of KwaZulu-Natal for the financial support to conduct my research.

Abstract

The Human Immunodeficiency Virus is a global public health concern. The Joint United Nations Programme on HIV/AIDS estimated that 36.9 million people were infected with HIV globally at the end of 2017. Almost 20% of these resided in South Africa, making this the highest global HIV burden held by any one country. It is thus important that HIV infection be detected early as this may have important implications in the control of the pandemic. The early recognition of acute HIV infection could present early treatment options that could alter the natural history of the disease, or even eliminate infection. Detecting acute infection early could also provide a unique opportunity to understand HIV transmission and pathogenesis, including early host-virus interactions. In the present study, blood samples were collected from 18-23 year old HIV-1 subtype C acutely infected women from Umlazi Township in KwaZulu-Natal, South Africa, that had participated in a study called Females Rising through Education, Support and Health (FRESH). Eleven blood samples from this cohort, collected within 24 hours of onset of plasma viremia, were used for this study. The aim of the present research was to identify sites within *pol* that were experiencing positive selective pressure and the likely implications of these mutations on viral functional domains and host cytotoxic T-lymphocyte (CTL) epitopes. The study also sort to observe the loss of drug resistant mutations (DRM) in the viral sequences of participants who had multiple timepoints and to correlate mutation loss to structural changes. Datamonkey and Phylogenetic Analysis by Maximum Likelihood (PAML) were used to detect positively selected sites. Putative functional domains were detected using Prosite and CTL epitopes were identified using the Los Alamos Molecular Immunology Database. Ancestral reconstruction was performed using PAML and Bayesian Evolutionary Analysis by Sampling Trees (BEAST) was used to calculate the time to the most recent common ancestor. Altogether 16 unique positively selected sites were identified in this cohort. Putative functional domains were highly conserved in protease, while positive mutations in reverse transcriptase resulted in either a loss of functional domains in conserved regions or in the gain of functional sites in non-conserved regions. Owing to the important role that protease plays in viral maturation and infectivity, mutations within these conserved regions could possibly lead to defective viral particles with reduced viral infectivity. The K103N in reverse transcriptase, observed in one participant, was the only DRM inherited from its common ancestor. The major limitation of this study was the small sample size.

Table of Contents

Preface	ii
Declaration	iii
Ethical Approval	iv
Acknowledgments	v
Abstract	vi
Table of Contents	vii
Table of Figures	x
List of Tables	xi
Appendices	xii
Abbreviations and Acronyms.....	xiii
CHAPTER 1: : Literature Review.....	1
1.1 Introduction.....	1
1.2 Origins of HIV/AIDS	1
1.3 HIV-1 Classification	3
1.4 Diversity and Geographical Spread of HIV-1 Group M.....	4
1.5 HIV-1 Structure and Genome	6
1.6 HIV-1 Lifecycle	8
1.7 Acute HIV Infection.....	10
1.8 Acute Infection and the FRESH cohort	12
1.9 Development of HIV Drug Resistance	14
1.10 Types of HIV Drug Resistance Mutations	16
1.10.1 HIV drug resistance in Reverse Transcriptase	17
1.10.2 HIV drug resistance in Protease.....	19
1.11 Detection of Drug Resistance Mutations and Testing for Minority Variants	19
1.11.1 Ultradeep PyroSequencing (UDPS).....	21
1.12 Aim and Objectives of the Study	21
1.12.1 Aim	21
1.12.2 Objectives.....	21

CHAPTER 2: Bioinformatics Analysis Toolkit.....	23
2.1 Introduction.....	23
2.2 HyPhy package in Datamonkey for Detecting Sites under Positive Selective Pressure	23
2.3 Phylogenetic Analysis.....	26
2.3.1 Overview of Phylogenetic Trees.....	26
2.3.2 Methods for Estimating Phylogenetic Trees	28
2.3.3 Models of Nucleotide Substitution.....	32
2.3.4 Phylogenetic Analysis by Maximum Likelihood (PAML).....	37
2.3.5 Bayesian Evolutionary Analysis by Sampling Trees (BEAST).....	38
2.4 HIV Analysis Online Tools.....	43
2.4.1 Stanford University HIV Drug Resistance Database	43
2.4.2 Los Alamos HIV Database.....	44
2.4.3 Prosite.....	44
 CHAPTER 3: Detection of Positive Selection Pressure in Acute Phase HIV Positive Treatment	
Naïve Sequence Alignments	46
3.1 Introduction.....	46
3.2 Sample Description	46
3.2.1 Sequence Quality Control	47
3.2.2 Sequence Alignment	47
3.3 Methods.....	48
3.3.1 Best-fit Model of Substitution Selection using j-Modeltest.....	48
3.3.2 Tree Construction using PAUP*	49
3.3.3 Phylogenetic Analysis by Maximum Likelihood (PAML) for detection of positively selected sites.....	50
3.3.4 Datamonkey specifications for positive selection.....	51
3.3.5 Online Analyses of positively selected sites	51
3.4 Results.....	52
3.4.1 Best-fit Models of Substitution Selected by j-Modeltest and PAUP* Tree Construction.....	52
3.4.2 Sites under Positive Selection Pressure detected by PAML	54
3.4.3 Sites under Positive Selection using Datamonkey	56
3.4.4 Positive selection at functional sites in Pr and RT	59
3.4.5 Positively selected sites in HIV subtype C epitopes	61
3.5 Discussion	62

CHAPTER 4: Ancestral Sequence Reconstruction using PAML and tMRCA estimation using BEAST	68
4.1 Introduction	68
4.2 Method	68
4.2.1 tMRCA tree construction using BEAST	68
4.2.2 Ancestral reconstruction using PAML	69
4.3 Results	70
4.3.1 Analysis of BEAST results in Tracer	70
4.3.2 Trees reflecting tMRCA and evolutionary rates from reconstructed ancestral sequences	73
4.4 Discussion	79
CHAPTER 5: General Discussion	83
CHAPTER 6: Conclusion	86
Bibliography	87
Appendices	102

Table of Figures

Figure 1.1: Worldwide prevalence of HIV-1 group M subtypes and CRF	5
Figure 1.2: HIV-1 structure and genome	7
Figure 1.3: HIV transmission and the establishment of HIV reservoirs	9
Figure 1.4: Progression of HIV-1 after infection	12
Figure 2.1: Basic tree structure and nomenclature	27
Figure 2.2: Transition and transversion substitution matrix.....	30
Figure 2.3: Hierarchy of nucleotide substitution models	34
Figure 3.1: Positive selection detection for Sanger and UDP Sequences	56
Figure 3.2: REL, MEME and iFEL positively selected sites in <i>pol</i> across F81, HKY85, TrN93 and REV models of substitution	57
Figure 3.3: Conserved putative functional sites in Pr and RT from the FRESH study cohort.....	59
Figure 4.1: Comparison of Marginal Densities of three runs of alignment 271 with associated summary estimates and trace plots for the posterior statistic	70
Figure 4.2: Comparison of Marginal Densities of three runs of alignment 271 with associated summary estimates and trace plots for the treeheight statistic	72
Figure 4.3: Time to the most recent common ancestor (tMRCA) for 036	74
Figure 4.4: Time to the most recent common ancestor (tMRCA) for 079	75
Figure 4.5: Time to the most recent common ancestor (tMRCA) for 267	77
Figure 4.6: Time to the most recent common ancestor (tMRCA) for 271	78

List of Tables

Table 1.1: Summary of HIV types and groups.....	3
Table 1.2: HIV RT and Protease Mutations for Drug-Resistance Surveillance.....	16
Table 2.1: Site by site selection models	24
Table 2.2: Classification of phylogenetic analysis methods and their strategies	29
Table 2.3: Some commonly used nucleotide models of substitution and summary of parameters..	33
Table 2.4: Associated programs used by BEAST	39
Table 2.5: General components of evolutionary models in BEAST	39
Table 3.1: j-Modeltest and PAUP* output for SS and UDP alignments.....	52
Table 3.2: Likelihood Ratio Test for Comparison of ‘best fit’ models of substitution for SS and UDP using standard Chi-squared distribution.....	53
Table 3.3: LRT for evidence of positive selection in SS and UDP alignments using standard Chi- squared distribution.....	54
Table 3.4: Likelihood values and parameter estimates under models of variable ω ratios among sites for Sanger and UDP alignments.....	54
Table 3.5: Positive selection using PAML.....	55
Table 3.6: Positive selection sites in Pr and RT for REL, MEME and iFEL for SS and UDP sequences and nearest known Drug Resistance Site	58
Table 3.7: Putative functional sites in FRESH cohort sequences	60
Table 3.8: Putative CTL epitopes and HLA's in Pr and RT from the FRESH cohort.....	61
Table 4.1: Summary statistics of interest from BEAST output.....	73

Appendices

Appendix 1: Codeml control file for positive selection	102
Appendix 2: Purifying selection sites identified by the REL method in Datamonkey.....	103
Appendix 3: Tracer of posterior statistic for 036	104
Appendix 4: Tracer of tree height statistic for 036	104
Appendix 5: Tracer of posterior statistic for 079	105
Appendix 6: Tracer of tree height statistic for 079	105
Appendix 7: Tracer of posterior statistic for 267	106
Appendix 8: Tracer of tree height statistic for 267	106
Appendix 9: Turnitin Report.....	107

Abbreviations and Acronyms

Acronym	Description
AA	Amino Acid
ACT	Auto Correlation Time
AIC	Akaike Information Criterion
AICc	Corrected Akaike Information Criterion
AIDS	Acquired Immunodeficiency Syndrome
ART	Antiretroviral Therapy
ASPCR	Allele Specific Polymerase Chain Reaction
AVA	Amplicon Variant Analyser
BEAST	Bayesian Evolutionary Analysis by Sampling Trees
BEAUti	Bayesian Evolutionary Analysis Utility
BEB	Bayes Empirical Bayes
BIC	Bayesian Information Criterion
CCR5	Chemokine Receptor Type 5
CDC	Centre for Disease Control
CD8 ⁺ /CTL	Cytotoxic T-Lymphocyte
cDNA	Complimentary DNA
CI	Confidence Interval
CKII-P	Casein Kinase II Phosphorylation
codeml.ctl	Codeml Control File
CRFs	Circulating Recombinant Forms
CXCR4	Chemokine Receptor Type 4
<i>dN</i>	Nonsynonymous
DNA	Deoxyribosenucleic Acid
dNTP	Deoxy-nucleotide triphosphates
DRMs	Drug Resistance Mutations
<i>dS</i>	Synonymous
DT	Decision Theory
ESS	Effective Sample Sizes
F81	Felsenstein Model
FEL	Fixed Effects Likelihood
FIs	Fusion Inhibitors
FRESH	Females Rising through Education, Support and Health
GTR	General Time Reversible

HIV	Human Immunodeficiency Virus
HIVdb	HIV Database
HKY85	Hasagawa, Kishino and Yano 85
HLA	Human Leukocyte Antigen
hLRT	Hierarchical Likelihood Ratio Tests
HPD	Highest Posterior Density
HTLVIII	Human T Lymphotropic Virus Type III
HTUs	Hypothetical Taxonomic Units
HyPhy	Hypothesis Testing Using Phylogenies
iFEL	Internal Branch FEL
IN	Integrase
INIs	Integrase Inhibitors
JC69	Jukes and Cantor
K2P	Kimura 2 Parameter
KS	Kaposi's Sarcoma
LAV	Lymphadenopathy Associated Virus
LRT	Likelihood Ratio Test
LTR	Long Terminal Repeat
MCMC	Markov Chain Monte Carlo
MEME	Mixed Effects Model of Evolution
ML	Maximum Likelihood
NC	Nucleocapsid
NEMs	Nucleotide Excision Mutations
NJ	Neighbour Joining
NNI	Nearest Neighbor Interchange
NNRTIs	Non Nucleoside Reverse Transcriptase Inhibitors
NRTIs	Nucleotide Reverse Transcriptase Inhibitors
OPV	Onset of plasma viremia
ORFs	Open Reading Frames
OTUs	Operational Taxonomic Units
PAML	Phylogenetic Analysis by Maximum Likelihood
PCP	<i>Pneumocystis Carinii Pneumonia</i>
PCR	Polymerase Chain Reaction
PIs	Protease Inhibitors
PID	Participant Identifier
PKC-P	Protein Kinase C Phosphorylation

PK-P	Protein Kinase Phosphorylation
PMTCT	Prevention of Mother to Child Transmission
PR	Protease
PUMs	Primer Unblocking Mutations
REL	Random Effect Likelihood
RT	Reverse Transcriptase
RT	Reverse Transcriptase
RTIs	Reverse Transcriptase Inhibitors
SC	Standard Cloning
SDRMs	Surveillance Drug Resistance Mutations
SGS	Single Genome Sequencing
SIV	Simian Immunodeficiency Virus
SLAC	Single Likelihood Ancestor Counting
SPR	Subtree Pruning and Regrafting
SS	Sanger Sequences
ssRNA	Single Stranded Ribonucleic Acid
SYM	Symmetrical Model
TAMs	Thymidine Analogue Mutations
TAR	Transactivation Response Element
TDR	Transmitted Drug Resistance
TIM	Transition Model
tMRCA	time to Most Recent Common Ancestor
TP	Time Point
TVM	Transversion Model
UDPS	Ultra Deep Pyrosequencing
UNAIDS	Joint United Nations Programme on HIV/AIDS
URF	Unique Recombinant Form
URLN	Uncorrelated Relaxed Log Normal
VMMC	Voluntary Medical Male Circumcision
WHO	World Health Organization
wt	Wild type

CHAPTER 1:

Literature Review

1.1 Introduction

The Human Immunodeficiency Virus (HIV) is a global public health concern. The Joint United Nations Programme on HIV/AIDS (UNAIDS) estimated that 36.9 million people were infected with HIV at the end of 2017 globally (UNAIDS, 2018a)(UNAIDS, 2018a). UNAIDS also estimated that since the start of the pandemic, 35.4 million people had died from AIDS-related illnesses. Regionally, Eastern and Southern Africa (ESA) (as per UNAIDS regional divisions) had the highest number of people living with HIV (approximately 19.6 million HIV positive people). Of the 1.8 million new HIV infections reported globally in 2017, the ESA region had the highest new infection rate (UNAIDS, 2018a). South Africa was estimated to have 7.2 million people living with HIV at the end of 2017 (UNAIDS, 2018b)(UNAIDS, 2018b). On the global arena, this ranks South Africa as having the highest number of people infected with HIV.

The HIV infection rate in South Africa decreased by 49% and AIDS-related deaths decreased by 29% since 2010 (UNAIDS, 2018b). The decrease in the South African HIV mortality rate can be attributed to several initiatives introduced by its government. These include prevention programmes such as prevention of mother to child transmission (PMTCT), condom use and distribution, voluntary medical male circumcision (VMMC), HIV education and HIV awareness. In addition to the preventative measures, the South African government has also introduced what is reputed to be the largest antiretroviral therapy (ART) rollout programme globally (AVERT, 2016; Steegen *et al.*, 2016). However, regardless of the government's investment of R14.2 million over the three years spanning 2015 to 2018 (Mapumulo, 2016), only 48% of the HIV infected population were receiving ART (AVERT, 2016).

1.2 Origins of HIV/AIDS

Acquired Immunodeficiency Syndrome (AIDS) was a term first used in 1981 by the Atlanta based Centre for Disease Control (CDC) to describe a group of disease entities observed in patients who presented with severely compromised cell-mediated immunity, rare malignancies and opportunistic infections (Center for Disease Control, 1981; Friedman-Kien, 1981; Gottlieb MS *et al.*, 1981; Marx, 1982). At the time, disease prevalence was

mainly observed among the male homosexual community, which presented with increasing levels of Kaposi's Sarcoma¹ (KS), and *Pneumocystis carinii* pneumonia² (PCP). Shortly afterward, prevalence was also detected among intravenous drug users, blood transfusion recipients, sexual partners, and children. This suggested that the aetiological agent for the condition was likely transmitted by body fluids (Freed, 2007). In 1983, Dr. Luc Montagnier and Dr. Francois Barre-Sinoussi from the Pasteur Institute in France were able to isolate a retrovirus believed to be the cause of AIDS, which at that time was called the Lymphadenopathy-Associated Virus (LAV) (Barre-Sinoussi *et al.*, 1983). The following year, Dr. Robert Gallo isolated a virus called the Human-T-Lymphotropic Virus Type III (HTLV-III) (Gallo *et al.*, 1984). Unbeknown to both sets of researchers, the LAV, and HTLV-III were, in fact, the same virus and was later named the Human Immunodeficiency Virus (HIV). In 1986, HIV was then established as the causative agent of AIDS and was renamed HIV type 1 (HIV-1) to distinguish it from HIV-2, a related and less prevalent AIDS-causing virus (Clavel *et al.*, 1986).

Although HIV was characterised in the 1980s, its presence in the human population dates back to the period between 1900 and the early 1920s (Korber *et al.*, 2000; Worobey *et al.*, 2008; Santos and Soares, 2010). The devastating effect that AIDS has had on humanity makes it one of worst pandemics in history, only paralleled to the Spanish influenza pandemic of 1918 estimated to have killed between 20 to 40 million people (Taubenberger *et al.*, 2001).

HIV belongs to the Lentivirus genus of the Retroviridae family. This genus includes both HIV-1 and HIV-2, as well as a significant number of simian immunodeficiency viruses (SIV) that infect different non-human primate species in the African continent (Locatelli *et al.*, 2008; Liégeois *et al.*, 2009; Santos and Soares, 2010). In their natural host, the SIV rarely causes the immune system to collapse. However, this is not the case extra-host. The origin of HIV in humans, for example, is traceable to multiple zoonotic infections, thought to be caused by the exposure that hunters and butchers had to the corporal fluids of non-human primates infected with SIV (Wolfe *et al.*, 2004; Kalish *et al.*, 2005). In terms of ancestry, the HIV-1 strains are closer “related” to the SIV that naturally infect chimpanzees (SIVcpz)

¹ Karposi's Sarcoma is a rare form of cancer that presents as purple lesions on the skin. It is normally rare in young individuals and makes them look older.

² PCP is a rare, but treatable lung infection that a healthy person can normally fend off easily.

(Gao *et al.*, 1999), while HIV-2 strains are closely related to the SIV from sooty mangabeys (SIVsm) monkeys (Gao *et al.*, 1992; Santos and Soares, 2010; Lihana *et al.*, 2012).

Santos and Soares (2010) state that it is uncertain how many times the SIVs crossed the species barrier and infected humans with success. This, they argue, is due to several factors that may influence the success of establishing infection, including “the efficiency of transmissibility, the capacity of avoiding the immunologic system; a successful replication in the new host and a pathogenic potential to guarantee human to human passage.” (Santos and Soares, 2010, p504). It has, however, been postulated that three such cross-species infection events have occurred in the past shaping HIV to its “present” state (Kandathil *et al.*, 2005; Lihana *et al.*, 2012).

1.3 HIV-1 Classification

HIV-1 is broadly grouped into four different phylogenetic lineages. These are groups M (Major group), N (New), O (Outlier) and P (Santos and Soares, 2010). The classification of groups M, N and O are based on phylogenetic sequences from the HIV-1 genome (*pol*, *gag*, *env*) (Wainberg, 2004; Jülg and Goebel, 2005; Paraschiv *et al.*, 2007). Group P is the most recent of the four groups to be discovered. It was first described in Cameroon in 2009 and to date has not shown any evidence of recombination with other HIV-1 subtypes (San Mauro and Agorreta, 2010; Lihana *et al.*, 2012). The types and groups of HIV are summarised in Table 1.1.

Table 1.1: Summary of HIV types and groups

Type	Group	Origin	Epidemiology	Comments
HIV-1	M	SIVcpz	All continents with exception of Antarctica	Major group responsible for the AIDS pandemic; more fit than HIV-1 group O and HIV-2.
	O	SIVgor or SIVcpz	Majorly found in Central and West Africa	Naturally resistant to NNRTI; less fit than group HIV-1
	N	Recombinant group M ancestor / SIVcpz	Only found in Cameroon	Very rare epidemically; few studies on drug resistance published.
	P	SIVgor	Cameroon	First described in 2009 in a Cameroonian woman. The actual number of infections is unknown.
HIV-2	-	SIVsm	Mainly found in Western and Central Africa; some cases in Western Europe, India, United States, Brazil and Japan	Apparently slower progression to AIDS; less susceptible to some anti-HIV-1 drugs; naturally resistant to NNRTI.

Source: Adapted from Santos and Soares (2010, p505)

HIV-1 group O is the most divergent group (Table 1.1) and has been implicated in having its origin from the SIV that infected wild gorillas (SIVgor)(Van Heuverswyn *et al.*, 2006). The epidemic pattern of HIV-1 group O has been restricted to West and Central Africa

(Santos and Soares, 2010). Group N was only identified in 1998 with its origin traced back to a recombination event between the ancestor of group M and the SIV that infected chimpanzees (SIVcpz) (Takehisa *et al.*, 2009). Group N is very rare epidemiologically. HIV-1 group M is responsible for more than 95% of the AIDS pandemic and contains the majority of HIV subtypes responsible for the disease (Lihana *et al.*, 2009; Santos and Soares, 2010).

1.4 Diversity and Geographical Spread of HIV-1 Group M

Although authors have contended that HIV-1 group M has 11 “pure” subtypes, namely subtypes A-K (Peeters and Sharp, 2000; Thomson, Perez Alvarez and Najera, 2002; Lihana *et al.*, 2009), the current classification, however, holds that there are nine such subtypes (i.e. excluding subtypes E and I) as proposed by earlier authors such as Robertson *et al.* (2000). Support for the latter classification was because no evidence was found to support the existence of a “pure” subtype E. Additionally, it was discovered that subtype I was actually a complex recombinant of subtypes A, G and I (Gao *et al.*, 1998; Nasioulas *et al.*, 1999; Santos and Soares, 2010). Subsequently, the “pure” subtypes E and I were reclassified as CRF01_AE and CRF06_cpx, respectively (Santos and Soares, 2010).

HIV-1 group M thus consists of nine different “pure” subtypes or non-recombinant forms (i.e. A-D, F-H, J, and K) as well as circulating recombinant forms. At the time of writing, group M had approximately 96 circulating recombinant forms (CRFs) (Los Alamos National Laboratory, 2018a; Recordon-Pinson *et al.*, 2018). The growth of the CRFs has more than doubled from 43 in 2009 (Lihana *et al.*, 2009). Subtype A is further subdivided into seven sub-subtypes, namely A1-A5, F1 and F2 (Lihana *et al.*, 2009, Santos and Soares, 2010). There also exists intersubtype forms, which are divided into two categories, namely CRF and URF (unique recombinant form). When found in a population that has at least three individuals without any epidemiological link and with the same intersubtype breakpoints, this form is called a CRF. However, if it is found in only a single patient, it is classified as a URF (Santos and Soares, 2010).

The global prevalence of group M subtypes and CRFs, can only reliably be traced to reports dating back to 2007 given by Hemelaar *et al.* (2011). As recently as 2017, authors such as Daw *et al.* (2017) still quoted Hemelaar *et al.* 's (2011) work. Literature also shows little or no evidence of attempts to update Hemelaar *et al.* 's work despite the plethora of isolated country / continent-specific epidemiological research that is in circulation. Notwithstanding,

subtype C is reported as the most virulent of the group M subtypes with 48% of the global HIV infections over the 2004 to 2007 period being accredited to it (Hemelaar *et al.*, 2006; Hemelaar *et al.*, 2011). This subtype is predominantly found in Southern and Eastern Africa, India and in the southern region of Brazil (see Figure 1.1).

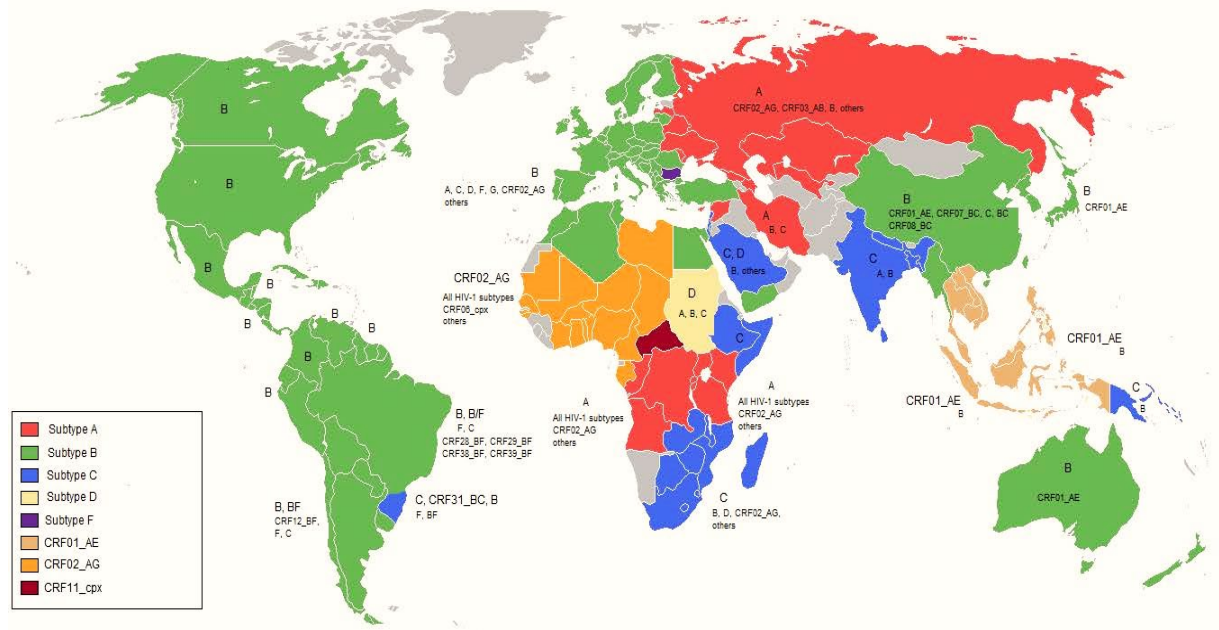


Figure 1.1: Worldwide prevalence of HIV-1 group M subtypes and CRF

Source: Santos and Soares (2010, p507)

The second most infectious “pure” strain is subtype A with approximately 12% of infections attributed to it (Hemelaar *et al.*, 2006; Hemelaar *et al.*, 2011; Lihana *et al.*, 2012). Subtype A is prevalent in Central Africa, Iran, Eastern Europe, and Central Asia. Within this group, however, subtypes A2 and A3 are primarily found in Africa and hardly ever in Europe. Subtype B is the most disseminated group M variant claiming 11% of global HIV infections (Hemelaar *et al.*, 2006; Hemelaar *et al.*, 2011). This group predominates in the developed world, in countries like the United States of America, countries of Western Europe, Japan and Australia (Soares *et al.*, 2007; Santos and Soares, 2010; Lihana *et al.*, 2012). The remaining six groups (i.e. D, F, G, H, J, and K) jointly represent around 8% of infections (Hemelaar *et al.*, 2006; Santos and Soares, 2010; Hemelaar *et al.*, 2011; Lihana *et al.*, 2012). Subtype D is also found in East Africa, while subtype F (i.e. F1 and F2) occurs in Central Africa, South America and Eastern Europe (Wainberg, 2004; Jülg and Goebel, 2005). Subtype G and A/G recombinants also occur in Eastern Africa and in Central Europe,

subtypes H and K only occur in Central Europe, and subtype J has been found in Central America (Wainberg, 2004).

Apart from the influence that “pure” group M variants have on HIV infection, some CRFs have a more substantial impact on local AIDS epidemics, accounting for almost 17% of HIV infections (Santos and Soares, 2010; Lihana *et al.*, 2012). For instance, CRF01_AE in Southeast Asia and CRF02_AG in Western Africa contributed 8% and 5% to the global infection rate respectively (Hemelaar *et al.*, 2006; Hemelaar *et al.*, 2011). CRF06_cpx is considered the second most prevalent recombinant form in West Africa (Santos and Soares, 2010). Other CRFs and URFs are each responsible for 4% of global infections. This brings the total percentage of CRFs to 17% and the overall percentage of all recombinants (i.e. CRFs plus URFs) to 21% (Hemelaar *et al.*, 2006; Hemelaar *et al.*, 2011).

1.5 HIV-1 Structure and Genome

Morphologically, HIV-1 virions are spheroid shaped and measure between 100-120nm in diameter. The structural proteins that form the viral core are Matrix (MA/p17), capsid (CA/p24), Nucleocapsid (NC/p7) and p6. The HIV-1 genome consists of two copies of non-covalently linked, positive sense single-stranded Ribonucleic Acids (ssRNA), which are tightly bound to p7 and enclosed within p24. The p24 capsid also contains the late assembly protein (p6); viral enzymes protease (Pr), reverse transcriptase (RT) and integrase (IN); as well as viral proteins Vpu, Vif, Vpr and Nef. The MA surrounds the cone-shaped capsid giving the virion integrity and forming a shell that connects directly to the inner side of the membrane. The NC is involved in the formation and stabilization of the genomic RNA dimers and in the nucleocapsid assembly. The p6 serves as the domain of p55 and is essential for the last stage of viral assembly as well as the release of the Vpr protein into the assembled virion. As the capsid buds off the host cell, it retains part of the host cell membrane. This forms an envelope that surrounds the capsid. Anchored within the envelope are proteins from the host cell as well as exterior and transmembrane glycoprotein (gp120 and gp41 respectively). The glycoproteins allow the virus to fuse and attach to target cells initiating the infectious replication cycle (Sierra, Kupfer and Kaiser, 2005; Fanales-Belasio *et al.*, 2010; Marsden and Zack, 2013). The structural features of HIV-1 are shown in Figure 1.2.

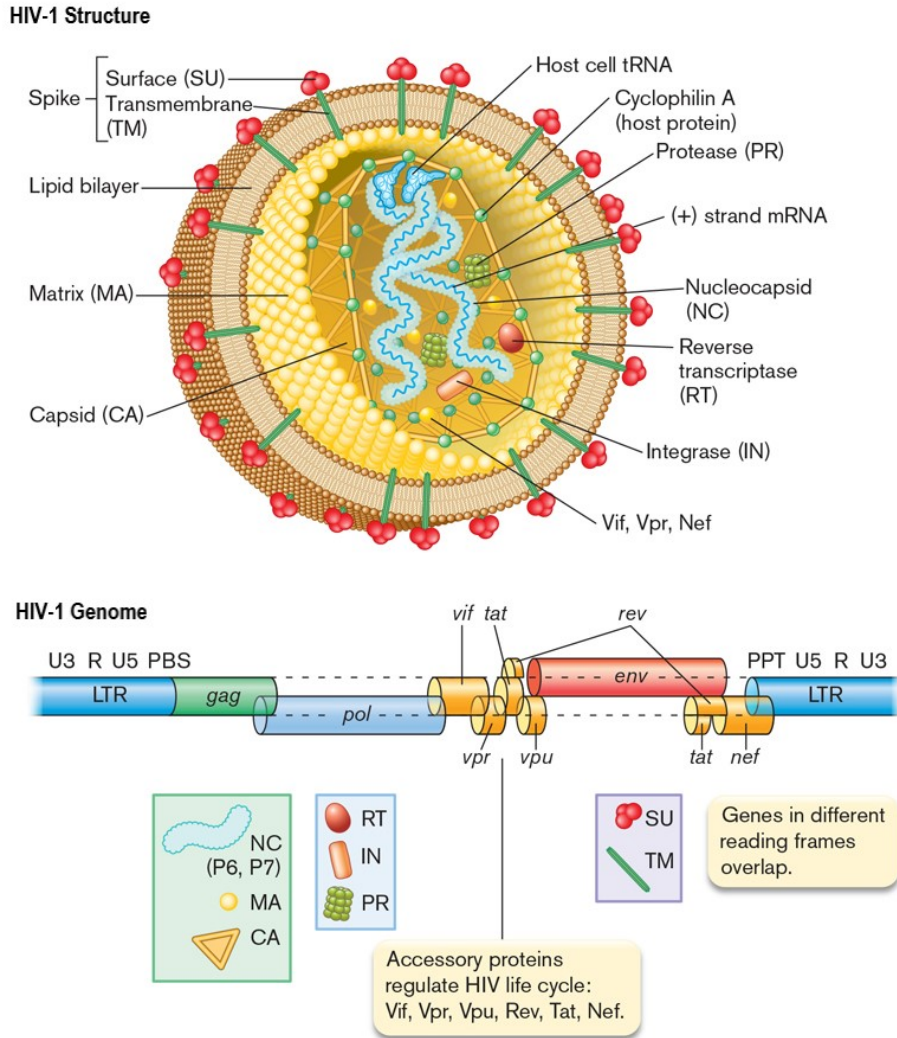


Figure 1.2: HIV-1 structure and genome

Source: Adapted from Slonczewski and Foster (2017)

The 9.2kb ssRNA molecules are comprised of four main regions, namely the Long Terminal Repeat (LTR) region, *gag-pol*, *env* and accessory genes (see Figure 1.2, HIV-1 genome) (Sierra *et al.*, 2005). The last three regions jointly contain nine Open Reading Frames (ORFs) that produce 15 proteins (Coffin, Hughes and Varmus, 1997; Frankel and Young, 1998; Watts *et al.*, 2009). The precursor to the *gag* polyprotein is proteolytically processed to produce the M, NC, CA and p6 proteins. The Gag-Pol polyprotein adds Pr, RT and IN. The *env* gene encodes a 30 AA signal peptide (SP), gp120 and gp41. Auxiliary proteins are coded by the additional sequences *tat*, *rev*, *rev*, *nef*, *vif*, *vpr* and *vpu* (coloured in orange in Figure 1.2, HIV-1 genome). These proteins control HIV infection of cells and viral replication. The LTR region, flanking both edges of the coding regions, is regulatory and contains the U3, R and U5 regions. The U3 region is a unique non-coding region, approximately 200-1200nt,

which forms the 5' end of the provirus following reverse transcription. It also contains binding sites for cellular transcription factors. The much shorter (18-250nt) R region forms direct repeats at both sides of the genome and contains the Transactivation Response Element (TAR), which is essential for tat-mediated transactivation. The first part of the viral genome to be reverse transcribed is found in the U5 non-coding region (approx. 75-250nt), thus forming the 3' end of the provirus genome (Coffin *et al.*, 1997; Frankel and Young, 1998; Lodish *et al.*, 2000; Watts *et al.*, 2009).

1.6 HIV-1 Lifecycle

The first step in the HIV-1 life cycle is the viral attachment to the host cell, which is facilitated by the binding of HIV-1 gp120 to host CD4 cell receptor. Upon binding to the target cell receptor, the gp120 undergoes a conformational change that causes it to bind to either Chemokine receptor type 5 (CCR5) or Chemokine receptor type 4 (CXCR4) found on the membrane of the target cell (see Figure 1.3). In addition, the receptor binding also causes a conformational change in gp120 exposing a hydrophobic domain on the gp41. This conformational change facilitates fusion with the cell membrane allowing the uncoating of viral core into the host cell cytoplasm. The RT enzyme then converts the ssRNA into dsDNA (i.e. complementary DNA) (Singh *et al.*, 2010; Arts and Hazuda, 2012). The complementary DNA (cDNA) is then transported to the nucleus where the viral DNA is integrated with the host cell DNA by viral enzyme IN (Melikyan *et al.*, 2000; Sierra *et al.*, 2005; Singh *et al.*, 2010; Arts and Hazuda, 2012; Craigie and Bushman, 2012).

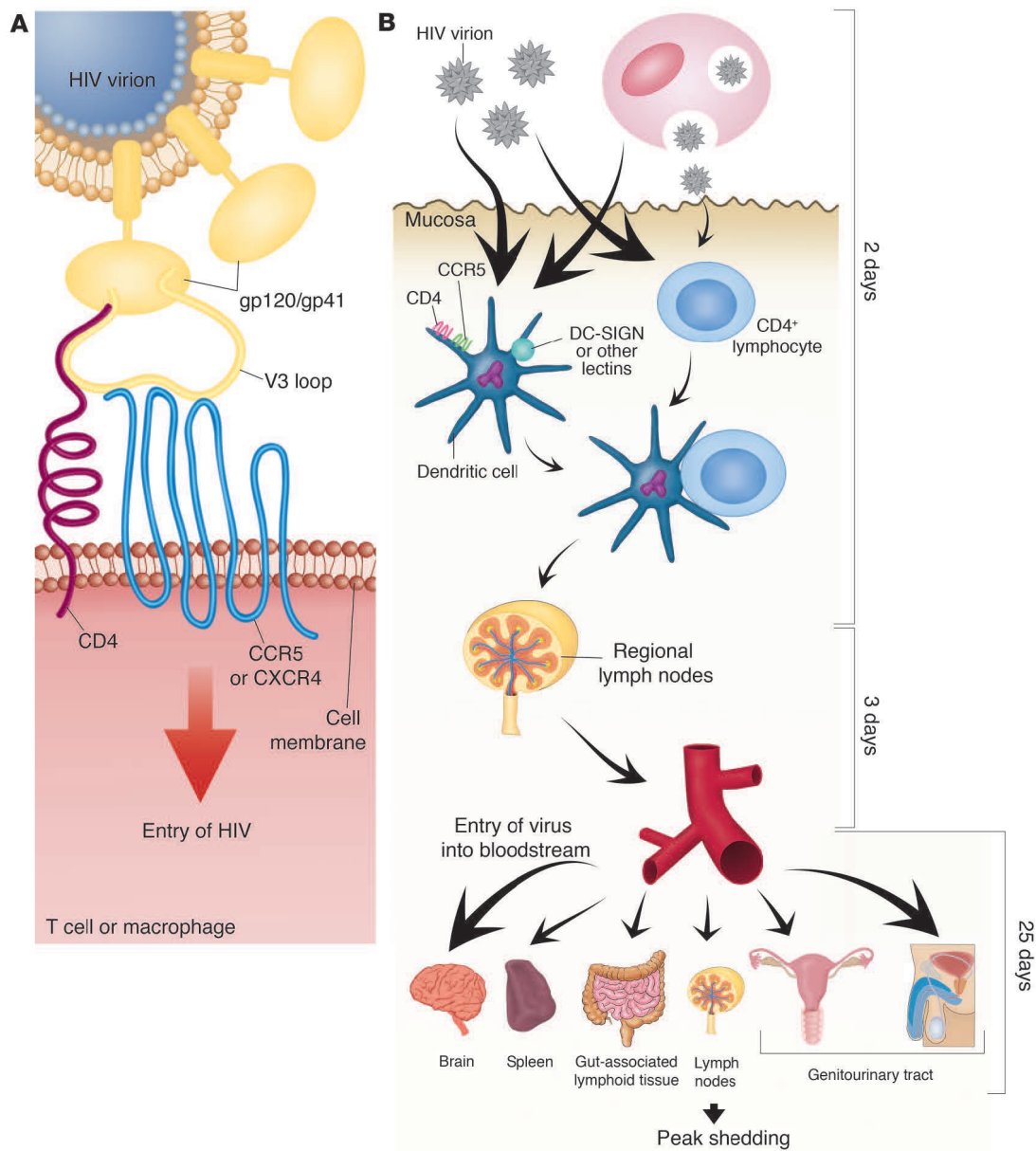


Figure 1.3: HIV transmission and the establishment of HIV reservoirs

(A) Interactions of HIV envelope glycoproteins, CD4, and CCR5 or CXCR4 co-receptors trigger fusion and entry of HIV. (B) Outline of the sequence and time course of events involved in viral dissemination. Source: Pilcher et al (2004, p940)

After integration, the RNA polymerase II transcribes the provirus into mRNAs. The mRNAs are translated into structural components, enzymes, and genomic RNA, which are subsequently transported to the cellular membrane and assembled into immature virions and released (Sierra et al 2005). The viral Pr cleaves the Gag and Gag-pol polyprotein into mature Gag and Pol proteins (Sierra *et al.*, 2005).

1.7 Acute HIV Infection

Acute infection refers to the initial stage of HIV infection up to the point where HIV is considered to be chronic (Fiebig *et al.*, 2003). It occurs approximately 1.5 weeks after HIV-1 exposure and ends when antibodies to HIV-1 are produced (Bassett *et al.*, 2011; Cohen *et al.*, 2011; Maartens, Celum and Lewin, 2014; McMichael *et al.*, 2010). It lasts on average 12 weeks and consists of an eclipse phase and five Fiebig stages, which are characterised by the progressive appearance of viral markers and antibodies in the blood (Figure 1.4) (Fiebig *et al.*, 2003; Cohen *et al.*, 2011; McMichael *et al.*, 2010).

The first 10 days after HIV-1 transmission, referred to as the eclipse phase, is the period during which the virus starts to establish itself in local tissue at the site of exposure. During this time HIV-1 RNA is undetectable and reservoirs of virions are established within latently infected CD4⁺ memory T-lymphocytes cells and macrophages (Cohen *et al.*, 2011; Chun and Fauci, 2012; Riou *et al.*, 2012; McMichael *et al.*, 2010). Latently infected cells are able to carry HIV-1 without expressing surface antigens thus escaping host immune recognition and resistance to virus-induced cytopathic effects (Chun and Fauci, 2012).

By the end of the eclipse phase, cell-free viruses and infected cells reach the draining lymph node where they encounter additional CD4⁺ cells to infect. Some of the viral particles are internalised by dendritic cells and presented to activated CD4⁺ T-lymphocytes, further increasing infection (McMichael *et al.*, 2010). The subsequent increased interaction between HIV-1 and cells expressing CD4⁺ receptors results in increased cellular infection and viral spread into the blood and lymphoid tissues, particularly the gut-associated lymphoid tissue (GALT) where a significant portion of CD4⁺ T-lymphocytes reside (Brenchley *et al.*, 2004; Pilcher *et al.*, 2004; Brenchley *et al.*, 2008). During this early phase of acute infection, viral concentrations in the blood and genital fluids peak, owing to increased HIV replication that is unrestrained by immune responses (Pilcher *et al.*, 2004; Bassett *et al.*, 2011; Ndhlovu *et al.*, 2015). HIV-1 in the GALT and other lymphoid tissues exponentially increases plasma viremia, which peaks at approximately 1 million copies of virus per ml, between 21-28 days after infection (Figure 1.4) (Pilcher *et al.*, 2004; McMichael *et al.*, 2010).

In response to peak viremia, acutely infected host's immune system mounts an intense inflammatory response that is characterized by high cytokine and chemokine levels (i.e. "cytokine storm") (Stacey *et al.*, 2009). Both the adaptive and innate immune responses are

activated, jointly contributing towards decreasing the viral load. CD8⁺ T-lymphocytes (CTL) start to kill productively infected CD4⁺ cells as part of the adaptive immune response shortly after infection. However, some viruses develop mutations in various epitopes and successfully escape immune selection (Llano, Frahm and Brander, 2009; Llano *et al.*, 2013; Pereyra *et al.*, 2014).

Upon CD4⁺ T-lymphocyte decline, infected individuals may become symptomatic (e.g. experience flu-like symptoms), which occurs around 26-35 days after initial infection (Pilcher *et al.*, 2004; Bassett *et al.*, 2011). The viral load continues declining until the viral set-point (i.e. point of stabilization) is reached, which usually marks the starting point of chronic HIV-1 infection (McMichael *et al.*, 2010). Viral set-point is individual-specific and dependent on viral replication and host immune responses. Consequently, higher viral set-points are associated with faster disease progression, while lower viral set-points are associated with slower disease progression, with immune escape mutations thought a major contributor towards maintaining higher viral set-points (McMichael *et al.*, 2010).

Fiebig *et al.* (2003) developed a six-stage model for stratifying the different stages of acute infection based on routine laboratory detection assays (see figure 1.4). They described these stages as:

Stage I : HIV present in blood samples, only RNA assay positive

Stage II : RNA and HIV-1 p24 antigen tests positive, antibody EIA non-reactive

Stage III : RNA, HIV-1 antigen and HIV IgM-sensitive EIA reactive, but Western blot without HIV-1-specific bands

Stage IV : Stage III plus indeterminate Western blot pattern, i.e. the presence of HIV-1-specific Western blot bands that fail to meet interpretative criteria for reactive Western blot defined by the USA Food and Drug Administration as reactivity to two of the following three bands: p24, gp 41, gp 120 / 160

Stage V : Stage IV, but reactive Western blot pattern, except lacking p31 (*pol*) reactivity

Stage VI : Stage V, but full Western blot reactivity including a p31 band

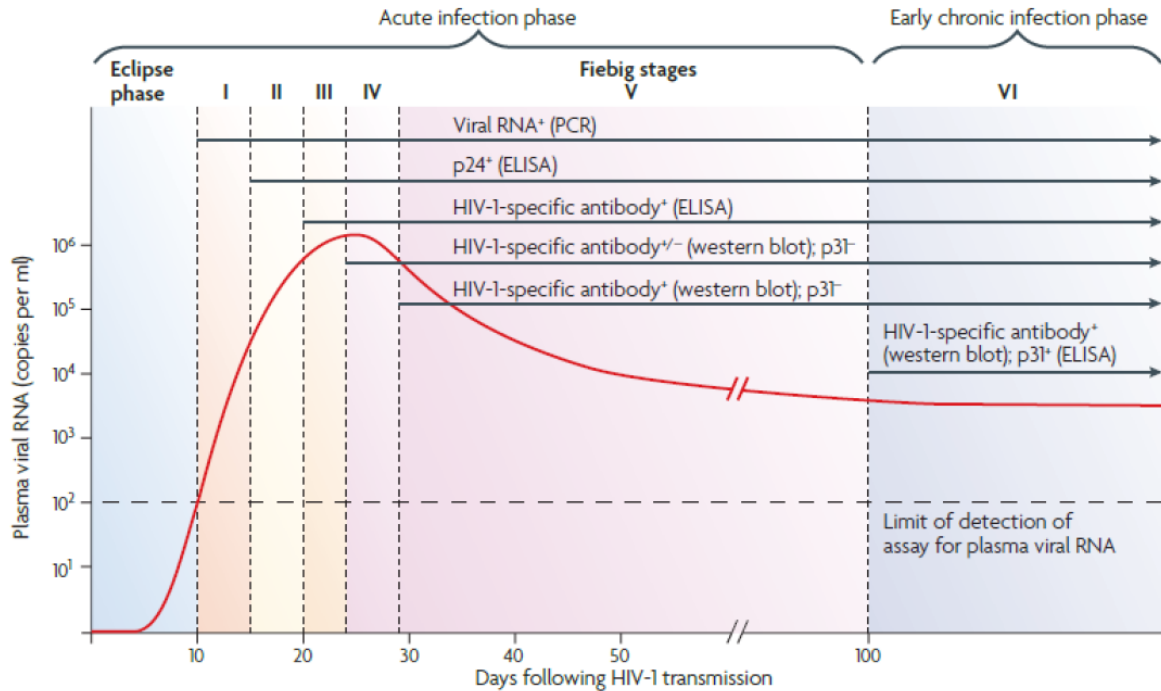


Figure 1.4: Progression of HIV-1 after infection

According to Fiebig *et al.* (2003), each of the stages from I-IV are relatively brief, lasting 3-5 days on average. Stage V is estimated to last an average of 69.5 days. Stage VI, on the other hand, has no set duration, except that its end point marks what can be thought of as either the recent or early phase of chronic infection (Fiebig *et al.*, 2003).

1.8 Acute Infection and the FRESH cohort

Pilcher *et al.* (2004) state that recognizing acute HIV infection is important for at least three reasons. Firstly, prevention strategies directed at individuals with acute HIV infection could potentially have a great impact. Secondly, very early recognition could present early treatment options that could alter the natural history of the disease, or even eliminate infection. Thirdly, it provides a unique opportunity to understand HIV transmission and pathogenesis, including early host-virus interactions, which lay the foundations for further study.

Early host-virus interactions are known to shape HIV variant populations in systemic and tissue compartments (Pilcher *et al.*, 2004). The number of replication events occurring during acute HIV infection is vast, and opportunities for host factors to exert pressure on existing or newly mutated viral variants is enormous. The host factors that exert early

selective pressure may include the cells that were initially infected, or microenvironments in the mucosa, submucosa, and lymph nodes, which present a wide spectrum of innate host defences, including IFNs and/or other molecules [(Shugars and Wahl, 1998; Shugars *et al.*, 1999) as cited by Pope and Haase (2003) and Pilcher *et al.* (2004)]. Host selective pressures, therefore, may either increase or decrease the diversity of the infecting-virus population (Pilcher *et al.*, 2004).

According to Pilcher *et al.* (2004), less than 0,002% of HIV positive individuals were diagnosed within the first month of infection globally. One of the main reasons for this is that early acute infection does not have a specific and recognisable acute retroviral syndrome (Pilcher *et al.*, 2004). Instead, early acute infection symptoms are non-specific and overlap many common febrile syndromes, such as influenza, malaria and rickettsial diseases (Pilcher *et al.*, 2004; Bassett *et al.*, 2011). Consequently, the true diagnosis of acute HIV is rarely considered at an initial patient encounter (Pilcher *et al.*, 2004; Ndhlovu *et al.*, 2015).

A second reason for failure to diagnose HIV infections is because routine antibody tests normally remain negative 1 to 2 weeks after acute retroviral symptoms appear (i.e. symptoms normally appear 26-35 days after initial infection) (Pilcher *et al.*, 2004). Diagnosis at this stage is through the presence of HIV p24 antigen (detected by ELISA) or HIV RNA (detectable by more sensitive nucleic acid amplification) (Pilcher *et al.*, 2004). Another challenge is that shortly before seroconversion, the p24 antigen tends to become undetectable due to the formation of early antibody-antigen complexes. A secondary antibody-negative, p24 antigen-negative period is sometimes observed (Pilcher *et al.*, 2004).

In the present study, blood samples were serially collected from 18-23 year old HIV uninfected sexually active women from Umlazi Township in KwaZulu-Natal, South Africa, that participated in a study called Females Rising through Education, Support and Health (FRESH) programme from November 2012 (Ndhlovu *et al.*, 2015). Baseline blood from these participants were cryopreserved and each participant screened bi-weekly by finger-prick plasma HIV RNA testing for evidence of acute HIV infection. Therefore, the acute blood samples used for this study were collected from participants with evidence of plasma viremia, usually within 24 hours of onset of plasma viremia (OPV) and continued being collected at regular intervals thereafter (Ndhlovu *et al.*, 2015). The FRESH acute cohort is unique because participants were followed prior to acquiring HIV and thus HIV infection

was identified at its earliest detectable point, making this cohort an excellent reflection of early acute infection. This cohort thus presents an ideal opportunity to observe host-virus dynamics at one of the earliest points of HIV infection.

1.9 Development of HIV Drug Resistance

HIV has a very high mutation rate owing to its error-prone RT enzyme that lacks a proofreading mechanism. It is estimated that one nucleotide mutation occurs with each replication cycle (Tang and Shafer, 2012). Although people are normally infected by one or a few original variants of the virus, it is estimated that 10^{10} virions are produced daily in untreated individuals giving rise to numerous variants or quasispecies (Tang and Shafer, 2012; Smith *et al.*, 2016). Quasispecies commonly refers to a complex distribution of variants that are genetically distinct but closely related (Ramirez *et al.*, 2013). This enables the virus to evade the immune system and lays the foundation for the development of ART resistance (Tang and Shafer, 2012). According to Tang and Shafer (2012), drug resistance can be attained by three different ways. Drug resistance can either be acquired (i.e. through the selection of drug resistance mutations (DRMs)), transmitted (i.e. from person to person), or occur naturally (in drug naïve viruses). Naturally occurring mutations, however, are considered rare (Tang and Shafer, 2012).

According to AIDS info (2015), the main reason for virological treatment failure (defined by the WHO (2015) as a viral load ≥ 1000 cpm) was due to incomplete patient adherence to their treatment regimens. The WHO defines adherence as “...the extent to which a person’s behaviour (i.e. taking medication, following a diet and/or changing lifestyle) corresponds with agreed recommendations from a health worker.” (WHO, 2016, pxiv). Adherence to therapy, in particular, is critical for full viral suppression and for optimal immune reconstitution (Conradie *et al.*, 2012; Tang and Shafer, 2012). The most notable repercussion of non-adherence is the development of DRMs in patients on ART. This does not only have negative implications for the non-adhering individuals but may potentially have a negative cascade effect on the people that they might transfer the drug-resistant virus to (Conradie *et al.*, 2012). As stated by Imaz, Falco and Ribera (2011), the emergence of drug resistance does not only have an impact on the ART, but reduces the treatment opportunities for both ART-experienced and ART-naïve patients due to the cross-resistance to available drugs (Imaz *et al.*, 2011; Tang and Shafer, 2012). Primary resistance (i.e. the transmission of resistant variants to uninfected individuals) thus raises serious clinical and public health

consequences and could potentially impact the ability to treat HIV in the near future (Alencar *et al.*, 2013). The emergence of HIV drug resistance, therefore, poses a major threat to sustaining the benefits of ART (Imaz *et al.*, 2011).

Several empirical studies have investigated the presence of transmitted drug resistance (TDR) mutations among treatment naïve patients. Afonso *et al.* (2012), for example, concluded after testing for TDR mutations among 86 Angolan patients who were diagnosed HIV positive not longer than a year ago, and who were confirmed ART-naïve, that an unexpectedly high frequency of DRMs towards RT inhibitors was found among these patients. Specifically, they found that of these recently infected treatment naïve patients, 14 (16.3%) displayed at least one DRM, 12 (14%) patients had DRM to NNRTIs, 9 (10.5%) had DRM related to NRTIs and seven (8%) showed DRM to both classes of RT inhibitors. However, they did not observe DRMs to PIs among any of the Angolan patients. Similarly, a study performed by Alencar *et al.* (2013) on 303 Brazilian blood donors that were recently identified HIV infected and were drug naïve, detected primary drug resistance in 36 (11.8%) HIV strains. The majority of these samples (31) were resistant to one drug class (17 to NNRTIs, eight to PIs, and 6 to NRTIs), four samples to two drug classes (NNRTIs and NRTIs) and one to all three drug classes.

Studies from other countries, despite finding TDR mutations among recently infected treatment naïve patients, found the prevalence to be moderate to low. Avila-Rios *et al.* (2011), for example, enrolled 145 treatment naïve patients in Guatemala over a 6 month period and found that the prevalence of TDR mutations was 8.3% (i.e. detected TDR mutations in 12 patients). Ten (83.3%) of these patients had TDR mutations against NNRTIs. Transmitted drug resistance mutations towards NRTIs and PIs were found to be less than one percent for both drug classes. Similarly, Chen *et al.* (2012) genotyped 299 *pol* sequences from blood samples collected over a two year period from newly HIV diagnosed Chinese patients who were all ART naïve. They found TDR mutations in 13 (4.3%) of these patients. Approximately 1.3% of TDR mutations were related to PIs, 0.3% to NRTIs, and 2.7% to NNRTIs.

To assist in standardising, guiding and facilitating the screening of TDR mutations globally, Bennett *et al.* (2009) compiled a list of 93 mutations that could be referred to in testing for TDR. These surveillance drug resistance mutations (SDRMs) were a consensus of reported

major DRMs that featured on the lists of three or more of five expert lists. The five expert lists used by Bennett *et al.* (2009) were the ANRS Drug Resistance Interpretation Algorithm (2008.07), HIVdb Drug Resistance Interpretation Algorithm (4.3.7), IAS-USA Mutations Associated With Drug Resistance (March / April 2008), Los Alamos National Laboratories HIV Sequence database (2007) and the Rega Institute Drug Resistance Interpretation Algorithm (7.1.1). The 93 mutations on the compiled SDRM list included 34 NRTI-resistance mutations at 15 RT positions, 19 NNRTI-resistance mutations at 10 RT positions and 40 PI-resistance mutations at 18 Pr positions (see Table 1.2). This list, however, is not a comprehensive list of DRMs but is a consensus list of the more frequently reported TDR mutations. It also continues to be used and referred to in other updated DRMs lists, such as Stanford University HIVdb (2015, 2017).

Table 1.2: HIV RT and Protease Mutations for Drug-Resistance Surveillance

Reverse Transcriptase Mutations				Protease Mutations					
NRTIs		NNRTIs		PIs					
M41	L	Q151	M	L100	I	L23	I	G73	STCA
K65	R	M184	VI	K101	EP	L24	I	L76	V
D67	NGE	L210	W	K103	NS	D30	N	V82	ATSFLCM
T69	D Ins	T215	YFSCDEIV	V106	AM	V32	I	N83	D
K70	RE	K219	QENR	V179	F	M46	IL	I84	VAC
L74	VI			Y181	CIV	I47	VA	I85	V
V75	MTAS			Y188	LCH	G48	VM	N88	DS
F77	L			G190	ASE	I50	VL	L90	M
Y115	F			P225	H	F53	FY		
F116	Y			M230	L	I54	VLMTAS		

Source: Bennett *et al.* (2009) as summarised by Stanford HIV Drug Resistance Database. 2015. *Major HIV-1 Drug Resistance Mutations* [Online]. Available: <http://hivdb.stanford.edu>

1.10 Types of HIV Drug Resistance Mutations

Drug resistance mutations are normally classified based on their resistance to three broad categories of drugs that act on either the virus' RT enzyme, its Pr enzyme and its IN enzyme. Of these, it is resistance to RT inhibitors (RTIs) and Pr inhibitors (PIs) that are considered the most important (Khalid and Sezerman, 2016). This is partly due to these treatments playing a vital role in first-line and second-line therapeutic defenses against HIV.

Mutations that cause resistance to RTIs and PIs are classified as either primary or secondary mutations. Primary mutations are single mutations that are first to appear. They give rise to low sensitivity to one or more inhibitors. Secondary mutations, on the other hand, develop later on and increase viral resistance and viral fitness, especially when in combination with other mutations. In other words, they do not cause resistance on their own (Nyombi *et al.*, 2008; Clavel and Mammano, 2010). Two distinct groups of drugs typically target the RT enzyme. These are either NRTIs or NNRTIs (Smith *et al.*, 2016). The next section briefly discusses the mechanisms involved in the development of RT and Pr inhibitor resistance. It will also briefly discuss IN inhibitor resistance.

1.10.1 HIV drug resistance in Reverse Transcriptase

The RT enzyme is involved in both RNA-dependent DNA polymerization and DNA-dependent DNA polymerization (Shafer, 2002; Shafer *et al.*, 2007; Amiel *et al.*, 2011; Smith *et al.*, 2016). Reverse transcriptase inhibitors take the form of both nucleoside and nucleotide analogues that get incorporated into the growing chain of the viral DNA by the RT enzyme causing premature chain termination (Shafer, 2002). Smith *et al.* (2016) expand on this stating that the DNA polymerase activity of the RT enzyme is the most frequent target for RT inhibitors. They explain that this is achieved by either interfering with the DNA polymerase's ability to complete the growing viral DNA strand or by preventing DNA polymerase from binding onto its binding site.

1.10.1.1 Resistance to NRTIs

Nucleoside Reverse Transcriptase Inhibitors lack the 3'-OH that is present on the deoxyribose of normal nucleosides. The incorporation of the NRTIs into the growing viral DNA strand causes premature chain termination (Smith *et al.*, 2016). There are two mechanisms by which drug resistance to NRTIs arise. The first mechanism is caused by mutations that prevent the NRTIs from being incorporated into the growing DNA chain during synthesis (Smith *et al.*, 2016). This method of mutagenesis gives rise to what is referred to as discriminatory mutations. They are so named because these particular mutations enable RT to differentiate between dideoxy-NRTI chain terminators and the cell's natural deoxynucleotide triphosphates (dNTPs) and in doing so, prevent NRTIs from being incorporated into the growing viral DNA chain (Tang and Shafer, 2012). The most common

discriminatory mutations include M184V/I, K65R, K70E/G, L74V, Y115F and the Q151M complex of mutations (Tang and Shafer, 2012).

The second mechanism of NRTI resistance is caused by the nucleotide excision mutations (NEMs) that remove NRTIs, consequently allowing DNA synthesis to continue (Shafer, 2002). These mutations are sometimes referred to as primer unblocking mutations (PUMs) and operate by the phosphorylytic excision of NRTI-triphosphates that were added to the growing viral DNA chain. In addition to being referred to as PUMs, NEMs are also known as thymidine analogue mutations (TAMs) because they are selected by the thymidine analogues Zidovudine and Stavudine (Tang and Shafer, 2012). The most common excising mutations are M41L, D67N, K70R, L210W, K219Q/E and T251N/F (Singh *et al.*, 2010; Tang and Shafer, 2012). TAMs are classified into two pathways namely type 1 or type 2 (Tang and Shafer, 2012). Type 1 includes mutations M41L, L20I, and T215Y, while Type II includes the D67N, K70R, T215F and K219Q/I/E (Tang and Shafer, 2012). In addition, the presence of M184V together with the TAMs has been reported to be the most common pattern that causes resistance to all NRTIs (Tang and Shafer, 2012).

1.10.1.2 Resistance to NNRTIs

The NNRTIs prevent HIV replication by binding in a small hydrophobic pocket that is approximately 10 Å from the polymerase active site. The binding pocket is located underneath the bound double-stranded nucleic acid substrate. The subsequent binding of an NNRTI distorts RT and in so doing affects the alignment of the primer terminus and the polymerase active site, blocking the chemical step of viral DNA synthesis (Shafer, 2002; Wright *et al.*, 2013; Smith *et al.*, 2016). Inhibitor binding, therefore, affects the flexibility of RT and in the process prevents the synthesis of DNA (Shafer, 2002; Wright *et al.*, 2013; Smith *et al.*, 2016).

The mutations that give rise to NNRTI resistance are located in the hydrophobic pocket that binds the inhibitors. These mutations typically alter the dimensions of the NNRTI binding pocket to an extent that NNRTIs can no longer bind (Smith *et al.*, 2016). These mutations often result in high levels of resistance to one or more NNRTIs (Shafer, 2002; Wright *et al.*, 2013; Smith *et al.*, 2016). NNRTI resistance usually develops when the NNRTIs are used in the presence of incomplete suppression of viral replication (Shafer 2002).

1.10.2 HIV drug resistance in Protease

The Pr enzyme is responsible for the formation of viral structural proteins and enzymes by the mechanism of post-translational processing of the viral Gag and Gag-pol encoded polyproteins (Shafer, 2002; Shafer *et al.*, 2007; Clavel and Mammano, 2010). Protease inhibitors are similar in structure to the Pr substrate and thus compete with it for binding to the Pr enzyme's active site (Shafer 2002). Resistance to PIs is caused by mutations in the substrate cleft, which decreases the binding between the mutant Pr and the PI (Shafer, 2002; Toor *et al.*, 2011; Su *et al.*, 2016).

There are two types of AA mutations that are associated with the substrate cavity, termed primary and secondary mutations. Primary mutations affect the AAs found inside the substrate cavity and cause resistance by decreasing the binding affinity between the PI and the mutant Pr enzyme (Su *et al.*, 2016). Secondary mutations, on the other hand, impact on the AAs located outside the substrate cavity and are able to either compensate for the mutations found at the active site or decrease the activity of the mutant Pr (Shafer, 2002; Clavel and Mammano, 2010; Amiel *et al.*, 2011). Drug resistance to PIs does not only develop at the Pr gene but also occurs in Pr cleavage sites consisting of Gag and Gag-pol polyproteins (Clavel and Mammano, 2010).

1.10.2.1 Gag mutations related to Protease resistance

Cleavage of the Gag and Gag-pol polyproteins is a critical step in the replication and infectivity of HIV (Clavel and Mammano, 2010; Kozisek *et al.*, 2012). The Gag polyprotein is the main substrate for Pr binding and it has been discovered that mutations located in the NC/SP2/p6 gag region play a role in the development of PI resistance (Dam *et al.*, 2009; Clavel and Mammano, 2010; Kozisek *et al.*, 2012; Su *et al.*, 2016). In addition, some gag mutations occur despite the absence of detectable or observable Pr mutations and may be indicative that HIV can cause resistance to PIs by altering the Pr substrate instead of Pr itself (Nijhuis *et al.*, 2007; Ghosn *et al.*, 2011).

1.11 Detection of Drug Resistance Mutations and Testing for Minority Variants

There are two different methods used to detect DRMs. These are broadly classified as either genotypic or phenotypic methods. Genotypic methods involve sequencing HIV *pol* gene from the HIV RNA population present in plasma to detect DRMs (Gianella and Richman,

2010). Specifically, genotypic resistance methods test for the presence of DRMs in the viral enzymes targeted by ARV drugs. These include Pr, RT and IN enzymes. Standard genotypic methods are, however, incapable of detecting minority variants and are thus limited to detecting mutations in major (dominant) viral populations (Gianella and Richman, 2010). Phenotypic methods, on the other hand, directly measure drug susceptibility based on viral replication and assess the effects of mutations contained in the tested sample(s) (Gianella and Richman, 2010). Similar to genotypic techniques, this method also detects mutations in major viral populations and is limited in detecting minority variants (Metzner *et al.*, 2005; Gianella and Richman, 2010; Halvas *et al.*, 2010).

Minority variants arise due to HIV's high replication rate and the high RT error rate (Abram *et al.*, 2010; Halvas *et al.*, 2010). They form a small portion of quasispecies that evolve as a consequence of high error-prone viral replication. Minority variants are important because they impact on ART and have been found to later emerge as the major viral population (Metzner *et al.*, 2005). This can either be attributed to these variants developing more resistance mutations or could be the result of partially suppressive ART, which allows the minority population to have a higher growth advantage over the majority population (Charpentier *et al.*, 2004).

To compensate for the shortcomings of standard genotypic techniques, more sensitive methods have been developed to detect minority variants. These include standard cloning (SC), single genome sequencing (SGS), allele-specific polymerase chain reaction (ASPCR) and ultradeep pyrosequencing (UDPS) (Paredes i Deiros, 2009; Halvas *et al.*, 2010). The allele-specific polymerase chain reaction is a very sensitive and/or specific technique that can detect specific minority variants. This technique, however, does not provide information on other DRMs that might be present in an individual (Gianella and Richman, 2010). In contrast, SGS and UDPS allow for the analysis of the entire gene with the sequencing of single virus particles obtained from the original HIV particle (Palmer *et al.*, 2005). This allows for the analysis of genetic linkages of each detected mutation. Cloning and sequencing of multiple clones also allows for genetic linkages of each detected mutation and can determine whether mutations are present on different variants in the HIV population (Gianella and Richman, 2010; Ramirez *et al.*, 2013).

1.11.1 Ultradeep PyroSequencing (UDPS)

This method utilizes a combination of emulsion polymerase chain reaction (PCR) with massive parallel pyrosequencing techniques where many sequences of individual molecules generated from RT PCR products are sequenced in a single run (Simen *et al.*, 2009). This technology produces a massive number of sequences which makes possible the detection of multiple DRMs. The technology's detection sensitivity for minority variants is dependent on the coverage and depth attained (i.e. the average number of times a gene has been sequenced during a run). In addition, its sensitivity to detect minority variants can be as low as 0.5 - 1% in viral populations (Paredes i Deiros, 2009). Ultradeep pyrosequencing can also sequence individual templates and determine genetic linkages within the same viral genome in the same way as cloning and SGS (Gianella and Richman, 2010). When compared to Sanger sequencing, research has found UDPS to be more accurate and more reproducible (Stelzl *et al.*, 2011), more sensitive, more efficient and more reliable (Liang *et al.*, 2011; Samuel *et al.*, 2016). Although UDPS remains one of the most sensitive techniques, it is very expensive, is more time consuming and laborious (especially setting up), and it generates a significant amount of data that require extensive bioinformatics expertise to fully analyze (Paredes i Deiros, 2009; Ramirez *et al.*, 2013; Garcia-Diaz *et al.*, 2014; Mohamed *et al.*, 2014).

1.12 Aim and Objectives of the Study

The research question that the study aimed to address was “What mutation patterns can be observed in the HIV sequences of acutely infected individuals?”

1.12.1 Aim

The aim of the study was to identify sites experiencing positive selective pressure in the HIV of acutely infected individuals and the likely implications of these mutations to viral functional domains and host epitopes.

1.12.2 Objectives

- To evaluate the consistency of HyPhy methods in Datamonkey for identifying positively selected sites
- To identify likely sites for positive selection
- To scan for positively selected mutations within putative functional domains

- To locate positively selected mutations within epitopes
- To identify most recent common ancestors by ancestral reconstruction
- To compare drug resistance mutations between the founder virus and virus quasispecies
- To determine the evolution of the loss of mutations
- To correlate loss of mutations to structural changes

CHAPTER 2:

Bioinformatics Analysis Toolkit

2.1 Introduction

The sequences in this study were firstly analysed for sites experiencing positive selective pressure. Sequences that had multiple timepoints were tracked to observe the structural changes accompanying the loss of DRMs and to detect if any DRMs spontaneously arose. For the first set of analyses, the HyPhy package in Datamonkey and the codeml package in the Phylogenetic Analysis by Maximum Likelihood (PAML) software was used to test for sites under positive selection pressure. Nucleotide models of substitution that best fitted the datasets were determined using j-Modeltest and maximum likelihood (ML) trees were drawn in PAUP* and PhyML for codeml analysis. To trace the acquisition or loss of DRMs over time, ancestral sequence reconstruction was performed using codonml of the PAML package, with timescaled phylogenies constructed using Bayesian Evolutionary Analysis by Sampling Trees (BEAST). Descriptions of the bioinformatics tools that were used for analyses in this study are given in the following sections.

2.2 HyPhy package in Datamonkey for Detecting Sites under Positive Selective Pressure

Hypothesis testing using phylogenies (HyPhy) is a software package in Datamonkey (a web-based suite of phylogenetic analysis tools for use in evolutionary biology) that performs likelihood-based analyses to study patterns of sequence evolution (Kosakovsky Pond and Frost, 2005; Delpont *et al.*, 2010). Positive selection analysis is one of the standard analyses methods contained within the HyPhy package. The classical version of Datamonkey has three traditional methods available to test for diversifying and purifying selection acting at a single codon site, namely Single Likelihood Ancestor Counting (SLAC), Fixed Effects Likelihood (FEL) and Random Effect Likelihood (REL). Two additional methods that are an extension of FEL are also used to detect positive selection. These are the Internal Branch FEL (iFEL) and the Mixed Effects Model of Evolution (MEME) methods (Delpont *et al.*, 2010). These methods are summarised in Table 2.1.

Table 2.1: Site by site selection models

Method	Description
SLAC	<ul style="list-style-type: none"> - the fastest and most conservative method - used for large datasets (≥ 50 sequences) and to obtain substitution maps at each site (a useful feature for visualizing the evolutionary process) - uses a combination of maximum-likelihood (ML) and counting approaches to infer nonsynonymous (dN) and synonymous (dS) substitution rates on a per-site basis for a given coding alignment and corresponding phylogeny - like FEL (see below), this method assumes that the selection pressure for each site is constant along the entire phylogeny
FEL	<ul style="list-style-type: none"> - the best overall method in terms of the trade-off between statistical performance and computational expense - used for intermediate to large datasets (≥ 50 sequences) and to obtain good site-by-site substitution rate estimates to infer dN and synonymous dS substitution rates on a per-site basis for a given coding alignment and corresponding phylogeny - assumes that the selection pressure for each site is constant along the entire phylogeny
iFEL	<ul style="list-style-type: none"> - can detect sites that are positively selected at the level of a population - iFEL can be used to test for sitewise selection on internal branches of the tree.
REL	<ul style="list-style-type: none"> - is an extension of familiar codon-based selection analyses pioneered by Nielsen and Yang and implemented in PAML - allows synonymous rate variation - is often the only method that can infer selection from small (5-15 sequence) or low divergence alignments - makes the most assumptions compared to the other methods and is therefore susceptible to high rates of false positives in extreme cases
MEME	<ul style="list-style-type: none"> - MEME combines fixed effects at the level of a site with random effects at the level of branches - is an extension of FEL, where the ω ($dN/dS > 1$) values are allowed to vary along branches according to a 2-bin distribution (i.e. some branches may be under positive selection while others under negative selection) - employs a mixed-effects ML approach to test the hypothesis that individual sites have been subject to episodic positive or diversifying selection (i.e. MEME aims to detect sites evolving under positive selection under a proportion of branches) - is most appropriate to detect episodic diversifying selection affecting individual codon sites

Source: Adapted from Datamonkey (www.datamonkey.org) and Poon, Frost and Kosakovsky Pond (2009)

The methods used to detect diversifying and purifying selection at the site level estimate the rate of nonsynonymous (dN) and synonymous (dS) changes occurring at each site in the sequence alignment (Kosakovsky Pond, Frost and Muse, 2005; Poon *et al.*, 2009; Delpont *et al.*, 2010). Of the methods in Datamonkey, the SLAC method is considered the most conservative counting method that involves reconstruction of the ancestral sequences using a single most likely ancestral reconstruction that considers all possible ancestral reconstructions or sampling from ancestral reconstructions (Kosakovsky Pond and Frost,

2005; Poon *et al.*, 2009; Delpont *et al.*, 2010). The FEL method estimates the substitution rate of dN and dS at each site of a sequence alignment. The iFEL method, on the other hand, determines the selection pressure that occurs on the internal branches of a tree (Kosakovsky Pond and Frost, 2005). This method tests for population-level selective pressures that are restricted to the interior branches of the tree (Poon *et al.*, 2009, Delpont *et al.*, 2010). Generally, the FEL and iFEL are the best methods in terms of statistical performance and computational expense (Poon *et al.*, 2009; Datamonkey, www.datamonkey.org).

The REL method is an extensive codon-based selection analysis technique that allows for the dN and dS rate variation with the selection pressure at individual sites (Kosakovsky Pond and Frost, 2005). This method is similar to the popular likelihood methods used in the PAML package, with some important additions (e.g. synonymous rate variation). Rather than directly estimating dN and dS at each site, REL estimates the parameters for discretized distributions of dN and dS (with three rate categories for a total of nine possible rate combinations) from the whole alignment and then infers which of these sites is most likely to be under positive selection. REL tends to be the most powerful of the three tests (i.e. SLAC, FEL and REL) because it uses the entire alignment to make inferences about rates at each site. However, it tends to have the highest rate of false positives. This is because the distribution of rates to be fitted have to be defined *a priori*, and it may not satisfactorily model the unobserved distribution of rates. The MEME method uses a mixed-effects ML approach (i.e. ω values are allowed to vary along branches according to a 2-bin distribution in that some branches may be under positive selection while others under negative selection) to test for episodic positive or diversifying selection at the individual site level (Datamonkey, www.datamonkey.org).

Each of the methods implemented by Datamonkey uses a nucleotide substitution model to estimate branch lengths and nucleotide substitution biases (e.g. transition/transversion biases) of the tree from sequence alignments. In general, Datamonkey can use one of 203 time-reversible nucleotide substitution models. The most supported time reversible model (denoted as REV) is comprised of eight free parameters (3 nucleotide frequencies + 5 substitution rates). Four of the most frequently used models are F81, HKY85, TrN93, and REV, which are predefined as “named” options within Datamonkey (Poon *et al.*, 2009; Datamonkey, www.datamonkey.org).

2.3 Phylogenetic Analysis

Phylogeny is an illustration of the evolutionary relationship between genes and organisms. Similar to a genealogy, it depicts which genes or organisms are closely related (Vandamme, 2009). To represent this relationship, a phylogenetic tree is constructed. A phylogenetic tree presents an intuitive approach to inferring relationships among copies of a gene or among loci of a multigene family (Bos and Posada, 2005). Historically, the principal interest in constructing phylogenetic trees was to observe the pattern of the evolutionary relationships (i.e. topology) of the tree. Recent applications have extended the use of phylogenetic trees to serve as a source from which emanates information regarding the processes for observed patterns of evolutionary relationships. In addition, the tree topology becomes the framework from which further inferences are drawn. Phylogenetics, therefore, makes it possible to analyse evolutionary rates, gene duplications, recombinations, polymorphisms, lineage divergence, and population demographics. Consequently, phylogenetic analysis provides useful tools to calculate the time to the most recent common ancestor (tMRCA) for all the extant alleles/genes (Vandamme, 2009). In addition, divergence time calculations are often used when investigating the origin of species. As the basis for inference, accurate estimates of evolutionary parameters often centre on the validity of a single phylogenetic reconstruction. Typically, “inaccurate estimation of trees may lead to biased results and erroneous inference of processes or mechanism of evolution” (Bos and Posada, 2005, 212).

2.3.1 Overview of Phylogenetic Trees

Phylogenetic trees are diagrams used to illustrate evolutionary relationships among genes and organisms and help to indicate which genes or organisms are most closely related. Phylogenetic trees are so termed because they resemble the structure of a tree (Figure 2.1), and the terms used to refer to the various parts of the diagram (i.e. root, branch, node, and leaf) are also reminiscent of trees (Vandamme, 2009).

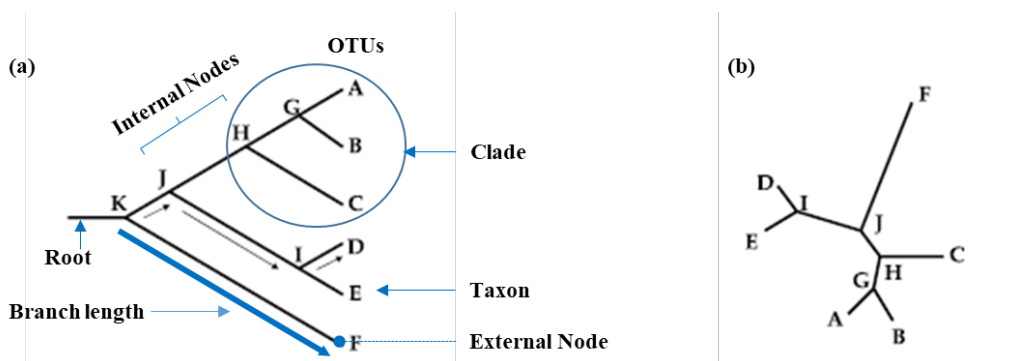


Figure 2.1: Basic tree structure and nomenclature

Structure of (a) a rooted and (b) an unrooted phylogenetic tree. Both trees have the same topology. A rooted tree is usually drawn with the root to the left. A, B, C, D, E, and F are external nodes or operational taxonomic units. G, H, J, K, and I are internal nodes or hypothetical taxonomic units, with K as root node. The unrooted tree does not have a root node. The lines between the nodes are branches. The arrow indicates the direction of evolution in the rooted tree (e.g. from root K to external node D). The direction of evolution is not known in an unrooted tree. Source: Adapted from Vandamme (2009), descriptions are taken verbatim.

A phylogenetic tree is typically composed of branches (edges) and nodes (Figure 2.1). Branches connect nodes, which are the points at which two or more branches diverge (e.g. A-G and G-B are branches that connect at node G in Figure 2.1). Both the branches and the nodes can either be internal or external. External (terminal) nodes or leaves (see A to F) represent extant (existing) taxa and are frequently called operational taxonomic units (OTUs), which is a generic term that can represent many types of comparable taxa (e.g. a family of organisms, individuals, or virus strains from a single species or from different species). Similarly, internal nodes are often referred to as hypothetical taxonomic units (HTUs) to highlight that they are the hypothetical ancestors of OTUs (Baldauf, 2003; Vandamme, 2009). A group of taxa that share a common branch have a monophyletic origin and are called a cluster or clade (e.g. in Figure 2.1 taxa A, B, and C form a cluster, with H as a common ancestor) (Freed, 2001; Vandamme, 2009). In contrast, C, D, and E do not form a cluster without including additional strains. They are, therefore, not of monophyletic origin and are called paraphyletic instead.

The topology of a tree refers to its branching pattern (i.e. the order of the nodes). Phylogenetic trees can either be rooted or unrooted (Figure 2.1). Rooted trees have a root that indicates the common ancestor of all the OTUs, thus illustrating the direction of the evolutionary process while the unrooted tree only positions the taxa relative to each other without showing the direction of the evolutionary process (Baldauf, 2003; Rizzo and Rouchka, 2007; Vandamme, 2009). In addition, phylogenetic trees are drawn with

proportional branch lengths corresponding to the amount of evolution. Thus the longer the branch length the more diversity in the sequence relative to the other branches (Baldauf, 2003).

2.3.2 Methods for Estimating Phylogenetic Trees

Methods for constructing phylogenetic trees from molecular data can be broadly classified according to either the kind of data that they use or to an underlying algorithm (Vandamme, 2009). The former approach (i.e. kind of data) refers to discrete character states or to distance matrices of pairwise dissimilarities. Character-state methods utilise any set of discrete characters, such as physiological properties, morphological characters, restriction maps, or sequence data. When applied to sequences, a ‘character’ refers to each sequence position in an alignment, while the nucleotides or AAs at that position are the ‘states’. Character-state methods preserve the original taxa character status and can thus be used to reconstruct the character state of ancestral nodes (Vandamme, 2009). Distance matrix methods, on the other hand, infer phylogenetic relationships from a pairwise distance matrix that is derived from calculating dissimilarities between each pair of taxa. These methods, however, cannot reconstruct character states of ancestral nodes because they discard the original character state of the taxa. The major advantage of distance methods is that they are generally computationally inexpensive, which is important when many taxa have to be analysed (Vandamme, 2009).

The algorithmic approach, in contrast, uses either a clustering algorithm (which usually produces one tree estimate) or an optimality criterion to evaluate different tree topologies (Vandamme, 2009). The optimality (or goodness-of-fit) criterion is used to evaluate different tree topologies for a given number of taxa in search of a tree that optimizes the predefined criteria. Maximum likelihood methods, for example, are a form of optimality approach that uses statistical criteria by considering the probability that a specific tree gave rise to the observed data (i.e. the aligned sequences) given a specific evolutionary model. This allows for comparison and relative support of different phylogenetic trees within a statistical framework. Clustering methods, on the other hand, avoid evaluating different trees by gradually clustering taxa into one tree (Vandamme, 2009). Generally, most distance-matrix methods use clustering algorithms to compute the best tree, while most character-state methods employ an optimality criterion. The classification of these methods is summarised in Table 2.2.

Table 2.2: Classification of phylogenetic analysis methods and their strategies

	Optimality search criterion	Clustering
Character state	Maximum parsimony (MP)	
	Maximum likelihood (ML)	
	Bayesian inference	
Distance matrix	Fitch-Margoliash	UPGMA
		Neighbour-joining

Source: Adapted from Vandamme (2009, p25)

Three popular methods used to estimate phylogenetic trees will be discussed briefly in the coming paragraphs. These are Neighbour-Joining (NJ), Maximum Likelihood (ML) and Bayesian methods. Each of these relies on explicit statistical models of evolution to reconstruct evolutionary trees.

2.3.2.1 Neighbour-Joining

The NJ method is a distance method that uses the genetic distance between sequences to construct a phylogenetic tree (Bos and Posada, 2005). Neighbour-joining tree construction typically involves sequentially finding pairs of OTU's that are connected by a single interior node. The algorithm sequentially connects every possible OTU pair and finally joins the OTU pair that yields the shortest tree (Vandamme, 2009).

Genetic distance in an NJ tree is based on the hypothesis that the difference between two sequences is directly related to their phylogenetic relationship (San Mauro and Agorreta, 2010). The difference between the sequences is due to a number of changes that have occurred along the branches (i.e. the evolutionary distance) (San Mauro and Agorreta, 2010). The genetic distance can also be the differences between a pair of sequences based on their transition (ti) or transversion (tv) substitution rates (Posada, 2009; San Mauro and Agorreta, 2010). Transitions are substitutions between nucleotides that are structurally similar (e.g. A↔G, which are both purines, or C↔T, which are both pyrimidines). Transversions, on the other hand, occur between nucleotides that are structurally dissimilar (e.g. A↔C or G↔T or A↔T or G↔C, which are substitutions between pyrimidines and purines and *vice versa*) (see Figure 2.2) (Bos and Posada, 2005).

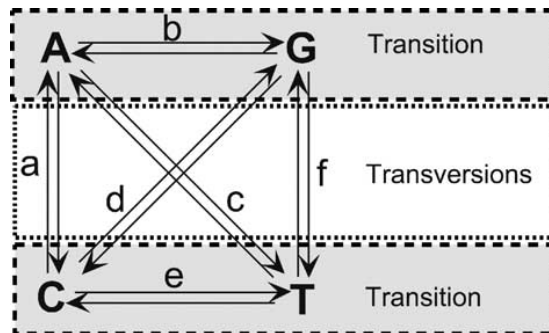


Figure 2.2: Transition and transversion substitution matrix

The substitution matrix illustrates the different rates of evolution of two possible transitions ($A \leftrightarrow G$, $C \leftrightarrow T$) and four possible transversions ($A \leftrightarrow T$, $A \leftrightarrow C$, $G \leftrightarrow C$, and $G \leftrightarrow T$). In this substitution matrix, substitution parameters are reversible, so that the rate of change from nucleotide i to nucleotide j is the same as the rate of change from j to i . Source: Bos and Posada (2005, p212).

Transitions are often observed at more than twice the rate of transversions ($t_i : t_v > 2$) despite there being twice as many possible transversions for any given nucleotide site. This trend towards more transitions occurs because a mutation to a similar nucleotide is more likely to be tolerated than a dissimilar one along a DNA sequence (Bos and Posada, 2005). The redundant nature of the genetic code is one of the main factors that dictate nucleotide substitution tolerance at various positions in a codon. Typically, similar substitutions at the third nucleotide position of an AA often does not alter the AA. As such, mutations at this position are more tolerated and restrictions on t_v or t_i are a little more flexible. This contrasts with mutations in the second and third codon positions of an AA in that changes in these are more likely to change or alter the underlying AA. These positions, therefore, tend to be more conservative and intolerant to mutations or substitutions.

In addition, protein structure is critical to its function. It is therefore important that structural integrity is preserved. As such, certain regions or domains within the protein are highly intolerant to mutations that will disrupt protein structural integrity and hence function. Substitution rates will, therefore, differ in different regions of the DNA sequence correlating to different domains in the protein (i.e. among codons rather than within codons) and can cause different parts of a gene to support different trees. The variation in substitution rates among different nucleotides in a sequence (as opposed to codon) is referred to as substitution rate heterogeneity or among-site rate variation (Li, 1997; Bos and Posada, 2005).

Overall, distance methods are relatively quicker compared to the other methods. However, they do not give information about the sequences as they are distance based (San Mauro and

Agorreta, 2010). In addition, the clustering algorithm used by NJ does not attempt to find clusters of OTUs that are most closely related. Rather, it shortens the length of the internal branches, and hence the length of the entire tree (Vandamme, 2009).

2.3.2.2 Maximum Likelihood

In maximum likelihood (ML) a tree that best explains the data, based on the specific substitution model, is derived directly from sequence data (Bos and Posada, 2005). A tree's likelihood is the probability of observing the data provided, with the tree and a model of evolution (San Mauro and Agorreta, 2010). Likelihood can, therefore, be estimated for different substitution models and the ML found. The ML has more statistical power in comparison to genetic distance methods, consequently providing a robust way of estimating phylogenies and understanding sequence evolution (Bos and Posada, 2005; San Mauro and Agorreta, 2010). Maximum likelihood methods are, however, more computationally intensive, especially when large numbers of sequences are being analysed (Bos and Posada, 2005).

2.3.2.3 Bayesian Inference

Bayesian methods are character-state methods that use an optimality criterion. They differ from ML in that they do not try to search only for the single best tree. However, Bayesian methods also use likelihood by searching for a set of plausible trees or hypotheses for the data by targeting a probability distribution of trees (i.e. posterior distribution), which inherently holds a confidence estimate of any evolutionary relationship (Vandamme, 2009). However, Bayesian methods require prior estimation of model parameters. These are often based on a prior belief, which is formalized as a prior distribution on the model parameters, for example, substitution model parameters, branch lengths, and tree topology (Huelsenbeck and Ronquist, 2001; Drummond and Rambaut, 2007; Vandamme, 2009). The comparative evidence present in the data then serves as a reference point for evaluating how one should update prior beliefs. Bayesian inference uses the Markov Chain Monte Carlo (MCMC) sampling technique to explore trees and estimate posterior probabilities based on them (Drummond and Rambaut, 2007; Vandamme, 2009). The posterior probability of a tree is the probability that the tree is correct, assuming that the model is correct (Huelsenbeck and Rannala, 2004). The advantage of Markov chains is their tendency to converge towards an equilibrium state regardless of the starting point (Ronquist, van der Mark and Huelsenbeck,

2009). In each step of the Bayesian inference, the likelihood ratio and prior ratio is calculated for the new state relative to the current state (Vandamme, 2009). After an initial convergence to a set of probable model/tree solutions, it is hoped that this stochastic algorithm samples from the posterior probability distribution. The frequency by which a particular tree topology is sampled is then proportional to its posterior probability. The result is a consensus tree or *maximum a posteriori* tree.

2.3.2.4 Summary of Estimation Methods

The NJ algorithm differs from ML and Bayesian methods because it calculates pairwise genetic distances between sequences and reconstructs a topology based on those distances. Bayesian and ML methods reconstruct a tree directly from sequence data. They thus use information in specific nucleotide differences instead of summarizing changes with a genetic distance. Due to these differences, ML methods offer better statistical properties in comparison to genetic distance-based methods. However, ML methods tend to be much more computationally intensive. The Bayesian method, like the ML method, utilizes the likelihood function. Using Bayesian statistics to reconstruct a phylogeny, however, results in the preferred outcome being one that maximizes the posterior probability, which is determined by the prior distribution and the likelihood of that tree. While NJ and ML produce a single best estimate of evolutionary relationship ignoring any uncertainty of the final tree, Bayesian methods, in contrast, produce a set of trees from which the one with the highest posterior probability is selected as the preferred / best tree (Bos and Posada, 2005; Schmidt and von Haeseler, 2009). Bayesian methods are generally considered to be faster than ML methods, and also offer the advantage of automatically incorporating an estimate of phylogenetic uncertainty (Bos and Posada, 2005). There is some contention, however, on the relative speed of ML versus Bayesian methods. Drummond and Rambaut (2007), for example, believe that the speed of Bayesian methods has erroneously been considered to be faster than heuristic searches that are based on ML.

2.3.3 Models of Nucleotide Substitution

Nucleotide substitution models provide an outline for phylogenetic reconstruction estimates for parameters used to find the best-fit tree. Models, however, differ from each other based on the number of parameters used (e.g. nucleotide frequencies, among-site variation, *inter alia*, see Figure 2.2) to represent evolutionary change. Furthermore, other models have been

derived from existing models by, for example, combining parameters (Bos and Posada, 2005). Consequently, models often share common or overlapping features. Among the most commonly known and used nucleotide substitution models are the Jukes and Cantor (JC69) model, Felsenstein (F81) model, the Kimura 2-parameter (K2P), K80 and K81 models; the Hasegawa, Kishino and Yano (HKY85) model, the symmetrical model (SYM), the transition model (TIM), the transversion model (TVM), and the General Time Reversible (GTR) model (Posada and Crandall, 1998; Bos and Posada, 2005).

Briefly, the JC69 model is the simplest of these models as it considers all possible nucleotide substitutions to have equal probability (Bos and Posada, 2005). Felsenstein’s F81 model bases the probability of nucleotide change on the nucleotides’ equilibrium frequencies. The K2P model uses a substitution matrix that permits ti and tv rates (Figure 2.3). The HKY85 model, on the other hand, allows for different substitution rates in nucleotide pairs. The GTR model has other models nested within it. This model allows for up to six different substitution rates and also permits different nucleotide substitutions (Lio and Goldman, 1998; Gatto, Catanzaro and Milinkovitch, 2006). A summary of some of the most commonly used nucleotide substitutions with their parameters is shown in Table 2.3.

Table 2.3: Some commonly used nucleotide models of substitution and summary of parameters

Model	Parameters		
	Number of parameters	Nucleotide frequencies	Substitution rate
JC69	1	Not included	$a=b=c=d=e=f$
F81	4	$\pi_A, \pi_C, \pi_G, \pi_T$	Not included
K80	2	Not included	$a=c=d=f, b=e$
K81	3	Not included	$a=f, b=e, c=d$
HKY85	6	$\pi_A, \pi_C, \pi_G, \pi_T$	$a=c=d=f, b=e$
SYM	6	Not included	a, b, c, d, e, f
TrN	7	$\pi_A, \pi_C, \pi_G, \pi_T$	$a=c=d=f, b, e$
GTR	10	$\pi_A, \pi_C, \pi_G, \pi_T$	a, b, c, d, e, f

Parameters of these models can include four different base frequencies and up to six substitution rates. The flexibility of models is such that invariable sites and/or a gamma distribution can simply be added to incorporate rate variation. It should be noted that the number of parameters or free parameters sometimes differ from author to author [cf. Figure 2.3 by Strimmer and von Haeseler (2009). Source: Bos and Posada (2005, p215)]

As mentioned earlier, models of evolution are sets of assumptions about the process of nucleotide or AA substitution. They thus “describe different probabilities of change from one nucleotide or AA to another along a phylogenetic tree, allowing us to choose among

different phylogenetic hypotheses to explain the data at hand” (Posada, 2009, 345). This makes them indispensable when estimating phylogenetic relationships among taxa in DNA and protein sequences.

Generally, more complex models are preferred over simpler ones because they fit the data better by virtue of having more parameters (Posada, 2009). The ideal use of models is to include as much model complexity as needed and no more in order to avoid over- or underfitting a model (Bos and Posada, 2005; Posada, 2009). Using complex models, however, requires that a large number of parameters be estimated. This has a number of disadvantages. Firstly, it makes the analysis of data computationally difficult and requires a significant amount of time. Secondly, more errors are introduced with the addition of more parameters due to the increase in the number of estimation that is required per parameter (Posada, 2009) (see Figure 2.3 for illustration of model complexity of some of the most frequently used models).

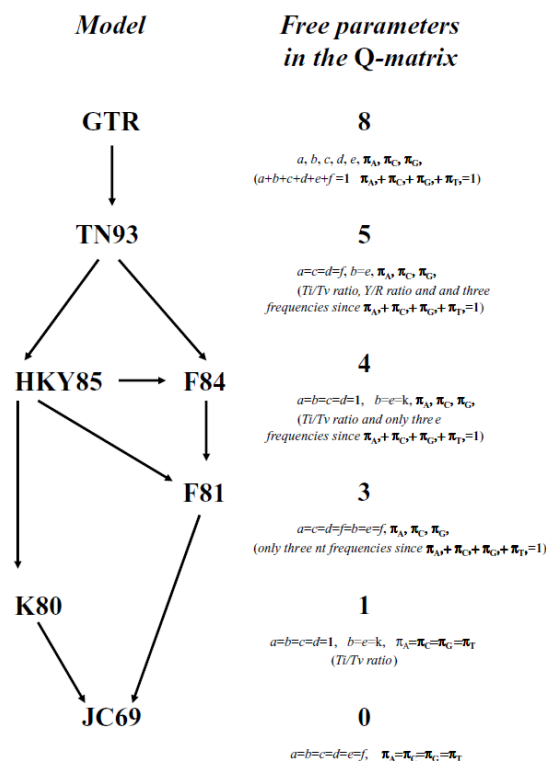


Figure 2.3: Hierarchy of nucleotide substitution models

Model complexity decreases from GTR down to JC69. Model complexity is dependent on the number of parameters that the model has i.e. the more parameter-rich, the more complex. Source: Strimmer and von Haeseler (2009, p124)

The best-fit model of evolution for a particular data set can be selected using trusted statistical techniques such as hierarchical likelihood ratio tests (hLRTs), information criteria and Bayesian or performance-based approaches. Regardless of the model selection strategy chosen, the fit of a model can be measured through the likelihood function, which is proportional to the probability of the data (D) given a model of evolution (M), a vector of K model parameters (θ), a tree topology (τ), and a vector of branch lengths (ν):

$$L = P(D|M, \theta, \tau, \nu) \quad \dots \text{Equation 1}$$

When the goal is to compute the likelihood of a model, the parameter values and the tree affect the calculations and become somewhat of nuisance parameters (so referred to by Posada (2009)) especially since they are not necessarily what is desired to be inferred. A standard strategy to “remove” unwanted parameters that might influence calculations, such as trees and parameter values, is often to use maximum likelihood estimates (Posada, 2009). These are values that make the likelihood function as large as possible:

$$\hat{\theta}, \hat{\tau}, \hat{\nu} = \max_{\theta, \tau, \nu} L(\theta, \tau, \nu) \quad \dots \text{Equation 2}$$

To facilitate the computation, it is standard practice to work with the maximized log likelihood:

$$\ell = \ln P(D|M, \hat{\theta}, \hat{\tau}, \hat{\nu}) \quad \dots \text{Equation 3}$$

Within a Bayesian setting, it is possible to integrate the undesired parameters out and obtain the marginal probability of the data given only the model $P(D|M)$ using computationally intensive techniques like the MCMC. Integrating out the tree, branch lengths, and model parameters to obtain $P(D|M)$ is represented by:

$$P(D|M) = \int \int \int P(D|M, \theta, \tau, \nu) P(\theta, \tau, \nu | M) d\theta d\tau d\nu \quad \dots \text{Equation 4}$$

A standard way to compare how two models fit is to contrast their log likelihoods using the likelihood ratio test (LRT) statistic (Posada and Crandall, 1998; Bos and Posada, 2005; Posada, 2009):

$$LRT = 2(\ell_1 - \ell_0) \quad \dots \text{Equation 5}$$

Where:

ℓ_1 is the maximum log likelihood under the more parameter-rich, complex model (alternative hypothesis) and

ℓ_0 is the maximum log likelihood under the less parameter-rich simple model (null hypothesis)

A different approach for model selection is to simultaneously compare all competing models. To do that, a penalty is charged to the likelihood of each model based on the number of free parameters in the model (K). Therefore, more parameters result in a bigger penalty.

Yet another method to compare models is the Akaike Information Criterion or AIC (Akaike, 1974). It is an asymptotically unbiased estimator of the Kullback–Leibler information quantity (Kullback and Leibler, 1951), which measures the expected distance between the true model and the estimated model (Posada, 2009):

$$AIC = -2\ell + 2K \quad \dots \text{Equation 6}$$

The AIC could be thought of as the amount of information lost when we use a substitution model to approximate the real process of molecular evolution. It is the model with the smallest AIC that is preferred. The AIC has the advantage of comparing both nested and non-nested models. When sample size (n) is small compared with the number of parameters ($n/K < 40$) a corrected version of the AIC is recommended (Posada, 2009):

$$AIC_c = AIC + \frac{2K(K+1)}{n-K-1} \quad \dots \text{Equation 7}$$

2.3.3.1 Selecting a Best Model of Substitution using Modeltest

A model test is performed to test for the best model of DNA substitution. This is a software programme that compares the different DNA substitution models using a hierarchical hypothesis testing framework (Posada and Crandall, 1998). The programme calculates the LTR statistic, associated p -values as well as AIC values (Posada and Crandall, 1998). The contemporary version of Modeltest used to test for DNA substitution models is j-Modeltest (v2.1.10). This program is based on PhyML (Posada, 2008). Its advantages over the

traditional Modeltest (it should be noted that some researchers still prefer the old Modeltest) are that it is a simpler, faster and more accurate algorithm to estimate large phylogenies by maximum likelihood. The programme uses five different selection strategies, including *hierarchical and dynamical likelihood ratio tests* (hLRT), the *Akaike information criterion* (AIC), the *Bayesian information criterion* (BIC), and a *decision theoretic performance-based* approach. It also calculates the relative importance and model-averaged estimates of substitution parameters, including a model-averaged estimate of the phylogeny. The program implements three different information criteria, namely the AIC, the BIC, and a performance-based approach based on decision theory (DT). Under the AIC framework, there is also the possibility of using a corrected version for small samples (AICc) instead of the standard AIC (Posada, 2008). j-Modeltest implements all 203 types of GTR substitution matrices, which when combined with unequal / equal base frequencies, gamma-distributed among-site rate variation and a proportion of invariable sites makes a total of 1624 (i.e. 203 x 8) possible models (Guindon and Gascuel, 2003; Darriba *et al.*, 2012; Darriba and Posada, 2016).

Although j-Modeltest is able to draw the best trees based on PhyML, it tends to lack flexibility in regards to tree topology (e.g. specifying clustering). To compensate for this, other models that allow for specifying outroots and outgroups are used, e.g. PAUP*. Often the PAUP* block that is generated by j-Modeltest for the best substitution models serves as a starting point for best tree reconstruction using the Phylogenetic Analysis Using Parsimony (PAUP*) package (version 4.0a). The PAUP block generated by j-Modeltest is then added to the end of the appropriate sequence alignment file. The block contains information on the best model that was selected by j-Modeltest, given the data, and has accompanying substitution rates and other model estimates which PAUP* can use as a starting point to generate phylogenetic trees (Posada, 2008). PAUP* then tests the sequence alignment against 56 models of evolution and gives an output called model scores. It also draws the best tree based on, e.g. ML, which can be viewed in programmes such as Figtree.

2.3.4 Phylogenetic Analysis by Maximum Likelihood (PAML)

Phylogenetic Analysis by Maximum Likelihood (PAML) software is a package of programs for the phylogenetic analyses of DNA and protein sequences using ML. The programs contained in the PAML package include BASEML, BASEMLG, CODEML, EVOLVER,

PAMP, YN00, MCMCTREE, and CHI2 software programmes. PAML can be used to compare and test phylogenetic trees. The main strengths of the package lays in its rich repertoire of evolutionary models, which can be used to estimate parameters in models of sequence evolution and to test biological hypotheses. Common uses of the programs include:

- i. the estimation of dN and dS rates and the detection of positive selection in protein-coding DNA sequences (YN00 and CODEML)
- ii. comparing and testing phylogenetic trees (BASEML and CODEML), inferring positive selection through phylogenetic comparison of protein-coding genes (CODEML)
- iii. likelihood ratio tests (LRTs) of hypotheses through comparison of nested statistical models (BASEML, CODEML, CHI2)
- iv. estimation of species divergence times under global and local clock models using likelihood (BASEML and CODEML) and Bayesian (MCMCTREE) methods, and
- v. reconstruction of ancestral sequences using nucleotide, AA, and codon models (BASEML and CODEML) (Yang, 2007b).

2.3.5 Bayesian Evolutionary Analysis by Sampling Trees (BEAST)

Bayesian evolutionary analysis by sampling trees (BEAST) is a cross-platform program for Bayesian MCMC analysis of molecular sequences. It is orientated towards rooted, time-measured phylogenies inferred using strict or relaxed molecular clock models (Drummond *et al.*, 2007). Among the main goals of BEAST is to provide a method for reconstructing phylogenies and to serve as a framework for testing evolutionary hypotheses without conditioning on a single tree topology (Drummond *et al.*, 2007). Using MCMC, BEAST averages over tree space weighting trees proportionally to their posterior probabilities. Being based on a Bayesian statistical framework, BEAST requires prior knowledge in combination with the information provided by the data (Drummond and Rambaut, 2007). Input into BEAST is by means of an XML file. To setup and generate this file, BEAST uses a program called Bayesian Evolutionary Analysis Utility (BEAUti). This program allows for specifying priors for BEAST. A brief description of BEAST accompanying programs is described in Table 2.4. These include LogCombiner, TreeAnnotator, Tracer, and Figtree.

Table 2.4: Associated programs used by BEAST

Associated programmes required / used by BEAST	Description
BEAUti - Bayesian Evolutionary Analysis Utility	BEAUti is a utility program with a graphical user interface for creating BEAST and *BEAST input files which must be written in the eXtensible Markup Language (XML). This application provides a clear way to specify priors, partition data, calibrate internal nodes, etc.
LogCombiner	When multiple (identical) analyses are run using BEAST (or MrBayes), LogCombiner can be used to combine the parameter log files or tree files into a single file that can then be summarized using Tracer (log files) or TreeAnnotator (tree files). However, it is important to ensure that all analyses reached convergence and sampled the same stationary distribution before combining the parameter files.
TreeAnnotator	TreeAnnotator is used to summarize the posterior sample of trees to produce a maximum clade credibility tree and summarize the posterior estimates of other parameters that can be easily visualized on the tree (e.g. node height). This program is also useful for comparing a specific tree topology and branching times to the set of trees sampled in the MCMC analysis.
Tracer	Tracer is used for assessing and summarizing the posterior estimates of the various parameters sampled by the Markov Chain. This program can be used for visual inspection and assessment of convergence and it also calculates 95% credible intervals (which approximate the 95% highest posterior density intervals) and effective sample sizes (ESS) of parameters (http://tree.bio.ed.ac.uk/software/tracer).
FigTree	FigTree is an excellent program for viewing trees and producing publication-quality figures. It can interpret the node-annotations created on the summary trees by TreeAnnotator, allowing the user to display node-based statistics (e.g. posterior probabilities) in a visually appealing way (http://tree.bio.ed.ac.uk/software/figtree).

Source: Adapted from Heath (2015, p9-10)

2.3.5.1 BEAUti

The BEAUti – BEAST duo package is particularly useful for three main research focus areas. These areas are species phylogenies for molecular dating, coalescent-based population genetics and measuring evolving populations (ancient DNA or time-stamped viral sequence data sets) (Drummond *et al.*, 2007). There are typically five components of an evolutionary model for a set of aligned nucleotides in BEAST (Drummond and Rambaut, 2009). These are described briefly in Table 2.5.

Table 2.5: General components of evolutionary models in BEAST

Evolutionary model component	Description
Substitution model	The substitution model is a homogeneous Markov process that defines the relative rates at which different substitutions occur along a branch in the tree.
Rate model among sites	The rate model among sites defines the distribution of relative rates of evolutionary change among sites.
Rate model among branches	The rate model among branches defines the distribution of rates among branches and is used to convert the tree, which is in units of time, to units of substitutions. These models are important for divergence time estimation procedures and producing timescales on demographic reconstructions.
Tree	A model of the phylogenetic or genealogical relationships of the sequences.
Tree prior	The tree prior provides a parameterized prior distribution for the node heights (in units of time) and tree topology.

Source: Adapted from Drummond and Rambaut (2009, p569)

Before initialising BEAST, an XML file containing *a priori* estimates and parameters for the analysis is specified in the BEAUti programme. BEAUti relies strongly on prior knowledge about the behaviour of the data to be analysed. It is thus important that as much information about the sequences is known before creating an XML file for BEAST in BEAUti.

Sequences to be analysed by BEAST are uploaded into the BEAUti package. This can be done by either uploading a single alignment (which may contain several related/unrelated sequences) or multiple alignments. In the latter instance, sequences can be partitioned and parameters linked for similar partitions. For example, if sequence alignments from five different species are uploaded, three of which share similar characteristics, it is possible to link the site models and/or clock models of these three alignments. In some instances, it is also possible to link tree models. When partitions are linked, a similar set of parameters will apply to each of the linked partitions for the BEAST analyses. This also simplifies specifying priors for linked models in that priors set for one of the partitions will apply equally to the others (e.g. in the example above, one set of priors needs be specified and will apply to all three linked species).

BEAUti also provides the option of using time-stamped data from sequences to infer dates for the sequences through its “tip dating” option. In tip dating, time stamps on sequences are guessed by BEAUti or can be manually entered. Guessing the tip dates is only possible if some form of a date (either in years, months or days) is specified somewhere in the sequence name. BEAUti then extracts the dates based on the specifications given by the user. These dates serve to assist BEAST to analyse sequences within the right timeframe(s).

After tip dating, site models are selected for the sequences. The number of models specified depends on the number of partitions and on how many of them have their models linked. Linked models will share the same site model. Site model in BEAUti specifically refers to the substitution model selected by programmes such as jModeltest. The inherent site models in BEAUti are JC69, HKY, TN93, and GTR substitution models. Parameters for these models can either be estimated by BEAST or specified from prior knowledge (e.g. from PAML, PAUP*, jModeltest, *inter alia*).

Once site models have been specified, the next step in BEAUti is to specify the clock model(s) for analyses. This is done to estimate divergence times. An advantage of assuming a molecular clock is that it can simplify phylogenetic reconstruction and increase reconstruction accuracy (Posada, 2009). The available clock model options available in BEAUti include the strict clock, relaxed clock exponential, relaxed clock lognormal, and random local clock. The strict molecular clock model essentially refers to the molecular clock hypothesis holding true. The molecular clock hypothesis assumes that sequence divergence accumulates at a roughly constant rate over time (Vandamme, 2009). This assumes that the evolutionary rate among lineages in a phylogeny remains uniform or constant over the entire tree (i.e. global clock rate) (Drummond *et al.*, 2007; Drummond and Rambaut, 2007). Evolutionary rates, however, are dependent on many factors, including the underlying mutation rate, generation times, metabolic rates in a species, population sizes, and selective pressure (Vandamme, 2009). As such, real molecular data frequently violates a strict molecular clock assumption.

Relaxed molecular clock models, on the other hand, assume independent rates of substitution on different branches, with one or two parameters that define the distribution of rates across branches (Drummond *et al.*, 2007). For the reason that there does not exist any *a priori* correlation between a lineage's rate and that of its ancestor, BEAST's relaxed molecular clock models are called "uncorrelated" clock models (Drummond *et al.*, 2007). When using the relaxed molecular clock models, the rate for each branch is drawn from an underlying exponential or lognormal distribution (Drummond *et al.*, 2007). The local clock model assumes that each branch within a phylogeny has its own rate of substitution.

All the parameters and hyperparameters specific to the models defined in the site model(s) and clock model(s) are listed in a *priors* window. It is possible to set up the prior distributions on these parameters, to define calibration nodes and calibration densities, and to specify a tree model. According to Heath (2015), an important, yet often overlooked, prior is the *tree prior*. This model describes how speciation events are spread over time. When combined with a model for branch rate, this model allows *relative* divergence time estimation (Heath, 2015). Some of the models contained in this prior include the Calibrated Yule Model, Birth Death Model, Coalescent Constant Population Model, Coalescent Bayesian Skyline, Fossilized Birth Death Model, Sampled Ancestor Fossilized Birth Death Skyline Model, to mention a few (Stadler *et al.*, 2013; Bouckaert *et al.*, 2014; Heath, 2015). Based on the

selected method, the option to set the estimated time of origin to the last sample becomes available or the root height. This gives BEAST a realistic period from which to extrapolate divergence over time.

Once priors have been set, the last steps in BEAUti are to specify the length of the Markov chain, sample frequency and file names for logging tree files and trace logs for each MCMC iteration. Thereafter the BEAUti file is saved as an XML file to serve as input into BEAST.

2.3.5.2 Tracer

After the XML file from BEAUti is loaded onto BEAST and BEAST commences with the MCMC runs, two main output files are produced, namely, the trace log file and tree log files. These files contain records of each iteration of the MCMC run and thus have very large amounts of data. It is thus unfeasible to review the data contained in these files by simply opening them in a spreadsheet program or a tree viewing program (Heath, 2015). Fortunately, BEAST has a general utility program for summarising and visualising posterior samples from Bayesian inference using MCMC. Tracer is a cross-platform, java program for summarizing posterior samples of scalar parameters. This program is necessary for assessing convergence, mixing and determining an adequate burn-in.

The effective sample sizes (ESS) of parameters are important measures in an MCMC analysis. The ESS is a measure used to evaluate mixing behavior (Lemey, Salemi and Vandamme, 2009). It is an indication of the number of independent samples that the trace is equivalent to or the number of effectively independent draws from the posterior in the sample (Lemey *et al.*, 2009; Heath, 2015). This is calculated as the chain length (excluding the burn-in) divided by the auto-correlation time (ACT), which is the average number of states in the MCMC chain that two samples have to be separated by for them to be uncorrelated). A Low ESSs reflects a high degree of (auto)correlation among samples that may be caused by poor mixing. Ideally, Markov chains should be run long enough and sufficiently sample the stationary distribution so that the ESS values of parameters of interest are all high (i.e. ≥ 200) (Drummond and Rambaut, 2009; Heath, 2015).

When Tracer is first opened, the “posterior” trace is selected and various statistics of this trace are shown under the “estimates” tab. The right-hand side of the estimates tab contains

a table of calculated statistics for each selected trace. These statistics include (Drummond and Rambaut, 2009):

- i. *Mean*
The mean value of the samples (excluding the *burn-in*).
- ii. *Stdev*
The standard error of the mean. This takes into account the effective sample size so a small ESS will give a large standard error.
- iii. *Median*
The median value of the samples (excluding the burn-in).
- iv. *The highest posterior density (HPD)*
Is the shortest interval that contains 95% of the sampled values.
95% HPD Lower – The lower bound of the HPD interval.
95% HPD Upper – The upper bound of the HPD interval.
- v. *Auto-Correlation Time (ACT)*
The average number of states in the MCMC chain that two samples have to be separated by for them to be uncorrelated (i.e. independent samples from the posterior). The ACT is estimated from the samples in the trace (excluding the burn-in).
- vi. *Effective Sample Size (ESS)*
The effective sample size (ESS) is the number of independent samples that the trace is equivalent to. This is calculated as the chain length (excluding the burn-in) divided by the ACT.

When satisfied with the BEAST output in Tracer, trace logs and tree logs can be combined into a single trace log file and a single tree log file respectively, using a programme called Logcombiner. Tree topologies, branch rates, and node heights can then be summarised using the program TreeAnnotator, which produces a single summary tree from the BEAST run that can be visualized in FigTree (Heath, 2015).

2.4 HIV Analysis Online Tools

2.4.1 Stanford University HIV Drug Resistance Database

The Stanford University HIV Drug Resistance Database (HIVdb), also known as the HIV RT and Pr sequence database, is an online database used for the interpretation of drug resistance in HIV. It consists of Pr and RT sequences from published data on genotype-

treatment correlation, genotype-phenotype correlation and genotype-outcome correlations (Shafer, 2006). The Stanford University HIVdb (<https://hivdb.stanford.edu/>) aligns and compares submitted RT and Pr sequences to the HIV-1 subtype B reference strain (HXB2) (Tang and Shafer, 2012). The programme generates a report of SDRMs in Pr and RT and estimates the proportion of sequences that contain SDRMs in submitted sequence alignments. Additional features of the HIVdb is that it allows you to test for a specific DRM and it gives feedback about that DRM. Moreover, researchers are able to download resources, such as spreadsheets containing lists of all submitted DRMs with predicted resistance scores, for offline analysis.

2.4.2 Los Alamos HIV Database

The HIV databases at Los Alamos (<https://www.hiv.lanl.gov/>) contain comprehensive data on HIV genetic sequences and immunological epitopes. The website also grants access to a variety of analysis and visual tools for data analysis. These tools include the sequence database, vaccine database, immunology database, and other viruses. The HIV database consists of an updated reservoir of HIV sequence data from GenBank entries (Kuiken, Korber and Shafer, 2003). Briefly, GenBank® is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank (<https://www.ncbi.nlm.nih.gov/>) at NCBI. It is a comprehensive database that contains the nucleotide sequences of approximately 260 000 formally described species (Benson *et al.*, 2013). Sequences are obtained mainly “through submissions from individual laboratories and batch submissions from large-scale sequencing projects, including whole-genome shotgun and environmental sampling projects” (Benson *et al.*, 2013, p36). Submissions are predominantly made using the web-based BankIt or standalone Sequin programs, whereupon GenBank staff assign accession numbers to each dataset received. Information is shared or transferred daily between DDBJ, ENA and GeneBank (Benson *et al.*, 2013).

2.4.3 Prosite

Prosite (www.prosite.expasy.org) is a database that chiefly consists of protein families and domains. The database operates by grouping together different proteins based on sequence similarities into smaller families. By analysing the constant and variable properties of such groups of similar sequences, Prosite attempts to derive signatures for protein families or domains, which distinguishes family members from all other unrelated proteins. Currently,

Prosite contains patterns and profiles specific for over 1000 protein families or domains. Each of these signatures is accompanied by documentation that provides background information on their structure and function.

CHAPTER 3:

Detection of Positive Selection Pressure in Acute Phase HIV Positive Treatment Naïve Sequence Alignments

3.1 Introduction

Natural selection is accepted as one of the main principles used to organize biology. It is broadly considered as consisting of two main types, namely purifying selection and positive selection. Purifying selection acts to eliminate deleterious mutations. Positive (or Darwinian) selection, on the other hand, favours advantageous mutations which tend towards being fixed (directional selection) or towards maintaining a polymorphism (balancing neutral selection) (Hughes, 2007; Shen *et al.*, 2010). Kimura's neutral theory of 1968 hypothesised that most molecular level evolutionary changes were chance fixations of selectively neutral mutations (Kimura, 1968; Kimura, 1983; Hughes, 2007). The neutral theory thus predicts, "that most polymorphisms are selectively neutral and are maintained by genetic drift; and that most changes at the molecular level that are fixed over evolutionary time are selectively neutral and are fixed by drift." (Hughes, 2007, p365). Furthermore, Hughes (2007) argues that one of the most important predictions of neutral theory is that purifying selection will predominate in coding and functionally important regions. Consequently, functionally important sequences are conserved and evolve slowly.

The purpose of this chapter was to analyse the HIV-1 sequences from 11 acutely infected, treatment naïve participants from a cohort in Durban, South Africa. Specifically, the analyses focused on identifying the sites in Pr and parts of RT that were experiencing positive selection pressure across the viral sequences from this cohort, as it has previously been shown that positive selection is a driving force in the generation of mutations in the *pol* region of the HIV genome (Banke *et al.*, 2009).

3.2 Sample Description

The sequences analysed were obtained from a study called "Acquired and Transmitted Drug Resistance in HIV-1 Subtype-C: Implications of Novel Mutations on Replication Capacity, Cleavage and Drug Susceptibility" (Ethical approval was granted by the Biochemical Research Ethics Committee of the University of Kwa-Zulu Natal (ref. no. BE347/13)). Briefly, in the aforementioned study, fifteen samples were acquired from stored plasma

samples collected from acutely infected patients enrolled in a study entitled “Females Rising through Education, Support and Health (FRESH)”. These samples were sequenced using both Sanger and UDP techniques. A detailed description of methods used to extract both SS and UDP viral RNA and preparation of sequences for analysis can be found in Singh (2015) (<https://researchspace.ukzn.ac.za/xmlui/handle/10413/14651>).

3.2.1 Sequence Quality Control

AVA software was used to obtain mutation prevalence and unique consensus sequences, which were further processed. Short consensus sequences (less than 90% of the expected length) were filtered out. Amino acid variant calling was realized using in-house perl code on pairwise alignments over HXB2R reference sequence, discarding those sequences with in-frame STOP codons. In order to discard strand-dependent sequencing errors, only variants which were present both in the forward and reverse strand and presented a forward / reverse prevalence frequency ratio within 1 log were accepted. GSS scores were calculated using Stanford HIVdb guidelines (Noguera-Julian *et al.*, 2013). Verification of mutations was conducted by manually inspecting flowgrams at positions of interest. Additionally, the frequencies of mutations in both the forward and reverse strands were compared using a two-tailed Fischer’s exact test. If variants were detected disproportionately (i.e. $p < 0.001$) in one direction they were not considered as low-frequency DRMs (Wang *et al.*, 2007; Singh, 2015). Only mutations present at frequency $\geq 1\%$ were considered.

3.2.2 Sequence Alignment

The fifteen samples acquired from the FRESH study were sequenced using Sanger and UDP sequencing techniques. Three of the 15 FRESH samples did not work for Sanger sequencing. These were for participants 268, 272 and 312. Fourteen of the 15 samples sequenced using UDP produced sufficient reads for data analysis. Sample 271 was the only sample that did not produce sufficient reads for data analysis. This resulted in 12 consensus sequence alignments for the Sanger technique (SS) and 14 for UDP. Consensus sequences were exported in “Fasta” format into ClustalX (v2.1) (<http://www.clustal.org/>) (Thompson *et al.*, 1997; Larkin *et al.*, 2007) where sequences were automatically aligned against each other and the South African subtype C reference sequence Ref.C.ZA.04.04ZASK146.AY772699 (henceforth will be referred to as RefC.ZA). Aligned sequences were manually edited using BioEdit Sequence Alignment Editor (v7.2.0) (Ibis Biosciences, An Abbott Company, CA.

USA). To standardize analyses, the sequence alignments from both Sanger and UDP were each trimmed to 350 codons in length. This region encompassed to the whole of Pr and the first 251 codons of RT. In addition, only consensus sequences that were common to both the SS and UDP alignments were used in this study. This resulted in 11 SS and 11 UDP sequences being used for data analyses (unless stated otherwise). Pairwise alignments for sequence similarities were performed on each pair of sequences (e.g. 036SS and 036UDP) using the Blossum62 similarity matrix. Eight sequence pairs were identical in both SS and UDP alignments. Three sequence pairs, namely 079, 093 and 271 were 99.6% identical, differing by only one AA between each pair. For the purposes of this study, both sets of corresponding alignments are assumed identical.

3.3 Methods

Positive selection was performed on both the SS and UDP datasets using two different methods. In the first method, the PAML package was used. This method used a multi-step process to generate phylogenetic trees needed to accompany the sequences to perform the analyses. The appropriate models of substitution were selected using j-Modeltest. Thereafter, maximum likelihood trees were constructed in PAUP* to serve as input trees for PAML analyses. The second method used the HyPhy package in Datamonkey. The specifications for these methods are discussed in the next sections.

3.3.1 Best-fit Model of Substitution Selection using j-Modeltest

Sanger and UDP sequence alignments were subjected to model testing using the j-Modeltest (v2.1.10) package in Linux. The starting tree for each model was based on ML with four categories in the likelihood settings (nCat=4). These categories were the number of substitution schemes; the inclusion of models with equal/unequal base frequencies (+F), models with/without a proportion of invariable sites (+I), and models with/without rate variation among sites (+G). This resulted in 11 substitution schemes and 88 candidate models of substitution. The “best” base tree search topology was used. This search operation selected the best tree topology by computing both nearest neighbor interchange (NNI) and subtree pruning and regrafting (SPR) algorithms and picked the better of the two. Once likelihood scores were calculated, corrected Akaike Information Criterion (AICc) analyses were performed. The best models of substitution using AICc were then selected by j-Modeltest together with parameter estimates for both SS and UDP alignments (see Table 3.1

for estimates). Unrooted phylogenetic trees for both the SS and UDP alignments were then drawn by the PhyML package in j-Modeltest. Since PhyML lacked flexibility in tree construction, the PAUP* blocks containing tree construction parameters from j-Modeltest were used for better tree construction in PAUP*.

3.3.2 Tree Construction using PAUP*

The PAUP* blocks specifying the best models of substitution for SS and UDP alignments, as well as the ML criterion specifications produced by j-Modeltest for these alignments, were appended to the end of the nexus formatted sequences of each alignment cluster as per their specific best AICc models of substitution. Sequence alignments with PAUP* blocks attached to them were imported into PAUP*(v4). These served as starting points for PAUP* phylogenetic tree construction. Based on the criterion obtained from the PAUP block, PAUP* found models of substitution that corresponded to the TIM1+I+G model for SS and to the TIM3+I+G model of UDP. Once successfully imported, best tree construction was commenced within PAUP*. Initially, the distance criterion was selected and rates set to gamma. Using these settings, a NJ phylogenetic search was commenced. Two sets of PAUP* runs were performed for each of SS and UDP alignments. The first was based on the 11 SS of each alignment, while the second had RefC.ZA sequence included among both sequences. In the second set of runs, the outroot was set to monophyletic with RefC.ZA used as the root outgroup. Likelihood scores were then estimated and distance was set to ML and PAUP* executed. The output from this run was then fixed and served as entry parameters for the next PAUP* iteration. Estimating and fixing parameters was executed several times in succession until $-\ln L$ no longer improved (see Table 3.1 for final scores) The NJ tree that yielded the best fit tree, together with its branch lengths, was then saved.

Optimality criterion was then set to likelihood and the same estimation and fixing procedure repeated. Thereafter, a heuristic search was performed with the fixed parameters using NNI swapping and the best tree kept in the PAUP* repository. The same procedure was repeated, but this time changing the swapping from NNI to tree bisection and reconnection (TBR). The best ML trees for SS and UDP alignments were saved with their branch lengths. One thousand bootstrapping repetitions were then performed on the best ML tree for each alignment. This process was done for both sequence alignments. The trees were then viewed

in FigTree (v1.4.3) and visually analysed for peculiar clustering. The unrooted trees were saved in Newick format for input into PAML for positive selection analysis.

3.3.3 Phylogenetic Analysis by Maximum Likelihood (PAML) for detection of positively selected sites

The trees saved from PAUP*, together with their corresponding sequence alignments (in Fasta format), were analysed using PAML (v4.9d) in Linux. The codeml.ctl file was adjusted to incorporate the SS and UDP sequence data files and their respective tree structure files (see Appendix 1 for compressed codeml.ctl file). This was done for each of the two sequence alignments.

After specifying the sequence data files and the tree structure files in the codeml.ctl file, codon sequence type (seqtype) was selected, with the equilibrium codon frequencies in the codon substitution model (codonFreq) set to F3x4. Under this model codon frequencies are used as free parameters, comprising 9 codon parameters (Yang, 2007b). No clock was specified, thus allowing for rates to vary freely between branches. This suggests that for each of SS and UDP alignments 19 (2n-3) parameters (i.e. branch lengths) were estimated. This model was also thought appropriate as it has been found elsewhere that HIV-1 has a weak molecular clock (Suzuki, Yamaguchi-Kabata and Gojobori, 2000). Consequently, unrooted SS or UDP trees were included in the tree files (treefile) as per Yang (2007b)'s recommendation (i.e. unrooted trees from PAUP* were used for this analysis). Equal AA distances were assumed (aaDist). The run mode was set to perform ML estimation of dS and dN in pairwise comparisons of the codon sequences using the trees generated in PAUP* for both SS and UDP alignments (i.e. runmode=0). The one ω ratio model for all branches was selected (M=0) with site models 0, 1, 2, 7 and 8 specified (NSsites = 0 1 2 7 8). This option tested five models of ω variations among sites. Briefly, M0 is the standard one ω for all sites model; M1 and M7 have a fraction of sites with $\omega < 1$ and a fraction with $\omega = 1$; and M2 and M8 account for positive selection (in both models an extra class of sites with $\omega > 1$ is allowed). M1-M2 and M7-M8 form two pairs of models that can be used to test for the presence of positive selection using the likelihood ratio test (LRT). The two sets of models are complementary, differing only in how sites with $\omega < 1$ are treated.

The universal genetic code (icode=0) was specified. The partitioning model for codon substitution (Mgene) was set to zero. Under this model κ , ω and π are the same across genes, but proportional branch lengths (cs) differ. Alpha was fixed with a constant rate (fix_alpha=1, alpha=0) as recommended by Yang (2007b). These settings gave preference to the NSsites models specified in the previous paragraph (Yang, 2007b). The number of categories in dG (i.e. ω -distribution) of NSsites models was set to 10 (Yang *et al.*, 2000). The ancestral rate was set to option 1. This value forces codeml to calculate rates for individual sites along the sequence using the empirical Bayes procedure (Yang, Kumar and Nei, 1995; Yang, Wong and Nielsen, 2005) and to perform the empirical Bayesian reconstruction of ancestral sequences (Yang *et al.*, 1995; Yang, 2007b). Once all the codeml.ctl specifications were set, the codeml control file was executed in Linux. Codeml was executed three times for each of the alignments to test for output reproducibility. At each run, codeml was able to detect the same positively selected sites.

3.3.4 Datamonkey specifications for positive selection

Each of SS and UDP sequence alignments was analysed using the HyPhy package in Datamonkey. The specific methods used for detecting positive selection were SLAC, FEL, iFEL, REL, and MEME. Each of these methods was run three times in succession at the 95% confidence interval (CI) across four models of substitution, namely F81, HKY85, TrN93, and REV.

3.3.5 Online Analyses of positively selected sites

Sequences were also blasted on the Stanford University HIVdb to test for SDRMs in the sequence alignments. The HIV Molecular Immunology Database from Los Alamos was used to investigate whether any of the mutated positively selected sites fell within putative epitope domains and to identify the associated human leukocyte antigens (HLAs) specific to those epitopes. Furthermore, Prosite was used to identify functional motifs possibly impacted by mutations in the FRESH cohort sequences. The output from Prosite was visualised using the Genedoc (v2.6.001) software package.

3.4 Results

3.4.1 Best-fit Models of Substitution Selected by j-Modeltest and PAUP* Tree

Construction

The TIM1+I+G model was selected as the best fit for the SS alignments with a $-\ln L$ score of 3279,294. The best AICc model selected for UDP alignments was TIM3+I+G, with a $-\ln L$ of 3250,714. Each of SS and UDP had 28 optimized free parameters (K) (i.e. substitution parameters + 19 branch lengths + topology). Similar nucleotide frequencies were detected in both SS and UDP by their respective best model estimates. However, the nucleotide substitution rates ($R_x [YZ]$) for the SS and UDP alignments differ considerably. For instance, $R(e)[CT]$ is 9,750 for SS, while it is 17,727 for UDP alignments. These differences suggest that the SS and UDP alignments differ to some extent, and are therefore not identical.

Overall, the highest rates of nucleotide substitution were between $A \leftrightarrow G$ and $C \leftrightarrow T$ for both SS and UDP alignments. These were 6,961 and 11,635 $A \leftrightarrow G$ substitutions, and 9,750 and 17,727 $C \leftrightarrow T$ substitutions for SS and UDP alignments respectively. This indicates that the rates of transition substitutions were more favoured in these sequences over any of the transversion rates. This is in keeping with other findings that indicate that transition rates are generally more than twice that of transversion rates (Bos and Posada, 2005).

Table 3.1: j-Modeltest and PAUP* output for SS and UDP alignments

Parameter	Sanger		UDP	
	j-Modeltest	PAUP	j-Modeltest	PAUP
Model selected	TIM1+I+G	GTR+I+G	TIM3+I+G	GTR+I+G
$-\ln L$	3279,294	3276,802	3250,714	3249,402
K	28		28	
freq A	0,394	0,390	0,393	0,390
freq C	0,169	0,166	0,164	0,165
freq G	0,208	0,213	0,212	0,214
freq T	0,229	0,232	0,231	0,231
R(a) [AC]	1,000	2,392	1,999	2,225
R(b) [AG]	6,961	12,264	11,635	10,709
R(c) [AT]	0,568	1,072	1,000	0,863
R(d) [CG]	0,568	0,822	1,999	0,904
R(e) [CT]	9,750	17,545	17,727	15,443
R(f) [GT]	1,000	1,000	1,000	1,000
p-inv	0,512	0,498	0,512	0,501
gamma shape	0,719	0,678	0,687	0,646ichel

PAUP* identified the GTR+I+G model of substitution as the best fit models corresponding to the j-Modeltest proposed models for both SS and UDP. The $-\ln L$ scores for using PAUP increased slightly for both SS and UDP. Performing a likelihood ratio test (LRT) to compare whether the proposed model by PAUP* (H_a) was a better fit than the model proposed by j-Modeltest (H_o) was not supported using the chi-squared distribution (see Table 3.2). This is evident in the LRT scores falling within the acceptable range of the chi-squared distribution (i.e. LRT scores smaller than the critical value of 9,210 for two degrees of freedom at the 0.01 significance level). This indicates that statistically 99% confidence can be placed on the similarity of the ‘best fit’ models selected by j-Modeltest and PAUP*.

Table 3.2: Likelihood Ratio Test for Comparison of ‘best fit’ models of substitution for SS and UDP using standard Chi-squared distribution

Component	Sanger		UDP	
	Modeltest	PAUP	Modeltest	PAUP
Hypothesis	H_o	H_a	H_o	H_a
Model	TIM1+I+G	GTR+I+G	TIM3+I+G	GTR+I+G
$\ln L$	-3279,294	-3276,802	-3250,714	-3249,402
$\Delta \ln L$	2,492		1,312	
$2\Delta \ln L$ (i.e. LRT)	4,984		2,624	
p-value	0,083		0,269	

Critical value: $\chi^2_{2,1\%}=9,210$

Similar to j-Modeltest, the overall highest rates of nucleotide substitution were between $A \leftrightarrow G$ and $C \leftrightarrow T$ for both SS and UDP alignments. However, using PAUP* the differences between SS and UDP estimates were not that dissimilar. These were 12,264 and 10,709 $A \leftrightarrow G$ substitutions, and 17,545 and 15,443 $C \leftrightarrow T$ substitutions for SS and UDP respectively.

3.4.2 Sites under Positive Selection Pressure detected by PAML

Table 3.3: LRT for evidence of positive selection in SS and UDP alignments using standard Chi-squared distribution

Component	Sanger				UDP			
Model	M1	M2	M7	M8	M1	M2	M7	M8
Hypothesis	H _o	H _a	H _o	H _a	H _o	H _a	H _o	H _a
lnL	-3172,2	-3158,46	-3176,39	-3158,74	-3134,78	-3122,37	-3140,32	-3124,48
Δ lnL	13,742		17,651		12,403		15,839	
2 Δ lnL (i.e. LRT)	27,485		35,302		24,806		31,678	
p-value	2E-06		1E-06		5E-06		1E-06	

Critical value: $\chi^2_{2,1\%}=9.210$

Testing the models for positive selection (M2 and M8) against those that did not allow for positive selection (M1 and M7) provided strong support for a fraction of the sites evolving under positive selection (see Table 3.3 and 3.4). This was achieved by comparing the LRT's between M1 and M2, and between M7 and M8. Both LRT's were greater than the critical value of 9,210 (2df, 1%) allowed by the chi-squared distribution (Yang *et al.*, 2005; Yang, 2007b). Furthermore, the p-values provided strong statistical evidence in favour of positive selection (i.e. p-values ≤ 0.000001 , which are highly significant) (Bielawski and Yang, 2003; Yang, 2007b).

Table 3.4: Likelihood values and parameter estimates under models of variable ω ratios among sites for Sanger and UDP alignments

Model code	Sanger						UDP					
	-LnL	Kappa (ts/tv)	Parameter estimates				-LnL	Kappa (ts/tv)	Parameter estimates			
M0(1)	3242,6108	5,6913	ω	0,1236			3212,5121	5,3695	ω	0,1237		
M1(2)	3172,1992	5,8036	p	0,9150	0,0850		3134,7764	5,5245	p	0,9115	0,0885	
			ω	0,0439	1,0000				ω	0,0382	1,0000	
M2(4)	3158,4569	6,2366	p	0,9147	0,0789	0,0064	3122,3736	5,9251	p	0,9113	0,0818	0,0069
			ω	0,0465	1,0000	9,3378			ω	0,0405	1,0000	8,7355
M7(2)	3176,3859	5,8001	p	0,1013			3140,3231	5,4944	p	0,0876		
			q	0,6462					q	0,5733		
M8(4)	3158,7350	6,2024	p	0,9922	0,1448	0,0078	3124,4839	5,8857	p	0,9916	0,1225	0,0084
			ω			8,0546			ω			7,6481
			q		1,1123				q		0,9501	

M0 = 1 ω ratio; M1 = nearly neutral; M2 = positive selection; M7 = β ; M8 = β and ω . Numbers in brackets next to the model codes indicate the number of free parameters in the ω distribution

The one-ratio model (M0), which assumes one ω ratio for all sites, gave an average ω ratio of 0,1236 for SS and 0,1237 for UDP. This indicates that on average, purifying selection was the dominating force during the evolution of the sequence alignments. Model 2 indicates that less than 1% of sites in both SS and UDP were under positive selection with $\omega = 9,3378$ and $8,7355$ for SS and UDP alignments respectively. Similarly, estimates under model 8 (β and ω) also suggested that less than 1% of the sites in the alignments were under positive selection with $\omega = 8,0546$ for SS and $\omega = 7,6481$ for UDP.

Since the LRTs mentioned earlier were significant, the Bayes Empirical Bayes (BEB) (Yang *et al.* 2005) procedure for identifying positively selected sites based on posterior probabilities for site classes was used (Yang, 2007a). The BEB was implemented under M2 and M8. The sites identified for SS and UDP are shown in Table 3.5.

Table 3.5: Positive selection using PAML

Model 2: Positive Selection (3 categories)				
Codon site	SS	UDP	Mutation(s) [wt Δ mut*]	Nearest DRM site [wt Δ mut**]
Pr63	√†	√	L63HPSVT	Q58E
RT123	√†	√	G123SND	V118I
Model 8: beta & $\omega > 1$ (11 categories)				
Pr19		√	T19VIL	K20TV
Pr63	√†	√	L63HPSVT	Q58E
Pr67		√	C67Y	G73ACDSTV
Pr82		√	V82I	V82ACFLMST
RT39		√	E39DKT	E40F
RT123	√†	√	G123SND	V118I
RT169		√	E169DA	V179DEFL
RT174		√	K174QR	V179DEFL
RT211		√	R211KQ	L210W
RT214		√	F214L	T215YFISCDVE
RT245		√	Q245KE	K238T
RT251		√	S251ID	K238T

* wt AAs based on Ref.C.ZA.04.04ZASK146.AY772699 consensus sequence; **DRMs taken from Stanford University HIV Drug Resistance Database (2017); † Pr > 99%, otherwise Pr > 95% for the rest

Codeml M2 detected two sites for positive selection in both sets of alignments using BEB. It detected Pr63 and RT123 in both sequence alignments with p-values < 0.01. These two

sites were also detected using M8 for both SS and UDP. In addition, M8 also detected 10 more sites for positive selection in the UDP sequences than in the SS sequences when compared to M2. These ten additional sites in the UDP alignments were at Pr19, Pr67, Pr82, RT39, RT169, RT174, RT211, RT214, RT245, and RT251 (all at $p < 0.05$). Protease V82I was the only identified site for positive selection that was a known DRM based on the Stanford SDRMs list. In an earlier study, Pr63 was also identified as a DRM (Chen, Perlina and Lee, 2004).

3.4.3 Sites under Positive Selection using Datamonkey

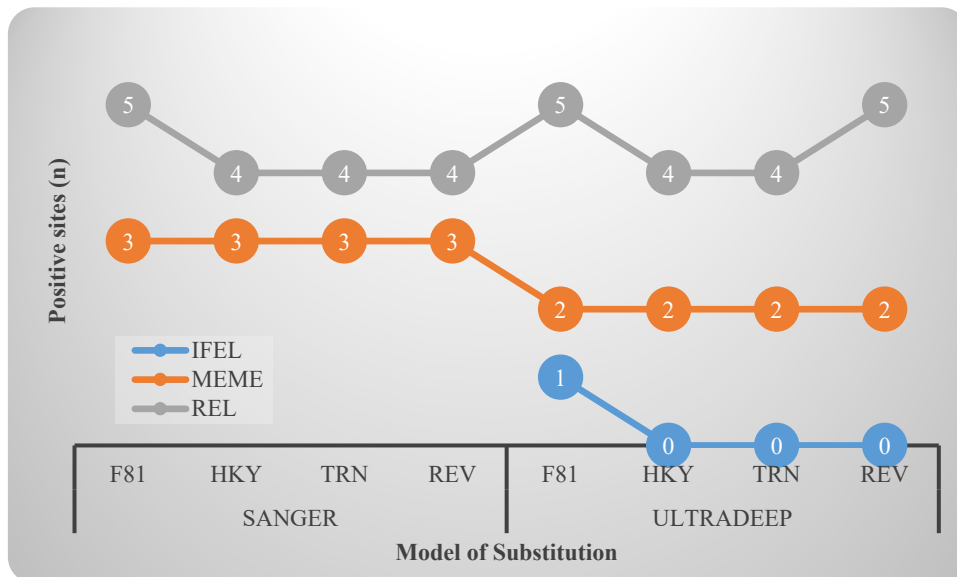


Figure 3.1: Positive selection detection for Sanger and UDP Sequences

numbers in the circles indicate the number of positive sites

Figure 3.1 shows that the REL method detected the most sites for positive selection of the conventional methods used by the HyPhy package in Datamonkey. It consistently did so for the SS and UDP sequence alignments and across all four models of substitution used to analyse the alignments at the 5% level of significance. In the SS alignments, REL identified five possible sites for positive selection using the F81 model, four sites for each of HKY85, TrN and REV models. For UDP alignments, REL identified five sites using F81 and REV; and four sites for each of HKY85 and TrN93, and an additional site in REV (i.e. five sites). Internal branch FEL (iFEL), on the other hand, only detected one site for positive selection in UDP alignments using MEME, but did not detect any sites in the SS alignments. The SLAC and FEL methods did not detect any sites for positive selection. A summary of the

positively selected sites identified in *pol* using REL, MEME and iFEL, together with the total frequencies detected by REL and MEME across the models of substitution, is illustrated in Figure 3.2.

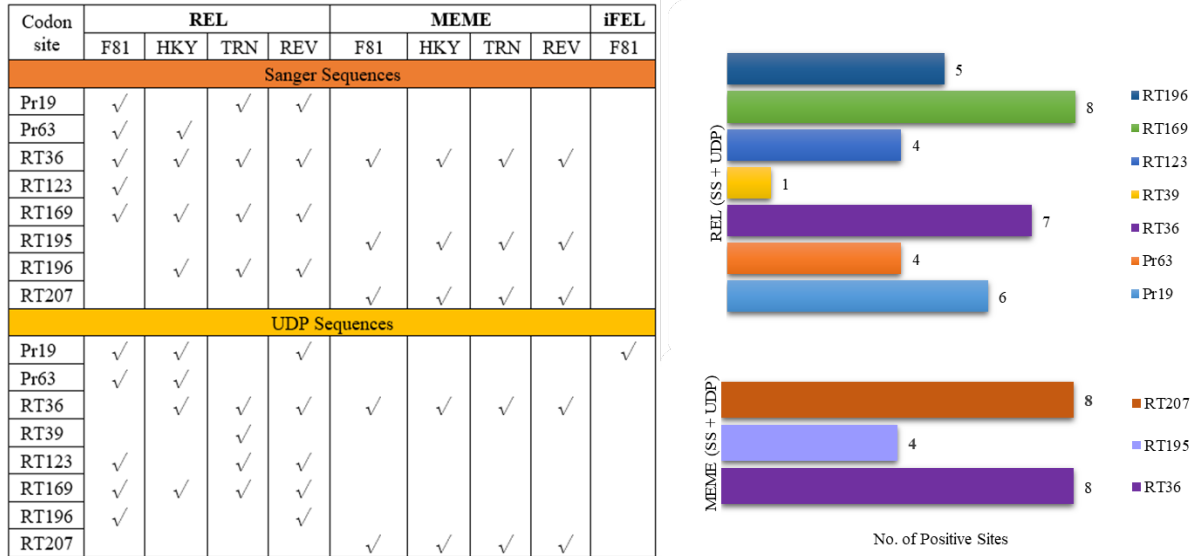


Figure 3.2: REL, MEME and iFEL positively selected sites in *pol* across F81, HKY85, TrN93 and REV models of substitution

- i) The table on the left has three broad categories of comparison (REL, MEME and iFEL in the top row). Each of these has the substitution models that detected positive sites for them as subcategories (F81, HKY, TRN and REV). These comparison categories were used for both SS and UDP sequence alignments (highlighted rows). The first column of the table indicates the codon sites that were identified by the entire HyPhy analysis at the 95% CI.
- ii) The figure on the left is a summary of all the positive sites detected by either the REL method or the MEME method across all four methods of substitution (i.e. combines both SS and UDP sites across the two methods).

Using the REL method, overall F81 detected six unique sites (i.e. a site that was common to both SS and UDP counted as one unique site) for positive selection (see Figure 3.2). Four of these, namely Pr19, Pr63, RT123 and RT169, were common to both SS and UDP alignments. Similar to F81, TrN also identified six unique sites for positive selection. However, it only detected RT positions 36 and 169 as common sites to both SS and UDP alignments. REV detected five unique sites for positive selection, with four common to both SS and UDP alignments, specifically Pr19, RT36, RT169 and RT196. HKY also identified five sites, three of which were common to both SS and UDP, namely Pr63, RT36 and RT169.

The MEME model had the same number of positively selected sites across all models of substitution (i.e. adding both SS and UDP sites). All of the identified sites were in RT with

positions 36 and 207 common to both SS and UDP alignments. Internal branch FEL, on the other hand, detected Pr19 in UDP alignments as the only site for diversifying selection.

Reverse transcriptase codon 169 was the only site selected by all substitution models using the REL method for both SS and UDP alignments (refer to the diagram on the right in Figure 4.2). This was followed by RT36, which was selected seven times (four times in SS and three times in UDP). Protease codon 19 was selected three times by both SS and UDP, while RT196 was selected 3 times in SS and twice in UDP. Protease 63 and RT123 were each identified four times.

In MEME, both RT36 and RT207 were identified as positive sites across all four models of substitution in both SS and UDP alignments. Reverse transcriptase codon 195 was also identified as a positive site, but only in the SS alignments.

Shown in Table 3.6 are the collective AA mutations in either Pr or RT corresponding to the AA sites identified in the preceding figure. It should be noted that although Table 3.6 may appear to depict similar data as Figure 3.2, it differs in that its purpose is to relate the observed positively selected sites to their closest known SDRM.

Table 3.6: Positive selection sites in Pr and RT for REL, MEME and iFEL for SS and UDP sequences and nearest known Drug Resistance Site

Codon Site	REL		MEME		iFEL	Mutation(s)	Nearest DRM site
	SS	UDP	SS	UDP	UDP	wt Δ mut*	wt Δ mut**
Pr19	✓	✓			✓	T19VIL	K20TV
Pr63	✓	✓				L63HPSVT	Q58E
RT36	✓	✓	✓	✓		A36E	E40F
RT39		✓				E39DKT	E40F
RT123	✓	✓				G123SND	V118I
RT169	✓	✓				E169DA	V179DEFL
RT195			✓			I195N	G190ASEQ
RT196	✓	✓				G196ER	G190ASEQ
RT207			✓	✓		E207ATK	L210W

* wt AAs based on Ref.C.ZA.04.04ZASK146.AY772699 consensus sequence; **DRMs taken from Stanford University HIV Drug Resistance Database (2017)

Observed mutations in Pr were T19VIL and L63HPSVT. Protease position 63 was the most highly variable locus of all the positively selected sites, toggling between six codons (i.e. wild-type plus the five mutations). Several mutations were observed in RT. These were at

A36E, E39DKT, G123SND, E169DA, I195N, G196ER and E207ATK. The more conserved sites were at position 36 and 195. Reverse transcriptase position 36 kept toggling between A and E, whereas there was only a single mutation in one of the SS alignments at position 195. The more volatile sites for positive selection in RT were E39DKT, G123SND and E207ATK.

3.4.4 Positive selection at functional sites in Pr and RT

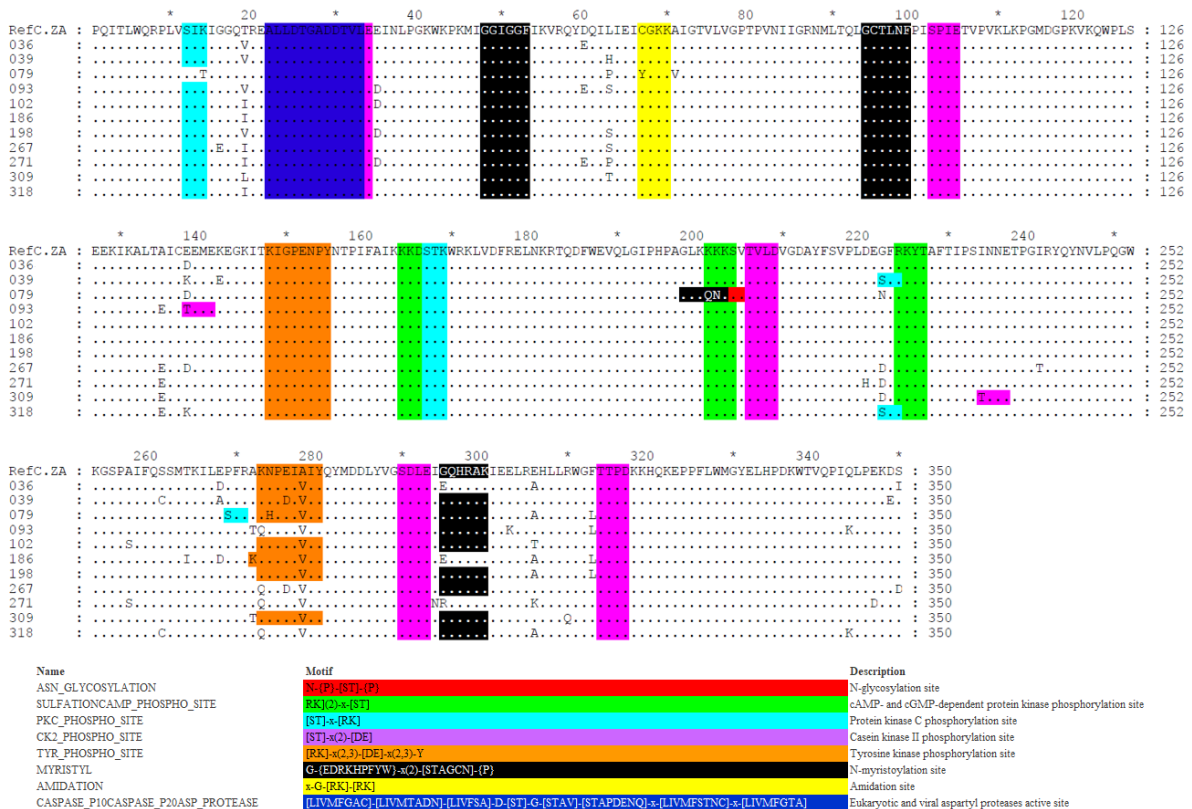


Figure 3.3: Conserved putative functional sites in Pr and RT from the FRESH study cohort

Figure 3.3 reflects the FRESH cohort’s putative functional sites and motifs in Pr and RT taken from the Prosite database and viewed in Genedoc. The putative functional regions are indicated by highlights. Conserved regions in the cohort can be seen by highlighted regions forming a band down the aligned sequences (i.e. are common across all or the majority of sequences). For example, the eukaryotic and viral aspartyl protease active site (highlighted in dark blue) was present in Pr of all the sequences, indicating that this site was a conserved region.

Table 3.7: Putative functional sites in FRESH cohort sequences

Functional Site	Positively selected sites		Negatively Selected sites		
	RT	Sequence	Pr	RT	Sequence
Eukaryotic and viral aspartyl proteases active site	-	-	23, 25, 30	-	All
Casein kinase II phosphorylation site	39	093		3	All
				42	093
				109	All
				218	All
cAMP- and cGMP-dependent protein kinase phosphorylation site	-	-		66	All
Protein kinase C phosphorylation site	123	039 & 318		124	039 & 318
				171	079
Tyrosine kinase phosphorylation site	174	All except 093, 267, 271 & 318		-	-
N-myristoylation site	196	All except 036, 186 & 271		198	All except 036, 186 & 271

Table 3.7 shows the functional sites in Pr and RT that experienced either positive or negative selection. Positive selection was based on the 16 positively selected sites identified using HyPhy and PAML, while negatively selected sites were taken from the REL method of the HyPhy package (see Appendix 2). There were no sites under positive selective pressure in Pr. However, purifying selection sites were detected in the aspartyl Pr active sites at codon positions 23, 25 and 30. One positively selected site was observed in a casein kinase II phosphorylation (CKII-P) site at RT39 (sequence 093). Most often, CKII-P sites were under purifying selection, with all the sequences having negatively selected sites at RT positions 3, 109 and 218. The only other negatively selected site in a CKII-P was at RT42 in the 093 sequence.

No positively selected sites were found in cAMP and cGMP-dependent protein kinase phosphorylation (PK-P) sites. These functional sites were conserved in all the sequences with purifying selective pressure acting only at RT66.

Sequences 039 and 318 both gained a protein kinase C phosphorylation (PKC-P) site, a consequence of their positively selected mutations at RT123. However, purifying selection was acting at the next codon position (RT124) within this “gained” PKC-P functional site. Another PKC-P functional site that had purifying selection acting on it was at RT171 in the 079 sequence. This sequence had also gained an N-myristoylation site with its DRM at RT103 (NB: this site was not a positively selected site). Moreover, positively selected sites were observed in the N-myristoylation functional sites at RT196 in 73% of the sequences

(i.e. was not detected in PID 036,186 and 271 sequences), although purifying selection was acting on the same functional sites two codons away (RT198) in the same sequences.

Positive selection was observed in 64% of the sequences in the tyrosine kinase phosphorylation (TK-P) site at RT174. Although no purifying pressure was acting on any of the TK-P functional sites, sequences 093, 267, 271 and 318 lost their TK-P functional sites due to their mutations in this region.

While neither positive nor negative selective pressure was observed at some of the functional sites, it is interesting to note that sequence 079 was the only sequence to have lost a PKC-P site in Pr (region 12-14) caused by a K14T mutation. However, this sequence was the only one to gain a PKC-P site (region RT170-172) with a P170S mutation, although negative pressure was detected at the next codon position (RT171). A CKII-P functional site was also gained by sequence 309 at the RT135-138 region by a I135T mutation in this sequence.

3.4.5 Positively selected sites in HIV subtype C epitopes

Table 3.8: Putative CTL epitopes and HLA's in Pr and RT from the FRESH cohort

Protein	AA positions within Protein	Epitope	Epitope site with mutation(s)	HLA
Pr	80-90	TPVNIIGRNML	V82I*	B*8101, B81
RT	116-124	FSVPLDEGF	G123SND	B*5702, B*5703
RT	118-127	VPLDEGFRKY	G123SND	B*3501
RT	244-252	IQLPEKDSW	Q245KE	B*5801
RT	246-254	LPEKDSWTV	S251ID	B7

Epitopes were identified based on the HXB2 reference sequence in the Los Alamos Molecular Immunology Database. Only epitopes with positively selected sites that had mutations referenced against the HXB2 consensus sequence in the Los Alamos Database are shown. The AA where mutations occurred are indicated in bold within the epitope sequence with the corresponding positive mutation in the column next to it. V82I* indicates that this mutation is a DRM. Source: Los Alamos National Laboratory (2018b)

Table 3.8 presents a summary of the positively selected sites that fell within putative epitope domains in Pr and RT and the HLAs that identify them as per the Los Alamos HIV Molecular Immunology Database. Only one epitope in Pr, TL11, contained a positively selected mutation (V82I). This epitope was recognised by the B81 HLA family. Four additional positively selected mutations were found in RT epitopes at RT116-124 (FF9), RT118-127

(VY10), RT244-252 (IW9), and RT246-254 (LV9). The HLAs specific for these epitopes are B*5702 and B*5703, B*3501, B*5801, and B7 respectively.

3.5 Discussion

This chapter presented an analysis of the intersequence divergence of 11 HIV-1 viral “strains” isolated from drug naïve participants in the FRESH study at their first blood draw following viral detection. The aim was to identify regions within *pol* (esp. Pr and RT) that were experiencing positive selective pressure. Testing for positive selection on the 22 sequences (11 SS and 11 UDP) was performed using PAML and HyPhy. A comparison of the methods and models contained in the HyPhy package for detecting sites under positive selective pressure was also performed. In addition, testing whether positively selected sites were within in functional domains and epitopes was also performed.

The HyPhy methods in Datamonkey compared were the SLAC, FEL, iFEL, REL and MEME. In each of these methods four sets of analyses were performed, each time using one of F81, HKY, TrN and REV models of substitution. Comparing the conventional methods for detecting site level positive selection (i.e. the first four methods), REL appeared to be the most sensitive to detecting diversifying selection. REL consistently detected the most sites (six) for diversifying selection across all the models of substitution used to analyse both the SS and UDP sequence alignments at the 5% level of significance. Consistency, in this instance, refers to detecting the same sites under positive selection in both the SS and UDP alignments. These alignments were at least 99.6% identical (see section 3.2.2). Consistency should be distinguished from reproducibility, which pertains to detecting the same sites with each run of analyses. The REL method is known to be the best method to infer selection from small (5-15 sequences) or low divergence alignments of the four methods because it uses the entire alignment to infer rates at each site (Poon *et al.*, 2009; Datamonkey, www.datamonkey.org). However, it is the most susceptible to selecting false positives because the distribution of rates to be fitted has to be defined *a priori*, and it may not satisfactorily model the undetected distribution of rates (Poon *et al.*, 2009; Datamonkey, www.datamonkey.org).

Using the REL method, the study found that F81 and TrN models of substitution were equally sensitive to detecting sites for positive selection. However, the F81 model was more consistent having 66.7% of its sites common to both sets of sequence alignments, compared

to the 33% of TrN. The REV method, although slightly less sensitive to detecting positive sites compared to F81 and TrN models of substitution, was the most consistent in that 80% of the sites that it detected were found in both UDP and SS sequence alignments. The HKY model had 60% consistency. The SLAC and FEL methods did not detect any sites for positive selection. This is in keeping with expectations, as these methods are known to best work with datasets ≥ 50 sequences. They also tend to be more conservative relative to the other methods (Poon *et al.*, 2009; Delpont *et al.*, 2010).

In PAML, testing models for positive selection (M2 and M8) against those that did not allow for positive selection (M1 and M7) provided strong support for a fraction of the sites evolving under positive selection for both the SS and UDP alignments. This was accomplished by comparing the LRT's between M1 and M2, and between M7 and M8 as per Bielawski and Yang (2003); Yang *et al.* (2005) and Yang (2007b). Specifically, M2 and M8 both found that less than 1% of the sites in the alignments were under positive selection (M2: $\omega_{SS} = 9.3378$ and $\omega_{UDP} = 8.7355$; M8: $\omega_{SS} = 8.0546$ and $\omega_{UDP} = 7.6481$). This also suggests that both the SS and UDP alignments were under strong purifying selection. This is in keeping with a study by Gordon *et al.* (2003), in which it was found that 95% of the sites in Pr and RT, in a cohort of 72 treatment naïve patients in KwaZulu-Natal, were under strong purifying selection.

The Bayes Empirical Bayes procedure for identifying positively selected sites based on posterior probabilities for site classes (Yang *et al.*, 2005; Yang, 2007b), implemented under M2 and M8 in PAML, identified 12 sites for diversification in the sequence alignments. Four of these were in Pr at positions 19, 63, 67 and 82; and the remaining eight were in RT at positions 39, 169, 123, 174, 211, 214, 245 and 251. Protease 63 and RT123 were detected by both M2 and M8 at a $p < 0.01$. These two codons were the only sites under positive selection pressure identified by M2. The remaining 10 sites were detected by M8 at a $p < 0.05$. The Hyphy package in Datamonkey detected nine sites for positive selection in *pol*. These were at Pr19, Pr63, RT36, RT39, RT123, RT169, RT195, RT196 and RT207 at $p < 0.05$.

Altogether 16 unique codons were identified as sites experiencing positive selective pressure using PAML and HyPhy. Approximately 69% of these sites were also identified in an earlier study by Gordon *et al.* (2003) to be under strong positive selection. These were at Pr sites

19 and 63, and RT sites 36, 39, 123, 174, 196, 207, 211, 214 and 245. In addition, Chen *et al.* (2004) also identified 50% of the 16 sites as experiencing positive selection in their study. These were at Pr19, Pr63, Pr82, RT39, RT123, RT211, RT214 and RT245. Despite the unique sequences in each alignment, purifying selection was the dominant force in the in this cohort as confirmed by PAML findings mentioned earlier. This is likely due to a large number of nucleotide substitutions being silent or synonymous, thus not altering the underlying AA sequences (Kimura, 1983; Gordon *et al.*, 2003; Hughes, 2007; Shen *et al.*, 2010).

Physical inspection of the sequences identified Pr63 as the most variable site of all the sites under positive selective pressure. This site toggled between six AAs (*wt* plus the five AAs viz, L63HPSVT). This lends strong support to the PAML and HyPhy packages selecting this site as one under strong Darwinian selective pressure. The second most variable site in Pr was T19VIL. This site was identified by iFEL as a site for diversifying evolution. The most conserved codon sites were in RT at positions 36 and 195 (viz, A36E and I195N). These sites were only detected in HyPhy and not by PAML as sites for positive selection. Reverse transcriptase position 36 kept toggling between A and E, whereas there was only a single mutation in one of the SS sequences at position RT195. The more volatile sites for positive selection in RT had mutations E39DKT, G123SND and E207ATK.

The only site under positive selective pressure that was a known DRM, as per Stanford University HIV Drug Resistance Database (2017), was in Pr (V82I). This site was detected using PAML. According to the Stanford University HIV Drug Resistance Database (2017), the V82I mutation is a highly polymorphic mutation that is the consensus AA in subtype G viruses and is not selected by PIs. Chen *et al.* (2004) also identified Pr63 (L63P) as a DRM site in their study of 40 000 HIV-1 sequences.

The extraordinary adaptability of HIV-1 to readily escape virus-specific CTL responses by selecting mutations that can hinder proper viral epitope processing, epitope binding to HLA molecules or detection by specific T cell receptors (Llano *et al.*, 2009; Llano *et al.*, 2013) has complicated effective vaccine development against the virus (Blanco-Heredia *et al.*, 2016). This has been typically understood to be caused by viral mutations in epitopes and the associated loss of targeting of immunodominant epitopes by protective HLAs, for example, B81 and B57 (Pereyra *et al.*, 2014). In a study by Acevedo-Sáenz *et al.* (2015),

which explored immune escape mutations in Pr and RT among 614 Columbian patients, four of the sites selected above were identified as mutations within CD8⁺/cytotoxic T-cell (CTL) epitopes. These positively selected sites were at Pr19 and Pr82 as well as RT39 and RT211. Protease codon 19, RT39 and RT211 were associated with inducing CTL responses, while Pr82 was implicated as an escape mutation.

Five epitopes containing positively selected mutations were detected in the present study, four of these were in RT and one in Pr. These epitopes were specific to HLAs B81, B*8101, B*5702, B*5703, B*3501, B*5801 and B7 (Los Alamos National Laboratory, 2018b). HLA-B57, B*5801 and B*8101 have been previously associated with low viremia (Kiepiela *et al.*, 2006; Ntale *et al.*, 2012) through the protective role that they played against HIV-1 replication (Kaslow *et al.*, 1996; Altfeld *et al.*, 2003; Kiepiela *et al.*, 2006; Altfeld and Goulder, 2011; Illing *et al.*, 2018). The HLA-B7 supertype, on the otherhand, has been associated with high viral loads and the rapid progression of the disease (De Groot *et al.*, 2008). Similarly, HLA-B35 has been associated with the rapid progression of the disease, however, HLA-B*3501 did not have any notable impact in this regard, except to prolong disease duration (Huang *et al.*, 2009).

According to Kaslow *et al.* (1996); Altfeld *et al.* (2003) and Altfeld and Goulder (2011), HLA-B57 family members are known for their association with the elite controller phenotype of HIV-infected individuals. It has been hypothesised that the protective effect of HLA-B57 is because of its more efficient presentation of immunogenic HIV peptides to antiviral CTLs than by non-protective HLA variants (Illing *et al.*, 2018). However, a study by Klooverpris *et al.* (2012) on a cohort of >2 000 HIV subtype C infected participants from southern Africa, found that small differences in HLA-B57 had significant impact on the immune control of HIV. They found that HLA-B*5703 was associated with a lower viral-load set point than HLA-B*5702 and HLA-B*5801. Moreover, HLA-B*5703 was associated with the lowest viral-load set point of the three, but also had the highest number of Gag epitopes targeted and the highest number of Gag escape mutations selected (Klooverpris *et al.*, 2012). Illing *et al.* (2018) confirmed this latter finding in their study exploring the impact of small changes in the peptide antigen presentation of HLA-B*5701, HLAB*5703 and HLA-B*5801. Their findings indicated that HLA-B*5703 more readily escaped the benefits of peptide editing within peptide-loading complexes. This may suggest that mutations in HLA-B*5703 epitopes restricted the potential benefits that positively

selected mutations might otherwise have passed on to the virus. Consequently, acquired mutations within HLA-B*5703 did not necessarily translate to viral replication advantage.

HLA-B35 has been shown to correlate with predisposition to faster HIV-1 progression towards AIDS (Al-Jabri, 2007). Although HLA-B*3501 is thought not to have any detectable impact on HIV-1 disease progression, it did, however, have some form of impact on the duration of HIV-1 (Huang *et al.*, 2009). A study by Huang *et al.* (2009), for example, found that HIV-1 infection was prolonged in the dendritic cells isolated from HIV-1 treatment-naïve participants who possessed HLA-B*3501 when compared to those that had HLA-B*3503.

There were no positively selected sites in any of the putative functional domains in Pr. However, three purifying mutations in the aspartyl Pr active site were observed. The functional domains in Pr were consistently conserved among sequences. Only participant 079 lost a PKC-P site (Pr12-14) through a K14T mutation. Overall, this suggests that in the absence of drug pressure, these domains might play an important role in acute viral survival against initial host immune responses known to occur at the early stage of viral infection in this cohort (Ndhlovu *et al.*, 2015). In addition, the conserved phosphorylation sites are likely regions where the virus modulates replication, as phosphorylation / dephosphorylation is known to be an important regulatory mechanism that activates / deactivates several enzymes (Ardito *et al.*, 2017). Phosphorylation thus plays an important role in regulating proteins that interact with nucleic acids, including RNA and DNA polymerase (Gordon *et al.*, 2003).

Although RT had several conserved functional domains, some sequences had either lost or gained functional sites. Sequence 093, for example, had a Thr mutation at a positively selected site (RT39), which gained it a CKII-P site at RT39-42. The gain in a CKII-P site may indicate that positive pressure acting on this viral strain pressured it to gain conformational dexterity when interacting with other molecules (e.g. change from hydrophobic apolar to hydrophilic polar due to phosphorylation) (Ardito *et al.*, 2017). This sequence, however, lost a TK-P site at RT174-181 at another positively selected site (RT174). Two other sequences also lost this functional site. Interestingly, this site was the only functional domain that did not have any codons that were under purifying selection. Given that cancer studies have shown that signaling pathways regulated by protein kinases contributed to the onset and progression of most cancers, one would expect that conserving

TK-P functional sites would be advantageous to the virus because of host-factor failure to modulate cell replication at these sites (Ardito *et al.*, 2017). It is, however, unclear what the consequences are to the virus losing these functional sites.

Three sequences also lost their N-myristoylation functional sites at RT196-201. It has been shown that N-myristoylation is an important evolutionarily conserved modification of proteins implicated in different physiological processes like cell cellular signalling, protein–protein interactions, targeting of proteins to endomembrane and plasma membrane systems, proliferation, differentiation, survival, and apoptosis (Wright *et al.*, 2010; Udenwobele *et al.*, 2017). Mutations in Gag that caused it to lose myristoylation functionality resulted in Gag mutants failing to bind tightly to the plasma membrane and subsequently unable to assemble into active viral particles [Li *et al.* (2007) as cited by Udenwobele *et al.* (2017)]. Loss of N-myristoylation functional sites, therefore, appear to come at a high cost for viral replication and survival.

PID 079 was the only sequence to eliminate a conserved functional site and replace it with parts of other functional sites. Although this site was not a positively selected site, this sequence replaced its cAMP and cGMP-dependent PK-P site (RT102-105) with parts of an N-myristoylation functional site and a partial N-glycosylation site. This sequence also inherited a K103N DRM, which may have been an advantageous mutation with which it both evaded drug pressure and found a way to replace a conserved site to maintain its survival. Assuming that 079 was infected by a person that was exposed to ARTs, it could also imply that this site was critical for viral functioning because the virus evolved a way to maintain its use despite the drug pressure within that host.

CHAPTER 4:

Ancestral Sequence Reconstruction using PAML and tMRCA estimation using BEAST

4.1 Introduction

The five participants that had multiple timepoints (TPs) and on which UDP was performed served as datasets for ancestral reconstruction and for estimating tMRCA. The number of sequences in each alignment were: Participant ID (PID) 036 (4 TPs), 079 (3 TPs); 267 (3 TPs), 268 (2 TPs) and 271 (4 TPs). BEAST software was used to estimate the tMRCA for each PID using the MCMC method and ancestral sequence reconstruction was performed using codeml in the PAML package.

4.2 Method

4.2.1 tMRCA tree construction using BEAST

A best-fit model of substitution was selected for four of these alignments in j-Modeltest. A substitution model could not be found for PID 268 as it only had two available timepoints. (Attempts at bypassing this in BEAST using the JC model of substitution did not work either because of the limited number of sequences in this alignment). The Bayesian Markov Chain Monte Carlo (MCMC) method, implemented in BEASTv2.4.8, was used to estimate phylogeny and the time to the most recent common ancestor (tMRCA).

The XML file for the BEAST analysis was created in the BEAUti programme. Each of the alignments had individual XML files prepared for use in BEAST and were uploaded as single alignments containing two or more sequences. Each sequence within the four datasets was tip labeled in “days since some time in the past”. This restricted the time scale to days rather than months or years. This was important because it made it possible to express the tMRCA estimate in days (esp. since the time between the first and last TPs for each participant did not exceed a year). The HKY85 model of nucleotide substitution was used as site model (as per jModeltest). All other site model parameters were estimated. An uncorrelated relaxed lognormal (URLN) clock was used for each dataset as recommended by Drummond *et al.* (2007). The URLN clock gives an indication of how clock-like the data is (as measured by the ucldstdev parameter).

The sampled ancestor fossilised Birth Death Skyline Model was used as tree prior with the origin set approximately 10 days longer than the last visit for each participant (i.e. approximately 10 days before the first sequence in each alignment partly to cater for the eclipse phase of acute infection illustrated in Figure 1.4, section 1.7). All other priors were estimated. The MCMC length was set to 60 million, with 10% burn-in. This discarded the first 6 million iterations of the MCMC run making it easier for convergence. Logging of parameter estimates and trees were set to be collected every 1 000 steps to build a posterior distribution of parameters.

The XML file was then uploaded into BEAST and two independent runs (except for PID 271, which was run three times) of 60 million steps each in the Markov chain were performed using BEAST. In addition, the second run for PID 036 was 456 million steps because ESS's were relatively low for some variables of interest (Convergence seemed unaffected by the increase in the MCMC chain from the 60m to 456m). Random samples of rooted phylogenetic trees from the posterior distribution from each dataset were generated by BEAST. Parameter estimates from each BEAST run were checked for convergence using Tracer (v1.7). Log files for identical runs of the Markov chain were combined with the program LogCombiner (v2.4.8) available in the BEAST package. A final maximum clade credibility tree, the tree in the posterior sample with the maximum sum of posterior clade probabilities, was determined for each dataset using TreeAnnotator (v2.4.7).

4.2.2 Ancestral reconstruction using PAML

The codeml programme from the PAML software package was used to reconstruct common ancestral sequences for each of the BEAST summary trees. Specifically, ancestral reconstruction was performed on PIDs 036, 079, 267 & 271 by codon maximum likelihood (codonml). A similar codeml control file setup was used as described in 3.3.3, except that in this instance, the BEAST-generated trees were used as input trees. Reconstructed ancestral sequences together with participant sequences were then aligned and inspected using BioEdit.

4.3 Results

4.3.1 Analysis of BEAST results in Tracer

Output log files from BEAST were viewed in the Tracer (v1.7) package. Each file produced statistics for parameters specified in the XML file created in BEAUti. As the output of these files was quite extensive, a brief explanation of some of the output is described using alignment 271. After the brief presentation, a summary table containing the main variables of interest (e.g. tMRCA) is presented and explained.

4.3.1.1 Tracelog for Alignment 271

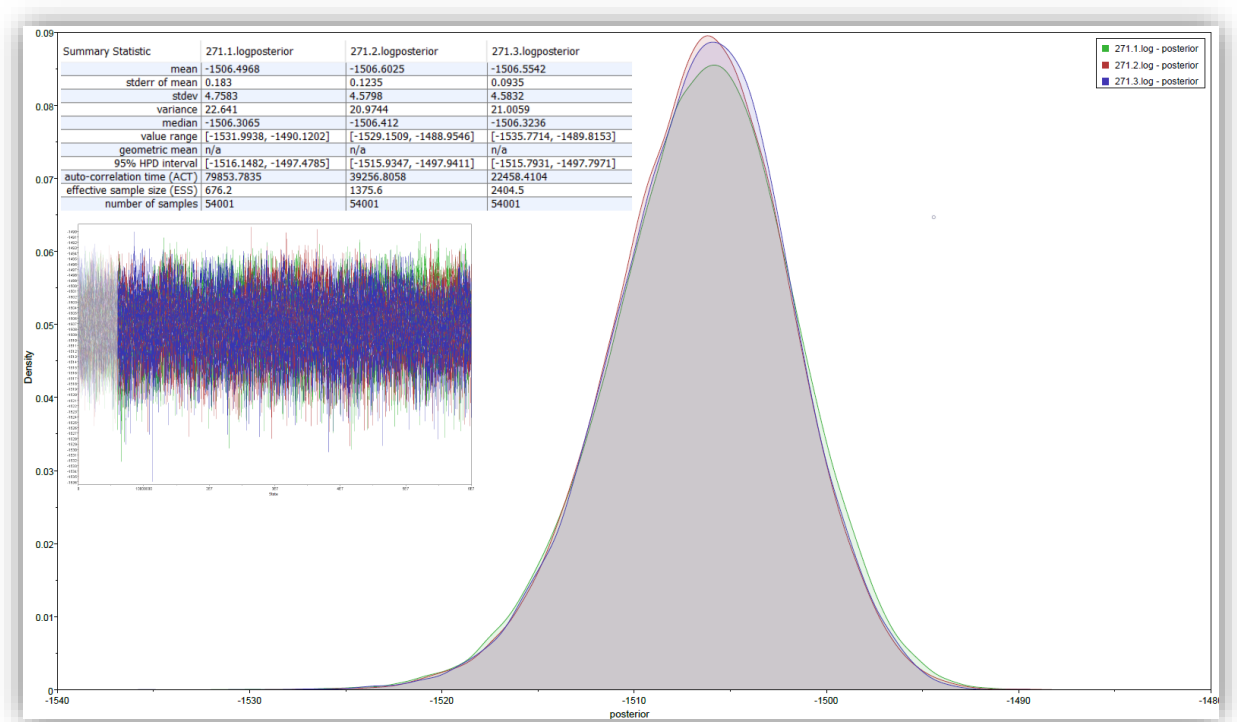


Figure 4.1: Comparison of Marginal Densities of three runs of alignment 271 with associated summary estimates and trace plots for the posterior statistic

The above figure is a compression of three main data analysis viewing pains in Tracer that accompany each parameter output generated by a BEAST MCMC run. The main graphic in the centre contains the Kernel Density Estimate (KDE) plots for the three independent MCMC runs of 60 000 000 each. The graphic that looks like a “hairy caterpillar” on the left-hand side of the main graph represents the raw trace logs for the three runs. The table on the top left-hand corner are the estimates for the parameter of interest (i.e. posterior statistic)

The first statistic for PID 271 alignments were the estimates for the posterior. As seen in the table insert in Figure 4.1, the effective sample sizes (ESS) for each BEAST run was above 200 for the posterior. The ESS value indicates the number of effectively independent draws

from the posterior in the sample and should ideally be > 200 (Drummond and Rambaut, 2009; Heath, 2015). This suggests that the MCMC runs produced adequate posterior estimates of divergence times and substitution model parameters for PID 271 sequence alignments for the posterior distribution (Heath, 2015). The raw trace plot for the posterior statistic corroborates this (see graphic immediately under the posterior estimates table). The raw trace is principally a diagnostic tool for inspecting convergence to the posterior, assessing the length of the burn-in and whether or not the chain is mixing well (Heath, 2015; Rambaut *et al.*, 2017; Rambaut *et al.*, 2018). The trace plot in the figure looks something like a hairy caterpillar (i.e. there are no obvious trends in the plot to suggest that the MCMC was still converging and there are also no large-scale fluctuations in the trace to suggest poor mixing) indicating that there was sufficient mixing in the MCMC runs. In addition, the density plots (i.e. graphic to the left of the raw trace plot) show that the marginal densities of the posterior parameter from the three independent runs [271.1.log (green), 271.2.log (red) and 271.3.log (blue)] are almost identical. This indicates that the three independent runs converged on the same stationary distribution. The highest posterior density region (HPD), for example the 271.1. logposterior, indicates that there is a 0.95 probability that the true value for the posterior lies in the range -1516.1482 to -1497.4785.

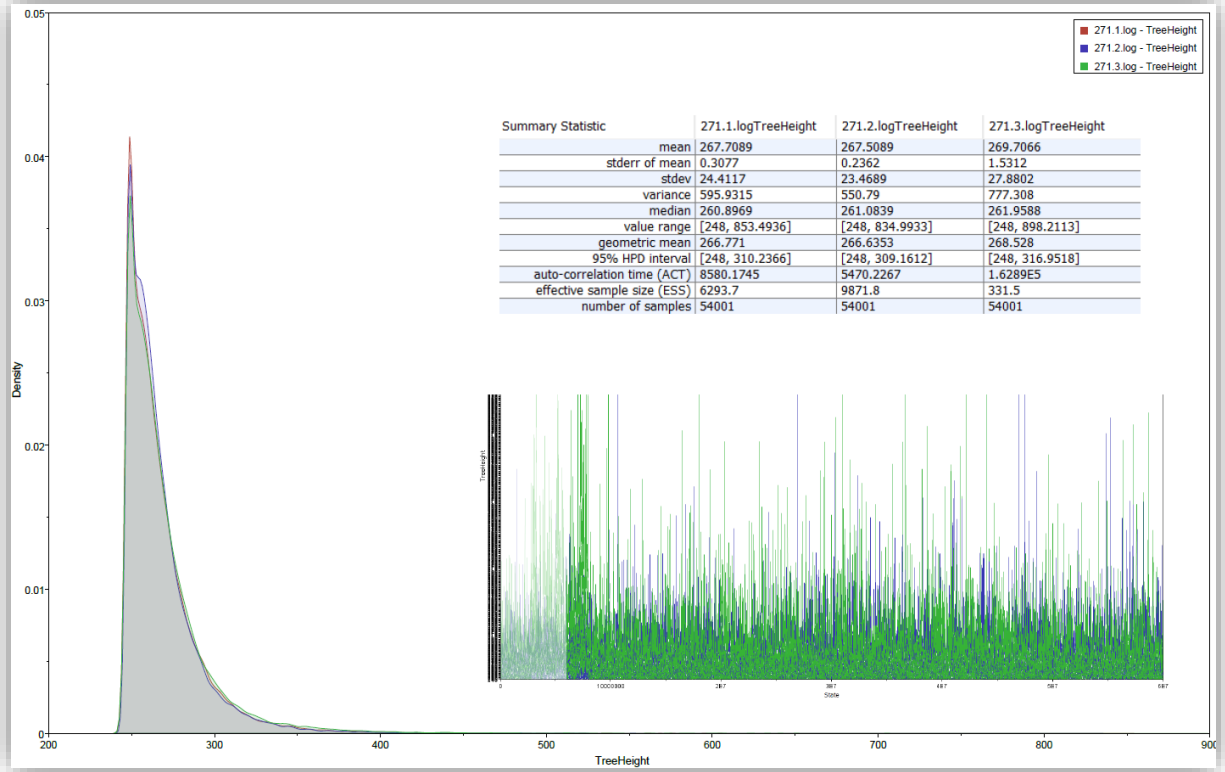


Figure 4.2: Comparison of Marginal Densities of three runs of alignment 271 with associated summary estimates and trace plots for the treeheight statistic

As seen from the tracelogs, marginal densities and the ESS, there was sufficient mixing in the MCMC runs for the treeheight (tMRCA) statistic given the data. The HPD interval indicates a 95% likelihood that the true tMRCA was somewhere in the range of 248 to 310 days (ref. to 271.1.logTreeHeight). The average tMRCA was 268 ± 24 days.

Each of the other alignments (i.e. PIDs 036, 079 and 267) showed strong evidence of sufficient mixing of their MCMC runs. This means that for runs in each group of alignments, the ESSs were greater than 200, marginal densities converged at the same stationary distribution and tracelogs resembled a “hairy caterpillar” (see appendices 3-8). The logs for each run in an alignment were combined into single logs per PID and viewed again in Tracer. Table 4.1 tabulates the summary estimates of the combined logs for each of the four alignments (i.e. PIDs 036, 079, 267 and 271).

Table 4.1: Summary statistics of interest from BEAST output

Statistic	Sequence Alignments							
	036		079		267		271	
	Mean	ESS	Mean	ESS	Mean	ESS	Mean	ESS
posterior	-1 444,13	4 706,00	-1 488,45	1 544,00	-1 433,32	5 145,00	-1 506,50	676,00
treeLikelihood	-1 408,55	8 609,00	-1 461,41	342,00	-1 406,53	6 632,00	-1 457,53	2 234,00
TreeHeight	366,37	45 435,00	345,21	33 959,00	346,29	33 274,00	267,71	6 294,00
Origin	455,40	140 730,00	459,03	36 425,00	459,64	42 545,00	332,24	19 492,00
uclDStdev	0,22	754,00	0,21	26 903,00	0,23	418,00	3,91	1 402,00

Table 4.1 shows that the ESSs for posterior, treelikelihood, treeheight, origin and uclDStdev are all above 200 for the four alignments (Drummond and Rambaut, 2009; Heath, 2015). The uclDStdev gives an indication of how clock-like the data is, with estimates closer to zero indicating that the data is quite clock-like. The evolutionary rates of the sequences of alignments 036, 079 and 267 were approximately uniform over their respective phylogenies (Drummond *et al.*, 2007; Drummond and Rambaut, 2007; Vandamme, 2009). Sequences in alignment 271, however, displayed substantial rate heterogeneity among lineages (i.e. uclDStdev value was much greater than 1.0) (Drummond *et al.*, 2007).

The treeheight statistic represents the tMRCA for the taxa subsets. This statistic denotes the estimated divergence time of the node representing the MRCA of the given taxa (i.e. it gives the marginal posterior distribution of the age of the root of the entire tree) (Drummond *et al.*, 2007). The table thus shows that the estimated tMRCA for 036 was 367 days, 079 was 346 days, 267 was 347 days and 271 was 268 days.

4.3.2 Trees reflecting tMRCA and evolutionary rates from reconstructed ancestral sequences

Aligned sequences and ancestors for each of the four UDP alignments are shown in the next four figures. Deviations from common ancestry were depicted on the time-scaled ML summary trees obtained from BEAST. This suggests that the length of the tree branches are in relation to time from a common ancestor and not based on evolutionary rates. The branch lengths in the inserts, however, are reflective of the evolutionary rate measured in nucleotide substitutions per codon.

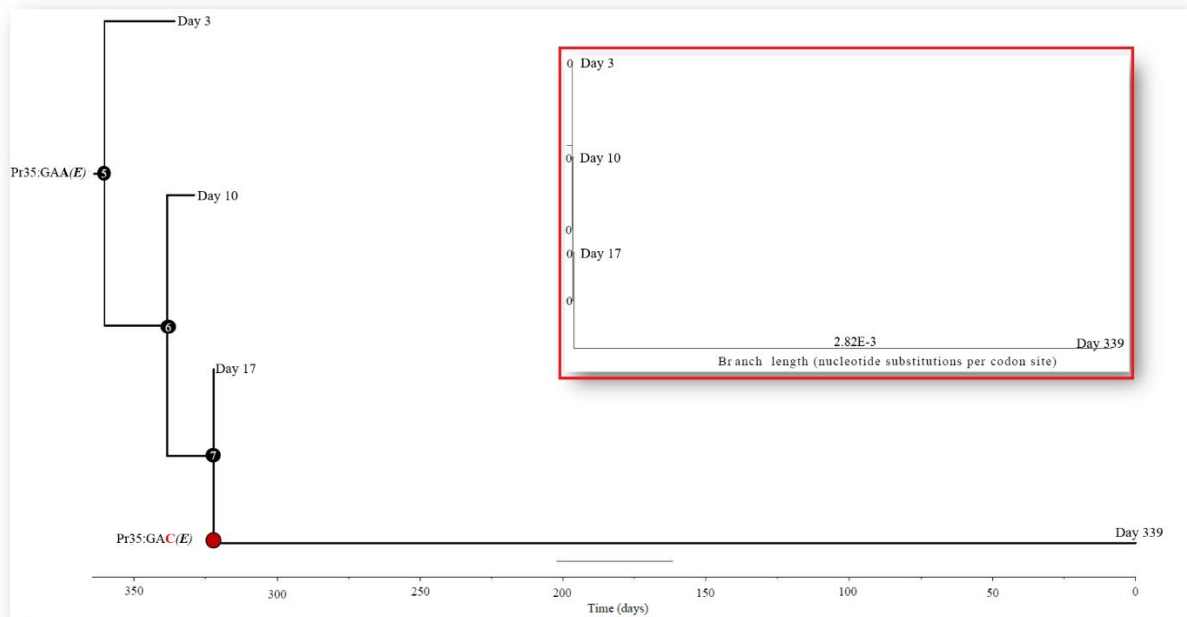


Figure 4.3: Time to the most recent common ancestor (tMRCA) for 036

The figure indicates the evolution of HIV in PID 036 over four timepoints (TPs). The scale is reverse ordered to illustrate time to the most recent common ancestor of the sequence alignments. The figure also gives a topological view of the relatedness of sequences. Nodes of the same colour indicate identical nucleotide sequences. Diversification events are indicated by a change of colour along the sequence (e.g. ●) with corresponding nucleotide change(s) indicated at that point.

Figure 4.3 shows that the estimated median tMRCA for the 036 phylogeny was approximately 360 days. Ancestral nodes 5 through to 7 are circled in black to indicate that they have identical sequences. After node 7, a diversification event occurs in the Day 339 sequence at Pr35, where a mutation at the third codon position in Pr35 (GAA to GAC) resulted in an AA change from Glu (E) to Asp (D). To indicate that divergence from a common ancestor has occurred, a red circle is used instead of a black circle. It should be noted that branch lengths are an indication of passage through time and not evolutionary rates. Evolutionary rates are, however, captured in the branch lengths of the inset (framed in red) in nucleotide substitutions per codon (as per codeml). The sequence alignments at Days 3, 10 and 17, as well as the ancestral sequences (nodes 5, 6 and 7) did not evolve over the two week period in which the samples were collected. The evolutionary rate from the ancestral node 5 to the most recent alignment (day 339) was $2.82e^{-3}$ nucleotide substitutions per codon in codeml (see inset Figure 4.3).

Physical examination of the sequence alignments against known major and minor DRMs from the Stanford University HIVdb (2015; 2017) did not detect any inherited or derived

DRMs. To verify, sequence alignments were blasted against the SDRMs on the Stanford University HIVdb and no surveillance DRMs detected.

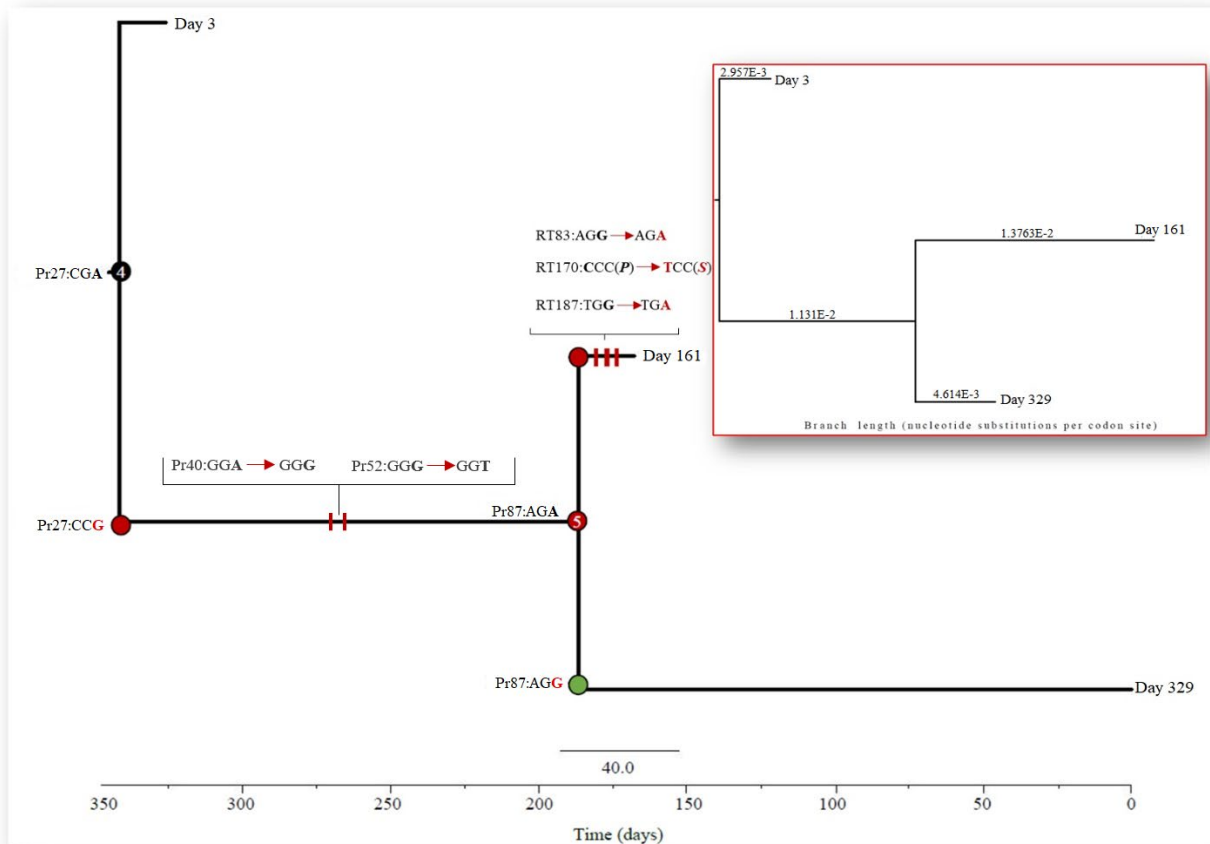


Figure 4.4: Time to the most recent common ancestor (tMRCA) for 079

BEAST estimated the tMRCA for all the 079 alignments to be 342 days (see Figure 4.4). The closest related taxon to the common ancestor of the phylogeny was at Day 3. A diversification event occurred at Pr27 with an A to G transition mutation (CGA to CGG) causing Day 161 and 329 sequences to break away from Day 3 and the common ancestor (node 4). Two additional mutations occurred along the Day 161 and Day 329 before the diversification event at the node 5 common ancestor. The first of these was a transition mutation at Pr40 (GGA to GGG) and the second a transversion mutation at Pr52 (GGG to GGT).

A third transition mutation at Pr87 (AGA to AGG) resulted in the Day 329 sequence deviated from its shared ancestry with the Day 161 sequence at node 5. The sequence at Day 161 derived three additional transition mutations in RT. These were at RT83 (AGG to AGA),

RT170 (CCC to TCC) and RT187 (TCG to TGA). The mutation at RT170 occurred at the 1st nucleotide position of the codon and resulted in an AA change from a Pro (P) to a Ser (S).

The evolutionary rate from the origin (node 4) to the common ancestor of Day 161 and Day 329 (node 5) was estimated to be $1.131e^{-2}$ nucleotide substitutions per codon. The evolutionary rate along Day 267 from the ancestor at node 5 was $1.376e^{-2}$ substitutions per codon and from node 5 to Day 329, $4.614e^{-3}$ nucleotide substitutions per codon. This indicates that the sequence that evolved the most from ancestry was at Day 161. However, the derived mutations at this TP were lost at Day 329 (168 days later).

Physical examination of the sequence alignments against surveillance DRMs from the Stanford University HIVdb (2015; 2017) detected one DRM in RT position 103 where a Lys (K) mutated to an Asn (N). This mutation was also detected when sequence alignments were blasted against the SDRMs on the Stanford HIV Drug Resistance Database (2017). The K103N mutation is a known NNRTI antagonist (Chen *et al.*, 2004; Halvas *et al.*, 2010; Chen *et al.*, 2012; Chen *et al.*, 2014). Tracing the mutation along the phylogenetic tree revealed that it was present in the original ancestor and remained conserved throughout the evolution of the 079 viral strain (i.e. was present at sequences at each timepoint). As such, this mutation was an inherited one and not spontaneously derived.

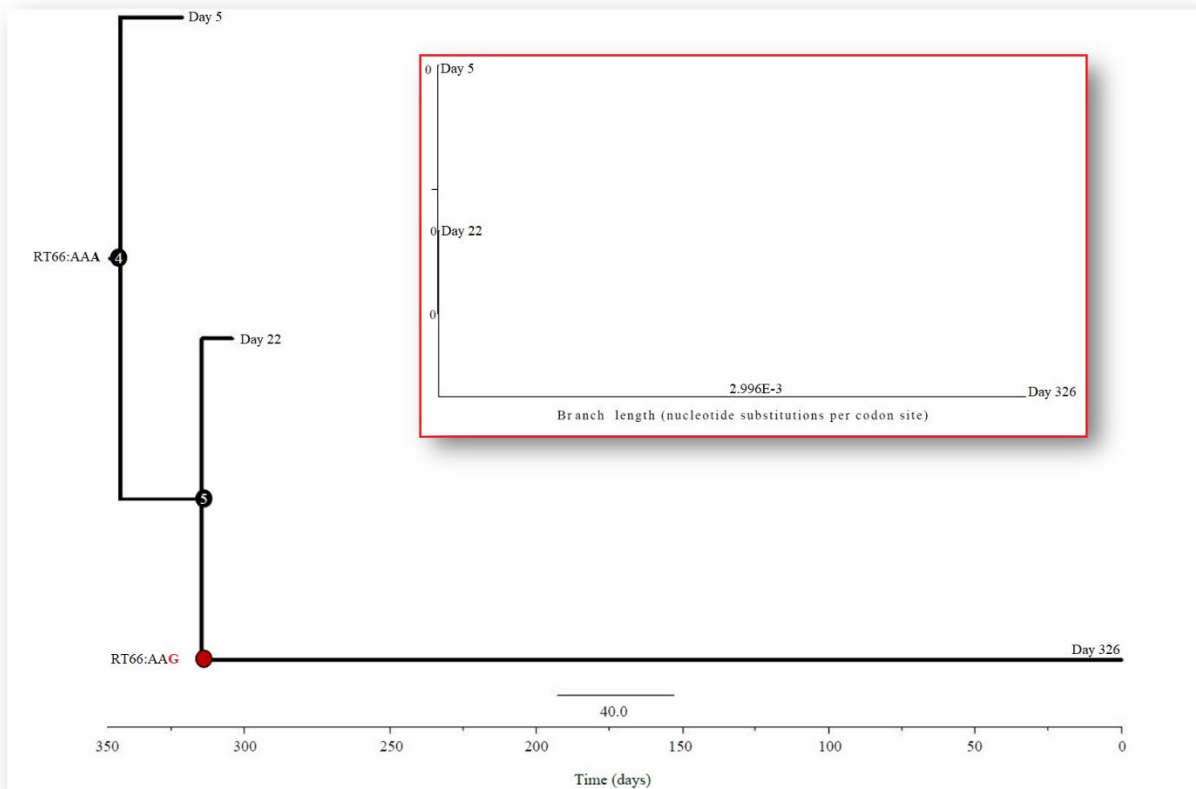


Figure 4.5: Time to the most recent common ancestor (tMRCA) for 267

The estimated tMRCA for 267's phylogeny was 342 days. Sequences at Day 5 and Day 22 were identical to the common ancestor at node 4, despite the 17 days that elapsed between them (see Figure 4.5). The only diversification event occurred at RT66 with an A to G transition mutation (AAA to AAG) causing the Day 326 sequence to break away from the Day 22 sequence at the common ancestor (node 5). The evolutionary rate along Day 326 from the ancestor at node 5 was $2.996e^{-3}$ nucleotide substitutions per codon. Physical examination and blasting of the sequence alignments against SDRMs from the Stanford University HIVdb (2015; 2017) did not detect any DRMs.

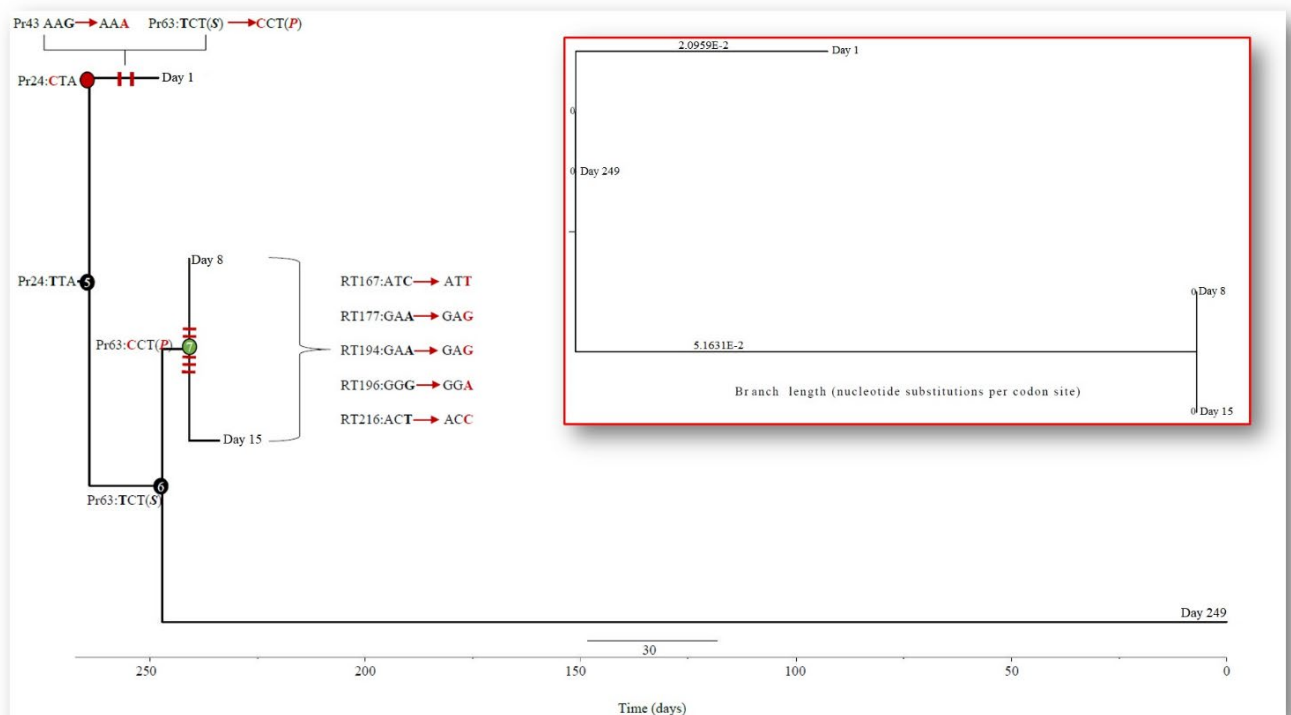


Figure 4.6: Time to the most recent common ancestor (tMRCA) for 271

The estimated tMRCA for 271 was 264 days. The closest sequence to the original ancestor (node 5) was Day 249. This TP shared identical sequences to the common ancestor at node 5. The first timepoint (Day 1) deviated from the original ancestor very early in the evolution of 271 with a transition mutation at Pr24 (TCT to CCT). Although this mutation occurred at the 1st nucleotide position of the codon it did not alter the *wt* AA. Two additional transition mutations were derived at Day 1. The first was at Pr43 (AAG to AAA) and the second at Pr63 (TCT to CCT). The latter mutation altered the AA from a Ser (S) to a Pro (P).

Day 8 and Day 15 sequences shared a common ancestor at node 7. This clade shared common ancestry with Day 249 at node 6. Sequences at Day 8 and Day 15 deviated from ancestral node 6 at Pr63 (TCT to CCT). These sequences were identical and derived five additional transition mutations in RT along their co-evolution. These were at RT167 (ATC to ATT), RT177 (GAA to GAG), RT194 (GAA GAG), RT196 (GGG to GGA) and RT216 (ACT to ACC). The Pr63 mutation observed at Day 1 was still present up to Day 15. The sequence at the last timepoint (Day 249) did not evolve relative to the common ancestor.

The evolutionary rate along Day 1 from the ancestral node 5 was $2.096e^{-2}$ nucleotide substitutions per codon and the evolutionary rate from node 5 through to Day 8 and Day 15 was $5.163 e^{-2}$ nucleotide substitutions per codon. Physical examination and blasting of the sequence alignments against SDRMs from the Stanford HIV Drug Resistance Database did not detect any DRMs.

4.4 Discussion

The aim of this section was to “reverse” trace any observed DRMs to the MRCA to investigate whether these DRMs were present in the MRCA or whether these DRMs were spontaneously acquired. In the event that DRMs were inherited from the MRCA, to observe the structural changes occurring at those sites as the DRMs were being lost over time. The samples used in this study were collected less than a week after participants were diagnosed. This created a unique window through which to observe how inherited DRMs were lost as minority variants competed for replicative dominance in newly infected treatment naïve patients. To achieve this aim, ancestral sequences were reconstructed from “extant” viral sequences, mapped onto time-scaled trees and evaluated.

Ancestral reconstruction was performed for each of the four UDP alignments that had more than two timepoints. The *codonml* programme in the PAML package was used to reconstruct ancestral sequences. Time-scaled trees for mapping divergence rates were constructed using BEAST. Participant 036’s sequences indicated that the *wt* strain remained dominant for the first 17 days post diagnosis. The only diversification event was detected 322 days later. This suggests that at some point between Day 17 and Day 339 a variant strain with a mutation at the third codon position in Pr35 became more dominant. This mutation altered the AA at this position from a Glu to Asp. The estimated evolutionary rate from the 036 phylogeny’s common ancestor to the most recent sequence (Day 339) was $2.82e^{-3}$ nucleotide substitutions per codon.

The sequences in alignment 079 showed a bit of variability. A diversification event occurred at some time along the *pol* gene in the 158 days between the first blood draw and the second. At Day 161 four mutations were observed, one of which altered the “resident” AA. These were at Pr27 (CGA to CGG), RT83 (AGG to AGA), RT170 (CCC to TCC) and RT187 (TCG to TGA), with the mutation in RT170 altering the AA from a Pro to a Ser. Three of

the mutations detected at Day 161 in RT positions 83, 170 and 187 were undetectable 168 days later at Day 329. The strain detected at this TP deviated from Day 3 and the common ancestor at mutations in Pr27 and Pr87. The evolutionary rates from common ancestors indicate that Day 161 had evolved the most at a rate of $1.376e^{-2}$, approximately 158 days from the 1st timepoint.

Similar to 036, the first two timepoints of 267 had identical sequences within the first 22 days of HIV⁺ diagnosis. A diversification event was detected at the last timepoint approximately 304 days after the second TP with an A to G transition mutation (AAA to AAG) at RT66. The evolutionary rate along TP3 from the ancestor at node 5 was $2.996e^{-3}$ nucleotide substitutions per codon.

Of the four alignments, the sequences in 271 displayed substantial rate heterogeneity among lineages as per BEAST (Drummond *et al.*, 2007). This could possibly be a consequence of the tMRCA being shorter between 1st and last timepoints. It may also indicate that minority variants were still competing to establish dominance. The other alignments were more clock-like, indicating that a global molecular clock could describe their rates of evolution over their respective phylogenies (Drummond *et al.*, 2007; Drummond and Rambaut, 2007; Vandamme, 2009).

Sequences in the 271 alignment were quite variable within the first 15 days post diagnosis (i.e. Day 1, Day 8 and Day 15). One day after diagnosis, two mutations were detected at Pr43 and Pr63. The mutation at Pr63 altered the AA from a Ser to a Pro. The Pr63 mutation remained conserved at the second and third timepoints, which were seven days apart. The Pr43 mutation, however, was undetectable from the second timepoint seven days after Day 1. The sequences at Day 8 and Day 15 were identical and acquired five more transition mutations in RT sites 167, 177, 194, 196 and 216. The sequence at the last timepoint (Day 249) shared identical sequences with the common ancestor. This suggests that this variant either remained in the background while the variants at Day 1 and Day 8 vied for dominance in the first few days of infection or, it could suggest that during the first few days of infection initial host immune responses were heightened, thereby forcing the virus to mutate to survive under this pressure. With the increase in viral load and diminished immune attack, the virus no longer required the earlier acquired mutations and reverted back to its original *wt* strain.

Thus, after approximately 249 days post-HIV⁺ diagnoses, the *wt* strain re-emerged with the mutations detected in the first 15 days being undetectable.

The Pr63 mutation was conserved and still present up to 2 weeks post diagnosis (Day 8 and Day 15). This could indicate that this site may play a significant role in conferring fitness advantage against the initial host immune response typical of this early stage of viral infection (Ndhlovu *et al.*, 2015). During the week between Day 8 and Day 15, five additional mutations arose in RT. These mutations were subsequently lost 234 days later at Day 249. This could suggest that despite the observed mutations, none of them were fixed over the 248 days between the 1st TP and the last TP as the dominant species eventually reverted back to *wt*. It may also suggest that due to the bottleneck effect, mutations earlier on in the evolution of the 271 quasispecies were dominant, but with the passage of time, these mutations did not confer enough competitive advantage to get fixed in the population. Consequently, reversion to the ancestral or *wt* sequence occurred. In addition, the rapid CD8⁺ T cell apoptosis and the inability of these cells to leave phenotypic markers for CD8⁺ T memory cells may have provided a window for viral replication in the absence of intense immune attack (Ndhlovu *et al.*, 2015). As such, some of the advantageous mutations were either no longer critical to viral fitness, or the other quasispecies (such as *wt* in this instance) were able to establish dominance in the absence of host immune response pressure.

Physical examination of the sequence alignments against surveillance DRMs from the Stanford University HIV Drug Resistance Database (2015) for all four alignments only detected a DRM in 079. This was mutation K103N in RT. This mutation was also the only one detected when sequences were blasted against the SDRMs on the Stanford HIV Drug Resistance Database for all four alignments. The K103N mutation is known to be an NNRTI antagonist. Tracing the mutation along the phylogenetic tree revealed that it was present in the original ancestor and remained conserved throughout the evolution of the 079 viral strain (i.e. was present at sequences at each timepoint). As such, this mutation was an inherited one and not spontaneously derived. Consequently, it was not possible to observe DRM loss, nor was it possible to investigate the structural changes that accompanied the loss.

Since no structural observations could be made due to the absence of DRMs, each alignment's sequences were also tested for inherited/acquired mutations in functional sites using Prosite and Genedoc. The findings indicated that functional domains remained

conserved from the MRCA to the last TP in each sequence per alignment (results not shown). This suggests that each sequence inherited and maintained functional sites from their common ancestor and did not spontaneously acquire any mutations that altered their functional states in both Pr and RT.

CHAPTER 5:

General Discussion

Both the HyPhy package in Datamonkey and the codeml package in PAML, despite performing analyses on essentially duplicate sequences, were inconsistent in detecting sites for positive selection common to both sets of alignments. Apart from sites that overlapped across methods, it was difficult to place confidence in any one method as the better method for detecting sites that were “truly” under positive evolutionary pressure (at $p < 0.05$). A more reliable approach perhaps would be to use several different methods for the analysis and create a pool of positively selected sites, with sites that appeared more frequently considered as having stronger support for Darwinian evolution.

Nonetheless, in Datamonkey, the REL method appeared to be the most appropriate for the small datasets used in this study. This is in keeping with Poon *et al.*'s (2009) recommendation of this method for alignments containing 5 -15 sequences. The F81 model of substitution was on average the more reliable model for detecting positive sites with a fair amount of consistency (66.7%). The REV model of substitution, although not that sensitive to detecting sites that were under positive selective pressure, was the more consistent model in that it detected 80% of sites common to both sets of alignments.

Overall, the sequences were under strong purifying selection with less than 1% of the sites in each alignment being under positive selection as per PAML. This is in keeping with a study by Gordon *et al.* (2003) on a treatment naïve cohort in KwaZulu-Natal. Altogether 16 sites for positive selection were identified by the HyPhy and PAML packages. Sixty-nine percent of these sites were also detected by Gordon *et al.* (2003) and 50% by Chen *et al.* in their studies. It is highly probable that the viruses in these participants were forced to evolve to evade host HIV-specific CTL responses shown to occur at Fiebig stage 1 in this cohort by Ndhlovu *et al.* (2015). In addition, it is also possible that early pressure to evolve could be due to other host factors. These include the types of cells infected, microenvironments in the mucosa and lymph nodes, which present a wide spectrum of innate host defenses, including IFNs and/or other molecules (Shugars and Wahl, 1998; Shugars *et al.*, 1999). Host-selective

pressures, therefore, may either increase or decrease the diversity of the virus population (Pilcher *et al.*, 2004).

In Pr, L63HPSVT was the most variable site under positive selection. However, this site was not found in any HLA epitope nor did it fall within any Pr functional domain. [Although no longer on the Stanford University's HIV Database of SDRMs, the L63P mutation was once considered a DRM (Chen *et al.*, 2004)]. The only positively selected site in Pr that was a DRM was V82I. This site had a mutation in the epitope of HLA-B*810. This HLA has been associated with low viremia and suppressing viral replication (Kaslow *et al.*, 1996; Altfeld *et al.*, 2003; Kiepiela *et al.*, 2006; Altfeld and Goulder, 2011; Illing *et al.*, 2018). This suggests that mutations in this epitope may be advantageous to viral replication and proliferation.

The most volatile sites for positive selection in RT were at E39DKT, G123SND and E207ATK. The E39T and G123S mutations enabled sequences that possessed them to gain a putative CKII-P functional site in an unconserved region in RT. The G123 site was also in the B*5702, B*5703 and B*3501 HLA epitopes. Mutations in B*5703 epitopes, however, did not benefit the virus because HLA-B*5703 was inherently error-prone in proofreading its isotopes (Illing *et al.*, 2018), thus negating mutations. Nonetheless, HLA-B*5703 and HLA-B*5702 have been associated with suppressing viral replication (Kloverpris *et al.*, 2012). Mutations in B*5702 may thus be advantageous to viral replication. Mutations in the HLA-B*3501 epitope may, however, present a fitness cost to the virus as this HLA has been found to serve no real function apart from naturally prolonging the duration of HIV (Huang *et al.*, 2009).

Ancestral reconstruction of sequences that had MTPs detected a DRM in only PID 079's alignments. This participant's sequences had the RT K103N mutation, which was traced back to its MRCA. This mutation remained fixed over time and was detected at each subsequent TP. Therefore, it was not possible to observe functional changes between TPs where DRMs were either lost or gained.

Apart from an AA mutation in one sequence each of PID 079 (RT170) and PID 271 (Pr63), there was no intrahost sequence diversity in any of the MTP alignments at the AA level. In addition, tracing changes in functional domains within each alignment found that functional

domains were conserved from the MRCA up to the last TP. This could suggest that the viruses underwent a tremendous amount of host factor pressure during the eclipse phase of acute infection. Although the sequences used in this analysis were collected less than a week post infection, early CTL responses were mounted at Fiebig Stage 1 (Nghlovu *et al.*, 2015). This could imply that by the earliest point of viral detection, the virus had already undergone minute general non-specific host immune attack, sufficient for the virus to adapt and establish a dominant or *wt* strain. This *wt* strain then remained conserved in the absence of drug pressure.

CHAPTER 6:

Conclusion

It emerges from this study that pooling results from different techniques is important in identifying sites under positive selective pressure. This adds a holistic element to identifying sites because the underlying algorithms of the different packages at times omitted sites that were detected by other algorithms.

The putative functional domains within this cohort remained relatively conserved. The gains and losses of functional domains, however, may suggest that the selective pressures acting on viruses were host-specific. Losses of conserved putative functional sites, however, seemed counterintuitive at times because these sites conferred viral resilience and fitness. Similarly, mutations in some identified epitopes did not confer meaningful viral advantage, mostly because the HLAs that were specific to them either naturally prolonged HIV progression or were inherently error-prone.

Observing the mutations of intrahost viral sequences over time found that viral sequences remained conserved from the common ancestor up to the last TP in each participant. Intrahost sequences also conserved all their putative functional sites inherited from their recent common ancestors. This could suggest that, in the absence of drug pressure, the viral quasispecies that survived early host immune attack during the eclipse phase of acute infection, became the dominant (or *wt*) strain that was preserved throughout. This may be the case because the earliest studied immune response in this cohort was at Fiebig Stage I.

A major limitation of this study was the relatively small sample size used for the analyses. However, the FRESH acute cohort is unique since HIV negative participants were enrolled and screened bi-weekly until HIV-1 infection was detected. HIV infection was thus detected as early as Fiebig Stage 1. As such, the sequences used in this study are truly reflective of acute infection. In addition, it is generally very difficult to obtain sequences so early in HIV infection.

Bibliography

- Abram, M. E., Ferris, A. L., Shao, W., Alvord, W. G. & Hughes, S. H. 2010. Nature, position, and frequency of mutations made in a single cycle of HIV-1 replication. *J Virol*, 84, 9864-78.
- Acevedo-Sáenz, L., Ochoa, R., Rugeles, M. T., Olaya-García, P., Velilla-Hernández, P. A. & Diaz, F. J. 2015. Selection pressure in cd8+ t-cell epitopes in the pol gene of HIV-1 infected individuals in colombia. A bioinformatic approach. *Viruses*, 7, 1313-1331.
- Afonso, J. M., Bello, G., Guimaraes, M. L., Sojka, M. & Morgado, M. G. 2012. HIV-1 genetic diversity and transmitted drug resistance mutations among patients from the north, central and south regions of angola. *PLoS One*, 7, e42996.
- AIDS info. 2015. *Management of the treatment experienced patient - virologic failure* [Online]. Available: <https://aidsinfo.nih.gov/guidelines/html/1/adult-and-adolescent-arv-guidelines/15/virologic-failure> [Accessed 19 April 2017].
- Al-Jabri, A. A. 2007. Mechanisms of host resistance against HIV infection and progression to aids. *Sultan Qaboos University medical journal*, 7, 82-96.
- Alencar, C. S., Sabino, E. C., Carvalho, S. M., Leao, S. C., Carneiro-Proietti, A. B., Capuani, L., Oliveira, C. L., Carrick, D., Birch, R. J., Goncalvez, T. T., Keating, S., Swanson, P. A., Hackett, J., Jr. & Busch, M. P. 2013. HIV genotypes and primary drug resistance among HIV-seropositive blood donors in brazil: Role of infected blood donors as sentinel populations for molecular surveillance of HIV. *J Acquir Immune Defic Syndr*, 63, 387-92.
- Altfeld, M., Addo, M. M., Rosenberg, E. S., Hecht, F. M., Lee, P. K., Vogel, M., Xu, G. Y., Draenert, R., Johnston, M. N. & Strick, D. 2003. Influence of hla-b57 on clinical presentation and viral control during acute HIV-1 infection. *Aids*, 17, 2581-2591.
- Altfeld, M. & Goulder, P. J. 2011. *The step study provides a hint that vaccine induction of the right cd8+ t cell responses can facilitate immune control of HIV*. Oxford University Press.
- Amiel, C., Charpentier, C., Desire, N., Bonnard, P., Lebrette, M. G., Weiss, L., Pialoux, G. & Schneider, V. 2011. Long-term follow-up of 11 protease inhibitor (pi)-naïve and pi-treated HIV-infected patients harbouring virus with insertions in the HIV-1 protease gene. *HIV medicine*, 12, 138-144.
- Ardito, F., Giuliani, M., Perrone, D., Troiano, G. & Muzio, L. L. 2017. The crucial role of protein phosphorylation in cell signaling and its use as targeted therapy (review). *Int J Mol Med*, 40, pp271-280.
- Arts, E. J. & Hazuda, D. J. 2012. HIV-1 antiretroviral drug therapy. *Cold Spring Harbor perspectives in medicine*, 2, a007161.
- Averting HIV and AIDS (AVERT). 2016. *HIV and aids in South Africa* [Online]. Available: www.avert.org/professionals/hiv-around-world/sub-saharan-africa/south-africa#footnote1_kjns9eh [Accessed 14 March 2017].

- Avila-Rios, S., Mejia-Villatoro, C. R., Garcia-Morales, C., Soto-Nava, M., Escobar, I., Mendizabal, R., Giron, A., Garcia, L. & Reyes-Teran, G. 2011. Prevalence and patterns of HIV transmitted drug resistance in Guatemala. *Rev Panam Salud Publica*, 30, 641-8.
- Baldauf, S. L. 2003. Phylogeny for the faint of heart: A tutorial. *Trends Genet*, 19, 345-51.
- Banke, S., Lillemark, M. R., Gerstoft, J., Obel, N. & Jorgensen, L. B. 2009. Positive selection pressure introduces secondary mutations at gag cleavage sites in human immunodeficiency virus type 1 harboring major protease resistance mutations. *J Virol*, 83, 8916-24.
- Bassett, I. V., Chetty, S., Giddy, J., Reddy, S., Bishop, K., Lu, Z., Losina, E., Freedberg, K. A. & Walensky, R. P. 2011. Screening for acute HIV infection in South Africa: Finding acute and chronic disease. *HIV medicine*, 12, 46-53.
- Bennett, D. E., Camacho, R. J., Otelea, D., Kuritzkes, D. R., Fleury, H., Kiuchi, M., Heneine, W., Kantor, R., Jordan, M. R., Schapiro, J. M., Vandamme, A.-M., Sandstrom, P., Boucher, C. A. B., van de Vijver, D., Rhee, S.-Y., Liu, T. F., Pillay, D. & Shafer, R. W. 2009. Drug resistance mutations for surveillance of transmitted HIV-1 drug-resistance: 2009 update. *PLOS ONE*, 4, e4724.
- Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. 2013. Genbank. *Nucleic acids research*, 41, D36-D42.
- Bielawski, J. P. & Yang, Z. 2003. Maximum likelihood methods for detecting adaptive evolution after gene duplication. *Genome evolution*. Springer.
- Blanco-Heredia, J., Lecanda, A., Valenzuela-Ponce, H., Brander, C., Ávila-Ríos, S. & Reyes-Terán, G. 2016. Identification of immunogenic cytotoxic t lymphocyte epitopes containing drug resistance mutations in antiretroviral treatment-naïve HIV-infected individuals. *PloS one*, 11, e0147571-e0147571.
- Bos, D. H. & Posada, D. 2005. Using models of nucleotide evolution to build phylogenetic trees. *Dev Comp Immunol*, 29, 211-27.
- Bouckaert, R. R., Heled, J., Kuehnert, D., Vaughan, T. G., Wu, C. H., Xie, D., Suchard, M. A., Rambaut, A. & Drummond, A. J. 2014. BEAST 2: A software platform for Bayesian evolutionary analysis. *PLoS Computational Biology*, 10(4).
- Brenchley, J. M., Paiardini, M., Knox, K. S., Asher, A. I., Cervasi, B., Asher, T. E., Scheinberg, P., Price, D. A., Hage, C. A., Kholi, L. M., Khoruts, A., Frank, I., Else, J., Schacker, T., Silvestri, G. & Douek, D. C. 2008. Differential th17 cd4 t-cell depletion in pathogenic and nonpathogenic lentiviral infections. *Blood*, 112, pp2826-2835.
- Brenchley, J. M., Schacker, T. W., Ruff, L. E., Price, D. A., Taylor, J. H., Beilman, G. J., Nguyen, P. L., Khoruts, A., Larson, M., Haase, A. T. & Douek, D. C. 2004. Cd4+ t cell depletion during all stages of HIV disease occurs predominantly in the gastrointestinal tract. *J Exp Med*, 200, pp749-759.
- Center for Disease Control 1981. Kaposi's sarcoma and pneumocystis pneumonia among homosexual men - New York City and California. *Morbidity & Mortality Weekly Report*, 30, pp305-308.

- Charpentier, C., Dwyer, D. E., Mammano, F., Lecossier, D., Clavel, F. & Hance, A. J. 2004. Role of minority populations of human immunodeficiency virus type 1 in the evolution of viral resistance to protease inhibitors. *J Virol*, 78, 4234-47.
- Chen, L., Perlina, A. & Lee, C. J. 2004. Positive selection detection in 40,000 human immunodeficiency virus (HIV) type 1 sequences automatically identifies drug resistance and positive fitness mutations in HIV protease and reverse transcriptase. *J Virol*, 78, 3722-32.
- Chen, M., Ma, Y., Duan, S., Xing, H., Yao, S., Su, Y., Luo, H., Yang, L., Chen, H., Fu, L., Qu, A., Ou, C. Y., Jia, M. & Lu, L. 2012. Genetic diversity and drug resistance among newly diagnosed and antiretroviral treatment-naive HIV-infected individuals in western yunnan: A hot area of viral recombination in china. *BMC Infect Dis*, 12, 382.
- Chen, Y., Chen, S., Kang, J., Fang, H., Dao, H., Guo, W., Lai, C., Lai, M., Fan, J., Fu, L., Andrieu, J. M. & Lu, W. 2014. Evolving molecular epidemiological profile of human immunodeficiency virus 1 in the southwest border of china. *PLoS One*, 9, e107578.
- Chun, T. W. & Fauci, A. S. 2012. HIV reservoirs: Pathogenesis and obstacles to viral eradication and cure. *AIDS*, 26, pp1261-1268.
- Clavel, F. & Mammano, F. 2010. Role of gag in HIV resistance to protease inhibitors. *Viruses*, 2, 1411-1426.
- Coffin, J. M., Hughes, S. H. & Varmus, H. E. 1997. *The interactions of retroviruses and their hosts*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY).
- Cohen, M. S., Shaw, G. M., McMichael, A. J. & Haynes, B. F. 2011. Acute HIV-1 infection. *N Engl J Med*, 364, pp1943-1954.
- Conradie, F., Wilson, D., Basson, A., De Oliveira, T., Hunt, G., Joel, D., Papathanasopoulos, M., Preiser, W., Klausner, J., Spencer, D., Stevens, W., Venter, F., Van Vuuren, C., Levin, L., Meintjes, G., Orrell, C., Sunpath, H., Rossouw, T. & Van Zyl, G. 2012. The 2012 southern african arv drug resistance testing guidelines : Guidelines. *Southern African Journal of HIV Medicine*, 13, 162-167.
- Craigie, R. & Bushman, F. D. 2012. HIV DNA integration. *Cold Spring Harbor perspectives in medicine*, 2.
- Dam, E., Quercia, R., Glass, B., Descamps, D., Launay, O., Duval, X., Kräusslich, H.-G., Hance, A. J., Clavel, F. & Group, A. S. 2009. Gag mutations strongly contribute to HIV-1 resistance to protease inhibitors in highly drug-experienced patients besides compensating for fitness loss. *PLoS Pathog*, 5, e1000345.
- Darriba, D. & Posada, D. 2016. Jmodeltest 2 manual v0.1.10. pp1-27.
- Darriba, D., Taboada, G., Doallo, R. & Posada, D. 2012. Modeltest 2: More models, new heuristics and parallel computing. *Nature Methods*, 9(8), 772.
- Datamonkey. www.datamonkey.org. *Methods* [Online]. Available: www.datamonkey.org [Accessed 4 April 2018].

- Daw, M. A., El-Bouzedi, A., Ahmed, M. O., Dau, A. A., In association with the Libyan Study Group of, H. & Hiv 2017. Molecular and epidemiological characterization of HIV-1 subtypes among libyan patients. *BMC research notes*, 10, 170-170.
- De Groot, A. S., Rivera, D. S., McMurry, J. A., Buus, S. & Martin, W. 2008. Identification of immunogenic hla-b7 "achilles' heel" epitopes within highly conserved regions of HIV. *Vaccine*, 26, 3059-3071.
- Delport, W., Poon, A. F., Frost, S. D. & Kosakovsky Pond, S. L. 2010. Datamonkey 2010: A suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics*, 26, pp2455-2457.
- Drummond, A. J., Ho, S. Y., Rawlence, N. & Rambaut, A. 2007. A rough guide to BEAST 1.4.
- Drummond, A. J. & Rambaut, A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, 7, 214.
- Drummond, A. J. & Rambaut, A. 2009. Bayesian evolutionary analysis by sampling trees. In: Lemey, P., Salemi, M. & Vandamme, A. (eds.) *The phylogenetic handbook: A practical approach to phylogenetic analysis and hypothesis testing*. 2nd ed. New York: Cambridge University Press.
- Fanales-Belasio, E., Raimondo, M., Suligoj, B. & Butto, S. 2010. HIV virology and pathogenetic mechanisms of infection: A brief overview. *Ann Ist Super Sanita*, 46, pp4-15.
- Fiebig, E. W., Wright, D. J., Rawal, B. D., Garrett, P. E., Schumacher, R. T., Peddada, L., Heldebrant, C., Smith, R., Conrad, A., Kleinman, S. H. & Busch, M. P. 2003. Dynamics of HIV viremia and antibody seroconversion in plasma donors: Implications for diagnosis and staging of primary HIV infection. *Aids*, 17, 1871-9.
- Frankel, A. D. & Young, J. A. 1998. HIV-1: Fifteen proteins and an rna. Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA.
- Freed, E. O. 2001. HIV-1 replication. *Somat Cell Mol Genet*, 26, 13-33.
- Friedman-Kien, A. E. 1981. Disseminated kaposi's sarcoma syndrome in young homosexual men. *J Am Acad Dermatol*, 5, pp468-471.
- Gao, F., Bailes, E., Robertson, D. L., Chen, Y., Rodenburg, C. M., Michael, S. F., Cummins, L. B., Arthur, L. O., Peeters, M., Shaw, G. M., Sharp, P. M. & Hahn, B. H. 1999. Origin of HIV-1 in the chimpanzee pan troglodytes troglodytes. *Nature*, 397, pp436-441.
- Gao, F., Robertson, D. L., Carruthers, C. D., Li, Y., Bailes, E., Kostrikis, L. G., Salminen, M. O., Bibollet-Ruche, F., Peeters, M. & Ho, D. D. 1998. An isolate of human immunodeficiency virus type 1 originally classified as subtype i represents a complex mosaic comprising three different group m subtypes (a, g, and i). *Journal of Virology*, 72, 10234-10241.

- Gao, F., Yue, L., White, A. T., Pappas, P. G., Barchue, J., Hanson, A. P., Greene, B. M., Sharp, P. M., Shaw, G. M. & Hahn, B. H. 1992. Human infection by genetically diverse sivsm-related HIV-2 in west africa. *Nature*, 358, pp495-499.
- Garcia-Diaz, A., McCormick, A., Booth, C., Gonzalez, D., Sayada, C., Haque, T., Johnson, M. & Webster, D. 2014. Analysis of transmitted HIV-1 drug resistance using 454 ultra-deep-sequencing and the deepchek((r))-HIV system. *J Int AIDS Soc*, 17, 19752.
- Gatto, L., Catanzaro, D. & Milinkovitch, M. C. 2006. Assessing the applicability of the gtr nucleotide substitution model through simulations. *Evol Bioinform Online*, 2, 145-55.
- GeneBank. <https://www.ncbi.nlm.nih.gov/genbank/>. *Genebank overview* [Online]. [Accessed 20 December 2018].
- Ghosn, J., Delaugerre, C., Flandre, P., Galimand, J., Cohen-Codar, I., Raffi, F., Delfraissy, J. F., Rouzioux, C. & Chaix, M. L. 2011. Polymorphism in gag gene cleavage sites of HIV-1 non-b subtype and virological outcome of a first-line lopinavir/ritonavir single drug regimen. *PLoS One*, 6, e24798.
- Gianella, S. & Richman, D. D. 2010. Minority variants of drug-resistant HIV. *Journal of Infectious Diseases*, 202, 657-666.
- Gordon, M., De Oliveira, T., Bishop, K., Coovadia, H. M., Madurai, L., Engelbrecht, S., Janse van Rensburg, E., Mosam, A., Smith, A. & Cassol, S. 2003. Molecular characteristics of human immunodeficiency virus type 1 subtype c viruses from kwazulu-natal, South Africa: Implications for vaccine and antiretroviral control strategies. *Journal of Virology*, 77, 2587-2599.
- Gottlieb MS, Schroff R, Schanker HM, Weisman JD, Fan PT, Wolf RA & A, S. 1981. Pneumocystis carinii pneumonia and mucosal candidiasis in previously healthy homosexual men: Evidence of a new acquired cellular immunodeficiency. *N Engl J Med*, 305, pp1425-1431.
- Guindon, S. & Gascuel, O. 2003. A simple, fast and accurate method to estimate large phylogenies by maximum-likelihood. *Systematic Biology*, 52, 696-704.
- Halvas, E. K., Wiegand, A., Boltz, V. F., Kearney, M., Nissley, D., Wantman, M., Hammer, S. M., Palmer, S., Vaida, F., Coffin, J. M. & Mellors, J. W. 2010. Low frequency nonnucleoside reverse-transcriptase inhibitor—resistant variants contribute to failure of efavirenz-containing regimens in treatment-experienced patients. *Journal of Infectious Diseases*, 201, 672-680.
- Heath, T. A. Divergence time estimation using BEAST v2. 2.0. Source URL: <http://treehinkers.org/tutorials/divergence-time-estimation-using-beast/>: Tutorial written for workshop on applied phylogenetics and molecular evolution, Bodega Bay California, 2015. 1-44.
- Hemelaar, J., Gouws, E., Ghys, P. D. & Osmanov, S. 2006. Global and regional distribution of HIV-1 genetic subtypes and recombinants in 2004. *Aids*, 20, W13-23.

Hemelaar, J., Gouws, E., Ghys, P. D., Osmanov, S., Isolation, W.-U. N. f. H. & Characterisation 2011. Global trends in molecular epidemiology of HIV-1 during 2000-2007. *AIDS (London, England)*, 25, 679-689.

Huelsenbeck, J. & Rannala, B. 2004. Frequentist properties of bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. *Systematic Biology*, 53.

Huelsenbeck, J. P. & Ronquist, F. 2001. Mrbayes: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17.

Hughes, A. L. 2007. Looking for darwin in all the wrong places: The misguided quest for positive selection at the nucleotide sequence level. *Heredity*, 99, 364.

Illing, P. T., Pymm, P., Croft, N. P., Hilton, H. G., Jovic, V., Han, A. S., Mendoza, J. L., Mifsud, N. A., Dudek, N. L., McCluskey, J., Parham, P., Rossjohn, J., Vivian, J. P. & Purcell, A. W. 2018. Hla-b57 micropolymorphism defines the sequence and conformational breadth of the immunopeptidome. *Nature communications*, 9, 4693-4693.

Imaz, A., Falco, V. & Ribera, E. 2011. Antiretroviral salvage therapy for multiclass drug-resistant HIV-1-infected patients: From clinical trials to daily clinical practice. *AIDS Rev*, 13, 180-93.

Jülg, B. & Goebel, F. D. 2005. HIV genetic diversity: Any implications for drug resistance? *Infection*, 33, 299-301.

Kalish, M. L., Wolfe, N. D., Ndongmo, C. B., McNicholl, J., Robbins, K. E., Aidoo, M., Fonjungo, P. N., Alemnji, G., Zeh, C., Djoko, C. F., Mpoudi-Ngole, E., Burke, D. S. & Folks, T. M. 2005. Central african hunters exposed to simian immunodeficiency virus. *Emerg. Infect. Dis.*, 11, pp1928-1930.

Kandathil, A. J., Ramalingam, S., Kannangai, R., David, S. & Sridharan, G. 2005. Molecular epidemiology of HIV. *Indian J Med Res*, 121, pp333-344.

Kaslow, R. A., Carrington, M., Apple, R., Park, L., Munoz, A., Saah, A., Goedert, J. J., Winkler, C., O'Brien, S. J. & Rinaldo, C. 1996. Influence of combinations of human major histocompatibility complex genes on the course of HIV-1 infection. *Nature medicine*, 2, 405.

Khalid, Z. & Sezerman, O. U. 2016. Prediction of HIV drug resistance by combining sequence and structural properties. *IEEE/ACM Trans Comput Biol Bioinform*.

Kiepiela, P., Ngumbela, K., Thobakgale, C., Ramduth, D., Honeyborne, I., Moodley, E., Reddy, S., de Pierres, C., Mncube, Z., Mkhwanazi, N., Bishop, K., van der Stok, M., Nair, K., Khan, N., Crawford, H., Payne, R., Leslie, A., Prado, J., Prendergast, A., Frater, J., McCarthy, N., Brander, C., Learn, G. H., Nickle, D., Rousseau, C., Coovadia, H., Mullins, J. I., Heckerman, D., Walker, B. D. & Goulder, P. 2006. Cd8+ t-cell responses to different HIV proteins have discordant associations with viral load. *Nature Medicine*, 13, 46.

Kimura, M. 1968. Evolutionary rate at the molecular level. *Nature*, pp624-626.

- Kimura, M. 1983. *The neutral theory of molecular evolution*, Cambridge, Cambridge University Press.
- Kloverpris, H. N., Stryhn, A., Harndahl, M., van der Stok, M., Payne, R. P., Matthews, P. C., Chen, F., Riddell, L., Walker, B. D., Ndung'u, T., Buus, S. & Goulder, P. 2012. Hla-b*57 micropolymorphism shapes hla allele-specific epitope immunogenicity, selection pressure, and HIV immune control. *Journal of virology*, 86, 919-929.
- Korber, B., Muldoon, M., Theiler, J., Gao, F., Gupta, R., Lapedes, A., Hahn, B. H., Wolinsky, S. & Bhattacharya, T. 2000. Timing the ancestor of the HIV-1 pandemic strains. *Science*, 288, pp1789-1796.
- Kosakovsky Pond, S. L. & Frost, S. D. W. 2005. Not so different after all: A comparison of methods for detecting amino acid sites under selection. *Molecular Biology and Evolution*, 22, 1208-1222.
- Kosakovsky Pond, S. L., Frost, S. D. W. & Muse, S. V. 2005. Hyphy: Hypothesis testing using phylogenies. *Bioinformatics*, 21, 676-679.
- Kozisek, M., Henke, S., Saskova, K. G., Jacobs, G. B., Schuch, A., Buchholz, B., Muller, V., Krausslich, H. G., Rezacova, P., Konvalinka, J. & Bodem, J. 2012. Mutations in HIV-1 gag and pol compensate for the loss of viral fitness caused by a highly mutated protease. *Antimicrob Agents Chemother*, 56, 4320-30.
- Kuiken, C., Korber, B. & Shafer, R. W. 2003. HIV sequence databases. *AIDS reviews*, 5, 52-61.
- Kullback, S. & Leibler, R. A. 1951. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1).
- Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J. & Higgins, D. G. 2007. Clustal w and clustal x version 2.0. *Bioinformatics*, 23, pp2947-2948.
- Lemey, P., Salemi, M. & Vandamme, A. 2009. *The phylogenetic handbook: A practical approach to phylogenetic analysis and hypothesis testing (2nd)*, New York, Cambridge University Press.
- Li, H., Dou, J., Ding, L. & Spearman, P. 2007. Myristoylation is required for human immunodeficiency virus type 1 gag-gag multimerization in mammalian cells. *J Virol Methods*, 81(23), pp12899-12910.
- Li, W. H. 1997. *Molecular evolution*, Sunderland, MA, Sinauer.
- Liang, B., Luo, M., Scott-Herridge, J., Semeniuk, C., Mendoza, M., Capina, R., Sheardown, B., Ji, H., Kimani, J., Ball, B. T., Van Domselaar, G., Graham, M., Tyler, S., Jones, S. J. & Plummer, F. A. 2011. A comparison of parallel pyrosequencing and sanger clone-based sequencing and its impact on the characterization of the genetic diversity of HIV-1. *PLoS One*, 6, e26745.

- Liégeois, F., Lafay, B., Formenty, P., Locatelli, S., Courgnaud, V., Delaporte, E. & Peeters, M. 2009. Full-length genome characterization of a novel simian immunodeficiency virus lineage (sivolv) from olive colobus (*procolobus verus*) and new sivwrcpbb strains from western red colobus (*piliocolobus badius badius*) from the tai forest in ivory coast. *J. Virol.*, 83, pp428-439.
- Lihana, R. W., Khamadi, S. A., Lwembe, R. M., Kinyua, J. G., Muriuki, J. K., Lagat, N. J., Okoth, F. A., Makokha, E. P. & Songok, E. M. 2009. HIV-1 subtype and viral tropism determination for evaluating antiretroviral therapy options: An analysis of archived kenyan blood samples. *BMC Infect Dis*, 9, 215.
- Lihana, R. W., Ssemwanga, D., Abimiku, A. & Ndembi, N. 2012. Update on HIV-1 diversity in africa: A decade in review. *AIDS Rev*, 14, 83-100.
- Lio, P. & Goldman, N. 1998. Models of molecular evolution and phylogeny. *Genome Res*, 8, 1233-44.
- Llano, A., Frahm, N. & Brander, C. 2009. How to optimally define optimal cytotoxic t lymphocyte epitopes in HIV infection. *HIV molecular immunology*, 2009, 3-24.
- Llano, A., Williams, A., Olvera, A., Silva-Arrieta, S. & Brander, C. 2013. Best-characterized HIV-1 ctl epitopes: The 2013 update. *HIV molecular immunology*, 2013, 3-25.
- Locatelli, S., Lafay, B., Liegeois, F., Ting, N., Delaporte, E. & Peeters, M. 2008. Full molecular characterization of a simian immunodeficiency virus, sivwrcpbt from temminck's red colobus (*piliocolobus badius temminckii*) from abuko nature reserve, the gambia. *Virology*, 376, pp90-100.
- Lodish, H., Berk, A., Zipursky, S., Matsudaira, P., Baltimore, D. & Darnell, J. 2000. Collagen: The fibrous proteins of the matrix. *Molecular Cell Biology*, 4.
- Los Alamos National Laboratory. 2018a. *The circulating recombinant forms (crfs)* [Online]. Available: [<http://www.hiv.lanl.gov/content/sequence/HIV/CRFs/CRFs.html>] [Accessed 10 November 2018].
- Los Alamos National Laboratory. 2018b. *Ctl/cd8+ epitope summary* [Online]. Available: [https://www.hiv.lanl.gov/content/immunology/tables/ctl_summary.html] [Accessed 19 December 2018].
- Los Alamos National Laboratory. [<https://www.hiv.lanl.gov/content/index>]. *HIV databases* [Online]. [Accessed 14 December 218].
- Maartens, G., Celum, C. & Lewin, S. R. 2014. HIV infection: Epidemiology, pathogenesis, treatment, and prevention. *Lancet*, 384, pp258-271.
- Mapumulo, Z. 2016. Budget woes won't stop new arv programme. *City Press* [Online]. Available: [www.city-press.news24.com/News/budget-woes-wont-stop-new-arv-programme-20160901-3] [Accessed 14 March 2017].
- Marsden, M. D. & Zack, J. A. 2013. HIV/aids eradication. *Bioorg Med Chem Lett*, 23, pp4003-4010.

- Marx, J. L. 1982. New disease baffles medical community. *Science* 217, pp618-621.
- McMichael, A. J., Borrow, P., Tomaras, G. D., Goonetilleke, N. & Haynes, B. F. 2010. The immune response during acute HIV-1 infection: Clues for vaccine development. *Nat Rev Immunol*, 10, pp11-23.
- Melikyan, G. B., Markosyan, R. M., Hemmati, H., Delmedico, M. K., Lambert, D. M. & Cohen, F. S. 2000. Evidence that the transition of HIV-1 gp41 into a six-helix bundle, not the bundle configuration, induces membrane fusion. *J Cell Biol*, 151, pp413-423.
- Metzner, K. J., Rauch, P., Walter, H., Boesecke, C., Zollner, B., Jessen, H., Schewe, K., Fenske, S., Gellermann, H. & Stellbrink, H. J. 2005. Detection of minor populations of drug-resistant HIV-1 in acute seroconverters. *AIDS*, 19, 1819-25.
- Mohamed, S., Penaranda, G., Gonzalez, D., Camus, C., Khiri, H., Boulme, R., Sayada, C., Philibert, P., Olive, D. & Halfon, P. 2014. Comparison of ultra-deep versus sanger sequencing detection of minority mutations on the HIV-1 drug resistance interpretations after virological failure. *Aids*, 28, 1315-24.
- Nasioulas, G., Paraskevis, D., Magiorkinis, E., Theodoridou, M. & Hatzakis, A. 1999. Molecular analysis of the full-length genome of HIV type 1 subtype i: Evidence of a/g/i recombination. *AIDS research and human retroviruses*, 15, 745-758.
- Ndhlovu, Z., Kanya, P., Mewalal, N., Kløverpris, Henrik N., Nkosi, T., Pretorius, K., Laher, F., Ogunshola, F., Chopera, D., Shekhar, K., Ghebremichael, M., Ismail, N., Moodley, A., Malik, A., Leslie, A., Goulder, Philip J. R., Buus, S., Chakraborty, A., Dong, K., Ndung'u, T. & Walker, Bruce D. 2015. Magnitude and kinetics of cd8+ t cell activation during hyperacute HIV infection impact viral set point. *Immunity*, 43, 591-604.
- Nijhuis, M., van Maarseveen, N. M., Lastere, S., Schipper, P., Coakley, E., Glass, B., Rovenska, M., de Jong, D., Chappey, C., Goedegebuure, I. W., Heilek-Snyder, G., Dulude, D., Cammack, N., Brakier-Gingras, L., Konvalinka, J., Parkin, N., Kräusslich, H.-G., Brun-Vezinet, F. & Boucher, C. A. B. 2007. A novel substrate-based HIV-1 protease inhibitor drug resistance mechanism. *PLoS Med*, 4, e36.
- Noguera-Julian, M., Casadellà, M., Pou, C., Rodríguez, C., Pérez-Álvarez, S., Puig, J., Clotet, B. & Paredes, R. 2013. Stable HIV-1 integrase diversity during initial HIV-1 rna decay suggests complete blockade of plasma HIV-1 replication by effective raltegravir-containing salvage therapy. *Virology Journal*, 10, 350.
- Ntale, R., Chopera, D., Ngandu, N., de Rosa, D. A., Zembe, L., Gamielien, H., Mlotshwa, M., Werner, L., Woodman, Z. & Mlisana, K. 2012. Temporal association of hla-b* 81: 01 and b* 39: 10 mediated HIV-1 p24 sequence evolution with disease progression. *Journal of virology*, JVI. 00539-12.
- Nyombi, B. M., Holm-Hansen, C., Kristiansen, K. I., Bjune, G. & Muller, F. 2008. Prevalence of reverse transcriptase and protease mutations associated with antiretroviral drug resistance among drug-naïve HIV-1 infected pregnant women in kagera and kilimanjaro regions, tanzania. *AIDS Res Ther*, 5, 13.
- Palmer, S., Kearney, M., Maldarelli, F., Halvas, E., Bixby, C., Bazmi, H., Rock, D., Falloon, J., Davey, R. & Dewar, R. 2005. Multiple, linked human immunodeficiency virus

type 1 drug resistance mutations in treatment-experienced patients are missed by standard genotype analysis. *J Clin Microbiol*, 43, 406 - 413.

Paraschiv, S., Otelea, D., Dinu, M., Maxim, D. & Tinischi, M. 2007. Polymorphisms and resistance mutations in the protease and reverse transcriptase genes of HIV-1 f subtype romanian strains. *Int J Infect Dis*, 11, 123-8.

Paredes i Deiros, R. 2009. Clinical implications of minority HIV-1 resistant variants.

Peeters, M. & Sharp, P. M. 2000. Genetic diversity of HIV-1: The moving target. *AIDS*, 14.

Pereyra, F., Heckerman, D., Carlson, J. M., Kadie, C., Soghoian, D. Z., Karel, D., Goldenthal, A., Davis, O. B., DeZiel, C. E., Lin, T., Peng, J., Piechocka, A., Carrington, M. & Walker, B. D. 2014. HIV control is mediated in part by cd8⁺ t-cell targeting of specific epitopes. *Journal of Virology*, 88, 12937-12948.

Pilcher, C. D., Eron, J. J., Galvin, S., Gay, C. & Cohen, M. S. 2004. Acute HIV revisited: New opportunities for treatment and prevention. *Journal of Clinical Investigation*, 113, 937-945.

Poon, A. F., Frost, S. D. & Kosakovsky Pond, S. L. 2009. Detecting signatures of selection from DNA sequences using datamonkey. *Bioinformatics for DNA Sequence Analysis*, 163-183.

Pope, M. & Haase, A. T. 2003. Transmission, acute HIV-1 infection and the quest for strategies to prevent infection. *Nat. Med.*, 9.

Posada, D. 2008. Jmodeltest: Phylogenetic model averaging. *Molecular Biology and Evolution*, 25, 1253-1256.

Posada, D. 2009. Selecting models of evolution. In: Lemey, P., Salemi, M. & Vandamme, A. (eds.) *The phylogenetic handbook: A practical approach to phylogenetic analysis and hypothesis testing*. 2nd ed. New York: Cambridge University Press.

Posada, D. & Crandall, K. A. 1998. Modeltest: Testing the model of DNA substitution. *Bioinformatics*, 14, 817-8.

Prosite. www.prosite.expasy.org. *Expasy bioinformatics resource portal* [Online]. Available: www.prosite.expasy.org [Accessed 18 November 2018].

Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. 2017. *Analysing BEAST output using tracer* [Online]. Available: http://beast.community/analysing_beast_output [Accessed 24 October 2018].

Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. 2018. *Tracer v1.7* [Online]. Available: <http://beast.community/tracer> [Accessed 18 October 2018].

Ramirez, C., Gregori, J., Buti, M., Tabernero, D., Camos, S., Casillas, R., Quer, J., Esteban, R., Homs, M. & Rodriguez-Frias, F. 2013. A comparative study of ultra-deep pyrosequencing and cloning to quantitatively analyze the viral quasispecies using hepatitis b virus infection as a model. *Antiviral Res*, 98, 273-83.

- Recordon-Pinson, P., Alves, B. M., Tumiotto, C., Bellecave, P., Bonnet, F., Neau, D., Soares, E. A., Soares, M. A. & Fleury, H. 2018. A new HIV-1 circulating recombinant form (crf98_cpx) between crf06_cpx and subtype b identified in southwestern france. *AIDS research and human retroviruses*.
- Riou, C., Ganusov, V. V., Champion, S., Mlotshwa, M., Liu, M. K., Whale, V. E., Goonetilleke, N., Borrow, P., Ferrari, G., Betts, M. R., Haynes, B. F., McMichael, A. J. & Gray, C. M. 2012. Distinct kinetics of gag-specific cd4+ and cd8+ t cell responses during acute HIV-1 infection. *J Immunol*, 188, 2198-206.
- Rizzo, J. & Rouchka, E. C. 2007. Review of phylogenetic tree construction. *University of Louisville Bioinformatics Laboratory Technical Report Series*.
- Robertson, D. L., Anderson, J. P., Bradac, J. A., Carr, J. K., Foley, B., Funkhouser, R. K., Gao, F., Hahn, B. H., Kalish, M. L., Kuiken, C., Learn, G. H., Leitner, T., McCutchan, F., Osmanov, S., Peeters, M., Pieniazek, D., Salminen, M., Sharp, P. M., Wolinsky, S. & Korber, B. 2000. HIV-1 nomenclature proposal. *Science*, 288, pp55-56.
- Ronquist, F., van der Mark, P. & Huelsenbeck, J. P. 2009. Bayesian phylogenetic analysis using mrbayes. In: Lemey, P., Salemi, M. & Vandamme, A. (eds.) *The phylogenetic handbook: A practical approach to phylogenetic analysis and hypothesis testing*. 2nd ed. New York: Cambridge University Press.
- Samuel, R., Julian, M. N., Paredes, R., Parboosing, R., Moodley, P., Singh, L., Naidoo, A. & Gordon, M. 2016. HIV-1 drug resistance by ultra-deep sequencing following short course zidovudine, single-dose nevirapine, and single-dose tenofovir with emtricitabine for prevention of mother-to-child transmission. *J Acquir Immune Defic Syndr*, 73, 384-389.
- San Mauro, D. & Agorreta, A. 2010. Molecular systematics: A synthesis of the common methods and the state of knowledge. *Cell Mol Biol Lett*, 15, 311-41.
- Santos, A. F. & Soares, M. A. 2010. HIV genetic diversity and drug resistance. *Viruses*, 2, 503-531.
- Schmidt, H. A. & von Haeseler, A. 2009. Phylogenetic inference using maximum likelihood methods. In: Lemey, P., Salemi, M. & Vandamme, A. (eds.) *The phylogenetic handbook: A practical approach to phylogenetic analysis and hypothesis testing*. 2nd ed. New York: Cambridge University Press.
- Shafer, R. W. 2002. Genotypic testing for human immunodeficiency virus type 1 drug resistance. *Clin Microbiol Rev*, 15, 247-77.
- Shafer, Robert W. 2006. Rationale and uses of a public HIV drug-resistance database. *Journal of Infectious Diseases*, 194, S51-S58.
- Shafer, R. W., Rhee, S. Y., Pillay, D., Miller, V., Sandstrom, P., Schapiro, J. M., Kuritzkes, D. R. & Bennett, D. 2007. HIV-1 protease and reverse transcriptase mutations for drug resistance surveillance. *AIDS*, 21, 215-23.
- Shen, X.-L., Li, S.-Z., Li, Y.-Q. & Chen, X. 2010. Episodic sitewise positive selection on the signal recognition particle protein fflh in actinobacteria. *FEBS Letters*, 584, 3975-3978.

- Shugars, D. C., Alexander, A. L., Fu, K. & Freel, S. A. 1999. Endogenous salivary inhibitors of human immunodeficiency virus. *Arch. Oral Biol.*, 44, pp445-453.
- Shugars, D. C. & Wahl, S. M. 1998. The role of the oral environment in HIV-1 transmission. *J. Am. Dent. Assoc.*, 129, pp851-858.
- Sierra, S., Kupfer, B. & Kaiser, R. 2005. Basics of the virology of HIV-1 and its replication. *Journal of Clinical Virology*, 34, 233-244.
- Simen, B. B., Simons, J. F., Hullsiek, K. H., Novak, R. M., Macarthur, R. D., Baxter, J. D., Huang, C., Lubeski, C., Turenchalk, G. S., Braverman, M. S., Desany, B., Rothberg, J. M., Egholm, M. & Kozal, M. J. 2009. Low-abundance drug-resistant viral variants in chronically HIV-infected, antiretroviral treatment-naïve patients significantly impact treatment outcomes. *J Infect Dis*, 199, 693-701.
- Singh, K., Marchand, B., Kirby, K. A., Michailidis, E. & Sarafianos, S. G. 2010. Structural aspects of drug resistance and inhibition of HIV-1 reverse transcriptase. *Viruses*, 2, 606-638.
- Singh, U. 2015. *Acquired and transmitted drug resistance in HIV-1 subtype-c: Implications of novel mutations on replication capacity, cleavage and drug susceptibility*. PhD Thesis, University of KwaZulu-Natal.
- Slonczewski, J. & Foster, J. W. 2017. *Microbiology : An evolving science (4ed)*, New York, W.W. Norton & Company.
- Smith, S. J., Pauly, G. T., Akram, A., Melody, K., Rai, G., Maloney, D. J., Ambrose, Z., Thomas, C. J., Schneider, J. T. & Hughes, S. H. 2016. Rilpivirine analogs potently inhibit drug-resistant HIV-1 mutants. *Retrovirology*, 13, 11.
- Soares, E. A. J. M., Santos, A. F. A., Sousa, T. M., Sprinz, E., Martinez, A. M. B., Silveira, J., Tanuri, A. & Soares, M. A. 2007. Differential drug resistance acquisition in HIV-1 of subtypes b and c. *PLoS ONE*, 2, e730.
- Stacey, A. R., Norris, P. J., Qin, L., Haygreen, E. A., Taylor, E., Heitman, J., Lebedeva, M., DeCamp, A., Li, D., Grove, D., Self, S. G. & Borrow, P. 2009. Induction of a striking systemic cytokine cascade prior to peak viremia in acute human immunodeficiency virus type 1 infection, in contrast to more modest and delayed responses in acute hepatitis b and c virus infections. *J Virol Methods*, 83, pp3719-3733.
- Stadler, T., Kühnert, D., Bonhoeffer, S. & Drummond, A. J. 2013. Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis c virus (hcv). *Proceedings of the National Academy of Sciences*, 110, 228-233.
- Stanford University HIV Drug Resistance Database. 2015. *Major HIV-1 drug resistance mutations* [Online]. Available: <http://hivdb.stanford.edu> [Accessed 12 May 2017].
- Stanford University HIV Drug Resistance Database. 2017. *Major HIV-1 drug resistance mutations* [Online]. Available: <https://hivdb.stanford.edu/assets/media/resistance-mutation-handout-Dec2017.b8f72e32.pdf> [Accessed 14 Mar 2018].

Stanford University HIVdb. <https://hivdb.stanford.edu/>. *Major HIV-1 drug resistance mutations* [Online]. Available: <https://hivdb.stanford.edu/assets/media/resistance-mutation-handout-Dec2017.b8f72e32.pdf> [Accessed 22 August 2018].

Steege, K., Carmona, S., Bronze, M., Papathanasopoulos, M. A., van Zyl, G., Goedhals, D., MacLeod, W., Sanne, I. & Stevens, W. S. 2016. Moderate levels of pre-treatment HIV-1 antiretroviral drug resistance detected in the first south african national survey. *PLoS One*, 11, e0166305.

Stelzl, E., Proll, J., Bizon, B., Niklas, N., Danzer, M., Hackl, C., Stabentheiner, S., Gabriel, C. & Kessler, H. H. 2011. Human immunodeficiency virus type 1 drug resistance testing: Evaluation of a new ultra-deep sequencing-based protocol and comparison with the trugene HIV-1 genotyping kit. *J Virol Methods*, 178, 94-7.

Strimmer, K. & von Haeseler, A. 2009. Genetic distances and nucleotide substitution models. In: Lemey, P., Salemi, M. & Vandamme, A. (eds.) *The phylogenetic handbook: A practical approach to phylogenetic analysis and hypothesis testing*. 2nd ed. New York: Cambridge University Press.

Su, C. T., Ling, W. L., Lua, W. H., Haw, Y. X. & Gan, S. K. 2016. Structural analyses of 2015-updated drug-resistant mutations in HIV-1 protease: An implication of protease inhibitor cross-resistance. *BMC Bioinformatics*, 17, 500.

Suzuki, Y., Yamaguchi-Kabata, Y. & Gojobori, T. 2000. *Nucleotide substitution rates of hiv1*.

Takehisa, J., Kraus, M. H., Ayouba, A., Bailes, E., Van Heuverswyn, F., Decker, J. M., Li, Y., Rudicell, R. S., Learn, G. H., Neel, C., Ngole, E. M., Shaw, G. M., Peeters, M., Sharp, P. M. & Hahn, B. H. 2009. Origin and biology of simian immunodeficiency virus in wild-living western gorillas. *J. Virol.*, 83, pp1635-1648.

Tang, M. W. & Shafer, R. W. 2012. HIV-1 antiretroviral resistance: Scientific principles and clinical applications. *Drugs*, 72, e1-25.

Taubenberger, J. K., Reid, A. H., Janczewski, T. A. & Fanning, T. G. 2001. Integrating historical, clinical and molecular genetic data in order to explain the origin and virulence of the 1918 spanish influenza virus. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, 356, pp1829-1839.

The Joint United Nations Programme on HIV/AIDS (UNAIDS). 2016a. *Fact sheet - latest statistics on the status of the aids epidemic (global fact sheet)* [Online]. Available: <http://www.unaids.org/en/resources/fact-sheet> [Accessed 12 April 2017 2017].

The Joint United Nations Programme on HIV/AIDS (UNAIDS). 2016b. *South Africa HIV and aids estimates (2015)* [Online]. Available: <http://www.unaids.org/en/regionscountries/countries/southafrica/> [Accessed 12 April 2017 2017].

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. 1997. The clustalx windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research*, 25, pp4876-4882.

- Thomson, M. M., Perez Alvarez, L. & Najera, R. 2002. Molecular epidemiology of HIV-1 genetic forms and its significance for vaccine development and therapy. *Lancet Infect Dis*, 2.
- Toor, J. S., Sharma, A., Kumar, R., Gupta, P., Garg, P. & Arora, S. K. 2011. Prediction of drug-resistance in HIV-1 subtype c based on protease sequences from art naive and first-line treatment failures in north india using genotypic and docking analysis. *Antiviral Res*, 92, 213-8.
- Udenwobele, D. I., Su, R.-C., Good, S. V., Ball, T. B., Shrivastav, S. V. & Shrivastav, A. 2017. Myristoylation: An important protein modification in the immune response. *Front. Immunol.* 8:751, 8:751, pp1-16.
- UNAIDS. 2018a. *Global HIV & aids statistics — 2018 fact sheet* [Online]. Available: <http://www.unaids.org/en/resources/fact-sheet> [Accessed].
- UNAIDS. 2018b. *South Africa HIV and aids estimates* [Online]. Available: <http://www.unaids.org/en/regionscountries/countries/southafrica> [Accessed 28 November 2018].
- Van Heuverswyn, F., Li, Y., Neel, C., Bailes, E., Keele, B. F., Liu, W., Loul, S., Butel, C., Liegeois, F., Bienvenue, Y., Ngolle, E. M., Sharp, P. M., Shaw, G. M., Delaporte, E., Hahn, B. H. & Peeters, M. 2006. Human immunodeficiency viruses: SIV infection in wild gorillas. *Nature*, 444, 164.
- Vandamme, A.-M. 2009. Basic concepts of molecular evolution. In: Lemey, P., Salemi, M. & Vandamme, A. (eds.) *The phylogenetic handbook: A practical approach to phylogenetic analysis and hypothesis testing*. 2nd ed. New York: Cambridge University Press.
- Wainberg, M. A. 2004. HIV-1 subtype distribution and the problem of drug resistance. *AIDS*, 18, S63-S68.
- Wang, C., Mitsuya, Y., Gharizadeh, B., Ronaghi, M. & Shafer, R. W. 2007. Characterization of mutation spectra with ultra-deep pyrosequencing: Application to HIV-1 drug resistance. *Genome Res*, 17, 1195-201.
- Watts, J. M., Dang, K. K., Gorelick, R. J., Leonard, C. W., Bess, J. W., Jr., Swanstrom, R., Burch, C. L. & Weeks, K. M. 2009. Architecture and secondary structure of an entire HIV-1 rna genome. *Nature*, 460, 711-716.
- Wolfe, N. D., Switzer, W. M., Carr, J. K., Bhullar, V. B., Shanmugam, V., Tamoufe, U., Prosser, A. T., Torimiro, J. N., Wright, A., Mpoudi-Ngole, E., McCutchan, F. E., Birx, D. L., Folks, T. M., Burke, D. S. & Heneine, W. 2004. Naturally acquired simian retrovirus infections in central african hunters. *Lancet*, 363, pp932-937.
- World Health Organisation. 2016. *Consolidated guidelines on the use of antiretroviral drugs for treating and preventing HIV infection: Recommendations for a public health approach* [Online]. Available: <http://www.who.int/hiv/pub/arv/arv-2016/en/> [Accessed 04 May 2017].
- Worobey, M., Gemmel, M., Teuwen, D. E., Haselkorn, T., Kunstman, K., Bunce, M., Muyembe, J.-J., Kabongo, J.-M. M., Kalengayi, R. M., Van Marck, E., Gilbert, M. T. P. &

- Wolinsky, S. M. 2008. Direct evidence of extensive diversity of HIV-1 in kinshasa by 1960. *Nature*, 455, 661-664.
- Wright, D. W., Deuzing, I. P., Flandre, P., van den Eede, P., Govaert, M., Setiawan, L., Coveney, P. V., Marcelin, A. G., Calvez, V., Boucher, C. A. & Beerens, N. 2013. A polymorphism at position 400 in the connection subdomain of HIV-1 reverse transcriptase affects sensitivity to nrtis and rnaseh activity. *PLoS One*, 8, e74078.
- Wright, M. H., Heal, W. P., Mann, D. J. & Tate, E. W. 2010. Protein myristoylation in health and disease. *J Chem Biol*, 3, pp19-35.
- Yang, Z. 2007a. Molecular phylogenetics: Principles and practice. In: D.J. Balding, M. Bishop & C. Cannings (eds.) *Handbook of statistical genetics*. John Wiley & Sons, Ltd.
- Yang, Z. 2007b. PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24, 1586-1591.
- Yang, Z., Kumar, S. & Nei, M. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics*, 141, 1641-1650.
- Yang, Z., Nielsen, R., Goldman, N. & Pedersen, A.-M. K. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, 155, 431-449.
- Yang, Z., Wong, W. S. W. & Nielsen, R. 2005. Bayes empirical bayes inference of amino acid sites under positive selection. *Molecular Biology and Evolution*, 22, 1107-1118.

Appendices

Appendix 1: Codeml control file for positive selection

```

seqfile      = /path/to/sequence.alignment/filename.fas * sequence data file name
treefile     = /path/to/tree.file/treefilename.re      * tree structure file name
outfile      = desired.file.name.for.output.mlc        * main result file name

noisy        = 3                                     * 0,1,2,3,9: how much rubbish on the screen
verbose      = 0                                     * 0: concise; 1: detailed; 2: too much
runmode      = 0                                     * 0: user tree

seqtype      = 1                                     * 1:codons; 2:AAs; 3:codons-->AAs
CodonFreq    = 2                                     * 0:1/61 each, 1:F1X4, 2:F3X4, 3:codon table
ndata        = 10
clock        = 0                                     * 0:no clock
aaDist       = 0                                     * 0:equal

aaRatefile   = /usr/lib/paml/data/dat/jones.dat      * only used for aa seqs with model=empirical(_F)
                                                    * dayhoff.dat, jones.dat, wag.dat, mtmam.dat, or your own

Model        = 0                                     * models for codons:
NSsites      = 0 1 2 7 8                             * 0:one w;1:neutral;2: positive selection; ;7:beta;8:beta&w

icode        = 0                                     * 0:universal code
Mgene        = 0                                     * codon: 0: rates

fix_kappa    = 0                                     * 1: kappa fixed, 0: kappa to be estimated
kappa        = 2                                     * initial or fixed kappa
fix_omega    = 0                                     * 1: omega or omega_1 fixed, 0: estimate
omega        = .4                                    * initial or fixed omega, for codons or codon-based AAs

fix_alpha    = 1                                     * 0: estimate gamma shape parameter; 1: fix it at alpha
alpha        = 0                                     * initial or fixed alpha, 0:infinity (constant rate)
Malpha       = 0                                     * different alphas for genes
ncatG        = 10                                    * # of categories in dG of NSsites models

getSE        = 0                                     * 0: don't want them
RateAncestor = 1                                     * (0,1,2): rates (alpha>0) or ancestral states (1 or 2)

Small_Diff   = .5e-6
Cleandata    = 1                                     * remove sites with ambiguity data (1:yes, 0:no)?
fix_blength  = -1                                    * -1: random
method       = 0                                     * Optimization method 0: simultaneous

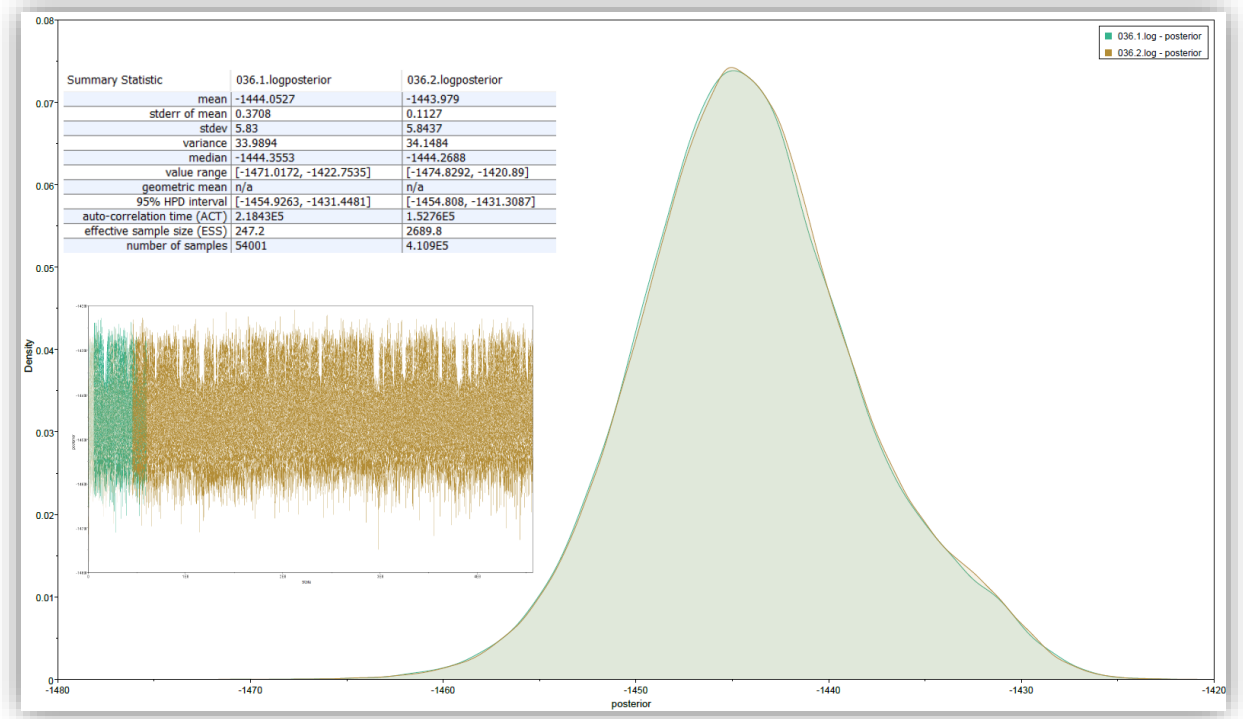
* Genetic codes: 0:universal
* These codes correspond to transl_table 1 to 11 of GENE BANK.

```

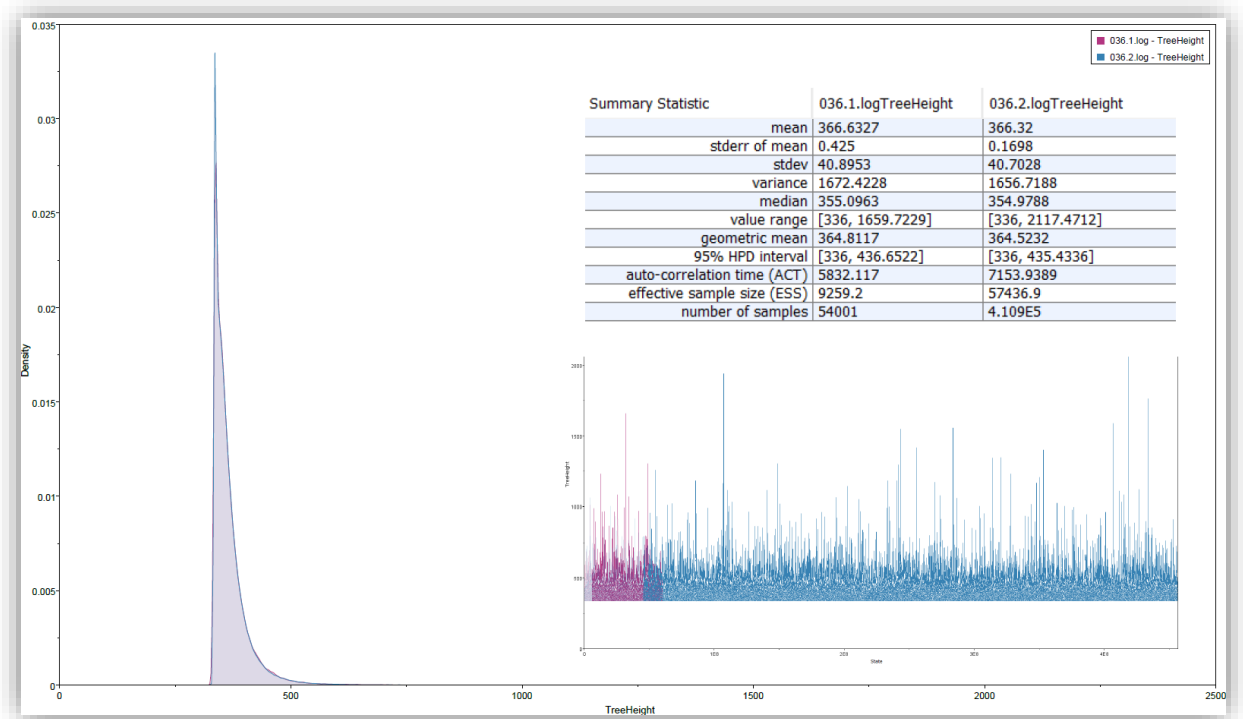
Appendix 2: Purifying selection sites identified by the REL method in Datamonkey

REL analysis results using F81					
Found 28 negatively selected sites (95 significance level)					
Codon	E[dS]	E[dN]	Normalized E[dN-dS]	Posterior Probability	Bayes Factor
7	2.46326	0.0288509	-2.43441	0.999361	117.858
18	4.51679	0.0287941	-4.48799	0.999971	2561.66
23	2.02875	0.0302932	-1.99845	0.999507	152.77
25	3.06369	0.0268972	-3.03679	0.999469	141.754
30	3.88802	0.0269068	-3.86111	0.999673	230.464
58	5.19986	0.0287919	-5.17107	0.999993	10839.5
75	2.87167	0.0306235	-2.84105	0.999282	104.81
88	5.11828	0.027185	-5.09109	0.999974	2944.74
95	4.54557	0.0274564	-4.51812	0.999941	1268.66
102	5.15827	0.0264304	-5.13184	0.999984	4774.75
111	5.10913	0.0328469	-5.07628	0.999992	9539.1
116	3.78384	0.0269007	-3.75694	0.99965	215.495
141	5.10767	0.029579	-5.07809	0.999964	2072.68
156	5.08579	0.0271984	-5.05859	0.999991	8237.15
165	3.25441	0.0302154	-3.22419	0.999624	200.361
188	4.19314	0.0295737	-4.16357	0.999557	169.927
208	2.44206	0.0314669	-2.41059	0.999885	657.823
223	2.55281	0.0262277	-2.52658	0.999426	131.303
244	4.7313	0.0288041	-4.7025	0.999918	917.865
248	4.83184	0.0299949	-4.80185	0.999999	130027
259	2.55785	0.0262281	-2.53162	0.999428	131.592
270	3.84606	0.0262107	-3.81985	0.99975	300.833
285	2.54257	0.0269335	-2.51563	0.999209	95.1662
297	3.82937	0.026536	-3.80283	0.9997	251.028
302	2.48577	0.0296386	-2.45614	0.999262	102.024
304	5.05199	0.0328673	-5.01912	0.999991	8376.37
309	3.24018	0.0321804	-3.208	0.9998	376.851
317	4.64792	0.0268754	-4.62105	0.999949	1473.87

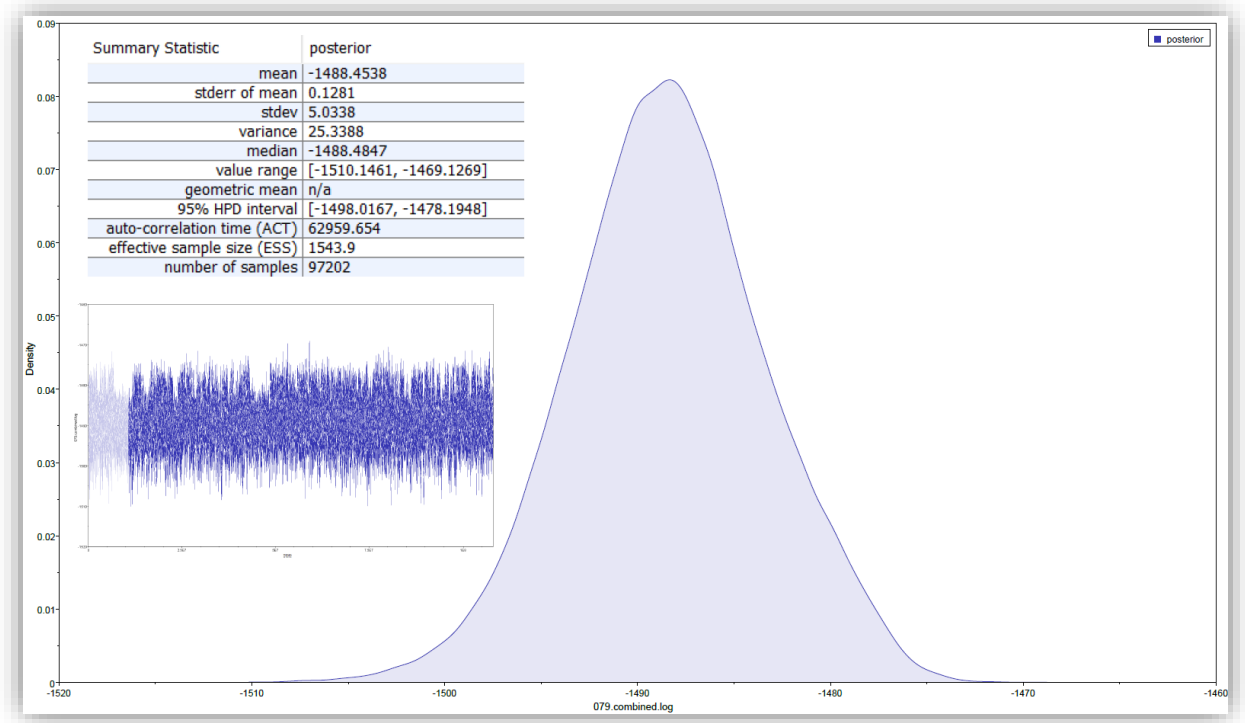
Appendix 3: Tracer of posterior statistic for 036



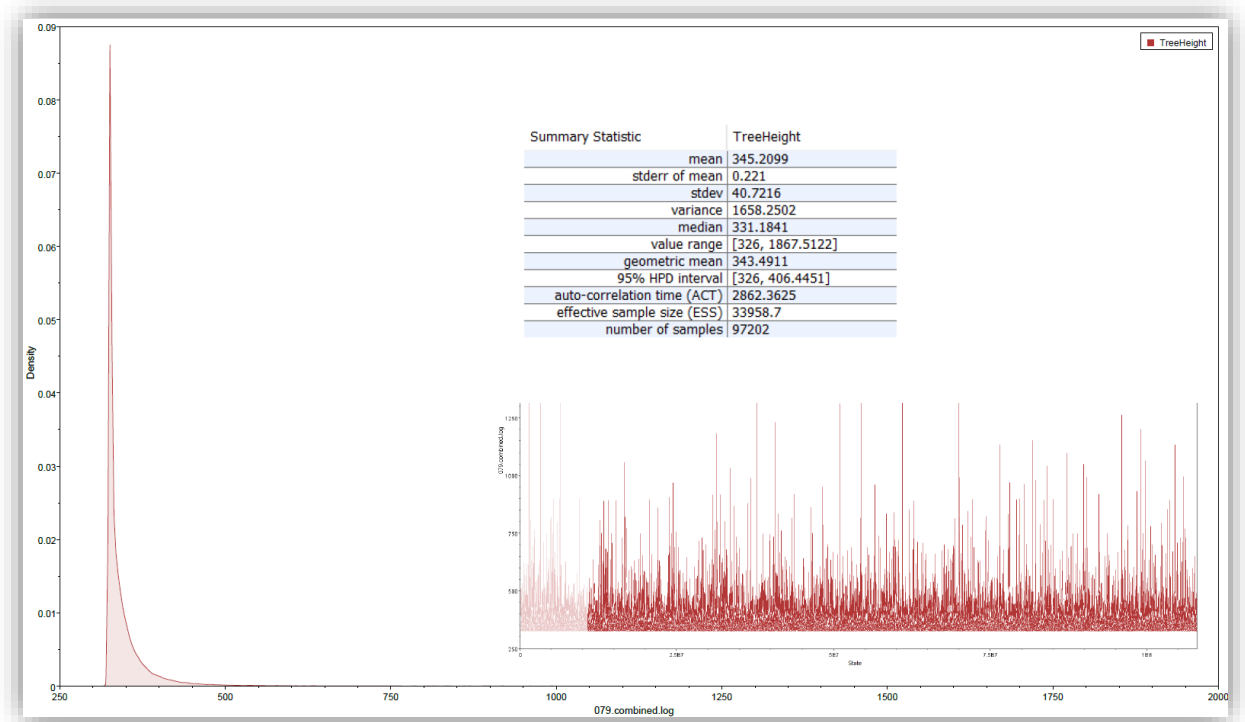
Appendix 4: Tracer of tree height statistic for 036



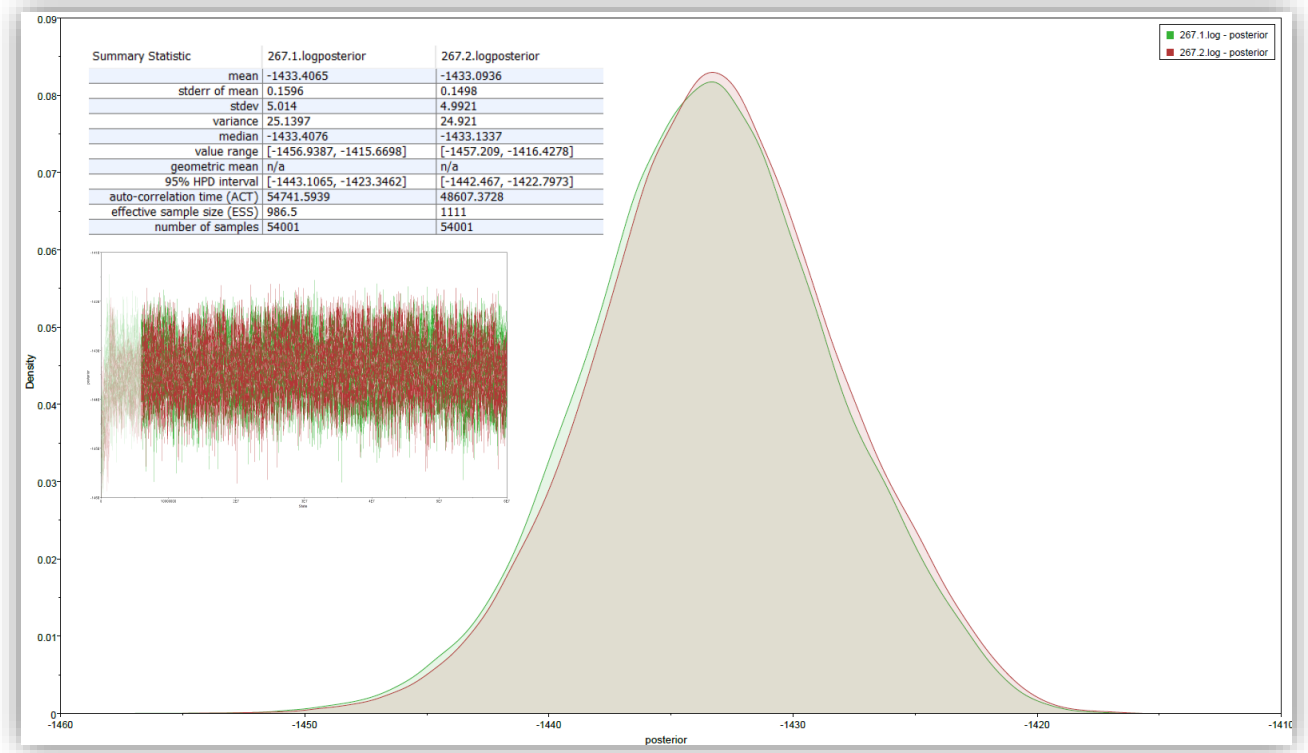
Appendix 5: Tracer of posterior statistic for 079



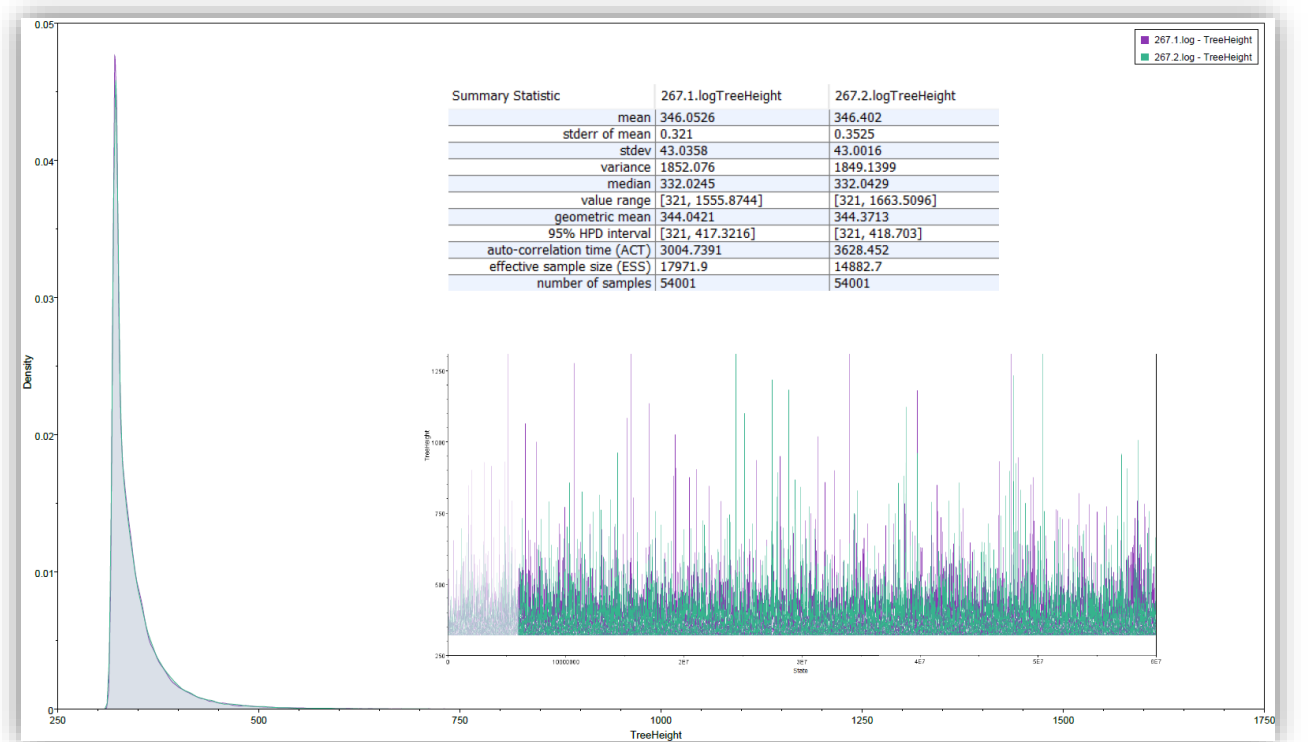
Appendix 6: Tracer of tree height statistic for 079



Appendix 7: Tracer of posterior statistic for 267



Appendix 8: Tracer of tree height statistic for 267



Appendix 9: Turnitin Report

Sequence analysis of an HIV-1 subtype C acutely infected cohort from Durban, South Africa

ORIGINALITY REPORT

7 %	8 %	5 %	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	w3.ualg.pt Internet Source	1 %
2	ir.canterbury.ac.nz Internet Source	1 %
3	mdpi.com Internet Source	1 %
4	docplayer.net Internet Source	1 %
5	datamonkey.org Internet Source	1 %
6	beast.community Internet Source	1 %
7	www.jci.org Internet Source	1 %
8	beast-mcmc.googlecode.com Internet Source	1 %