Evolution of HIV-1 Subtype C gp120 *envelope* Sequences in the Female Genital Tract and Blood Plasma during Acute and Chronic Infection

KAVISHA RAMDAYAL

UNIVERSITY OF THE WESTERN CAPE

Evolution of HIV-1 Subtype C gp120 *envelope* Sequences in the Female Genital Tract and Blood Plasma during Acute and Chronic Infection

Thesis presented in partial fulfilment of the requirements for the Degree of *Philosophiae Doctor in Bioinformatics* at the South African National Bioinformatics Institute, Faculty of Life Sciences, University of the Western Cape, South Africa

by

Kavisha Ramdayal

December 2014

Supervisor: Dr. Gordon Harkins, PhD Bioinformatics

Co-Supervisor: Prof. Alan Christoffels, PhD Bioinformatics

Dedication

This work is dedicated to my Mom, Dad, Brother and Sister, who have supported me through every step of this long and demanding journey, and to the five family members that I lost during the course of this PhD – one of whom died at the hands of AIDS.

"In the end, it's not the years in your life that count. It's the life in your years", Lincoln (nd).

Acknowledgements

Thanks to Prof. Lynn Morris and Dr. Bronwen Lambson for their provision and permission to use the HIV sequences analysed in this study. Thanks to Dr. Gordon Harkins and Prof. Alan Christoffels for their on-going support, patience and mentorship throughout the duration of this project and to my colleagues and friends at the South African National Bioinformatics Institute, University of Cape Town and Clickatell for their understanding, support and encouragement. I am especially grateful to my partner, Dr. Faisel Mosoval for his undying support and reassurance. There are no words to describe his role in helping me overcome the many obstacles I faced. Thank you.

Keywords

Bayesian
Blood plasma
Compartmentalization
Co-receptor
CVL
Env
Glycosylation
HIV-1
Longitudinal
Monotypic
Phylogenetics
Recombination
Slatkin-Maddison
tMRCA
Variable-loop

Abstract

Evolution of HIV-1 Subtype C gp120 *envelope* Sequences in the Female Genital Tract and Blood Plasma during Acute and Chronic Infection

K. Ramdayal

PhD Thesis, South African National Bioinformatics Institute, University of the Western Cape, South Africa

Heterosexual transmission of HIV-1 via the female genital tract is the leading route of HIV infection in sub-Saharan Africa. Viruses then traffic between the cervical compartment and blood ensuring pervasive infection. Previous studies have however reported the existence of genetically diverse viral populations in various tissue types, each evolving under separate selective pressures within a single individual, though it is still unclear how compartmentalization dynamics change over acute and chronic infection in the absence of ARVs. To better characterize intrahost evolution and the movement of viruses between different anatomical tissue types, statistical and phylogenetic methods were used to reconstruct temporal dynamics between blood plasma and cervico-vaginal lavage (CVL) derived HIV-1 subtype C gp120 *envelope* sequences. A total of 206 cervical and 253 blood plasma sequences obtained from four treatment naïve women enrolled in the CAPRISA Acute Infection study cohort in South Africa were evaluated for evidence of genotypic and phenotypic differences between viral populations from each tissue type up to 3.6 years post-infection. Evidence for tissue-specific differences in genetic diversity, V-loop length variation, codon-based selection, co-receptor usage, hypermutation, recombination and potential N-linked glycosylation (PNLG) site accumulation were investigated.

Of the four participants studied, two anonymously identified as CAP270 and CAP217 showed evidence of infection with a single HIV-1 variant, whereas CAP177 and CAP261 showed evidence of infection by more than one variant. As a result, genetic diversity, PNLGs accumulation and the number of detectable recombination events along the gp120 *env* region were lowest in the former patients and highest in the latter. Overall, genetic

diversity increased over the course of infection in all participants and correlated significantly with viral load measurements from the blood plasma in one of the four participants tested (i.e. CAP177).

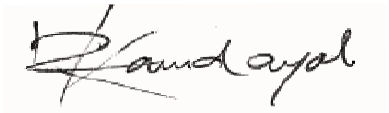
Employing a structured coalescent model approach, rates of viral migration between anatomical tissue types on time-measured genealogies were also estimated. No persistent evidence for the existence of separate viral populations in the cervix and blood plasma was found in any of the participants and instead, sequences generally clustered together by time point on Bayesian Maximum Clade Credibility (MCC) trees. Clades that were monophyletic by tissue type comprised mostly of low diversity or monotypic sequences from the same time point, consistent with bursts of viral replication. Tissue-specific monophyletic clades also generally contained few sequences and were interspersed among sequences from both tissue-types. Tree and distance-based statistical tests were employed to further evaluate the degree to which cervical and blood plasma viruses clustered together on Bayesian MCC trees using the Slatkin-Maddison (S-M), Simmonds Association index (AI), Monophyletic Clade (MC), Wright's measure of population subdivision (F_{ST}) and Hudson's Nearest Neighbour (Snn) statistics, in the presence and absence of monotypic and low diversity sequences. Statistical evidence for the presence of tissue-specific population structure disappeared or was greatly reduced after the removal of monotypic and low diversity sequences, except in CAP177 and CAP217, in 3/5 of longitudinal tree and distance-based tests.

Analysis of phenotypic differences between viral populations from the blood plasma and cervix revealed inconsistent tissue-specific patterns in genetic diversity, codon-based selection, co-receptor usage, hypermutation, recombination, V-loop length variation and PNLG site accumulation during acute and chronic infection among all participants. There is therefore no evidence to support the existence of distinct viral populations within the blood plasma and cervical compartments longitudinally, however slightly constrained populations may exist within the female genital tract at isolated time points, based on the statistical findings presented in this study.

02 December 2014

Declaration

I declare that "Evolution of HIV-1 Subtype C gp120 *envelope* Sequences in the Female Genital Tract and Blood Plasma during Acute and Chronic Infection" is my own work and that it has not been submitted for a degree or examination at any other university. All the resources I have used are quoted and all work that was the result of a joint effort has been indicated and acknowledged by complete references.



02 December 2014

Kavisha Ramdayal

Date

Contents

Abbreviations

aa	Amino acid
AI	Association Index
AIDS	Acquired Immunodeficiency Syndrome
APOBEC3	Apolipoprotein B RNA-editing Catalytic polypeptide-like 3
ART	Anti-retroviral Therapy
BaTS	Bayesian Tip-Significance
BEAST	Bayesian Evolutionary Analysis of Sampling Trees
C1-C5	Conserved regions 1-5
CAPRISA	Centre for the AIDS Programme of Research in South Africa
CCR5	Chemokine Co-Receptor 5
CD4	Cluster of Differentiation 4 Receptor
cDNA	Complementary-Deoxyribonucleic Acid
CI	Confidence Interval
CVL	Cervico-vaginal Lavage
CXCR4	Chemokine Co-Receptor 4
DNA	Deoxyribonucleic Acid
ESS	Effective Sampling Size
gp41, gp120	Glycoproteins 41 and 120
GTR+G	General Time-Reversible model + Gamma distribution
HIV	Human Immunodeficiency Virus
HPD	Highest Posterior Density
HPM	Hierarchical Phylogenetic Model
INDELs	Insertions and Deletions
NAbs	Neutralizing antibodies
NICD	National Institute for Communicable Diseases
MCC	Maximum Clade Credibility
MCMC	Markov chain Monte Carlo
MH	Metropolis-Hastings
PAUP	Phylogenetic Analysis Using Parsimony
PBS	Phosphate-buffered Saline

PCR	Polymerase Chain Reaction
p.i.	Post-infection
PNLGs	Potential N-linked Glycosylation Sites
PS	Parsimony Score
RNA	Ribonucleic Acid
STIs	Sexually Transmitted Infections
SGA	Single Genome Amplification
S-M	Slatkin-Maddison
SVM	Support-Vector Machines
tMRCA	Time to the Most Recent Common Ancestor
V1-V5	Variable loops 1-5
WebPSSM	Web Position Specific Scoring Matrix
XML	Extensible Markup Language

List of Figures

List of Tables

Definitions

AI	Association index is a statistic that assesses the population structure within a sample by weighting the contribution of each internal node based on how deep it is positioned on a phylogenetic tree (Wang *et al*., 2001; Gantt *et al*., 2010).
BF	Bayes factor is a theoretical Bayesian framework that is most often used for the comparison of multiple models, usually to determine which model better fits the data (Jeffreys, 1961; Drummond & Rambaut, 2007).
F_{ST}	The F_{ST} test is a distance-based statistic that assesses the correlation of randomly chosen alleles within a subpopulation relative to the total population (Wright, 1943; Weir & Cockerham, 1984; Holsinger & Weir, 2009).
Likelihood	The likelihood of the data is the probability of the observed genotypes based on the specified model parameters (Wilson & Rannala, 2003).
Mann-Whitney U test	The non-parametric Mann-Whitney U test compares the distributions of two unmatched populations and estimates if one of two populations is randomly larger than the other (Mann & Whitney, 1947).
MC	Maximum (single-state) clade size is a statistic that quantifies the observation that stronger phylogeny-trait associations should produce larger monophyletic clades that share a single trait (Gantt *et al*., 2010).

MCC tree	A maximum clade credibility tree is the phylogenetic tree within a posterior distribution of trees with the highest product of clade posterior probabilities that is generated using Bayesian phylogenetic inference methods (Presti, 2010; Dictionary 3.0, 2011).
MCMC	Markov chain Monte Carlo methods are a class of algorithms that sample from a probability distribution based on a Markov chain construction (Robert & Casella, 2004).
Outliers	Values in a set of data that are so extreme relative to the other data in the sample that they appear to not form part of the sampled data set (Zar, 1999).
PS	Parsimony score, also known as the Slatkin-Maddison test, reconstructs character states at ancestral nodes in a tree or posterior distribution of trees using a parsimony approach, where the significance of the observed number of state changes within a phylogeny or phylogenies are calculated and compared to a null distribution of PS statistics for randomized phylogenies (Slatkin & Maddison, 1989; Parker *et al*., 2008; Gantt *et al*., 2010).

Research Motivation

The overeaching goal of this study was to determine if HIV-1 evolves separately within different tissue types in patients that have not been exposed to antiretroviral therapy, over long-term HIV-1 infection. The biological and clinical implications of independently evolving viruses influences three important disease management issues including mother-to-child transmission, female-to-male transmission and targeted therapeutic, prophylactic or eradication strategies. Therefore defining anatomical compartments, in which independent viral evolution may be occurring due to compartmentalization of viral populations, is crucial and an increasing amount of work has gone into this area of research.

To investigate if viral compartmentalization did exist during acute and chronic infection, the aims of this study were as follows:

i. Determine whether genotypic or phenotypic differences exist between viral populations in the cervix and blood plasma.
ii. Statistically assess the degree of viral compartmentalization between the sampled tissue types.
iii. Statistically assess the flow of viruses between the cervix and blood plasma in each participant.
iv. Statistically evaluate if genotypic patterns correlate with phenotypic traits such as viral loads, CD4 cell counts or other clinical indicators associated with disease progression.

Based on the above aims, the following objectives were undertaken:

i. Assess if viral evolution was compartmentalized by tissue type through phylogenetic and statistical analyses.
ii. Determine the amount of homologous recombination that has occurred between viruses sampled in the cervix and blood plasma for each participant, and characterize any detectable hypermutation patterns.
iii. Estimate the most likely number of viral variants that each participant was infected with and the degree to which viral populations were compartmentalized by tissue type between participants infected by single and multiple transmitted variants.

iv. Estimate the direction, timing and relative rates of viral migration between the cervix and blood plasma.

v. Statistically evaluate whether viral populations in the cervix and blood plasma demonstrated any evidence of phenotypic differences, i.e. in genetic diversity, variable-loop lengths, co-receptor usage, potential N-linked glycosylation sites or selection pressures.

Chapter **1**
Introduction

The HIV/AIDS epidemic is a crisis of enormous spiritual, social, economic and political proportions that may be the most devastating health disaster in human history, (UNICEF, 2003; Lamptey *et al*., 2006; Afonso *et al*., 2012). In sub-Saharan Africa, the area most severely affected by the epidemic, 25 million people were reported to be living with HIV/AIDS from the total of 35.3 million infected globally at the end of 2012 (Rossenkhan *et al*., 2012; UNAIDS, 2013). South Africa accounts for 6.4 million of these cases where there is an estimated 14.4% HIV prevalence in women and 9.9% prevalence in men (Shisana *et al*., 2014). Although the incidence of HIV infections have been on the decline in recent years (UNAIDS, 2012), young women in this region between the ages of 15 and 34 remain on average, approximately eight times more likely to be HIV positive than men in the same age group (Shisana *et al*., 2014). The majority of infections in this region are caused by HIV-1 subtype C viruses that have been transmitted heterosexually through the genital mucosa (Karim *et al*., 2010; Buonaguro *et al*., 2007; van Harmelen *et al*., 1997; Kemal *et al*., 2003). Globally, approximately 90% of HIV-1 infections are transmitted this way since the genital mucosa is the initial point of contact for the majority of exposed individuals (Kemal *et al*., 2003; Hladik & Hope, 2009).

The female genital tract therefore plays a central role in both sexual and perinatal transmission of HIV and is generally the source of systemic infection in women, however the complex biological processes that influence transmission dynamics and the persistence of HIV infection within this anatomical compartment remain incompletely understood (Karim *et al*., 2010; Kaushic, 2010). For example, since the genital tract plays a central role in mother-to-child and female-to-male sexual transmission of HIV, it is important to understand if it serves as a genetically distinct compartment from the blood plasma and if so, what the clinical and biological implications of this phenomenon could be (Bull *et al*., 2013). Compartmentalization has also been reported to exist at a cellular level among $CD4^+$ and $CD8^+$ T cells between different tissue types (Potter *et al*., 2006; Delobel *et al*., 2005; Fulcher *et al*., 2004; Heeregrave *et al*., 2009), suggesting the presence of differential immunological pressures between anatomical compartments.

Moreover, since the female genital immune system is responsible for protection of the reproductive system from pathogens and being tolerant to antigens such as sperm and embryo's, specialized cells and antibodies responsible for this type of protection are required and have been reported to challenge HIV-1 viruses too (Hirbod & Broliden, 2007). Effective disease management is therefore greatly dependent on our knowledge of whether or not unique viruses are developing and persisting within different tissues, which in turn informs microbicide and vaccine development strategies that should ideally target viral strains in all tissue types equally (Bull *et al*., 2013; Hladik & Hope, 2009; Andreolétti *et al*., 2007).

This study consequently focuses on the investigation of HIV-1 subtype C viral evolution in female participants longitudinally (up to 3.6 years post-infection), encompassing both acute and chronic stages of infection. A key aim of this study is to contribute towards our understanding of comparative HIV evolutionary dynamics in the female genital tract and blood stream over a lengthier period, compared to a cross-sectional view of these characteristics. Importantly all four participants studied here were not on any anti-retroviral medications throughout the duration of this study, which reduced a large possibility of drug-resistant mutations arising in this particular data set. Given that selection mechanisms likely vary at least slightly between different tissues and cell types, the possibility that these differences might yield genetically distinct viral sub-populations in different anatomical compartments was investigated.

Compartmentalization of viruses by tissue type is thought to arise from gradual diversification of the transmitted initial viral variant/s under differential tissue-specific selection pressures resulting in separate populations with distinct phenotypic characteristics (Sturdevant *et al*., 2012; Fulcher *et al*., 2004; Si-Mohamed *et al*., 2000). However distinct viral populations can only exist if there is no or limited mixing of viruses between anatomical compartments, which could only be possible in the presence of a barrier between the genital tract and blood stream, restricting the migration of viruses between these tissues. The mucosal linings of the female genital tract have been described as this sort of barrier; composed of squamous epithelial cells, these linings form a protective layer that serve to prevent HIV transmission from occurring (Hladik & Hope, 2009; Hirbod & Broliden, 2007).

To investigate this hypothesis, an additional aim of this study was to estimate the degree of viral movements longitudinally between tissue types, to evaluate evidence for the existence of such a barrier. Simply quantifying the relative rates at which viruses are migrating to and from the genital tract in different individuals would be a major step towards understanding patterns of HIV transmission *in vivo*. Framing these dispersal dynamics within the context of viral evolution during acute and chronic stages of infection could also provide additional potentially valuable information on the exact sites at which (and possibly even likely situations under which) important events leading to the loss of HIV-1 immunological control occurs.

Previous studies on compartmentalization in the female genital tract have reported strong evidence of compartmentalization between the genital tract and blood plasma, however many of these have been cross-sectional in design, considering viruses sampled at a single time point only (Kemal *et al*., 2003; Bull *et al*., 2009; Andreolétti *et al*., 2007; De Pasquale *et al*., 2003; Philpott *et al*., 2005; Tirado *et al*., 2005; Poss *et al*., 1998; Uvin & Caliendo, 1997). There have also been studies that found no evidence of distinct viral populations between the genital tract and blood plasma (Chomont *et al*., 2007). More recently, in a longitudinal study by Bull *et al*. (2013) that compared plasma RNA, PBMC DNA, genital RNA and DNA of HIV-1 found no statistical evidence for compartmentalization between these tissues and no evidence of tissue-specific differences in amino acid residues, co-receptor usage or immune selection signatures. Furthermore, it was shown that the presence of monotypic and low diversity variants inflated statistical measures of compartmentalization that when corrected for, significantly reduced the statistical support for evidence of distinct viral populations in these tissues (Bull *et al*., 2013). Rapid population turnover that results in relatively low intra-patient viral genetic diversity is a commonly observed pattern over the course of HIV-1 infection, with sequence variants from latter stage replicating viruses often rising to predominance and obscuring previously selected viral variants (Bull *et al*., 2009). As a consequence cross-sectional studies that offer only a momentary snapshot of sequence diversity through time (usually during the acute stage) are often not representative of the entire course of infection. To address this drawback, longitudinal studies that sample during acute and chronic stages of infection are necessary to reliably resolve the question of whether distinct viral populations do actually exist in the female genital tract of those infected.

1.1 The human immunodeficiency virus

1.1.1 History and epidemiology of HIV

The HIV/AIDS pandemic is one of the most highly researched and globally funded socioeconomic, medical and governance affecting issues, with an impact on a large range of disciplines, including those of the natural sciences, economics, social sciences and medicine (Shaeffer, 1994; Over, 2001; Tawfik & Kinoti, 2006). The subject of HIV is a topic with more questions than answers, many uncorroborated theories and interests dating as far back as the late 1970's (Gupta *et al*., 2003; Seiter *et al*., 2011; Lihana *et al*., 2012; Mousseau & Valente, 2012; Pantaleo, 2000). Although the aquired immune deficiency syndrome (AIDS) was first noted in communities of gay men, due to the presence of Kaposi's sarcoma (a type of cancer) and *pneumocystis carinii pneumonia* (PCP), common in old people with compromised immune systems (Gupta *et al*., 2003; Seiter *et al*., 2011; SADC, 2013), and subsequently observed in heterosexual networks, drug users and mother-to-child transmissions, it was only in 1984 that HIV (initially known as HTLV-III) infection was discovered to be the result of a retrovirus capable of destroying lymphocytes (Gallo & Montagnier, 2003; Gupta *et al*., 2003). Various theories on the origin of HIV exist, including the *OPV theory*, which proposed that HIV transmission occurred through contamination of oral polio vaccines that were cultivated from chimpanzee tissues (Worobey *et al.,* 2004), however this theory was refuted shortly after its release (Garrett, 1995; Karlen, 1995). HIV's origins have been traced to zoonotic or cross-species transmission of SIV between humans and African primates, with studies in 1994 and 1998 reporting the presence of HIV-1 in blood samples dating back to 1959 and the early 1960's respectively (Sharp & Hahn, 2011; Requejo, 2006; Buonaguro *et al*., 2007).

The globally accepted use of penicillin and other antibiotics for treatment of sexual diseases (Lihana *et al*., 2012), coupled with a significant increase in the migration of people resulted in the formation of riskier sexual networks (Lihana *et al*., 2012; Rambaut *et al*., 2004; Pantaleo, 2000), allowing the virus to spread in humans through three prime transmission routes i.e. sexual, parenteral (blood-borne) and perinatal, creating many separate epidemics in distinct geographic origins (Gupta *et al*., 2003; SADC, 2013; Lihana *et al*., 2012; Rambaut *et al*., 2004; Pantaleo, 2000).

Many different types and frequencies of risk behaviors and practices such as unprotected sex with multiple partners or sharing drug injection equipment followed suite (Seiter *et al*., 2011; Engelman & Cherepanov, 2012; Mousseau & Valente, 2012; Rambaut *et al*., 2004; Pantaleo, 2000). Individual factors such as biological, demographic and behavioural risk profiles also played a role in influencing HIV acquisition and spread. Social networks compounded the spread of the disease, adding to the diverse challenges that hindered interventions in epidemic expansion (Rothenburg *et al*., 2001).

Recent studies estimate at least 35 million people are currently living with HIV/AIDS worldwide, 25 million of whom reside in sub-Saharan Africa (UNAIDS, 2013), yet it is this very continent that is still lacking the infrastructure required to effectively handle this pandemic (Visser, 2004; Lihana *et al*., 2012). On the African continent, South Africa is among the countries with the highest HIV prevalence rates in the world, however rates of citizen testing remains relatively low despite knowledge of testing services being high (Luseno & Wechsberg, 2009). In 2007, UNAIDS estimated 33 million people living with HIV/AIDS, where more than 96% of new HIV infections occur in middle and poorer income countries (Amo *et al*., 2010). Prevalence statistics have since increased to approximately 25.2 million in sub-Saharan Africa, where approximately one fifth of South African women in reproductive age groups (i.e. between 15 and 49) are infected with HIV (UNAIDS, 2012; Shisana *et al*., 2014). In the general population, women are reported to have an increased risk of HIV transmission, which is negatively influenced by the presence of sexually transmitted diseases (STDs) (N'Galy *et al*., 1998; Broliden, 2010).

1.1.2 HIV among women

Early in the epidemic few women were diagnosed with HIV/AIDS, as many infected women that contracted HIV through injection drug use were not accurately diagnosed (CDC, 2008). At the end of 2002 more than 42 million people were living with HIV/AIDS (UNICEF, 2003) and by 2004 HIV infection was the fifth and sixth leading cause of death among women between 35 – 44 and 25 – 34 years respectively (CDC, 2008). The number of HIV infected women has been steadily rising on a global scale since then (National Institutes of Health, 2008) with more than 90% of HIV infections having occurred through heterosexual intercourse, a mode of transmission that women are more vulnerable to because of the substantial mucosal exposure to seminal fluids (National Institutes of Health, 2008).

Pregnancy and childbirth further compound health risks faced by women, which require careful and consistent health care (World Health Organization, 2009). Furthermore, women have been at greater risk due to discrimination through inadequate education, poor pay, violence, abuse and exploitation by men (UNICEF, 2003; World Health Organization, 2009; UNFPA, 2013) making them more susceptible to HIV infection than men in the longer term. Men, both young and old with multiple sexual partners have also compounded the problem by having sexual relations with younger women, posing a greater risk to them than to older women (UNICEF, 2003).

Although there have been several efforts in the past to promote discussion, strengthen social values and provide support, HIV prevalence and mortality statistics have continued to grow. Staggeringly, a recent study has reported that on average 50 young women are newly infected with HIV every hour (UNAIDS, 2013). The report went on to state that from the 50 countries that participated in the study, between 9 and 60% of women aged between 15 and 49 years had experienced violence in an intimate relationship within the last year, and these women were 50% more likely to acquire HIV than women who had not experienced violence (UNAIDS, 2013). A high prevalence of rape, physical violence and other forms of abuse were reported in several different countries frequently among sex workers, generally however women in "conflict-affected situations" tended to be particularly more vulnerable to sexual abuse (UNAIDS, 2013).

1.1.3 HIV classification and distribution

HIV-1 belongs to the Lentivirus genus of the *Retroviridae* family, and is classified into four groups, i.e. M, N, O and P, with M being the main source of the global AIDS pandemic (Requejo, 2006; Abecasis *et al*., 2013). HIV-2 contains eight subtypes (A to H), however only subtypes A and B are linked to the HIV epidemic seen mainly in West Africa (Marx *et al*., 2001; Santiago *et al*., 2005). HIV subtypes make it possible to track the course of the epidemic through the identification of specific markers in samples, or by detecting specific antibodies that bind to the virus (Requejo, 2006).

There has been an unprecedented and non-uniform increase in HIV-1 subtype C prevalence historically compared to all other subtypes (Novitsky *et al*., 2002). Subtype C is the most prevalent globally and represents approximately 52% of infections worldwide (Jacobs *et al*., 2014). Predominantly found in Southern Africa, India and China, subtype C is the only subtype that accounts for prevalence rates as high as 20 to 40% in the general HIV positive population (Novitsky *et al*., 2002), while subtype B is the main genetic form in the northern hemisphere (Abecasis *et al*., 2013; Buonaguro *et al*., 2007).

The geographic spread of HIV-1 subtypes C and B (considered to be the main cause of the HIV epidemic) is crucial for epidemiological HIV vaccine design and testing strategies, however viral diversity poses a major problem in this area as it significantly affects the "specificity and/or sensitivity of serological and molecular" tests (Buonaguro *et al*., 2007).

Although multiple factors play a role in the rate of disease progression, patients infected with HIV have been classified into three main types of progressors (Kumar, 2013):

- Rapid progressors – AIDS develops within 3 years after seroconversion.
- Intermediate progressors – AIDS develops gradually between approximately 3 to 10 years after seroconversion.
- Long-term non-progressors – HIV infected individuals retain high $CD4^+$ and $CD8^+$ cell counts and remain off antiretroviral therapy until $CD4^+$ cell counts decline, which can take more than 20 years.

About 5 to 10% of HIV positive people are rapid progressors who display high levels of viral replication and a steep decline in $CD4^+$ cell counts within two years of the initial HIV infection event, leading to the development of AIDS (Geubbels & Bowie, 2006). Another 5 to 10% that are classified as long-term non-progressors are able to effectively control HIV, sometimes maintaining low viral loads and a competent immune system for more than 20 years (Geubbels & Bowie, 2006), while the remaining proportion of HIV positive population progress from HIV to AIDS in approximately 8 to 10 years, dependent on socio-economic factors such as malnutrition or access to health care for example.

1.1.4 HIV cell structure and replicative cycle

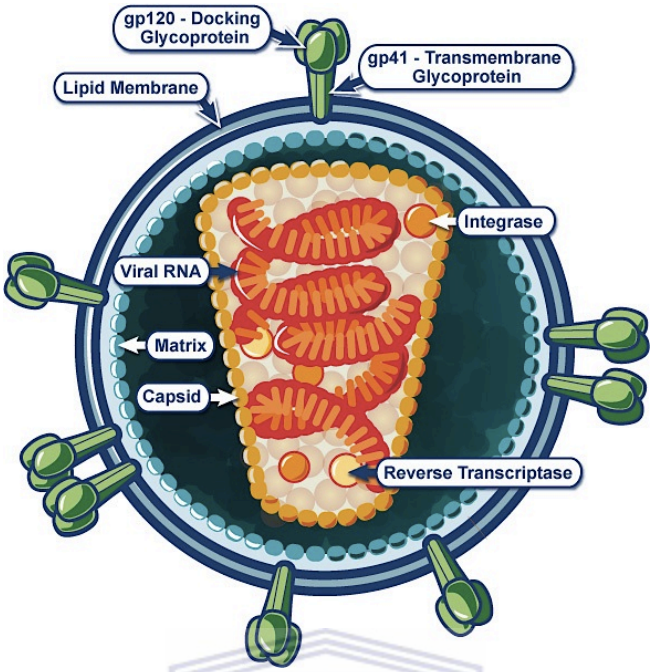


Figure 1.1 An overview of key components making up the HIV virion (adapted from NIAID, 2012). HIV is composed of a capsid (core) that encapsulates two single stranded RNA molecules, three structural genes (*gag*, *pol* & *env*), six regulatory genes (*tat, rev, nef, vif, vpr & vpu*), three enzymes (reverse transcriptase, integrase & protease) and viral proteins (p24 & p17) enclosed by a lipid membrane containing several glycoproteins from both the host cell and the virus (NIAID, 2012; Arora & Seth, 2003; Requejo, 2006).

Each of these components plays a role in the infection process eventually forcing host cells into manufacturing new copies of the virus. The viral envelope, is composed of two layers of fatty molecules known as lipids that are taken from the membrane of previously infected human CD4 cells when newly formed viruses bud from the host cell (NIAID, 2012). On average there are approximately 72 copies of a complex HIV protein known as *Env* embedded throughout the viral envelope that protrude through the surface of the the virus that consists of two glycoproteins, gp120 and gp41 (collectively referred to as the gp160 region), both of which have been the focal point of extensive vaccine development research (NIAID, 2012).

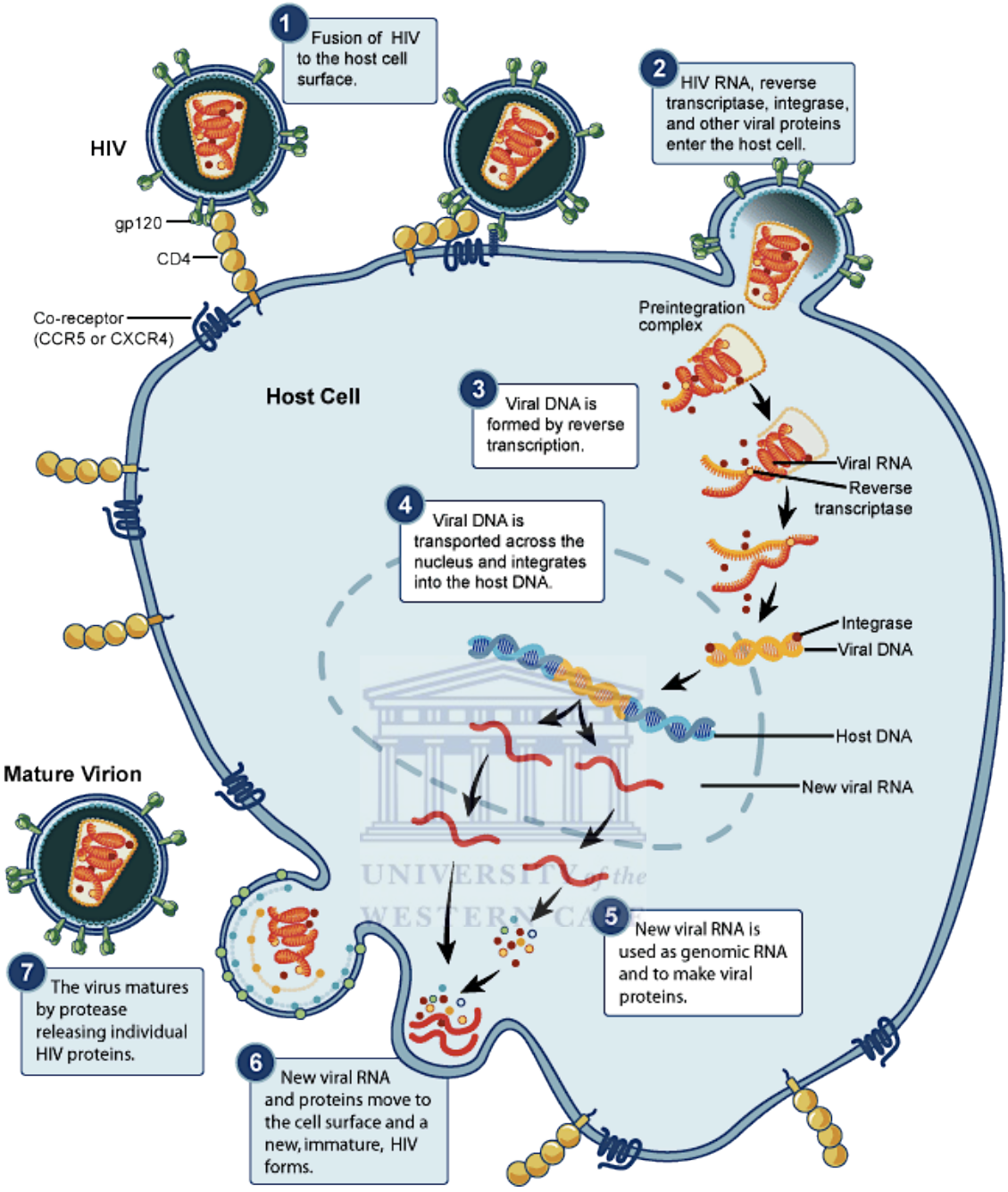


Figure 1.2 Overview of the stages, enzymes and proteins involved in the HIV replicative cycle (adapted from NIAID, 2012). HIV replication takes place through seven basic steps as follows (Seiter *et al*., 2011; Engelman & Cherepanov, 2012):

1. **Binding** – The gp120 glycoprotein on the HIV envelope binds to a $CD4^+$ receptor and one co-receptor (either CCR5 or CXCR4) on the surface of a $CD4^+$ cell.
2. **Fusion** – The virus then fuses to the host $CD4^+$ cell and releases its RNA into the $CD4^+$ cell.

3. **Reverse transcription** – The reverse transcriptase enzyme converts the HIV single-stranded RNA to a double-stranded DNA particle, also known as a pre-integration complex (PIC).
4. **Integration** – The double-stranded DNA produced by the reverse transcriptase enzyme enters the $CD4^+$ cell's nucleus. PIC-associated integrase and host chromatin-binding protein lens epithelium-derived growth factor (LEDGF) then assists in the integration of the double-stranded DNA into the host cell's DNA, forming a provirus.
5. **Transcription** – When the host cell ($CD4^+$) receives a signal to become active, the provirus then uses RNA polymerase II (RNA Pol II), a host enzyme and positive transcription elongation factor b (P-TEFb) to create copies of HIV's genomic material and different sizes of RNA strands known as messenger RNA (mRNA), which are used to create chains of HIV proteins.
6. **Assembly and budding** – In the assembly stage, an HIV enzyme called protease cuts the long chains of HIV proteins into smaller individual proteins which combine with copies of HIV RNA to form new HIV particles. Newly assembled (immature) viruses then push their way out of the host cell in a process known as budding, mediated by ESCRT (endosomal sorting complex required for transport) complexes. These new viruses are encapsulated in part of the host cell's outer envelope and are studded with glycoproteins (i.e. protein or sugar combinations), which allow the virus to bind to $CD4^+$ receptors and co-receptors later on in the reproductive cycle.
7. **Maturation** – In the final stage of replication, the protease enzyme that played a role in step 5 of the replicative process, completes cutting of HIV protein chains into individual proteins that combine to make a new mature/virulent HIV particle.

This basic process ensures the ongoing persistence of HIV even in the presence of antiretroviral therapies (Sigal *et al*., 2011), which has been proposed to occur through two main routes, i.e. infection by *cell-free* virions or through direct *cell-to-cell* transmission (Sattentau, 2008, Monel *et al*., 2012). The replication process explained above describes the cell-to-cell mode of transmission, which is known to be more rapid and helps the virus avoid "several biophysical, kinetic and immunological barriers" within the host (Carr *et al*., 1995; Chen *et al*., 2007; Dimitrov *et al*., 1993; Sourisseau *et al*., 2007).

Alternatively, infection by *cell-free* virions has been proposed to occur by viruses that may have escaped from CD4 cells through pores that arise along the plasma membrane, instead of through extracellular budding (Monel *et al*., 2012; Sato *et al*., 1997), however this mode of replication is yet to be experimentally verified.

1.1.5 HIV *envelope* region

HIV *envelope* proteins have a significant role to play in how the immune system responds to invading immunogens and the production of infectious viral particles (Kantanen *et al*., 1995). More specifically, the gp160 envelope protein plays a fundamental role in the infection process, as it is responsible for HIV binding to the host cell's surface (Land *et al*., 2003).

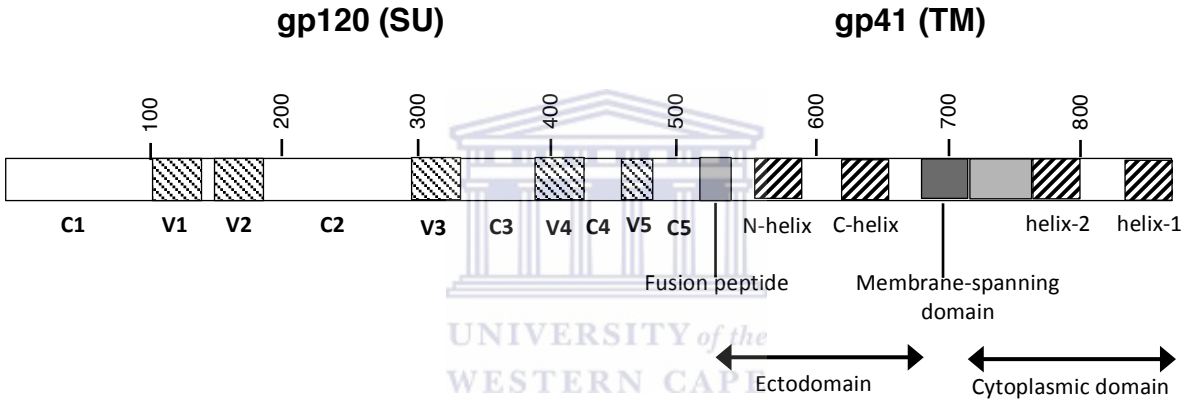


Figure 1.3 Genetic sub-regions within the HIV gp160 *envelope* protein (Ganser-Pornillos *et al*., 2008). The gp160 region consists of two non-covalent sub-units: (1) gp120 (SU), that encodes for a surface protein which is soluble and binds to the chemokine receptor as well as the CD4 receptor on the host cell, and (2) gp41 (TM), that encodes for a transmembrane protein which anchors the glycoprotein to the viral membrane and mediates cell fusion (Land & Braakman, 2001). The surface protein gp120 (SU) assists the virus in attaching to the host cell by binding to the primary receptor (CD4), while forming a structural arrangement that creates a "high affinity binding site" for a chemokine co-receptor such as CCR5 or CXCR4 (Miyauchi *et al*., 2009). The SU protein therefore, determines the tropism and specificity of the virus for a single receptor molecule (Coffin *et al*., 1997).

The gp120 region includes five relatively conserved regions (C1-C5) that make up the core of the gp120 protein, and five hypervariable regions (V1-V5) that form flexible loop structures on the surface of the gp120 protein (Bunnik, 2010).

The V3-loop is the primary determinant of co-receptor tropism, while the V1V2-loops have been associated with the accumulation of potential N-linked glycosylation sites (PNLGs) during chronic infection, which directly contributes towards viral resistance and evasion from HIV-specific neutralizing antibodies (NAbs) (Yuan *et al*., 2013; Malherbe *et al*., 2013; van den Kerkhof *et al*., 2013; Go *et al*., 2009; Blay *et al*., 2006; Frost *et al*., 2005).

1.2 HIV evolution and viral fitness

1.2.1 Asparagine-linked glycosylation

Eukaryotic cells contain enzymes that catalyse the construction of glycoproteins on the cells' surface and assist in cellular processes including immune response and intercellular recognition (Connor & Imperiali, 1996). These glycoproteins are transported to the Golgi complex where they undergo progressive modification at the terminal carbohydrates (Dean, 1999). In the HIV *envelope* glycoproteins, both Asparagine (N-linked) and Oxygen (O-linked) glycosylations occur (Agnihotri, 2008). N-linked glycosylations occur as the peptide sequence is synthesised from mRNA on the ribosomes, and insertion into the endosplasmic reticulum is directed by the hydrophobic amino terminal sequence of the polypeptide chain (Andeweg *et al*., 1995). The glycans specifically form a complex that mediates receptor and co-receptor binding and the membrane fusion events that permit virus entry into host cells (Mascola & Montefiori, 2003).

Potential N-linked glycosylation sites (PNLGs), also known as sequons, require the context of the amino acid pattern N-X-[S or T] where X can be any amino acid, followed by a Serine (S) or Threonine (T) (Zhang *et al*., 2004). A sequon will not be glycosylated if it contains or is followed by a Proline, and glycosylation may be inhibited by certain combinations of N-X-S or when followed by specific amino acids (Zhang *et al*., 2004). The proteins that are targeted for N-linked glycosylation on the HIV-1 virus structure are the gp120 proteins, of which N-linked glycans constitute 50% (Hart *et al*., 2003). This high concentration of N-linked glycans blocks antibody access to the conserved regions of gp120 making them a common focal point for studies investigating anti-viral strategies.

In 1992, Lee *et al*. investigated 24 N-linked glycosylation sites along the gp120 region of HXB2 to determine the relative importance of each site with respect to viral infectivity. Of the 24 sites studied, the majority were found to be "dispensable" or unimportant in terms of viral infectivity, whereas five sites that were linked to infectivity, were found to be located on the amino-terminal half of the *env* gp120 region leading to the conclusion that PNLG sites that infer infectivity are actually not randomly distributed along the gp120 region (Lee *et al*., 1992).

In a site-directed mutagenesis study by Ogert *et al*. (2001) in which specific N-linked glycosylation sites were compromised and subsequently evaluated in the context of viral infectivity, PNLG sites that were found to be "functionally critical" in the replicative process included N135, N141, N156, N160 and N301. Glycosylation sites near the V1V2-loops (i.e. N135, N141, N156, N160) were found to compromise the use of both CCR5 and CXCR4 co-receptor usage, whereas a mutation at N301 resulted in a total loss of CCR5 binding activity while the use of CXCR4 remained at 50% (Ogert *et al*., 2001). In a similar approach Li *et al*. (2008) discovered three sites (i.e. N406, N448 and N463) that were of particular importance in the HIV replicative cycle with the loss of a glycan at PNLG site N448, resulting in a structural change within the C4 region, making it more resistant to proteolytic cleavage. Even though the distant C1C2 sites were not affected by the loss of a glycan, the highly localised effect still reduced CD4 cell recognition and therefore represented an important finding with respect to the development of potential future biological therapeutic agents (Li *et al*., 2008).

The findings of Li *et al*. (2008) were consistent with those of François and Balzarini (2011), who observed that under site-directed mutagenesis, compromise of the highly conserved N260 site caused a significantly lower expression of gp120 and gp41 in the virus particle, leading to the recommendation that this site is a suitable target for therapeutic development. Previous studies had observed variation in the net number of sequons in the variable loop regions associated with patterns of co-receptor usage, and showed that a particular sequon in the V3-loop was highly conserved in CCR5-tropic viruses whereas other PNLG sites in the V1V2-loop appear to influence the use of the CXCR4 co-receptor usage (Zhang *et al*., 2004).

In another study by Toma *et al*. (2011), it was shown that the loss of N-linked glycosylation sites in the V5-loop region was associated with the CD4 antibody Ibalizumab, which was introduced to patients on failing drug regimens. Cells with fewer N-linked glycosylation sites in the V5-loop were found to be less susceptible to the antibody while those with more N-linked sites were more susceptible, indicating that the glycosylation sites in the V5-loop may further act as target sites for therapeutics.

Moreover, Balzarini *et al*. (2004) observed that the removal of several glycosylation sites along the gp120 region resulted in a more infectious virus compared with the parental wild-type, and that mutation of the glycosylation site in the V3-loop (N301) can lead to a co-receptor switch as a result of a higher affinity for the CXCR4 co-receptor. Thus, demonstrating that targeting the glycosylation sites may not necessarily result in a reduction in HIV-1 activity, highlighting the importance of long-term studies of glycosylation site patterns to infer their effect on HIV-1 activity.

In a lipoprotein-specific study that investigated two asparagine-linked glycosylation sites at the amino and carboxy terminals of lipoprotein lipase and hepatic lipase (Ben-Zeev *et al*., 1994), it was observed that glycosylation at the C-terminal domain is not essential for the expression of active lipases, but was imperative for active lipase secretion at the conserved N-linked glycosylation sites for these enzymes. What this indicates is that the conserved regions may contain N-linked glycosylation sites that can actually be compromised by targeting the enzymes associated with these sites. Furthermore, insertions at N-linked glycosylation sites and the resultant observations have become a growing area of research. One such study reported an insertion of arginine at amino acid site 11 on the gp120 region, which resulted in a loss of the N-linked glycosylation site responsible for acquiring CXCR4-tropism (Tsuchiya *et al*., 2013).

Glycosylation analysis however, is becoming less feasible to perform through wet-bench analysis due to cost and time limitations, and since there is such a high volume of sequence data that is presently available, this has led to the development of several bioinformatics methods and tools for *in silico* glycosylation analysis on the gp120 region (Poon *et al*., 2007; Shaw & Zhang, 2013). Studies that employed these and other methods of analysing glycosylation sites have led to the hypothesis that the adaptive repertoire of the HIV-1 glycan shield is limited by functional interactions between the N-linked glycans (Poon *et al*., 2007). However, all algorithms have their respective strengths and weaknesses in their accuracy of predicting sites for viral targeting and therefore longitudinal immune pressures that may affect sequon accumulation should ideally be confirmed through wet lab experimentation.

1.2.2 Recombination in HIV

Recombination has long been described as one of the most important mechanisms through which retroviruses generate genetic diversity and rapidly rejuvenate sequences in a way than is unlikely through "stepwise accumulation of point mutations" (Simon-Loriere *et al*., 2010; Santoro & Perno, 2013). It is estimated that more than 20% of HIV-1 infections in Africa involve recombinant strains (Van der Kuyl & Cornelissen, 2007). Because HIV is uniquely diploid, each virion carries two complete RNA genomic strands that can undergo a series of recombination events, which along with other mutational events can uniquely modify the viral strain, often leading to fitter viral forms (Burke, 1997; Santoro & Perno, 2013). Even when just a single viral variant initiates an infection and the HIV population that arises within an individual during the acute phase of infection is usually relatively homogenous, as the infection progresses to the chronic phase however, viral diversity can increase rapidly (Herbeck *et al*., 2011; Kearney *et al*., 2009; Shankarappa *et al*., 1999). This diversification process is facilitated by HIV's high mutation and recombination rates coupled with the persistence of powerful continually shifting selection pressures favouring the expansion of viral lineages that harbour viruses with the capacity to evade host humoral and cellular immune responses (Neher & Leitner, 2010; Abram *et al*., 2010; Levy *et al*., 2004; Armitage *et al*., 2012; Zhuang *et al*., 2002; Lau & Wong, 2013; Onafuwa-Nuga & Telesnitsky, 2009; Mostowy *et al*., 2011; Streeck *et al*., 2008).

Less fit viruses can also be produced through the division of favorable mutational combinations, since not all mutations lead to the production of fitter viruses, alternatively fitter viruses could also be produced from parental genomes with an overall lesser fitness (Santoro & Perno, 2013).

Through recombination, infected host cells containing different proviral RNA strands give rise to a heterozygous, recombinant genomic strand capable of subsequent rounds of infection and outgrowth (Abrahams *et al*., 2009). The complexity of HIV diversification dynamics following multiple variant transmission events also increases with increasing genetic divergence of the founder viruses. Since the RNA genome allows for extensive template switching, this also makes it capable of giving rise to multiple subtypes of a virus across population groups (Carr *et al*., 2010). This phenomenon, coupled with the occurrence of polymorphisms present in viral genomes, has the potential to give rise to a completely new viral genome within a single round of replication. In cases where degrees of population-wide genetic diversity are high from the moment of transmission, recombination between the founder variants has the potential to increase diversity far faster than in single variant transmissions (Novitsky *et al*., 2011). Early recombination events contributed to the emergence of current 'pure' HIV-1 subtypes while later recombination events resulted in circulating recombinant forms (CRFs) that exist today (Carr *et al*., 2010), many of which are commonly found across populations in India, China, Myanmar and Central Africa (Neogi *et al*., 2012). There are now nine official 'pure' subtypes and at least 58 validated CRFs that have been identified and categorized through phylogenetic analyses, in existence today (Carr *et al*., 2010; Santoro & Perno, 2013).

1.2.3 Selective pressures acting on HIV

Although the clinical course of HIV is generally well-defined, considerable variability in rates of disease progression exist between patients, indicating differences in the nature of evolutionary paths between the viral population and the host immune system (Lemey *et al*., 2007). Given that the nature of selection likely varies at least slightly between different tissues and cell types, it is plausible that these differences could yield genetically distinct viral sub-populations within different anatomical compartments. This enormous potential for evolutionary change in HIV populations occurs within a single individual (Lemey *et al*., 2007).

Neutralizing antibodies (nAb) have also been reported to induce strong selective pressures that act on the *env* gene and control viral replication and disease progression (Chun *et al*., 2014).

Other factors affecting the rate of HIV evolution and replication include adaptive cytotoxic T cell (CTL) responses which have been reported to partially control HIV replication *in vivo*, and human leukocyte antigen (HLA) class I allelles, that have been linked to the control of disease progression (Koup *et al*., 1994; Carrington *et al*., 1999; Trachtenberg *et al*., 2003). One reported disadvantage of CTL responses however is its ability to frequently target epitopes in other viral genes more strongly (e.g. *nef* and *gag*), resulting in the proliferation of viral variants that escape CTL responses and consequently become more 'virally fit' (Lemey *et al*., 2007; Goulder & Watkins, 2004; Martinez-Picado *et al*., 2006; Leslie *et al*., 2004), however evolutionary escape is associated with a gain in *relative* fitness and not actually *absolute* fitness (Rambaut *et al*., 2004).

Antiretroviral therapies also play a significant role in the development of drug resistant mutations that infer viral fitness over long-term infection (Mesplède *et al*., 2013). Amazingly, HIV is able to rapidly adapt to changing conditions within any individual, successfully evading host responses and therapeutic measures by accumulating numerous amino acid mutations, usually within the variable-loop regions on the *env* gene, while being able to maintain all other replicative functionalities (Lemey *et al*., 2007).

1.2.4 Chemokine receptors CCR5 and CXCR4

To bind to and gain entry into host cells, HIV is dependent on the CD4 glycoprotein that it uses as a receptor and a 7-transmembrane-domain chemokine receptor such as CCR5 and/or CXCR4 that it uses as co-receptors (Cecilia *et al*., 2000; Schuitemaker *et al*., 2011). Co-receptor usage has been shown to correlate with *syncytium-inducing* (SI) and *non-syncytium-inducing* viruses, where CCR5 usage is linked to the later phenotype while multiple others including CXCR4, CCR3, CCR2B, CCR8, GPR15, STRL33, CX3CR1, and dual-tropic viruses (capable of using either CXCR4 or CCR5 co-receptors) are associated with the SI phenotype (Cecilia *et al*., 2000).

This variability in co-receptor usage modulates viral transmission and treatment response and there have been several studies that have associated a switch in co-receptor usage from CCR5 to CXCR4 with disease progression (Cecilia *et al*., 2000; Lusso, 2006; Rambaut *et al*., 2004).

It has also been reported that viruses tend to use CXCR4 co-receptors during the late stages of infection which has been linked to CD4 T cell depletion and rapid disease progression from HIV to AIDS (Connor *et al*., 1997; Penn *et al*., 1999). However, although multiple co-receptors interact with HIV *in vitro*, CCR5 and CXCR4 have been found to be of particular relevance clinically (Schuitemaker *et al*., 2011; Lusso, 2006).

1.3 Naturally occurring HIV restriction factors

Although anti-retroviral therapy has dramatically reduced morbidity and mortality rates in HIV infection, there remains a great demand for alternative clinical strategies since ART entails lifelong therapy, toxicity issues, considerable costs and the emergence of drug-resistant strains (Abdel-Mohsen *et al*., 2013; Alter & Moody, 2010). Naturally occurring HIV restriction factors have been shown to suppress HIV to undetectable levels in the absence of ART by interfering with the virus' replicative cycle at various steps, and therefore represent a promising direction for a practical cure especially if combined with existing therapies to eradicate viral reservoirs (Chan *et al*., 2014; Abdel-Mohsen *et al*., 2013). Among these restriction factors are the widely studied inhibitors of HIV-1 known as APOBEC3G, TRIM5, Tetherin and SAMHD1 (Malim & Bieniasz, 2012; Santa-Marta *et al*., 2013; Chan *et al*., 2014).

1.3.1 APOBEC-induced hypermutation

Viral replication is dependent on the compromise of human intracellular defence mechanisms and the protection of viral gene products that are essential to viral invasion into host cells (Mangeat *et al*., 2003). It is therefore essential for a virus to negate the effect of these families of proteins known for their protective properties in host cells. Hypermutation is observed to be the result of mutational pressure, where a general predisposition to be adenine rich in association with reverse transcription is hypothesized (Müller & Bonhoeffer, 2005).

Guanine to adenine (G to A) hypermutation has a widespread association with significantly reduced plasma HIV RNA levels *in vivo* (Pace *et al*., 2006), highlighting that APOBEC has a similarly significant impact on viral loads as other retroviral restriction factors, which has been reported to be equivalent in response to infection on the same scale as therapies such as Zidovudine for example (Pace *et al*., 2006).

The APOBEC proteins belong to a family of cytidine deaminases expressed in many cell types (Kourteva *et al*., 2012; Strebel, 2005; Noguchi *et al*., 2005), with APOBEC3G and APOBEC3F reported to be responsible for converting cytidine to uridine (Amoêdo *et al*., 2011), which reflects as mutations from Guanine to Adenine (Noguchi *et al*., 2005). The APOBEC3G (A3G) and APOBEC3F (A3F) proteins are frequently reported as inhibitors of HIV-1 (Kourteva *et al*., 2012; Yu *et al*., 2003) where reported targets for APOBEC3G are sequences that are complementary to CGG and TGG, wherein it specifically focuses on mutating G's to A's (Yu *et al*., 2003).

On average hypermutation occurs in 20% of proviral sequences but significant patient-to-patient variability has been reported. Nevertheless, although A3G mutation patterns are present in 95% of hypermutated sequences, hypermutation alone cannot explain the control of viremia and viral suppression is thought to occur through a combination of other viral and host factors, including CTL responses and other immunological factors (Ghandi *et al*., 2008). The question of whether A3G-induced mutations are always lethal to the virus or may occur at sub-lethal frequencies was investigated by Armitage *et al*. (2012), who found that even a single incorporated "A3G-unit" is likely to cause extensive and inactivating levels of HIV hypermutation, and that hypermutation is typically a discrete phenomenon.

1.3.2 Broadly cross neutralizing antibodies

HIV infection inherently elicits a humoral immune response in the human body, prompting the generation of a complex mixture of antibody isotopes with varying specificities (Tomaras & Haynes, 2009). However unlike other antibody responses, HIV-specific antibodies have been reported to take a longer time to develop, usually arising long after HIV latency has been established (Tomaras & Haynes, 2009).

Although the present study does not explore this particular aspect of disease progression, previous work in this domain has recently been conducted for one of the participants studied here (i.e. CAP177) in whom well-developed broadly cross-neutralizing (BCN) antibodies were found to target glycans at amino acid site 332 on the gp120 *env* gene (Moore *et al*., 2012). Viruses that lacked a glycan at amino acid position 332 were found to be resistant to the BCN antibody PGT128, whereas viruses that harboured a glycan at this position were sensitive to PGT128 (Moore *et al*., 2012).

In a larger scale investigation by Gray *et al*. (2011) where other participants from the same cohort studied here were analysed, antibodies that neutralized more than 40% of viruses were discovered in 7 out of 40 participants tested. Longitudinal analysis also revealed that the extent of cross-neutralizing antibody activity developed gradually over time from around the second year post-infection. Two participants showed particularly high neutralization activity, i.e. CAP257 and CAP256, both of whom demonstrated variant neutralization breadths of 82% and 77% respectively, however neutralization activity was found to peak at 4 years post-infection after which there was no increase in activity (Gray *et al*., 2011). Simliar to Moore *et al*. (2012), Gray *et al*. (2011) also reported evidence of antibody activity dependent on the N-linked glycan present at amino acid site 332 on the gp120 *env* region, in addition to another antibody (similar to PG9/16) that was found to be linked to a glycan at aa site 160. Furthermore, in a more recent study of participants from the same cohort studied here, four PNLG sites linked to antibody neutralization sensitivity with strong statistical support in participant CAP177 at amino acid sites 209, 332, 334 and 683 were reported (Lacerda *et al*., 2013). However, despite all of these findings Gray *et al*. (2011) concluded that less than 20% of subtype C infections lead to the development of cross-neutralizing antibodies and if antibodies do arise, they tend to target different regions of the HIV-1 *envelope* region, including regions that are yet to be fully characterized.

1.3.3 Other well known restriction factors

Several other cellular proteins capable of impeding viral replication have been identified, aside from the APOBEC protein family, which act on different steps of the HIV replicative cycle. The TRIM (Tripartite Motif) family of proteins is one of these and is characterized by a highly conserved structure on the amino-terminal region (Santa-Marta *et al*., 2013). Several TRIM family members incl. TRIM5, TRIM11, TRIM15, TRIM19, TRIM22, etc. have been reported to be involved in numerous biological processes such as innate immunity, cell differentiation, and transcriptional regulation (Santa-Marta *et al*., 2013). Among these TRIM5 is the most widely studied protein, as it blocks HIV infection at an early stage in infection by binding to the virus's capsid and inducing premature disassembly before reverse transcription occurs (Stremlau *et al*., 2004). The mechanism through which TRIM5 does this however, is still not fully understood.

Tetherin is another naturally occurring restriction factor that is widely known for its antiviral activity against numerous enveloped virus families, whose main target is the lipid bilayer on the outermost layer of the host cell (Swiecki *et al*., 2012). Tetherin physically binds to budding viral cells preventing the release of new viruses by anchoring them to the host cell (Chan *et al*., 2014). It has however, also been suggested that it could likewise play the opposite role in cell-to-cell transmission by bringing viruses and host cells closer to each other, since viruses have previously demonstrated the ability to escape tetherin restriction (Santa-Marta *et al*., 2013).

Finally, a factor that has been reported to reduce the intracellular dNTP (deoxynucleotide triphosphate) levels to the point where it becomes too low to support HIV-1 reverse transcription, is the SAMHD1 (sterile alpha motif and histidine-aspartate domain-containing) hydrolase (Santa-Marta *et al*., 2013). SAMHD1 is a dNTP hydrolase that is activated when GTP binds to it, causing it to cleave into deoxynucleoside and triphosphate products. Once this happens, the level of dNTP immediately declines in the host cell, limiting HIV-1 reverse transcription during the replicative cycle (Santa-Marta *et al*., 2013). Although there have been studies that focused on therapeutic manipulation of dNTP pools and the development of SAMHD1 inhibitors as potential anti-retroviral therapies, there remains a lack of clinical analyses in this direction (Santa-Marta *et al*., 2013).

1.4 Viral reservoirs

1.4.1 HIV Latency

Current drug regimens suppress HIV replication, but are unable to eradicate the virus completely (Chun & Fauci, 2012). They render the infected CD4 T-cells in a state of 'rest', but these infected cells are still carried in infected individuals. Latent HIV reservoirs are the principal barriers preventing the eradication of HIV infection (Bandyopadhyay *et al*., 2006) and pose as a constant danger to infected individuals as viral replication may potentially be re-initiated, especially when immature T cells (thymocytes) are exported into the blood (Brooks *et al*., 2001). Latency also provides a mechanism through which viruses escape immune recognition (Siliciano & Greene, 2011).

Viral reservoirs typically constitute less than 1 per $1x10^6$ cells (Brooks, 2001), but may be stable for the entire natural lifespan of the host (McNamara & Collins, 2011; Alexaki *et al*., 2008). It is therefore imperative to understand how latency is able to persist, given that cellular lifespans may be very short and may be susceptible to intrusion by other cellular invaders such as bacteria or other viruses. The spread of HIV requires only one cell to give rise to an infectious virus (Erdmann, 2010), however HIV does not have a gene that codes for latency, it simply "remembers" certain antigens encountered even in their latent state and undergoes polymorphisms to evade these antigens when replicating (Erdmann, 2010). To eradicate HIV reservoirs, latent viruses need to be reactivated and eliminated through the body's immune response system and to do this there are several pharmacological compounds that have been reported to reverse latency without activating T cell activation across the host (Shan *et al*., 2012). Nonetheless HIV is known to actively replicate throughout the course of infection and the major mechanism by which it escapes host immune responses is through rapid evolution of mutations more frequently than through latency (Siliciano & Greene, 2011).

1.4.2 Persistent viremia

The large flux of new information relating to HIV after the introduction of antiretroviral therapy has improved the outlook for controlling the HIV pandemic. Even though it is still impossible to definitively cure HIV/AIDS, barring a few exceptions, it is possible to suppress viral loads, partially restoring the body's immune function and inhibiting the progression of HIV (Laprise *et al*., 2013).

Persistent viral infection brings about an almost constant dysfunction of the T-cell lymphocytes (Brockman *et al*., 2009), which in turn inhibits the immune response and compromises the host's ability to protect itself against the invasion of replicating HIV. The body's "natural killer (NK) cell phenotype" that is responsible for lowering viral loads to undetectable levels within the host, becomes less expressed during ongoing viral replication (Brunetta *et al*., 2010). More specifically, it has been found that eliminating a cells ability to produce an enzyme known as "NKG2A" decreases its capability to lower viremia levels (Brunetta *et al*., 2010).

When antiviral therapies are introduced, HIV RNA undergoes a decay (Iglesias-Ussel & Romerio, 2011), initially at a rapid pace coinciding with the half-life of the virus within host cells, and then at a slower pace, which coincides with the lessening of longer lived infected cells. A study by Murray *et al*., (2007) as cited by Iglesias-Ussel and Romerio (2011) went on to report that under monotherapy with Raltegravir for example, there was actually no second phase decay of HIV RNA levels, which led to the belief that this second phase was due to new infections arising from long lived cells or activated latently infected cells. Iglesias-Ussel and Romerio (2011) hypothesized that HIV infection may occur when infected cells escape the cytopathic effect of virus-immune responses and clonal expansion and then become latent memory cells containing the provirus, however this theory remains controversial. In practice, long-term therapy has been associated with a loss of antibodies, as was discovered in the Visconti cohort, while the absence of therapy was linked to a rapid viral rebound and disease progression in patients with a low-level of viremia (Sáez-Cirión *et al*., 2013). Only in rare individuals has post-therapy control been observed, mostly in patients that received therapy during early infection (Sáez-Cirión *et al*., 2013).

Evidence of long-term viral supression has also previously been associated with a lack of viral replication, suggesting that low level viremia is most likely the result of virus production from reservoirs as opposed to new rounds of replication (Dinoso *et al*., 2009). Ultimately however, only through cell death is a viral reservoir depleted, therefore it becomes important for the host to undergo virologic suppression to protect itself from developing drug resistant viruses (Laprise *et al*., 2013) even though HAART antiviral therapy may assist in lowering viral loads in infected individuals, it is unable to prevent the establishment of viral reservoirs over time (Iglesias-Ussel & Romerio, 2011).

1.5 Tissue-specific viral compartmentalization

HIV resides in a wide variety of tissues and compartmentalization of distinct viral populations has been reported in both the male and female genital tracts (Zhu *et al*., 1996; Kemal *et al*., 2003; Philpott *et al*., 2005; Diem *et al*., 2008; Bull *et al*., 2009; Avery *et al*., 2013; Bull *et al*., 2013), the digestive tract (Zhang *et al*., 2002; van Marle *et al*., 2007; Imamichi *et al*., 2011), the brain (Korber *et al*., 1994; Ohagen *et al*., 2003; Ritola *et al*., 2005; Pillai *et al*., 2006; Salemi *et al*., 2005; Smith *et al*., 2010; Harrington *et al*., 2011; Sturdevant *et al*., 2012), the lungs (Itescu *et al*., 1994; Heath *et al*., 2009; Lewis *et al*., 2013), the spleen (Wong *et al*., 1997), the breast (Gantt *et al*., 2010) and the kidneys (Marras *et al*., 2002). If compartmentalization arises from the diversification of transmitted viral variant(s) under differential tissue-specific selection pressures such as those exerted by the immune system, host cell availability and cellular differences in viral replication rates, then it might be expected that viruses derived from different anatomical compartments might possess distinct phenotypic characteristics (Pillai *et al*., 2006; Harrington *et al*., 2011; Heath *et al*., 2009; Schnell *et al*., 2010). Such characteristics potentially include host cell preferences, the extent of glycosylation and degrees of drug resistance (Zhu *et al*., 1996; Bull *et al*., 2009).

Tissue-specific viruses are characterized by host cell characteristics, anti-retroviral drug penetration and region-specific immunological pressure among other stimuli (Pillai *et al*., 2006). In the absence of viral migration it is expected that anatomical compartments serve as sanctuary sites for viruses, where they evolve independently and adapt to local immunological or cellular characteristics (Clarke *et al*., 2000; Bull *et al*., 2009).

As a result effective prevention, treatment or control strategies are dependent on an accurate understanding of how HIV evolves in specialized tissues, and whether compartment-specific pressures contribute to the diversity and fitness of viral subpopulations (Kemal *et al*., 2003; Clarke *et al*., 2000).

Although most published studies on viral compartmentalization in the female genital tract have been cross-sectional in design, considering viruses sampled at only a single time point, compartmentalization of viruses in the female genital tract has previously been detectable in around 33% of chronically infected women (Overbaugh *et al*., 1996; Kemal *et al*., 2003; Philpott *et al*., 2005; Andreolétti *et al*., 2007; Boeras *et al*., 2011; Chaudhary *et al*., 2012) and has been reported in studies of both subtype B and C infections (Adal *et al*., 2005; Chomont *et al*., 2007). Philpott *et al*. (2005) for example found that among the five women studied, strong evidence of compartmentalization was found in four of the women using phylogenetic methods even though no statistically significant difference in co-receptor usage or N-linked glycosylation patterns between cervical and plasma derived viruses was evident. Similarly, Chaudhary *et al*. (2012) found that among the six women studied, although no tissue-specific difference in co-receptor usage was discovered, sequences clustered together by tissue type on maximum likelihood phylogenetic trees in women with higher CD4 T cell counts, suggesting that the immune response plays a fundamental role in the compartmentalization of viral populations.

Despite these findings however, the only three published longitudinal studies involving viruses sampled between 1.5 and 3.5 years post-infection have still been unable to conclusively establish whether compartmentalization is the rule (detectable in 3/3 individuals studied by Poss *et al*., 1998) or the exception (detectable in 4/14 individuals studied by Sullivan *et al*., 2005) using rigorous analytical methods (Bull *et al*., 2013). Among the longitudinal studies that investigated both genotypic and phenotypic differences between viral populations from the cervix and blood stream, contradictory results were reported (Poss *et al*., 1998; Sullivan *et al*., 2005; Bull *et al*., 2013). In the most recent longitudinal study by Bull *et al*. (2013) in which co-receptor usage, migration events and PNLG site accumulation was analysed in cervical and blood plasma sequences from eight women, no definitive tissue-specific differences were reported. Although a greater number of PNLGs were discovered within cervical sequences in six of the eight women analysed, when PNLG sites were

excluded from the sequences and reanalysed, the clustering of sequences on phylogenetic trees did not change (i.e. sequences remained intermingled on all clades) (Bull *et al*., 2013). Sullivan *et al*. (2005) and Poss *et al*. (1998) focused on genetic diversity, divergence and evolution of cervical and blood plasma sequences, and found that there was no clear distinction between viruses from either tissue type in 10 of the 17 data sets analysed, however the methodologies employed in these studies were not extensive and came before the introduction of advanced phylogenetic methods.

Numerous approaches have been developed to predict the importance of "sanctuary sites", HIV persistence and virus trafficking between compartments including mathematical and statistical models (Clarke *et al*., 2000). In a review of the most appropriate statistical methods to detect viral compartmentalization, Zárate *et al*. (2007) describes phylogenetic methods as fundamental to establishing whether viruses migrate between compartments, especially as the viral load increases, since there is a likely sampling bias when samples are taken during periods when viral loads are high. Zárate *et al*. (2007) further observed that the compartmentalization of HIV was not confined to specific disease stages and that there is actually no ideal method to detect compartmentalization. Phylogenetic tree-based methods however, such as the Slatkin-Maddison (SM) test (Slatkin & Maddison, 1989) and Simmonds association index (AI) (Wang *et al*., 2001) are generally considered to have greater power in detecting compartmentalization than distance-based methods such as the Wright's F_{ST} (Wright, 1943) and AMOVA tests (Excoffier *et al*., 1992; Zárate *et al*., 2007). However, it is also recommended if a phylogenetic approach is adopted, that screening for recombination and the use of reliable phylogenetic models is imperative to obtain a better understanding of viral compartmentalization and infection dynamics, as these factors could influence phylogenetic inferences (Zárate *et al*., 2007).

In a study employing phylogenetic and Slatkin-Maddison analyses, Gantt *et al*. (2010), found limited compartmentalization of HIV-1 in breast milk specimens, suggesting extensive interchange of viruses between breast milk and the blood plasma. HIV-1 viruses were compartmentalized within breast milk in 35% of the sequences obtained from nine women through Slatkin-Maddison testing, while phylogenetic analyses demonstrated extensive mixing between viral sequences from milk and blood plasma samples.

Monotypic sequences were found to be overrepresented in HIV-1 populations from milk specimens and accounted for half of the inferred compartmentalization, however local HIV-1 production within the breast was later linked to inflammation (Gantt *et al*., 2010). In another investigation that analysed sequence data for evidence of tissue-specific compartmentalization, Marras *et al*. (2002) compared peripheral blood mononuclear cells (PBMC's) and kidney-derived sequences from two patients sampled cross-sectionally and found strong evidence of tissue-specific HIV evolution in the kidney. Phylogenetic analyses revealed kidney-derived sequences formed tissue-specific "subclusters" that were surrounded by blood-derived sequences. This pattern combined with evidence of HIV-1-specific proviral DNA and mRNA that was detected in kidney cells only, pointed towards the existence of a "renal viral reservoir", (Marras *et al*., 2002).

Research into viral compartmentalization patterns in other tissue types such as the liver has also suggested that tissue-specific selection pressures that drive viral adaptation in the liver's microenvironment may be responsible for tissue-specific compartmentalization in this organ compared to viral populations sampled from the blood stream (Blackard *et al*., 2011). HIV compartmentalization has been exhibited in the brain as well. In a study by Salemi *et al*. (2005) in which a local-molecular-clock analysis was performed, the authors reported evidence that HIV-1 subpopulations in the meninges and temporal lobe respectively evolve about 30 to 100 times faster than viral populations in other parts of the brain. In contrast, Imamichi *et al*. (2011) found no evidence of viral compartmentalization between viruses in the gut colon and ileum or between the gut and peripheral blood, and the presence of HIV-1 sequences in both the gut and peripheral blood at the same time points tested suggested ongoing viral replication in both compartments.

In another study that investigated the existence of genetically distinct HIV-1 populations within different tissue compartments including the central nervous system, the reproductive tract and gastrointestinal mucosa, Potter *et al*. (2004) found no convincing evidence for this phenomenon suggesting that the emergence of phylogenetic structure in different compartments over time in HIV-infected individuals is likely to arise from a continual process of migration of infected cells into each compartment, followed by localized expansion and evolution of each population.

Nevertheless, although well-documented differences between HIV populations from different tissue-specific compartments in the same individual have been reported, including between the genital tract and blood plasma (Philpott *et al*., 2005; Kemal *et al*., 2003; Zhu *et al*., 1996; Diem *et al*., 2008), it has been suggested that many of these studies failed to rigorously test for compartmentalization (Bull *et al*., 2009) and results from a more recent longitudinal study (Bull *et al*., 2013) that tested for this phenomenon, has cast doubt on the generality of previous findings, suggesting that further investigation into the dynamics of viral subpopulations over long-term infection are required. This study therefore aimed to address the methodological gap in previous studies through the use of both distance and tree based methods (including advanced phylogenetic reconstruction models as implemented in BEAST, tMRCA and substitution rate estimation, etc.) while contributing to the limited knowledge on compartmentalization dynamics over long-term HIV infection in treatment naïve patients.

Chapter **2**
Methodology

2.1 Participant background

HIV-1 *env* gp120 sequences analysed in this study, were obtained in collaboration with the National Institute for Communicable Diseases (NICD) and the Centre for the AIDS Programme of Research in South Africa (CAPRISA). Cervico-vaginal lavage (CVL) and blood plasma samples were obtained from participants enrolled in the CAPRISA 002 Acute Infection study in Durban, Kwa-Zulu Natal. The cohort, originally containing 245 uninfected high-risk women was initiated in 2004 to investigate the natural history of HIV-1 during acute infection (van Loggerenberg, 2008). The subset of participants included in this study consisted of four women anonymously identified as CAP177, CAP217, CAP261 and CAP270, all of whom had no previous exposure to anti-retroviral therapy. All CVL samples obtained from these women were collected only when they were not menstruating, during their luteal phase. Participants were aged between 18 and 37 years old and were also tested for sexually transmitted infections (incl. vaginosis, HSV type 2, syphilis, gonorrhoea, etc.) in addition to having their viral loads and CD4 cell counts in the blood plasma measured at each clinical visit.

2.2 Ethics

All sequences analysed in this study have been passed through an ethics committee and approved through the CAPRISA cohort. Ethics approval has been received from three South African Universities (University of KwaZulu-Natal Ref: E013/04, 5 July 2004; University of the Witwatersrand Ref: MM040202, 22 April 2004; University of Cape Town Ref: 025/2004CA, 23 June 2004), and all samples were anonymously recorded.

2.3 Sample processing

All HIV-1 samples were sent to the AIDS Virus Research Unit at the National Institute for Communicable Diseases (NICD) where RNA extraction, cDNA synthesis, PCR and sequencing were conducted. In order to obtain the CVL sample, both the vagina and cervix were washed repeatedly with 10ml phosphate buffered saline (PBS) before centrifuging the PBS to separate cellular components from supernatant. The supernatant was then checked for blood contamination with Roche Cobas combur dipsticks before being aliquoted into 2ml portions and stored at -70^{0}C. Although CVL samples were taken during acute and chronic infection, it was not possible to amplify *env* sequences from every sample taken possibly due to inconsistent HIV-1 shedding in the female genital tract.

RNA was extracted between 0.6 and 1.6ml of the CVL supernatant and 140μl of plasma using the Qiagen QIAmp viral RNA MiniKit according to the manufacturer's protocol. For extraction from CVL samples, an on-the-column DNase digestion step was included to ensure that all amplified *env* sequences were obtained from cDNA and not DNA during RNA purification. cDNA synthesis and the complete HIV-1 envelope were amplified in a nested PCR reaction according to the protocol described by Salazar-Gonzalez *et al*. (2008). The protocol for single genome amplification (SGA) was then followed, as sequences from individual viruses were required for the purposes of this study. PCR reactions were continued until at least 15 amplicons were obtained for each sample or until all the cDNA was used. As CVL samples contain relatively very few viruses compared to the blood plasma, it was often not possible to amplify the required number of amplicons.

PCR products were cleaned with the Qiagen PCR product purification kit and directly sequenced with the ABI PRISM Big Dye Terminator V3.1 Cycle Sequencing Ready Reaction Kit (Applied Biosystems) and resolved on a 3100 Automatic Capillary Sequencer (Applied Biosystems). Sequence analysis was then performed using the Sequencher V4.5 program to form contiguous sequences for the complete HIV-1 *env* region, where amplicons containing double peaks or interrupted reading frames were excluded from downstream analyses.

2.4 Multiple sequence alignment

HIV-1 *env* sequences from participants CAP177, CAP217, CAP261 and CAP270 were obtained from the NICD and were arranged into four sets of alignments per participant for subsequent analyses (Table 2.1). Intrapatient sequence alignments were generated using ClustalX v2 (Larkin *et al*., 2007) in multiple alignment mode with HXB2 as the reference sequence. ClustalX is an open-source software package that aligns sequences in pairs to generate a distance matrix that is used to construct a rough initial phylogenetic tree of the sequences, from which a multiple sequence alignment is progressively created based on the branching order of a guide tree (Thompson *et al*., 1997). Subsequent visual inspection and manual editing of the alignments was then performed using the Se-Al v2.0a11 (http://tree.bio.ed.ac.uk/software/seal/) sequence alignment editor (Rambaut, 2002; Abecasis *et al*., 2007). Intrapatient sequence alignments were then codon aligned in Se-Al to avoid the inclusion of DNA encoding frame shifts and mistranslated sequences that would affect the results of downstream analyses.

Table 2.1 Intrapatient multiple sequence alignments organized into four qualitative data sets categorized by *env* region and presence or absence of monotypic and low diversity sequences. The first data set included all HIV-1 *env* gp120 sequences while the second excluded monotypic sequences and third excluded both monotypic and low diversity sequences. The fourth data set contained separated V-loop regions from each intrapatient alignment within data set 1 only.

Data set	Content	Description
1	Intrapatient full-length HIV-1 *env* gp120 alignments	Included all sequences.
2		Excluded monotypic sequences (included low diversity sequences).
3		Excluded both monotypic and low diversity sequences.
4	Intrapatient HIV-1 *env* V-loop alignments	Variable loop regions V1-V5 from all sequences (including monotypic and low diversity sequences).

2.4.1 Monotypic and low diversity sequence removal

The presence of identical (monotypic) and genetically similar (low diversity) sequences has been shown to bias statistical tests for compartmentalization towards false positives (Bull *et al*., 2009; Gantt *et al*., 2010), and monotypic sequences have also been proposed to be associated with recent bursts of replicating viruses, or with proliferation of provirus containing host cells in discrete tissue types (Bull *et al*., 2009; Boeras *et al*., 2011; Gantt *et al*., 2010). Monotypic and low diversity sequences were therefore reduced to a single representative sequence to eliminate any potential false positives in the compartmentalization results.

A total of 101 monotypic *env* sequences were detected among all participants during assembly of data sets excluding monotypic sequences (i.e. data sets 2 and 3; Table 2.1), the majority of which were found in participant CAP177 and were distributed across all six time points sampled 14 (n=13), 28 (n=17), 378 (n =3), 560 (n=4), 924 (n=2) and 1295 (n= 10) days post-infection. In CAP217, 32 monotypic sequences were detected at four sampling time points 14 (n=15), 63 (n = 11), 770 (n = 3) and 1316 (n= 3), with only six monotypic sequences found in CAP261 at 63 days post-infection. In CAP270, 14 monotypic sequences were detected at 56 (n = 8), 147 (n =2), 406 (n =2) and 903 (n =2) days post-infection. A single sequence was used to represent several monotypic sequences, resulting in a total of 23 representative sequences that remained and 78 monotypic sequences that were omitted.

Low diversity sequences were defined as sequences with a less than 0.1% pairwise genetic distance between other sequences within each participant data set. A total of 50 low diversity sequences were identified in participants CAP177 and CAP217, whereas none were present in CAP261 and CAP270. More than half of these sequences were detected in CAP177 at 14 (n=5), 28 (n=10), 196 (n=3), 378 (n=7) and 1295 (n=3) days post-infection, while in CAP217 low diversity sequences were present at 14 (n=3), 63 (n=15) and 420 (n=4) days post-infection. Similar to the way monotypic sequences were handled, a single sequence was used to represent multiple low diversity sequences, resulting in a total of 12 representative sequences that remained (i.e. in data set 3; Table 2.1) while 38 low diversity sequences were omitted. Alignments were then analysed via the pipeline described in Figure 2.1.

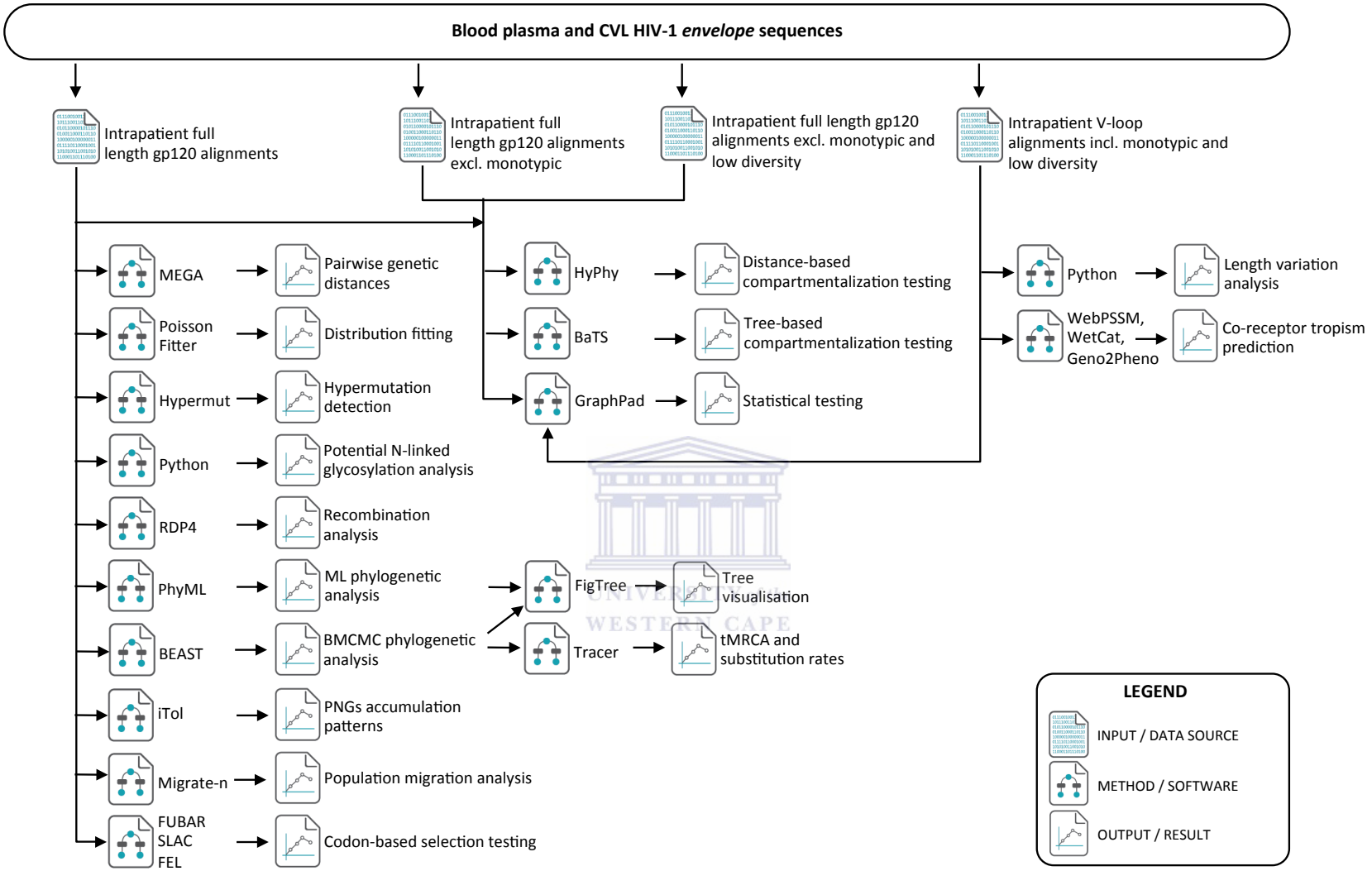


Figure 2.1 Graphical representation of the analysis pipeline followed for each qualitative data set. Blood plasma and CVL sequence alignments, prepared using ClustalX and Se-Al (section 2.4) were analysed through four main streams as illustrated above. Full-length *env* gp120 sequences were analysed for pairwise genetic distances, Poisson distributions, hypermutation, potential N-linked glycosylation, recombination, ML and BMCMC phylogenetic structure, substitution rates, time to the Most Recent Common Ancestor, PNLGs accumulation patterns, viral population migration and codon-based selection. Three of the four intrapatient alignments were also used for distance-based and tree-based tissue-specific viral compartmentalization analyses, while V-loop regions were evaluated for co-receptor tropism and length variation.

2.5 Phylogenetic evaluation of viral compartmentalization

Phylogenies were constructed using Bayesian methods for longitudinal analysis and Maximum Likelihood (ML) methods for cross-sectional analysis. Bayesian methods included Markov Chain Monte Carlo (BMCMC) inference as implemented in the BEAST v1.8.0 (Bayesian Evolutionary Analysis of Sampling Trees, http://code.google.com/p/beast-mcmc/downloads/list/) software package (Drummond *et al.*, 2002, Drummond & Rambaut, 2007) and MrBayes v3.2 (http://mrbayes.sourceforge.net/, Huelsenbeck & Ronquist, 2001). Bayesian analysis uses what is scientifically known about the data with predefined parameters known as 'priors' to determine probability assessments of the hypothesis, known as the 'posterior', through the use of Bayes theorem (Newton & Raftery, 1994; Drummond & Rambaut, 2007; Kitchen *et al.*, 2004):

$$p(\theta \mid Y) = p(Y \mid \theta) \times p(\theta) / p(Y)$$

Bayes theorem states that the posterior distribution (θ) is proportional to the sampling density of the data, given θ multiplied by the prior distribution of θ, where the prior distribution incorporates any scientific knowledge about θ before observing the data, while the denominator $p(Y)$ represents a normalizing factor that ensures the posterior density totals 1 (Weiss *et al.*, 1997; Suchard *et al.*, 2001; Kitchen *et al.*, 2004).

Unlike the Maximum Likelihood approach that provides a single estimate of the most likely tree, the Bayesian MCMC method produces a posterior distribution of trees, from which samples are taken using either the Gibbs or Metropolis-Hastings (MH) MCMC sampling methods. For this analysis, the MH algorithm was used (as implemented in the BEAST v1.8.0 software package), which samples from the distribution of trees by making use of a "full joint density function and (independent) proposal distributions" for each variable being studied. Although the Gibbs sampling method is more frequently used, one of the limitations of this technique is its inefficiency in handling local regions with high densities, where "mixing" of the Gibbs sampling chain is very slow (Gelfand *et al.*, 1990; Lynch, 2007; Yildirim, 2012). Sampling chains are considered to have limited mixing when successive samples are highly correlated, resulting in a very slowly changing value from one sample to the next.

Alternatively, the MH algorithm is designed around two main constraints, (1) "the sampler should not tend to visit higher probability areas under the full joint density" and (2) "the sampler should explore the space and avoid getting stuck at one site", (Yildirim, 2012). The MH method ensures that there is a sampling balance by satisfying these constraints, which guarantees that the resulting stationery distribution of the MH algorithm is actually the target posterior distribution of trees that is required for this type of analysis.

For this analysis, all phylogenies were constructed under the best-fit nucleotide substitution model as identified by ModelTest.

2.5.1 Nucleotide substitution model selection

The best-fit nucleotide substitution model was identified using ModelTest v3.7, which uses log likelihood scores to determine the most appropriate model of DNA evolution that fits the data (Posada & Crandall, 1998). A total of twelve nucleotide substitution models were tested using full-length HIV-1 *env* gp120 intrapatient alignments (data set one; Table 2.1) as input, before the GTR+G model was determined as the best-supported model for all participants.

2.5.2 Evolutionary model selection

Model selection is a crucial step in molecular phylogenetics and Bayesian estimation methods as it influences estimated substitution rates, phylogenies and posterior probabilities among other model parameters (Posada & Buckley, 2004). Model selection however, is simply a way of approximating evolutionary relationships, rather than identifying them explicitly, choosing an appropriate evolutionary model is essential for confidence in the results obtained from phylogenetic analysis (Wang *et al*., 2001; Posada & Buckley, 2004).

To determine the best evolutionary model for each intrapatient alignment in data set one (Table 2.1), four different evolutionary models were investigated in BEAST including non-parametric (Bayesian skyline plot) and parametric (constant population size) demographic models, as well as the strict and uncorrelated lognormal relaxed molecular clock models. The Markov chain length was set to 200 million iterations with trees sampled at every 20,000 steps, producing a posterior distribution of 10,000 trees.

Resultant log files from BEAST were viewed in Tracer v1.5, where burn-in limits were adjusted individually to ensure that effective sample size (ESS) scores of all model parameters were above 200, indicating ample mixing of the Markov chain and convergence to the stationery distribution. Bayes Factor (BF) testing was then applied to identify the evolutionary model with the best statistical support for each alignment (Drummond & Rambaut, 2007). The models with the highest $\log_{10}$ BF score (i.e. closest to 100) were selected as the best evolutionary models for each of the data sets analysed, and all Bayesian phylogenetic analyses were performed on the supercomputing cluster at the South African National Bioinformatics Institute (SANBI).

2.5.3 Bayesian probabilistic modeling of viral evolution

For full intrapatient *env* alignments (data set one; Table 2.1) a posterior distribution of phylogenetic trees were inferred using the best fit nucleotide substitution and evolutionary models (identified in sections 2.5.1 and 2.5.2) with BEAST v1.8.0 to estimate (1) the ancestral relationships, (2) the time to the most recent common ancestor (tMRCA) and (3) nucleotide substitution rates for HIV-1 *env* sequences from the blood plasma and CVL from each participant. BEAST input files were created in BEAUti v1.8.0, where the Markov chain length was set to 100 million steps and trees were sampled at intervals of 10,000. Resulting log files were viewed in Tracer v1.5 to determine burn-in limits and improve overall ESS scores as described above. Maximum clade credibility trees were then interpreted from the posterior distribution of 10,000 trees after discarding the first 1,000 trees as burn-in, using the TreeAnnotator v1.8.0 program. All trees were then further annotated in FigTree v1.4.0 to illustrate tissue-specific evolutionary patterns.

2.5.4 Maximum likelihood phylogenetic reconstruction

Maximum likelihood (ML) phylogenetic structure was estimated using sequences from the earliest time point where sequences from both tissue types were available for each participant, in a cross-sectional sample design with PhyML (http://www.atgc-montpellier.fr/phyml/). The GTR+G nucleotide substitution model was used for this analysis with nodal support assessed through bootstrapping (i.e. 1,000 replicates). All ML trees were then annotated in FigTree v1.4.0 (http://tree.bio.ed.ac.uk/software/figtree/) and remained unrooted.

2.6 Statistical evaluation of viral compartmentalization and population migration

2.6.1 Tree-based compartmentalization testing

One of the most common phylogenetic methods used to test the number of state changes between compartments is the parsimony approach implemented in the Fitch algorithm that calculates the parsimony score (PS) and association index (AI) statistics (Slatkin & Maddison, 1989; Fitch, 1971; Parker *et al*., 2008). The Slatkin-Maddison (S-M) test for compartmentalization was used to evaluate viral segregation between blood plasma and cervical tissues using a Bayesian Markov chain Monte Carlo approach as implemented in the Bayesian Tip-association Significance testing software tool BaTS v2 (Esbjörnsson *et al*., 2011; Gantt *et al*., 2010; Parker *et al*., 2008; Huelsenbeck & Ronquist, 2001). This method has been previously used to detect intra-host compartmentalization of HIV between various tissue types (McGrath *et al*., 2001, Kemal *et al*., 2003, Gantt *et al*., 2010).

BaTS takes as input either a single ML tree or a posterior sample of trees and calculates three metrics of phylogeny-trait correlations; (1) the parsimony score of Slatkin and Maddison (1989) (PS) which represents the number of times that a trait under investigation is gained or lost (Fitch 1971), (2) the Simmonds association index (AI) which assesses the degree of population structure within a phylogenetic tree (Slatkin & Maddison, 1989; Wang *et al*., 2001; Zárate *et al*., 2007), and (3) the monophyletic clade (MC) statistic (Salemi *et al*., 2005) which quantifies the maximum clade size on the MCC tree comprised of samples that share a single trait such as tissue type. Associated p-values for these three statistics are determined by permuting sequences between tissue compartments (Parker *et al*., 2008). Since the trait of interest (host-tissue type) can be considered a "location" within the participant, a tight phylogeny-trait correlation (low AI and PS values and a high MC value) suggests low lineage dispersal between the cervix and blood plasma (Parker *et al*., 2008), indicative of viral compartmentalization.

For this analysis, Bayesian MCMC trees were generated with MrBayes v3.2.1 (Huelsenbeck & Ronquist, 2001) using the GTR+G substitution model for all participants (data set one; Table 2.1) where the Markov chain length was set to one million steps and trees were sampled at intervals of 1,000. A 50% burn-in limit was then applied to the posterior sample of trees to exclude phylogenies with high p-values from the distribution. A posterior sample of 5,000 trees was then analysed with BaTS using 1,000 null distributions to test the significance of the observed phylogeny-trait correlations between viral populations in the blood plasma and cervix in each participant. Observed phylogeny-trait correlations were compared to those found in 1,000 randomly generated trees in which the sequences had been randomly arranged and considered significant when less than 1% of the randomized trees required the same or a fewer number of migration events as for the sample data.

2.6.2 Distance-based compartmentalization testing

While tree-based methods are dependent on the structure of a phylogenetic tree, which can be confirmed by subjectively examining the trees, distance-based tests are good indicators of genetic similarity between tissue-specific samples that are independent of the evolutionary path (Heath *et al*., 2009). Although there have been multiple studies that simply examined clustering of tissue-specific clades on phylogenetic trees to classify compartmentalization patterns, this feature is sometimes not obvious and it is vital to confirm results from tree-based methods with mathematical distance-based methods (**Zá**rate *et al*., 2007; Heath *et al*., 2009). To do this, *Hudson's Nearest Neighbor* (Snn) and *Wright's measure of population subdivision* (F_{ST}) statistics were used (Hudson, 2000; Wright, 1943; Slatkin, 1993; **Zá**rate *et al*., 2007; Heath *et al*., 2009; Edo-Matas *et al*., 2010).

The Snn statistic was used to compare genetic distances within and between sequences from different tissue types independent of phylogenetic inferences (Bull *et al*., 2009) by measuring "how often the nearest neighbor (in sequence space) sequences are from the same locality in geographic space", (Heath *et al*., 2009), where geographic space refers to the sampled tissue type and locality is a particular region within the *env* gene. Whereas, the F_{ST} statistic compares the mean pairwise genetic distance between sequences from different tissue types to mean distances between sequences from the same tissue type (**Zá**rate *et al*., 2007; Edo-Matas *et al*., 2010).

Both Snn an F_{ST} tests for viral compartmentalization were performed longitudinally and by time point with data sets that included or excluded monotypic and low diversity sequences (data sets one, two and three; Table 2.1). Alignments were analysed using the HyPhy v2.2 software package (Pond *et al*., 2005) and tested with 1,000 permutations, to determine if significant evidence of compartmentalized structure existed for viruses sampled from the two tissue types investigated (Bull *et al*., 2009; Pond *et al*., 2005).

2.6.3 Structured coalescent-based migration estimation

The coalescence-based program Migrate-n v3.3.0 (http://popgen.sc.fsu.edu/Migrate/Migrate-n.html) was used to estimate the amount of gene flow between viral populations in the cervix and blood plasma (Beerli & Felsenstein, 1999). Migrate-n uses *Bayes Factors* to compute the marginal likelihoods of two hypotheses (i.e. state 1 and state 2) and addresses complex questions such as whether explicit migration between different populations exists (Beerli & Palczewski, 2010). In this study a full model with two population sizes and two migration rates (i.e. in and out of the blood plasma (state 1) and cervical tissue (state 2) was used. Migrate-n input files were created for each participant from full-length gp120 sequences (data set one; Table 2.1) using a Python script before being tested for significance with 1,000 null replicates.

2.7 Calculating average pairwise genetic distances

The degree of sequence variation between different gene loci can help to determine what types of evolutionary mechanisms and mutational factors affect specific genomic regions (Castresana, 2002). Average pairwise genetic distances were calculated using the p-uncorrected model in MEGA v5.2.2 (http://www.megasoftware.net/). MEGA estimates pairwise genetic distances by calculating the proportion of nucleotide differences between pairs of sequences (Tamura *et al*., 2013). Changes in tissue-specific genetic diversity in each participant over time were then visualized using boxplots.

2.8 Poisson distribution fitting

Although most HIV transmission events involve a single viral variant, up to 50% of transmissions are estimated to represent multiple variant infections (Gottlieb *et al*., 2008; Keele *et al*., 2008; Ritola *et al*., 2004). The clinical implications of multiple variant infections are important to the rate of disease progression, as more diverse viral populations have previously been associated with rapid disease progression (Gottlieb *et al*., 2004; Grobler *et al*., 2004; Sagar *et al*., 2003; Abrahams *et al*., 2009). To test whether each of the participants analysed here were likely infected by more than one HIV variant, the approach described by Keele *et al*. (2008) and Abrahams *et al*. (2009) was used, in which multiple variant infections did not follow a Poisson distribution, suggesting that infection with multiple variants does not usually occur through independent events, but rather to high transmission rates (Abrahams *et al*., 2009).

When formulating a hypothesis about the specific distribution of a variable of interest (in this case the number of transmitted variants), variables whose values are determined by an infinite number of independent random events are said to follow a normal distribution, for example the height of a person is the result of many independent factors such as genetic predisposition, nutrition, diseases, etc. therefore height tends to be normally distributed within population groups (StatSoft, 2013). Alternatively, if a variable's values are the result of rare events, then the variable will be distributed according to the Poisson distribution, for example an accident that is the result of a series of unlikely events (StatSoft, 2013). As a result the Poisson distribution is sometimes referred to as the distribution of rare events and is defined as follows:

$$f(x) = (?^{x} * e^{-?})/x!, \text{ for } x = 0,1,2,\ldots, 0 < ?$$

where **?** = (lambda) is the expected value of **x** (the mean), and **e** is the base of the natural algorithm sometimes referred to as "Euler's e" (StatSoft, 2013).

For this analysis, to determine the pattern of modality, where a *bimodal* distribution (i.e. the presence of two or more peaks in the Poisson distribution) was hypothesized to be an indication of HIV infection by more than one viral variant, and a *unimodal* distribution (i.e. the presence of a single peak in the Poisson distribution) was hypothesized to indicate infection by a single HIV variant, the Poisson Fitter tool (http://www.hiv.lanl.gov/) was used to estimate distribution patterns as implemented in the Los Alamos HIV Sequence Database (Rose & Korber, 2000; Dixit & Perelson, 2004), which analyses the frequency of Hamming distances by calculating the best fitting distribution using the Goodness of Fit test (GOF) and maximum likelihood method (Novitsky *et al*., 2011). Mean substitution rates obtained from BEAST analysis (in section 2.5.3) were used as input values per participant and Poisson distributions were visualized as a line charts with Hamming distances plotted as histograms.

2.9 Potential N-linked glycosylation site prediction

The number and distribution of PNLGs were determined using an in-house Python v2.7.2 script (Appendix I) that combined the algorithm implemented in the HIV LANL N-glycosite tool (http://www.hiv.lanl.gov/) along with additional criteria from the literature describing the identification of N-linked glycosylation sequons. While the N-glycosite tool searches for sequons satisfying the tri-peptide consensus of NX(S/T), where N is asparagine, S is serine, T is threonine and X is any amino acid except proline, the Python script used in this study searched for sequons conforming to the NX(S/T)Y consensus, where X & Y respresent any amino acid except proline, and where in the NX(S) context, X could not be tryptophan (W), aspartic acid (D), glutamic acid (E) or leucine (L) (Zhang *et al*., 2004; Kasturi *et al*., 1997; Rao & Bernd, 2010). These additional criteria were specified to ensure the exclusion of sequons that have previously been reported to be poorly glycosylated *in vitro* (Rao & Bernd, 2010).

HXB2 was used as the reference sequence for this analysis, since PNLGs have already been biochemically defined for this HIV subtype (Zhang *et al*., 2004). Once PNLG sites were predicted, the total number and distribution of PNLGs across the full-length gp120 and V-loop regions were statistically compared between tissue types at all sampling points over the course of infection in each participant.

Non-parametric Mann-Whitney and Wilcoxin signed-rank tests were used to assess whether the number of PNLGs in *env* subregions displayed significant differences between the tissue types investigated, and were considered significant if p-values were less than or equal to 0.05.

2.9.1 Clustering of potential N-linked glycosylation sites

To determine if PNLGs accumulated consistenly or emerged at specific time points or infection stages, in both or a single tissue type throughout the course of infection, Bayesian MCC trees (produced in section 2.5.3), were used to illustrate the presence of selected glycosylation sites that demonstrated an increasing occurrence in blood plasma or CVL viruses in each participant. Annotation files containing binary data indicating the presence or absence of a specific PNLG site were first created using Python before being used as an annotation file in the Interactive Tree of Life (iTol) v2 tool (Letunic & Bork, 2011). Trees were then examined visually before selected trees were reannotated in Figtree v1.4.0 to illustrate closely related sequences that shared specific PNLGs, some of which are known to have an impact on neutralizing antibody response. This information was then linked to predicted recombination events to determine whether recombination played a role in the accumulation of PNLGs and/or if PNLGs emerging later in the course of infection explicitly coincided with predicted recombination events or recombinant regions within blood plasma or CVL sequences.

2.10 Recombination detection

Detection of potential recombinant sequences, identification of likely parental sequences and localization of possible recombination breakpoints within full-length gp120 sequences (data set one; Table 2.1) was achieved using the RDP, Geneconv, Bootscan, Maximum Chi Square, Chimaera, Sister Scan and 3Seq recombination detection methods as implemented in RDP4 (http://darwin.uvigo.es/rdp/rdp.html) (Martin *et al*., 2010).

The RDP method scans through multiple sequence alignments searching for evidence of recombination based on three fundamental sets of criteria (Martin *et al*., 2010; Varsani *et al*., 2006):

- All non-informative sites are discarded, e.g sites that are identical between any three sequences, different in all three sequences, or unique to the three sequences that is also not present in the reference sequence or any other sequences in the alignment.
- Informative sites are analysed by a sliding window that scans through one nucleotide at a time and then calculates an *average percentage identity* between paired sequences in a set of three sequences, for example where sequence A is compared to sequence B, sequence B compared to sequence C, and sequence C compared to sequence A. Regions are considered potentially recombinant if the average percentage identity between sequence A and B or B and C is higher than that of C and A.
- Using the binomial distribution, the probability that nucleotide identities might have occurred by chance is also assessed and a p-value is calculated from this probability by multiplying it with the total number of unique windows examined. The p-value is then Bonferroni-corrected by multiplying it with the total number of nucleotide triplets (codons) examined in the alignment.

For this analysis, the default parameter settings were used for all detection methods except RDP, where a window size of 15 nucleotides was set and only events detected by three or more methods were considered as credible evidence of recombination. The breakpoint positions and recombinant sequence(s) inferred for every potential recombination event was then manually checked and adjusted where necessary using the extensive phylogenetic and recombination signal analysis features available in RDP4.

All full-length gp120 intrapatient sequences were analysed for patterns based on the distribution of recombination breakpoints across tissue types and sampled timepoints. Recombinants were considered *unique* if only a single sequence showed evidence of a breakpoint, and *enriched* if more than one sequence contained the same breakpoint.

2.11 Codon-based selection analysis

One of the fundamental processes driving the rapid evolution of HIV is natural selection, which plays a pivotal role in viral diversity, differentiation and adaptation (Poon *et al*., 2009). Most recent codon-based techniques that infer selection are based on a probabilistic approach that check whether the "nonsynonymous substitution rate at a specific site is faster or slower than the neutral rate, which is typically set to the synonymous rate at the same site (or to the mean synonymous rate for the entire alignment)", (Murell *et al*., 2013; Pond & Frost, 2005). To determine if there were any specific codons that were under positive or negative selection in blood plasma or CVL-derived viruses during disease progression, the *Fast Unconstrained Bayesian AppRoximation* (FUBAR), *Single Likelihood Ancestor Counting* (SLAC), *Mixed Effects Model of Evolution* (MEME) and *Fixed Effect Likelihood* (FEL) methods were used to analyse full-length gp120 intrapatient alignments (data set one; Table 2.1) as implemented on the Datamonkey web server (Pond & Frost, 2005; Murrell *et al*., 2013). Intrapatient alignments were first run through the *Genetic Algorithm for Recombination Detection* (GARD) tool to infer recombination breakpoints using the HKY85 model, with all other parameters kept at their default settings. FUBAR, FEL and MEME analyses were conducted using GARD inferred trees and the HKY85 model, where default probability cut offs were set for all analyses. Based on recommendations from other authors, only sites that were identified by at least three methods with a p-value less than or equal to 0.05 were considered as credible evidence for positive or negative selection (Wlasiuk & Nachman, 2010; de Matos *et al*., 2013; Castel *et al*., 2014).

2.12 Genotypic HIV-1 co-receptor tropism prediction

The entry of viruses into human host cells is dependent on specific protein interactions that occur during the binding of the HIV-1 gp120 protein to the CD4 cellular receptor and co-receptor proteins. The type of co-receptor protein that is used by the virus, usually the CCR5 or CXCR4 chemokine receptors, has a "prognostic value", which has been previously associated with disease progression (Dybowski *et al*., 2010). Patients with viruses that use the CXCR4 co-receptor protein have been linked to faster disease progression and accelerated CD4 cell decline, compared to those with CCR5-tropic viruses (Paraschiv *et al*., 2014; Shepherd *et al*., 2008). Co-receptor usage has also been reported to shift from CCR5 co-receptor usage in the early stages of infection, to CXCR4 usage during the later stages of infection (Esbjörnsson *et al*., 2010).

The identification of phenotypically distinct viruses is therefore crucial in our understanding of HIV pathogenesis and treatment options, and as a result the identification of co-receptor tropism through predictive sequence-based algorithms has improved dramatically in recent years (Schuitemaker *et al*., 2011; Dybowski *et al*., 2010).

For this analysis, co-receptor tropism was predicted from the translated HIV-1 *env* V3-loop sequences of all blood plasma and CVL viruses using three different tools (Table 2.2). Input files complying with the formatting requirements for Geno2pheno and Wetcat were produced using Python v2.7.2 (http://python.org/).

Table 2.2 Summary of the tools and methods used in the analysis of co-receptor tropism prediction. Three alternative tools were used for result comparison, each comprising machine learning and rule-based methods.

Tool	URL	Method/s
Geno2pheno (Sing *et al.*, 2007)	http://coreceptor.bioinf.mpi-inf.mpg.de/index.php	FPR 5% FPR 10%
WebPSSM (Jensen *et al.*, 2003)	http://indra.mullins.microbiol.washington.edu/webpssm/	PSSM
Wetcat (Pillai *et al.*, 2003)	http://genomiac2.ucsd.edu:8080/wetcat/v3.html	SVM ChargeRule

Geno2Pheno has been reported by several studies to have a high sensitivity compared to other prediction methods as it is capable of predicting co-receptor tropism in all HIV-1 genotypes and includes adjustable cut-offs (Simon *et al*., 2010; Crous *et al*., 2012). However different prediction tools have demonstrated varying strengths and weaknesses in identifying R5 and X4 viruses. In a review of genotypic prediction methods by Cheuca *et al*. (2009), the SVM (Support Vector Machine) method showed a high sensitivity to detecting CXCR4 viruses while its specificity was fairly low whereas the PSSM method displayed a higher specificity than sensitivity in detecting CCR5 viruses. For this analysis, the original "g2p" co-receptor tool was used combined with two different false positive rates (Table 2.2). FPR (false positive rate) scores below 10% (i.e. for an analysis with a FPR cutoff of 10%) indicated CXCR4 co-receptor usage, and FPR scores above 10% indicated CCR5 co-receptor usage.

Alternatively, WebPSSM predicts HIV-1 co-receptor usage by using one of two available position-specific scoring matrices (PSSM) that searches for particular motifs on the V3-loop sequence. The selected PSSM uses a null model that serves as a baseline model and consists of sequences with known properties which WebPSSM uses to compare to query sequences (Jensen *et al*., 2003). The comparison generates a score that indicates the likelihood of the query sequence possessing the known property. WebPSSM classifies viruses associated with CXCR4 usage in a matrix containing V3-loop sequences for synctium-inducing viruses (SI), and for CCR5 usage in a matrix containing V3-loop sequences for nonsynctium-inducing viruses with higher scores, implying closer relation to known CXCR4 or CCR5-trophic viruses (Jensen *et al*., 2003).

Because of the difficulty of finding appropriate cut-off values for co-receptor prediction, a quantitative phenotypic prediction by Support Vector Machines (SVM) in Wetcat was also used. This machine learning technique is used for regression problems with many free variables (in this case, sequence positions) and a target variable (resistance factor) subject to considerable noise (Fouchier *et al*., 1992). Lastly, the ChargeRule method is based on statistical analysis of the V3-loop and its phenotypic characteristics, and suggests that the presence of a positively charged residue at position 11 and/or 25 base position confers the ability to bind with CXCR4, however if these conditions are not met, CCR5 co-receptor usage is inferred (Fouchier *et al*., 1992).

2.13 APOBEC-induced hypermutation detection

All blood plasma and CVL sequences were analysed for evidence of APOBEC-induced hypermutation using a combination of both the original and Hypermut v2.0 software tool (http://www.hiv.lanl.gov/) as implemented in the Los Alamos HIV Sequence Database (Rose & Korber, 2000). Hypermut identifies G to A mutations among the many regularly occurring mutations in both a dinucleotide and codon context by comparing every sequence in the alignment to the first sequence while searching for changes between neighbouring locations (Rose & Korber, 2000; Janini, 2001). For this analysis, a consensus sequence was first generated for each intrapatient alignment and saved in FASTA format. The conservative default settings were then used to analyse each intrapatient alignment (data set one; Table 2.1). Results were then exported in tab-delimited format for downstream analysis.

2.14 Length variation within the *env* V-loop regions

Although interactions between the HIV *env* gene and host cells are complex and have not yet been comprehensively described, it is well documented that this gene changes considerably over the course of infection in a single individual (Curlin *et al*., 2010). Insertions and deletions (indels) that are subject to positive or negative selection occur regularly throughout the *env* gene and particularly within the variable loop regions, which subsequently affects V-loop length variation (Sagar *et al*., 2006; Dosenovic *et al*., 2009; Curlin *et al*., 2010). Since variable regions have been reported to influence co-receptor affinity, cellular tropism and sensitivity to neutralizing antibodies, changes within these regions are potentially important markers in increasing our understanding of selective pressures and their role in viral evolution (Brown *et al*., 2011; Curlin *et al*., 2010).

To determine if there were any tissue-specific changes within the *env* V-loop regions longitudinally, intrapatient sequence alignments were divided into subregions V1 – V5 using Se-Al v2.0a11 (data set four; Table 2.1) based on UniProt guidelines for the *env* glycoprotein gp160 region (http://www.uniprot.org/uniprot/P04578) as described in Table 2.3.

Table 2.3 Nucleotide and amino acid base position ranges (relative to HXB2) outlining the V-loop regions that were analysed for length variation in blood plasma and CVL sequences along the course of infection.

HIV-1 *env* V-loop region	Position relative to HXB2	
	Nucleotide (bp)	Amino acid (bp)
V1	391 – 468	131 – 156
V2	469 – 588	157 – 196
V3	886 – 990	296 – 330
V4	1153 – 1254	385 – 418
V5	1381 – 1413	461 – 471

All V-loop sequences were translated from nucleotide to protein bases before the lengths were calculated using a Python v2.7.2 script that counted the number of amino acid bases in each V-loop region per sequence from all participants, while excluding gap regions (Appendix II). V-loop lengths were then statistically tested (section 2.15) to determine if there were any significant region or tissue-specific differences between V-loop lengths in blood plasma and CVL viruses over the sampling periods.

2.15 Significance testing

Non-parametric tests are prescribed when sample sizes of below 100 are available per variable and when nothing about the parameters of interest is known in the population (StatSoft, 2013). Based on methods described in other compartmentalization studies, genotypic and phenotypic differences between viral populations were statistically assessed using the Mann–Whitney U test, which compares medians between two unpaired groups of data and assumes that each population follows the same distribution shape (Evering *et al*., 2014; Ramirez *et al*., 2009; Pillai *et al*., 2006; Ottander *et al*., 1997; Motulsky, 2003). Intrapatient tissue-specific differences were also assessed using the Wilcoxon signed-rank test, which is used to compare differences between two paired groups, assuming all samples are independent of each other and differences are equally distributed around the median per group (Motulsky, 2003; Pillai *et al*., 2006). Since samples were obtained from the blood plasma and cervix independently of each other, associations between V-loop length variation, pairwise genetic distances, and potential N-linked glycosylation sites were assessed using both of these tests as implemented in the Graphpad Prism v6 software package (Parker *et al*., 2008).

Correlation calculations were also performed using Spearman's rank order correlation and the Fisher's Exact test to evaluate tissue-specific differences in phenotypic traits. However, "while it is customary to adjust p-values for multiple comparisons when making positive claims, the uncorrected p-values reported here are conservative with respect to the negative conclusions derived in this study" (Heath *et al.,* 2009), i.e. that compartmentalization does not exist over long-term HIV infection.

Chapter **3**
Results

3.1 Study population and clinical indicators

HIV-1 samples were obtained from the cervix and blood plasma of four female participants sampled anonymously as part of the CAPRISA cohort in Durban, Kwa-Zulu Natal (Table 3.1). All participants involved in this study were heterosexually infected with HIV-1 subtype C and remained treatment naïve throughout the sampling period.

Table 3.1 Overview of participant sequences and time points sampled longitudinally during acute and chronic infection stages. A total of 459 individual HIV-1 *env* sequences were amplified, 206 generated from cervical and 253 from blood plasma-derived samples between 14 and 1316 days post-infection. Time points where it was not possible to amplify viruses are indicated with a "–" symbol.

Participant ID	Age (t_0)	Fiebig stages	Days p.i.	Number of sequences		Total unique sequences
				Blood plasma	CVL	
CAP177	37	I/II; III	14	18	–	5
			28	12	18	13
			196	11	–	11
			378	11	19	27
			560	8	5	9
			924	4	1	3
			1295	18	19	27
CAP217	20	IV; VI	14	18	–	17
			63	20	20	29
			420	6	5	11
			770	6	5	8
			1316	19	19	35
CAP261	18	VI	63	19	19	32
			413	11	12	23
			945	10	10	20
CAP270	24	V	56	11	14	11
			105	–	1	1
			147	8	8	14
			231	14	3	17
			406	12	11	21
			595	7	5	12
			903	10	12	20

Although there were time points at which the *env* gene could not be amplified and sequenced from samples, viral load, CD4 cell counts and STI information was still available for these time points. Further analysis conducted by collaborators at the NICD later confirmed that three of the four female participants studied here were defined as intermediate progressors while the fourth participant, CAP270, was classified as a rapid progressor and subsequently placed on antiretrovirals after 903 days post-infection.

To avoid repetition, please note that all references to sampled time points in this chapter refer to the time in days post-infection.

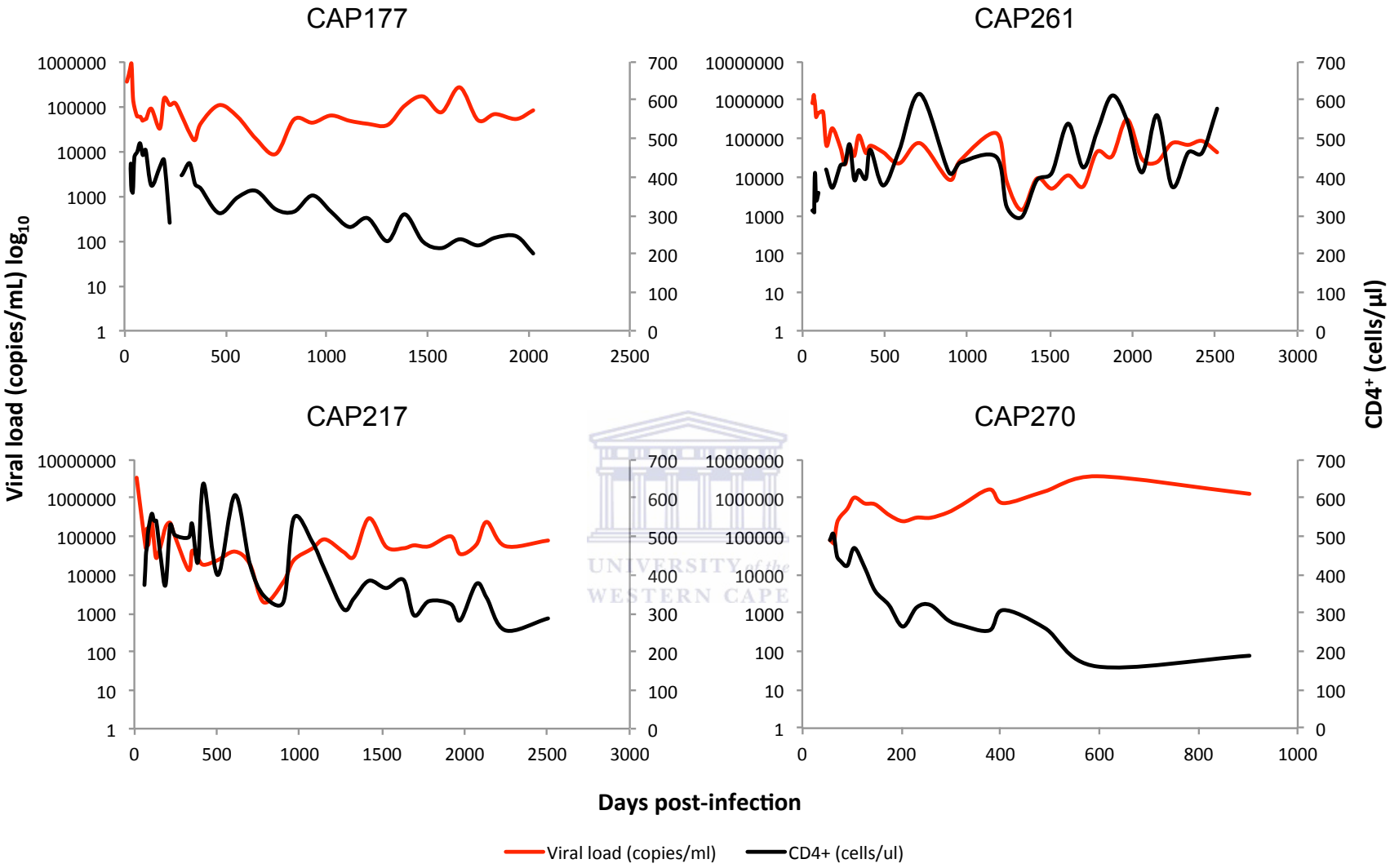


Figure 3.1 Line charts indicating changes in viral load and CD4 cell counts during the sampling period in each participant. Viral load counts are shown on the primary y-axis (red) with CD4 cell counts on the secondary y-axis (black) set at a major limit of 700 cells/μl) in all participants. Breaks in line charts for CAP177 and CAP261 indicate time points where CD4 cell count data was not available, at 252 and 126 days respectively.

In all participants except CAP261, there was a noticeable decline in CD4 cells as the infection progressed. At the baseline, viral loads were considerably higher in CAP217 (3,260,000 copies/mL), CAP261 (821,000 copies/mL) and CAP177 (359,000 copies/mL) compared to CAP270 (76,800 copies/mL), however as the infection progressed viral loads increased substantially (by 98%) to 3,710,000 copies/mL in CAP270 at 595 days, coinciding with the lowest recorded CD4 cell count (161 cells/μl) in this participant over the entire sampling period. Generally, although viral loads and CD4 cell counts appeared to fluctuate in all participants during the course of infection, viral loads remained relatively constant in CAP177, CAP217 and CAP261.

Disease progression was more pronounced in CAP270 than any of the other participants, as evidenced by the rapid rise in viral loads and decline of CD4 cells, eventually leading to this participant being defined as a rapid progressor, however it is unclear what occurred between 595 and 903 days that could have led to the decline in viral loads seen at 903 days, as there were no sequences available between these time points, i.e. the least number of sampled time points were available for CAP270 (n=19), compared to CAP177 (n=36), CAP217 (n=37) and CAP261 (n=36). Overall CAP261 demonstrated the greatest control of HIV infection in terms of CD4 cell count and viral load stability compared to all other participants, having started with a CD4 cell count of 311 cells/μl at 63 days that gradually rose to 578 cells/μl at the final sampling time point (2513 days), whereas CD4 cell counts continued to decline in all other participants.

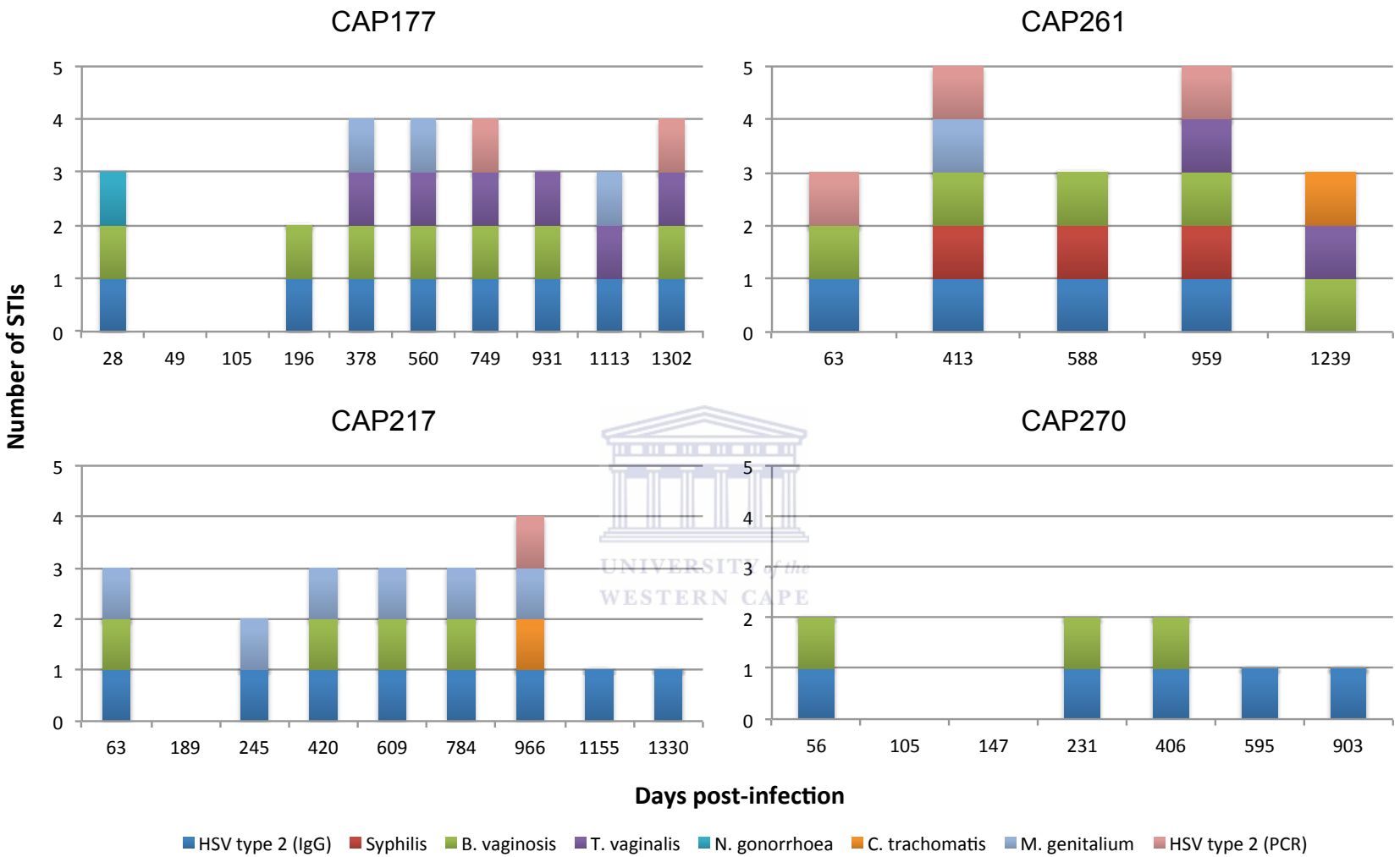


Figure 3.2 Stacked column charts illustrating all sexually transmitted infections that each participant tested positive for during the course of infection, where different colours represent different STIs. Participants were tested for the presence of *HSV type 2* (using two different methods), *Syphilis*, *B. vaginosis*, *T. vaginalis*, *N. gonorrhoea*, *C. trachomatis* and *M. genitalium*. Time points at which participants tested negative for all STIs are indicated by a space on the graph.

Participant CAP177 tested positive for STIs at all but two of the time points sampled, with a consistent *HSV type 2* and *B. Vaginosis* infection present at all time points except 1113 days (Figure 3.2). *T. Vaginalis* and *M. Genitalium* infections appeared at 378 and 749 days, along with the *HSV* and *Vaginosis*, establishing both viral and bacterial infections in CAP177 during almost the entire course of HIV infection.

CAP217 tested positive for STIs at all time points, except 189 days (Figure 3.2). Similar to CAP177, *HSV type 2* infection persisted in CAP217, accompanied by the presence of *M. genitalium* during six out of the nine time points sampled. In addition to this, *B. vaginosis* was detected at four time points, three of which were during consecutive visits (i.e. 420, 609, 784 days). An isolated case of *C. trachomatis* was also found at 966 days, demonstrating simultaneous bacterial and viral infections in this participant too.

CAP261 tested positive for STIs at all time points, with a consistent *B. vaginosis* and *HSV type 2* infection present at all time points except 1239 days (Figure 3.2). Rare cases of *M. genitalium* and *C. trachomatis* infections were also detected at 413 and 1239 days respectively, and a *Syphilis* infection arising around 413 days that remained present until 959 days, after which another new infection (*T. vaginalis*) was found in CAP261.

Alternatively, CAP270 tested positive for only two of the eight STI's that were screened for, i.e. *HSV type 2* and *B. vaginosis* (Figure 3.2). No STI's were found at 105 and 147 days, whereas dual infections were observed at 56, 231 and 406 days. CAP270 thereafter remained infected with *HSV type 2* at 595 and 903 days, while *B. vaginosis* was no longer detectable.

3.2 Phylogenetic evaluation of viral compartmentalization

3.2.1 Nucleotide substitution model selection

For all four HIV-1 subtype C *env* datasets the *general time reversible* nucleotide substitution) model with *gamma* distributed rate variation 4 (GTR + G_4) was identified as the best-fit nucleotide substitution model using ModelTest (Posada & Crandall, 1998).

3.2.2 Evolutionary model selection

Bayes factor tests (Kass & Rafferty, 1995), based on the ratio of the marginal likelihoods of the alternative models provided by BEAST, revealed that an uncorrelated lognormal relaxed-clock, constant population size model provided the best fit to the data for three of the four alignments tested. For sequences obtained from participant CAP217, a lognormal relaxed-clock Bayesian skyline plot model was identified as the best fit.

3.2.3 Bayesian phylogenetic reconstruction

To determine if the female genital tract represented a distinct compartment during acute and chronic HIV infection, *env* sequences from all sampled time points were analysed per participant. It was hypothesized that if the female genital tract did serve as a distinct viral compartment, CVL-derived sequences should have clustered together as monophyletic clades on trees with high statistical support, independent from blood plasma sequences at multiple time points over the course of infection.

On time-scaled maximum clade credibility (MCC) trees, viral sequences generally clustered together by sampling time point and displayed substantial divergence between time points (Figures 3.3a-d). A few exceptions to this pattern were seen in CAP177, CAP217 and CAP270. In participant CAP177 blood plasma sequences from time points at 196 and 560 days clustered together with ≥70% posterior support. In participant CAP217 genital tract sequences from successive sampling points at 770 and 1316 days clustered together on the tree with ≥90% posterior support while for participant CAP270 blood plasma sequences sampled at 147 and 231 days clustered together on MCC trees with <70% posterior support.

In participant CAP177, strong evidence of a second highly divergent low frequency HIV variant (*96-cvl-28days*) was found in the CVL at 28 days.

This sequence was distinctly separated from all other blood plasma and CVL sequences obtained from this participant during the early stages of infection by an extremely long branch length, more clearly illustrated on the cross-sectional tree for this time point (Appendix III, Figure 3.4a). Furthermore it was clear that a closely related variant of *96-cvl-28days* established a systemic infection that yielded at least two main genetically distinct lineages (>90% posterior probability) with closely related variants in both the blood plasma and CVL that persisted until at least 560 days (Figure 3.3a). While the other main lineage persisted and continued to diversify up until the final sampling point at 1295 days, the descendant members of the lineage founded by the putative ancestor of *96-cvl-28days* was no longer detectable in either tissue type. It was also apparent that by 28 days, the main lineage that persisted until 1295 days also initially comprised of more than one viral variant (Figure 3.3a).

Similar patterns were observed in participant CAP261, where at 63 days (the earliest sampling point in this participant) the presence of at least two distinct HIV variants forming separate well-supported clades was clearly evident (Appendix III, Figure 3.4b) and descendants from only one of the highly divergent founding lineages persisted through to the final sampling point at 945 days (Figure 3.3b). Phylogenetic analyses for CAP217 and CAP270 revealed an interspersion of blood plasma and CVL sequences on both trees (Figures 3.3c and 3.3d) and sequences generally clustered together by time point instead of tissue type in a ladder-like topology, indicative of infection by a single HIV variant (Appendix III, Figures 3.4c and 3.4d) (Novitsky *et al*., 2011).

In all participants, monophyletic clades generally contained few sequences and were mainly interspersed among sequences from both tissue-types, similar to patterns observed by Bull *et al*. (2013). MCC tree topologies differed substantially between inferred single and multiple HIV variant infected participants. For participants CAP217 and CAP270, sequences yielded trees with classic ladder-like topologies and poorly supported short interior branches, characteristic of patterns generated by a single variant transmission event and subsequent diversification (Keele *et al*., 2008). In contrast, MCC trees generated from CAP177 and CAP261 sequences displayed long, well supported internal branches sprouting from the basal nodes, a pattern characteristic of an initial transmission event involving more than one viral variant (Keele *et al*., 2008; Novitsky *et al*., 2011).

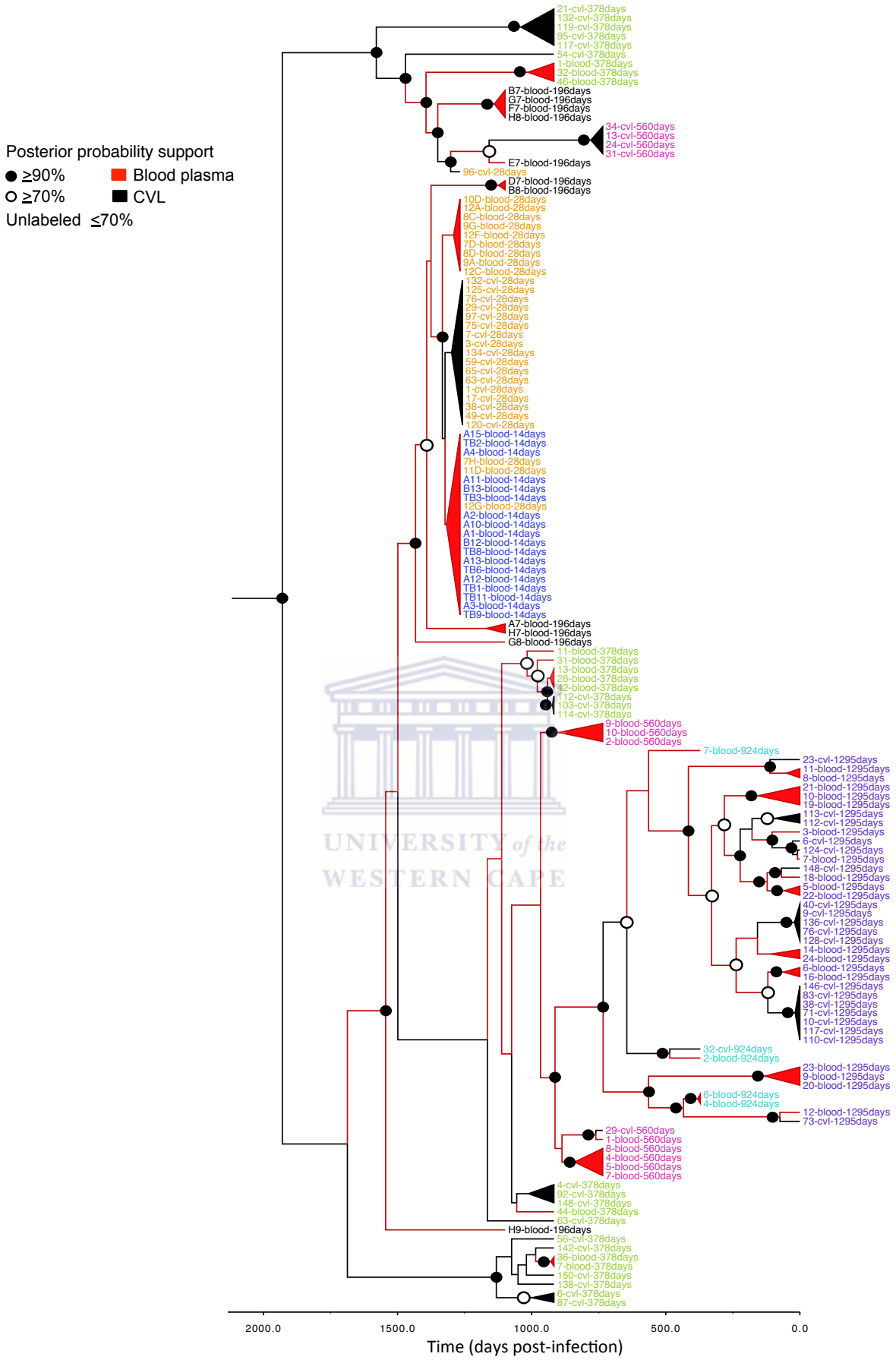


Figure 3.3a Time-scaled Bayesian maximum clade credibility tree for CAP177, constructed under the GTR + G_4 substitution model and a constant population size relaxed-clock evolutionary model. Branches are coloured according to the most probable state of their tissue origin where red represents viruses from the blood plasma and black indicates viruses from the cervix. Posterior probabilities ≥ 90% are indicated by a filled circled and ≥ 70% by an open circle at the nodes, with branch labels coloured according to the time points sampled.

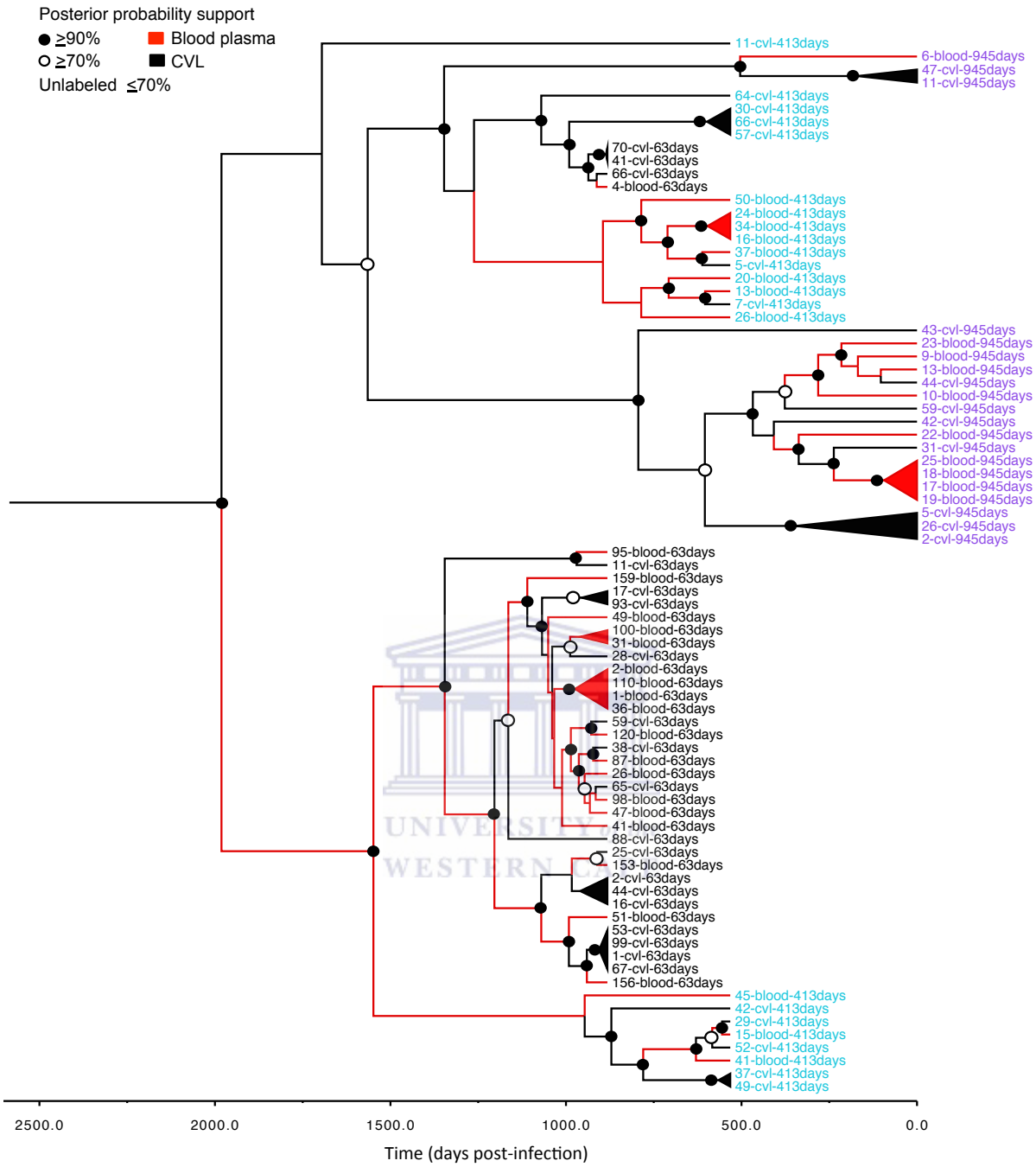


Figure 3.3b Time-scaled Bayesian maximum clade credibility tree for CAP261, constructed under the GTR + G_4 substitution model and a constant population size relaxed-clock evolutionary model. Branches are coloured according to the most probable state of their tissue origin where red represents viruses from the blood plasma and black indicates viruses from the cervix. Posterior probabilities $\geq$ 90% are indicated by a filled circled and $\geq$ 70% by an open circle at the nodes, with branch labels coloured according to the time points sampled.

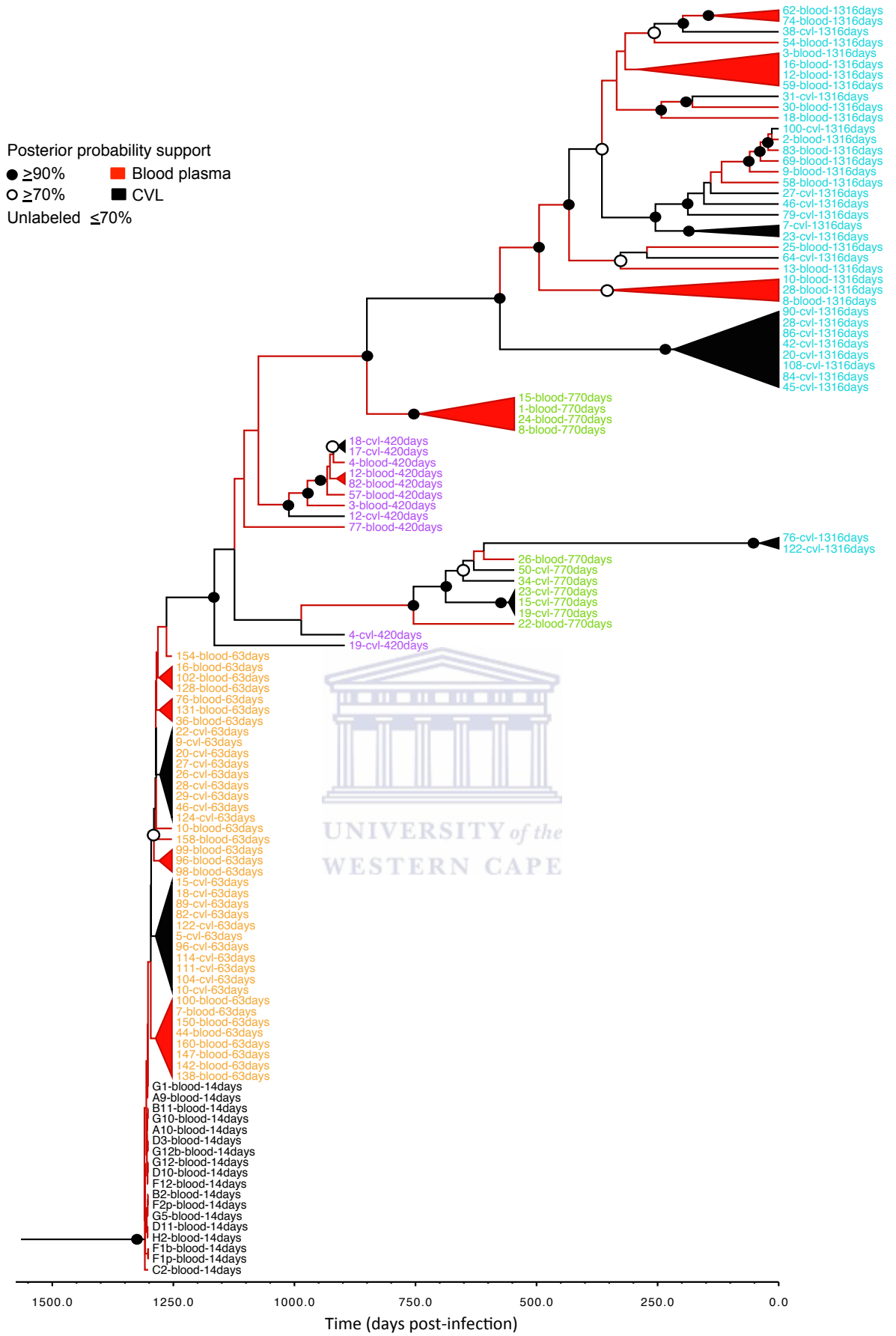


Figure 3.3c Time-scaled Bayesian maximum clade credibility tree for CAP217, constructed under the GTR + G_4 substitution model and a Bayesian skyline plot relaxed-clock evolutionary model. Branches are coloured according to the most probable state of their tissue origin where red represents viruses from the blood plasma and black indicates viruses from the cervix. Posterior probabilities ≥ 90% are indicated by a filled circled and ≥ 70% by an open circle at the nodes, with branch labels coloured according to the time points sampled.

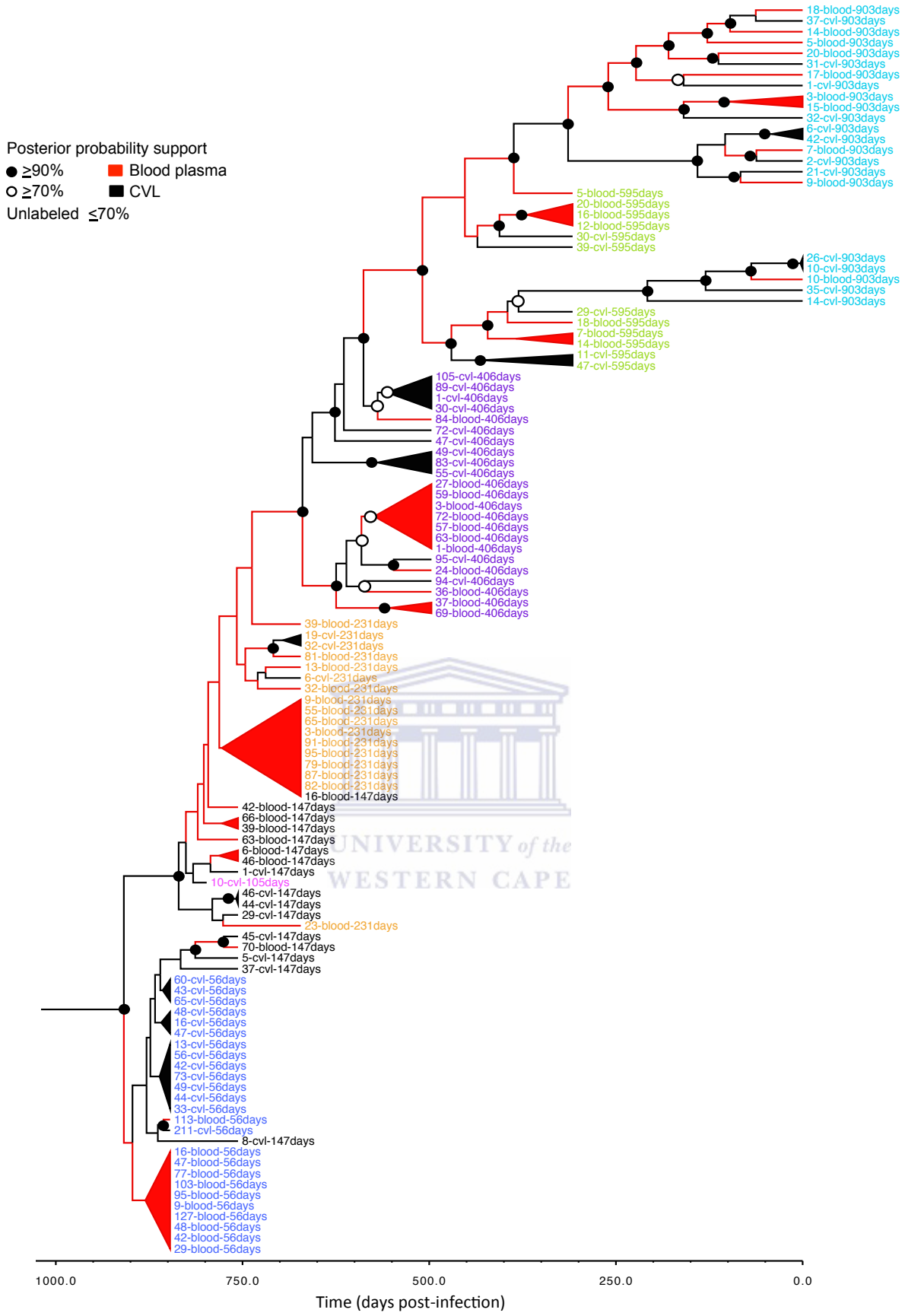


Figure 3.3d Time-scaled Bayesian maximum clade credibility tree for CAP270, constructed under the GTR + G_4 substitution model and a constant population size relaxed-clock evolutionary model. Branches are coloured according to the most probable state of their tissue origin where red represents viruses from the blood plasma and black indicates viruses from the cervix. Posterior probabilities ≥ 90% are indicated by a filled circled and ≥ 70% by an open circle at the nodes, with branch labels coloured according to the time points sampled.

The highest mean nucleotide substitution rates were estimated in participants CAP177 (3.09 x 10^{-2} substitutions per site per year; 95% HPDs = 2.44 x 10^{-2} – 3.77 x 10^{-2}) and CAP270 (2.20 x 10^{-2} substitutions per site per year; 95% HPDs = 1.77 x 10^{-2} – 2.67 x 10^{-2}), while the lowest substitution rates were estimated in CAP217 (9.58 x 10^{-3} substitutions per site per year; 95% HPDs = 7.95 x 10^{-3} – 1.13 x 10^{-2}) and CAP261 (1.09 x 10^{-2} substitutions per site per year; 95% HPDs = 7.26 x 10^{-3} – 1.46 x 10^{-2}). All rates were within the range of previously reported rates of intrahost HIV evolution (Novitsky *et al*., 2013; Carvajal-Rodríguez *et al*., 2008; Aulicino *et al*., 2011; Lemey *et al*., 2007; Lukashov & Goudsmit, 1997). Overall, considering *env* sequence datasets derived from all four participants collectively, approximately 74 tissue-specific clades were identified longitudinally, comprised of highly similar sequences with mean pairwise genetic diversities ranging between 0% in CAP217 and 2.30% in CAP177 on MCC trees.

3.2.4 Estimation of the time to the Most Recent Common Ancestor

Estimates of the mean time to the most recent common ancestor (tMRCA) for participants CAP217 (mean = 1311 days; 95% HPDs = 1303–1324 days) and CAP270 (mean = 914 days; 95% HPDs = 878–964 days) accurately reflected the known time of infection as estimated using laboratory staging (Table 3.2), however for CAP177 (mean = 1977 days; 95% HPDs = 1590–2454 days) and CAP261 (mean = 2029 days; 95% HPDs = 1482–2687 days) the mean tMRCA's predated the known time of infection by 53% and 115% respectively, a pattern which is characteristic of an infection founded by more than one HIV-1 variant (Novitsky *et al*., 2011; Sturdevant *et al*., 2012).

Table 3.2 TMRCA and known time post-infection comparisons in each participant over chronic and acute infection stages. The known time of infection is shown in blue text with the estimated mean tMRCA's shown in red text. All values quoted below represent the time in days post-infection.

Statistic	CAP177	CAP261	CAP217	CAP270
Known time post-infection	1295	945	1316	903
Mean tMRCA	1977	2029	1311	914
Std error of mean	5.97	9.60	0.20	0.32
Median	1930	1981	1309	909
Geometric mean	1963	2005	1311	914
95% HPD lower	1590	1482	1303	878
95% HPD upper	2454	2687	1324	964

It is notable that the lower and upper 95% HPD parameters for participants CAP217 and CAP270 were found to be within close proximity to the known time of infection, whereas in participants CAP177 and CAP261 they were not. Furthermore, based on rigorous methods followed by Novitsky *et al*. (2011) in the identification of single and multiple HIV variants in which the geometric means of the tMRCA estimate was interpreted, transmission of multiple variants in participants CAP177 and CAP261 were evident.

3.3 Statistical evaluation of viral compartmentalization and population migration

3.3.1 Tree-based compartmentalization analysis

In phylogenetic tree-based statistical analyses performed using BaTS (Parker *et al*., 2008) on longitudinally sampled data from each participant, the null hypothesis of panmixis (i.e. free movement of viruses) between the blood plasma and cervical compartments was rejected in three of the four participants by all tests applied (Table 3.3). Overall, significantly lower than expected values were obtained for the PS and AI statistics and significantly higher than expected values for the MC (Blood) statistic, given a model of random movement of viruses between compartments (i.e. a movement model with unrestricted migration between compartments). In participants CAP177 and CAP217, the MC (CVL) statistic was also significantly higher than expected while in participant CAP261, only the PS statistic was significantly lower than that expected (Table 3.3).

When monotypic (identical) sequences were removed from longitudinal data sets for CAP261 (n = 6) and CAP270 (n = 14), almost no evidence of statistically significant structure by tissue type remained however when monotypic sequences were similarly removed from CAP177 (n = 49) and CAP217 (n = 32) data sets, the PS and AI statistics remained significantly different from null expectations (Table 3.3).

The removal of both monotypic and low diversity (<0.1% divergent) sequences (such that low diversity, tissue-specific clades were collapsed to a single sequence) in CAP177 (n = 28) and CAP217 (n = 22) made almost no difference compared to the structure detected when only monotypic sequences were excluded, with the exception of statistically significant structure becoming visible in the MC (CVL) in CAP177 (Table 3.3). To further investigate if any of the significant evidence for compartmentalization was influenced by sequences from mid or later stages of infection in each participant, data sets were separated by time point and analysed independently.

Table 3.3 Longitudinal compartmentalization results using tree-based statistics AI, PS and MC in BaTS for all participants. Data set types are indicated by roman numerals I, II and III (where I = monotypic and low diversity sequences included; II = monotypic sequences removed; III = monotypic and low diversity (<0.01% divergence) sequences removed), with significant p-vlaues shown in red text.

Participant ID	Data set	n (Blood)	n (CVL)	Statistic	Observed mean	Lower 95% CI	Upper 95% CI	Null mean	Lower 95% CI	Upper 95% CI	p-value
CAP177	I	68	62	AI	2.50	1.81	3.23	6.25	5.21	7.30	0.0000
				PS	23.64	21.00	26.00	42.08	37.94	45.92	0.0000
				MC (Blood)	7.16	4.00	9.00	4.30	3.20	5.94	0.0320
				MC (CVL)	7.22	7.00	9.00	3.25	2.41	4.34	0.0050
	II	60	42	AI	2.86	2.26	3.52	4.85	3.84	5.85	0.0010
				PS	23.31	21.00	26.00	32.24	28.31	35.73	0.0000
				MC (Blood)	3.94	3.00	4.00	4.15	3.09	5.68	0.5400
				MC (CVL)	4.08	4.00	5.00	3.00	2.09	4.16	0.1380
	III	50	33	AI	2.74	2.28	3.25	3.78	2.89	4.70	0.0190
				PS	20.99	19.00	23.00	25.88	22.59	28.90	0.0090
				MC (Blood)	3.86	3.00	4.00	4.04	3.00	5.95	0.4330
				MC (CVL)	3.64	3.00	4.00	2.73	2.00	4.02	0.0740
CAP217	I	66	49	AI	3.20	2.26	4.16	5.22	4.30	6.11	0.0000
				PS	30.00	27.00	33.00	37.10	33.78	40.14	0.0010
				MC (Blood)	4.74	4.00	7.00	4.38	3.41	6.64	0.5670
				MC (CVL)	8.00	8.00	8.00	3.14	2.41	4.08	0.0020
	II	52	38	AI	2.68	1.900	3.48	4.14	3.27	5.02	0.0000
				PS	24.11	22.00	27.00	28.81	25.48	31.75	0.0100
				MC (Blood)	4.54	4.00	7.00	4.19	3.11	6.21	0.4710
				MC (CVL)	6.01	6.00	6.00	2.96	2.17	4.05	0.0150
	III	45	29	AI	2.25	1.61	2.91	3.32	2.52	4.10	0.0190
				PS	19.13	17.00	21.00	22.88	19.83	25.61	0.0180
				MC (Blood)	4.20	4.00	5.00	4.36	3.02	6.47	0.5530
				MC (CVL)	5.67	5.00	6.00	2.68	1.97	4.01	0.0070

Participant ID	Data set	n (Blood)	n (CVL)	Statistic	Observed mean	Lower 95% CI	Upper 95% CI	Null mean	Lower 95% CI	Upper 95% CI	p-value
CAP261	I	40	41	AI	2.81	2.31	3.32	3.55	2.51	4.47	0.1100
				PS	22.39	21.00	23.00	26.90	23.00	30.48	0.0200
				MC (Blood)	4.00	4.00	4.00	3.25	2.08	5.77	0.2360
				MC (CVL)	4.00	4.00	4.00	3.31	2.17	5.52	0.2400
	II	40	37	AI	3.36	2.78	3.92	3.56	2.57	4.56	0.3890
				PS	23.31	22.00	24.00	25.64	21.94	29.15	0.1200
				MC (Blood)	4.00	4.00	4.00	3.42	2.13	5.97	0.2350
				MC (CVL)	3.02	3.00	3.00	3.19	2.02	5.03	0.6110
CAP270	I	62	57	AI	4.61	3.79	5.45	5.73	4.65	6.78	0.0440
				PS	34.36	32.00	37.00	39.46	35.53	43.42	0.0120
				MC (Blood)	6.68	6.00	7.00	3.94	2.88	5.54	0.0080
				MC (CVL)	3.84	3.00	5.00	3.65	2.59	5.11	0.3040
	II	61	46	AI	4.54	3.76	5.35	5.05	4.07	6.03	0.1910
				PS	31.18	29.00	34.00	34.52	30.72	37.91	0.0600
				MC (Blood)	6.44	9.00	9.00	4.32	3.17	5.80	0.0360
				MC (CVL)	3.52	3.00	5.00	3.19	2.23	4.65	0.5880

In cross-sectional analyses where the data for each time point was treated as fully independent and analysed separately, there was strong evidence of compartmentalization at all sampling time points in HIV sequences derived from participant CAP177 whereas, in CAP261 and CAP217, statistically significant structure by tissue type was only detected at the final sampling points (945 and 1316 days respectively, Table 3.4). In participant CAP270 the only statistics that were significantly different to null expectations were MC (Blood) at 147 days, and MC (CVL) and MC (Blood) at 406 days (Table 3.4)

The removal of monotypic sequences from CAP177 led to a reduction in the number of significant results obtained at 28 and 378 days but not at 560 days, and completely removed all statistical evidence for compartmentalization at the final sampling point taken at 1295 days (Table 3.4), whereas in CAP217 and CAP261, removal of monotypic sequences had no effect on the results obtained and significant structure by tissue type remained detectable at 1316 and 945 days respectively.

In contrast, the removal of both monotypic and low diversity sequences in CAP217, resulted in compartmentalization no longer being detectable at any of the time points containing low diversity sequences (i.e. 63 and 420 days), whereas in participant CAP177 statistically significant structure was still detectable at 28 days within CVL populations (Table 3.4).

Table 3.4 Cross-sectional compartmentalization results using tree-based statistics AI, PS and MC in BaTS for all participants. Data set types are indicated by roman numerals I, II and III (where I = monotypic and low diversity sequences included; II = monotypic sequences removed; III = monotypic and low diversity (<0.01% divergence) sequences removed), with significant p-vlaues shown in red text.

Participant ID	Data set	Days p.i.	n (Blood)	n (CVL)	Statistic	Observed mean	Lower 95% CI	Upper 95% CI	Null mean	Lower 95% CI	Upper 95% CI	p-value
CAP177	I	28	12	18	AI	0.22	0.00	0.48	1.32	1.00	1.57	0.0000
					PS	2.94	2.00	3.00	9.34	8.02	10.35	0.0000
					MC (Blood)	4.31	2.00	8.00	2.00	1.65	2.48	0.0180
					MC (CVL)	0.22	0.00	0.48	1.32	1.00	1.57	0.0000
		378	11	19	AI	0.53	0.15	0.94	1.05	0.64	1.44	0.0140
					PS	7.35	6.00	8.00	8.93	7.06	10.42	0.0460
					MC (Blood)	2.92	2.00	3.00	1.70	1.12	2.89	0.0220
					MC (CVL)	4.00	4.00	4.00	3.31	1.84	6.00	0.2750
		560	8	5	AI	0.27	0.27	0.27	0.53	0.25	0.86	0.1870
					PS	2.00	2.00	2.00	3.88	2.80	4.81	0.0140
					MC (Blood)	2.00	2.00	2.00	2.13	1.19	4.00	0.6770
					MC (CVL)	4.00	4.00	4.00	1.46	1.00	2.00	0.0090
		1295	18	19	AI	1.06	0.83	1.19	1.99	1.34	2.62	0.0070
					PS	8.30	8.00	9.00	12.26	9.93	14.34	0.0020
					MC (Blood)	4.01	3.00	5.00	2.74	1.98	4.06	0.0660
					MC (CVL)	7.00	7.00	7.00	2.91	2.03	4.54	0.0080
	II	28	12	18	AI	0.22	0.00	0.48	1.32	1.00	1.57	0.0000
					PS	2.94	2.00	3.00	9.34	8.02	10.35	0.0000
					MC (Blood)	4.31	2.00	8.00	2.00	1.65	2.48	0.0180
					MC (CVL)	0.22	0.00	0.48	1.32	1.00	1.57	0.0000
		378	11	17	AI	0.39	0.10	0.68	1.00	0.58	1.38	0.0130
					PS	6.04	6.00	7.00	8.63	6.80	10.31	0.0230
					MC (Blood)	2.94	3.00	3.00	1.82	1.10	3.33	0.0690
					MC (CVL)	4.00	4.00	4.00	3.10	1.65	5.01	0.2970
		560	7	4	AI	0.30	0.30	0.30	0.51	0.05	0.83	0.2310
					PS	2.00	2.00	2.00	3.14	2.00	4.00	0.0820
					MC (Blood)	2.00	2.00	2.00	2.15	1.32	4.00	0.7490
					MC (CVL)	3.00	3.00	3.00	1.33	1.00	2.00	0.0200

Participant ID	Data set	Days p.i.	n (Blood)	n (CVL)	Statistic	Observed mean	Lower 95% CI	Upper 95% CI	Null mean	Lower 95% CI	Upper 95% CI	p-value
		1295	18	11	AI	1.09	0.84	1.31	1.50	0.87	2.08	0.1250
					PS	8.32	8.00	9.00	8.70	6.94	10.26	0.2460
					MC (Blood)	3.73	3.00	5.00	3.25	2.00	4.93	0.7030
					MC (CVL)	3.00	3.00	3.00	2.03	1.07	3.20	0.1620
	III	28	3	5	AI	0.21	0.01	0.38	0.35	0.21	0.40	0.1060
					PS	1.89	1.00	2.00	2.28	1.88	2.53	0.3790
					MC (Blood)	1.00	1.00	1.00	1.20	1.00	1.39	1.0000
					MC (CVL)	2.78	1.00	5.00	1.80	1.38	2.77	0.0010
		378	9	14	AI	0.55	0.44	0.79	0.77	0.34	1.20	0.1320
					PS	6.06	6.00	7.00	7.03	5.28	8.59	0.1370
					MC (Blood)	2.89	2.00	3.00	1.63	1.00	3.12	0.0600
					MC (CVL)	2.64	2.00	3.00	2.75	1.24	4.64	0.4220
		1295	18	9	AI	1.13	0.89	1.35	1.29	0.69	1.85	0.3260
					PS	8.15	8.00	9.00	7.46	5.87	8.93	0.7540
					MC (Blood)	3.90	3.00	5.00	3.53	2.05	5.30	0.7520
					MC (CVL)	1.75	1.00	2.00	1.75	1.00	3.00	0.2780
CAP217	I	63	20	20	AI	1.23	0.55	1.93	1.91	1.44	2.29	0.0120
					PS	11.55	9.00	14.00	13.24	11.48	14.56	0.1080
					MC (Blood)	3.55	3.00	6.00	2.75	2.25	3.50	0.2270
					MC (CVL)	3.17	2.00	5.00	2.76	2.25	3.49	0.2350
		420	6	5	AI	0.31	0.15	0.55	0.53	0.15	0.80	0.0930
					PS	3.68	3.00	4.00	3.51	2.00	4.59	0.8810
					MC (Blood)	2.11	2.00	3.00	1.85	1.05	3.04	0.3730
					MC (CVL)	1.87	1.00	2.00	1.51	1.00	3.00	0.1950
		770	6	5	AI	0.40	0.10	0.57	0.47	0.12	0.66	0.4070
					PS	2.81	2.00	3.00	3.55	2.15	4.63	0.2510
					MC (Blood)	1.00	1.00	1.00	1.74	1.00	3.04	1.0000
					MC (CVL)	3.00	3.00	3.00	1.48	1.00	3.00	0.0480
		1316	19	19	AI	1.06	0.81	1.41	1.92	1.23	2.58	0.0230
					PS	8.32	7.00	9.00	12.67	10.19	15.00	0.0030
					MC (Blood)	3.34	2.00	4.00	2.71	1.91	4.23	0.0530
					MC (CVL)	10.00	10.00	10.00	2.76	1.93	4.93	0.0020

Participant ID	Data set	Days p.i.	n (Blood)	n (CVL)	Statistic	Observed mean	Lower 95% CI	Upper 95% CI	Null mean	Lower 95% CI	Upper 95% CI	p-value
	II	63	19	13	AI	0.81	0.29	1.36	1.52	1.05	1.91	0.0050
					PS	8.01	6.00	10.00	10.03	8.61	11.33	0.0190
					MC (Blood)	4.38	3.00	8.00	3.11	2.50	4.02	0.0520
					MC (CVL)	2.88	2.00	5.00	2.15	1.59	3.17	0.0680
		770	6	3	AI	0.47	0.41	0.61	0.27	0.01	0.47	0.7530
					PS	2.82	2.00	3.00	2.39	1.87	3.00	1.0000
					MC (Blood)	1.00	1.00	1.00	1.89	1.00	5.00	1.0000
					MC (CVL)	1.01	1.00	1.00	1.11	1.00	1.80	1.0000
		1316	19	17	AI	1.06	0.81	1.40	1.81	1.15	2.51	0.0320
					PS	8.30	7.00	9.00	11.86	9.32	14.21	0.0100
					MC (Blood)	3.35	2.00	4.00	2.84	1.99	5.02	0.0940
					MC (CVL)	8.00	8.00	8.00	2.55	1.62	4.98	0.0020
	III	63	14	5	AI	0.56	0.11	1.02	0.71	0.44	1.01	0.2200
					PS	4.48	3.00	5.00	4.35	3.54	4.92	1.0000
					MC (Blood)	3.63	2.00	6.00	3.64	2.66	7.42	0.8120
					MC (CVL)	1.34	1.00	2.00	1.31	1.05	2.08	1.0000
		420	5	4	AI	0.21	0.15	0.47	0.44	0.05	0.69	0.0910
					PS	3.00	3.00	3.00	2.81	2.00	3.95	0.6930
					MC (Blood)	1.99	2.00	2.00	1.69	1.00	3.00	0.2100
					MC (CVL)	1.85	1.00	2.00	1.43	1.00	2.33	0.0940
CAP261	I	63	19	19	AI	1.69	1.24	2.05	1.77	1.13	2.38	0.4500
					PS	12.09	11.00	13.00	12.62	10.07	14.82	0.3120
					MC (Blood)	2.59	2.00	3.00	2.57	1.98	4.09	0.1420
					MC (CVL)	3.99	4.00	4.00	2.56	1.98	4.09	0.0680
		413	11	12	AI	0.82	0.75	0.99	1.05	0.51	1.52	0.1840
					PS	6.99	7.00	7.00	7.51	5.21	9.10	0.3600
					MC (Blood)	3.00	3.00	3.00	2.20	1.05	4.00	0.2110
					MC (CVL)	4.00	4.00	4.00	2.45	1.10	4.01	0.1200
		945	10	10	AI	0.38	0.32	0.41	0.94	0.40	1.47	0.0510
					PS	3.95	4.00	4.00	6.60	4.95	8.04	0.0100
					MC (Blood)	4.00	4.00	4.00	2.17	1.00	4.00	0.0950
					MC (CVL)	4.99	5.00	5.00	2.18	1.00	4.00	0.0160

Participant ID	Data set	Days p.i.	n (Blood)	n (CVL)	Statistic	Observed mean	Lower 95% CI	Upper 95% CI	Null mean	Lower 95% CI	Upper 95% CI	p-value
	II	63	19	15	AI	2.12	1.68	2.47	1.61	0.96	2.25	0.9180
					PS	12.40	11.00	13.00	11.15	8.97	13.09	0.9420
					MC (Blood)	2.53	2.00	3.00	2.78	2.00	4.19	0.3450
					MC (CVL)	2.17	2.00	3.00	2.25	1.25	3.39	0.7310
CAP270	I	56	11	14	AI	0.99	0.50	1.50	1.18	0.76	1.58	0.1580
					PS	7.54	6.00	9.00	8.09	7.07	9.14	0.5210
					MC (Blood)	2.25	2.00	3.00	2.09	1.59	2.55	0.5280
					MC (CVL)	3.04	2.00	5.00	2.64	2.00	3.08	0.1510
		147	8	8	AI	0.46	0.38	0.63	0.73	0.28	1.16	0.1020
					PS	3.00	3.00	3.00	5.20	3.66	6.91	0.0150
					MC (Blood)	4.34	4.00	5.00	2.03	1.01	3.87	0.0250
					MC (CVL)	2.33	2.00	3.00	2.03	1.01	3.99	0.5200
		231	14	3	AI	0.44	0.15	0.68	0.45	0.14	0.74	0.4780
					PS	2.77	2.00	3.00	2.70	2.00	3.00	1.0000
					MC (Blood)	3.07	2.00	4.00	4.34	2.68	8.97	0.8330
					MC (CVL)	1.14	1.00	2.00	1.10	1.00	1.69	1.0000
		406	12	11	AI	0.80	0.57	0.94	1.21	0.66	1.75	0.1550
					PS	5.04	5.00	5.00	7.52	5.63	9.30	0.0190
					MC (Blood)	6.00	6.00	6.00	2.50	1.57	4.30	0.0120
					MC (CVL)	3.90	3.00	4.00	2.27	1.30	3.95	0.0490
		595	7	5	AI	0.46	0.07	0.65	0.52	0.20	0.82	0.2580
					PS	3.89	3.00	4.00	3.76	2.66	4.87	0.7040
					MC (Blood)	1.78	1.00	3.00	1.97	1.09	3.97	1.0000
					MC (CVL)	1.98	2.00	2.00	1.44	1.00	2.21	0.0770
		903	10	12	AI	1.00	0.82	1.31	1.01	0.49	1.54	0.5990
					PS	7.86	7.00	8.00	7.19	5.13	8.96	0.7700
					MC (Blood)	1.47	1.00	2.00	2.03	1.00	4.00	1.0000
					MC (CVL)	2.00	2.00	2.00	2.48	1.12	5.00	0.6250
	II	56	11	8	AI	0.89	0.43	1.36	0.87	0.57	1.16	0.5420
					PS	6.25	5.00	8.00	6.01	5.08	6.78	0.5200
					MC (Blood)	2.44	2.00	4.00	2.47	1.85	3.13	0.9100
					MC (CVL)	1.84	1.00	3.00	1.80	1.39	2.36	0.3200

Participant ID	Data set	Days p.i.	n (Blood)	n (CVL)	Statistic	Observed mean	Lower 95% CI	Upper 95% CI	Null mean	Lower 95% CI	Upper 95% CI	p-value
		147	8	7	AI	0.47	0.38	0.64	0.64	0.27	0.99	0.1530
					PS	3.00	3.00	3.00	4.87	3.45	6.00	0.0260
					MC (Blood)	4.32	4.00	5.00	2.03	1.02	3.99	0.0430
					MC (CVL)	1.36	1.00	2.00	1.71	1.01	3.87	1.0000
		406	11	11	AI	0.80	0.57	0.92	1.19	0.63	1.75	0.1740
					PS	5.03	5.00	5.00	7.30	5.17	8.94	0.0230
					MC (Blood)	5.00	5.00	5.00	2.38	1.36	3.96	0.0290
					MC (CVL)	3.90	3.00	4.00	2.32	1.34	3.96	0.0460
		903	10	11	AI	1.25	1.06	1.56	0.99	0.47	1.49	0.8530
					PS	7.87	7.00	8.00	6.92	5.04	8.67	0.8540
					MC (Blood)	1.47	1.00	2.00	2.09	1.01	4.00	1.0000
					MC (CVL)	2.00	2.00	2.00	2.37	1.11	4.70	0.5820

3.3.2 Distance-based compartmentalization analysis

To further test the hypothesis that HIV-1 is compartmentalized in the cervix or blood plasma, distance-based tests were also performed on gp120 *env* sequences from each participant using Wright's measure of population subdivision (F_{ST}) and Hudson's Nearest-Neighbor (Snn) statistics.

Table 3.5 Longitudinal compartmentalization results using distance-based statistics F_{ST} and Snn in HyPhy in all participants. Data set types are indicated by roman numerals I, II, and III (where I = monotypic and low diversity sequences included; II = monotypic sequences removed; III = monotypic and low diversity (<0.01% divergence) sequences removed). Statistically significant evidence of compartmentalization for the F_{ST} test was determined based on guidelines described by Josefsson *et al*. (2013), where p-value <0.05 and bootstrap value >0.95, whereas for the Snn test p-values below 0.05 were considered to be significant. Significant p-vlaues are shown in red text.

Participant ID	Data set	n (Blood)	n (CVL)	Statistic	Observed value	p-value	Bootstrap mean	Bootstrap median	Bootstrap standard deviation	Bootstrap 95% CI
CAP177	I	68	62	F_{ST}	0.013	0.049	0.019	0.017	0.011	0.005 – 0.054
				Snn	0.743	0.008	0.845	0.845	0.028	0.791 – 0.907
	II	60	42	F_{ST}	0.001	0.314	0.011	0.009	0.007	0.001 – 0.025
				Snn	0.715	0.010	0.864	0.864	0.033	0.797 – 0.924
	III	50	33	F_{ST}	-0.001	0.458	0.011	0.010	0.007	0.001 – 0.028
				Snn	0.687	0.003	0.857	0.855	0.037	0.783 – 0.928
CAP217	I	66	49	F_{ST}	0.009	0.196	0.018	0.016	0.015	-0.004 – 0.064
				Snn	0.654	0.009	0.847	0.844	0.030	0.800 – 0.919
	II	52	38	F_{ST}	-0.000	0.361	0.011	0.009	0.010	-0.003 – 0.036
				Snn	0.583	0.108	0.843	0.845	0.038	0.766 – 0.917
	III	45	29	F_{ST}	0.006	0.181	0.019	0.017	0.013	0.001 – 0.054
				Snn	0.566	0.261	0.852	0.852	0.043	0.764 – 0.932
CAP261	I	40	41	F_{ST}	0.001	0.308	0.014	0.012	0.008	0.003 – 0.032
				Snn	0.584	0.145	0.839	0.840	0.041	0.761 – 0.916
	II	40	37	F_{ST}	0.004	0.197	0.017	0.015	0.010	0.003 – 0.043
				Snn	0.498	0.564	0.827	0.831	0.044	0.738 – 0.909

Participant ID	Data set	n (Blood)	n (CVL)	Statistic	Observed value	p-value	Bootstrap mean	Bootstrap median	Bootstrap standard deviation	Bootstrap 95% CI
CAP270	I	62	57	F_{ST}	0.004	0.149	0.013	0.012	0.006	0.003 – 0.029
				Snn	0.560	0.154	0.840	0.841	0.033	0.777 – 0.902
	II	61	46	F_{ST}	0.006	0.162	0.015	0.014	0.009	0.000 – 0.039
				Snn	0.522	0.456	0.822	0.822	0.036	0.755 – 0.894

In distance-based compartmentalization analyses performed on longitudinally sampled data from each participant using HyPhy (Pond *et al*., 2005), statistically significant evidence of genetic differentiation between blood plasma and CVL viral populations was found in CAP177 and CAP217 with the Snn test, whereas no such evidence was detected in CAP261 and CAP270 by either test (Table 3.5). Although the F_{ST} test predicted a single p-value below 0.05 in CAP177, the mean bootstrap value did not exceed 0.95, therefore implying a lack of evidence to support the existence of genetically distinct viral populations between the two tissue types. In total, the Snn and F_{ST} tests predicted fewer significant p-values compared to tree-based tests (Table 3.3) as expected, since distance-based tests have been reported to be less sensitive than tree-based tests (Zárate *et al*., 2007), however significant compartmentalization was detected in participants CAP177 and CAP217, consistent with tree-based tests (Tables 3.3 and 3.5).

When monotypic and low diversity sequences were removed from longitudinal data sets for CAP177 (n = 49) and CAP217 (n = 32), statistical evidence to support compartmentalized structure was no longer detected in CAP217, although in CAP177 statistical structure remained present (Table 3.5). As a result, cross-sectional analyses were performed on data sets from CAP177, to determine if compartmentalized structure was driven by sequences from particular time points only.

Table 3.6 Cross-sectional compartmentalization results using distance-based statistics F_{ST} and Snn in HyPhy for participant CAP177. Data set types are indicated by roman numerals I, II, and III (where I = monotypic and low diversity sequences included; II = monotypic sequences removed; III = monotypic and low diversity (<0.01% divergence) sequences removed). Statistically significant evidence of compartmentalization for the F_{ST} test was determined based on guidelines described by Josefsson *et al*. (2013), where p-value <0.05 and bootstrap value >0.95, whereas for the Snn test p-values below 0.05 were considered to be significant. Significant p-vlaues are shown in red text.

Participant ID	Data set	Days p.i.	n (Blood)	n (CVL)	Statistic	Observed value	p-value	Bootstrap mean	Bootstrap median	Bootstrap standard deviation	Bootstrap 95% CI
CAP177	I	28	12	18	F_{ST}	-0.072	0.392	0.028	-0.059	0.134	-0.087 – 0.297
					Snn	0.938	0.032	0.970	0.967	0.023	0.933 – 1.000
		378	11	19	F_{ST}	0.001	0.377	0.034	0.029	0.023	-0.000 – 0.097
					Snn	0.817	0.001	0.871	0.880	0.060	0.753 – 0.967
		560	8	5	F_{ST}	0.283	0.006	0.358	0.339	0.117	0.166 – 0.574
					Snn	0.846	0.015	0.948	0.962	0.059	0.846 – 1.000
		1295	18	19	F_{ST}	0.048	0.001	0.077	0.073	0.025	0.040 – 0.141
					Snn	0.757	0.001	0.877	0.865	0.048	0.795 – 1.000
	II	28	12	18	F_{ST}	-0.071	0.437	0.031	-0.021	0.098	-0.070 – 0.229
					Snn	0.904	0.032	0.916	0.915	0.042	0.839 – 0.990
		378	11	17	F_{ST}	0.013	0.194	0.053	0.047	0.028	0.009 – 0.140
					Snn	0.821	0.001	0.894	0.893	0.064	0.780 – 1.000
		560	7	4	F_{ST}	0.228	0.017	0.299	0.277	0.131	0.094 – 0.609
					Snn	0.818	0.067	0.923	0.909	0.072	0.818 – 1.000
		1295	18	11	F_{ST}	-0.029	0.977	0.000	-0.003	0.015	-0.022 – 0.041
					Snn	0.690	0.053	0.831	0.828	0.061	0.707 – 0.983
	III	28	3	5	F_{ST}	-0.153	0.395	0.057	-0.026	0.191	-0.158 – 0.445
					Snn	0.688	0.111	0.835	0.875	0.107	0.625 – 1.000
		378	9	14	F_{ST}	0.000	0.388	0.043	0.034	0.030	0.004 – 0.110
					Snn	0.783	0.005	0.870	0.870	0.070	0.739 – 1.000
		1295	18	9	F_{ST}	-0.040	0.996	-0.006	-0.009	0.018	-0.039 – 0.051
					Snn	0.648	0.168	0.812	0.815	0.072	0.636 – 0.963

In cross-sectional distance-based compartmentalization analyses for participant CAP177, where sequences from each time point was treated as fully independent and analysed separately, significant structure was detected at all time points when monotypic and low diversity sequences were included (Table 3.6). After the removal of monotypic sequences statistically significant evidence for compartmentalization was no longer detectable at 560 and 1295 days, following which, the removal of both low diversity and monotypic sequences resulted in almost no statistical evidence for compartmentalized structure between blood plasma and CVL populations, except at 378 days ($p \leq 0.05$).

3.3.3 Structured coalescent-based migration analysis

In maximum likelihood-based migration analysis performed using Migrate-n (http://popgen.sc.fsu.edu/Migrate/Migrate-n.html), unequal migration rates between anatomical compartments were found in all participants during acute and chronic infection stages (Table 3.7).

Table 3.7 Longitudinal analyses of viral migration patterns between matched blood plasma and cervical compartments in each participant**.**

Participant ID	n (Blood)	n (CVL)	Median Slice Sample (M)		Mean Slice Sample (M)	
			Blood to CVL	CVL to Blood	Blood to CVL	CVL to Blood
CAP177	68	62	273.0	761.0	272.7	760.8
CAP217	66	49	876.3	318.3	861.9	331.8
CAP261	40	41	515.0	609.0	516.4	611.3
CAP270	62	57	215.7	871.7	217.7	866.4

In CAP261, an almost equal rate of movement was predicted between viruses in the blood plasma and cervical compartment, while in CAP177 and CAP270 almost four times as many viruses were estimated to migrate from the cervical compartment into the blood plasma compared to the number migrating into the cervical compartment from the blood plasma (Table 3.7). In contrast, higher migration rates were predicted to have occurred from the blood plasma into the cervix in participant CAP217.

To further investigate if the presence of monotypic sequences caused a decrease in estimates of the *effective migration rates* between tissue types, monotypic sequences were excluded from all data sets and reanalysed using Migrate-n.

Table 3.8 Longitudinal analyses of viral migration patterns between blood plasma and cervical compartments in each participant, after the exclusion of monotypic sequences.

Participant ID	n (Blood)	n (CVL)	Median Slice Sample (M)		Mean Slice Sample (M)	
			Blood to CVL	CVL to Blood	Blood to CVL	CVL to Blood
CAP177	60	42	424.3	300.3	439.3	302.7
CAP217	52	38	201.0	735.0	221.3	739.6
CAP261	40	37	720.3	879.0	716.9	869.0
CAP270	61	46	942.3	827.0	926.0	825.8

After the removal of monotypic sequences there was a reversal of migration patterns in CAP177, CAP217 and CAP270, despite the estimated number of migrants in either direction being within close proximity to previous estimations for CAP177 and CAP270 (i.e. in the presence of monotypic sequences). A drastic change in migratory patterns was observed in CAP217 where a considerable number of viruses appeared to be migrating from the cervix to blood plasma, while an almost equal proportion of viruses were predicted to migrate between the blood plasma and cervical compartments in CAP261 once again (Table 3.8).

To determine if migration patterns varied between acute and chronic stages of infection, sequences were separated by time point and analysed independently.

Table 3.9 Cross-sectional analyses of viral migration patterns between blood plasma and cervical compartments in each participant.

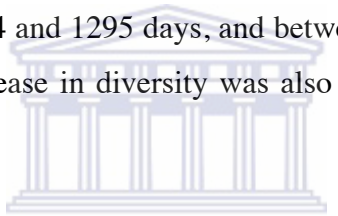
Participant ID	Days p.i.	n (Blood)	n (CVL)	Median Slice Sample (M)		Mean Slice Sample (M)	
				Blood to CVL	CVL to Blood	Blood to CVL	CVL to Blood
CAP177	28	12	18	245.0	223.7	280.8	265.2
	378	11	19	344.3	73.7	372.2	123.3
	560	8	5	316.3	203.0	373.8	264.0
	1295	18	19	811.0	210.3	800.2	212.4
CAP217	63	20	20	112.3	918.3	168.2	902.5
	420	6	5	744.3	329.0	718.7	389.5
	770	6	5	225.7	589.7	318.5	551.6
	1316	19	19	55.0	407.0	58.7	427.2
CAP261	63	19	19	206.3	884.3	236.7	846.6
	413	11	12	424.3	203.7	476.3	363.6
	945	10	10	528.3	197.0	516.9	202.6
CAP270	56	11	14	198.3	881.7	272.8	854.9
	147	8	8	277.0	893.0	344.2	864.0
	231	14	3	761.0	130.3	728.2	208.2
	406	12	11	619.0	349.7	623.6	396.3
	595	7	5	209.7	698.3	293.1	676.6
	903	10	12	117.0	620.3	125.1	633.1

Although the sample size was greatly reduced in cross-sectional analyses, the accuracy of estimates obtained based on the coalescent approach implemented in Migate-n is more dependent on the number of "independent loci than on sample size" (Bradic *et al*., 2012; Felsenstein, 2006). When sequences were analysed individually per time point, at all four times where sequences from both tissues were available for CAP177, a greater number of viruses were predicted to migrate from the blood plasma into the cervical compartment (Table 3.9), consistent with patterns observed in longitudinal analyses after the removal of monotypic sequences (Table 3.8). More viruses were estimated to migrate from the blood plasma into the cervical compartment in CAP261 too, at two of the three time points sampled (Table 3.9), however this was not consistent with longitudinal migration patterns (Tables 3.7 and 3.8).

In CAP217 and CAP270, although migratory patterns varied between sampling points, generally a larger number of viruses were estimated to migrate from the cervix into the blood plasma (i.e. at approximately 70% of time points tested) (Table 3.9), unlike longitudinal migration patterns for these participants (Table 3.8). Nevertheless, exchange of viruses between anatomical compartments was evident in all participants both longitudinally and per time point, in the presence and absence of monotypic sequences.

3.4 Intrapatient viral diversity

Pairwise genetic distances between full-length gp120 blood plasma and CVL-derived sequences increased over the infectious period measured in all participants (Figure 3.5). CAP261 and CAP177 displayed higher pairwise genetic distances in their viral populations compared to CAP270 and CAP217 throughout the entire infection period. As a result the highest increases in genetic diversity between consecutive sampling time points were observed in CAP177 between 924 and 1295 days, and between 413 and 945 days in CAP261 (Figure 3.5). Another major increase in diversity was also detected in CAP177 between 28 and 196 days, (Figure 3.5).



Levels of genetic diversity in CVL viral populations from CAP177 fluctuated between 378 and 1295 days, where viruses that were noticeably more divergent than the mainstream population appeared as outliers at 28, 378 and 1295 days (Figure 3.5). Diversities varied between blood plasma-derived viruses too, where there was a large rise in diversity between 28 and 1295 days in this participant, coinciding with outliers at 28, 560 and 924 days. No differences in diversity between viral populations in the blood plasma and cervix was evident in CAP177, however generally pairwise genetic distances were considerably lower between viruses in the cervix than those present in blood plasma towards the later stages of infection (Figure 3.5).

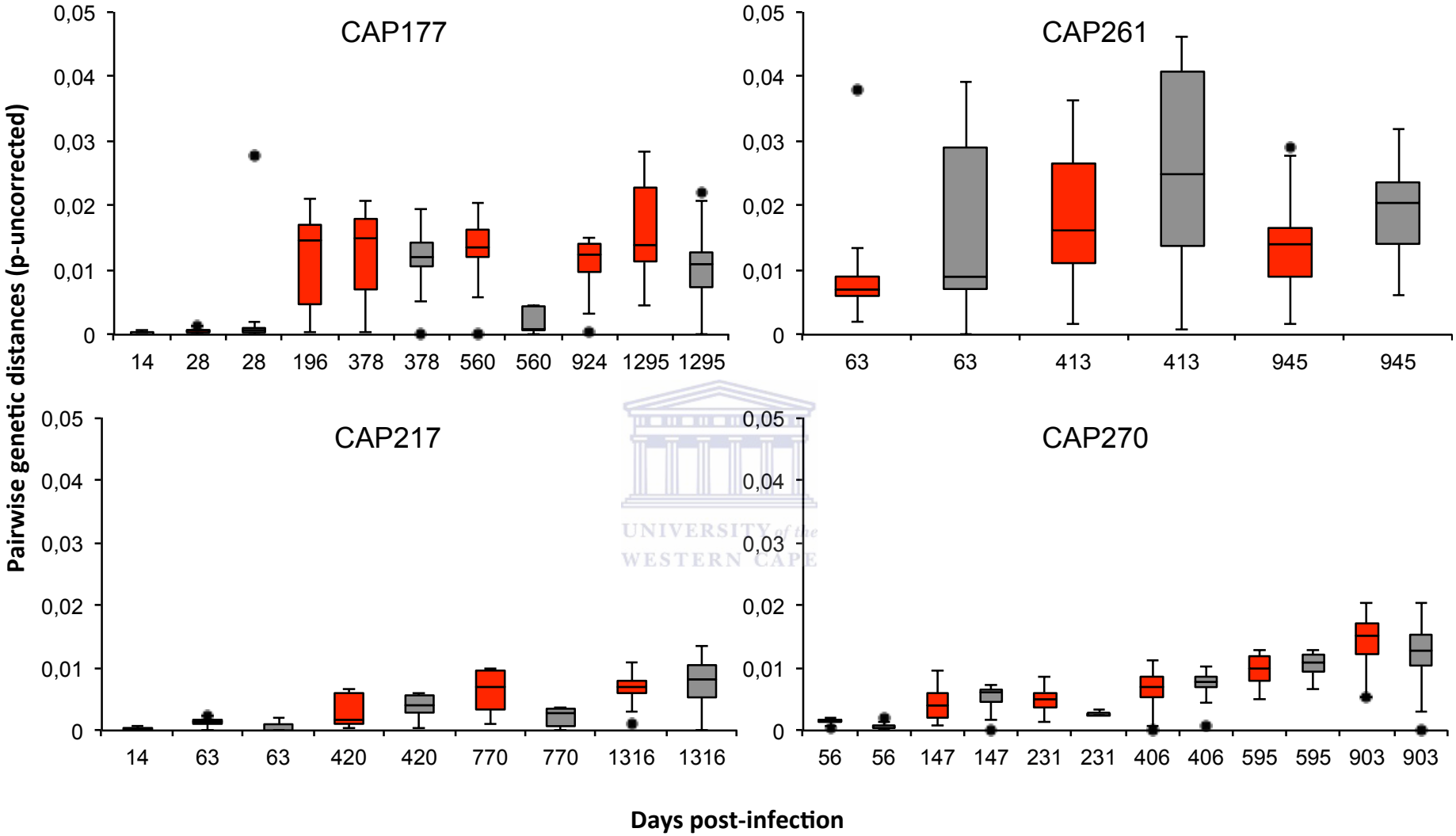


Figure 3.5 Box-and-whisker plots depicting intrapatient HIV-1 pairwise genetic distances between blood plasma (red) and CVL (grey) derived viruses in each participant over the course of infection. The median pairwise genetic distance is indicated by a bold horizontal line at the centre of the box, with the upper and lower limits of each box corresponding to the 75th and 25th percentiles, and whiskers extending to the upper and lower adjacent values respectively, stratified by sampling time points per tissue type as indicated on the x-axis. Outliers are represented by filled circles above and below box plots.

There was a steady increase in genetic diversity in CAP217, predominantly between 63 and 420 days in both blood plasma and CVL sequences, however viruses in the cervix were generally less diverse than those from the blood plasma throughout the infection period until 1316 days, when the diversity in CVL viruses overtook the diversity in blood plasma viruses (Figure 3.5). Upper and lower limits of pairwise genetic distances overlapped between blood plasma and CVL sequences over all time points except 770 days in CAP217 indicating a large degree of similarity between tissue-specific samples. No consistent tissue-specific differences were seen in CAP217, except at 770 days when the diversity in CVL-derived viruses was noticeably lower than viral populations from the blood plasma.

Although there were only three time points at which matched samples were available for CAP261 (i.e. sequences from both tissue types), a more diverse viral population was clearly evident in the cervix at 63 days (Figure 3.5). At 413 days, viral diversity in both tissue types increased, until 945 days when both populations were genetically very similar. Despite these observations, viruses in the cervix of CAP261 were more diverse than those in the blood plasma at all time points sampled.

In CAP270 a large rise in diversity was evident at 147 days, which continued to increase steadily in both tissue types at all subsequent time points, with the exception of CVL viruses sampled at 231 days, however the lower diversity in CVL viruses at 231 days was likely due to the limited number of viruses amplified (n=6) at that particular time point (Figure 3.5).

Viral populations harboured significant tissue-specific differences in pairwise genetic distances at 10 of the 17 time points where samples from both tissue types were obtained as identified through statistical testing (Table 3.10).

Table 3.10 Statistical assessment of tissue-specific differences between pairwise genetic distances, using the Mann-Whitney and Wilcoxin signed-ranked non-parametric tests. Time points where significant differences in genetic diversity between blood plasma and CVL viruses were determined by both statistical tests are indicated in red text.

Participant ID	Days p.i.	n (Blood)	n (CVL)	Mann-Whitney U test		Wilcoxin signed-rank	
				U value	p-value	W value	p-value
CAP177	28	12	18	5414	0.3433	–	–
	378	11	19	5348	0.0667	1227	< 0.0001
	560	8	5	269	< 0.0001	45	0.0039
	1295	18	19	19817	< 0.0001	11781	< 0.0001
CAP217	63	20	20	25513	< 0.0001	4465	< 0.0001
	420	6	5	95	0.2825	-45	0.0039
	770	6	5	123	0.0076	45	0.0039
	1316	19	19	17897	0.0003	-7546	< 0.0001
CAP261	63	19	19	17517	0.0015	-9086	< 0.0001
	413	11	12	2447	0.0010	-1414	< 0.0001
	945	10	10	1391	0.0022	-1035	< 0.0001
CAP270	56	11	14	4166	< 0.0001	1485	< 0.0001
	147	8	8	500	0.0748	-178	0.0003
	231	14	3	254	0.0103	-3	0.5000
	406	12	11	2107	0.1237	-1225	< 0.0001
	595	7	5	120	0.5321	-55	0.0020
	903	10	12	1927	0.0078	1035	< 0.0001

Significant tissue-specific differences between population diversity was detected in all participants at isolated time points in CAP177, CAP217 and CAP270, and at all time points in CAP261, although data from only three time points were available in this participant. In contrast the least number of significant tissue-specific differences in pairwise genetic distances was detected in the participant with the most number of sampled time points, i.e. CAP270.

The mean and median pairwise genetic distances between viral populations in the blood plasma and CVL did not demonstrate any consistent differences in pairwise genetic diversity, however blood plasma viruses from CAP177 appeared to be more diverse than those sampled from the CVL, whereas in CAP261 the opposite pattern was observed (Table 3.11).

Table 3.11 Mean and median pairwise genetic distances between blood plasma and CVL viruses from each participant over the entire sampling period.

Participant ID	Days p.i.	Blood plasma		CVL	
		Mean	Median	Mean	Median
CAP177	28	0.0005	0.0003	0.0035	0.0007
	378	0.0124	0.0149	0.0117	0.0120
	560	0.0135	0.0135	0.0022	0.0009
	1295	0.0162	0.0139	0.0097	0.0109
CAP217	63	0.0012	0.0012	0.0007	0.0010
	420	0.0030	0.0018	0.0040	0.0040
	770	0.0062	0.0070	0.0022	0.0028
	1316	0.0068	0.0070	0.0078	0.0082
CAP261	63	0.0100	0.0070	0.0143	0.0090
	413	0.0180	0.0162	0.0265	0.0250
	945	0.0141	0.0140	0.0192	0.0203
CAP270	56	0.0014	0.0012	0.0006	0.0004
	147	0.0042	0.0041	0.0052	0.0061
	231	0.0049	0.0049	0.0027	0.0024
	406	0.0063	0.0069	0.0075	0.0078
	595	0.0097	0.0100	0.0105	0.0109
	903	0.0145	0.0153	0.0124	0.0128

3.4.1 Association between viral diversity and viral loads

To determine if there were any significant associations between viral genetic diversity and viral loads, average pairwise genetic distances were calculated for each time point sampled and compared to viral load measurements from the corresponding time points. CAP177 was the only participant in whom a statistically significant correlation between average pairwise genetic diversity and viral load was found ($p \leq 0.0141$). In all other participants, p-values were not significant ($p > 0.05$) using Spearman's rank correlation (Table 3.12).

Table 3.12 Correlation between average genetic distances and viral loads in all participants over the sampling period. Participants in whom a significant correlation between viral loads and viral diversity was found are indicated in red text.

Participant ID	Days p.i.	Plasma viral load	Average genetic distance	Spearman's rho	p-value	95% CI	DF
CAP177	14	359000	0.0004	-0.8555	0.0141	-0.9783 to -0.2877	6
	28	698000	0.0036				
	196	152000	0.0172				
	378	42100	0.0182				
	560	59200	0.0180				
	924	43800	0.0174				
	1295	38800	0.0212				
CAP217	14	3260000	0.0001	-0.6119	0.2727	-0.9703 to 0.5878	4
	63	75600	0.0014				
	420	18400	0.0058				
	770	1940	0.0088				
	1316	30000	0.0130				
CAP261	63	821000	0.0115	-0.9893	0.0931	Requires more than 3 data points	2
	413	65400	0.0203				
	945	27500	0.0193				
CAP270	56	76800	0.0009	0.5164	0.2942	-0.5083 to 0.9358	5
	147	687000	0.0051				
	231	310000	0.0051				
	406	738000	0.0083				
	595	3710000	0.0120				
	903	1290000	0.0156				

The strongest associations were detected in CAP177 ($p \leq 0.0141$) and CAP261 ($p \geq 0.0931$) where there was a clear rise in genetic diversity as participants progressed into the chronic phase of infection (suggesting a more rapid rate of disease progression in these participants), whereas the weakest associations were noted in CAP217 ($p > 0.2727$) and CAP270 ($p > 0.2942$) (Table 3.12).

3.5 Poisson distribution fitting

To determine the number of transmitted viral variants that likely infected each participant, frequency distributions illustrating intersequence Hamming distances (difference in base positions between two genomes (Keele *et al*., 2008)) were obtained and analysed per participant, using the Poisson Fitter tool (Rose & Korber, 2000). Blood plasma and CVL sequences from the earliest time points sampled were analysed and deviation from the Poisson model was considered as evidence of infection by more than one viral variant. CAP177 and CAP261 were among the participants that deviated from the Poisson distribution, whereas CAP217 and CAP270 appeared to follow a unimodal distribution (Figure 3.6), consistent with patterns by single variant founded infections (Novitsky *et al*., 2011).

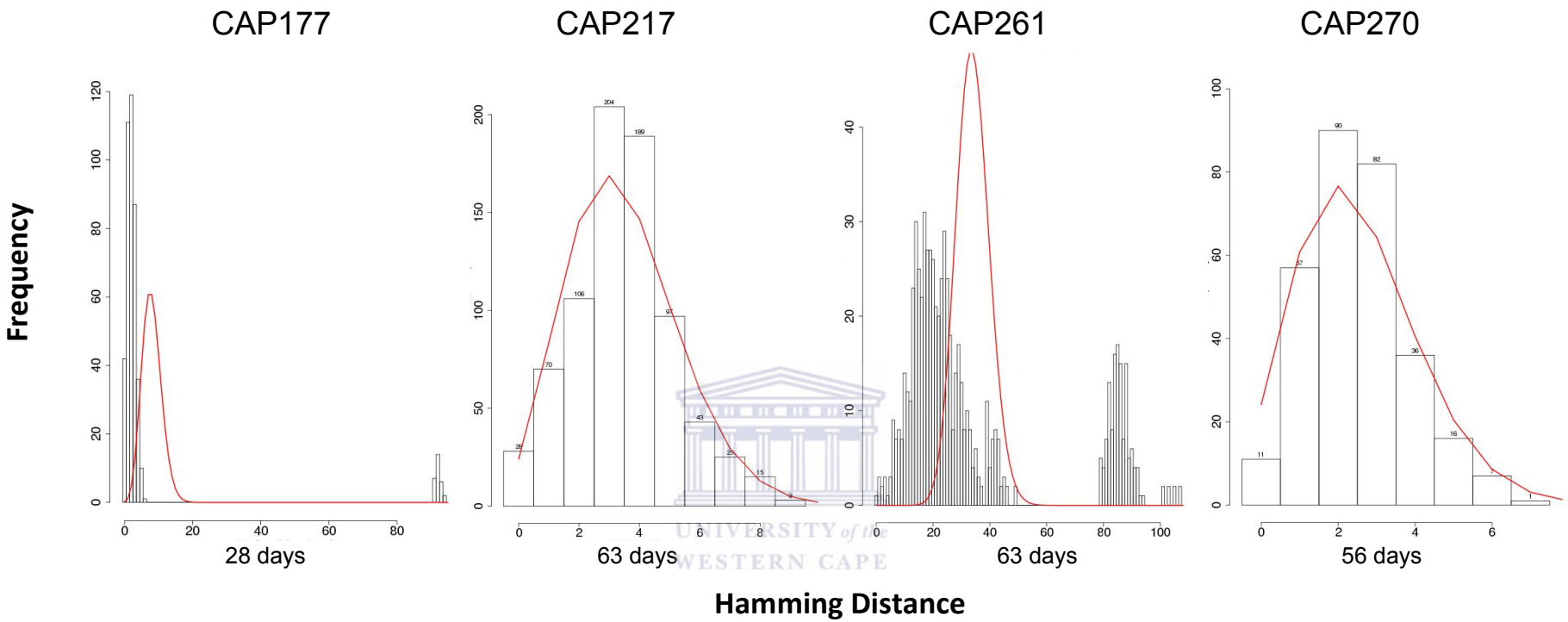


Figure 3.6 Bar charts illustrating Hamming distances between blood plasma and CVL sequences at the earliest sampling time point where sequences from both tissues were available for each participant. Although unimodal lines were fitted along frequency distributions predicted in all participants, multiple peaks were clearly visible in CAP177 and CAP261. This is because the Poisson-Fitter provides a null model that assumes early HIV-1 infection initiated by a single variant by default, prior to the onset of intrahost selection (Wolinsky *et al*., 1992; Novitsky *et al*., 2011). Well-defined peaks were observed in CAP177 and CAP261 at 28 and 63 days respectively, indicative of infection by more than one viral variant, whereas in CAP217 and CAP270 single unimodal distributions were evident at 63 and 56 days, suggesting infection by a single viral variant in these participants (Novitsky *et al*., 2011).

3.6 Accumulation of potential N-linked glycosylation sites

In total the frequencies of around 42.90% of the inferred PNLG's did not change (i.e. sites that remained either present or absent in all *env* gp120 sequences consistently), 23.70% increased in frequency (i.e. sites that began to *appear* in an increasing number of sequences with time), and 10.33% decreased in frequency (i.e. sites that were no longer detectable in an increasing number of sequences over time) between the first and last time point sampled, while 23.06% fluctuated randomly along the course of infection (Figure 3.7).

Table 3.13 Summary of the average number of PNLG sites predicted in the V1V2, V4 and V5-loops of blood plasma and CVL sequences among the four participants over time. Sampling time points where sequences from the blood plasma or CVL were not available are indicated by the "–" symbol.

Participant ID	n	Days p.i.	V1V2		V4		V5	
			Blood plasma	CVL	Blood plasma	CVL	Blood plasma	CVL
CAP177	18	14	4.05	–	3.05	–	2.00	–
	30	28	4.00	4.05	3.00	3.00	2.00	1.94
	11	196	3.82	–	3.00	–	2.00	–
	30	378	5.10	4.47	2.90	3.00	1.80	1.68
	13	560	5.38	4.20	1.88	1.20	2.00	1.20
	5	924	6.75	7.00	2.75	2.00	1.75	3.00
	37	1295	7.71	8.68	2.39	3.58	1.61	1.37
CAP217	18	14	5.84	–	3.05	–	2.00	–
	20	63	5.80	5.50	2.90	3.00	2.00	1.85
	11	420	4.33	4.40	3.00	3.00	1.00	1.20
	11	770	4.83	4.00	3.00	3.00	1.17	1.80
	38	1316	6.32	5.95	2.89	3.00	1.00	1.11
CAP261	38	63	6.80	6.84	3.37	3.58	0.95	1.00
	23	413	7.45	6.00	4.27	3.92	0.55	0.50
	20	945	7.80	8.00	4.30	4.70	1.00	1.00
CAP270	25	56	6.91	7.00	4.90	4.80	1.00	1.00
	1	105	–	7.00	–	5.00	–	1.00
	16	147	6.75	6.75	3.50	3.50	1.00	1.00
	17	231	7.00	6.83	3.93	4.17	1.00	1.00
	23	406	6.92	6.73	4.00	4.00	1.00	1.00
	12	595	6.86	6.80	4.00	4.00	1.00	1.00
	22	903	6.30	6.25	3.90	3.75	0.90	0.91

Both the largest PNLGs frequency increases and decreases occurred in the V-loop regions in all participants (range = 64 – 86%), and of these, more increased (range = 36 – 67%) than decreased (range = 13 – 29%) over the sampling periods analysed (Figure 3.7). Generally however, the number of PNLGs within the V-loop regions did not differ consistently between blood plasma and CVL sequences in any of the participants over time (Table 3.13).

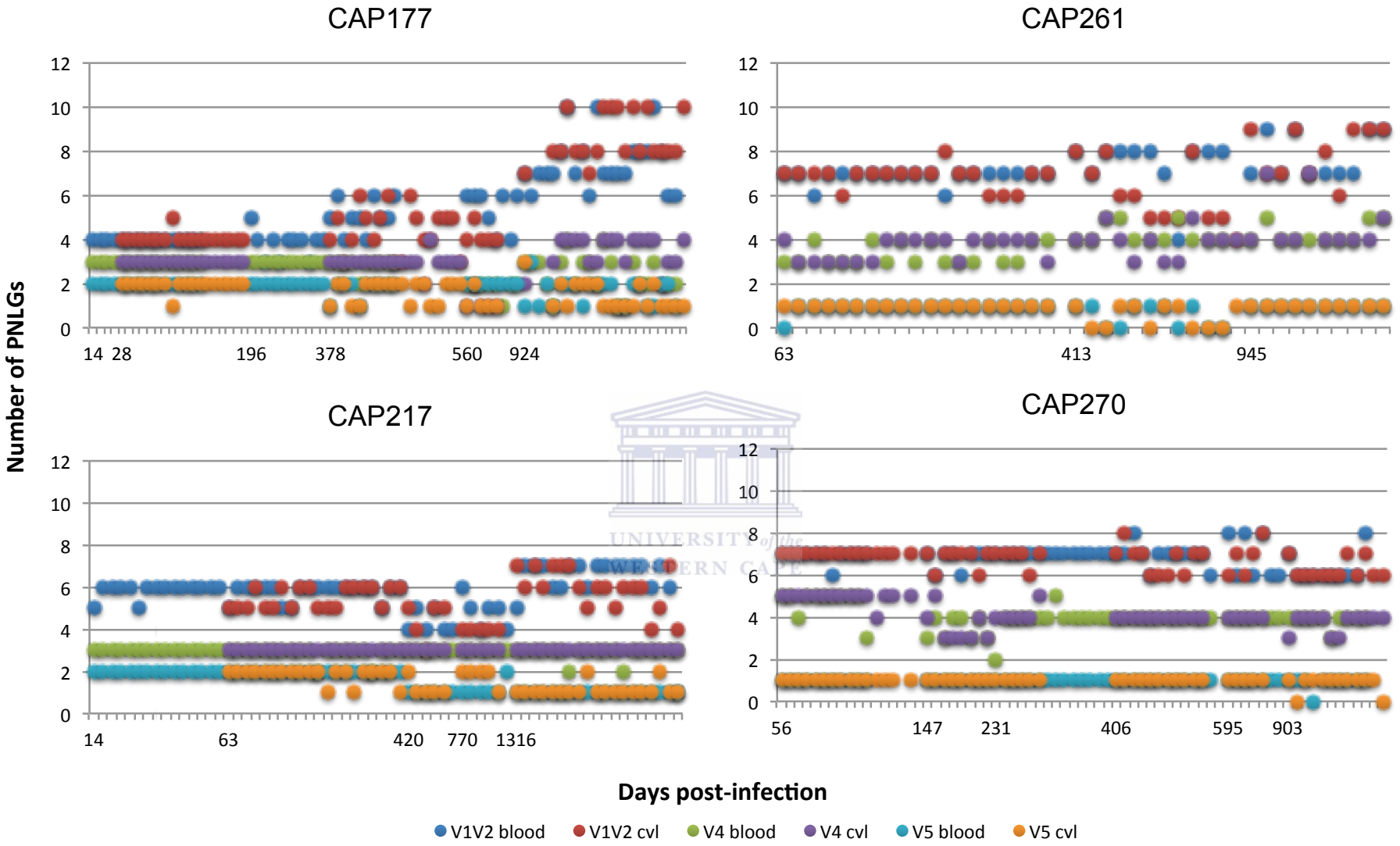


Figure 3.7 Total numbers of PNLG sites within the V1V2, V4 and V5-loops of blood plasma and CVL sequences throughout the sampling period among all participants. The actual number of PNLG sites predicted per sample is presented on the scatterplot, separated by tissue type for each V-loop region, where the number of PNLG sites is shown on the y-axis and the time in days post-infection on the x-axis.

Changes in the frequency of selected PNLG sites varied considerably between participants during their respective infectious periods. While some PNLGs in all of the participants appeared to randomly fluctuate in frequency around the population mean, others displayed systematic and often quite substantial changes in frequency between the first and last sampling time points (Figure 3.7). For example, in CAP270, the PNLG site located at N400 in the V4-loop, was present in 96% of the sequences at 56 days but no longer detected at 903 days, whereas a PNLG site at N407 also in the V4-loop that was present in all sequences at 56 days in this participant remained present in only 33% of sequences by the final sampling point at 903 days. Similarly in CAP217, the presence of a PNLG site at N463 varied in blood plasma and CVL sequences over the sampling period, decreasing from a frequency of 100% to 5% between 14 and 1316 days, while a PNLG site at N131 increased from a presence of 0% to 49% during the same period.

In CAP217, PNLGs at positions N130, N131 and N142 were detected for the first time at the final sampling point (1316 days) in approximately 29% of the sequences obtained from this time point. In CAP270, several PNLGs at N130, N135, N186, N187, N413, N460, N637 and N674 that were not present at the first sampling time point (i.e. 56 days) increased in frequency over the sampling period in both blood plasma and CVL viruses from this participant, where approximately 67% of these PNLGs occurred in the V-loop regions, with the exception of PNLGs at N130, N460, N637 and N674 that were located in the C1, C4 and C5 subregions respectively.

As was previously reported, the presence and frequency of PNLGs at N332 and N334 were inversely related over the course of infection in participant CAP177 (Moore *et al*., 2012) (Figure 3.8). All sequences obtained at 14 and 28 days from both tissue types showed the presence of a PNLGs at N334 and absence of a PNLGs at N332 until 196 days, when both the N332 and N334 PNLGs were present in 9% of all sequences. By 378 days, PNLGs at N334 were no longer detected in any of the sequences, while PNLGs at N332 were present in 91% of the sequences, and by 560 days had reached complete fixation before declining again to being no longer detected in any of the sequences at 1295 days (Figure 3.8). Over the same period, the presence of PNLGs at N234 increased from a 0% to 95% frequency in blood plasma and CVL sequences from CAP177.

In CAP261, PNLGs that occurred outside of V-loop regions included N356, which increased in frequency from being present in 53% of sequences at 63 days to 100% at 945 days, while PNLGs at sites N460 and N743 decreased from complete fixation at 63 days to complete extinction at 945 days. Although there were some distinct changes in the frequency of PNLGs, sites that remained present in all four participants included N156, N160, N301 and N386, with similar patterns observed only at PNLG sites N156 (V1-loop) and N301 (V3-loop), both of which remained present in blood plasma and CVL sequences throughout the sampling period.

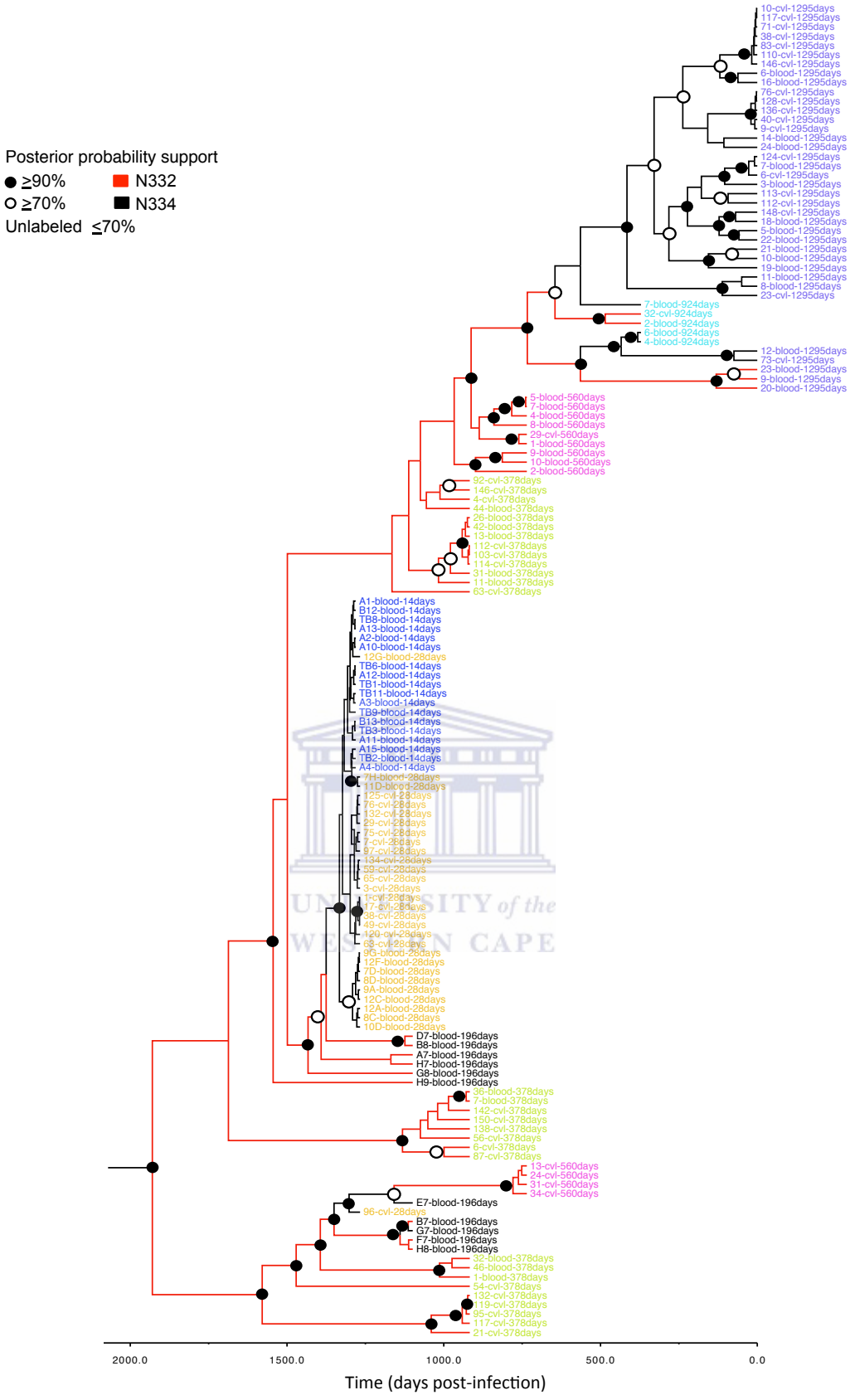


Figure 3.8 Time-ordered Bayesian maximum clade credibility (MCC) tree for CAP177 under the GTR + G_4 substitution model and a constant population size relaxed clock evolutionary model, with branches coloured according to the presence of a PNLG site at position N332 (red) or N334 (black) on the translated gp120 region. Posterior probabilities ≥ 90% are indicated by a filled circled and ≥ 70% by an open circle at the nodes, with branch labels coloured according to the time points sampled.

PNLG changes were initially evaluated within the V-loop regions, as this is where considerable changes in glycosylation accumulation were expected. To do this, the total number of PNLGs in blood plasma sequences was compared to those in CVL sequences at each of the time points sampled per participant (Table 3.14).

Table 3.14 Statistical assessment of tissue-specific differences in the total number of PNLGs within the V-loop regions, using the Mann-Whitney and Wilcoxin signed-ranked non-parametric tests. Tests where the standard deviation was equal to zero or where an "invalid floating point operation" error was received (i.e. where GraphPad was unable to fit the data using the selected model and options (Motulsky, 2003)), are indicated by the "–" symbol. P-values below 0.05 for both tests were considered as credible evidence of significant tissue-specific differences between PNLGs numbers.

Participant ID	HIV-1 *env* region	Days p.i.	Mann-Whitney U test		Wilcoxin signed-rank	
			U value	p-value	W value	p-value
CAP177	V1V2	28	–	–	–	–
		378	128	0.1170	21	0.2500
		560	42	0.0191	10	0.1250
		1295	254	0.0078	-61	0.0327
	V4	28	–	–	–	–
		378	104	0.4606	–	–
		560	34	0.0244	–	–
		1295	289	0.0002	-105	0.0001
	V5	28	–	–	–	–
		378	106	0.5340	1	> 0.9999
		560	–	–	–	–
		1295	213	0.1503	18	0.2500
CAP217	V1V2	63	260	0.0515	27	0.0547
		420	17	0.7621	0	> 0.9999
		770	–	–	–	–
		1316	241	0.0600	51	0.2435
	V4	63	–	–	–	–
		420	–	–	–	–
		770	–	–	–	–
		1316	–	–	–	–
	V5	63	–	–	–	–
		420	–	–	–	–
		770	25	0.0578	–	–
		1316	–	–	–	–

Participant ID	HIV-1 *env* region	Days p.i.	Mann-Whitney U test		Wilcoxin signed-rank	
			U value	p-value	W value	p-value
CAP261	V1V2	63	191	0.6646	3	0.8438
		413	102	0.0216	26	0.0313
		945	55	0. 7071	-4	0.6250
	V4	63	219	0.2058	-22	0.2754
		413	85	0.1771	10	0.1875
		945	53	0.8146	-4	0.5000
	V5	63	–	–	–	–
		413	69	0.8590	3	0.8125
		945	–	–	–	–
CAP270	V1V2	56	–	–	–	–
		147	32	0.9443	0	> 0.9999
		231	–	–	–	–
		406	78	0.4203	9	0.3125
		595	18	0.9294	4	0.6250
		903	62	0.9281	0	> 0.9999
	V4	56	86	0.4096	–	–
		147	34	0.8577	0	> 0.9999
		231	49	0.4256	-3	0.5000
		406	–	–	–	–
		595	–	–	–	–
		903	69	0.4020	3	0.5000
	V5	56	–	–	–	–
		147	–	–	–	–
		231	–	–	–	–
		406	–	–	–	–
		595	–	–	–	–
		903	64	0.6981	0	> 0.9999

Significant differences in PNLGs between blood plasma and CVL viruses in the V-loop regions were found at five isolated time points among all participants, as confirmed through Mann-Whitney and Wilcoxin signed-rank testing. In CAP177, significant tissue-specific differences in the number of PNLGs were detected in the V1V2 and V4-loops of viruses sampled at 560 and 1295 days. Similar differences were identified in CAP261 within the V1V2-loops at 413 days, whereas no such differences were detected in participants CAP217 and CAP270.

When comparing PNLGs across the full-length gp120 region however, significant differences between PNLGs numbers in blood plasma and CVL sequences were evident in all participants predominantly towards the later stages of infection (Table 3.15).

Table 3.15 Statistical assessment of tissue-specific differences in the total number of PNLGs along the HIV-1 *env* gp120 region, using the Mann-Whitney and Wilcoxin signed-ranked non-parametric tests. Tests where the standard deviation was equal to zero or where an "invalid floating point operation" error was received (i.e. where GraphPad was unable to fit the data using the selected model and options (Motulsky, 2003)), are indicated by the "–" symbol. P-values below 0.05 for both tests were considered as credible evidence of significant tissue-specific differences between PNLGs numbers.

Participant ID	HIV-1 *env* region	Days p.i.	Mann-Whitney U test		Wilcoxin signed-rank	
			U value	p-value	W value	p-value
CAP177	Full-length gp120	28	–	–	–	–
		378	127	0.1331	23	0.0469
		560	38	0.0092	15	0.0625
		1295	282	0.0007	-126	0.0003
CAP217	Full-length gp120	63	247	0.1556	30	0.2661
		420	18	0.6233	-2	0.8750
		770	18	0.6411	2	0.7500
		1316	249	0.0382	67	0.1540
CAP261	Full-length gp120	63	197	0.6295	-6	0.8311
		413	113	0.0032	36	0.0078
		945	69	0.1456	-31	0.0234
CAP270	Full-length gp120	56	88	0.4090	-3	0.5000
		147	48	0.0676	15	0.1563
		231	45	0.8069	-1	> 0.9999
		406	109	0.0041	21	0.0313
		595	19	0.8453	0	> 0.9999
		903	96	0.0152	31	0.0234

Significant differences in the number of PNLGs between tissue-types was found at 4 of the 17 time points where samples from both tissues were available (Table 3.15), most of which (± 3) occurred within the later sampling time points among all participants. Generally, there were more PNLGs gained than lost during disease progression and tissue-specific differences became evident during later sampling points, when viruses were more diverse than populations present at earlier time points as expected.

3.6.1 Clustering of PNLG sites on phylogenetic trees

To gain an insight into the evolutionary relationship between PNLGs accumulation and disease progression, PNLG sites were mapped to Bayesian MCC trees (previously produced in section 3.2.3). While it was evident that there was an overall increase in the number of glycosylation sites among all participants over the sampling period, it was unclear if PNLG site accumulation occurred discretely or within genetically related sequences. To evaluate this, PNLG sites that appeared in an increasing number of sequences over time, were assessed for clustering on phylogenetic trees.

Table 3.16 PNLG sites showing evidence of increasing frequency within gp120 subregions in blood plasma and CVL viruses over the course of infection. PNLGs indicated in red text were of particular interest as they showed a rise in occurrence in more than one participant during the sampling period, whereas PNLGs in black showed an increase in a single participant only.

Participant ID	HIV-1 *env* region	PNLGs relative to HXB2	Total PNLGs
CAP177	C1	N130	1
	V1V2	N137, N139, N141, N142, N186	5
	C2	N234	1
	V4	N393	1
	C4	N442, 460	2
	C5	N674	1
CAP217	V1V2	N131, N142, N190	3
CAP261	V1V2	N141, N187, N190	3
	C3	N356	1
	V4	N386, N408, N413	3
CAP270	C1	N130	1
	V1V2	N135, N186, N187	3
	V4	N413	1
	C5	N637, N674	2

From the total of 28 PNLGs that showed a clear rise in frequency over time, 16 were detected in more than one participant, 10 of which were located in the V1V2-loop. When PNLG sites where then mapped to Bayesian MCC trees, glycosylation sites appeared in monophyletic clades containing two or more sequences for all 16 PNLGs that were mapped (Figures 3.9a-d), consistent with patterns of selection (Cherry *et al*., 2009).

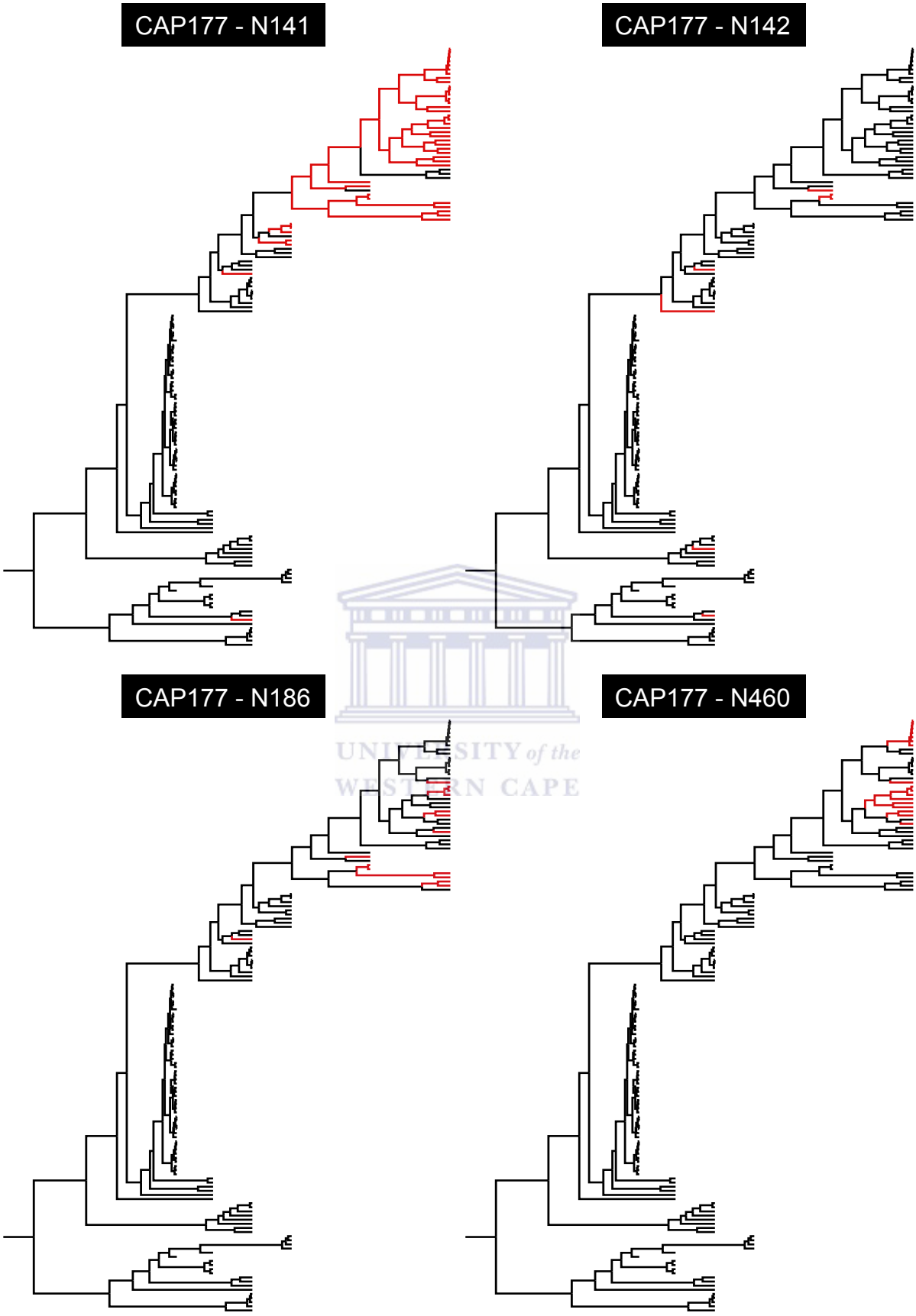


Figure 3.9a Time-ordered Bayesian maximum clade credibility trees illustrating potential N-linked glycosylation sites N141, N142, N186 and N460 in participant CAP177. Participant ID and the PNLG site that is mapped per tree are shown in the black box above each tree, with red branches indicating sequences that contained the PNLG site listed above.

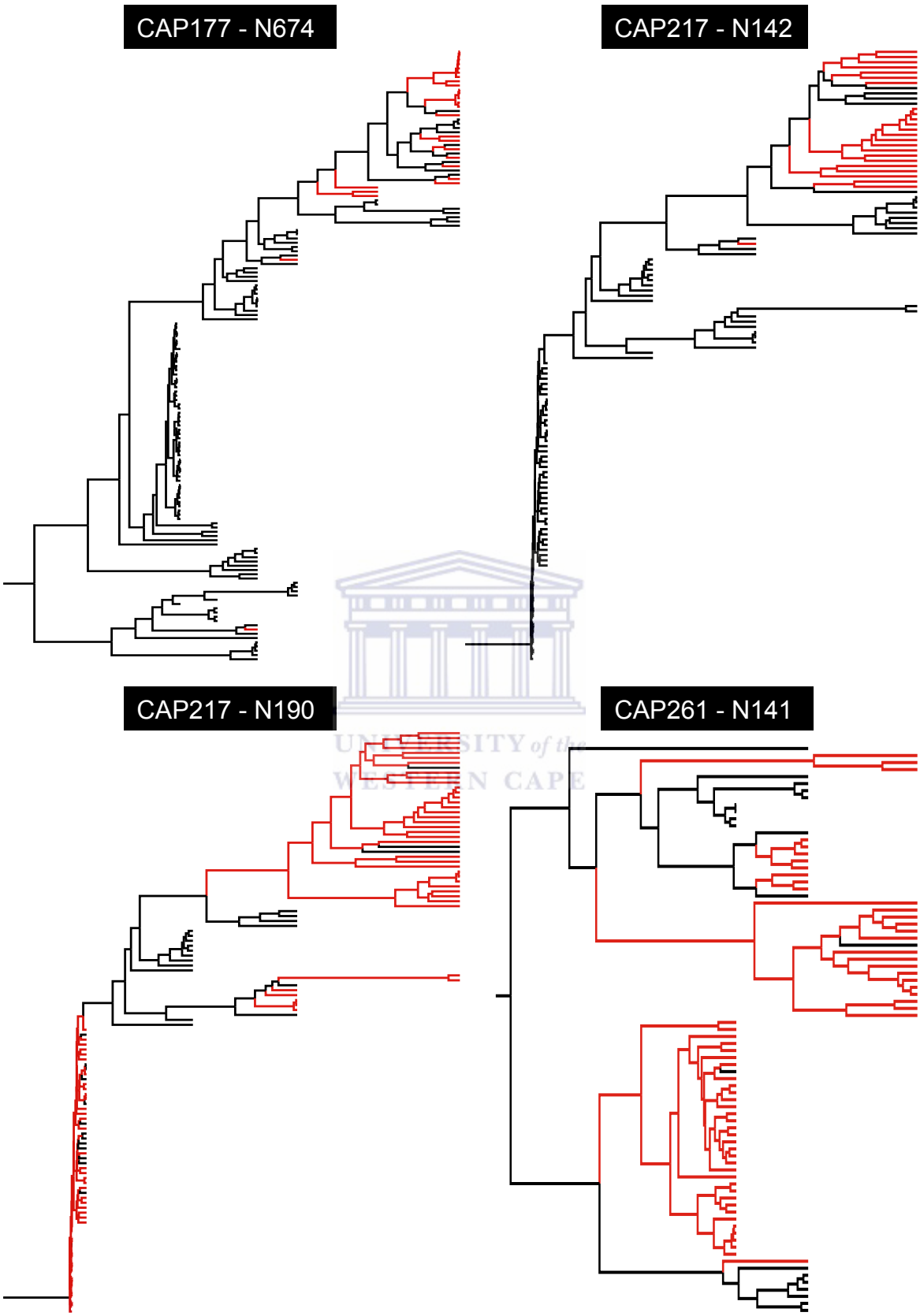


Figure 3.9b Time-ordered Bayesian maximum clade credibility trees illustrating potential N-linked glycosylation sites N674, N142, N190 and N141 in participants CAP177, CAP217 and CAP261 respectively. Participant IDs and the PNLG site that is mapped per tree are shown in the black box above each tree, with red branches indicating sequences that contained the PNLG site listed above.

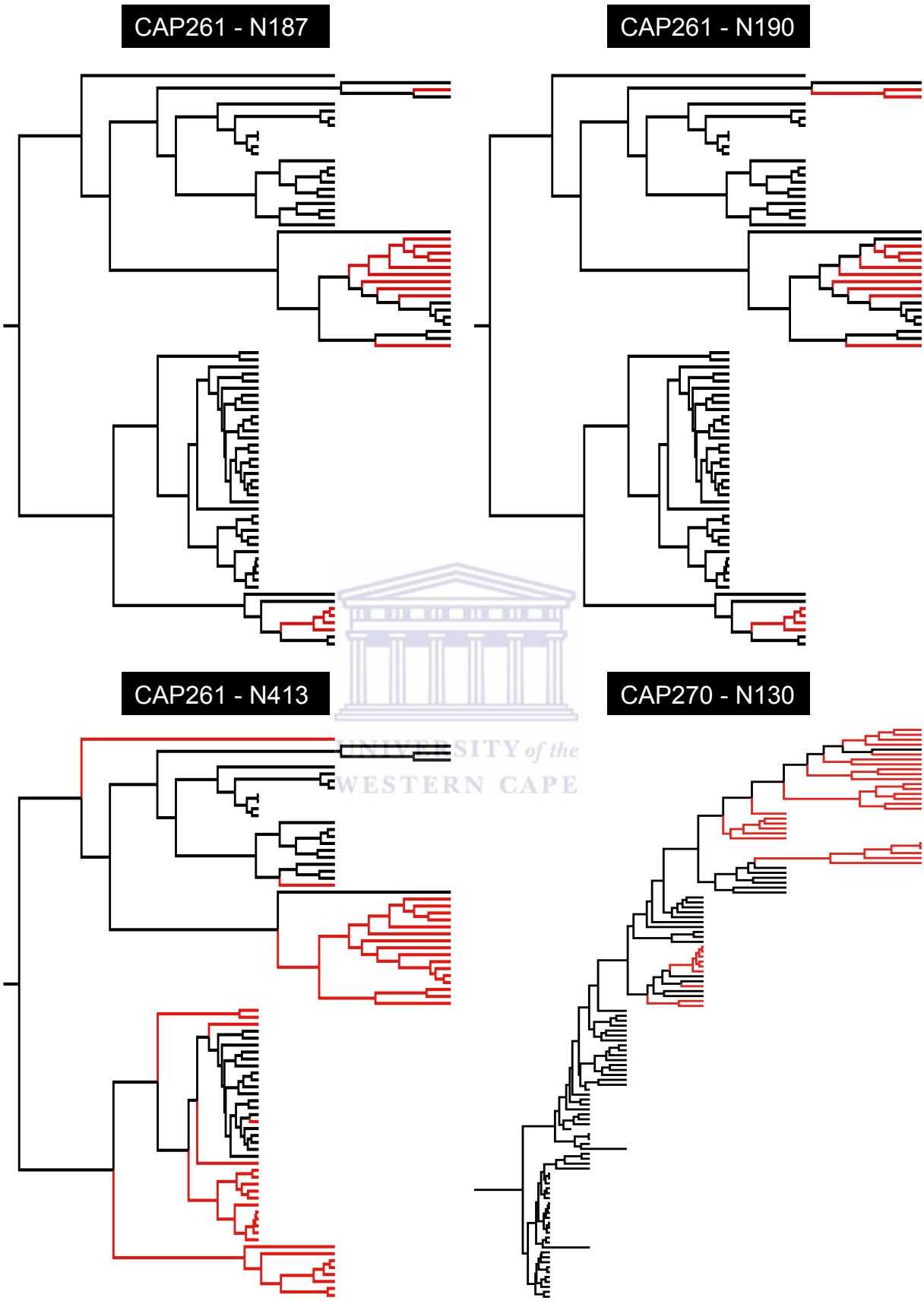


Figure 3.9c Time-ordered Bayesian maximum clade credibility trees illustrating potential N-linked glycosylation sites N187, N190, N413 and N130 in participants CAP261 and CAP270. Participant IDs and the PNLG site that is mapped per tree are shown in the black box above each tree, with red branches indicating sequences that contained the PNLG site listed above.

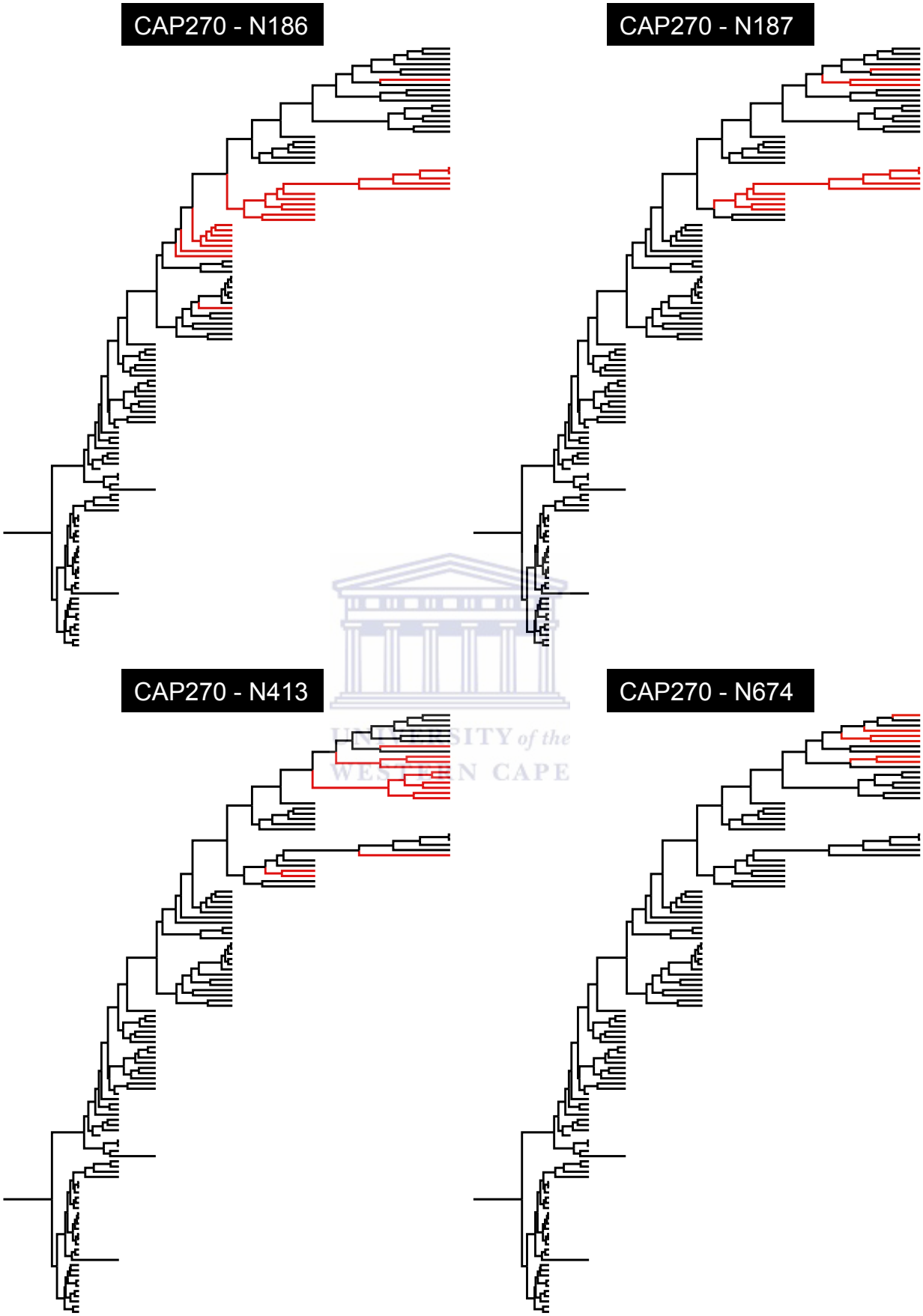


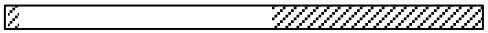
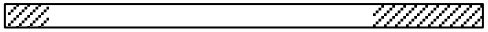
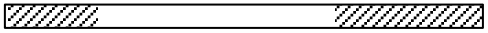
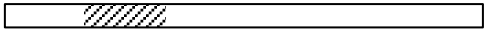
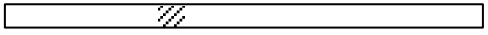
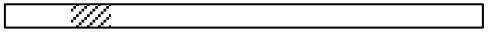
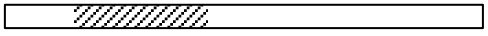
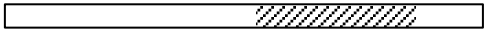
Figure 3.9d Time-ordered Bayesian maximum clade credibility trees illustrating potential N-linked glycosylation sites N186, N187, N413 and N674 in participant CAP270. Participant ID and the PNLG site that is mapped per tree are shown in the black box above each tree, with red branches indicating sequences that contained the PNLG site listed above.

3.7 Intrapatient recombination detection

To identify potential recombinants, full-length gp120 *env* blood plasma and CVL sequences were analysed from all participants. Evidence of recombination was detected in three of the four participants analysed using RDP4, however since samples from the initial time of infection were not available for any of the participants, the transmission of recombinants at the point of infection could not be ruled out (Abrahams *et al*., 2009).

A total of 17 unique, statistically significant recombination events (p-value $\leq$ 0.05 after Bonferroni correction) were identified using RDP4 (Table 3.17), approximately 71% of which were identified in sequences from mid to late sampling points in participants CAP177 (Figure 3.10a) and CAP261 (Figure 3.10b), whereas all recombinants detected in CAP270 occurred during the last sampling point at 903 days post-infection (Figure 3.10c). Table 3.17 shows an overview of unique recombination events detected in CAP177, CAP261 and CAP270. Although no recombination breakpoints were detected in viral populations from CAP217, recombination is believed to be present at an undetectable level in this participant.

Table 3.17 Summary of unique intrapatient recombination events detected among CAP177, CAP217 and CAP261 *env* sequences. RDP4 methods with supporting p-values and recombination block graphs are shown. Striped regions within the recombination block graphs represent fragments of the gp120 sequence where recombination was predicted, while the absence of p-values signify recombination events that were not detected by the selected method.

	Event number	Recombination block pattern	Recombination detection analysis p-value						
			RDP	GENECOV	BootScan	MaxChi	Chimaera	Siscan	3Seq
CAP177	1		–	–	4.014×10^{-04}	1.599×10^{-02}	2.147×10^{-02}	6.185×10^{-07}	1.981×10^{-05}
	2		2.260×10^{-02}	1.358×10^{-02}	2.240×10^{-03}	–	–	–	9.287×10^{-05}
	3		–	–	1.120×10^{-02}	3.293×10^{-03}	–	–	9.119×10^{-02}
	4		–	–	–	1.083×10^{-02}	9.366×10^{-04}	3.345×10^{-05}	1.750×10^{-05}
	5		1.506×10^{-05}	2.582×10^{-06}	5.607×10^{-06}	3.881×10^{-03}	1.129×10^{-03}	5.133×10^{-05}	1.568×10^{-04}
	6		8.449×10^{-03}	4.250×10^{-02}	1.626×10^{-02}	8.182×10^{-04}	2.622×10^{-03}	2.261×10^{-06}	2.182×10^{-02}
	7		–	–	–	3.751×10^{-04}	5.892×10^{-04}	2.437×10^{-04}	2.254×10^{-06}
	8		–	–	–	-	4.402×10^{-02}	5.722×10^{-04}	5.366×10^{-03}
CAP261	1		4.089×10^{-03}	1.588×10^{-02}	4.475×10^{-03}	1.403×10^{-02}	3.447×10^{-02}	1.144×10^{-02}	1.119×10^{-04}
	2		1.540×10^{-04}	2.038×10^{-03}	6.184×10^{-04}	3.625×10^{-07}	9.851×10^{-07}	6.988×10^{-10}	2.632×10^{-08}
	3		3.507×10^{-03}	–	3.544×10^{-03}	–	–	–	1.671×10^{-02}
	4		2.554×10^{-04}	3.699×10^{-05}	1.256×10^{-05}	1.775×10^{-05}	4.521×10^{-04}	1.858×10^{-06}	3.865×10^{-08}
	5		2.264×10^{-02}	–	1.553×10^{-02}	3.281×10^{-05}	4.693×10^{-05}	3.045×10^{-04}	1.300×10^{-04}
	6		–	9.283×10^{-03}	8.117×10^{-04}	2.797×10^{-01}	–	4.006×10^{-04}	–
	7		–	–	–	6.889×10^{-04}	1.289×10^{-02}	–	9.173×10^{-03}
	8		4.637×10^{-05}	7.656×10^{-03}	4.671×10^{-05}	8.297×10^{-05}	3.302×10^{-03}	1.266×10^{-07}	–
CAP270	1		–	–	3.644×10^{-02}	9.973×10^{-01}	–	6.894×10^{-04}	–

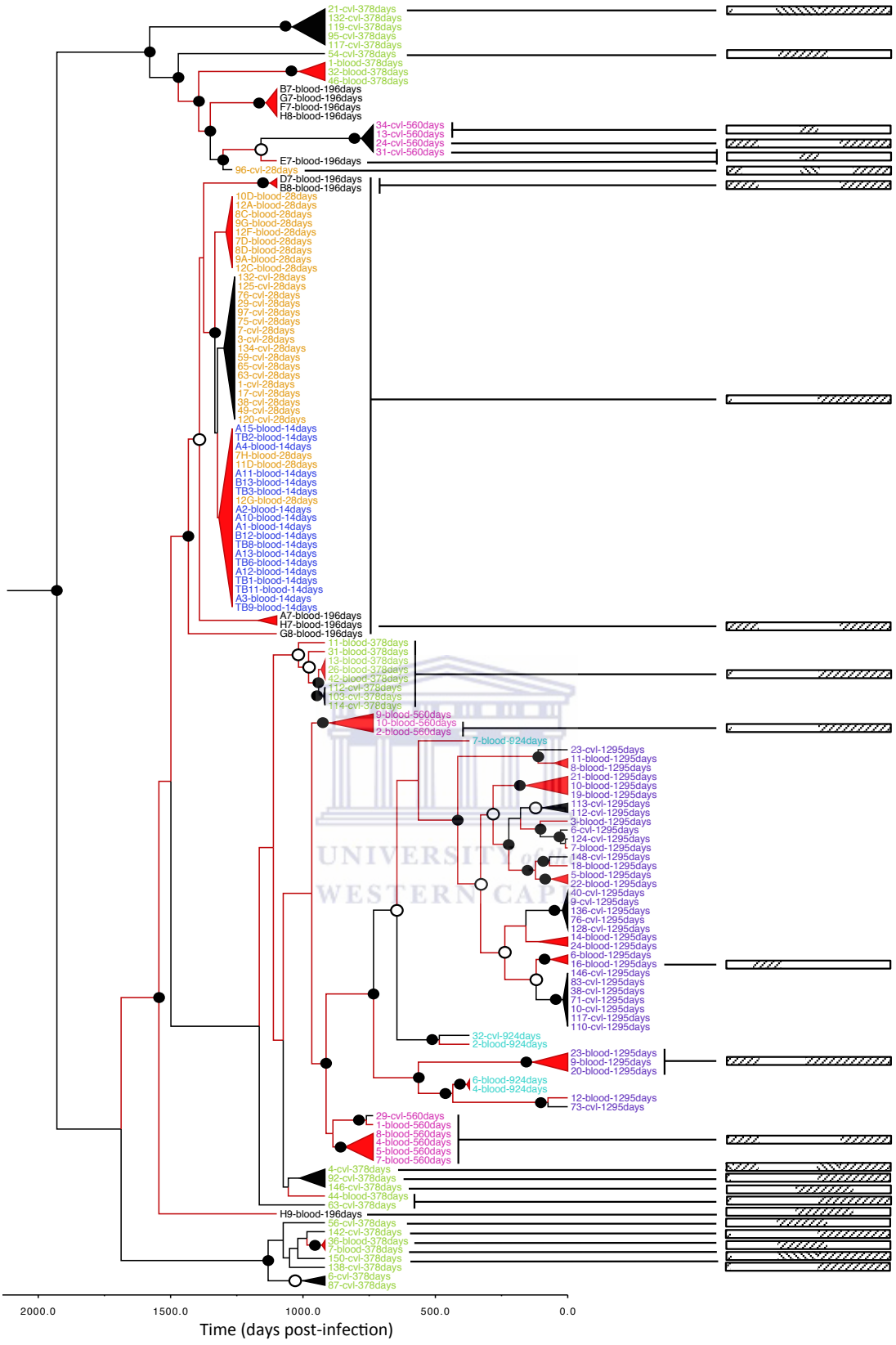


Figure 3.10a Time-scaled Bayesian MCC tree for CAP177 indicating sequences in which recombination breakpoints were detected using RDP4. Recombination block graphs displayed to the right of taxon names indicate unique recombination events as listed in Table 3.17. Branches are coloured according to the most probable state of their tissue origin where red represents viruses from the blood plasma and black indicates viruses from the cervix. Posterior probabilities ≥ 90% are indicated by a filled circled and ≥ 70% by an open circle at the nodes, with branch labels coloured according to the time points sampled.

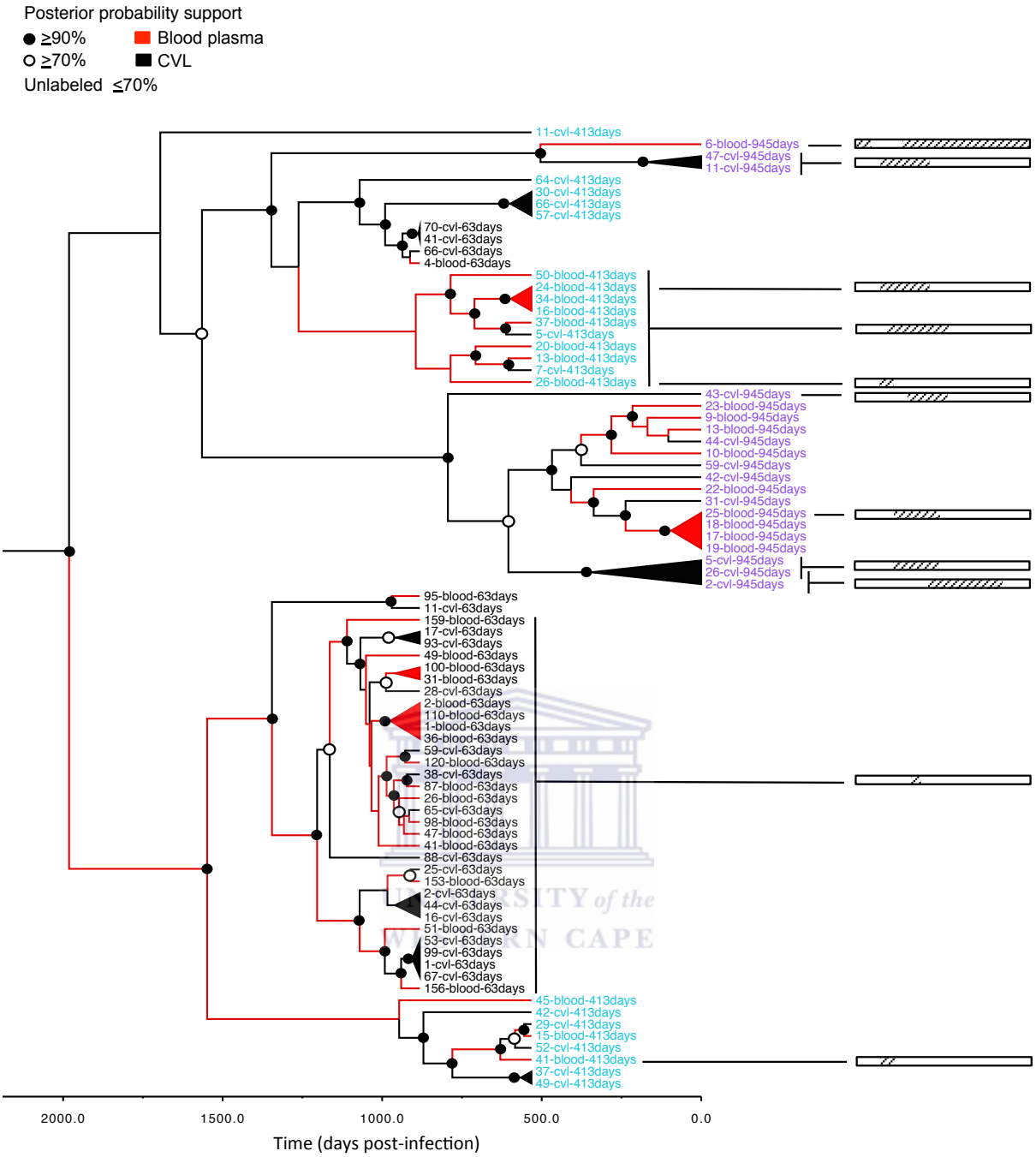


Figure 3.10b Time-scaled Bayesian MCC tree for CAP261 indicating sequences in which recombination breakpoints were detected using RDP4. Recombination block graphs displayed to the right of taxon names indicate unique recombination events as listed in Table 3.17. Branches are coloured according to the most probable state of their tissue origin where red represents viruses from the blood plasma and black indicates viruses from the cervix. Posterior probabilities ≥ 90% are indicated by a filled circled and ≥ 70% by an open circle at the nodes, with branch labels coloured according to the time points sampled.

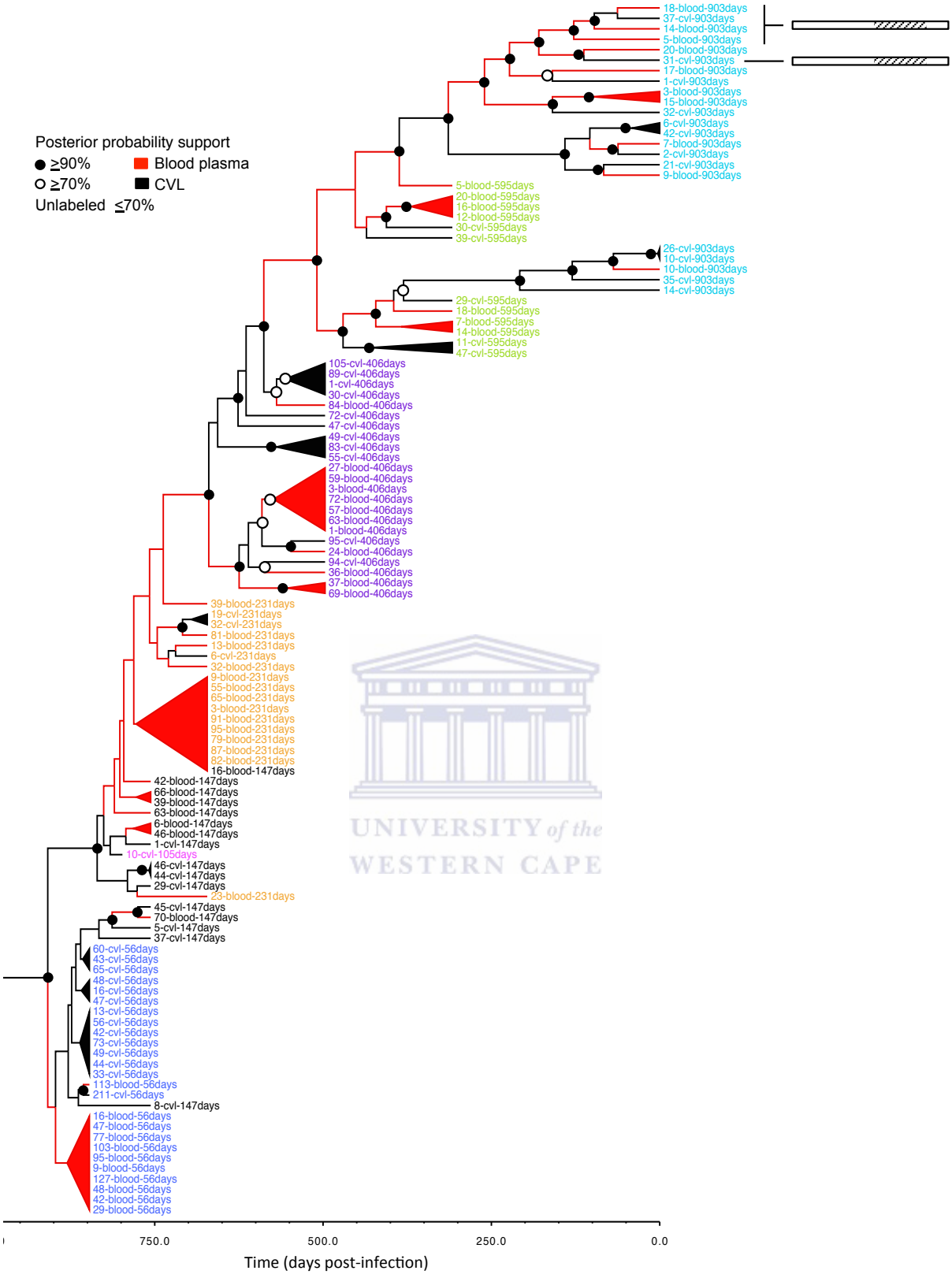


Figure 3.10c Time-scaled Bayesian MCC tree for CAP270 indicating sequences in which recombination breakpoints were detected using RDP4. Recombination block graphs displayed to the right of taxon names indicate unique recombination events as listed in Table 3.17. Branches are coloured according to the most probable state of their tissue origin where red represents viruses from the blood plasma and black indicates viruses from the cervix. Posterior probabilities ≥ 90% are indicated by a filled circled and ≥ 70% by an open circle at the nodes, with branch labels coloured according to the time points sampled.

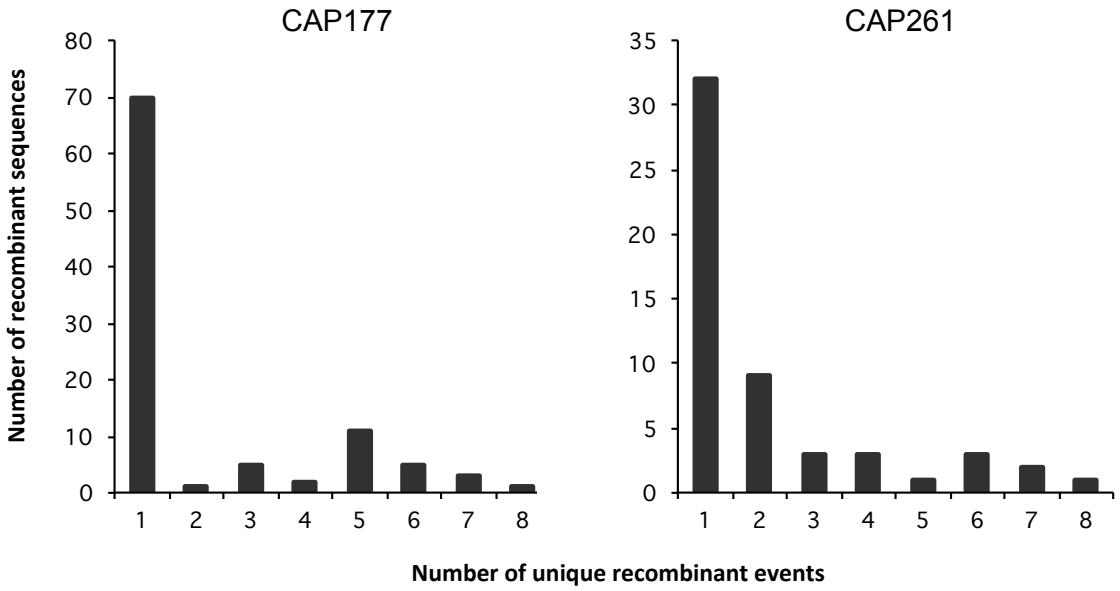


Figure 3.11 Summary of unique recombination events and recombinant sequences detected in participants CAP177 and CAP261. In CAP177, two unique recombination events were found in the majority of sequences, i.e. event 1 (47%) and event 5 (7%) whereas in CAP261, recombination events 1 and 2 were detected in the majority of sequences (40% and 11% respectively). There were also a few sequences that contained more than one recombination event in CAP177 (n = 8) and CAP261 (n = 2), however none of these recombinant sequences were unique to a particular tissue type or stage of infection.

The first recombination event in CAP177 (event 1) was detected at 14 days and subsequently at 28, 196 and 378 days, after which sequences displaying this event were no longer, found (Figure 3.10a). In sequences sampled at 28 days two further unique recombination events were detected (events 2 and 3) followed by another two at 196 days (events 4 and 5), then a single new event at 378 days (event 6) and finally, two new events at 1295 days (events 7 and 8).

In CAP261, the first recombination event (event 1) was detected in sequences sampled at 63 days after which it was no longer found. Three further unique recombination events then emerged at 413 days (events 2, 3 and 4) followed by another four at 945 days (events 5, 6, 7 and 8) (Figure 3.10b). Whereas in CAP270, the first and only detectable recombination event was found at 903 days, which was the last time point sampled for this participant (Figure 3.10c).

3.7.1 Tissue-specific recombination signatures

When the total numbers of recombinants were tallied per tissue type and time point no tissue-specific differences were evident. In CAP177, there were no CVL sequences available for comparison with blood plasma sequences at 14 and 196 days, whereas at 924 days only a single HIV-1 amplicon from the CVL was available compared to a total of four sequences from the blood plasma (Figure 3.12).

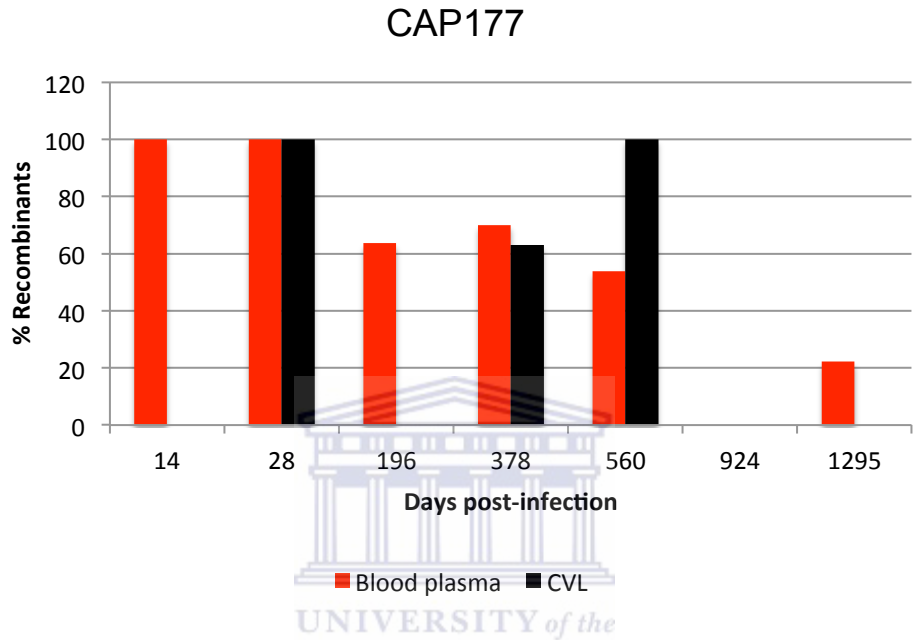


Figure 3.12 Percentage of recombinant viruses detected in blood plasma and CVL sequences from CAP177 over the sampling period. Time points where no recombinant sequences were detected within blood plasma or CVL samples are indicated by whitespaces.

The total number of recombinants appeared to fluctuate in CVL viral populations through time while the number of detectable recombinants in the blood plasma declined over the sampling period.

Patterns of recombination were further evaluated by analysing data related to parental sequences, including the tissue of origin and time points at which each parental sequence was detected. The percentage of total recombinant sequences with major and minor parents from the same or an earlier time point increased from 0% at 14 days to 79% at 378 days in CAP177, while recombinants with at least one parent from a later time point decreased from 100% at 14 days to 0% at 1295 days. No recombinants with a parent from a later time point could have been estimated at 1295 days as this was the last time point sampled in CAP177.

The number of recombinants with parents from the same tissue type also remained low (0 – 3%) at all time points, except at 378 days (17%), while the number of recombinants with parents from different tissue types was much higher but fluctuated from 100% at 14 days to 62% at 378 days, and finally to 8% at 1295 days. Nevertheless at all time points tested, there were more recombinants comprised of parents from different tissue types compared to those with parents from the same tissue type in this participant. One of the weaknesses of this approach however was that parents of recombinant sequences could have been unsampled at earlier time points, therefore if a major parent was found to be sampled after the time point when the recombinant sequence was detected for example, it is likely that the actual parent of the recombinant sequence could have been a relative of the predicted parental sequence, as estimated by RDP4.

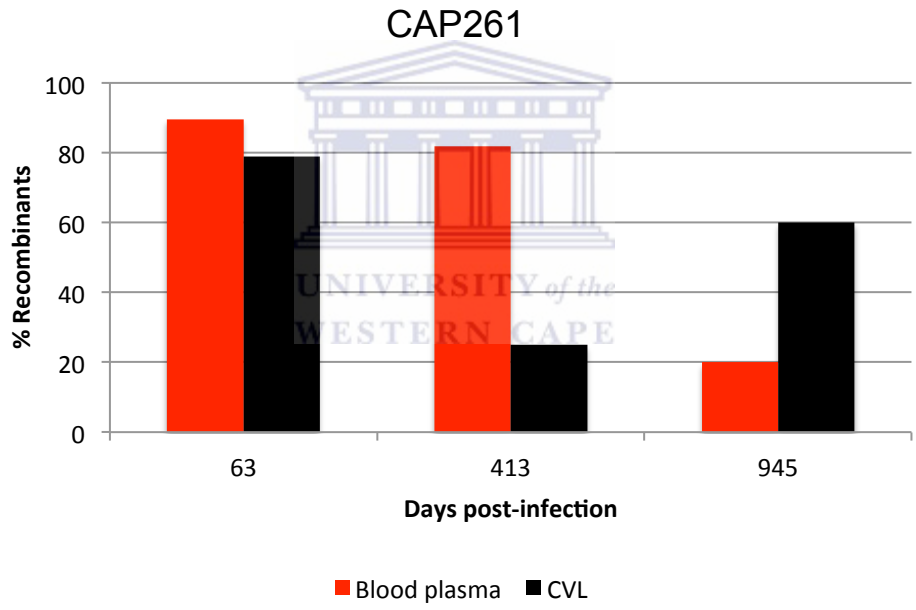


Figure 3.13 Percentage of recombinant viruses detected in blood plasma and CVL samples sequences from CAP261 over the sampling period. Although there were only sequences available from three time points for CAP261, the total number of recombinants fluctuated in frequency between 63 and 945 days in CVL sequences, whereas the number of recombinants in the blood plasma declined in frequency over the same period. Among the recombinant variants found in CAP261, 83% were comprised of a major and minor parent from the same tissue type whereas only 17% were made up of parents from different tissue types. A total of 36% were recombinant variants comprised of parents from the same or an earlier time point, while 64% were variants with at least one parent from a later time point.

When recombinants were analysed per time point the number of recombinants with major and minor parents from the same or an earlier time point increased from an absence in all sequences at 63 days to a presence in all sequences at 945 days. Recombinants with at least one parent from a later time point decreased from 100% at 63 days to 0% at 945 days, however no recombinants with a parent from a later time point could have been estimated at 945 days as this was the last time point sampled in CAP261. The total percentage of recombinants with parents from the same tissue type also decreased from 100% at 63 days, to 37% at 945 days, whereas the number of recombinants with parents from different tissue types increased from 0% at 63 days to 25% at 413 days, and finally to 75% at 945 days.

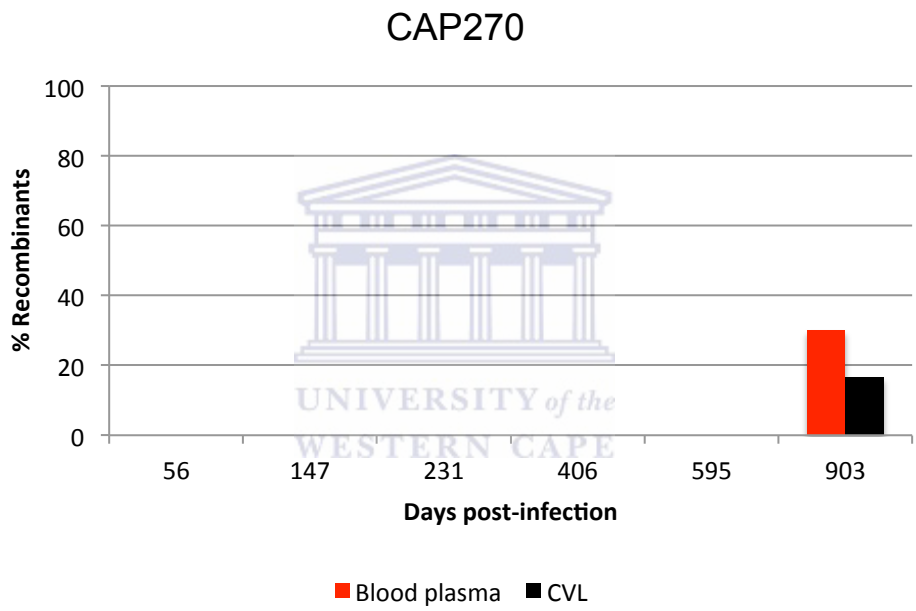


Figure 3.14 Percentage of recombinant viruses detected in blood plasma and CVL sequences from CAP270 over the sampling period. Time points where no recombinant sequences were detected within blood plasma or CVL samples are indicated by whitespaces.

Despite the availability of sequences from six time points for CAP270, only five recombinant sequences were detected at 903 days, three in the blood plasma and two in the CVL. All detected recombinant viruses were inferred to have minor and major parents from different tissue types, and showed no evidence of tissue-specific recombination.

3.7.2 Relationship between glycosylation and recombination

To test the hypothesis that the accumulation of potential N-linked glycosylation sites (PNLGs) along the *env* gene may be associated with recombination, all recombinant sequences were analysed for the location of recombination breakpoints. Although the detection of precise breakpoint positions are estimations only (Lamers *et al*., 2009), probable breakpoint positions from RDP4 were translated to amino acid base positions and compared to PNLG sites (predicted in section 3.6), across the gp120 region for each participant.

Table 3.18 Average number of PNLG sites within HIV-1 *env* subregions before and after the occurrence of a recombination event in CAP177, CAP217 and CAP270. Only regions where recombination events were detected between matched samples are shown below. Regions where the average number of PNLGs increased after the occurrence of a recombination event is shown in red text, with regions where the average number of PNLGs decreased after the occurrence of a recombination event shown in blue text. Regions where no change in the average number of PNLGs occurred before or after a recombination event are indicated in black text.

Participant ID	HIV-1 *env* region	Average number of PNLGs	
		Before or at first recombinant	After or at last recombinant
CAP177	Full-length gp120	23	31
	V1V2	4	8
	V4	3	3
	V5	2	1
	C1	1	2
	C2-C3	8	9
	C4	1	3
	C5	4	5
CAP261	Full-length gp120	29	32
	V1V2	7	8
	V4	3	4
	V5	1	1
	C1	1	1
	C2-C3	10	10
	C4	3	2
	C5	5	4

Participant ID	HIV-1 *env* region	Average number of PNLGs	
		Before or at first recombinant	After or at last recombinant
CAP270	Full-length gp120	31	30
	C1	1	2
	C5	4	4

A distinct difference in the average number of PNLGs before and after the occurrence of a recombination event was observed in all participants where recombination was detected. In 53% of the *env* subregions analysed, there was an increase in the number of PNLGs after the occurrence of a recombination event, whereas in only 21% there was a decrease in the number of PNLGs following recombination (Table 3.18). In 26% of the regions analysed, no change in PNLGs accumulation was observed before or after a recombination event.

In CAP177, there was an increase in the number of PNLGs in all regions analysed except in the V3 and V5-loops (Table 3.18), however differences between PNLGs numbers before and after the occurrence of a recombination event were not statistically significant (Mann-Whitney p-value $\geq$ 0.5604). In CAP261, although there were *env* regions where the number of PNLGs increased, decreased or remained unchanged following a recombination event, overall there was a greater number of regions where there was an increase in the number of PNLGs after recombination (Mann-Whitney p-value > 0.9999). Unsurprisingly, regions where there was no change in the number of PNLGs were mostly limited to regions of low variability, i.e. conserved regions C1, C2, C3, C5 and the V4-loop (Table 3.18).

In CAP270 all of the recombinant regions detected occurred in the conserved regions of the *env* gene (C1 and C5), and although a slight rise in the number of PNLGs was evident in sequences from this participant within the C1 region, bearing the signature of this recombination event the increase was not significant (Mann-Whitney p-value $\geq$ 0.8248).

3.8 Codon-based selection analysis

Selection analysis was performed using four different methods including the Single Likelihood Ancestor Counting (SLAC) method, the Fixed Effect Likelihood (FEL) method, the Mixed Effects Model of Evolution (MEME) and the Fast Unbiased Bayesian AppRoximation (FUBAR) method. Based on recommendations from other authors, only sites that were identified by at least three of these methods with significant statistical support ($p \leq 0.05$ or Posterior probability ≥ 0.9) were considered as credible evidence of positive or negative selection (Wlasiuk & Nachman, 2010; Castel *et al*., 2014).

Table 3.19 Percentage of amino acid (aa) sites identified as being under positive or negative selection along the gp120 region in blood plasma and CVL sequences using FUBAR, FEL and SLAC methods. Sites under negative selection were not tested for by MEME and have therefore been excluded from the table below. The "(+)" symbol indicates the total percentage of sites identified as being under positive selection, with "(-)" indicating the percentage of sites under negative selection.

Method	Participant ID							
	CAP177		CAP217		CAP261		CAP270	
	(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)
FUBAR	48%	52%	60%	40%	54%	46%	37%	63%
FEL	39%	61%	27%	73%	30%	70%	22%	78%
SLAC	38%	62%	19%	81%	15%	85%	29%	71%

Selection analyses revealed between 40% (with FUBAR) and 85% (with SLAC) of negatively selected sites in CAP217 and CAP261 respectively, and between 15% (with SLAC) and 60% (with FUBAR) of positively selected sites in CAP261 and CAP217 (Table 3.19). Overall, there was more evidence of sites under negative selection (65%) than positive selection (35%) among all participants, consistent with selection signatures observed by others (Bazykin *et al*., 2006; Frank, 2002). Nevertheless, strong evidence of positive selection was detected for a number of sites in each participant, by all four methods (Table 3.20).

Table 3.20 Codon-based sites along the translated gp120 *env* region identified as being under positive selection in each participant using SLAC, FEL, MEME and FUBAR methods. The "*" symbol indicates sites with significant statistical support for positive selection that coincided with a PNLG site on the translated *env* gene, with sites that were predicted by all four methods indicated in red text.

Participant ID	Method			
	FUBAR	FEL	MEME	SLAC
CAP177	82, 139, 158, 281, 294, 304, 306, 308, 333, 342, 349*, 363, 366, 367, 419, 424*, 438, 453, 664, 687, 756, 820, 823, 825, 829, 876	158, 281, 294, 304, 306, 308, 333, 342, 349*, 363, 366, 367, 419, 424*, 438, 453, 664, 687, 756, 820, 823, 825, 829, 876	82, 158, 281, 294, 304, 306, 308, 333, 342, 349*, 363, 366, 367, 419, 424*, 438, 453, 664, 687, 756, 820, 823, 825, 829, 876	82, 158, 294, 306
CAP217	27, 129, 354, 472, 474*, 509, 530, 800, 833	27, 129, 354, 472, 474*, 509, 530, 800, 833	27, 129, 354, 472, 474*, 509, 530, 800, 833	27, 474*
CAP261	132, 189*, 359, 396, 412*, 515, 655, 789, 831	132, 189*, 359, 396, 412*, 515, 655, 789, 831	132, 189*, 359, 396, 412*, 515, 655, 789, 831	189*, 359, 396, 412*, 789
CAP270	129, 132, 154*, 180, 182, 188, 293, 374, 866	129, 132, 180, 182, 188, 293, 374, 866	129, 132, 154*, 180, 182, 188, 293, 374, 866	129, 132, 154*, 180, 182, 188, 374, 866

FUBAR, FEL and MEME results were mainly in agreement in their prediction of sites, whereas the SLAC method was more conservative. Evidence of positive selection was found in both conserved and variable regions along the gp120 *env* region in all participants. A total of 17 positively selected sites were detected by all four methods, three of which (aa 474 in CAP217, and aa sites189 and 412 in CAP261) were found to coincide with a previously predicted PNLG site. Overall, at least one site that was predicted to be under positive selection, coincided with a PNLG site in each participant (Table 3.20).

To assess if there were any tissue-specific differences in selection patterns, blood plasma and CVL sequences were analysed separately per participant (Figure 3.15).

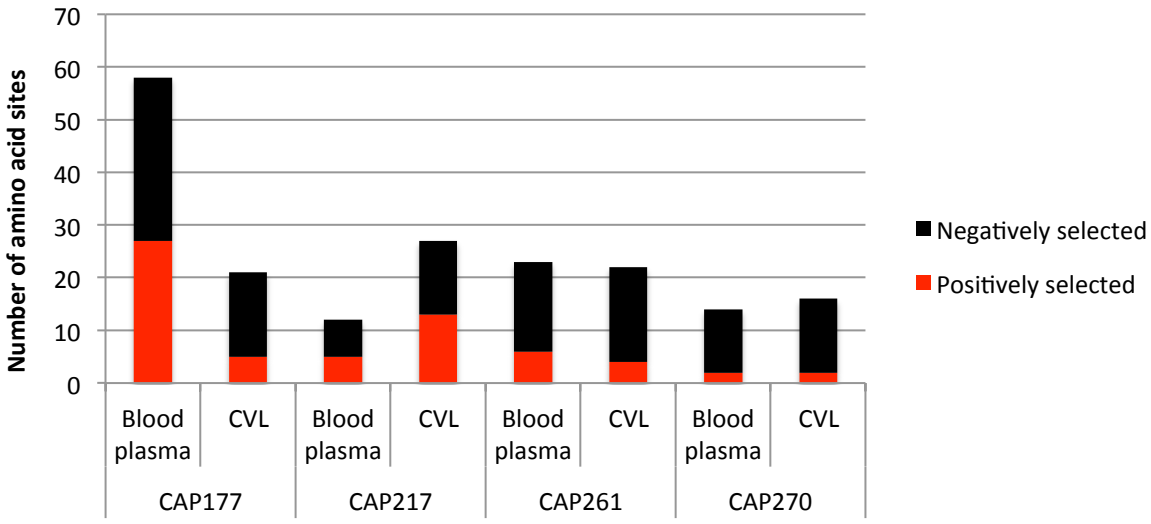


Figure 3.15 Number of sites identified as being under positive or negative selection in blood plasma and CVL sequences with significant statistical support by at least three different methods.

In participants CAP261 and CAP270 almost identical selection signatures were seen in both tissues (i.e. between 93 and 98% equivalent), whereas in CAP177 and CAP217 noticeably larger differences were evident (Figure 3.15). In CAP177, 47% more sites were found to be under both positive and negative selection in sequences isolated from the blood plasma compared to the amount found in CVL sequences. In CAP217 there was a slightly smaller tissue-specific difference (38%), with more sites predicted in CVL sequences instead of blood plasma.

3.9 Prediction of HIV-1 co-receptor tropism

Genotypic analysis of HIV-1 co-receptor tropism performed on the V3-loop region of all sequences obtained from the blood plasma and cervix showed that all except one sequence from CAP177 sampled at 378 days were CCR5-trophic. Upon closer inspection this sequence contained a deletion in the V3-loop, which resulted in a coding frame shift and an incorrect prediction, and was subsequently excluded from any further analyses. None of the participants experienced tropism changes between chronic and acute infection stages, despite this phenomenon having been previously associated with disease progression (Connor *et al*., 1997; Burger & Hoover, 2008; Cecilia *et al*., 2000).

To further evaluate the role of the V3-loop in modulating gp120 structure, *net charges* were analysed. It has been reported that the V3-loops of CCR5-tropic viruses are usually less positively charged than CXCR4-tropic viruses and an increase in net charge has been linked to the conversion of CCR5-tropic viruses into CXCR4-tropic (Yokoyama *et al*., 2012; Speck *et al*., 1997; Fouchier *et al*., 1992; Kato *et al*., 1999). Positively charged amino acids at positions 11 and 25, increases in total net charge or decreases in the number of PNLGs in the V3-loop have also been associated with CXCR4 co-receptor usage (De Jong *et al*., 1992; Hoffman *et al*., 2002; Fouchier *et al*., 1995).

On average, net charge analysis revealed a slightly higher V3-loop charge in viruses from the blood plasma compared to viruses from the cervix in all participants. From the total of 18 time points where sequences were available from both tissue types among all participants, 7 showed a higher V3-loop charge in viruses from the blood plasma, whereas an equal charge in both blood plasma and CVL viruses was seen at 9 time points, and 2 showed a higher charge in CVL viruses (Table 3.21).

Table 3.21 Average V3-loop net charges of HIV-1 blood plasma and CVL *env* sequences. Time points at which the average V3-loop charge was higher in the blood plasma is indicated in red text, with higher charges in CVL sequences shown in blue.

Participant ID	Days p.i.	Average V3-loop net charge	
		Blood plasma	CVL
CAP177	14	4.00	–
	28	4.00	4.00
	196	3.82	–
	378	3.70	3.68
	560	4.25	3.20
	924	5.00	5.00
	1295	4.06	4.06
CAP217	14	6.00	–
	63	6.00	6.00
	420	6.83	6.40
	770	6.00	5.80
	1316	3.84	3.95
CAP261	63	3.95	3.84
	413	3.91	3.58
	945	3.90	3.90

Participant ID	Days p.i.	Average V3-loop net charge	
		Blood plasma	CVL
CAP270	56	4.00	4.00
	105	–	4.00
	147	4.00	4.00
	231	4.00	4.00
	406	4.00	4.00
	595	4.00	4.00
	903	3.80	3.83

Although there were sequences available from more time points in some participants than others, overall considerably higher V3-loop net charges were found in blood plasma sequences. No consistent tissue-specific patterns in V3-loop net charges could be inferred in each participant, however based on reports from literature, there is a more likely possibility of co-receptor switching in viruses present in the blood plasma compared to those present in the cervix further into disease progression, based on the higher positive V3-loop net charges within sequences from this tissue.

3.10 Identification of APOBEC-induced hypermutation

No evidence for the presence of hypermutation was found in any of the sequences from the blood plasma or CVL in any participant (Fisher's Exact test $p > 0.05$), however, a higher number of G to A mutations were detected in sequences from all participants and in both tissue types, with the exception of CVL sequences from participant CAP177 (Figure 3.16a). A larger number of A to G mutations were seen in CVL sequences from CAP177 at all time points where sequences from both tissues were available, throughout the sampling period. Interestingly, the opposite pattern was observed in blood plasma sequences from CAP177, where G to A mutations were consistently higher throughout the course of infection (Figure 3.16a).

In participants CAP217 and CAP270, both G to A and A to G mutation patterns were alike in both tissues (Figure 3.16a-b). Sequences from the blood plasma showed only a slight difference in the average numbers of G to A and A to G mutations in CAP261, whereas in CVL sequences, considerable differences between all sampled time points were detected (Figure 3.16b). Generally, hypermutation patterns remained consistent in all participants and in both blood plasma and CVL sequences, i.e. either more G to A mutations or more A to G mutations that persisted throughout the sampling period.

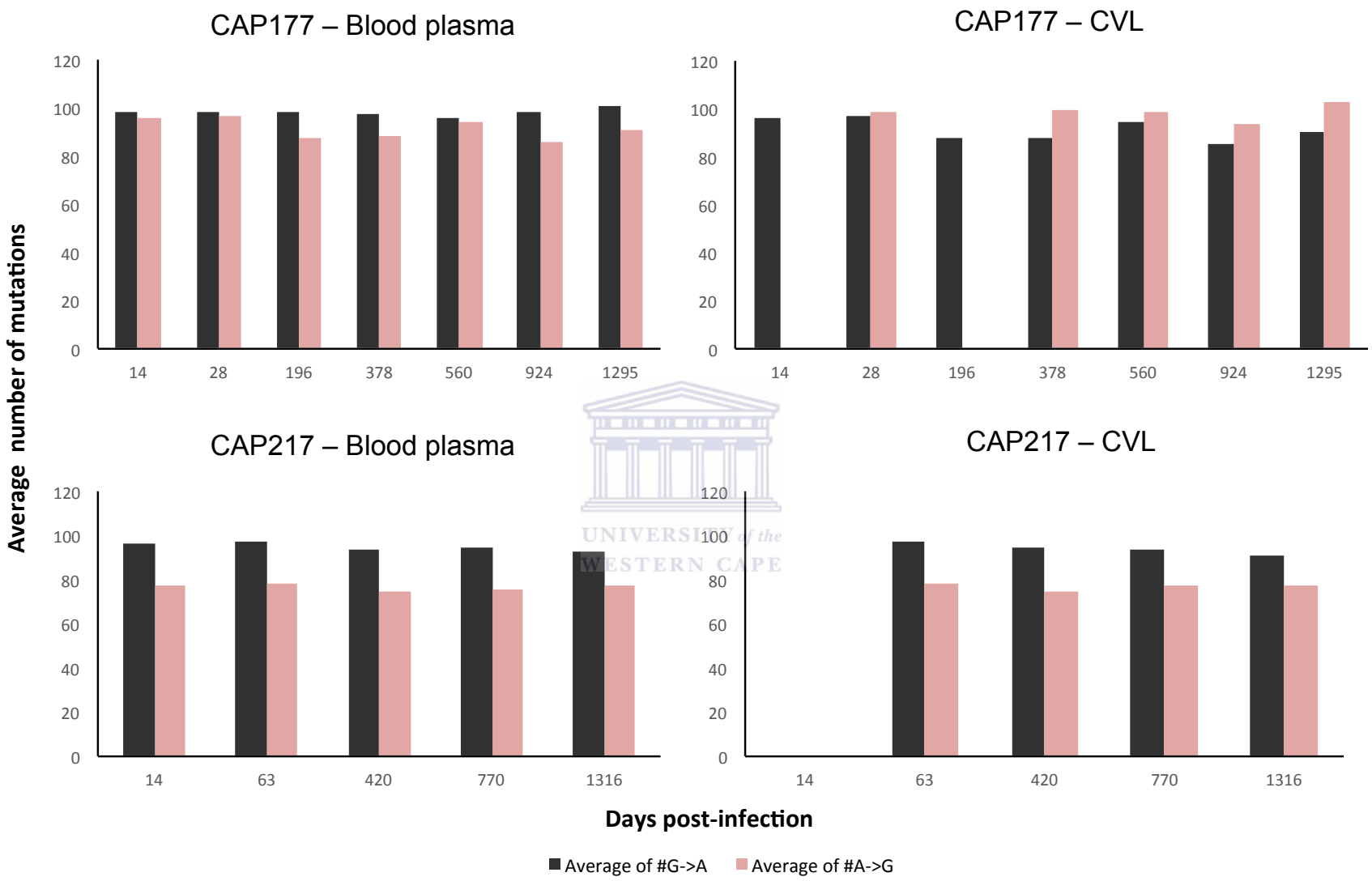


Figure 3.16a Column chart depicting hypermutation patterns within blood plasma and CVL sequences from participants CAP177 and CAP217 over the course of infection. The average numbers of Guanine to Adenine (G to A) mutations are shown in black, with the average numbers of Adenine to Guanine (A to G) mutations shown in red. Time points where G to A or A to G mutations were not detected within blood plasma or CVL sequences are indicated by whitespaces.

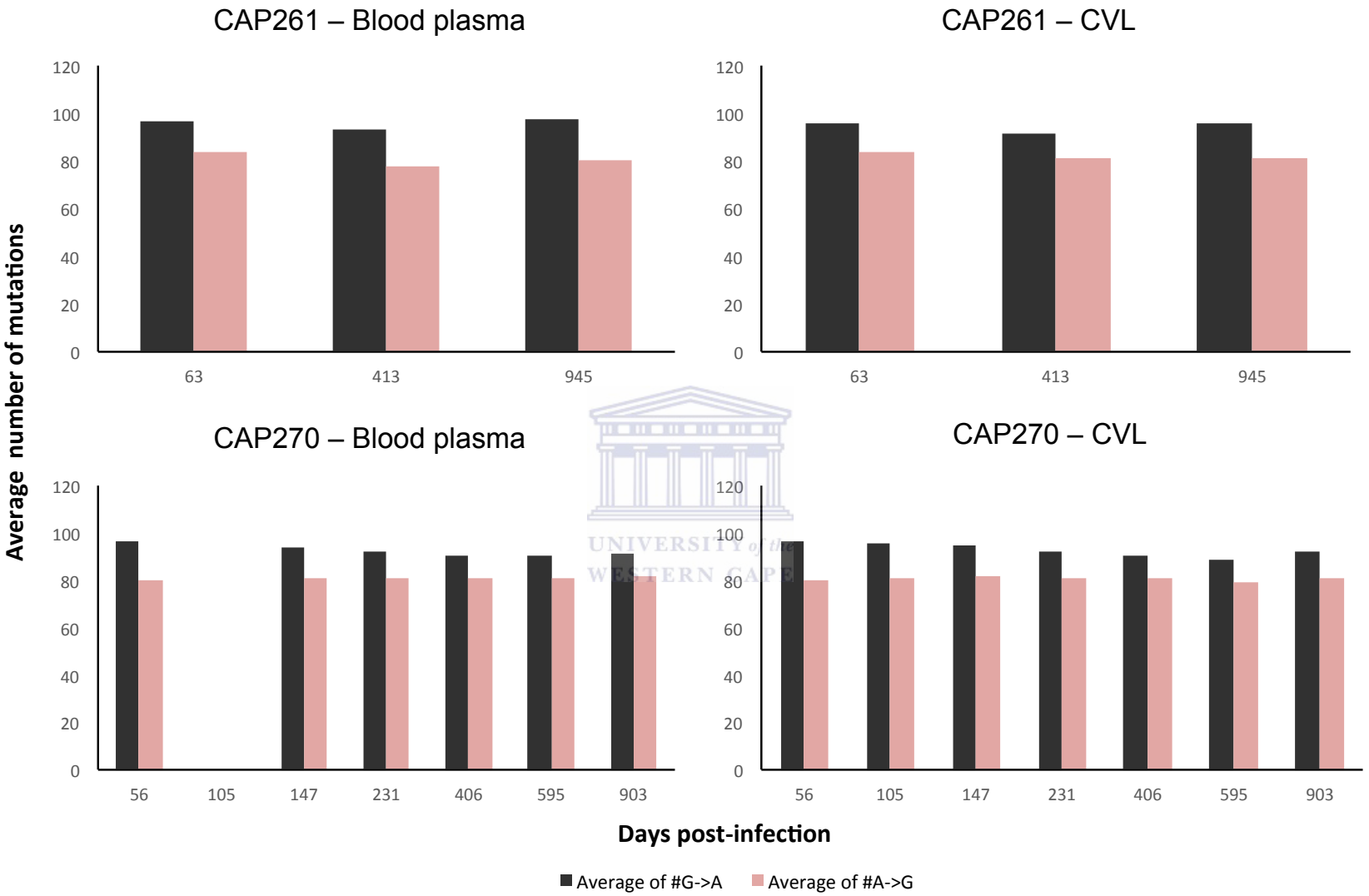


Figure 3.16b Column chart depicting hypermutation patterns within blood plasma and CVL sequences from participants CAP261 and CAP270 over the course of infection. The average numbers of Guanine to Adenine (G to A) mutations are shown in black, with the average numbers of Adenine to Guanine (A to G) mutations shown in red. Time points where G to A or A to G mutations were not detected within blood plasma or CVL sequences are indicated by whitespaces.

3.11 Length variation within the *env* V-loop regions

V-loop lengths generally increased over the sampling period in all of the V-loop regions except in V3 and V5-loops in all participants' sequences, irrespective of tissue type (Figure 3.17). The greatest increases in length were observed in the V1V2-loops, which ranged from 56 amino acid (aa) bases in CAP217 to 81 bases in CAP270, while the V4-loop ranged from 18 aa bases in CAP177 to 35 bases in CAP261, and in the V5-loop from 9 aa bases in CAP270 to 17 bases in CAP217 (Table 3.22).

Table 3.22 Summary of the minimum, maximum, median and mean V1V2, V4 and V5-loop lengths in blood plasma and CVL sequences among all participants. Values shown in the table below indicate loop length in amino acid bases.

Participant ID	Statistic	V1V2		V4		V5	
		Blood plasma	CVL	Blood plasma	CVL	Blood plasma	CVL
CAP177	Mean	63	64	20	21	13	13
	Median	60	62	20	20	13	13
	Maximum	80	75	26	29	15	15
	Minimum	57	57	18	18	11	11
CAP217	Mean	71	69	23	22	14	14
	Median	74	67	23	23	17	15
	Maximum	78	74	23	23	17	17
	Minimum	56	59	20	20	10	10
CAP261	Mean	72	71	28	28	14	13
	Median	73	73	28	28	14	12
	Maximum	77	77	29	35	17	17
	Minimum	65	60	28	22	10	10
CAP270	Mean	77	76	23	24	12	12
	Median	76	76	22	22	12	12
	Maximum	81	81	28	28	12	15
	Minimum	66	66	19	19	9	10

CAP177 and CAP261 displayed similar average V-loop lengths (CAP177 mean = 32, range = 11-80; CAP261 mean = 38, range = 10-77 respectively) over the sampling period compared to participants CAP217 and CAP270 (CAP217 mean = 35, range = 10-78; CAP270 mean = 37, range = 9-81 respectively). Nevertheless, significant tissue-specific differences between V-loop lengths were discovered at isolated timepoints in the V1V2 and V4-loops in sequences from CAP177, CAP217 and CAP261, but not CAP270 (Table 3.23).

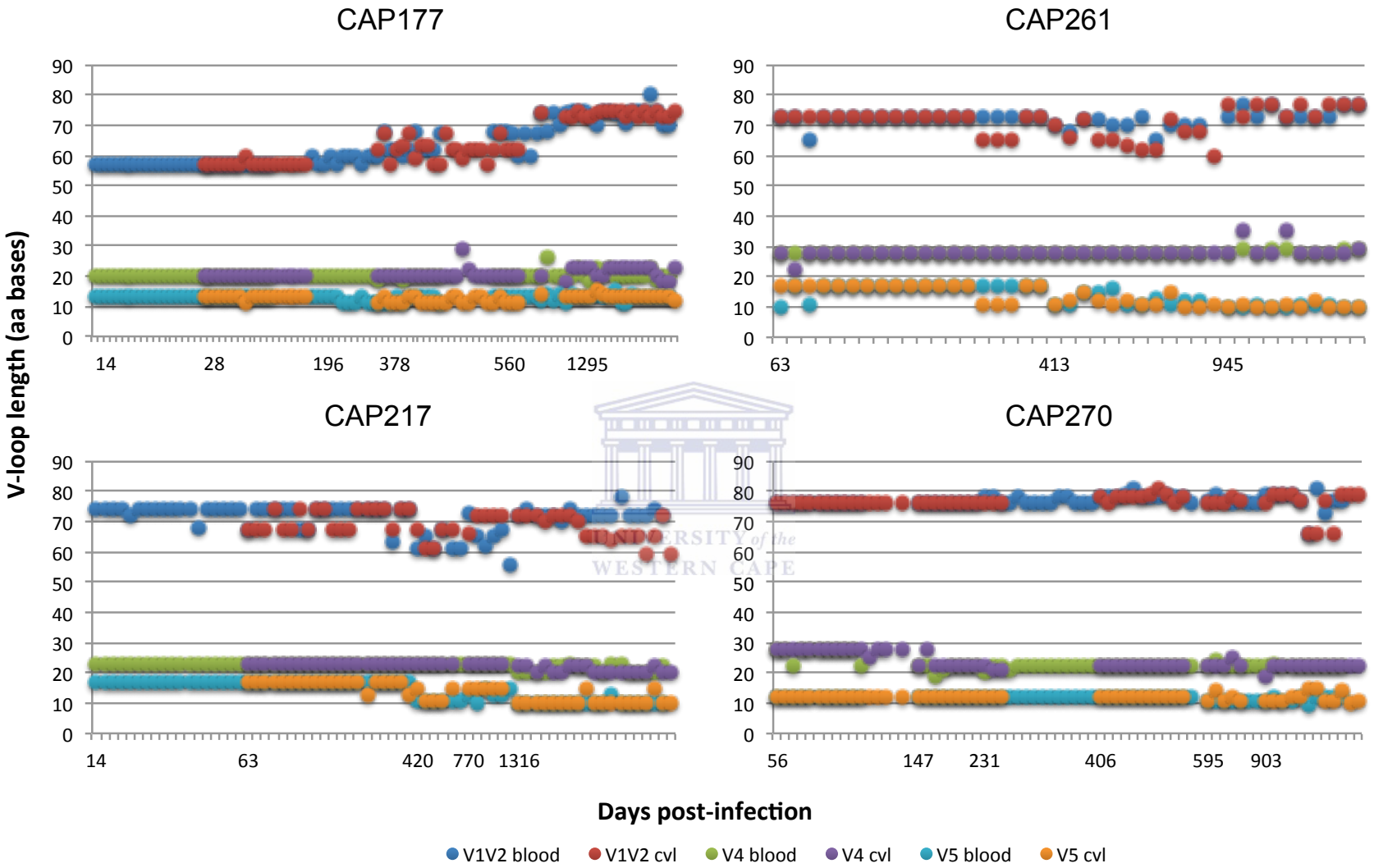


Figure 3.17 Variable-loop length changes in the V1V2, V4 and V5-loop regions among all participants over the sampling period. The actual V-loop lengths per sample are presented on the scatterplot, separated by tissue type for each V-loop region, where the V-loop length is shown on the y-axis in amino acid bases and the time in days post-infection on the x-axis.

Table 3.23 Statistical assessment of tissue-specific differences in V-loop lengths using the Mann-Whitney and Wilcoxin signed-ranked non-parametric tests. Tests where the standard deviation was equal to zero or where an "invalid floating point operation" error was received (i.e. where GraphPad was unable to fit the data using the selected model and options (Motulsky, 2003)), are indicated by the "–" symbol. P-values below 0.05 for both tests were considered as credible evidence of significant tissue-specific differences between V-loop lengths.

Participant ID	HIV-1 *env* region	Days p.i.	Mann-Whitney U test		Wilcoxin signed-rank	
			U value	p-value	W value	p-value
CAP177	V1V2	28	–	–	–	–
		378	125	0.3798	15	0.4922
		560	29	0.2250	11	0.1875
		1295	183	0.7191	-7	0.7646
	V4	28	–	–	–	–
		378	133	0.0452	–	–
		560	–	–	–	–
		1295	234	0.0394	-69	0.0295
	V5	28	–	–	–	–
		378	123	0.3671	9	0.3125
		560	–	–	–	–
		1295	172	0.9845	-13	0.3828
CAP217	V1V2	63	255	0.0763	32	0.0547
		420	21	0.3061	-4	0.5000
		770	24	0.1114	-9	0.3125
		1316	289	0.0001	99	0.0006
	V4	63	–	–	–	–
		420	–	–	–	–
		770	–	–	–	–
		1316	204	0.2616	22	0.3652
	V5	63	–	–	–	–
		420	–	–	–	–
		770	–	–	–	–
		1316	181	0.5634	-4	0.5000
CAP261	V1V2	63	200	0.3098	5	0.3750
		413	104	0.0187	40	0.0117
		945	65	0.2040	-9	0.3125
	V4	63	–	–	–	–
		413	–	–	–	–
		945	55	0.6960	-4	0.6250
	V5	63	189	0.7090	0	> 0.9999
		413	84	0.2710	16	0.3125
		945	53	0.8213	0	> 0.9999

Participant ID	HIV-1 *env* region	Days p.i.	Mann-Whitney U test		Wilcoxin signed-rank	
			U value	p-value	W value	p-value
CAP270	V1V2	56	–	–	–	–
		147	–	–	–	–
		231	–	–	–	–
		406	70	0.8331	7	0.5625
		595	20	0.7235	4	0.6250
		903	68	0.6247	7	0.4375
	V4	56	87	0.3827	–	–
		147	43	0.1058	-6	0.2500
		231	31	0.1063	0	> 0.9999
		406	–	–	–	–
		595	19	0.8028	-1	> 0.9999
		903	71	0.1864	3	0.5000
	V5	56	–	–	–	–
		147	–	–	–	–
		231	–	–	–	–
		406	–	–	–	–
		595	–	–	–	–
		903	60	0.9720	-13	0.2969

Significant differences in V-loop lengths between blood plasma and CVL sequences were only observed at three isolated time points (i.e. 3 of the 51 paired samples tested) among all participants. None of the estimated significant differences coincided with early stages of infection, and two of the three tests that were significant for tissue-specific differences in V-loop lengths occurred within the V1V2-loops (Table 3.23), a region that is known to be highly variable. Generally however, although V-loops gradually increased in base length over the course of infection in all participants (Figure 3.17), changes were mostly similar in both blood plasma and CVL viruses (Table 3.23).

3.11.1 Relationship between V-loop expansion and PNLGs accumulation

In CAP177 both the number of PNLGs in the V1V2-loops and the lengths of these loops increased over time, ranging from 5 PNLGs and an average length of 57 amino acid bases at 28 days, to 15 PNLGs and 74 bases in length at 1295 days, while V4-loop lengths increased moderately from 3 to 5 PNLGs between 14 and 1295 days (Figure 3.17).

A similar pattern was observed in CAP261, with increases in both the number of PNLGs and base length of the V1V2 and V4-loops over time, ranging from 8 to 11 PNLGs and an average length of 72 to 75 amino acid bases between 63 and 945 days in the V1V2-loop, and from 6 to 9 PNLGs and 22 to 35 bases between 63 and 945 days in the V4-loop (Figure 3.17).

In CAP217 and CAP270 all detectable PNLG site gains or losses occurred in the V1V2-loops in the former participant while almost half (47%) occurred in the latter. Overall there was an increase in V1V2-loop lengths ranging from 76 amino acid bases at 56 days to 81 bases at 903 days in CAP270. In CAP217 V1V2-loop lengths ranged from 71 amino acid bases at 63 days to 70 bases at 1316 days. Similarly in CAP270, increases in the length of the V4-loops ranged from an average of 27 amino acid bases at 56 days to 22 bases at 903 days, corresponding with PNLGs losses in this region at positions N400 and N407, both of which appeared to decline from being present in 82 and 100% of sequences at 56 days to 0 and 42% respectively at 903 days.

Table 3.24 Comparison of significant tissue-specific differences in the number of PNLG sites and V-loop lengths using the Mann-Whitney non-parametric test. Time points where significant differences in both the number of PNLG sites and V-loop lengths between blood plasma and CVL sequences were estimated are shown in red text.

Participant ID	HIV-1 *env* region	Days p.i.	Mann-Whitney p-values	
			PNLGs	V-loop lengths
CAP177	V1V2	28	–	–
		378	0.1170	0.3798
		560	0.0191	0.2250
		1295	0.0078	0.7191
	V4	28	–	–
		378	0.4606	0.0452
		560	0.0244	–
		1295	0.0002	0.0394
	V5	28	–	–
		378	0.5340	0.3671
		560	–	–
		1295	0.1503	0.9845

Participant ID	HIV-1 *env* region	Days p.i.	Mann-Whitney p-values	
			PNLGs	V-loop lengths
CAP217	V1V2	63	0.0515	0.0763
		420	0.7621	0.3061
		770	–	0.1114
		1316	0.0600	0.0001
	V4	63	–	–
		420	–	–
		770	–	–
		1316	–	0.2616
	V5	63	–	–
		420	–	–
		770	0.0578	–
		1316	–	0.5634
CAP261	V1V2	63	0.6646	0.3098
		413	0.0216	0.0187
		945	0. 7071	0.2040
	V4	63	0.2058	–
		413	0.1771	–
		945	0.8146	0.6960
	V5	63	–	0.7090
		413	0.8590	0.2710
		945	–	0.8213
CAP270	V1V2	56	–	–
		147	0.9443	–
		231	–	–
		406	0.4203	0.8331
		595	0.9294	0.7235
		903	0.9281	0.6247
	V4	56	0.4096	0.3827
		147	0.8577	0.1058
		231	0.4256	0.1063
		406	–	–
		595	–	0.8028
		903	0.4020	0.1864
	V5	56	–	–
		147	–	–
		231	–	–
		406	–	–
		595	–	–
		903	0.6981	0.9720

Significant tissue-specific differences in both PNLGs numbers and V-loop lengths were only evident in two participants at isolated time points ((Table 3.24) and in different V-loop regions (i.e. within the V4-loop in CAP177 sequences and V1V2-loop in CAP261 sequences), however there was no evidence of similar differences in 96% of the paired samples tested, suggesting a relatively similar pattern in PNLGs accumulation and V-loop length variation in viruses from both tissue types.

3.11.2 Relationship between V-loop length variation and recombination

To test the hypothesis that recombination may be associated with V-loop length variation, average V-loop lengths were compared between sequences taken before and after the occurrence of a recombination event in all participants that exhibited evidence of recombination (section 3.7).

Table 3.25 Average V-loop lengths before and after the first and last occurrence of a recombination event in participants CAP177, CAP217 and CAP270. Regions where the V-loop length increased after the occurrence of a recombination event is shown in red text, with regions where the average number of PNLGs decreased after the occurrence of a recombination event shown in blue text. Regions where no changes in V-loop lengths were found before or after a recombination event occurred are indicated in black text.

Participant ID	V-loop region	Average loop length (amino acid bases)	
		Before or at first recombinant	After or at last recombinant
CAP177	V1V2	57	73
	V3	34	34
	V4	20	21
	V5	13	13
CAP261	V1V2	72	75
	V3	34	34
	V4	28	29
	V5	16	10
CAP270	V1V2	76	76
	V3	34	34
	V4	27	22
	V5	12	12

V-loop lengths were found to have increased in the V1V2- and V4-loops within viruses from participants CAP177 and CAP261, the two participants that harboured the most recombinant viral populations (Table 3.17), whereas no such increase was noted in CAP270. V3- and V5-loop lengths remained consistent among all participants despite the presence of recombination, with the exception of the V5-loop in CAP261 sequences, where an average decrease in length was discovered. Generally however, the occurrence of a recombination event could not conclusively be linked to V-loop length expansion, but could potentially be associated with multi-variant infections.

Chapter 4
Discussion

Defining the characteristics of HIV compartmentalization and evolution in specific tissue types may offer important clues to our understanding of transmission dynamics and disease persistence (Heath *et al*., 2009; Santoro & Perno, 2013), yet despite its importance in the control of HIV infection, only a few studies have characterized HIV evolution in the female genital tract over long-term infection. Many studies that have investigated tissue-specific viral compartmentalization and its role in disease persistence have reported conflicting results (Philpott *et al*., 2005; Kemal *et al*., 2003; Bull *et al*., 2013; Zhu *et al*., 1996; Diem *et al*., 2008). From those that have reported the existence of distinct viral populations within the female genital tract, compartmentalization patterns were defined mainly on visual inspection of phylogenetic trees, while other studies focused on *env* subregions or reported evidence for compartmentalization at isolated time points during the course of infection.

Furthermore, some studies did not employ measures to avoid the misrepresentation of compartmentalization patterns caused by monotypic and low diversity sequences, or account for the number of transmitted viral variants, or the presence of recombination. In contrast, studies that did account for test biases and employed both tree and distance-based methods found no persistent tissue-specific compartmentalization, although despite the lack of consistent structure over long-term infection, reports of "transient", "partial", and "subcompartmentalization" patterns were described. This study therefore sought to investigate HIV evolution between the female genital tract and blood plasma in a treatment naïve cohort over acute and chronic infection, using recommended methods for compartmentalization detection (Zárate *et al*., 2007).

Given the lower quantities of viruses present in the genital tract, it was hypothesized that compartmentalization of HIV viruses could exist within the genital tract. HIV-1 subtype C sequences were subsequently amplified, cloned and sequenced from CVL and blood plasma samples obtained from four treatment naïve women using single genome amplification methods that minimized template resampling and PCR recombination.

Contaminated samples, such as CVL samples containing any trace of blood were also omitted from further analysis to eliminate the presence of potential variants from the blood plasma, similar to guidelines followed by Imamichi *et al*. (2011).

To accurately detect viral compartmentalization, confounding factors such as the presence of single or multiple HIV-1 variants within a single participant, had to be identified and to do this, information on the shape and arrangement of phylogenetic trees, the estimated time to the most recent common ancestor (tMRCA), Poisson distributions, pairwise genetic distances and results from statistical tests were analysed. Differences in V-loop length variation, potential N-linked glycosylation site accumulation, recombination patterns and clinical indicators were also investigated to assess differences between viral populations from the blood plasma and cervical compartments within each participant as well as between inferred single and multiple-variant infected participants. To this end, the number of transmitted viral variants was first examined, as it greatly affects the elucidation of almost all other downstream analyses performed in this study.

4.1 Estimating the number of transmitted viral variants

Of the four participants studied here, two (CAP270 and CAP217) were likely infected by a single viral variant and two (CAP177 and CAP261) by multiple HIV variants. As a result, estimates of pairwise genetic distances and the number of detectable recombination events were significantly lower in the former participants compared to the latter, consistent with previous studies (Novitsky *et al*., 2011; Carvajal-Rodríguez *et al*., 2008; Aulicino *et al*., 2011; Lemey *et al*., 2007; Lukashov & Goudsmit, 1997). TMRCA estimates for participants CAP177 and CAP261 also exceeded the known time of infection substantially, compared to estimates for CAP217 and CAP270, characteristic of infection by more than one viral variant (Keele *et al*., 2008; Abrahams *et al*., 2009). Strong evidence for the presence of multiple variants in CAP177 and CAP261 was also discovered in phylogenetic analyses, where multiple lineages were visible on cross-sectional maximum likelihood and longitudinal Bayesian MCC trees, similar to phylogenetic structure observed by others (Novitsky *et al*., 2011; Abrahams *et al*., 2009).

Although CAP177 and CAP261 are considered to be examples of multi-variant infections, the question of whether these women were initially infected by multiple HIV variants at the onset of transmission or were superinfected later on, remains unclear. Since sequences from only the blood plasma were available at 14 days post-infection for CAP177, the presence of a highly divergent minor variant in the CVL, that was not present in the blood plasma at 28 days could also be explained by the transmission of a second viral variant in the period separating these two sampling points, suggesting superinfection in this participant (Van der Kuyl & Cornelissen, 2007). However, given that only four clones were sequenced from the blood plasma at 14 days in CAP177 during acute infection when the viral load was high, it is more plausible that the sample size was simply too small to reliably detect viral variants that accounted for less than 30% of the population at this time-point. In principle, the same reasoning could be applied to the other participant, CAP261, which has been described as likely involving a multi-variant transmission, since the first viruses to be successfully amplified from both the blood plasma and CVL were from samples collected at 63 days post-infection, i.e. two months after the initial transmission event.

In both CAP177 and CAP261 despite the presence of multiple variants, not all lineages persisted to subsequent sampling time points, possibly due to a selective advantage as it has been reported that some viral variants with greater competitive ability or superior fitness actually displace viruses from the original infection (Kiwelu *et al*., 2013; Templeton *et al*., 2009; Van der Kuyl & Cornelissen, 2007).

4.2 Study population and clinical indicators

Blood plasma viral loads during chronic and acute infection is known to be a strong predictor of viral loads within the female genital tract (Kovacs *et al*., 2001; Lavreys *et al*., 2006) and accordingly, the four women from whom it was possible to amplify CVL viruses had among the highest blood plasma viral loads from the 18 originally analysed. Only five of the original 18 women in the CAPRISA cohort were sampled within 30 days of infection, however amplification of *env* genes was not possible from many samples taken and the only participant for whom it was possible to amplify viruses from all time points sampled was CAP270.

HIV infection is normally characterized by an acute phase during which viral loads are high, followed by a decrease as host immunity develops eventually leading to the chronic phase of infection, which can last several years (depending on the progressor type), until the onset of AIDS. Among the participants studied here, all exhibited typical signs of disease progression. Viral loads increased during the acute phase and remained steady in participants CAP177, CAP261 and CAP217, whereas in CAP270 disease progression was substantially faster.

Despite the rise in viral loads, CD4 counts remained stable in CAP261, indicative of patterns observed in "HIV-1 controllers" or long-term non-progressors, which has also previously been observed within multiple-variant infected patient groups (Lamine *et al*., 2007; Casado *et al*., 2007; Van der Kuyl & Cornelissen, 2007). Viral load control was also evident in CAP217 and CAP177 to a lesser degree, where despite slight fluctuations in viral loads, CD4 cell counts appeared to decline at a slow rate, while viral loads appeared to remain relatively stable over the sampling period in both participants. In CAP270 however, the only participant in whom viral loads were exponentially higher than CD4 cell counts, CD4 cell counts declined rapidly as viral loads grew. This participant was later classified as a rapid progressor and was subsequently placed on ARVs following the sampling period analysed here. Alternatively in CAP177, the only participant in whom a significant correlation between viral genetic diversity and viral load was detected (Table 3.12), it is likely that this result was confounded by time since early infections are associated with high viral loads and a largely homologous founder population followed by a period of strong immune response, which results in a viral load decrease and selection of immune escape variants associated with an increase in diversity (van Deutekom *et al*., 2013; Troyer *et al*., 2005).

In all four women there was also consistent evidence of coinfection with single or multiple STIs, which has previously been associated with an increase in inflammation and leukocytes in the genital tract, factors that have been reported to significantly increase HIV shedding (Johnson & Lewis, 2008; Anderson & Cu-Uvin, 2008). Overall there was a high prevalence of *HSV type 2* and *B. vaginosis* coinfection among all participants, with other STIs arising along the sampling period that did not appear to be associated with HIV progression, since CAP270 (i.e. the participant that progressed to AIDS faster than all other participants analysed), only tested positive for two STIs throughout the entire sampling period.

The implications of HIV coinfection has previously been studied by others and is reported to alter the course of both HIV and STI disease progression (Funnye & Akhtar, 2003), while complicating therapeutic efforts (Griemberg *et al*., 2006).

Furthermore, some researchers have investigated if the presence of STIs actually facilitates the transmission or acquisition of HIV, which has been reported to be possible in the case of *Syphilis* coinfection as explained by Funnye and Akhtar (2003). Other STIs, capable of causing genital ulcers have also been associated with an increased risk of HIV transmission, including HSV-2 (Funnye & Akhtar, 2003; Galvin & Cohen, 2004) which is estimated to increase the chance of HIV acquisition by approximately 3-fold in men and women, and up to 6-fold in high-risk individuals (Sheth *et al*., 2008). Shockingly, in sub-Saharan Africa HSV-2 coinfection is estimated to be present in approximately half of all HIV infections (Wald, 2004). This not only complicates HIV disease progression, but also reduces HIV-specific immune responses (Sheth *et al*., 2008) thereby limiting natural host responses against HIV. Among the participants studied here however, the number of STIs appeared to be associated with the number of transmitted variants, as participants CAP177 and CAP261, both of whom were inferred to be infected by more than one viral variant, consistently tested positive for more STIs than participants CAP217 and CAP270 throughout the sampling period.

This could potentially be explained by the hypothesis described by Haaland *et al*. (2009), who suggests that the presence of some STIs, that occur before HIV is acquired, creates inflammatory genital conditions which substantially increases the risk of becoming infected by multiple HIV variants within a single transmission event. It may be possible, based on this hypothesis that CAP177 and CAP261 could have been infected with STIs before HIV, although this could equally have been possible with participants CAP217 and CAP270 too.

4.3 Phylogenetic evaluation of viral compartmentalization

Tree-based methods for compartmentalization detection are known to be sensitive to topological uncertainty and the presence of recombination, and therefore place much weight on phylogenetic segregation achieved via poorly supported short interior branches (Zárate *et al*., 2007; Rousseau *et al*., 2007; Posada & Crandall, 2002).

There was little evidence for the latter effect here as all participants demonstrated well supported (posterior support ≥70% or ≥90) phylogenetic structure on longitudinal Bayesian MCC trees, where monophyletic compartmentalization patterns among all participants was largely driven by diverse sequences from the final sampling time points, during chronic infection.

Visual examination of longitudinal trees showed the presence of blood plasma and CVL samples intermingled between clades in cross-sectional and longitudinal trees among all participants, suggesting that viruses were readily mixing between anatomical compartments, consistent with recent findings by Bull *et al*. (2013). Similar branching topologies and clustering patterns were also evident on Bayesian MCC trees, between single and multi-variant HIV infected participants. In all participants, clades that were monophyletic by tissue type comprised mostly of sequences from the same time point and can be explained by either rapid turnover of population genetic diversity through time, transient bursts of viral replication (Bull *et al*., 2013; Marchant *et al*., 2006; Ostrowski *et al*., 1998) or a short HIV life span, since the half-life of infected cells is estimated to be around 1.6 days (Heeregrave *et al*., 2009).

Although cross-sectional studies are inherently limited, particularly in the estimation of temporal dynamics (Sturdevant *et al*., 2012), when data sets were analysed individually per time point a clear indication of multiple lineages was visible at 28 days in CAP177 and 63 days in CAP261, supporting the conclusion that these participants were infected by more than one viral variant. Furthermore, phylogenetic structure among these participants was consistent with those observed in other multi-variant infections (Kearney *et al*., 2009). These findings are in contrast to a study by Moore *et al*. (2012), in which data sets "with evidence of dual infection" were claimed to be excluded, however CAP177 was included among the study participants analysed, despite the vast evidence of multiple variants in this participant, presented here. Alternatively, in participants CAP217 and CAP270 outgrowth of a single viral variant was evident on both cross-sectional and longitudinal phylogenetic trees.

Nevertheless, tissue-specific viral compartmentalization was not evident during long-term infection, regardless of population homogeneity or the number of transmitted variants in any of the participants studied.

4.4 Statistical evaluation of compartmentalization and population migration

Despite at least three quarters of statistical tree-based tests for compartmentalization suggesting significant evidence for the existence of distinct viral populations in the cervix and blood plasma, in three of the four participants (CAP177, CAP217, CAP270), tissue-specific lineages did not evolve and persist over time in any of these participants, consistent with previous studies (Bull *et al*., 2013; Kemal *et al*., 2003). Transient bursts of viral replication, visible on Bayesian MCC trees, are known to bias the results of these tests through the production of numerous monotypic and low diversity sequences (Bull *et al*., 2009). While the exclusion of identical sequences from the longitudinal datasets removed almost all statistical evidence of populations being structured by tissue type in CAP270, tissue specific population structure was retained in CAP177 and CAP217 in 2/3 and 1/3 of tree-based tests respectively.

Results from cross-sectional analyses revealed three different compartmentalization patterns among the four participants. An example of the first pattern was demonstrated by CAP177 whereby statistically significant evidence for compartmentalization was detected at all sampling time points including sequences collected in the acute and chronic stages of infection. An example of the second pattern was seen in CAP270, where statistical evidence for compartmentalization was only transiently observed at particular, but not successive time points during the chronic stage of infection, while the third pattern, evident in CAP261 and CAP217 exhibited statistical evidence for compartmentalization during the later stages of chronic infection. Although the removal of monotypic sequences led to a reduction in the number of significant results in CAP177 using tree-based tests at 28, 328 and 560 days post-infection, all evidence supporting the existence of compartmentalization was absent at the final sampling point (1295 days) in this participant. There was no effect on results from cross-sectional analyses following the removal of monotypic sequences in CAP217 and CAP270 where, in the later time points, significant evidence of compartmentalization remained detectable at 1316 and 406 days, respectively.

When both monotypic sequences and low diversity tissue-specific sequences were removed from CAP177 and CAP217 datasets, evidence for significant compartmentalization was only detectable at 28 days in CAP177.

In distance-based compartmentalization testing, both F_{ST} and Snn statistics estimated fewer statistically significant results to support compartmentalized structure between blood plasma and cervical populations, than tree-based tests. After analysis of gp120 regions including and excluding monotypic and low diversity sequences, the Snn test predicted strong significance for the presence of compartmentalization in CAP177, whereas no such structure was found in all other participants. As expected however, after the removal of monotypic and low diversity sequences, almost no evidence of compartmentalization remained, except at 378 days in CAP177. The detection of significant structure with F_{ST} and Snn tests however, may be due in part to the small viral population sizes analysed per participant (i.e. limited effective population size) (Gantt *et al*., 2010) or the lesser power of distance-based tests (Zárate *et al*., 2007).

The flow of viral populations between anatomical compartments was then analysed to determine if viruses were trafficking between the blood plasma and cervix, and to assess the rate at which this movement was occurring using a structured coalescent model. Analyses revealed substantial mixing of viruses between the blood plasma and cervical compartments in all participants during chronic and acute infection stages, indicating "an equilibrium of viral quasiespecies between the two compartments", (Imamichi *et al*., 2011). An almost equal rate of viral movement was estimated between the blood plasma and cervical compartments in CAP261, while in CAP270 and CAP177 four times the amount of viruses were predicted to migrate from the cervix to the blood plasma compared to movement in the opposite direction. Unlike all other participants however, there was a higher rate of viral movement from the blood plasma to the cervix in CAP217. After the exclusion of monotypic sequences, bi-directional movement was again observed in all participants with an almost equal rate of migration in CAP177, CAP261 and CAP270, whereas in CAP217 more viruses were estimated to migrate from the cervix to the blood plasma. When sequences were analysed per time point, there was no consensus on migration direction and migratory patterns fluctuated between acute and chronic infection.

In summary, exchange of viruses between the blood plasma and cervical tissues was evident in all participants both longitudinally and cross-sectionally, suggesting a continual mixing of populations throughout the infection period.

4.5 Intrapatient viral diversity

The extensive genetic variability of HIV is the product of various mechanisms including host factor immune pressures, mutations that occur through the process of reverse transcription, rapid turnover rates, recombination between replicating viruses and drug resistant mutations, among other pressures (Luft *et al*., 2011; Santoro & Perno, 2013). Not only do these mechanisms lead to extensive variability within the HIV genome, but they also contribute to the creation of millions of variants within a single individual, directly affecting response to therapy, disease progression and viral transmission (Santoro & Perno, 2013).

In intrapatient diversity analysis, genetic distances increased over the course of infection in all participants, consistent with previously published data on variability within the *env* gp120 region (Rossenkhan *et al*., 2012). Viral diversification was more pronounced in participants CAP177 and CAP261 when compared to CAP217 and CAP270, regardless of tissue origin, suggesting an association between diversity and the number of transmitted viral variants (Sagar *et al*., 2003). Although patterns of increased viral diversity were similar in blood plasma and CVL viruses, they appeared to diverge more distinctly within the acute stage of infection in all participants, in a linear manner, similar to observations by Shankarappa *et al*. (1999). This linear pattern in diversity was more pronounced in CAP217 and CAP270, possibly due to the presence of a single founder variant in these participants, although the number of sampled time points where sequences from both tissues were available per participant likely played a role in this observation.

In a study that looked at HIV diversity over a much longer time frame than that studied here (between 6 and 9 years), three distinct phases of diversity were reported, the first of which is consistent with results reported here, i.e. phase one – corresponding to a linear increase in diversity at a rate of about 1% per year (Shankarappa *et al*., 1999).

In this study, the greatest increases in diversity were found in CAP270, CAP177 and CAP261 all of whom harboured populations that evolved at rates of 1.29%, 1.27% and 1.20% per year respectively, whereas populations in CAP217 diversified at approximately 0.73% per year. This can be associated with the rate of disease progression in these participants, since CAP270 (i.e. the participant with the highest diversification rate) was later classified as a rapid progressor, while CAP177 and CAP261 were inferred to have been infected by multiple viral variants (reported to progress to AIDS faster than those infected by single founder viruses) and finally CAP217, who was inferred to be infected by a single viral variant (Novitsky *et al*., 2011; Van der Kuyl & Cornelissen, 2007), possessed the lowest rate of evolution per year, as expected.

4.6 Accumulation of potential N-linked glycosylation sites

Examination of glycosylation accumulation patterns revealed substantial changes in the frequency of PNLGs within *env* subregions over the sampling period. PNGLs had a tendency to emerge midway or at later time points as the infection progressed however these changes were not specifically associated with viruses from the cervix or blood plasma. PNLG site changes in the V1V2, V4 and V5-loops among viruses from CAP270, CAP261 and CAP217 could potentially have been driven in response to neutralizing antibodies, however *env* antibody binding sites in these individuals presently remain unmapped. In CAP177, it is known that the emergence of a PNLG site at amino acid site 332 and associated loss of a PNLGs at amino acid site 334 one year post-infection allowed viruses to escape the neutralizing effect of autologous monoclonal antibodies targeting an epitope at site 332 (Gray *et al*., 2011; Moore *et al*., 2012).

No tissue-specific differences in PNLGs accumulation was evident in any participant and viruses appeared to accumulate PNLG sites discretely with time, although when PNLGs data was analysed per time point, significant differences in PNLGs numbers between blood plasma and CVL viruses was found in 5 out of the 51 paired samples tested (i.e. per V-loop region and sample time point). Nevertheless, these differences did not correspond to acute or chronic stages of infection and were not consistent between single and multiple variant infections.

In the analysis of PNLG site accumulation and its association with viral evolution, PNLGs that appeared in an increasing number of viruses over time were mapped to longitudinal Bayesian MCC trees in search of PNLG sites that occurred locally to a small subset of closely related branches/viruses. In doing so, several sites that appeared to accumulate in an increasing number of closely related viruses from both blood plasma and CVL sequences were found, i.e. N130, N186, N141, N142 and N190. Five of these six PNLGs appeared predominantly during the mid to later stages of infection pointing towards the evolution of fitter viruses, while three of the five PNLGs were found to be within the V1V2-loop region, an area of the *env* gene that has previously been associated with host immune escape (Van Gils *et al*., 2011).

4.7 Intrapatient recombination

In intrapatient recombination analysis, evidence of *env* gp120 quasispecies in three of the four participants (CAP177, CAP261 and CAP270) was found. In CAP177 and CAP261 continuous recombination was evident throughout the entire infection period, however no evidence of recombination was found in CAP217 with only a single recombination event detected in CAP270 at the latest time point sampled (903 days). It was possible that recombinants were not detected in CAP270 due to "low levels of replication, or substitution with non-recombinant viruses due to lower fitness" as suggested by Kiwelu *et al*. (2013).

The average incidence of recombination was higher in multi-variant infected participants (present in 57.64% of all sequences) than single-variant (present in 3.89% of all sequences) infections within the same *env* region. These findings are consistent with other studies in participants infected with HIV-1 subtype C viruses (Kiwelu *et al*., 2013; Novitsky *et al*., 2011). The relatively high rate of recombination in multi-variant HIV infections has previously been attributed to the virus's ability to combine the functionality of the reverse transcriptase enzyme between RNA templates in each variant during the transcription stage of replication, making it easier to accumulate recombinants this way (Kiwelu *et al*., 2012; Coffin, 1979; Kiwelu *et al*., 2013). Nevertheless, intrapatient recombination does not require coinfection to produce recombinants, because as the infection progresses "intrahost diversification produces a pool of quasispecies that can be used as distinct templates" for recombination (Kiwelu *et al*., 2013).

When recombination breakpoints were analysed, the distribution of breakpoints across the gp120 region was consistent with results reported by Kiwelu *et al*. (2013) and Lamers *et al*. (2009). Although there were similar recombination patterns in participants CAP177 and CAP261, all identified recombinants were unique and none of the participants shared estimated recombination breakpoints, suggesting ongoing recombination within populations in the blood plasma and cervix. Analysis of the number of HIV-1 recombinants over chronic and acute infection indicated that the proportion of detectable recombinants decreased from an average presence of 92% to 25% in combined blood plasma and CVL sequences among all participants where recombination was detected, while the number of unique recombination events increased from approximately 0.67 events at the earliest time points sampled to 2.67 events at the last time points sampled in all participants.

When changes in V-loop lengths were compared to recombination patterns, differences between single and multiple variant infections were noted, particularly in sequence length increases within the V1V2-loops of multi-variant infected participants CAP177 and CAP261. Transmission of more than one HIV variant has previously been linked to faster disease progression, which may explain this difference (Novitsky *et al*., 2011). Minor increases in sequence lengths were also observed in the V4 and V5-loops in these participants, however the presence of five recombinant sequences in CAP270 did not affect any sequence length changes in any of the V-loop regions. Similarly when comparing the average number of PNLGs present in the same *env* subregions before and after the occurrence of a recombination event, there were considerable differences between single and multiple variant infections. Recombinant regions were spread across multiple subregions in blood plasma and CVL sequences isolated from participants CAP177 and CAP261, where increases in PNLGs numbers were found in almost all variable loops (excluding the V5-loop). Non-significant increases in PNLGs were also present in the conserved regions of all participants, specifically in the C1 and C5 regions of CAP270 sequences, i.e. the regions in which the single recombination event was detected in this participant.

4.8 Codon-based selection

Molecular and purifying selection are caused by a range of pressures acting on a viral variant including those induced by the host itself, and are both responsible for selectively applying evolutionary forces that preserve certain amino acid residues while allowing nucleotide sequence variation to persist in other parts of a sequence (Castel *et al*., 2014). In selection analyses, both blood plasma and CVL viruses were found to be subject to positive and negative selection pressures along the gp120 *env* region using at least three different methods. Although more evidence of negative than positive selection was discovered among all participants, significant evidence of positive selection was found for a total of 17 amino acid sites using FUBAR, MEME, FEL and SLAC methods. The FUBAR method provided results on sites that were under diversifying selection, i.e. sites that occurred locally to a small subset of viruses on phylogenetic trees. In tissue-specific selection analysis, to test the hypothesis that viruses in the cervix were under different selective pressures than those sampled from the blood plasma, sequences were separated and analysed individually by tissue type per participant.

Viruses from CAP261 and CAP270 showed a similar number of sites under positive/negative selection, while viruses from CAP177 and CAP217 showed very different numbers of sites under selection between tissue types. This may have been partly due to the sample sizes acquired from each participant or to the adaptation of viral populations to their host environment and the result of evolutionary pressures exerted on them by the host immune system as suggested by Castel *et al*. (2014). Although there is little information concerning the exact role of selective pressures on PNLGs accumulation, the number of amino acid sites identified as being under positive selection was compared to PNLG sites along the translated gp120 *env* gene in search of potential overlaps between the two. Interestingly, six sites identified with significant evidence of positive selection coincided with previously identified PNLG sites at positions N349 & N424 in CAP177, N474 in CAP217, N189 & N412 in CAP261 and N154 in CAP270.

More PNLG sites were found to be under positive selection in multi variant infected participants CAP177 and CAP261 than single variant infected participants CAP217 and CAP270, suggesting a possible selective advantage in viral populations within multi variant infections. In studies that investigated adaptive and selective mutations, it was found that slower disease progression was actually linked to a larger number of positively selected sites and faster *env* adaptation rates (Lemey *et al*., 2007), consistent with patterns found in CAP177 and CAP261, both of whom have previously been identified as long term non-progressors.

4.9 Co-receptor tropism

Although several studies have previously reported that the V3-loop (i.e. the primary determinant for co-receptor usage) is a highly variable region (Fernandez *et al*., 1995; Almond *et al*., 2010), V3-loops remained consistent in both blood plasma and CVL viruses from all participants, which may be due to the absence of antiretrovirals in this cohort, based on a 2011 study in which Waters *et al*. demonstrated the influence of ARVs on co-receptor usage, by suppressing planned treatment and retesting co-receptor tropism, wherein 2 of the 37 participants showed evidence of co-receptor switching.

Detection of CCR5 and CXCR4-tropic viruses is usually expected where HIV recombination has been observed according to Mild *et al*. (2007) however in this data set no CXCR4-topic viruses were evident despite the presence of recombination in three of the four participants. Many authors have also hypothesized that a shift in co-receptor tropism from CCR5 to CXCR4 usage tends to occur within the later stages of infection (Mild *et al*., 2007; Lusso, 2006; Cecilia *et al*., 2000), however this was not the case in any of the participants analysed here either.

Net charge analyses revealed an overall higher possibility of co-receptor switching from CCR5 to CXCR4-tropic viruses in the blood plasma further into disease progression, based on the higher net charges observed in viruses from this tissue type at 6 of the 8 time points where differences were noted. Tissue-specific differences in net charge could not be attributed to V-loop length variation or PNLGs distribution however, as these traits remained constant in the V3-loop in both blood plasma and CVL viruses throughout the course of infection in all participants.

The change in co-receptor usage is also dependent on the presence of recombination and the number of transmitted viral variants, hence viral populations in participants CAP177 and CAP261 are more likely to exhibit shifts in co-receptor usage as the infection progresses, since these participants have both been inferred to be infected by multiple viral variants and have demonstrated extensive evidence of recombination in both blood plasma and CVL viruses.

4.10 Hypermutation signatures

To determine if viruses from each tissue type were differentiated by characteristics associated with hypermutation, gp120 blood plasma and CVL sequences were analysed for G-to-A and A-to-G mutations, resulting in no statistically significant evidence of hypermutation in any of the isolated sequences ($p > 0.05$). Despite this finding more G-to-A mutations were observed in both blood plasma and CVL viruses with the exception of viruses isolated from the CVL in participant CAP177. Conflicting hypermutation patterns were evident in viruses from the blood plasma and CVL in CAP177, suggesting the possibility of differential selection pressures in each anatomical compartment within this participant. However, despite the larger number of G-to-A mutations among all participants (excl. CAP177 CVL viruses), no significant evidence of a host immune response in the form of APOBEC3-induced hypermutation was found. A study by Armitage *et al*. (2012) describes this as a "discrete all or nothing phenomenon" stating that not all G-to-A mutations result in a lethal effect on the virus and in simulated studies "sub-lethal" activity has been shown to occur, resulting in increased genetic diversities as opposed to virus inactivation.

4.11 Length variation within the *env* V-loop regions

After systematic examination of length variation within the V-loop regions, V1V2, V4 and V5-loops displayed length expansions in addition to an increase in PNLGs accumulation during disease progression among all participants, most predominantly in the V1V2-loops, a region that is widely known to be important in inducing neutralizing antibody response (Gorny *et al*., 1994; Scanlan *et al*., 2002; Thali *et al*., 1993; Xiang *et al*., 2002; Mild *et al*., 2007). Minimal length variation was found in the V5-loop while no changes occurred in the V3-loops of both blood plasma and CVL viruses from all participants, consistent with findings by Novitsky *et al*. (2009) and Coetzer *et al*. (2007).

V1V2, V4 and V5 loop lengths increased gradually over the course of infection in all participants, a trend that has previously been associated with the accumulation of PNLGs and induction of neutralizing antibodies, where transmitted viruses with shorter loops capable of containing fewer PNLGs slowly increase in length to confer viral fitness, although this is yet to be confirmed experimentally (Novitsky *et al*., 2009). When tissue-specific differences were analysed per *env* subregion, significant differences in loop lengths and the number of PNLGs between blood plasma and CVL viruses were found at isolated time points in participants CAP177, CAP217 and CAP261, once again mainly in the V1V2-loops during mid to late infection stages, similar to findings by Curlin *et al*. (2010). Despite this finding however, no significant tissue-specific differences between PNLGs numbers and loop lengths were evident in the majority (86%) of the paired samples tested.

4.12 Study limitations

One of the key limitations affecting this study was during the amplification of CVL-derived viruses, where there was a presence of microscopic blood contamination in CVL samples. Almost three quarters of acute and chronic infection samples contained microscopic blood and were subsequently excluded from further analysis, leading to a smaller sample size from this tissue compared to the number of samples obtained from the blood plasma. Other drawbacks included the total number of sequences that were eventually amplifiable in addition to the long follow up periods between sampling points in some of the participants studied here, which played an important role in the outcome of statistical test results and other downstream analyses. More specifically in participants CAP261 and CAP270, there was a lack of samples from the first 50 days post infection, which was a major limitation in the analysis and comparison of viral sequences between the chronic and acute infection stages. Comparisons between genomic and clinical data also potentially contained a few drawbacks. Firstly, all clinical testing was performed outside of this study and the same individual may not have conducted laboratory testing of all samples obtained from each participant over the 3.6-year sampling period. Consistency of clinical measurements may therefore be difficult to assume, however international guidelines were followed in the processing of all samples and therefore any slight inconsistencies that may exist are regarded as negligible.

Conclusions

This study is one of a few that reports on the evolution of HIV-1 in treatment naïve female participants over acute and chronic infection stages in which both single and multiple subtype C variant infections were surveyed within a subtype C dominated region. Given all of the findings presented here, the presence of recombinant strains, viral diversity, PNLG sites, selection signatures and other confounding factors, pose major challenges to treatment strategies and the design of a suitable HIV vaccine. Nevertheless, this study has demonstrated that over long-term HIV infection viral populations are not compartmentalized by tissue type, despite the presence of statistical support at isolated time points. From the four participants analysed here, CAP270 and CAP217 showed evidence of infection with a single viral variant whereas, CAP177 and CAP261 indicated infection by more than one variant. As a result, genetic diversity, selection signatures and the number of detectable recombination events in viral populations were significantly lower in the former participants compared to the latter.

Genetic diversity increased over the course of infection in all four participants, while CD4 cell counts declined and viral loads grew steadily. Although there was a jump in viral loads in CAP270, there was no other evidence to suggest that this participant was superinfected since the tMRCA was quite accurately reconstructed to the known time post infection and the Poisson distribution suggested infection by a single variant. Despite these results however, samples from the first 50 days were not available for CAP270, therefore we cannot definitively say whether this or any other participant studied here was superinfected.

No consistent evidence for the existence of separate populations in the cervix and blood plasma was found in any of the participants using phylogenetic and statistical measures. Significant test results were attributed to the presence of monotypic or low diversity sequences, and test sensitivity, wherein tree-based tests have been reported to be more reliable than distance-based tests (Zárate *et al*., 2007). Furthermore, sequences generally clustered together by time point on Bayesian maximum clade credibility (MCC) trees in longitudinal phylogenetic analyses. Clades that were monophyletic by tissue type comprised mostly of low diversity or monotypic sequences from the same time point, consistent with rapidly replicating viral populations.

Viral sequences from the cervix and blood plasma did not exhibit significant tissue-specific differences in recombination, co-receptor usage, V-loop length variation, codon-based selection or PNLGs accumulation patterns, although notable differences in these traits were observed between multiple and single-variant infected participants. Furthermore, although this project would have benefited significantly from an expansion of the study population, it is important to note that only three published longitudinal studies containing cervical and plasma derived sequences exist to date, with the largest data set consisting of 14 women. Ultimately however, while this study provides a new contribution on the subject of HIV compartmentalization in female seroconvertors, the presence of compartmentalized viral populations in a minority of women cannot be ruled out, since only four women were analysed.

Bibliography

Abdel-Mohsen, M., Raposo, R. A. S., Deng, X., Li, M., Liegler, T., Sinclair, E., et al. (2013). Expression profile of host restriction factors in HIV-1 elite controllers. *Retrovirology, 10*(106), 1–13. doi:10.1186/1742-4690-10-106.

Abecasis, A. B., Wensing, A. M. J., Paraskevis, D., Vercauteren, J., Theys, K., Van de Vijver, D. A. M. C., et al. (2013). HIV-1 subtype distribution and its demographic determinants in newly diagnosed patients in Europe suggest highly compartmentalized epidemics. *Retrovirology*, *10*(7), 1–13. doi:10.1186/1742-4690-10-7

Abecasis, A., Vandamme, A-M., and Lemey, P. (2007). Sequence alignment in HIV computational analysis. In T. Leitner, B. Foley, B. Hahn, P. Marx, F. McCutchan, J. Mellors, S. Wolinsky, and B. Korber (ed.), *HIV sequence compendium 2006/2007*. Theoretical Biology and Biophysics Group, Los Alamos, NM. p. 2-16.

Abrahams, M-R., Anderson, J. A, Giorgi, E. E., Seoighe, C., Mlisana, K., Ping, L-H., et al. (2009). Quantitating the multiplicity of infection with human immunodeficiency virus type 1 subtype C reveals a non-poisson distribution of transmitted variants. *Journal of Virology, 83*(8), 3556–67. doi:10.1128/JVI.02132-08.

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. *Journal of Virology, 84*(19), 9864–78. doi:10.1128/JVI.00915-10

Adal, M., Ayele, W., Wolday, D., Dagne, K., Messele, T., Tilahun, T., Berkhout, B., Mayaan, S., Pollakis, G., Dorigo-Zetsma, W. (2005). Evidence of genetic variability of human immunodeficiency virus type 1 in plasma and cervicovaginal lavage in ethiopian women seeking care for sexually transmitted infections. *AIDS Research and Human Retroviruses*. *21*(7), 649-53.

Afonso, J. M., Bello, G., Guimarães, 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*(8), e42996. doi:10.1371/journal.pone.0042996

Agnihotri, K. (2008). Review Of Literature. Phd Thesis.

Ajoge, H. O., Gordon, M. L., de Oliveira, T., Green, T. N., Ibrahim, S., Shittu, O. S., Olonitola, S. O., et al. (2011). Genetic characteristics, coreceptor usage potential and evolution of Nigerian HIV-1 subtype G and CRF02_AG isolates. *PloS one, 6*(3), e17865. doi:10.1371/journal.pone.0017865.

Alexaki, A., Liu, Y., & Wigdahl, B. (2008). Cellular Reservoirs of HIV-1 and their Role in Viral Persistence. *Current HIV Research*, *6*(5), 388–400.

Almond, D., Kimura, T., Kong, X., Swetnam, J., Zolla-pazner, S., & Cardozo, T. (2010). Structural conservation predominates over sequence variability in the crown of HIV type 1's V3 loop. *AIDS Research and Human Retroviruses, 26*(6), 717–23. doi:10.1089/aid.2009.0254

Alter, G., & Moody, M. A. (2010). The humoral response to HIV-1: new insights, renewed focus. *The Journal of Infectious Diseases, 202*(Suppl 2), S315–S322. doi:10.1086/655654.

Amo, J. del, Pérez-Cachafeiro, S., Hernando, V., González, C., Jarrin, I., & Bolúmar, F. (2010). Migrant health: Epidemiology of HIV and AIDS in migrant communities and ethnic minorities in EU/EEA countries (pp. 1–122).

Amoêdo, N. D., Afonso, A. O., Cunha, S. M., Oliveira, R. H., Machado, E. S., & Soares, M. A. (2011). Expression of APOBEC3G/3F and G-to-A hypermutation levels in HIV-1-infected children with different profiles of disease progression. *PLoS one, 6*(8), 1–10. doi:10.1371/journal.pone.0024118.

Anderson, B. L., & Cu-Uvin, S. (2008). Determinants of HIV shedding in the lower genital tract of women. *Current Infectious Disease Reports*. *10*(6), 505-511.

Anderson, J.A., Ping, L-H., Dibben, O., Jabara, C.B., Arney, L., Kincer, L., et al., (2010). HIV-1 Populations in Semen Arise through Multiple Mechanisms. *PLoS Pathogens, 6*(8). 1-12. doi:10.1371/journal.ppat.1001053.

Andeweg, A. C., Boers, P. H., Osterhaus, A. D., & Bosch, M. L. (1995). Impact of natural sequence variation in the V2 region of the envelope protein of human immunodeficiency virus type 1 on syncytium induction: a mutational analysis. *Journal of General Virology, 76*, 1901–7. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7636471

Andreolétti, L., Skrabal, K., Perrin, V., Chomont, N., Saragosti, S., Gresenguet, G., et al. (2007). Genetic and Phenotypic Features of Blood and Genital Viral Populations of

Clinically Asymptomatic and Antiretroviral-Treatment-Naive Clade A Human Immunodeficiency Virus Type 1-Infected Women. *Journal of Clinical Microbiology, 45*(6), 1838-1842. doi: 10.1128/JCM.00113-07.

Andrews, C. A. (2010) Natural Selection, Genetic Drift, and Gene Flow Do Not Act in Isolation in Natural Populations. *Nature Education Knowledge, 3*(10):5.

Archibald, D. W., Hebert, C. A., Gregory, K. L., & Lewis, G. K. (1993). Effects of Human Salivas Recombinant HIV-1 Proteins. *Critical Reviews in Oral Biology & Medicine, 4*(3/4), 475–478. doi:10.1177/10454411930040033101.

Armitage, A. E., Deforche, K., Chang, C., Wee, E., Kramer, B., John, J., et al. (2012). APOBEC3G-Induced Hypermutation of Human Immunodeficiency Virus Type-1 Is Typically a Discrete "'All or Nothing'" Phenomenon. *PLoS Genetics, 8*(3), 22–24. doi:10.1371/journal.pgen.1002550.

Arora, A., & Seth, P. (2003). Antigenicity and immunogenicity of HIV envelope gene expressed in baculovirus expression system. *Gene Therapy and Molecular Biology, 7*, 37–42.

Aulicino, P. C., Bello, G., Guimaraes, M. L., Ruchansky, D., Rocco, C., Mangano, A., et al. (2011). Longitudinal analysis of HIV-1 BF1 recombinant strains in vertically infected children from Argentina reveals a decrease in CRF12_BF pol gene mosaic patterns and high diversity of BF unique recombinant forms. *Infection, Genetics and Evolution*, *11*(2), 349–57. doi:10.1016/j.meegid.2010.11.008

Avery, L. B., Vanausdall, J. L., Hendrix, C. W., & Bumpus, N. N. (2013). Compartmentalization and Antiviral Effect of Efavirenz Metabolites in Blood Plasma, Seminal Plasma, and Cerebrospinal Fluid. *Drug Metabolism and Disposition*, *41*(February), 422–429.

Balzarini, J., Laethem, K. Van, Hatse, S., Vermeire, K., Clercq, E. De, Peumans, W., et al. (2004). Profile of Resistance of Human Immunodeficiency Virus to Mannose-Specific Plant Lectins. *Journal of Virology, 78*(19), 10617–10627. doi:10.1128/JVI.78.19.10617

Bandyopadhyay, S., Kelley, R., & Ideker, T. (2006). Discovering regulated networks during HIV-1 latency and reactivation. *Pacific Symposium on Biocomputing*. 354-66.

Baum, D. (2008). Reading a Phylogenetic Tree: The Meaning of Monophyletic Groups. *Nature Education, 1*(1), 190.

Bazykin, G. A, Dushoff, J., Levin, S. A, & Kondrashov, A. S. (2006). Bursts of nonsynonymous substitutions in HIV-1 evolution reveal instances of positive selection at conservative protein sites. *Proceedings of the National Academy of Sciences, 103*(51), 19396–401. doi:10.1073/pnas.0609484103

Beerli, P., & Felsenstein, J. (1999). Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics, 152*(2), 763–73.

Beerli, P., & Palczewski, M. (2010). Unified framework to evaluate panmixia and migration direction among multiple sampling locations. *Genetics, 185*(1), 313–26. doi:10.1534/genetics.109.112532.

Ben-Zeev, O., Stahnke, G., Liu, G., Davis, R. C., & Doolittle, M. H. (1994). Lipoprotein lipase and hepatic lipase: the role of asparagine-linked glycosylation in the expression of a functional enzyme. *Journal of Lipid Research, 35*(9), 1511–23. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7806965

Blackard, J. T., Ma, G., Martin, C. M., Rouster, S. D., Shata, M. T., & Sherman, K. E. (2011). HIV Variability in the Liver and Evidence of Possible Compartmentalization. *AIDS Research and Human Retroviruses, 27*(10), 1117–1126. doi:10.1089/aid.2010.0329

Blay, W. M., Gnanakaran, S., Foley, B., Doria-Rose, B. T., Korber, B. T., & Haigwood, N. L. (2006). Consistent Patterns of Change during the Divergence of Human Immunodeficiency Virus Type 1 Envelope from That of the Inoculated Virus in Simian/Human Immunodeficiency Virus-Infected Macaques. *Journal of Virology, 80*(2), 999–1014. doi:10.1128/JVI.80.2.999

Boeras, D. I., Hraber, P. T., Hurlston, M., Evans-Strickfaden, T., Bhattacharya, T., Giorgi, E. E., et al. (2011). Role of donor genital tract HIV-1 diversity in the transmission bottleneck. *Proceedings of the National Academy of Sciences*, *108*(46), E1156–63. doi:10.1073/pnas.1103764108

Bradic, M., Beerli, P., García-de León, F. J., Esquivel-Bobadilla, S., & Borowsky, R. L. (2012). Gene flow and population structure in the Mexican blind cavefish complex (Astyanax mexicanus). *BMC Evolutionary Biology*, *12*(9), 1471–2148. doi:10.1186/1471-2148-12-9

Brockman, M. A., Kwon, D. S., Tighe, D. P., Pavlik, D. F., Rosato, P. C., Sela, J., et al. (2009). IL-10 is up-regulated in multiple cell types during viremic HIV infection and reversibly inhibits virus-specific T cells. *Blood, 114*, 346–356. doi:10.1182/blood-2008-12-191296.

Broliden, K. (2010). Innate Molecular and Anatomic Mucosal Barriers against HIV Infection in the Genital Tract of HIV-Exposed Seronegative Individuals. *The Journal of Infectious Diseases*, *202*(Suppl 3), S351–5. doi:10.1086/655964

Brooks, D. G., Kitchen, S. G., Kitchen, C. M. R., Scripture-Adams, D. D., & Zack, J. A. (2001). Generation of HIV latency during thymopoiesis. *Nature Medicine, 7*(4), 459–464. Retrieved from http://dx.doi.org/10.1038/86531

Brown, R.J.P., Peters, P.J., Caron, C., Gonzalez-Perez, M.P., Stones, L., et al. (2011). Intercompartmental Recombination of HIV-1 Contributes to env Intrahost Diversity and Modulates Viral Tropism and Sensitivity to Entry Inhibitors. *Journal of Virology, 85*(12). 6024-6037, doi: 10.1128/JVI.00131-11.

Brunetta, E., Fogli, M., Varchetta, S., Bozzo, L., Hudspeth, K. L., Marcenaro, E., et al. (2010). Chronic HIV-1 viremia reverses NKG2A/NKG2C ratio on natural killer cells in patients with human cytomegalovirus co-infection. *AIDS, 24*, 27–34. doi:10.1097/QAD.0b013e3283328d1f.

Bull, M., Learn, G., Genowati, I., McKernan, J., Hitti, J., Lockhart, D., Tapia, K., et al. (2009). Compartmentalization of HIV-1 within the Female Genital Tract Is Due to Monotypic and Low-Diversity Variants Not Distinct Viral Populations. *PLoS ONE, 4*(9).

Bull, M. E., Heath, L. M., McKernan-Mullin, J. L., Kraft, K. M., Acevedo, L., Hitti, J. E., et al. (2013). Human immunodeficiency viruses appear compartmentalized to the female genital tract in cross-sectional analyses but genital lineages do not persist over time. *The Journal of Infectious Diseases*, *207*(8), 1206–15. doi:10.1093/infdis/jit016

Bunnik, E. M. (2010). HIV-1 neutralizing humoral immunity, viral evolution and disease progression.

Buonaguro, L., Tornesello, M. L., & Buonaguro, F. M. (2007). Human Immunodeficiency Virus Type 1 Subtype Distribution in the Worldwide Epidemic: Pathogenetic and

Therapeutic Implications. *Journal of Virology, 81*(19), 10209–19. doi:10.1128/JVI.00872-07.

Burger, H., & Hoover, D. (2008). HIV-1 tropism, disease progression, and clinical management. *The Journal of Infectious Diseases*, *198*(8), 1095–7. doi:10.1086/591624

Burke, D. S. (1997). Recombination in HIV: An Important Viral Evolutionary Strategy. *Emerging Infectious Diseases, 3*(3), 253–9. doi:10.3201/eid0303.970301

Carr, J. M., Hocking, H., Li, P., & Burrell, C. J. (1999). Rapid and efficient cell-to-cell transmission of human immunodeficiency virus infection from monocyte-derived macrophages to peripheral blood lymphocytes. *Virology*, *265*(2), 319–29. doi:10.1006/viro.1999.0047

Carr, J. K., Wolfe, N. D., Torimiro, J. N., Tamoufe, U., Eyzaguirre, L., Birx, D. L., et al. (2010). HIV-1 recombinants with multiple parental strains in low-prevalence, remote regions of Cameroon: evolutionary relics? *Retrovirology, 7*(39), 1–8. doi:10.1186/1742-4690-7-39

Carrington, M., Nelson, G. W., Martin, M. P., Kissner, T., Vlahov, D., Goedert, J. J., et al. (1999). HLA and HIV-1: Heterozygote Advantage and B*35-Cw*04 Disadvantage. *Science*, *283*(March), 1748–1752.

Carvajal-Rodríguez, A., Posada, D., Pérez-Losada, M., Keller, E., Abrams, E. J., Viscidi, R. P., & Crandall, K. A. (2008). Disease progression and evolution of the HIV-1 env gene in 24 infected infants. *Infection, Genetics and Evolution*, *8*(2), 110–20. doi:10.1016/j.meegid.2007.10.009

Casado, C., Pernas, M., Alvaro, T., Sandonis, V., García, S., Rodríguez, C., et al. (2007). Coinfection and Superinfection in Patients with Long-Term, Nonprogressive HIV-1 Disease. *The Journal of Infectious Diseases*, *196*(15 September), 895–9. doi:10.1086/520885

Casartelli, N., Guivel-Benhassine, F., Bouziat, R., Brandler, S., Schwartz, O., & Moris, A. (2009). The antiviral factor APOBEC3G improves CTL recognition of cultured HIV-infected T cells. *The Journal of Experimental Medicine*, *207*(1), 39–49. doi:10.1084/jem.20091933.

Castel, G., Razzauti, M., Jousselin, E., Kergoat, G. J., & Cosson, J-F. (2014). Changes in diversification patterns and signatures of selection during the evolution of murinae-associated hantaviruses. *Viruses, 6*(3), 1112–34. doi:10.3390/v6031112

Castresana, J. (2002). Estimation of genetic distances from human and mouse introns. *Genome Biology, 3*(6), p.1-7.

CDC. (2008). HIV/AIDS among Women (pp. 1–7).

Cecilia, D., Kulkarni, S. S., Tripathy, S. P., Gangakhedkar, R. R., Paranjape, R. S., & Gadkari, D. A. (2000). Absence of coreceptor switch with disease progression in human immunodeficiency virus infections in India. *Virology*, *271*, 253–8. doi:10.1006/viro.2000.0297

Chaillon, A., Gianella, S., Wertheim, J. O., Richman, D. D., Mehta, S. R., & Smith, D. M. (2014). HIV Migration Between Blood and Cerebrospinal Fluid or Semen Over Time. *The Journal of Infectious Diseases*, *209*(15 May), 1642–52. doi:10.1093/infdis/jit678

Chan, E., Towers, G. J., & Qasim, W. (2014). Gene therapy strategies to exploit TRIM derived restriction factors against HIV-1. *Viruses, 6*(1), 243–63. doi:10.3390/v6010243.

Chaudhary, S., Noel, R., Rodriguez, N., Collado, S., Munoz, J., Kumar, A., & Yamamura, Y. (2012). Correlation between CD4 T cell Counts and Virus Compartmentalization in Genital and Systemic Compartments of HIV-infected Females. *Virology*, *417*(2), 320–326. doi:10.1016/j.virol.2011.06.018.Correlation

Chen, P., Hübner, W., Spinelli, M. A., & Chen, B. K. (2007). Predominant mode of human immunodeficiency virus transfer between T cells is mediated by sustained Env-dependent neutralization-resistant virological synapses. *Journal of Virology*, *81*(22), 12582–95. doi:10.1128/JVI.00381-07

Cherry, J. L., Lipman, D. J., Nikolskaya, A., & Wolf, Y. I. (2009). Evolutionary Dynamics of N-Glycosylation Sites of Influenze Virus Hemagglutinin. *PloS Currents, 18*(1). doi: 10.1371/currents.RRN1001.

Chiu, Y-L., Soros, V. B., Kreisberg, J. F., Stopak, K., Yonemoto, W., & Greene, W. C. (2005). Cellular APOBEC3G restricts HIV-1 infection in resting CD4+ T cells. *Nature, 435*(7038), 108–14. doi:10.1038/nature03493.

Chomont, N., Hocini, H., Grésenguet, G., Brochier, C., Bouhlal, H., Andréoletti, L., et al. (2007). Early archives of genetically-restricted proviral DNA in the female genital tract after heterosexual transmission of HIV-1. *AIDS*, *21*(October 2005), 153–162.

Chueca, N., Garrido, C., Álvarez, M., Poveda, E., de Dios Luna, J., Zahonero, N., Hernández-Quero, J., Soriano, V., Maroto, C., de Mendoza, C. and García, F. (2009). Improvement in the determination of HIV-1 tropism using the V3 gene sequence and a combination of bioinformatic tools. *Journal of Medical Virology, 81*: 763–767. doi: 10.1002/jmv.21425

Chun, T., & Fauci, A. S. (2012). HIV reservoirs: pathogenesis and obstacles to viral eradication and cure. *AIDS, 26*, 1261–1268. doi:10.1097/QAD.0b013e328353f3f1.

Chun, T.-W., Murray, D., Justement, J. S., Blazkova, J., Hallahan, C. W., Fankuchen, O., et al. (2014). Broadly neutralizing antibodies suppress HIV in the persistent viral reservoir. *Proceedings of the National Academy of Sciences* , *111* (36), 13151–13156. doi:10.1073/pnas.1414148111

Clarke, J. R., White, N. C., & Weber, J. N. (2000). HIV Compartmentalization: Pathogenesis and Clinical Implications. *AIDS Reviews*, *2*, 15–22.

Coetzer, M., Cilliers, T., Papathanasopoulos, M., Ramjee, G., Karim, S. A., Williamson, C., et al. (2007). Longitudinal analysis of HIV type 1 subtype C envelope sequences from South Africa. *AIDS Research and Human Retroviruses, 23*, 316–321

Coffin, J. M. (1979). Structure, Replication, and Recombination of Retrovirus Genomes: Some Unifying Hypotheses. *Journal of General Virology, 42*, 1–26.

Coffin, J.M., Hughes, S.H., & Vamus, H.E. (1997). Retroviruses. Plainview, *N.Y: Cold Spring Harbor Laboratory Press*. ISBN 0-87969-497-1.

Collins, K. R., Quiñones-mateu, M. E., Wu, M., Luzze, H., Johnson, J. L., Hirsch, C., Toossi, Z., et al. (2002). Human Immunodeficiency Virus Type 1 (HIV-1) Quasispecies at the Sites of Mycobacterium tuberculosis Infection Contribute to Systemic HIV-1 Heterogeneity. *Journal of Virology, 76*(4), 1697–1706. doi:10.1128/JVI.76.4.1697.

Connor, B. R. I., Sheridan, K. E., Ceradini, D., Choe, S., & Landau, N. R. (1997). Change in Coreceptor Use Correlates with Disease Progression in HIV-1 Infected individuals. The *Journal of Experimental Medicine, 185*(4), 621–8.

Connor, S. E. O., & Imperiali, B. (1996). Modulation of protein structure and function by asparagine-linked glycosylation. *Review, 3*, 803–812.

Crous, S., Shrestha, R. K., & Travers, S. A. (2012). Appraising the performance of genotyping tools in the prediction of coreceptor tropism in HIV-1 subtype C viruses. *BMC Infectious Diseases*, *12*(203), 1–8. doi:10.1186/1471-2334-12-203

Curlin, M. E., Zioni, R., Hawes, S. E., Liu, Y., Deng, W., Gottlieb, G. S., Zhu, T., et al. (2010). HIV-1 envelope subregion length variation during disease progression. *PLoS pathogens, 6*(12), e1001228. doi:10.1371/journal.ppat.1001228.

Da, D., Gu, R. L., & Ratner, L. (1992). Role of asparagine-linked glycosylation in human immunodeficiency virus type 1 transmembrane envelope function. *Virology, 187*(1), 377–382.

De Jong, J. J., De Ronde, A., Keulen, W., Tersmette, M., & Goudsmit, J. (1992). Minimal Requirements for the Human Immunodeficiency Virus Type 1 V3 Domain To Support the Syncytium-Inducing Phenotype: Analysis by Single Amino Acid Substitution. *Journal of Virology, 66*(11), 6777–80.

De Pasquale, M. P., Brown, A. J. L., Uvin, S. C., Allega-ingersoll, J., Caliendo, A. M., Sutton, L., et al. (2003). Differences in HIV-1 pol Sequences From Female Genital Tract and Blood During Antiretroviral Therapy. *Journal of Acquired Immune Deficiency Syndromes*, *34*(1), 37–44. doi:10.1097/00126334-200309010-00005

de Matos, A. L., McFadden, G., & Esteves, P. J. (2013). Positive evolutionary selection on the RIG-I-like receptor genes in mammals. *PloS One, 8*(11), e81864. doi:10.1371/journal.pone.0081864

de Tolly, K., Skinner, D., Nembaware, V., & Benjamin, P. (2012). Investigation into the Use of Short Message Services to Expand Uptake of Human Immunodeficiency Virus Testing, and Whether Content and Dosage Have Impact. *Telemedicine and e-Health, 18*(1), 18-23.

Dean, N. (1999). Asparagine-linked glycosylation in the yeast Golgi. *Biochimica et Biophysica Acta, 1426*, 309–322.

Delobel, P., Sandres-Saune, K., Cazabat, M., L'Faqihi, F. E., Aquilina, C., Obadia, M., Pasquier, C., et al. (2005). Persistence of distinct HIV-1 populations in blood monocytes and naive and memory CD4 T cells during prolonged suppressive HAART.

AIDS, *19*, 1739–1750.

Dictionary 3.0. Credibility. Available from http://www.dictionary30.com/meaning/Credibility

Diem, K., Nickle, D., C., Motoshige, A., Fox, A., Ross, S., Mullins, J., I., Corey, L., Coombs R., W., & Krieger, J., N. (2008). Male genital tract compartmentalization of human immunodeficiency virus type 1 (HIV). *AIDS Research and Human Retroviruses*. *24*(4), 561-571.

Dimitrov, D. S., Willey, R. L., Sato, H., Chang, L., Blumenthal, R., & Martin, M. A. (1993). Quantitation of Human Immunodeficiency Virus Type 1 Infection Kinetics. *Journal of Virology*, *67*(4), 2182–2190.

Dinoso, J. B., Kim, S. Y., Wiegand, a M., Palmer, S. E., Gange, S. J., Cranmer, L., et al. (2009). Treatment intensification does not reduce residual HIV-1 viremia in patients on highly active antiretroviral therapy. *Proceedings of the National Academy of Sciences*, *106*(23), 9403–8. doi:10.1073/pnas.0903107106

Dixit, N. M., & Perelson, A. S. (2004). Multiplicity of Human Immunodeficiency Virus Infections in Lymphoid Tissue. *Journal of Virology, 78*(16), 8942–8945. doi:10.1128/JVI.78.16.8942.

Dosenovic, P., Chakrabarti, B., Soldemo, M., Forsell, M. N. E., Li, Y., Phogat, A., et al. (2009). Selective expansion of HIV-1 envelope glycoprotein-specific B cell subsets recognizing distinct structural elements following immunization. *Journal of Immunology, 183*, 3373–82. doi:10.4049/jimmunol.0900407.

Drummond, A. J., & Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology, 7*(214). doi:10.1186/1471-2148-7-214

Drummond, A. J., Nicholls, G. K., Rodrigo, A. G., & Solomon, W. (2002). Estimating Mutation Parameters, Population History and Genealogy Simultaneously From Temporally Spaced Sequence Data. *Genetics Society of America, 1320*(July), 1307–1320.

Drummond, A. J., Pybus, O. G., Rambaut, A., Forsberg, R., & Rodrigo, A. G. (2003). Measurably evolving populations. *Trends in Ecology & Evolution, 18*(9), 481–488. doi:10.1016/S0169-5347(03)00216-7.

Dybowski, J. N., Heider, D., & Hoffmann, D. (2010). Prediction of co-receptor usage of HIV-1 from genotype. *PLoS Computational Biology, 6*(4), e1000743. doi:10.1371/journal.pcbi.1000743.

Edo-Matas, D., Lemey, P., Tom, J.A., Serna-Bolea, C., van den Blink, A.E., et al. (2011). Impact of CCR5delta32 Host Genetic Background and Disease Progression on HIV-1 Intrahost Evolutionary Processes: Efficient Hypothesis Testing through Hierarchical Phylogenetic Models. *Molecular Biology and Evolution, 28*(5). 1605-1616. doi: 10.1093/molbev/msq326.

Edo-Matas, D., van Gils, M. J., Bowles, E. J., Navis, M., Rachinger, A., Boeser-Nunnink, B., et al. (2010). Genetic composition of replication competent clonal HIV-1 variants isolated from peripheral blood mononuclear cells (PBMC), HIV-1 proviral DNA from PBMC and HIV-1 RNA in serum in the course of HIV-1 infection. *Virology*, *405*(2), 492–504. doi:10.1016/j.virol.2010.06.029

Engelman, A., & Cherepanov, P. (2012). The structural biology of HIV-1: mechanistic and therapeutic insights. Nature Reviews. *Microbiology, 10*(4), 279–90. doi:10.1038/nrmicro2747.

Erdmann, J. (2010). Chasing HIV from Its Hiding Place. *Chemistry & Biology, 17*(8), 787–788. doi:10.1016/j.chembiol.2010.08.003

Esbjörnsson, J., Månsson, F., Martínez-Arias, W., Vincic, E., Biague, A. J., Silva, Z. J., et al. (2010). Frequent CXCR4 tropism of HIV-1 subtype A and CRF02 _ AG during late-stage disease - indication of an evolving epidemic in West Africa. *Retrovirology, 7*(23), 1–13.

Esbjörnsson, J., Mild, M., Månsson, F., Norrgren, H., & Medstrand, P. (2011). HIV-1 Molecular Epidemiology in Guinea-Bissau, West Africa: Origin, Demography and Migrations. *PloS one, 6*(2). doi:10.1371/journal.pone.0017025.

Evering, T. H., Kamau, E., St. Bernard, L., Farmer, C. B., Kong, X-P., & Markowitz, M. (2014). Single genome analysis reveals genetic characteristics of Neuroadaptation across HIV-1 envelope. *Retrovirology*, *11*(1), 65. doi:10.1186/s12977-014-0065-0

Excoffier, L., Smouse, P. E., & Quattro, J. M. (1992). Analysis of Molecular Variance Inferred From Metric Distances Among DNA Haplotypes: Application to Human Mitochondrial DNA Restriction Data. *Genetics*, *131*(June), 479–491.

Felsenstein, J. (2006). Accuracy of Coalescent Likelihood Estimates: Do We Need More Sites, More Sequences, or More Loci? *Molecular Biology and Evolution*, *23*(3), 691–700. doi:10.1093/molbev/msj079

Fernandez, M. H., Faith, A., Higgins, J. A., Weber, J., & M, R. A. D. (1995). The effect of a single amino acid substitution within the V3 loop of HIV-1 gp120 on HLA-DR1-restricted CD4 T-cell recognition. *Immunology, 85*, 176–183.

Fessel, W. J., Anderson, B., Follansbee, S. E., Winters, M. A., Lewis, S. T., Weinheimer, S. P., et al. (2011). The efficacy of an anti-CD4 monoclonal antibody for HIV-1 treatment. *Antiviral Research, 92(*3), 484–487. doi:10.1016/j.antiviral.2011.09.010

Fitch, W.M. (1971). Toward defining the course of evolution: minimal change for a specific tree topology. *Systematic Zoology, 20*, 406–416.

Fouchier, R. A. M., Brouwer, M., Broersen, S. M., & Schuitemaker, H. (1995). Simple determination of human immunodeficiency virus type 1 syncytium-inducing V3 genotype by PCR. *Journal of Clinical Microbiology, 33*(4), 906–11.

Fouchier, R. A. M., Groenink, M., Kootstra, N. A., Tersmette, M., Huisman, H. G., Miedema, F., & Schuitemaker, H. (1992). Phenotype-Associated Sequence Variation in the Third Variable Domain of the Human Immunodeficiency Virus Type 1 gp120 Molecule. *Journal of Virology, 66*(5), 3183–3187.

François, K. O., & Balzarini, J. (2011). The highly conserved glycan at asparagine 260 of HIV-1 gp120 is indispensable for viral entry. *The Journal of Biological Chemistry, 286*(50), 42900–10. doi:10.1074/jbc.M111.274456

Frank, S. A. (2002). Immunology and Evolution of Infectious Disease. Princeton (NJ): Princeton University Press. Chapter 15, Measuring Selection with Population Samples. Available from: http://www.ncbi.nlm.nih.gov/books/NBK2379/

Fraser, C. (2005). HIV recombination: what is the impact on antiretroviral therapy? *Journal of the Royal Society*, *2*, 489–503. doi:10.1098/rsif.2005.0064

Frost, S. D. W., Wrin, T., Smith, D. M., Kosakovsky Pond, S. L., Liu, Y., Paxinos, E., et al. (2005). Neutralizing antibody responses drive the evolution of human immunodeficiency virus type 1 envelope during recent HIV infection. *Proceedings of the National Academy of Sciences, 102*(51), 18514–9. doi:10.1073/pnas.0504658102

Fulcher, J. A., Hwangbo, Y., Zioni, R., Nickle, D., Lin, X., Heath, L., Mullins, J. I., et al. (2004). Compartmentalization of Human Immunodeficiency Virus Type 1 between Blood Monocytes and CD4+ T Cells during Infection. *Journal of Virology, 78*(15), 7883–7893. doi:10.1128/JVI.78.15.7883.

Funnye, A. S., & Akhtar, A. J. (2003). Syphilis and Human Immunodeficiency Virus Co-Infection. *Journal of the National Medical Association*, *95*(5), 363–382.

Gallo, R. C., & Montagnier, L. (2003). The discovery of HIV as the cause of AIDS. *The New England Journal of Medicine*, *349*(24), 2283–5. doi:10.1056/NEJMp038194

Gallo, R. C., Salahuddin, S. Z., Popvic, M., Shearer, G. M., Kaplan, M., Haynes, B. F., et al. (1984). Frequent detection and isolation of cytopathic retroviruses (HTLV-III) from patients with AIDS and at risk for AIDS. *Science*, *224*(4 May), 500–503.

Galvin, S. R., & Cohen, M. S. (2004). The role of sexually transmitted diseases in HIV transmission. *Nature Reviews Microbiology*, *2*(1), 33–42. Retrieved from http://dx.doi.org/10.1038/nrmicro794

Gandhi, S. K., Siliciano, J. D., Bailey, J. R., Siliciano, R. F., & Blankson, J. N. (2008). Role of APOBEC3G/F-Mediated Hypermutation in the Control of Human Immunodeficiency Virus Type 1 in Elite Suppressors. *Journal of Virology, 82*(6), 3125–30. doi:10.1128/JVI.01533-07.

Ganser-Pornillos, B. K., Yeager, M., & Sundquist, W. I. (2008). The structural biology of HIV assembly. *Current opinion in structural biology, 18*(2), 203-217. doi: 10.1016/j.sbi.2008.02.001

Gantt, S., Carlsson, J., Heath, L., Bull, M. E., Shetty, A. K., Mutsvangwa, J., Musingwini, G., et al. (2010). Genetic Analyses of HIV-1 env Sequences Demonstrate Limited Compartmentalization in Breast Milk and Suggest Viral Replication within the Breast That Increases with Mastis. *Journal of Virology, 84*(20), 10812–10819. doi:10.1128/JVI.00543-10.

Garrett, L. (1995). *The Coming Plague: Newly Emerging Diseases in a World Out of Balance*. New York: Farrar, Straus and Giroux [Thirteen printing].

Geleziunas, R. (2012). Gilead HIV Eradication Program Eradicate Latently Infected Cells (pp. 1–28).

Gelfand, A. E., Hills, S. E., Racine-Poon, A., & Smith, A. F. M. (1990). Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling. *Journal of the American Statistical Association*, *85*(412), 972–985.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis, 1*(3), 515–533.

Geubbels, E., & Bowie, C. (2006). Epidemiology of HIV/AIDS in adults in Malawi. *Malawi Medical Journal, 18*(3), 99–121.

Go, E. P., Chang, Q., Liao, H., Sutherland, L. L., Alam, S. M., Barton, F., & Desaire, H. (2009). Glycosylation Site-Specific Analysis of Clade C HIV-1 Envelope Proteins. *Journal of Proteome Research, 8*(9), 4231–4242. doi:10.1021/pr9002728.Glycosylation

Gorny, M. K., Moore, J. P., Conley, A. J., Karawowska, S., Sodroski, J., Williams, C., et al. (1994). Human Anti-V2 Monoclonal Antibody That Neutralizes Primary but Not Laboratory Isolates of Human Immunodeficiency Virus Type 1. *Journal of Virology, 68*(12), 8312–8320.

Gottlieb, G. S., Hawes, S. E., Wong, K. G., Raugi, D. N., Agne, H. D., Critchlow, C. W., Kiviat, N. B., et al. (2008). Envelope Viral Variation in the PBMC and Genital Tract of ARV-Naive Women in Senegal. *AIDS Research and Human Retroviruses, 24*(6), 857–864. doi:10.1089/aid.2008.0015.

Gottlieb, G. S., Nickle D. C., Jensen, M. A., Wong K. G., Grobler, J., Li, F., Liu, S-L., et al. (2004). Dual HIV-1 infection associated with rapid disease progression. *The Lancet, 363*(9409), 619-622. doi: 10.1016/S0140-6736(04)15596-7

Goulder, P. J. R., & Watkins, D. I. (2004). HIV and SIV CTL escape: implications for vaccine design. *Nature Reviews Immunology*, *4*(8), 630–640. Retrieved from http://dx.doi.org/10.1038/nri1417

Graci, J. D., Colacino, J. M., Peltz, S. W., Dougherty, J. P., & Gu, Z. (2008). Review HIV type-1 latency: targeted induction of proviral reservoirs. *Antiviral Chemistry & Chemotherapy, 19,* 177–187.

Gray, E. S., Madiga, M. C., Hermanus, T., Moore, P. L., Wibmer, C. K., Tumba, N. L., et al. (2011). The neutralization breadth of HIV-1 develops incrementally over four years and is associated with CD4+ T cell decline and high viral load during acute infection. *Journal of Virology*, *85*(10), 4828–40. doi:10.1128/JVI.00198-11

Griemberg, G., Bautista, C. T., Pizzimenti, M. C., Orfus, G., Alonso, B., Fernandez T., et al. (2006). High prevalence of syphilis-HIV co-infection at four hospitals of the City of Buenos Aires, Argentina. *Revista Argentina de Microbiología*, *38*, 134–136.

Grobler, J., Gray, C. M., Rademeyer, C., Seoighe, C., Ramjee, G., Karim, S. A., et al. (2004). Incidence of HIV-1 Dual Infection and Its Association with Increased Viral Load Set Point in a Cohort of HIV-1 Subtype C – Infected Female Sex Workers. *The Journal of Infectious Diseases*, *190*(7), 1355–9. doi:10.1086/423940

Gupta, G. R., Whelan, D., & Allendorf, K. (2003). Integrating Gender HIV/AIDS Programmes into. *WHO Library Cataloguing-in-Publication Data*, ISBN 9241590394.

Haaland, R. E., Hawkins, P. A., Salazar-Gonzalez, J., Johnson, A., Tichacek, A., Karita, E., et al. (2009). Inflammatory genital infections mitigate a severe genetic bottleneck in heterosexual transmission of subtype A and C HIV-1. *PLoS Pathogens*, *5*(1), e1000274. doi:10.1371/journal.ppat.1000274

Harrington, P. R., Schnell, G., Letendre, S. L., Ritola, K., Robertson, K., Hall, C., et al. (2011). Cross-sectional characterization of HIV-1 env compartmentalization in cerebrospinal fluid over the full disease course. *AIDS*, *23*(8), 115–124. doi:10.1016/j.virol.2009.11.032.Temporal

Hart, M. L., Saifuddin, M., & Spear, G. T. (2003). Glycosylation inhibitors and neuraminidase enhance human immunodeficiency virus type 1 binding and neutralization by mannose-binding lectin. *Journal of General Virology, 84*, 353–360. doi:10.1099/vir.0.18734-0.

Heath, L., Fox, A., McClure, J., Diem, K., van 't Wout, A. B., Zhao, H., et al. (2009). Evidence for limited genetic compartmentalization of HIV-1 between lung and blood. *PloS One, 4*(9), e6949. doi:10.1371/journal.pone.0006949.

Heeregrave, E. J., Geels, M. J., Brenchley, J. M., Baan, E., Ambrozak, D. R., van der Sluis, R. M., et al. (2009). Lack of in vivo compartmentalization among HIV-1 infected naive and memory CD4+ T cell subsets. *Virology*, *393*(1), 24–32. doi:10.1016/j.virol.2009.07.011.Lack

Henklein, P., Bruns, K., Sherman, M. P., Tessmer, U., Licha, K., Kopp, J., et al. (2000). Functional and structural characterization of synthetic HIV-1 Vpr that transduces cells,

localizes to the nucleus, and induces G2 cell cycle arrest. *The Journal of Biological Chemistry, 275*(41), 32016–26. doi:10.1074/jbc.M004044200.

Herbeck, J. T., Rolland, M., Liu, Y., McLaughlin, S., McNevin, J., Zhao, H., et al. (2011). Demographic processes affect HIV-1 evolution in primary infection before the onset of selective processes. *Journal of Virology, 85*(15), 7523–34. doi:10.1128/JVI.02697-10

Hiller, N. L., Ahmed, A., Powell, E., Martin, D. P., Eutsey, R., Earl, J., et al. (2010). Generation of genic diversity among Streptococcus pneumoniae strains via horizontal gene transfer during a chronic polyclonal pediatric infection. *PLoS Pathogens, 6*(9), e1001108. doi:10.1371/journal.ppat.1001108.

Hirbod, T., & Broliden, K. (2007). Mucosal immune responses in the genital tract of HIV-1-exposed uninfected women. *Journal of Internal Medicine*, *262*, 44–58. doi:10.1111/j.1365-2796.2007.01822.x

Hladik, F., & Hope, T. J. (2009). HIV infection of the genital mucosa in women. *Current HIV/AIDS Reports*, *6*(1), 20–28. doi:10.1007/s11904-009-0004-1

Hoffman, N. G., Seillier-Moiseiwitsch, F., Ahn, J., Walker, J. M., & Swanstrom, R. (2002). Variability in the Human Immunodeficiency Virus Type 1 gp120 Env Protein Linked to Phenotype-Associated Changes in the V3 Loop. *Journal of Virology, 76*(8), 3852–3864. doi:10.1128/JVI.76.8.3852

Holsinger, K. E., & Weir, B. S. (2009). Genetics in geographically structured populations: defining, estimating and interpreting FST. *Nature Reviews. Genetics*, *10*(9), 639–50. doi:10.1038/nrg2611

Huang, W., Eshleman, S. H., Toma, J., Fransen, S., Stawiski, E., Paxinos, E. E., et al. (2007). Coreceptor Tropism in Human Immunodeficiency Virus Type 1 Subtype D: High Prevalence of CXCR4 Tropism and Heterogeneous Composition of Viral Populations. *Journal of Virology, 81*(15), 7885–93. doi:10.1128/JVI.00218-07.

Hudson, R. R. (2000). A New Statistic for Detecting Genetic Differentiation. *Genetics*, *155*, 2011–2014.

Hué, S., Gray, E. R., Gall, A., Katzourakis, A., Tan, P. C., et al. (2010). Disease-associated XMRV sequences are consistent with laboratory contamination. *Retrovirology, 7*(111), doi:10.1186/1742-4690-7-11.

Huelsenbeck, J.P. & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogeny. *Bioinformatics, 17*, 754–755.

Iglesias-Ussel, M. D., & Romerio, F. (2011). HIV Reservoirs: The New Frontier. *AIDS Reviews, 13*, 13–29.

Imamichi, H., Degray, G., Dewar, R. L., Mannon, P., Yao, M., Chairez, C., et al. (2011). Lack of Compartmentalization of HIV-1 Quasispecies Between the Gut and Peripheral Blood Compartments. *The Journal of Infectious Diseases*, *2014*(15 July), 309–314. doi:10.1093/infdis/jir259

Itescu, S., Simonelli, P. F., Winchester, R. J., & Ginsberg, H. S. (1994). Human immunodeficiency virus type 1 strains in the lungs of infected individuals evolve independently from those in peripheral blood and are highly conserved in the C-terminal region of the envelope V3 loop. *Proceedings of the National Academy of Sciences*, *91*(24), 11378–82.

Jacobs, G. B., Wilkinson, E., Isaacs, S., Spies, G., de Oliveira, T., Seedat, S., & Engelbrecht, S. (2014). HIV-1 subtypes B and C unique recombinant forms (URFs) and transmitted drug resistance identified in the Western Cape Province, South Africa. *PloS One, 9*(3), e90845. doi:10.1371/journal.pone.0090845

Janini, M., Rogers, M., Birx, D.R. & McCutchan F.E (2001). Human Immunodeficiency Virus Type 1 DNA Sequences Genetically Damaged by Hypermutation Are Often Abundant in Patient Peripheral Blood Mononuclear Cells and May Be Generated during Near-Simultaneous Infection and Activation of CD4+ T Cells. *Journal of Virology, 75*(17), pp.7973–7986.

Jeffreys, H. (1961). Theory of Probability (3rd ed.), Oxford, UK: Oxford University Press.

Jensen, M. A., Li, F., Wout, A. B. Van, Nickle, D. C., Shriner, D., He, H., et al. (2003). Improved Coreceptor Usage Prediction and Genotypic Monitoring of R5-to-X4 Transition by Motif Analysis of Human Immunodeficiency Virus Type 1 env V3 Loop Sequences. *Journal of Virology, 77*(24), 13376–13388. doi:10.1128/JVI.77.24.13376.

Johansson, S. E., Rollman, E., Chung, A. W., Center, R. J., Hejdeman, B., Stratov, I., et al. (2011). NK Cell Function and Antibodies Mediating ADCC in HIV-1-Infected Viremic and Controller Patients. *Viral Immunology, 24*(5), 359–368. doi:10.1089/vim.2011.0025

Johnson, L. F., & Lewis, D. A. (2008). The effect of genital tract infections on HIV-1

shedding in the genital tract: a systematic review and meta-analysis. *Sexually Transmitted Diseases, 35*(11), 946-959.

Josefsson, L., Palmer, S., Faria, N. R., Lemey, P., Casazza, J., Ambrozak, D., et al. (2013). Single cell analysis of lymph node tissue from HIV-1 infected patients reveals that the majority of CD4+ T-cells contain one HIV-1 DNA molecule. *PLoS Pathogens*, *9*(6), e1003432. doi:10.1371/journal.ppat.1003432

Kaewmuangmoon, J., Suwanvijitr, T., Cherdshewasart, W. & Chanchao, C. (2010). Leaf morphometric and genetic variation of Butea superba in Thailand. *ScienceAsia, 36*, doi:10.2306/scienceasia1513-1874.2010.36.180.

Kaiser Family Foundation. (2013). The Global HIV/AIDS Epidemic: A Timeline of Key Milestones, Available at: http://kaiserfamilyfoundation.files.wordpress.com/2008/08/global-hiv-aids-timeline-050313.pdf

Kantanen, M. L., Leinikki, P., & Kuismanen, E. (1995). Endoproteolytic cleavage of HIV-1 gp160 envelope precursor occurs after exit from the trans-Golgi Network (TGN). *Archives of Virology, 140*(8). 1441-1449.

Karim, A. Q., Sibeko, S., & Baxter, C. (2010). Preventing HIV Infection in Women: A Global Health Imperative. *Clinical Infectious Diseases, 50*(Suppl 3), S122–9. doi:10.1086/651483.

Karlen, A. (1995). *Plague's Progress: A Social History of Man and Disease*. London: Victor Gollancz.

Kasturi, L., Chen, H. & Shakin-Eshleman, S.H. (1997). Regulation of N-linked core glycosylation: use of a site-directed mutagenesis approach to identify Asn-Xaa-Ser/Thr sequons that are poor oligosaccharide acceptors. *Biochemical Journal, 323*(Pt 2). 415-419.

Kato, K., Sato, H., & Takebe, Y. (1999). Role of Naturally Occurring Basic Amino Acid Substitutions in the Human Immunodeficiency Virus Type 1 Subtype E Envelope V3 Loop on Viral Coreceptor Usage and Cell Tropism. *Journal of Virology, 73*(7), 5520–6.

Kaushic, C. (2010). HIV-1 Infection in the Female Reproductive Tract: Role of Interactions between HIV-1 and Genital Epithelial Cells. *American Journal of Reproductive Immunology, 65*, 253–260. doi:10.1111/j.1600-0897.2010.00965.

Kearney, M., Maldarelli, F., Shao, W., Margolick, J. B., Daar, E. S., Mellors, J. W., et al. (2009). Human immunodeficiency virus type 1 population genetics and adaptation in newly infected individuals. *Journal of Virology, 83*(6), 2715–27. doi:10.1128/JVI.01960-08

Keele, B. F., Giorgi, E. E., Salazar-Gonzalez, J. F., Decker, J. M., Pham, K. T., Salazar, M. G., et al. (2008). Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proceedings of the National Academy of Sciences*, *105*(21), 7552–7. doi:10.1073/pnas.0802203105

Kemal, K.S., Foley, B., Burger, H., Anastos, K., Minkoff, H., Kitchen, C., Philpott, S.M., et al. (2003). HIV-1 in genital tract and plasma of women: Compartmentalization of viral sequences, coreceptor usage, and glycosylation. *Proceedings of the National Academy of Sciences, 100*(22), 12972-12977. doi: 10.1073/pnas.2134064100.

Kitamura, K., Wang, Z., Chowdhury, S., Simadu, M., Koura, M., & Muramatsu, M. (2013). Uracil DNA Glycosylase Counteracts APOBEC3G-Induced Hypermutation of Hepatitis B Viral Genomes: Excision Repair of Covalently Closed Circular DNA. *PLoS Pathogens, 9*(5), e1003361. doi:10.1371/journal.ppat.1003361.

Kitchen, C. M. R., Philpott, S., Burger, H., Weiser, B., Anastos, K., & Suchard, M. A. (2004). Evolution of Human Immunodeficiency Virus Type 1 Coreceptor Usage during Antiretroviral Therapy: a Bayesian Approach. *Journal of Virology, 78*(20), 11296–11302. doi:10.1128/JVI.78.20.11296.

Kiwelu, I. E., Novitsky, V., Margolin, L., Baca, J., Manongi, R., Sam, N., et al. (2013). Frequent intra-subtype recombination among HIV-1 circulating in Tanzania. *PloS One, 8*(8), e71131. doi:10.1371/journal.pone.0071131

Kiwelu, I. E., Novitsky, V., Margolin, L., Baca, J., Manongi, R., Sam, N., et al. (2012). HIV-1 subtypes and recombinants in Northern Tanzania: distribution of viral quasispecies. *PloS One, 7*(10), e47605. doi:10.1371/journal.pone.0047605

Korber, B. T., Kunstman, K. J., Patterson, B. K., Furtado, M., McEvilly, M. M., Levy, R., & Wolinsky, S. M. (1994). Genetic differences between blood- and brain-derived viral sequences from human immunodeficiency virus type 1-infected patients: evidence of conserved elements in the V3 region of the envelope protein of brain-derived sequences. *Journal of Virology*, *68*(11), 7467–81.

Koup, R. A., Safrit, J. T., Cao, Y., Andrews, C. A., Mcleod, G., Borkowsky, W., et al. (1994). Temporal Association of Cellular Immune Responses with the Initial Control of Viremia in Primary Human Immunodeficiency Virus Type 1 Syndrome. *Journal of Virology*, *68*(7), 4650–4655.

Kourteva, Y., De Pasquale, M., Allos, T., McMunn, C., & D'Aquila, R. T. (2012). APOBEC3G expression and hypermutation are inversely associated with human immunodeficiency virus type 1 (HIV-1) burden in vivo. *Virology, 430*(1), 1–9. doi:10.1016/j.virol.2012.03.018.

Kovacs, A., Wasserman, S. S., Burns, D., Wright, D. J., Cohn, J., Landay, A., et al. (2001). Determinants of HIV-1 shedding in the genital tract of women. *Lancet*, *358*(9293), 1593–601. doi:10.1016/S0140-6736(01)06653-3

Kumar, P. (2013). Long term non-progressor (LTNP) HIV infection. *The Indian Journal of Medical Research, 138*(3), 291–3.

Lacerda, M., Moore, P. L., Ngandu, N. K., Seaman, M., Gray, E. S., Murrell, B., et al. (2013). Identification of broadly neutralizing antibody epitopes in the HIV-1 envelope glycoprotein using evolutionary models. *Virology Journal, 10*(347), 1–18. doi:10.1186/1743-422X-10-347

Lamers, S. L., Salemi, M., Galligan, D. C., Oliveira, T. De, Fogel, G. B., Sara, C., et al. (2009). Extensive HIV-1 Intra-Host Recombination Is Common in Tissues with Abnormal Histopathology. *PLoS ONE, 4(*3), 1–11. doi:10.1371/journal.pone.0005065.

Lamine, A., Caumont-Sarcos, A., Chaix, M. L., Saez-Cirion, A., Rouzioux, C., Delfraissy, J. F., Pancino, G., & Lambotte, O. (2007). Replication-competent HIV strains infect HIV controllers despite undetectable viremia (ANRS EP36 study). *AIDS, 21*(8), 1043-5.

Land, A. M., Ball, T. B., Luo, M., Pilon, R., Sandstrom, P., Embree, J. E., et al. (2008). Human Immunodeficiency Virus (HIV) Type 1 Proviral Hypermutation Correlates with CD4 Count in HIV-Infected Women from Kenya. *Journal of Virology, 82*(16), 8172–8182. doi:10.1128/JVI.01115-08.

Land, A., & Braakman, I. (2001). Folding of the human immunodeficiency virus type 1 envelope glycoprotein in the endoplasmic reticulum. *Biochimie, 83*(8), 783–790.

Land, A., Zonneveld, D., & Braakman, I. (2003). Folding of HIV-1 Envelope glycoprotein involves extensive isomerization of disulfide bonds and conformation-dependent leader

peptide cleavage. *Federation of American Societies for Experimental Biology, 17*(9), 1058–1067.

Laprise, C., Pokomandy, A. De, Baril, J., Dufresne, S., & Trottier, H. (2013). Virologic Failure Following Persistent Low-level Viremia in a Cohort of HIV-Positive Patients: Results From 12 Years of Observation. *Clinical Infectious Diseases*, 1–8. doi:10.1093/cid/cit529.

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*.2947-2948.

Lau, K. A., & Wong, J. J. L. (2013). Current trends of HIV recombination worldwide. *Infectious Disease Reports, 5*(1S), 15–20. doi:10.4081/idr.2013.s1.e4

Lavreys, L., Baeten, J. M., Panteleeff, D. D., Richardson, B. A., McClelland, R. S., Chohan, V., Mandaliya, K., Ndinya-Achola, J. O., & Overbaugh, J. (2006). High levels of cervical HIV-1 RNA during early HIV-1 infection. *AIDS, 20*(18), 2389-2390.

Lee, W., Syu, W., Dut, B. I. N., Matsuda, M., Tan, S., Wolft, A., et al. (1992). Nonrandom distribution of gpl20 N-linked glycosylation sites important for infectivity of human immunodeficiency virus type 1. *Proceedings of the National Academy of Sciences of the United States of America, 89*(March), 2213–2217.

Lemey, P., Kosakovsky Pond, S. L., Drummond, A. J., Pybus, O. G., Shapiro, B., Barroso, H., et al. (2007). Synonymous substitution rates predict HIV disease progression as a result of underlying replication dynamics. *PLoS Computational Biology*, *3*(2), e29. doi:10.1371/journal.pcbi.0030029

Lemey, P., Rambaut, A., Drummond, A.J. & Suchard, M.A. (2009). Bayesian phylogeography finds its roots. *PLoS Computational Biology, 5*:e1000520.

Leslie, A. J., Pfafferott, K. J., Chetty, P., Draenert, R., Addo, M. M., Feeney, M., et al. (2004). HIV evolution: CTL escape mutation and reversion after transmission. *Nature Medicine*, *10*(3), 282–289. Retrieved from http://dx.doi.org/10.1038/nm992

Letunic, I. & Bork, P. (2011). Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Research, 39*(2), 475-478, doi:10.1093/nar/gkr201.

Levy, D. N., Aldrovandi, G. M., Kutsch, O., & Shaw, G. M. (2004). Dynamics of HIV-1 recombination in its natural target cells. *Proceedings of the National Academy of Sciences, 101*(12), 4204–9. doi:10.1073/pnas.0306764101

Lewis, M. J., Frohnen, P., Ibarrondo, F. J., Reed, D., Iyer, V., Ng, H. L., et al. (2013). HIV-1 Nef Sequence and Functional Compartmentalization in the Gut Is Not Due to Differential Cytotoxic T Lymphocyte Selective Pressure. *PloS One*, *8*(9). doi:10.1371/journal.pone.0075620

Li, H., Chien, P. C., Tuen, M., Visciano, M. L., Cohen, S., Blais, S., et al. (2008). Identification of an N-Linked Glycosylation in the C4 Region of HIV-1 Envelope gp120 That Is Critical for Recognition of Neighboring CD4 T Cell Epitopes. *The Journal of Immunology*, *180*(6), 4011–4021. doi:10.4049/jimmunol.180.6.4011

Lihana, R. W., Ssemwanga, D., Abimiku, A., & Ndembi, N. (2012). Update on HIV-1 Diversity in Africa: A Decade in Review. *AIDS Reviews, 14*(2), 83–100.

Love, R. P., Xu, H., & Chelico, L. (2012). Biochemical analysis of hypermutation by the deoxycytidine deaminase APOBEC3A. *The Journal of Biological Chemistry, 287*(36), 30812–22. doi:10.1074/jbc.M112.393181.

Luft, L. M., Gill, M. J., & Church, D. L. (2011). HIV-1 viral diversity and its implications for viral load testing: review of current platforms. *International Journal of Infectious Diseases*, *15*, e661–70. doi:10.1016/j.ijid.2011.05.013

Lukashov, V. V, & Goudsmit, J. (1997). Evolution of the Human Immunodeficiency Virus Type 1 Subtype-Specific V3 Domain Is Confined to a Sequence Space with a Fixed Distance to the Subtype Consensus. *Journal of Virology*, *71*(9), 6332–6338.

Luseno, W. K., & Wechsberg, W. M. (2009). Correlates of HIV testing among South Africa women with high sexual and substance-use risk behaviours. *AIDS Care, 21*(2), 178–184. doi:10.1080/09540120802017594.Correlates

Lusso, P. (2006). HIV and the chemokine system: 10 years later. *The EMBO Journal*, *25*(3), 447–56. doi:10.1038/sj.emboj.7600947

Lynch, S. M. (2007). *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*. (S. M. Lynch, Ed.) (pp. 1–359). New York, NY: Springer New York. doi:10.1007/978-0-387-71265-9

Malherbe, D. C., Sanders, R. W., van Gils, M. J., Park, B., Gomes, M. M., Schuitemaker, H., et al. (2013). HIV-1 Envelope Glycoprotein Resistance to Monoclonal Antibody 2G12 Is Subject-Specific and Context- Dependent in Macaques and Humans. *PloS One, 8*(9), e75277. doi:10.1371/journal.pone.0075277

Malim, M. H., & Bieniasz, P. D. (2012). HIV Restriction Factors and Mechanisms of Evasion. *Cold Spring Harbor Perspectives in Medicine, 2*(a006940). doi:10.1101/cshperspect.a006940.

Mangeat, B., Turelli, P., Caron, G., Friedli, M., Perrin, L., & Trono, D. (2003). Broad antiretroviral defence by human APOBEC3G through lethal editing of nascent reverse transcripts. *Nature, 424*(6944), 99–103. doi:10.1038/nature01709.

Mann, H. B., & Whitney, D. R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Annals of Mathematical Statistics, 18*(1), 50–60. doi:10.1214/aoms/1177730491.

Marchant, D., Neil, S. J. D., & McKnight, A. (2006). Human immunodeficiency virus types 1 and 2 have different replication kinetics in human primary macrophage culture. *The Journal of General Virology*, *87*(2), 411–8. doi:10.1099/vir.0.81391-0

Marras, D., Bruggeman, L. A., Gao, F., Tanji, N., Mansukhani, M. M., Cara, A., et al. (2002). Replication and compartmentalization of HIV-1 in kidney epithelium of patients with HIV-associated nephropathy. *Nature Medicine*, *8*(5), 522–526.

Martin, D. P., Lemey, P., Lott, M., Moulton, V., Posada, D., & Lefeuvre, P. (2010). RDP3: a flexible and fast computer program for analyzing recombination. *Bioinformatics (Oxford, England), 26(*19), 2462–3. doi:10.1093/bioinformatics/btq467.

Martinez-Picado, J., Prado, J. G., Fry, E. E., Pfafferott, K., Leslie, A., Chetty, S., et al. (2006). Fitness Cost of Escape Mutations in p24 Gag in Association with Control of Human Immunodeficiency Virus Type 1. *Journal of Virology*, *80*(7), 3617–3623. doi:10.1128/JVI.80.7.3617

Marx, P. A., Alcabes, P. G., & Drucker, E. (2001). Serial human passage of simian immunodeficiency virus by unsterile injections and the emergence of epidemic human immunodeficiency virus in Africa. *Philosophical Transactions of the Royal Society, 356* (1410), 911-920.

Mascola, J. R., & Montefiori, D. C. (2003). HIV-1: nature's master of disguise. *Nature Medicine, 9(*4), 393-4.

McGrath, K. M., Hoffman, N. G., Resch, W., Nelson, J. A., & Swanstrom, R. (2001). Using HIV-1 sequence variability to explore virus biology. *Virus Research*, *76*(2), 137–160. doi:10.1016/S0168-1702(01)00271-4

McNamara, L. A., & Collins, K. L. (2011). Hematopoitic stem/precursor cells as HIV reservoirs. *Current Opinion in HIV and AIDS*, *6*(1), 43–48. doi:10.1097/COH.0b013e32834086b3.Hematopoietic

Mesplède, T., Quashie, P. K., Osman, N., Han, Y., Singhroy, D. N., Lie, Y., et al. (2013). Viral fitness cost prevents HIV-1 from evading dolutegravir drug pressure. *Retrovirology*, *10*(22), 1–7. doi:10.1186/1742-4690-10-22

Mild, M., Esbjörnsson, J., Fenyö, E. M., & Medstrand, P. (2007). Frequent intrapatient recombination between human immunodeficiency virus type 1 R5 and X4 envelopes: Implications for coreceptor switch. *Journal of Virology, 81*(7), 3369–76. doi:10.1128/JVI.01295-06

Miyauchi, K., Kim, Y., Latinovic. O., Morozov. V., & Melikyan, G. B. (2009). HIV enters cells via endocytosis and dynamin-dependent fusion with endosomes. *Cell, 1*(137), 433-444. doi: 10.1016/j.cell.2009.02.046.

Monel, B., Beaumont, E., Vendrame, D., Schwartz, O., Brand, D., & Mammano, F. (2012). HIV Cell-to-Cell Transmission Requires the Production of Infectious Virus Particles and Does Not Proceed through Env-Mediated. *Journal of Virology*, *86*(7), 3924–33. doi:10.1128/JVI.06478-11

Moore, P. L., Gray, E. S., Wibmer, C. K., Bhiman, J. N., Nonyane, M., Sheward, D. J., et al. (2012). Evolution of an HIV glycan-dependent broadly neutralizing antibody epitope through immune escape. *Nature Medicine, 18*(11), 1688–92. doi:10.1038/nm.2985

Mostowy, R., Kouyos, R. D., Fouchet, D., & Bonhoeffer, S. (2011). The role of recombination for the coevolutionary dynamics of HIV and the immune response. *PloS one, 6*(2), e16052. doi:10.1371/journal.pone.0016052

Motulsky, H. (2003). *The InStat guide to choosing and interpreting statistical tests* (pp. 1–126).

Mousseau, G., & Valente, S. (2012). Strategies to Block HIV Transcription: Focus on Small Molecule Tat Inhibitors. *Biology, 1*(3), 668–697. doi:10.3390/biology1030668.

Müller, V., & Bonhoeffer, S. (2005). Guanine-adenine bias: a general property of retroid viruses that is unrelated to host-induced hypermutation. *Trends in Genetics, 21*(5), 264–268. doi:10.1016/j.tig.2005.03.002

Murray, J. M., Emery, S., Kelleher, A. D., Law, M., Chen, J., Hazuda, D. J., Nguyen, B. Y., Teppler, H., & Cooper, D. A. (2007). Antiretroviral therapy with the integrase inhibitor raltegravir alters decay kinetics of HIV, significantly reducing the second phase. *AIDS. 21*(17). 2315-21.

Murrell, B., Moola, S., Mabona, A., Weighill, T., Sheward, D., Pond, S. L. K., & Scheffler, K. (2013). FUBAR: A Fast, Unconstrained Bayesian AppRoximation for Inferring Selection. *Molecular Biology and Evolution*, *30*(5), 1196–205. doi:10.1093/molbev/mst030

N'Galy, B., Ryder, R., Bila, K., Mwandagalirwa, K., Colebunders, R. L., Francis, H., et al. (1998). Human immunodeficiency virus infection among employees in an African hospital. *New England Journal of Medicine, 319*, 1123–7.

National Institutes of Health (2008). NIH: National Institute of Allergy and Infectious Diseases, Available at: http://www.niaid.nih.gov/topics/hivaids/understanding/population%20specific%20information/pages/womenhiv.aspx

Neher, R. A., & Leitner, T. (2010). Recombination rate and selection strength in HIV intra-patient evolution. *PLoS Computational Biology, 6*(1), e1000660. doi:10.1371/journal.pcbi.1000660

Nei, M. & Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution, 3*, 418-426.

Neogi, U., Bontell, I., Shet, A., De Costa, A., Gupta, S., Diwan, V., et al. (2012). Molecular Epidemiology of HIV-1 Subtypes in India: Origin and Evolutionary History of the Predominant Subtype C. *PloS One, 7*(6), e39819. doi:10.1371/journal.pone.0039819

Newton, M. A., & Raftery, A. E. (1994). Approximate Bayesian Inference with the Weighted Likelihood Bootstrap. *Journal of the Royal Statistical Society*, *56*(1), 3–48.

NIAID. (2012). Biology of HIV, Available at: http://www.niaid.nih.gov/topics/hivaids/understanding/biology/Pages/biology.aspx

Noguchi, C., Ishino, H., Tsuge, M., Fujimoto, Y., Imamura, M., Takahashi, S., & Chayama, K. (2005). G to A hypermutation of hepatitis B virus. *Hepatology, 41*(3), 626–33. doi:10.1002/hep.20580.

Norman, J. M., Mashiba, M., McNamara, L. a, Onafuwa-Nuga, A., Chiari-Fort, E., Shen, W., & Collins, K. L. (2011). The antiviral factor APOBEC3G enhances the recognition of HIV-infected primary T cells by natural killer cells. *Nature Immunology, 12*(10), 975–83. doi:10.1038/ni.2087.

Novitsky, V., Lagakos, S., Herzig, M., Bonney, C., Kebaabetswe, L., Rossenkhan, R., et al. (2009). Evolution of proviral gp120 over the first year of HIV-1 subtype C infection. *Virology, 383*(1), 47–59. doi:10.1517/17425250903483207.Efavirenz

Novitsky, V., Smith, U. R., Gilbert, P., Mclane, M. F., Williamson, C., Ndung, T., et al. (2002). Human Immunodeficiency Virus Type 1 Subtype C Molecular Phylogeny: Consensus Sequence for an AIDS Vaccine Design? *Journal of Virology, 76*(11), 5435–5451. doi:10.1128/JVI.76.11.5435

Novitsky, V., Wang, R., Margolin, L., Baca, J., Rossenkhan, R., Widenfelt, E. Van, & Essex, M. (2011). Transmission of Single and Multiple Viral Variants in Primary HIV-1 Subtype C Infection. *PloS One, 6*(2), 1–17. doi:10.1371/journal.pone.0016714.

Ogert, R. A., Lee, M. K., Ross, W., Buckler-White, A., Martin, M. A., & Cho, M. W. (2001). N-Linked Glycosylation Sites Adjacent to and within the V1/V2 and the V3 Loops of Dualtropic Human Immunodeficiency Virus Type 1 Isolate DH12 gp120 Affect Coreceptor Usage and Cellular Tropism. *Journal of Virology, 75*(13), 5998–6006. doi:10.1128/JVI.75.13.5998.

Ohagen, a., Devitt, a., Kunstman, K. J., Gorry, P. R., Rose, P. P., Korber, B., et al. (2003). Genetic and Functional Analysis of Full-Length Human Immunodeficiency Virus Type 1 env Genes Derived from Brain and Blood of Patients with AIDS. *Journal of Virology*, *77*(22), 12336–12345. doi:10.1128/JVI.77.22.12336-12345.2003

Onafuwa-Nuga, A., & Telesnitsky, A. (2009). The remarkable frequency of human immunodeficiency virus type 1 genetic recombination. *Microbiology and Molecular Biology Reviews, 73*(3), 451–80. doi:10.1128/MMBR.00012-09

Ostrowski, M. A., Krakauer, D. C., Li, Y., Justement, S. J., Learn, G., Ehler, L. A., et al. (1998). Effect of Immune Activation on the Dynamics of Human Immunodeficiency Virus Replication and on the Distribution of Viral Quasispecies. *Journal of Virology, 72*(10), 7772–7784.

Ottander, U., Nakata, M., Backstrom, T., Liu, K., Ny, T., & Olofsson, J. I. (1997). Compartmentalization of human chorionic gonadotrophin sensitivity and luteinizing hormone receptor mRNA in different subtypes of the human corpus luteum. *Human Reproduction*, *12*(5), 1037–1042.

Over, M. (2001). Impact of the HIV/AIDS Epidemic on the Health Sectors of Developing Countries (pp. 311–344).

Overbaugh, J., Anderson, R. J., Ndinya-Achola, J. O., & Kreiss, J. K. (1996). Distinct but related human immunodeficiency virus type 1 variant populations in genital secretions and blood. AIDS Research and Human Retroviruses 12(2), 107–115

Pace, C., Keller, J., Nolan, D., James, I., Moore, C., & Mallal, S. (2006). Population Level Analysis of Human Immunodeficiency Virus Type 1 Hypermutation and Its Relationship with APOBEC3G and vif Genetic Variation. *Journal of Virology, 80*(18), 9259–9269. doi:10.1128/JVI.00888-06.

Palmer, S. (2013). Advances in detection and monitoring of plasma viremia in HIV-infected individuals receiving antiretroviral therapy. *New Advances in Immunological and Virological Monitoring, 8*(00). doi:10.1097/COH.0b013e32835d80af.

Pantaleo, G. (2000). Mechanisms of Human Immunodeficiency Virus (HIV) Escape from the Immune Response. *Preisverleihung/Stiftung Professor Dr. Max Cloëtta*, 27(28), 1–128.

Paraschiv, S., Bâțan, I., Bănică, L., Niculescu, I., & Oțelea, D. (2014). Baseline HIV-1 tropism prediction in advanced immune suppressed patients: evidence of CXCR4 viruses in IDUs infected with recombinant forms. *BMC Infectious Diseases, 14*(Suppl 4), O20. doi:10.1186/1471-2334-14-S4-O20

Parker, J., Rambaut, A., & Pybus, O. G. (2008). Correlating viral phenotypes with phylogeny: Accounting for phylogenetic uncertainty. *Infection, Genetics and Evolution, 8*, 239–246. doi:10.1016/j.meegid.2007.08.001.

Penn, M. L., Grivel, J.-C., Schramm, B., Goldsmith, M. A., & Margolis, L. (1999). CXCR4 utilization is sufficient to trigger CD4+ T cell depletion in HIV-1-infected human

lymphoid tissue. *Proceedings of the National Academy of Sciences*, *96*(January), 663–668.

Philpott, S., Burger, H., Tsoukas, C., Foley, B., Anastos, K., Kitchen, C. & Weiser, B., (2005). Human Immunodeficiency Virus Type 1 Genomic RNA Sequences in the Female Genital Tract and Blood: Compartmentalization and Intrapatient Recombination. *Journal of Virology, 79*(1). 353-363. doi: 10.1128/JVI.79.1.

Pillai, S.K., Pond, S.L.K., Liu, Y., Good, B.M., Strain, M.C., Ellis, R.J., Letendre, S., Smith, D.M., et al., (2006). Genetic attributes of cerebrospinal fluid-derived HIV-1 env. *Brain, 129*. 1872-1883. doi: 10.1093/brain/awl136.

Plantier, J.-C., Leoz, M., Dickerson, J. E., De Oliveira, F., Cordonnier, F., Lemee, V., et al. (2009). A new human immunodeficiency virus derived from gorillas. *Nature Medicine, 15*(8), 871–872.

Pond, S. L. K., & Frost, S. D. W. (2005). Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics (Oxford, England), 21*(10), 2531–3. doi:10.1093/bioinformatics/bti320.

Pond, S. L. K., Frost, S. D. W., & Muse, S. V. (2005). HyPhy: hypothesis testing using phylogenies. *Bioinformatics, 21*(5), 676–9. doi:10.1093/bioinformatics/bti079.

Poon, A. F. Y., Frost, S. D. W., & Pond, S. L. K. (2009). Detecting signatures of selection from DNA sequences using Datamonkey. *Methods in Molecular Biology, 537*, 163–183.

Poon, A. F. Y., Lewis, F. I., Pond, S. L. K., & Frost, S. D. W. (2007). An evolutionary-network model reveals stratified interactions in the V3 loop of the HIV-1 envelope. *PLoS Computational Biology, 3*(11), e231. doi:10.1371/journal.pcbi.0030231

Posada, D., & Buckley, T. R. (2004). Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Systematic Biology*, *53*(5), 793–808. doi:10.1080/10635150490522304

Posada, D. & Crandall, K.A. (1998). Modeltest: testing the model of DNA substitution. *Bioinformatics, 14*(9). 817-81.

Posada, D., & Crandall, K. A. (2002). The effect of recombination on the accuracy of phylogenetic estimation. *Journal of Molecular Evolution, 54*(3), 396–402.

Poss, M., Rodrigo, a G., Gosink, J. J., Learn, G. H., de Vange Panteleeff, D., Martin, H. L., Bwayo, J., et al. (1998). Evolution of envelope sequences from the genital tract and peripheral blood of women infected with clade A human immunodeficiency virus type 1. *Journal of Virology, 72*(10), 8240–51.

Potter, S. J., Lemey, P., Achaz, G., Chew, C. B., Vandamme, A-M., Dwyer, D. E., & Saksena, N. K. (2004). HIV-1 compartmentalization in diverse leukocyte populations during antiretroviral therapy. *Journal of Leukocyte Biology*, *76*(September), 562–570. doi:10.1189/jlb.0404234.http

Potter, S. J., Lemey, P., Dyer, W. B., Sullivan, J. S., Chew, C. B., Vandamme, A-M., et al. (2006). Genetic analyses reveal structured HIV-1 populations in serially sampled T lymphocytes of patients receiving HAART. *Virology*, *348*, 35–46. doi:10.1016/j.virol.2005.12.031

Presti, R. M. L. (2010). Geological vs. Climatological Diversification in the Mediterranean Area: Micro- and Macroevolutionary Approaches in Anthemis L. (Compositae, Anthemideae. *Logos Verlag Berlin GmbH*.

Rambaut A (2002). Se-Al: Sequence alignment editor, Version 2.0a11, Available at: http://tree.bio.ed.ac.uk/software/seal/.

Rambaut, A., Posada, D., Crandall, K. A., & Holmes, E. C. (2004). The Causes and Consequences of HIV Evolution. Nature Reviews. *Genetics, 5*(January), 52–61. doi:10.1038/nrg1246

Ramirez, S., Perez-Del-Pulgar, S., Carrion, J. A, Costa, J., Gonzalez, P., Massaguer, A., et al. (2009). Hepatitis C virus compartmentalization and infection recurrence after liver transplantation. *American Journal of Transplantation*, *9*(7), 1591–601. doi:10.1111/j.1600-6143.2009.02666.x

Rao, R.S.P. & Bernd, W. (2010). Do N-glycoproteins have preference for specific sequons. *Bioinformatician, 5*(5). 208-212.

Reddy, K., Winkler, C. A, Werner, L., Mlisana, K., Abdool Karim, S. S., & Ndung'u, T. (2010). APOBEC3G expression is dysregulated in primary HIV-1 infection and polymorphic variants influence CD4+ T-cell counts and plasma viral load. *AIDS, 24*(2), 195–204. doi:10.1097/QAD.0b013e3283353bba.

Requejo, H. I. Z. (2006). Worldwide molecular epidemiology of HIV. *Revista de Saúde Pública, 40*(2), 331–345.

Ritola, K., Robertson, K., Fiscus, S. A, Hall, C., & Swanstrom, R. (2005). Increased Human Immunodeficiency Virus Type 1 (HIV-1) env Compartmentalization in the Presence of HIV-1-Associated Dementia. *Journal of Virology*, *79*(16), 10830–4. doi:10.1128/JVI.79.16.10830-10834.2005

Robert, C. P., & Casella, G. (2004). Monte Carlo statistical methods (Second edition). New York: *Springer*. ISBN 0-387-21239-6

Rose, P.P. & Korber, B.T. (2000). Detecting hypermutations in viral sequences with an emphasis on G → A hypermutation. *Bioinformatics, 16*(4), pp.400–401.

Rossenkhan, R., Novitsky, V., Sebunya, T. K., Musonda, R., Gashe, B. A., & Essex, M. (2012). Viral Diversity and Diversification of Major Non-Structural Genes vif, vpr, vpu, tat exon 1 and rev exon 1 during Primary HIV-1 Subtype C Infection. *PloS One*, *7*(5), e35491. doi:10.1371/journal.pone.0035491

Rothenberg, R., Baldwin, J., Trotter, R., & Muth, S. (2001). The risk environment for HIV transmission: results from the Atlanta and Flagstaff network studies. *Journal of Urban Health: Bulletin of the New York Academy of Medicine, 78*(3), 419–32. doi:10.1093/jurban/78.3.419.

Rousseau, C. M., Learn, G. H., Bhattacharya, T., Nickle, D. C., Heckerman, D., Chetty, S., et al. (2007). Extensive intrasubtype recombination in South African human immunodeficiency virus type 1 subtype C infections. *Journal of Virology, 81*(9), 4492–500. doi:10.1128/JVI.02050-06

Ryan KJ, Ray CG (editors) (2004). Sherris Medical Microbiology (4th ed.). McGraw Hill. p. 881. ISBN 0-8385-8529-9.

SADC. (2013). Comprehensive Review of Biomedical, Behavioural, and Structural Determinants of HIV Risk and Protective Factors and Best Practices in HIV Prevention.

Sáez-Cirión, A., Bacchus, C., Hocqueloux, L., Avettand-Fenoel, V., Girault, I., Lecuroux, C., et al. (2013). Post-treatment HIV-1 controllers with a long-term virological remission after the interruption of early initiated antiretroviral therapy ANRS VISCONTI Study. *PLoS Pathogens*, *9*(3), e1003211. doi:10.1371/journal.ppat.1003211

Sagar, M., Lavreys, L., Baeten, J. M., Barbra, A., Mandaliya, K., Chohan, B. H., et al. (2003). Infection with Multiple Human Immunodeficiency Virus Type 1 Variants Is Associated with Faster Disease Progression. *Journal of Virology*, *77*(23), 12921–12926. doi:10.1128/JVI.77.23.12921

Sagar, M., Wu, X., Lee, S., & Overbaugh, J. (2006). Human Immunodeficiency Virus Type 1 V1-V2 Envelope Loop Sequences Expand and Add Glycosylation Sites over the Course of Infection, and These Modifications Affect Antibody Neutralization Sensitivity. *Journal of Virology, 80*(19), 9586–9598. doi:10.1128/JVI.00141-06

Salazar-Gonzalez, J.F., Bailes, E., Pham K.T., Salazar, M.G., Guffey, M.B., Keele, B.F. et al. (2008). Deciphering Human Immunodeficiency Virus Type 1 Transmission and Early Envelope Diversification by Single-Genome Amplification and Sequencing. *Journal of Virology, 82(*8), pp.3952–70.

Salazar-Gonzalez, J.F., Salazar, M.G., Learn, G.H., Fouda, G.G., Kang, H.H., et al. (2011). Origin and Evolution of HIV-1 in Breast Milk Determined by Single-Genome Amplification and Sequencing. *Journal of Virology, 85*(6). 2751-2763. doi: 10.1128/JVI.02316-10.

Salemi, M., Lamers, S. L., Huysentruyt, L. C., Galligan, D., Gray, R. R., & McGrath, M. S. (2009). Distinct Patterns of HIV-1 Evolution within Metastatic Tissues in Patients with Non-Hodgkins Lymphoma. *PloS one, 4*(12). doi:10.1371/journal.pone.0008153.

Salemi, M., Lamers, S. L., Yu, S., de Oliveira, T., Fitch, W. M., & McGrath, M. S. (2005). Phylodynamic Analysis of Human Immunodeficiency Virus Type 1 in Distinct Brain Compartments Provides a Model for the Neuropathogenesis of AIDS. *Journal of Virology, 79*(17), 11343–11352. doi:10.1128/JVI.79.17.11343

Santa-Marta, M., de Brito, P. M., Godinho-Santos, A., & Goncalves, J. (2013). Host factors and HIV-1 replication: clinical evidence and potential therapeutic approaches. *Frontiers in Immunology, 4*(October), 343. doi:10.3389/fimmu.2013.00343.

Santiago, M. L., Range, F., Keele, B. F., Li, Y., Bailes, E., Bibollet-Ruche, F. et al. (2005). Simian Immunodeficiency Virus Infection in Free-Ranging Sooty Mangabeys (Cercocebus atys atys) from the Taï Forest, Côte d'Ivoire: Implications for the Origin of Epidemic Human Immunodeficiency Virus Type 2. *Journal of Virology, 79*(19), 12515–12527. doi:10.1128/JVI.79.19.12515.

Santoro, M. M., & Perno, C. F. (2013). HIV-1 Genetic Variability and Clinical Implications. *ISRN Microbiology*, *2013*.

Sato, H., Orensteint, J., Dimitrov, D., & Martin, M. (1992). Cell-to-cell spread of HIV-1 occurs within minutes and may not involve the participation of virus particles. *Virology*, *186*(2), 712–724. doi:10.1016/0042-6822(92)90038-Q

Sattentau, Q. (2008). Avoiding the void: cell-to-cell spread of human viruses. *Nature Reviews Microbiology*, *6*(11), 815–826. Retrieved from http://dx.doi.org/10.1038/nrmicro1972

Scanlan, C. N., Pantophlet, R., Wormald, M. R., Saphire, E. O., Stanfield, R., Wilson, I. A., et al. (2002). The Broadly Neutralizing Anti-Human Immunodeficiency Virus Type 1 Antibody 2G12 Recognizes a Cluster of α 1 → 2 Mannose Residues on the Outer Face of gp120. *Journal of Virology, 76*(14), 7306–7321. doi:10.1128/JVI.76.14.7306

Schnell, G., Price, R. W., Swanstrom, R., & Spudich, S. (2010). Compartmentalization and Clonal Amplification of HIV-1 Variants in the Cerebrospinal Fluid during Primary Infection. *Journal of Virology, 84*(5), 2395–2407. doi:10.1128/JVI.01863-09.

Schuitemaker, H., van't Wout, A. B., & Lusso, P. (2011). Clinical significance of HIV-1 coreceptor usage. *Journal of Translational Medicine, 9*(1), doi:10.1186/1479-5876-9-S1-S5.

Seiter, J., Fass, M., Stanley, E., & Waterman, M. (2011). HIV/AIDS: Biology and Treatment. *Biology International, 49*, 86–95.

Shaeffer, S. (1994). The Impact of HIV/AIDS on Education (pp. 1–34).

Shan, L., Deng, K., Shroff, N. S., Durand, C. M., Rabi, S. A., Yang, H-C., et al. (2012). Stimulation of HIV-1-specific cytolytic T lymphocytes facilitates elimination of latent viral reservoir after virus reactivation. *Immunity, 36*(March 23), 491–501. doi:10.1016/j.immuni.2012.01.014

Shankarappa, R., Margolick, J. B., Gange, S. J., Rodrigo, A. G., Upchurch, D., Farzadegan, H., et al. (1999). Consistent Viral Evolutionary Changes Associated with the Progression of Human Immunodeficiency Virus Type 1 Infection. *Journal of Virology, 73*(12), 10489–502.

Sharp, P. M., & Hahn, B. H. (2011). Origins of HIV and the AIDS Pandemic. *Cold Spring Harbor Perspectives in Medicine*, *1*(a006841), 1–22. doi:10.1101/cshperspect.a006841

Shaw, T. I., & Zhang, M. (2013). HIV N-linked glycosylation site analyzer and its further usage in anchored alignment. *Nucleic Acids Research, 41*(Web Server), W454–8. doi:10.1093/nar/gkt472

Shepherd, J. C., Jacobson, L. P., Qiao, W., Jamieson, B. D., Phair, J. P., Piazza, P., et al. (2008). Emergence and persistence of CXCR4-tropic HIV-1 in a population of men from the multicenter AIDS cohort study. *The Journal of Infectious Diseases, 198*(15 October), 1104–12. doi:10.1086/591623.

Sheth, P. M., Sunderji, S., Shin, L. Y. Y., Rebbapragada, A., Huibner, S., Kimani, J., et al. (2008). Coinfection with Herpes Simplex Virus Type 2 Is Associated with Reduced HIV-Specific T Cell Responses and Systemic Immune Activation. *The Journal of Infectious Diseases*, *197*(15 May), 1394–401. doi:10.1086/587697

Shisana, O., Rehle, T., Simbayi, L. C., Zuma, K., Jooste, S., Zungu, N., et al. (2014). South African National HIV Prevalence, Incidence and Behaviour Survey, 2012. *HSRC Press*.

Sigal, A., Kim, J. T., Balazs, A. B., Dekel, E., Mayo, A., Milo, R., & Baltimore, D. (2011). Cell-to-cell spread of HIV permits ongoing replication despite antiretroviral therapy. *Nature*, *477*(7362), 95–98. Retrieved from http://dx.doi.org/10.1038/nature10347

Si-Mohamed, A., Kazatchkine, M. D., Heard, I., Goujon, C., Prazuck, T., Aymard, G., et al. (2000). Selection of drug-resistant variants in the female genital tract of human immunodeficiency virus type 1-infected women receiving antiretroviral therapy. *The Journal of Infectious Diseases*, *182*(1), 112–22. doi:10.1086/315679

Siliciano, R. F., & Greene, W. C. (2011). HIV latency. *Cold Spring Harbor Perspectives in Medicine*, *1*, a007096. doi:10.1101/cshperspect.a007096

Simon, B., Grabmeier-Pfistershammer, K., Rieger, A., Sarcletti, M., Schmied, B., & Puchhammer-Stöckl, E. (2010). HIV coreceptor tropism in antiretroviral treatment-naive patients newly diagnosed at a late stage of HIV infection. *AIDS*, *24*(13), 2051–8. doi:10.1097/QAD.0b013e32833c93e6

Simon-Loriere, E., Martin, D. P., Weeks, K. M., & Negroni, M. (2010). RNA Structures Facilitate Recombination-Mediated Gene Swapping in HIV-1. *Journal of Virology, 84*(24), 12675–82. doi:10.1128/JVI.01302-10

Slatkin, M. (1993). Isolation by Distance in Equilibrium and Non-Equilibrium Populations. *Evolution*, *47*(1), 264–279.

Slatkin, M. & Maddison, W., P. (1989). A cladistic measure of gene flow inferred from the phylogenies of alleles. *Genetics, 123*, 603–613.

Sourisseau, M., Sol-Foulon, N., Porrot, F., Blanchet, F., & Schwartz, O. (2007). Inefficient human immunodeficiency virus replication in mobile lymphocytes. *Journal of Virology*, *81*(2), 1000–12. doi:10.1128/JVI.01629-06

Smith, D. M., Zárate, S., Shao, H., Pillai, S. K., Letendre, S., Wong, J. K., Richman, D. D., et al. (2010). Pleocytosis Is Associated with Disruption of HIV Compartmentalization between Blood and Cerebral Spinal Fluid Viral Populations. *Virology, 385*(1), 204–208. doi:10.1016/j.virol.2008.11.010.Pleocytosis.

Speck, R. F., Wehrly, K., Platt, E. J., Atchison, R. E., Charo, I. F., Kabat, D., et al. (1997). Selective employment of chemokine receptors as human immunodeficiency virus type 1 coreceptors determined by individual amino acids within the envelope V3 loop. *Journal of Virology, 71*(9), 7136–9.

StatSoft (2013). Electronic Statistics Textbook. Tulsa, OK: StatSoft. Available at: http://www.statsoft.com/textbook/.

Strebel, K. (2005). APOBEC3G & HTLV-1: inhibition without deamination. *Retrovirology, 2*(37). doi:10.1186/1742-4690-2-37.

Streeck, H., Li, B., Poon, A. F. Y., Schneidewind, A., Gladden, A. D., Power, K. A., et al. (2008). Immune-driven recombination and loss of control after HIV superinfection. *The Journal of Experimental Medicine, 205*(8), 1789–96. doi:10.1084/jem.20080281

Stremlau, M., Owens, C. M., Perron, M. J., Kiessling, M., Autissier, P., & Sodroski, J. (2004). The cytoplasmic body component TRIM5alpha restricts HIV-1 infection in Old World monkeys. *Nature, 427*(6977), 848–53. doi:10.1038/nature02343.

Sullivan, S. T., Mandava, U., Evans-Strickfaden, T., Lennox, J. L., Ellerbrock, T. V, & Hart, C. E. (2005). Diversity, Divergence, and Evolution of Cell-Free Human Immunodeficiency Virus Type 1 in Vaginal Secretions and Blood of Chronically Infected Women: Associations with Immune Status. *Journal of Virology*, *79*(15), 9799–9809. doi:10.1128/JVI.79.15.9799

Sturdevant, C. B., Dow, A., Jabara, C. B., Joseph, S. B., Schnell, G., Takamune, N., et al. (2012). Central Nervous System Compartmentalization of HIV-1 Subtype C Variants Early and Late in Infection in Young Children. *PLoS Pathogens*, *8*(12).

doi:10.1371/journal.ppat.1003094

Suchard, M. A., Weiss, R. E., & Sinsheimer, J. S. (2001). Bayesian Selection of Continuous-Time Markov Chain Evolutionary Models. *Molecular Biology and Evolution*, *18*(6), 1001–1013.

Swiecki, M., Wang, Y., Gilfillan, S., Lenschow, D. J., & Colonna, M. (2012). Paradoxical roles of BST2/Tetherin in promoting IFN-I response and viral infection. *Journal of Immunology, 188*(6), 2488–2492. doi:10.1016/j.virol.2010.12.041.HIV

Swofford, D. (1999). PAUP: Phylogenetic Analysis Using Parsimony, Version 4.0b10, Available at: http://paup.csit.fsu.edu/.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., & Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Molecular Biology and Evolution, 30*(12), 2725–9. doi:10.1093/molbev/mst197.

Tawfik, L., & Kinoti, S. N. (2006). The impact of HIV/AIDS on the health workforce in developing countries (pp. 1–28).

Templeton, A. R., Kramer, M. G., Jarvis, J., Kowalski, J., Gange, S., Schneider, M. F., et al. (2009). Multiple-infection and recombination in HIV-1 within a longitudinal cohort of women. *Retrovirology, 6*(54), 1–12. doi:10.1186/1742-4690-6-54

Thali, M., Moore, J. P., Furman, C., Charles, M., Ho, D. D., Robinson, J., & Sodroski, J. (1993). Characterization of Conserved Human Immunodeficiency Virus Type 1 gpl20 Neutralization Epitopes Exposed upon gpl20-CD4 Binding. *Journal of Virology, 67*(7), 3978–3988.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. and Higgins, D. G. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research, 25*, 4876-82.

Tirado, G., Jove, G., Reyes, E., Sepulveda, G., Yamamura, Y., Singh, D. P., & Kumar, A. (2005). Differential evolution of cell-associated virus in blood and genital tract of HIV-infected females undergoing HAART. *Virology*, *334*(2), 299–305. doi:10.1016/j.virol.2005.01.030

Toma, J., Weinheimer, S. P., Stawiski, E., Whitcomb, J. M., Lewis, S. T., Petropoulos, C. J., & Huang, W. (2011). Loss of Asparagine-Linked Glycosylation Sites in Variable

Region 5 of Human Immunodeficiency Virus Type 1 Envelope Is Associated with Resistance to CD4 Antibody Ibalizumab. *Journal of Virology, 85*(8), 3872–3880. doi:10.1128/JVI.02237-10

Tomaras, G. D., & Haynes, B. F. (2009). HIV-1-specific antibody responses during acute and chronic HIV-1 infection. *Current Opinion in HIV/AIDS, 4*(5), 373–379. doi:10.1097/COH.0b013e32832f00c0

Trachtenberg, E., Korber, B., Sollars, C., Kepler, T. B., Hraber, P. T., Hayes, E., et al. (2003). Advantage of rare HLA supertype in HIV disease progression. *Nature Medicine*, *9*(7), 928–935. Retrieved from http://dx.doi.org/10.1038/nm893

Troyer, R. M., Collins, K. R., Abraha, A., Fraundorf, E., Moore, D. M., Krizan, R. W., et al. (2005). Changes in human immunodeficiency virus type 1 fitness and genetic diversity during disease progression. *Journal of Virology*, *79*(14), 9006–18. doi:10.1128/JVI.79.14.9006-9018.2005

Tsuchiya, K., Ode, H., Hayashida, T., Kakizawa, J., Sato, H., Oka, S., & Gatanaga, H. (2013). Arginine insertion and loss of N-linked glycosylation site in HIV-1 envelope V3 region confer CXCR4-tropism. *Scientific Reports, 3*(April 2013), 1–8. doi:10.1038/srep02389

UNAIDS (2010), Global Report. Available at: http://www.unaids.org/globalreport/documents/20101123_GlobalReport_full_en.pdf

UNAIDS. (2012). Global Fact Sheet (pp. 7–10).

UNAIDS. (2013). AIDS by the numbers (pp. 1–12).

UNFPA. (2013). The Role of Data in Addressing Violence against Women and Girls (pp. 1–46).

UNICEF. (2003). What Religious Leaders Can Do About HIV/AIDS: Action for Children and Young People (pp. 1–56).

USAID (2011). HIV/AIDS Health Profile: Sub-Saharan Africa. Available at: http://www.usaid.gov/our_work/global_health/aids/Countries/africa/hiv_summary_africa.pdf

Uvin, S. C., & Caliendo, A. M. (1997). Cervicovaginal human immunodeficiency virus secretion and plasma viral load in human immunodeficiency virus-seropositive women. *Obstetrics and Gynecology*, *90*(5), 739–43. doi:10.1016/S0029-7844(97)00411-0

van den Kerkhof, T. L. G. M., Feenstra, K. A., Euler, Z., van Gils, M. J., Rijsdijk, L. W. E., Boeser-Nunnink, B. D., et al. (2013). HIV-1 envelope glycoprotein signatures that correlate with the development of cross-reactive neutralizing activity. *Retrovirology, 10(*102), 1–19. doi:10.1186/1742-4690-10-102

Van der Kuyl, A. C., & Cornelissen, M. (2007). Identifying HIV-1 dual infections. *Retrovirology*, *4*(67), 1–12. doi:10.1186/1742-4690-4-67

van Deutekom, H. W. M., Wijnker, G., & de Boer, R. J. (2013). The rate of immune escape vanishes when multiple immune responses control an HIV infection. *The Journal of Immunology*, *191*(6), 3277–86. doi:10.4049/jimmunol.1300962

Van Gils, M. J., Bunnik, E. M., Boeser-Nunnink, B. D., Burger, J. A., Terlouw-Klein, M., Verwer, N., & Schuitemaker, H. (2011). Longer V1V2 region with increased number of potential N-linked glycosylation sites in the HIV-1 envelope glycoprotein protects against HIV-specific neutralizing antibodies. *Journal of Virology*, *85*(14), 6986–95. doi:10.1128/JVI.00268-11

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

van Marle, G., Gill, M. J., Kolodka, D., McManus, L., Grant, T., & Church, D. L. (2007). Compartmentalization of the gut viral reservoir in HIV-1 infected patients. *Retrovirology*, *4*(87), 1–14. doi:10.1186/1742-4690-4-87

Varsani, A., van der Walt, E., Heath, L., Rybicki, E. P., Williamson, A. L. & Martin, D. P. (2006). Evidence of ancient papillomavirus recombination. *Journal of General Virology*. *87*, 2527-2531.

Visser, M. (2004). The Impact of HIV/AIDS in Africa Changing Strategies in Addressing HIV/AIDS.

Wald, A. (2004). Herpes simplex virus type 2 transmission: risk factors and virus shedding. *Herpes, 11*(Suppl 3): 130A-7A.

Wang, T. H., Donaldson, Y. K., Brettle, R. P., Bell, J. E., & Simmonds, P. (2001). Identification of Shared Populations of Human Immunodeficiency Virus Type 1 Infecting Microglia and Tissue Macrophages outside the Central Nervous System. *Journal of Virology*, *75*(23), 11686–11699. doi:10.1128/JVI.75.23.11686

Wang, W., Nie, J., Prochnow, C., Truong, C., Jia, Z., Wang, S., et al. (2013). A systematic study of the N-glycosylation sites of HIV-1 envelope protein on infectivity and antibody-mediated neutralization. *Retrovirology, 10*(14), 1–14. doi:10.1186/1742-4690-10-14

Waters, L. J., Scourfield, A. T., Marcano, M., Gazzard, B. G., Bower, M., Nelson, M., & Stebbing, J. (2011). The Evolution of Coreceptor Tropism in HIV-infected Patients Interrupting Suppressive Antiretroviral Therapy. *Clinical Infectious Diseases, 52*(5), 671–3. doi:10.1093/cid/ciq198.

Weir, B. S., & Cockerham, C. C. (1984) Estimating F-statistics for the analysis of population structure. *Evolution*, *38*, 1358 –1370.

Weiss, R. E., Wang, Y., & Ibrahim J. G. (1997). Predictive model selection for repeated measures random effects models using Bayes factors. *Biometrics, 53*(2), 159–169.

Wilson, G. A., & Rannala, B. (2003). Bayesian Inference of Recent Migration Rates Using Multilocus Genotypes. *Genetics Society of America, 163*(March), 1177–1191.

Wlasiuk, G., & Nachman, M. W. (2010). Adaptation and constraint at Toll-like receptors in primates. *Molecular Biology and Evolution, 27*(9), 2172–86. doi:10.1093/molbev/msq104.

Wolinsky, S. M., Wike, C. M., Korber, B. T., Hutto, C., Parks, W. P., Rosenblum, L. L., et al. (1992). Selective transmission of human immunodeficiency virus type-1 variants from mothers to infants. *Science* , *255* (5048), 1134–1137. doi:10.1126/science.1546316

Wong, J. K., Ignacio, C. C., Torriani, F., Havlir, D., Fitch, N. J., & Richman, D. D. (1997). In Vivo Compartmentalization of Human Immunodeficiency Virus: Evidence from the Examination of pol Sequences from Autopsy Tissues. *Journal of Virology*, *71*(3), 2059–71.

World Health Organization. (2009). Women and Health: Today's Evidence, Tomorrow's Agenda (pp. 1–897).

Worobey, M., Santiago, M. L., Keele, B. F., Ndjango, J-B. N., Joy, J. B., Labama, B. L., et al. (2004). Origin of AIDS: Contaminated polio vaccine theory refuted. *Nature*, *428*(6985), 820. Retrieved from http://dx.doi.org/10.1038/428820a

Wright, S. (1943). Isolation by distance. *Genetics*, *28*(2), 114–138.

Xiang, S-H., Doka, N., Choudhary, R. K., Sodroski, J., & Robinson, J. E. (2002). Characterization of CD4-Induced Epitopes on the HIV Type 1 gp120 Envelope Glycoprotein Recognized by Neutralizing Human Monoclonal Antibodies. *AIDS Research and Human Retroviruses, 18*(16), 1207–1217.

Yildirim, I. (2012). Bayesian Inference: Metropolis-Hastings Sampling (pp. 1–6).

Yokoyama, M., Naganawa, S., Yoshimura, K., Matsushita, S., & Sato, H. (2012). Structural dynamics of HIV-1 envelope Gp120 outer domain with V3 loop. *PloS One, 7*(5), e37530. doi:10.1371/journal.pone.0037530

Yu, Q., Landau, N. R., & König, R. (2003). Vif and the Role of Antiviral Cytidine Deaminases in HIV-1 Replication. HIV Sequence Compendium, *Reviews*. 1–13.

Yuan, T., Li, J., & Zhang, M-Y. (2013). HIV-1 envelope glycoprotein variable loops are indispensable for envelope structural integrity and virus entry. *PloS One, 8*(8), e69789. doi:10.1371/journal.pone.0069789.

Zárate, S., Pond, S. L. K., Shapshak, P., & Frost, S. D. W. (2007). Comparative Study of Methods for Detecting Sequence Compartmentalization in Human Immunodeficiency Virus Type 1. *Journal of Virology, 81*(12), 6643–6651. doi:10.1128/JVI.02268-06.

Zhang, M., Gaschen, B., Blay, W., Foley, B., Haigwood, N., Kuiken, C., & Korber, B. (2004). Tracking global patterns of N-linked glycosylation site variation in highly variable viral glycoproteins: HIV, SIV, and HCV envelopes and influenza hemagglutinin. *Glycobiology*, *14*(12), 1229–46. doi:10.1093/glycob/cwh106.

Zhu, T., Wang, N., Carr, A., Nam, D. S., Moor-Jankowski, R., Cooper, D. A., & Ho, D. D. (1996). Genetic Characterization of Human Immunodeficiency Virus Type 1 in Blood and Genital Secretions: Evidence for Viral Compartmentalization and Selection during Sexual Transmission. *Journal of Virology*, *70*(5), 3098–3107.

Zhuang, J., Jetzt, A. E., Sun, G., Yu, H., Klarmann, G., Ron, Y., et al. (2002). Human Immunodeficiency Virus Type 1 Recombination: Rate, Fidelity, and Putative Hot Spots. *Journal of Virology*, *76*(22), 11273–11282. doi:10.1128/JVI.76.22.11273

Appendices

Appendix I – Python script to predict PNLG sites

```
#! /usr/bin/env python
import sys
import re
f=open("CAP177-string.txt")
HXB2="MRVKEKYQHLWRWGWRWGTMLLGMLMICSATEKLWVTVYYGVPVWKEATTTLFCASDAK
AYDTEVHNVWATHACVPTDPNPQEVVLVNVTENFNMWKNDMVEQMHEDIISLWDQSLKPCVKLTPL
CVSLKCTDLK--NDTNTNSS----------SGRMIMEKGEIKNCSFNISTSIRGKVQKEYAFFYKLDIIPIDND------
TTSYKLTSCNTSVITQACPKVSFEPIPIHYCAPAGFAILKCNNKTFNGTGPCTNVSTVQCTHGIRPVVSTQLL
LNGSLAEEEVVIRSVNFTDNAKTIIVQLNTSVEINCTRPNNNTRKRIRIQRGPGRAFVTIG-
KIGNMRQAHCNISRAKWNNTLKQIASKLREQFGNNKTIIFKQSSGGDPEIVTHSFNCGGEFFYCNSTQLF
NSTWFNSTWSTEGSNNTEGSDTITLPCRIKQIINMWQKVGKAMYAPPISGQIRCSSNITGLLLTRDGGN--
---SNNESE----
IFRPGGGDMRDNWRSELYKYKVVKIEPLGVAPTKAKRRVVQREKRAVGIGALFLGFLGAAGSTMGAAS
MTLTVQARQLLSGIVQQQNNLLRAIEAQQHLLQLTVWGIKQLQARILAVERYLKDQQLLGIWGCSGKLIC
TTAVPWNASWSNKSLEQIWNHTTWMEWDREINNYTSLIHSLIEESQNQQEKNEQELLELDKWASLWN
WFNITNWLWYIKLFIMIVGGLVGLRIVFAVLSIVNRVRQGYSPLSFQTHLPTPRGPDRPEGIEEEGGERDR
DRSIRLVNGSLALIWDDLRSLCLFSYHRLRDLLLIVTRIVELLGR-------
RGWEALKYWWNLLQYWSQELKNSAVSLLNATAIAVAEGTDRVIEVVQGACRAIRHIPRRIRQGLERILL--
---------*"
temp2=[]
def findtriplet(seqsnip):
    triplet=""
    for i in seqsnip:
        if i != "-" and len(triplet)!=4:
            triplet+=i
    return triplet
    triplet=""
def sequon(triplet):
    temp=[]
    for i in triplet:
        temp.append(i)
    if (temp[2]=="S") and (temp[1]!="P") and (temp[1]!="W") and (temp[1]!="D") and
(temp[1]!="E") and (temp[1]!="L") and (temp[3]!="P"):
        return "yes"
    elif (temp[2]=="T") and (temp[1]!="P") and (temp[3]!="P"):
        return "yes"
    else:
        return "no"
    temp=[]
def indexing(allpos):
    hxb2pos=""
```

```
    for i in allpos:
        count = HXB2[:i].count("-")
        hxb2pos+=","+str((int(i)-count))
        #print "alignment pos: ",i," HXB2 pos: ",int(i)-count
    return hxb2pos
for line in f:
    if ">" in line:
        temp2.append(line[:-1])
    else:
        sequence=line[:-1]
        for index in range(len(sequence)):
            if(sequence[index]=='N'):
seqsnip=sequence[index]+sequence[index+1]+sequence[index+2]+sequence[index+3]+sequ
ence[index+4]+sequence[index+5]+sequence[index+6]+sequence[index+7]+sequence[index+
8]+sequence[index+9]+sequence[index+10]+sequence[index+11]+sequence[index+12]+sequ
ence[index+13]+sequence[index+14]+sequence[index+15]+sequence[index+16]+sequence[i
ndex+17]+sequence[index+18]+sequence[index+19]+sequence[index+20]+sequence[index+
21]+sequence[index+22]+sequence[index+23]+sequence[index+24]+sequence[index+25]
                triplet=findtriplet(seqsnip)
                if sequon(triplet) == "yes":
                    temp2.append(index+1)
allpos=[]
for i in temp2:
    if (">" not in str(i)) and (i not in allpos):
        allpos.append(i)
allpos.sort()
tmp4=""
tmp3="Alignment pos:,"
tmp4+="HXB2 pos:"+indexing(allpos)
for i in allpos:
    tmp3+=str(i)+","
print tmp3+",Total"
print tmp4
all=""
for i in temp2:
    if ">" in str(i):
        all+="|"+str(i)
    else:
        all+=","+str(i)
def check(tmp):
    tmp2=""
    count=0
    for i in allpos:
        if str(i) in tmp:
            tmp2+=",N"
            count+=1
        else:
```

```
            tmp2+=", "
    print tmp[0],tmp2+",",", ",count
for i in all.split("|"):
    tmp=i.split(",")
    check(tmp)
f.close()
```

Appendix II – Python script to count V-loop aa base length

```
#!/usr/bin/python
name=''

def count(line):
    c=0
    for i in line:
        if i != "" and i != "-":
            c+=1
    return(c-1)

f=open('V1V2.txt')
for line in f:
    if ">" in line:
        name=line[:-1]
    else:
        print name,"=",count(line)

f.close()
```

Appendix III – Supplementary Figures

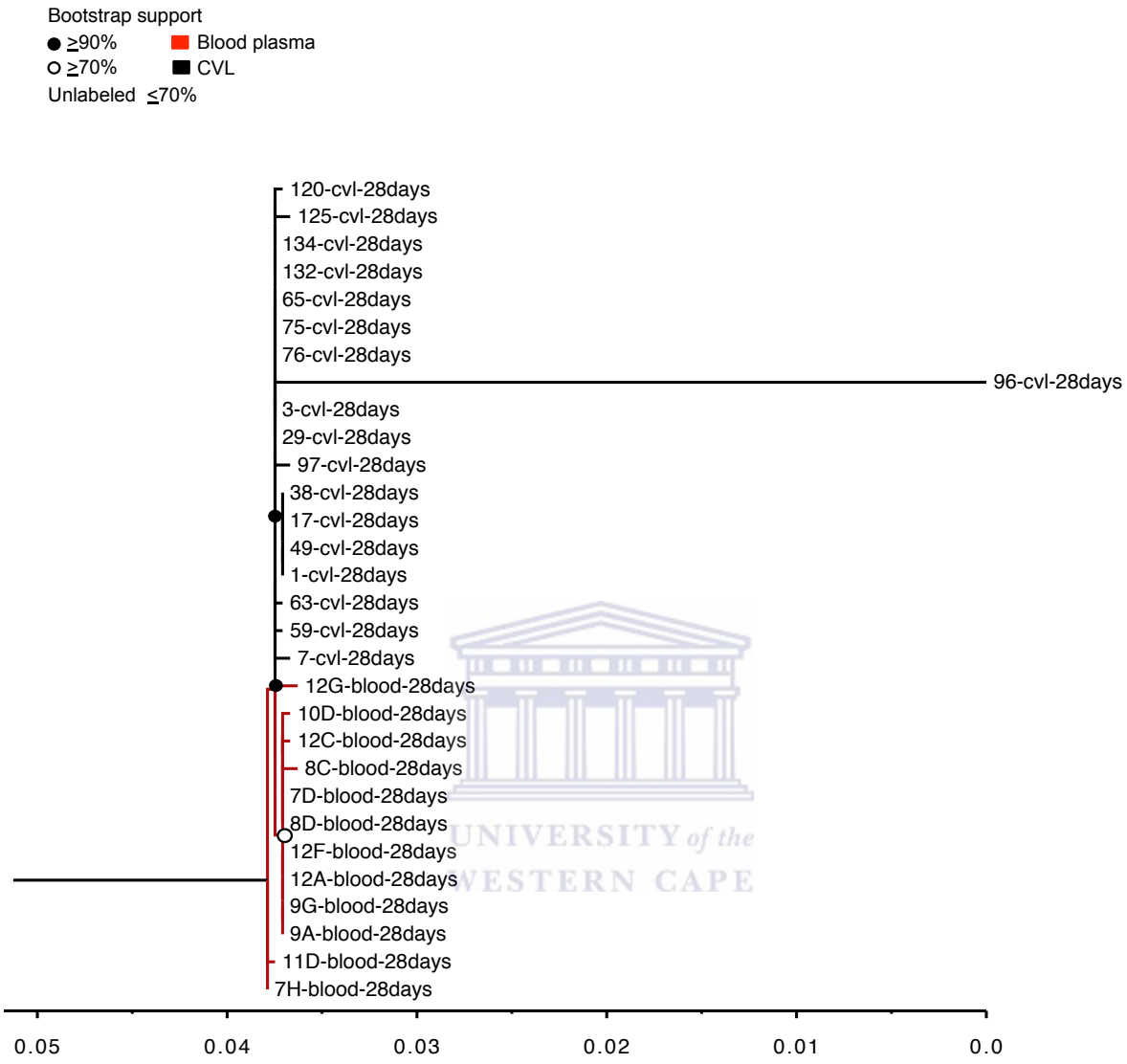


Figure 3.4a Cross-sectional maximum likelihood tree for HIV sequences from CAP177 at 28 days post-infection constructed under GTR + G substitution model using PHYML. Branches are coloured by tissue type where red = blood plasma and black = CVL, with bootstrap support greater than 90% indicated by a filled circle and greater than 70% by an open circle at the nodes.

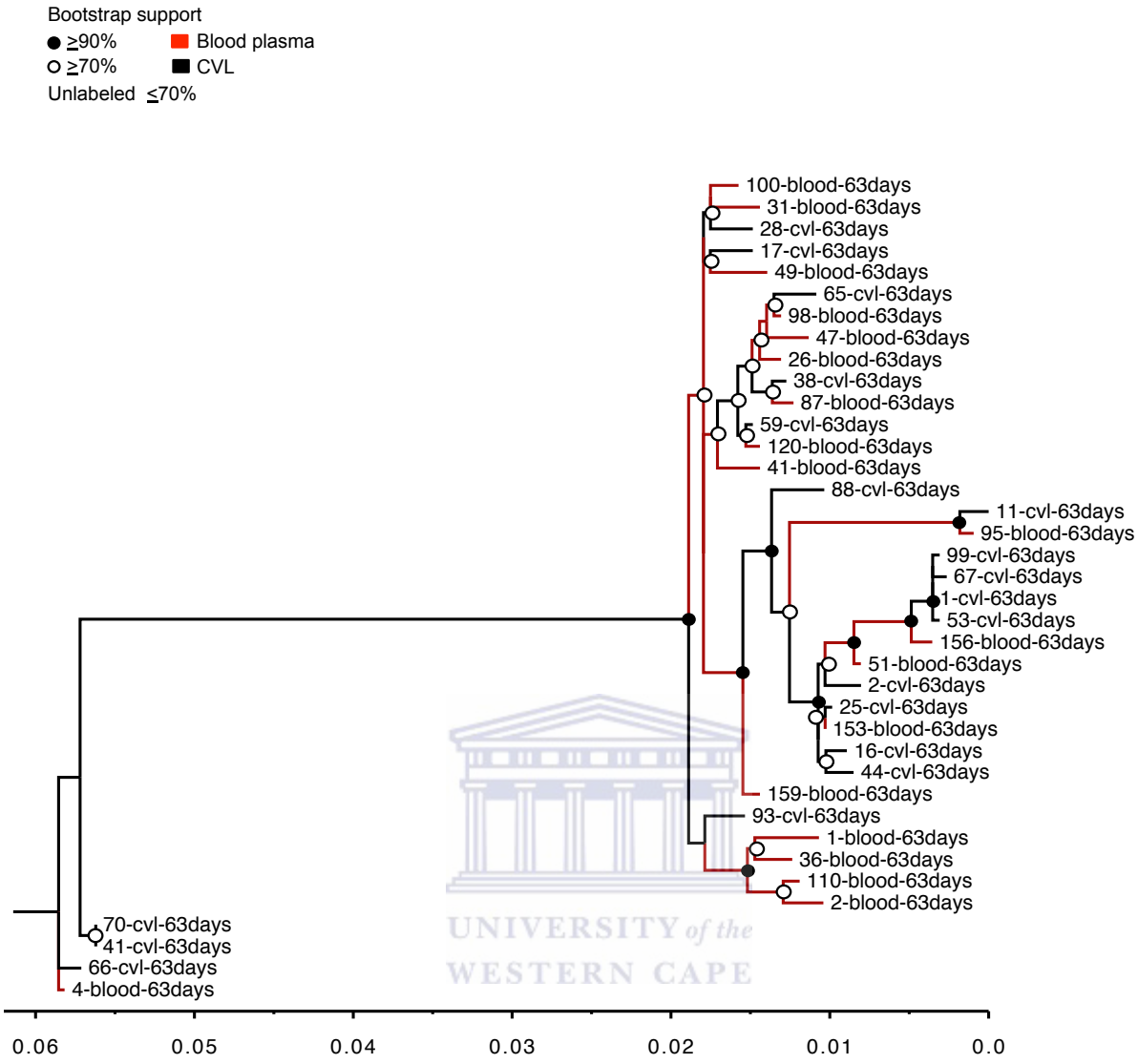


Figure 3.4b Cross-sectional maximum likelihood tree for HIV sequences from CAP261 at 63 days post-infection constructed under GTR + G substitution model using PHYML. Branches are coloured by compartment where red = blood plasma and black = CVL, with bootstrap support greater than 90% indicated by a filled circle and greater than 70% by an open circle at the nodes.

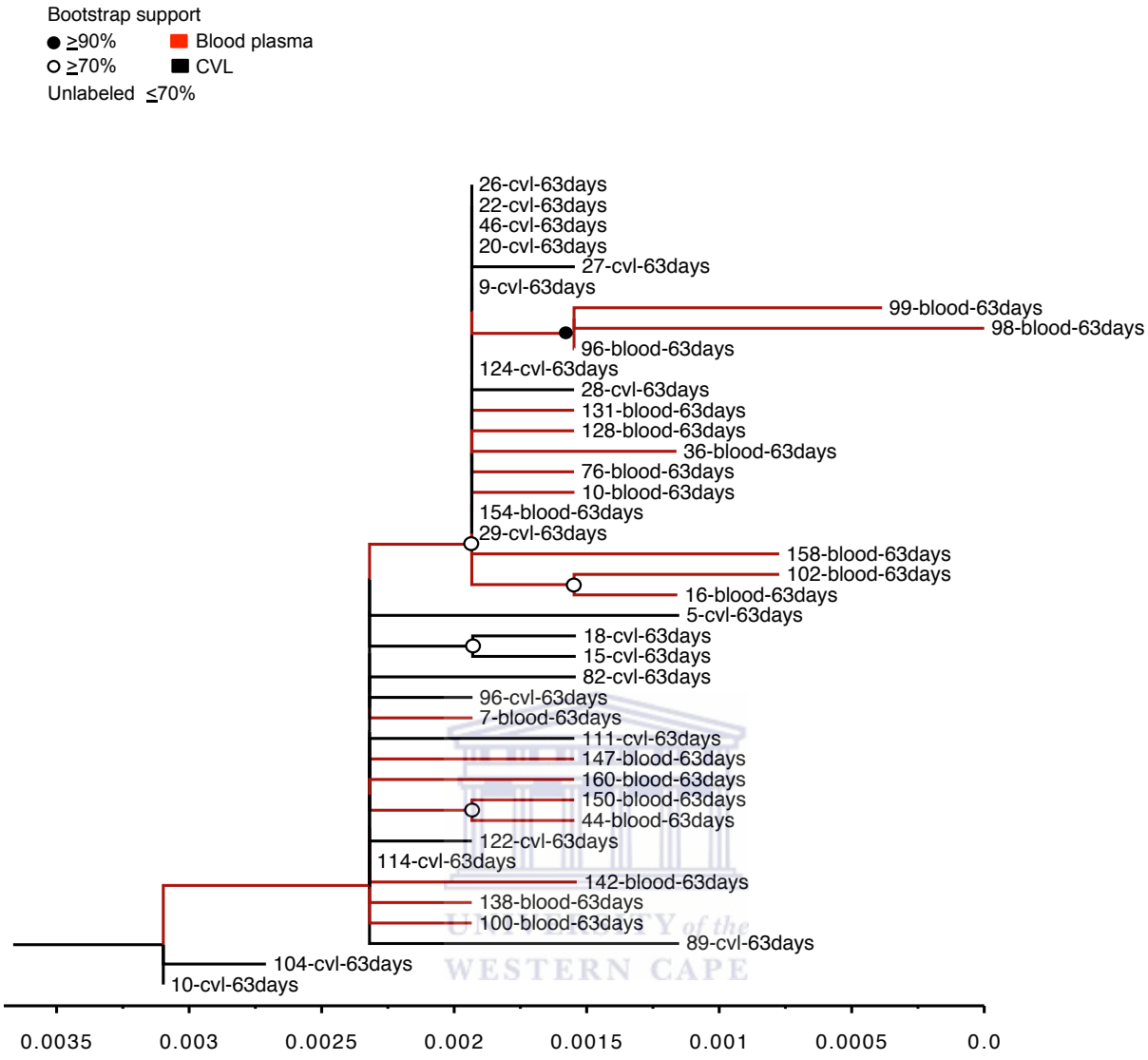


Figure 3.4c Cross-sectional maximum likelihood tree for HIV sequences from CAP217 at 63 days post-infection constructed under GTR + G substitution model using PHYML. Branches are coloured by compartment where red = blood plasma and black = CVL, with bootstrap support greater than 90% indicated by a filled circle and greater than 70% by an open circle at the nodes.

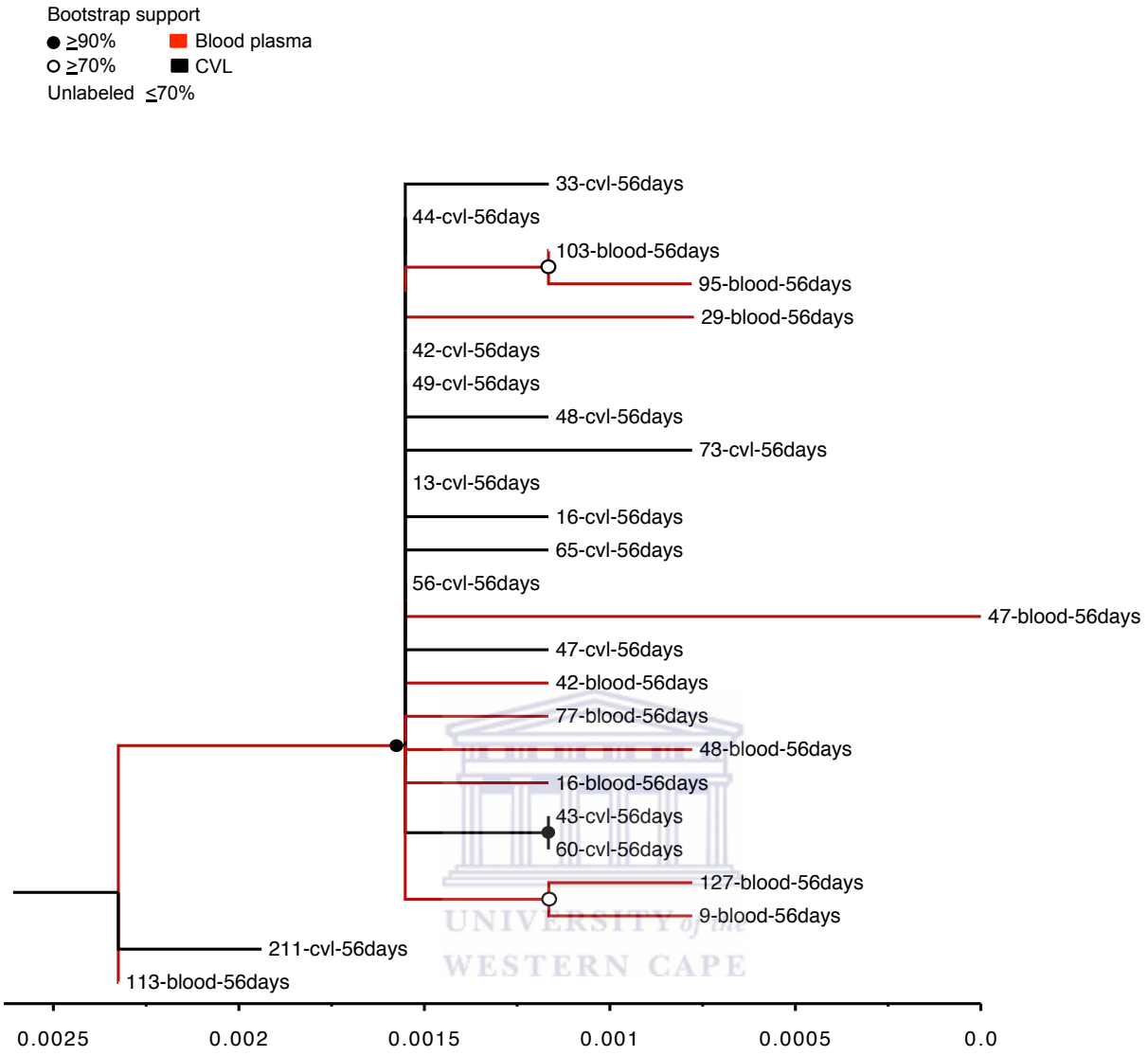


Figure 3.4d Cross-sectional maximum likelihood tree for HIV alignments from CAP270 at 56 days post-infection under GTR + G nucleotide substitution model predicted using PHYML. Branches are coloured by compartment where red = blood plasma and black = CVL, with bootstrap support greater than 90% indicated by a filled circle and greater than 70% by an open circle at the nodes