Evolutionary analyses of HIV-1 protein-coding sequences: adaptation to host cytotoxic immune responses and purifying selection at synonymous sites

By

Nobubelo Kwanele Ngandu

Thesis Presented for the Degree of
DOCTOR OF PHILOSOPHY
in the Department of Molecular and Cell Biology (Bioinformatics)
UNIVERSITY OF CAPE TOWN

June, 2009

Abstract

The detailed understanding of human immunodeficiency virus type 1 (HIV-1) pathogenesis can yield insights relevant to the design of more effective therapeutics against the virus in infected individuals. In this work, two aspects of HIV evolution relevant to pathogenesis were evaluated using Bioinformatics approaches. Firstly, the interaction of the virus with the cytotoxic T-lymphocyte immune response mediated by human leukocyte antigens. The cytotoxic immune response is one of the major forces exerting selection pressure on the HIV-1 protein sequence and a key predictor of disease progression. Here, (i) the relationship between the cytotoxic immune response and sequence variation, (ii) virus immune escape patterns across major infected populations and (iii) adaptive evolution of the virus following its transmission from chimpanzee to humans were analysed. The virus sequence and immune response data were obtained from public databases and collaborating laboratories. Phylogeny-based codon models and entropy were used to estimate selection pressure and sequence variability per site. Sequences of the immunogenic Nef and Gag proteins from HIV-1 subtype C, obtained from a Southern African cohort showed frequent immunogenicity in conserved regions. However, in comparison to an equivalent subtype B dataset, the human leukocyte antigens from this cohort were poorly characterized in public databases and their sequence binding preferences were poorly predicted by publicly available sequence motifs. Several previous studies have shown that HIV-1 protein sequences adapt to cytotoxic immune responses mediated by human leukocyte alleles that are common in host populations. Here, the relationship between HLA allele frequency and viral evolution was assessed using data from major infected populations. Results suggest that immune responses not only remove potential epitopes in unconstrained regions but also drive the evolution of functionally constrained regions of the virus. Evidence of adaptation of HIV-1 to two human leukocyte antigen alleles following cross-species transmission from chimpanzee was also observed.

Secondly, the extent of purifying selection pressure acting on synonymous sites of the virus nucleotide sequence was investigated. This has not been previously done despite the virus nucleotide sequence containing many functional motifs that regulate viral

lifecycle and hence are expected to be under pressure to remain conserved. Viral reference sequences from the major subtypes of HIV-1 were obtained from the Los Alamos HIV databases. Using phylogeny-based codon models to obtain the rate of synonymous substitutions, synonymous sites within twenty-three known functional and novel motifs were found to be under purifying selection pressure. This thesis exposes previously ignored evolutionary characteristics of the synonymous sites of virus nucleotide sequences and contributes new findings to the understanding of the evolution of HIV-1 in relation to the human immune response.

Acknowledgements

My appreciation goes to the Stanford-South Africa Biomedical Informatics Training Program, which is supported by the Fogarty International Center, part of the National Institutes of Health (grant no. 5D43 TW006993) and the South Africa National Bioinfomatics Network for providing grants to support this research work.

I extend my utmost and sincere gratitude to Prof Cathal Seoighe for his supervision, guidance and teaching over the course of my PhD studies. His immeasurable contribution towards my orientation and growth in the field of HIV Bioinformatics is much appreciated. Thank you to Dr Konrad Scheffler, Dr Wayne Delport, Rodger Duffett and all members of the National Bioinformatics Node (University of Cape Town) who were there to answer even the silliest question I had. My gratitude is also extended to my collaborators who contributed to make this work possible; Prof Carolyn Williamson, Dr Helba Bredell, Dr Clive Gray and Dr Penny Moore. Special thanks to the Stanford-South Africa Biomedical Informatics Network Executive Committee, Prof Russ Altman, Prof Winston Hide, Prof Cathal Seoighe and Dr Betty Cheng for their support and mentorship. I also thank my family (Mum, Ruth, Jenipher, Guido, Baldwin, Tanyaradzwa), close relatives and friends for their prayers and encouragement. Last and most importantly, I thank God for His love without which I would not have had this opportunity.

Contents

List of Publications

Nobubelo K. Ngandu, Cathal Seoighe, Konrad Scheffler (**2009). Evidence of HIV-1 adaptation to host HLA alleles following chimp-to-human transmission**. *Virology Journal,* 6:164

Nobubelo K. Ngandu, Konrad Scheffler, Penny Moore, Zenda Woodman, Darren Martin and Cathal Seoighe (**2008**). **Extensive purifying selection acting on synonymous sites in HIV-1 Group M sequences.** *Virology Journal* 5:160.

Nobubelo Ngandu, Helba Bredell, Clive M. Gray, Carolyn Williamson, Cathal Seoighe and the HIVNET028 STUDY TEAM (**2007**). **CTL Response to HIV Type 1 Subtype C Is Poorly Predicted by Known Epitope Motifs.** *AIDS Research and Human Retroviruses Journal* 23(8):1033-1041

Abbreviations

ω - dN/dS or omega

AIDS – acquired immunodeficiency syndrome

ANOVA – analysis of variance

AU – Australia

BR – Brazil

CRF – circulating recombinant form

CRS – cis-repressive sequence

CTL - cytotoxic T-lymphocyte

dN – nonysynonymous substitutions per nonysynonymous site

DNA – deoxyribonucleic acid

dS - synonymous substitutions per synonymous site

ELISpot – Enzyme-linked ImmunoSpot

ER – endoplasmic reticulum

ESE – exonic splicing enhancer

ESS – exonic splicing silencer

GARD - Genetic algorithm for recombination detection

HIV-1 – human immunodeficiency virus type 1

HLA - human leukocyte antigen

HyPhy - HYpothesis testing using PHYlogenies

IN – India

INS – inhibitory sequence

KE – Kenya

LTR – long terminal repeat

mRNA – messenger ribonucleic acid

PAML - Phylogenetic Analysis by Maximum Likelihood

PBMC – peripheral blood mononucleocytes

PPT – poly-purine tract

PSSM - position specific scoring matrices

RNA – ribonucleic acid

RNase - ribonuclease

RRE – Rev-responsive element

RT – reverse transcriptase

SIVcpz – simian immunodeficiency virus infecting chimpanzee

US – United States of America

ZA – South Africa

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Rationale

Even though AIDS was discovered more than 20 years ago (Masur *et al.*,1981), there is no cure for HIV-1 infection to date. Tireless efforts to fully understand the pathogenesis of the virus in humans and to develop a vaccine for the pandemic are in progress. The infection of humans with HIV-1 is puzzling in that the virus causes AIDS and subsequently death yet its natural hosts, the chimpanzee, from whom the zoonotic transmission of the virus to humans occurred, were until very recently found not to develop disease. There are obviously some differences in the pathogenesis of the virus between these two host species that cause the differences in disease outcome. In recent years, the hope for eradicating AIDS in humans has been on developing an anti-HIV vaccine that can boost immune responses that are able to prevent infection. An ideal vaccine reagent would be a highly conserved and immunodominant sequence region that is targeted by protective immune responses that stop the replication of the virus across all infected individuals.

Efforts towards understanding the pathogenesis of the virus as well as designing potential anti-HIV-1 vaccines are faced with two main challenges. The first is the high evolution rate of the HIV-1 genome sequence. The virus mutates and evolves rapidly due to (i) errors caused by the virus reverse transcriptase enzyme (Preston *et al.,*1988) (ii) recombination between viral genomes (Vartanian *et al.,*1991) and (iii) selection pressure acting on the viral sequences. The second challenge is the genetic variation between infected human hosts. With the exception of identical twins, each individual human has unique genetic make up. The genetic diversity causes inter-patient differences in the specific anti-HIV immune responses, resulting in variation in the diversifying selection pressure exerted upon the virus sequence. Hence different evolutionary patterns arise on the HIV-1 sequences thereby increasing the diversity of the virus. Various pathways of the immune system are involved in the response against HIV-1 infection (Levy,2007). The cytotoxic T-lymphocyte (CTL) immune

response in particular has been found to be very important in anti-HIV responses. The CTL immune response is mounted against viral peptides presented by human leukocyte antigen (HLA) molecules. The HLA molecules have been found to lead to varying selection pressures on the HIV-1 sequence and some of them correlated with HIV/AIDS disease progression. There is therefore need for detailed analysis of CTL immune responses-related selection pressures acting on the HIV-1 sequence and how it has influenced the evolution of the HIV-1 sequences since the zoonotic transmission from chimpanzee.

Much focus has been devoted to evaluating selection pressures acting on the protein sequences of HIV-1 in relation to immune responses. It had been assumed that the only selection pressure acting on the virus sequence is that which affects the encoded protein sequence, that is, affects codon sites that can alter amino acids. However, past evidence for the correlation of synonymous substitution rates and disease progression in HIV infection proved otherwise (Lemey *et al.*,2007). In addition, the virus has numerous motifs on the nucleotide sequence (Peterson *et al.*,1996; Van Lint *et al.*,1994; Wolff *et al.*,2003), which function to regulate gene expression and thus synonymous sites within such motifs can also be under some selection pressure. Therefore selection pressures affecting both the synonymous sites as well as those affecting the protein sequence need to be evaluated in order to improve the understanding of HIV-1 pathogenesis in humans. The detailed understanding of all aspects of the pathogenesis of the virus in humans can provide clues as to why AIDS develops and can subsequently lead to the development of better anti-HIV therapeutics.

1.2 Aims and Objectives

This PhD was done in order to contribute to scientific research aimed at understanding some aspects of HIV-1 pathogenesis in humans. The overall aim of the work was to evaluate the evolution of the virus sequence at both the synonymous sites and at the amino acid level. The early stages of the virus life cycle are regulated by a number of motifs of the nucleotide sequence and hence the rate of evolution at such motif regions could be restricted in order to preserve their functions. Thus it is interesting to determine the synonymous sites within these nucleotide motifs that are

under selection pressure. In the second case, the evolution of the amino acid sequence is evaluated in respect to selection pressures exerted by the CTL immune response. In particular, to understand in detail the evolution of the sequence sites that contain immunogenic epitopes, the overview of virus immune escape patterns in different infected populations and the extent of selection pressure exerted by HLA molecules since the time the virus first infected humans.

There are four analysis chapters (chapters 3 through to 6) in this thesis each reporting analyses carried out towards one of the following specific objectives:

(i) to evaluate the extent of selection pressure acting on the synonymous sites of HIV-1 protein-coding genes, i.e. sites that reflect selection pressure acting on the nucleotide sequence only and not that acting on the amino acid level

(ii) to determine the sequence variation and evolution at immunogenic regions of the HIV-1 sequence and relate the observed CTL immune responses to patient HLA genotypes

(iii) to investigate the adaptation of HIV-1 to specific CTL immune responses through immune escape mutations in different global human populations

(iv) to evaluate how HIV-1 has adapted to CTL immune responses directed by specific HLA alleles following its transmission from chimpanzee to humans

1.3 Thesis Layout

The background and a review of the literature related to areas covered in this thesis are presented in **Chapter 2** to form a basis for understanding the rest of the thesis. This chapter starts by introducing the evolution and diversity of HIV-1 followed by the biology of its lifecycle in an infected cell. The main causes of rapid evolution of the HIV-1 sequence as well as examples of commonly used methods for measuring the rate of evolution or determining selection pressure acting on the sequence are outlined. A summary of the human immune response, a major force exerting selection pressure on the HIV-1 sequence is given. An in-depth evaluation of the CTL immune

response pathway is laid out since it is one of the major determinants of HIV/AIDS disease progression and a main focus of this thesis. Methods for determining peptides that are targeted by the CTL response and previously published findings regarding immune escape and disease progression are also reviewed. This chapter concludes with challenges facing studies directed towards understanding the pathogenesis of HIV-1 and designing vaccines against HIV infection.

The evolution of synonymous sites of the nucleotide sequence is evaluated in **Chapter 3**. Here, for the first time, the extent of selection pressure acting on the synonymous sites across the whole HIV-1 genome is evaluated. As synonymous substitution rates have been shown to vary across sites, it is expected that sites that carry out important regulatory functions are under selection pressure to remain conserved in order to preserve their functions. Also, purifying selection pressure, which causes synonymous sites to be conserved, has been shown to cause false detection of positive selection. Therefore, this chapter presents work that was aimed at identifying all synonymous sites of the HIV-1 coding regions that are highly conserved across the 11 major subtypes of the HIV-1. Alignments of non-recombinant sequences of HIV-1 genes were downloaded from the Los Alamos HIV sequence database (Leitner *et al.*,2005). Known and novel regions under purifying selection pressure were identified and puplished to provide a reference source for studies that analyse positive selection pressure in HIV-1 sequences, so that false positive inference of positive selection resulting from purifying selection at synonymous sites can be avoided.

Chapter 4 investigates the relationship between the CTL response and the evolution and variation of the HIV-1 sequence. In particular, the relationship between the frequency of immune responses and sequence variability and the consistency of observed immune responses with patient HLA alleles. Ideal vaccine reagents are immunodominant peptides located in highly conserved regions of the virus sequence and eliciting protective immune responses across different individuals. The raw data used in this chapter were provided by two collaborating research groups headed by Professor Carolyn Williamson at the University of Cape Town and by Dr Clive Gray at the National Institute for Communicable Disease. The data comprised of CTL immune responses observed in a Southern African cohort infected with HIV-1

subtype C showing responses to the immunodominant Nef and Gag proteins of HIV-1. The autologous sequences isolated from the respective individuals were used for analysing conservation. Human leukocyte antigens (HLA) alleles, which are the key components that direct the CTL immune pathway, were provided from each patient. Two independent sequence analysis approaches were used to determine sequence variation along sites that are frequently recognised by the CTL immune response across the individuals in the cohort. The observed immune responses are also related to the corresponding patient HLA genotype and comparisons made with a comparative dataset from a subtype B infected North American cohort.

HIV-1 is known to mutate as a way to escape binding of HLA molecules and recognition by the CTL immune response (Kawashima *et al.*,2009). Immune escape mutations have been observed in regions bound by specific HLA alleles (Brumme *et al.*,2007). In **Chapter 5**, an overview of the way the virus sequence has adapted to the HLA alleles of different infected human populations is evaluated. The chapter explores whether HIV-1 mostly adapts to HLA alleles that occur more frequently in a population. Common HLA alleles are likely to mediate immune responses more frequently across different individuals. In such a case, if an escape mutation arises, the causative immune response in individuals who are subsequently infected by the mutant virus strain may prevent reversion of the mutant residue to wildtype and can cause the escape form to become fixed. The chapter also investigates whether adaptation mostly occurs in sites that have no functional constraints and hence can freely vary as compared to those sites that carry out important functions such that any mutation compromises viral fitness. The HIV-1 sequences from different population regions, the HLA frequencies in the corresponding populations as well as the sequence patterns required for HLA alleles to bind were obtained from public databases (Korber *et al.,*2006; Leitner *et al.,*2005; Meyer *et al.*,2007).

In **Chapter 6** a more detailed approach is used to determine HIV-1 adaptation to HLA alleles directing CTL immune responses following the cross-species transmission of the virus from chimpanzee to human. In this approach, CTL-directed selection pressure is investigated from the time the virus was first exposed to the human host. The objective was to identify positive selection that arose from immune escape soon after transmission from chimpanzee and may have resulted in the permanent loss of

some CTL epitopes. The sequences were obtained from the Los Alamos HIV sequence database (Leitner *et al.*,2005). A structure-based method, which uses amino-acid pairwise binding potentials was used to identify potential target sites for each HLA with an available crystal structure (Altuvia *et al.*,2004). The chapter therefore reports those HLA alleles with known crystal structures that are associated with positive selection pressure on CTL targeted regions of the HIV-1 sequence following the cross-species zoonosis event.

In **Chapter 7**, a summary of the conclusions of this study is given. Challenges and limitations of each project are presented as well as potential solutions.

Chapter 2

Background and Literature Review

2.1 Origin and Diversity of HIV-1

Scientific evidence shows that the human immunodeficiency virus type 1 (HIV-1) first infected humans in the early twentieth century through zoonotic transmission from the chimpanzee *Pan troglodytes troglodytes* species infected with simian immunodeficiency virus (SIVcpz). (Huet *et al.*,1990; Korber *et al.*,2000; Sharp *et al.*,1999; Sharp *et al.*,2001; Zhu *et al.*,1998). Further evidence shows that the cross-species transmission event took place in West-central Africa through blood contact between hunters and hunted chimpanzees (Hahn *et al.*,2000; Sharp *et al.*,1995). Multiple independent zoonosis events are thought to have occurred in the different parts of West-central Africa, which gave rise to phylogenetically distinct lineages of HIV-1 infecting humans (Gao *et al.*,1999; Hahn *et al.*,2000). Currently, there are three distinct lineages of HIV-1 sequences that are known to have come from these different cross-species events and are named group M (for Main), group O (for outlier) and group N (for non-M, non-O) (Figure 2.1) (De Leys *et al.*,1990; Jonassen *et al.*,1997; Simon *et al.*,1998). Of these, HIV-1 Group M is the most pathogenic accounting for most of the HIV-1 infections and the AIDS pandemic worldwide. (Gao *et al.*,1999; Hahn *et al.*,2000; Keele *et al.*,2006).

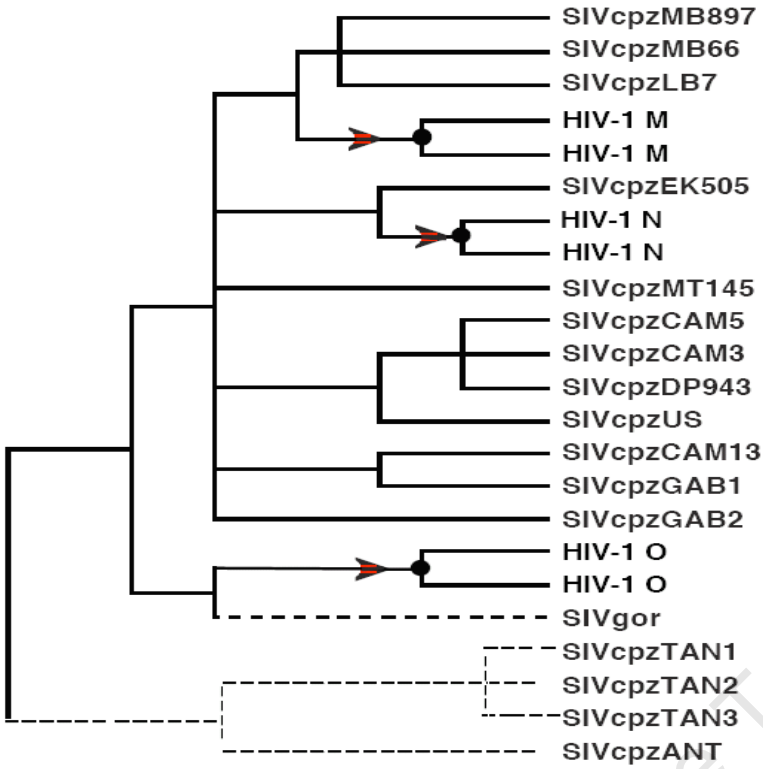


Figure 2.1: Adapted from (Wain *et al.*,2007). A phylogenetic tree showing the relationship between the three HIV-1 lineages M, N and O (with arrows) and the SIVcpz strains, dotted branches show other SIV that are not from chimpanzee *Pan troglodytes troglodytes* species, the countries of origin for some of the SIVcpz sequences are indicated as extensions to 'SIVcpz' and are CAM – Cameroon, US- United States of America, GAB- Gabon, TAN – Tanzania

2.1.1 The diversity and distribution of HIV-1 Group M

HIV-1 Group M has evolved rapidly since the 1930s resulting in diverse strains that have been grouped into subtypes based on sequence similarity. There are presently eleven major subtypes, A1, A2, B, C, D, F1, F2, G, H, J and K within the Group M lineage (Robertson *et al.*,2000). Inter-subtype differences of the envelope nucleotide sequences are up to 25-30% (Buonaguro *et al.*,2007; Korber *et al.*,1998). Most of these HIV-1 subtypes are found in the West-central African region where they originally diversified from the ancestral group M sequence (Peeters *et al.*,2000). The geographical distribution of the subtypes varies across different parts of the world.

The most prevalent subtype, subtype C, which accounts for almost 50% of world infections predominates in Sub-Saharan Africa and parts of Asia (Table 2.1.1, Figure 2.1.1) (Buonaguro *et al.,*2007; Geretti,2006). The rest of the subtypes are not comparably prevalent, with subtype B showing 10% prevalence and the A subtypes being 12% prevalent (Buonaguro *et al.*,2007). Subtype B is widely distributed in the developed world which includes North America, Europe and Australia (Peeters *et al.*,2000). Subtype A is found in parts of Europe and in Kenya and the surrounding East-African countries. Recombination between different subtypes has been observed and occurs when an HIV-1 positive individual is re-infected with a different subtype. This has resulted in recombinant sequences that cannot be clearly classified under any of the existing subtypes and are referred to as Circulating Recombinant Forms (CRFs) (Quinones-Mateu *et al.*,1999). There are currently more than 10 different CRFs and are distributed in different parts of the world (Leitner T *et al.*,2005).

Group M subtype	Percentage of world infections	Predominantly infected regions
A1 & A2	12.3	East Africa, East Europe, Central Asia
B	10.42	USA, Australia, West-Europe, Japan
C	49.91	Sub-Saharan Africa, India, China, Brazil
D	2.53	Central Africa
F1 & F2	0.59	South America, East Europe, Central Africa
G	6.32	Central Africa
H	0.17	Central Africa
J	0.14	Central Africa
K	0.04	Cameroon
CRF01_AE recombinant	4.69	South-east Asia
CRF02_AG recombinant	4.77	Central Africa
CRF03_AB recombinant	0.1	North-Europe
Other CRFs	8.02	Different parts of the world

Table 2.1.1: The HIV-1 group M subtype distribution in human populations. Percentages were taken from (Buonaguro *et al.*,2007). Geographical distribution of subtypes were taken from (Delgado *et al.,*2002; Geretti,2006; Leitner T *et al.,*2005).

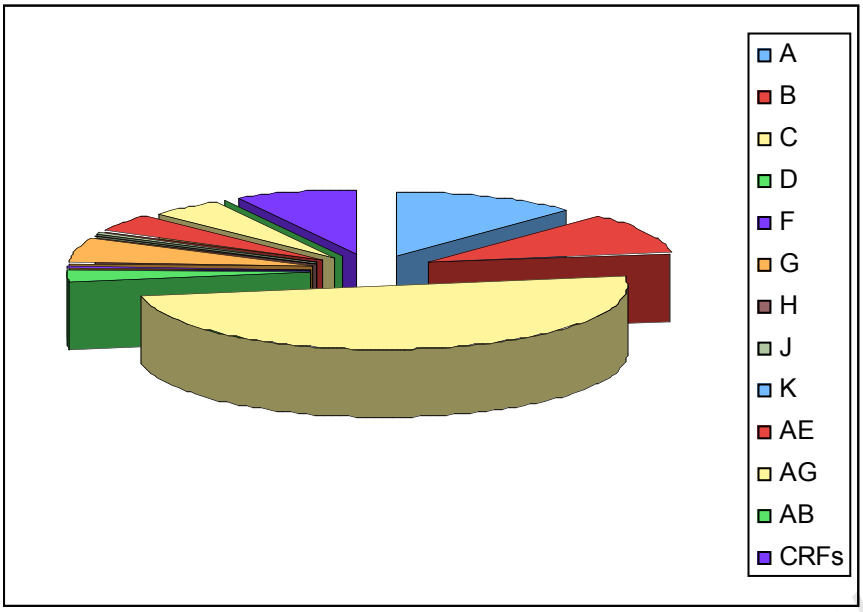


Figure 2.1.1 Percentages of HIV-1 group M subtype infections in the world. The numbers were taken from (Buonaguro *et al.*,2007).

2.1.2 Differences between SIVcpz and HIV-1 pathogenesis

Even though HIV-1 descended from SIVcpz, the pathogenesis of the two viruses differs between humans and chimpanzee respectively. The virus naturally replicates, reproduces and multiplies rapidly within a living host cell. Viral loads continue to increase in the host if the replication is un-interrupted or not suppressed. In humans, the immune response against HIV-1 is usually elevated upon infection causing a decrease and subsequent non-uniform fluctuations in viral load (Musey *et al.*,1997). In a host where protective immune responses are mounted, latent stages can be observed where viral load is reduced to very low levels (Betts *et al.*,1999). The exhaustion of the immune response however takes place with time resulting in uncontrollable increases in viral load and the development of acquired immunodeficiency syndrome (AIDS) and subsequently death (McMichael *et al.*,2001; Saksela *et al.*,1994; Shearer *et al.*,1997). By contrast, the chimpanzees *P t troglodytes* have in the past observed not to develop AIDS and remain asymptomatic throughout infection even in the presence of persistently high viral loads (Pandrea *et al.,*2008; Silvestri *et al.,*2007; Silvestri,2008). High viral loads were characteristic in

chimpanzees yet these hosts maintained low immune activation against SIVcpz (Silvestri,2008) unlike in humans who show stages of elevated immune responses. The high viral loads and low immune activation observed over the past years in natural hosts of HIV, the chimpanzees, could indicate a possibility of co-existence equilibrium between the virus and its host. However, a most recent study reported AIDS-related symptoms in SIV infected chimpanzee (Keele *et al*., 2009), which could indicate that the virus has successfully evaded its natural host but probably much slower than in humans, given the previous reports. These major differences observed between HIV and SIVcpz pathogenesis could be the result of a combination of both the differences in the two hosts' genetic make-up and sequence differences between the viruses themselves. Indeed there are genetic differences between SIVcpz and HIV-1 sequences since it is expected that HIV-1 underwent adaptive sequence changes to enable it to adapt to the recipient human host after the zoonosis event (Wain *et al.*,2007).

2.2 The Structure of HIV-1

HIV-1 is a virus classified under the retrovirus family and has a RNA genome which can be reverse transcribed into DNA (Levy,2007). The genome is a dimmer of two RNA molecules. It comprises of nine genes encoding proteins that carry out different functions. The nucleotide sequence also contains motifs most of which function to regulate gene expression (Pereira *et al.*,2000).

2.2.1 Genomic organisation of HIV-1

The HIV-1 genome consists of almost ten thousand base-pairs flanked by non-coding regions known as the Long-terminal repeats (LTRs), at each end thus forming the 5`LTR and the 3`LTR (Das *et al.*,1998). Each of the nine genes is translated using either the first, second or third reading frame (Pavesi *et al.*,1997). The reading frames overlap along some regions of the different genes to a maximum of 3 reading frames overlapping the same site (Pavesi *et al.*,1997). The genomic organisation of HIV-1 is shown in Figure 2.2.1 with the corresponding reading frames marked. The individual genes encode one of three types of proteins, classified according to types of function. The *gag*, *pol* and *env* genes encode structural proteins, the regulatory proteins are

encoded by the *tat* and *rev* genes and the rest of the genes, namely *nef*, *vif*, *vpr* and *vpu* code for accessory proteins (Frankel *et al.*,1998).

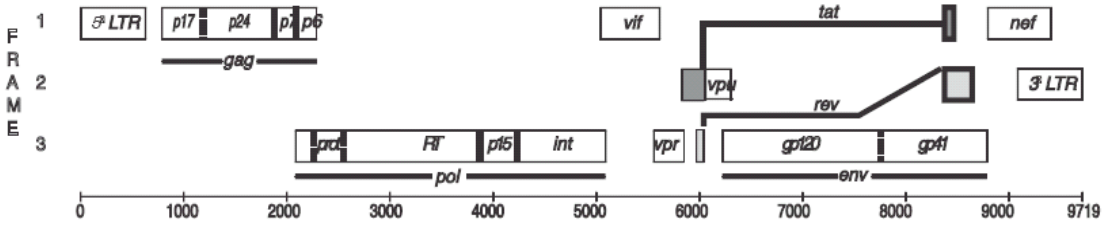


Figure 2.2.1: The genomic organization of the HIV-1 genome adapted from the Los Alamos database (Leitner *et al.*,2005). The bottom axis shows the position of bases numbered based on the HIV-1 HXB2 reference sequence from the Los Alamos sequence database (http://www.hiv.lanl.gov/content/sequence/). The reading frame is given at the left end of the diagram and the genes translated in each frame are drawn on the same row represented by rectangles drawn to scale. The names of proteins encoded by the structural genes are given within each rectangle above the corresponding coding region. The *tat* (black rectangles) and the *rev* (grey rectangles) exons are located in non-adjacent regions as indicated in the diagram.

2.2.2 Structural proteins

The *gag* gene codes for a poly-protein which is cleaved by the virus protease enzyme into four structural proteins, the matrix (p17, 132 amino acids) at the N-terminus, capsid (p24, 231 amino acids), nucleocapsid (p7, 55 amino acids) and p6 (51 amino acids) located at the C-terminus (Figure 2.2.2) (Frankel *et al.*,1998). The proteins form layers beneath the viral membrane and enclosing the RNA genome. The matrix is found beneath the lipid layer of the membrane and is important for the assembly of the envelope glycoproteins within the membrane (Gottlinger,2001). The capsid layer is directly below the matrix and forms a protective sheath around the genome and other proteins associated with the genome. The nucleocapsid directly interacts with regions of the RNA genome and facilitates its dimerisation and encapsidation during

assembly of virus particles (Berkowitz *et al.*,1995). The highly variable p6 protein has been found to play a role in the infectivity and release of new virions (Alexander *et al.*,2000; Gottlinger *et al.*,1991).

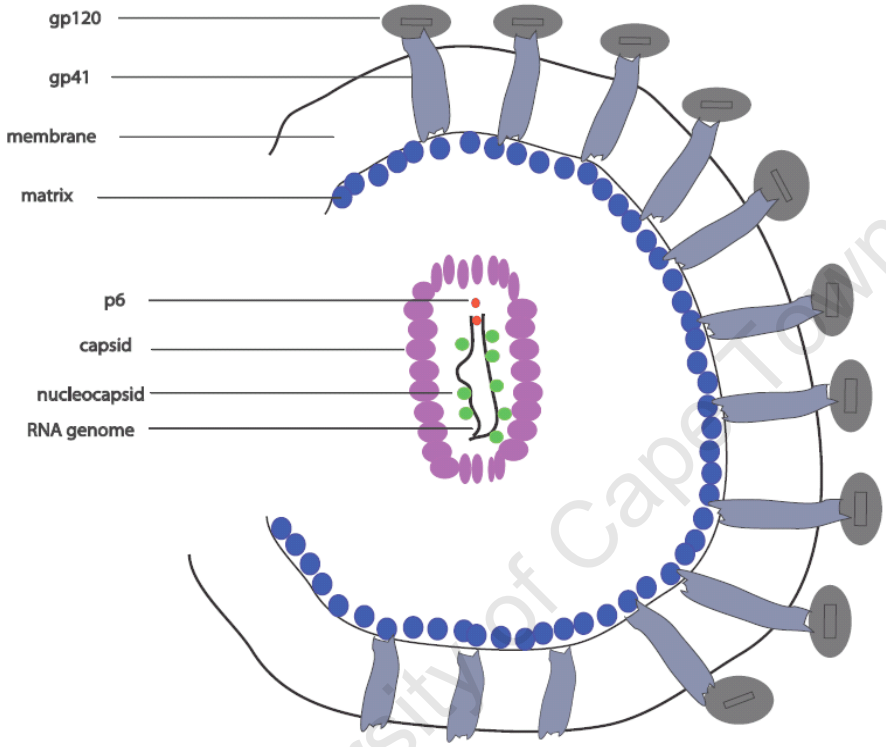


Figure 2.2.2 Schematic drawing of a section of the structure of a mature HIV-1 particle showing structural proteins encoded by the *gag* and *env* genes, based on (Levy,2007).

The *env* gene codes for the viral envelope glycoproteins (gp), gp120 and gp41 (Figure 2.2.2) (Wyatt *et al.*,1998). The gp120 molecules are the attachment points of the virus to CD4 molecules expressed on surfaces of some host cells of the immune system (Doms *et al,* 2000). The gp120-CD4 interaction triggers a conformational change on the gp120 thus fascilitating the binding to adjacent chemokine receptors, CCR5 or CXCR4 on the surface of the same host cell (Lapham *et al*., 1996). This then activates the fusion of the virus and host membranes thus enabling the virus to enter

the host cell (Cormier *et al.*,2001; Cormier *et al.*,2002). The pg120 surfaces are also sensitive to neutralizing antibodies of the immune response. Glycoprotein 41 has three main functions. The amino terminal domain has a critical role in the actual fusion of the virus membrane and the host cell membrane following the attachment of gp120 to the receptors on the surface of the host cell (Pascual *et al.*,2005). The trans-membrane domain spans across the cell membrane of the virus and firmly holds the gp120 in position (Helseth *et al.*,1991). The C-terminal region forms a tail which slightly extends into the interior of the virus particle and is important for incorporation of envelope glycoproteins in the formation of new virions (Murakami *et al.*,2000).

2.2.3 Regulatory proteins

The *tat* and *rev* genes code for regulatory proteins. The trans-activating protein (Tat, 86 amino acids long) interacts with a Tat responsive element within the 3` LTR in the regulation of virus replication (Greene *et al.*,2002). Not all the mRNA of HIV-1 genes is fully spliced and some genes like *gag* are translated unspliced into polyproteins (Kimura *et al*., 2000). Un-spliced mRNA cannot be normally transported through the nuclear membrane to the cytoplasm where translation takes place (Peterson *et al.,* 1996). Therefore, the regulator of viral protein expression (Rev, 116 amino acids long) through interaction with a Rev-responsive element, an RNA structure within the *env* gene, functions to facilitate the transport of the unspliced viral mRNA from the nucleus to the cytoplasm. (Peterson *et al.*,1996).

2.2.4 Accessory proteins

The accessory protein Nef (negative factor, 206 amino acids) had been observed to carry out a number of functions including down-regulation of expression of CD4 molecules by HIV infected cells (Foster *et al.*,2008; Lama *et al.*,1999). CD4 molecules are expressed on the surfaces of some white blood cells and contain receptors that attach to the viral envelope, together with chemokine receptors, enabling fusion and entry of viruses into cells. Therefore, the down-regulation of CD4 molecules by the Nef protein prevents super-infection of an infected cell, which can be disadvantageous to the endogenous virus. Another major function of the Nef

protein is the down-regulation of human leukocyte antigen molecules (HLA) (Greenberg *et al.*,1998). HLAs bind to virus peptides and activate anti-HIV immune responses therefore the Nef protein protects the infected cell from immune attack.

The virus infectivity factor (Vif, 192 amino acids) protein is essential for the assembly of new virus progeny and their ability to infect host cells (Levy,2007). The viral protein R (Vpr, 96 amino acids) is important in regulating replication and in the import of complexes required for the integration of the virus genome into the host cell DNA in the nucleus (Greene *et al.*,2002). The viral protein U (Vpu, 81 amino acids long) contributes to the function of the Nef protein by degrading CD4 molecules that are being transported to the cell surface within the endoplasmic reticulum of the infected cell (Frankel *et al.*,1998). The release of virus progeny from infected cells is facilitated by Vpu proteins that are bound to the membrane (Schubert *et al.*,1996).

2.2.5 Functions at the nucleotide sequence level

The genome of HIV-1 also contains a number of motifs that function at the nucleotide sequence level within the coding regions (Cochrane *et al.,*1991; Schneider *et al.,*1997; Schwartz *et al.,*1992). The non-coding LTR regions (including the regions of the 3'LTR which also encode for part of the Nef protein) do contain a number of transcription factor binding sites and motifs that interact with other proteins in the regulation of gene expression such as the Tat responsive element in the 3'LTR (Das *et al.,*1998; Pereira *et al.,*2000; Quinones-Mateu *et al.,*1998). Many splice sites and splicing regulatory elements are found overlapping with coding regions in other reading frames and regulate the alternative splicing of the more than twenty transcripts that can be produced from the viral genome (Figure 2.2.5) (Gog *et al.,*2007; Kammler *et al.,*2006). The *nef*, two *tat* exons and the two *rev* exons are fully spliced as the early products of gene expression (Swanson *et al.*,1998). The *env, vif, vpu, vpr* and *tat* can be partially spliced and require the Rev protein for efficient transport from the nucleus (Lassen *et al*., 2006) The unspliced *gag* and *pol* mRNA are the late products as they require threshold levels of the Rev protein to facilitate their nuclear transport to the cytoplasm (Kammler *et al.*,2006).

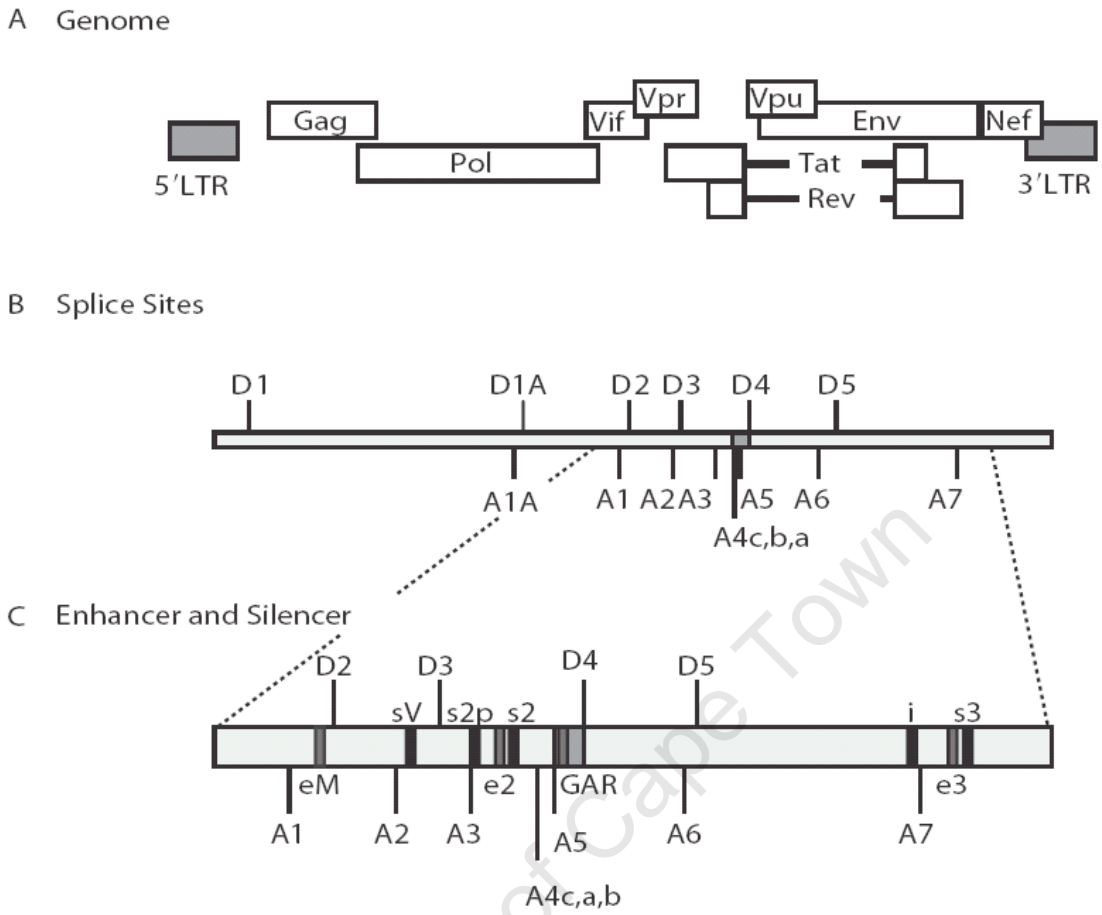


Figure 2.2.5 The organisation of splicing regulatory elements of the HIV-1 genome adapted from (Asang *et al.*,2008). The open reading frames of the HIV-1 genome. B: 5` splice donor sites labelled D1 to D5 and 3` splice acceptor sites labelled A1 to A7. C: Splice silencer elements in black and splicing enhancers in grey vertical bars.

In addition to splice acceptor and donor sites, inhibitory nucleotide sequence motifs (abbreviated as INS motifs) have been identified within the *gag* and *pol* (cis-repressive sequence) genes which play roles in negatively regulating gene expression (Cochrane *et al.,*1991; Schneider *et al.,*1997; Schwartz *et al.,*1992; Wolff *et al.,*2003). An intragenic nuclease hypersensitive regulatory region (also referred to as HS7) with functions similar to that of the LTR has been previously identified within the central *pol* gene and contains transcription factor binding sites (Goffin *et al.,*2005; Van Lint *et al.,*1994; Verdin *et al.,*1990). A highly conserved poly-purine tract in the central region of the *nef* nucleotide sequence has been found to be important in initiating of

plus-strand DNA synthesis (Miles *et al.*,2005; Rausch *et al.*,2004). The Rev-responsive element is a nucleotide sequence along the *env* gene that forms stem loops in its secondary structure and directly interacts with the Rev protein in regulating the transport of unspliced and partially spliced mRNA from the nucleus to the cytoplasm of the infected cell (Hadzopoulou-Cladaras *et al.,*1989; Hung *et al.,*2000; Peterson *et al.,*1996; Renwick *et al.,*1995). The HIV-1 genome is therefore subject to selection pressure as both the protein and nucleotide sequence levels.

2.3 The Lifecycle of HIV-1

HIV-1 is transmitted via blood contact and other body fluids such as genital fluid and breast milk. It infects cells of the immune system, which include T cells, macrophages and other cells containing appropriate receptors on the cell surface that can be bound by gp120 molecules of the virus envelope. However, in addition to the chemokine receptors that are bound by the virus, it also has a very high affinity for CD4 molecules hence HIV-1 mostly infects T cells that naturally express CD4 molecules on other surfaces, the CD4+ T cells.

2.3.1 Stages of the HIV-1 life cycle

Like other viruses, HIV-1 only replicates within a living host cell. The first stage of its lifecycle is attachment to the host cell and entry. The gp120 molecules on the virus envelope attach to chemokine receptors CCR5 or CXCR4 and CD4 molecules on the cell surface thus triggering fusion of the two membranes and release of the virus contents into the cytoplasm of the host cell (Levy,2007). The virus reverse transcriptase enzyme then reverse transcribes the dimeric RNA genome of the virus into DNA within the host cell cytoplasm. The double-stranded DNA is transported into the nucleus of the host cell where it is integrated into the host cell genome by the virus integrase enzyme (Wu,2004). This enables the virus to take advantage of the host cell machinery to express its genes. The host enzymes are used for the transcription and translation of the viral DNA. As mentioned in section 2.2.5, smaller exons from *nef, tat* and *rev* that are fully spliced are the early products of translation since their mRNA is easily transported from the nucleus to the cytoplasm where translation takes place. The rest of the partially spliced and unspliced mRNA is

translated as late products as they require adequate levels of the Rev protein to aid in their nuclear export. The protease enzyme is used to cleave poly-proteins such as Gag and Pol into individual functional protein products. Copies of the full genome unspliced RNA are also exported to the cytoplasm for the formation of new virus progeny (Wu,2004). The resulting virus proteins and genomic RNA dimmers are then assembled into new virus particles within the cytoplasm. The mature virus progeny are released from the infected cell via budding to the host cell after which they infected other cells. The main stages of viral replication are illustrated in Figure 2.3.1.

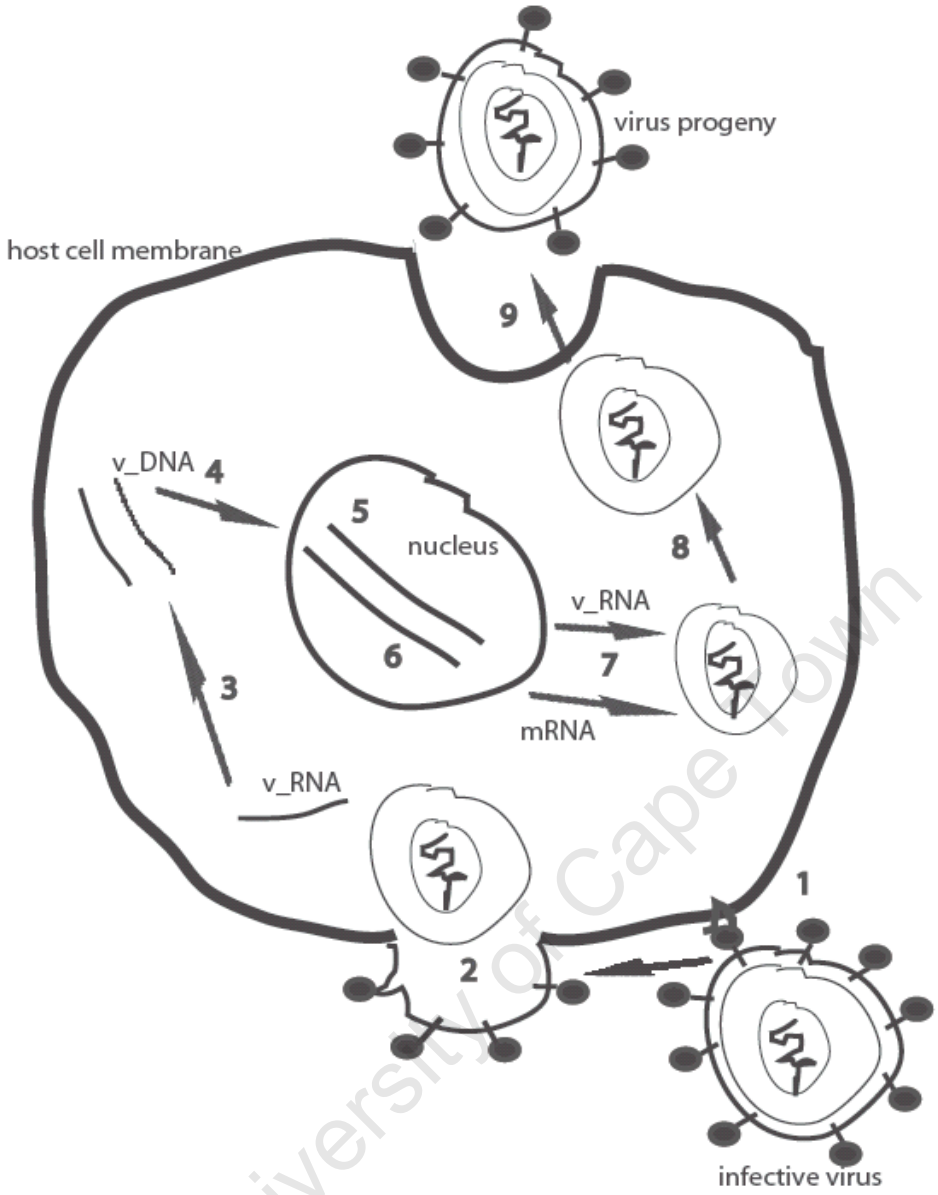


Figure 2.3.1 Stages of HIV-1 replication within a human host cell numbered in order from 1 through to 9 following the arrows. The stages were adapted from various sources (Grigorov *et al.,*2006; Levy,2007; Nisole *et al.,*2004; Wu,2004).1; attachment of gp120 molecules to receptors on cell surface, 2; membrane fusion and release of contents, 3: reverse transcription of viral genomic RNA (v_RNA) to double stranded DNA (v_DNA), 4; nuclear transport of v_DNA, 5; integration of v_DNA into host DNA, 6: transcription of v_DNA, 7; new viral v_RNA and mRNA transported to cytoplasm, translation of mRNA and assembly of new virions, 8; maturation 9; coating with envelop and release of mature progeny.

2.3.2 The HIV-1 rate of replication and evolution

HIV-1 has a very high turnover rate, producing an average of about two hundred infectious virions within a single infected host cell (Dimitrov *et al.*,1993). The average generation time is two and a half days during which about 10^{10} virions can be produced per day (Perelson *et al.*,1996). The mutation rate of the virus is very high with an approximated $1.7*10^{-3}$ nonsynonymous nucleotide mutations per site per year (Li *et al.*,1988). These mutations result mainly from errors made by the reverse transcriptase enzyme which is defective of proof-reading and makes approximately 0.2 errors per genome during each replication cycle (Preston *et al.,*1988). Therefore each virion genome produced after a single lifecycle is not exactly similar to the parent virus thus causing a continuous evolution of virus sequences. The HIV-1 sequences are also highly recombinogenic because of the dimeric structure of the RNA genome which influences recombination between non-adjacent regions of the dimmer or between genomes (Sharp *et al.*,1995; Vartanian JP *et al.*,1991). Recombination occurs more frequenctly during reverse transcription of the minus-strand when the reverse transcriptase switches templates (Negroni *et al.*, 2000). Recombination results in shuffling of sequences, producing new genomic RNA that is different from the parent.

The evolution of HIV-1 is also largely influenced by external diversifying selective pressures acting on the virus sequence. Regions of the sequence that are of functional or structural importance can be under purifying selection pressure to maintain their fitness. However, as the virus constantly experiences random mutations genome-wide, some of these mutations can confer a fitness advantage and enable the virus to evade external selective pressures such as immune attack and thus adapt to its environment. The host immune response in humans in particular does exert varying selection pressures of viral sequences and the latter have been observed to undergo immune escape mutations. In such cases, wildtype residues can be lost and the resultant mutants fixed if the latter enable the virus to adapt well to its new host without compromising its fitness and viability. Due to different immune responses between infected individuals and differences in sequence sites of the HIV-1 sequence that may be targeted, immune escape patterns and fixation of mutants can vary widely across different HIV-1 subtypes and within a single subtype. This therefore further increases

the evolution rate and diversity of the virus between infected individuals. The rate of sequence evolution and variability of HIV-1 sequences has been found to be directly correlated with its pathogenicity and disease progression (Quinones-Mateu *et al.*,1998; Shpaer *et al.*,1993).

2.4 Determination of the Rate of HIV-1 Evolution

Efforts are being made by scientists worldwide to find a cure for HIV-1 infection. One of the greatest challenges to designing effective treatment is keeping up with the ongoing rapid evolution of HIV-1 and its increasing diversity. The natural mutation of the virus, errors caused by reverse transcriptase, recombination and external selection pressure exerted on the virus all contribute to the evolution of HIV-1.

Different regions of the virus evolve at varying rates mainly due to their functional differences. Sites that have important functions can be under purifying selection pressure to preserve their functions and hence tend to be conserved and evolve slower. In contrast, regions of the sequence that have no critical functions such that losing the wildtype nucleotides or amino acid residues does not confer a fitness cost can evolve more rapidly. As the virus gets transmitted from host to host, it is constantly being exposed to new and different selection pressure hence different mutations that enable adaptation to recipient hosts arise (Casado *et al.*,2001). Different humans have different genetic make up (except for identical twins) hence adaptive evolution to host factors such as the immune response can differ between different viruses thus further increasing the diversity between virus sequences. Resistance mutations also arise against Anti-HIV therapeutics. Determining the evolution rate of the virus sequences is therefore important for studies aimed at understanding HIV-1 pathogenesis. Numerous methods have been developed and attempts made to evaluate the evolution rates of HIV-1 sequences. These methods enable the inference of evolutionary rates at both the nucleotide sequence level and the amino acid sequence level.

2.4.1 Codon models of sequence evolution

The greatest achievement in studying the evolution of protein-coding sequences has been the development of mathematical codon models that take into account the

phylogenetic relationships of sequences. Two major software packages, Phylogenetic Analysis by Maximum Likelihood (PAML) (Yang,2007) and HYpothesis testing using PHYlogenies (HyPhy) (Kosakovsky Pond *et al.*,2005d), are available within which various codon models and methods for analysing sequence evolution can be implemented. The codon models use phylogenetic trees representing evolutionary relationships between the sequences. The selection pressure acting upon each codon site of a sequence is usually assessed from the ratio of nonysynonymous (amino acid changing) substitutions per nonysynonymous site (dN) to the synonymous (silent mutations) substitutions per synonymous site (dS). The dN/dS ratio is also referred to as omega (ω). Positive selection is inferred when $\omega>1$, purifying selection when $\omega<1$ and $\omega = 1$ implies that sites are evolving neutrally. The instantantaneous transition rates used by codon models are summarized in Table 2.4.1a.

Substitution rate	Nucleotide differences between codon i and codon j
0	If i and j differ at two or three nucleotide positions
π_j	If i and j differ by one synonymous transversion
$\kappa\pi_j$	If i and j differ by one synonymous transition
$\omega^{(h)}\pi_j$	If i and j differ by one nonsynonymous transversion
$\omega^{(h)}\kappa\pi_j$	If i and j differ by one nonsynonymous transition

Table 2.4.1a: the instantaneous substitution rate from codon i to codon j in a sequence alignment with h number of codons, the approach used by codon models where $\kappa\pi_j$ is the equilibrium frequency of codon j, κ = transition/transversion rate ratio and $\omega^{(h)}$ is the ω for codon site h.

The early codon models were simpler, with fewer assumptions but usually less detailed and have been improved over the years to recent ones that are more biologically accurate especially for rapidly evolving sequences such as those of HIV-1 (examples in Table 2.4.1b and also reviewed in (Delport *et al.*,2009)). The early codon models that are commonly used are GY94 and MG94 (Goldman *et al.*,1994;

Muse *et al.*,1994). In the MG94, the subsitution rates are based on the target nucleotide within a codon and in the GY94 they are based on the target codon. Later, Yang and colleagues developed the 'Nonsynonymous rates model' i.e. allowing site-to-site rate heterogeneity of nonsynonymous substitution rates (Yang *et al.*,2000). However, dS was assumed to be constant and equal to one across all sites, thus the site-to-site variation in ω reflects only the variation in dN. The proportional model, which is the approach used is most previous models assumes that dS and dN are proportional such that dN can be obtained from ω*dS e.g. (Yang *et al.*,2000). Following the evidence that dS varies between codon sites (e.g. (Hurst *et al.*,2001)) the MG94 model was modified by (Kosakovsky Pond *et al.*,2005a) to allow for site-to-site variation in synonymous substitution rates. Improvements to the latter model also included independent and discrete distributions of dN and dS, termed the 'Dual' model where $dN \neq \omega * dS$. Therefore this latter model allows for estimation of selection pressure acting on the nucleotide sequence independent of that acting on the amino acid level, i.e. only affecting dS. Recent developments include a genetic algorithm that enables the determination of averaged ω acting on each branch of an phylogeny to be used when different lineages are expected to have undergone different selection pressures (Kosakovsky Pond *et al.*,2005b).

The PAML package includes nested models for the formation of a likelihood ratio test that compares the statistical fit of models that allow a subset of sites with positive selection to the fit of models without positive selection. Frequently used examples are the M1 (neutral) and M2 (selection) models developed by (Nielsen *et al.*,1998). M1 forms the null model in which ω_0 is fixed to 1 and sites assumed to be evolving neutrally. M2, for the detection of the selection, is the alternative model where ω is estimated from the data with the likelihood compared to that of M1. The models are implemented using the CODEML program available within the PAML package (Yang,2007). These nested models were later improved to M1a (near neutral) such that ω_0 is estimated from the data and $0<\omega_0<1$ and M2a (positive selection) which allows $\omega_0>1$ (Wong *et al.*,2004; Yang *et al.*,2005b).

Codon Model	Authors and Publication date	Main Features
MG94 Constant model	Muse & Gaut, 1994	Constant rates model, single dN across all sites, dS = 1 across sites, dN & dS are proportional i.e. dN = ω*dS
M1 (neutral) & M2 (selection)	Nielsen & Yang, 1998	Nested Models for likelihood comparison, M1 (fixed $\omega_0 = 0$, $\omega_1 = 1$) and M2 (ω_2 varies among sites & is >0).
Nonsynonymous	Yang et al, 2000	Site-to-site variation in dN, constant dS, proportional
M1a (near neutral) & M2a (positive selection)	Wong et al 2004	Improved M1 & M2, where for M1a (ω_0 varies freely between $0<\omega_0<1$)
MG94 improved Dual model	Kosakovsky Pond & Muse, 2005	Site-to-site variation in dN, Site-to-site variation in dS, dual discrete distributions of dN and dS i.e. dN ≠ ω*dS
Dual lineage variable rates model	Kosakosky Pond & Frost (2005)	As improved MG94 but lineage-specific, averaged ω for each branch on a phylogeny

Table 2.4.1b Examples of most frequently used codon models showing the stages of improvement that has been made over the years

2.4.2 Tackling the problem of recombination

Another challenge in the use of codon models was the problem of recombination, which is especially frequent in virus sequences. The main problems caused by recombination are that firstly the recombination breakpoints can cause a distortion in the residue of the sequence itself and may appear as a signal of positive selection. Secondly, recombined sequences are usually from different sequences, for example from different subtypes in the case of HIV-1, and hence have different phylogenetic histories. In this case, the phylogeny of each segment of the recombinant should be

inferred separately and the different segments should not be assumed to have the same tree (Scheffler *et al.*,2006). Methods for the detection of recombination have been developed in order to identify recombining sequences prior to analysis of selection pressure using phylogeny-based methods such as the codon models. One of these tools is the Genetic algorithm for recombination detection (GARD) (Kosakovsky Pond *et al.*,2006) available from the HyPhy package and implemented in the Datamonkey web server (Kosakovsky Pond *et al.*,2005c). In order to avoid the complication of analysing different recombined segments of a sequence alignment separately, Scheffler and colleagues (Scheffler *et al.*,2006) developed and algorithm that allows for use of codon models to detect selection using different tree topologies for each recombination segment of a single alignment in a single analysis by specifying recombination breakpoints and providing the corresponding trees as input to the analysis.

2.4.3 Methods for measuring evolution using amino acid sequences

Phylogenetic methods that measure the evolution rate of amino acid sites from alignments of amino acid sequences are also available. For example AAML is a PAML program with similar functions as CODEML but takes amino acid sequences as input (Yang,2007). Another example is the Rate4Site program developed by Mayrose and colleagues (Mayrose *et al.,*2004; Pupko *et al.,*2002) for determining conserved amino acid sites by calculating evolutionary rates at each site using either Bayesian or maximum likelihood methods. Sequence entropy, a measure of disorder per amino acid site in an alignment is also commonly used as a measure of amino acid variability per site e.g. in (Sato *et al.,*2007). It is a simplified approach for identifying protein regions that are highly conserved.

2.5 Major Pathways of the Human Immune Response

The major obstacle to curbing HIV-1 infection is that it infects cells of the immune system, predominantly the CD4+ T cells Chan *et al.,* 1998). Monocytes such as macrophages as well as dendritic cells are also infected but to a lesser extent than the CD4+ T cells (Levi, 2007). The immune system is the major tool with which the human host fights against HIV-1 pathogens. The human immune system is divided

into two main pathways, the innate and adaptive immune responses (Figure 2.5) (Todar, 2009). Different arms of the immune system are involved in targeting HIV-1 pathogens that are either circulating in the blood plasma (exogenous) or have infected cells (endogenous).

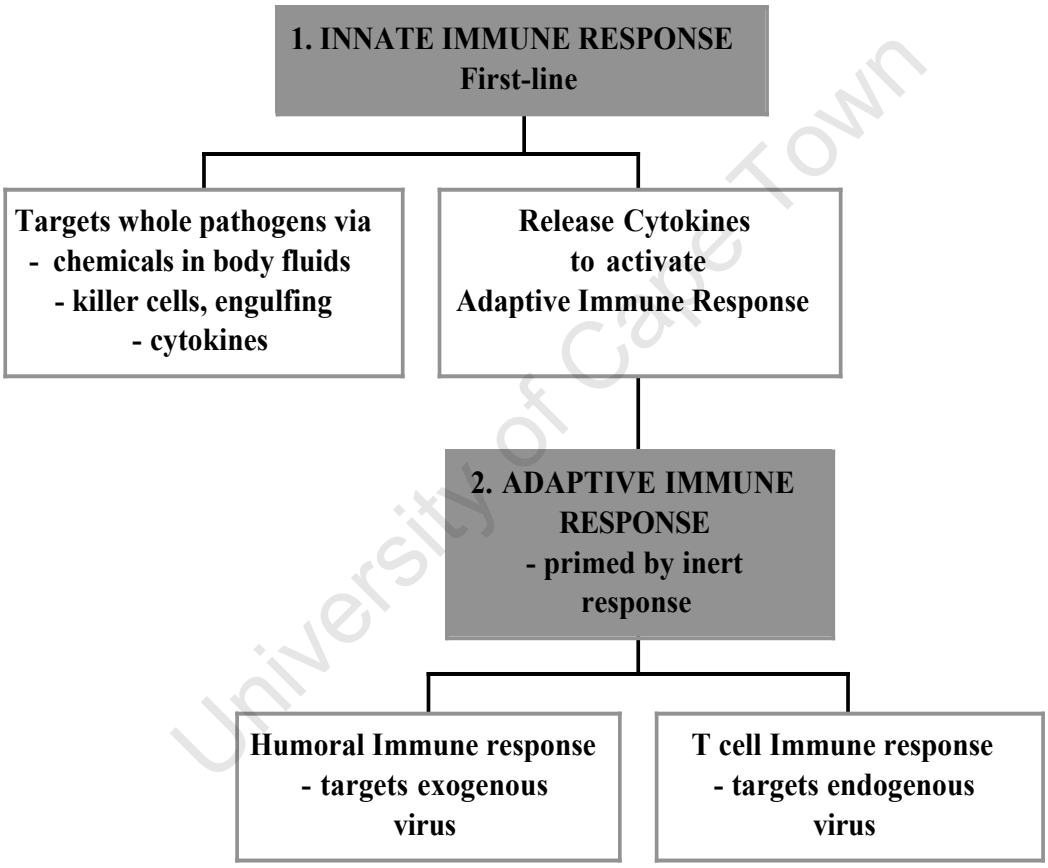


Figure 2.5: The main pathways of the human immune response are shaded in grey (Levy,2007).

2.5.1 The innate immune response

The innate immune response is the first-line pathway and provides immediate protection against pathogens upon infection (Todar, 2009). It therefore targets viral pathogens before they infect host cells, by recognising anti-self characteristics on the surfaces of the pathogen. Chemicals known as cytokines, produced by cells of the immune system circulate in body fluids such as blood plasma, provide a first line of attack against exogenous pathogens (Flajnik *et al.*,2004). Other fluids of the body containing both natural (born-with) and acquired immunity (received from other source such as vaccination or perinatal transmission) such as saliva, placental fluid, tears and milk can also contain immune chemicals that provide innate immune responses (Flajnik *et al.*,2004; Todar, 2009). Some cells of the immune response, which include neutrophils, natural killer cells and macrophages also attack whole pathogens by either release of chemicals or by engulfing the pathogen (Levy,2007). The other most important characteristic of the innate immune response is the activation of the cells of the adaptive immune response via cytokines, in response to the detection of a foreign pathogen (Flajnik *et al.*,2004).

2.5.2 The adaptive immune response

The adaptive immune response is a complex of two pathways that function inter-dependently. One is the humoral immune pathway, which mainly recognises and targets exogenous pathogens and the second is the T cell immune pathway that targets endogenous pathogens (Todar, 2009). The outline of the two pathways is shown in Figure 2.5.2 below. The cells involved in mediating adaptive immunity are produced by stem cells and developed either in the bone marrow or thymus gland.

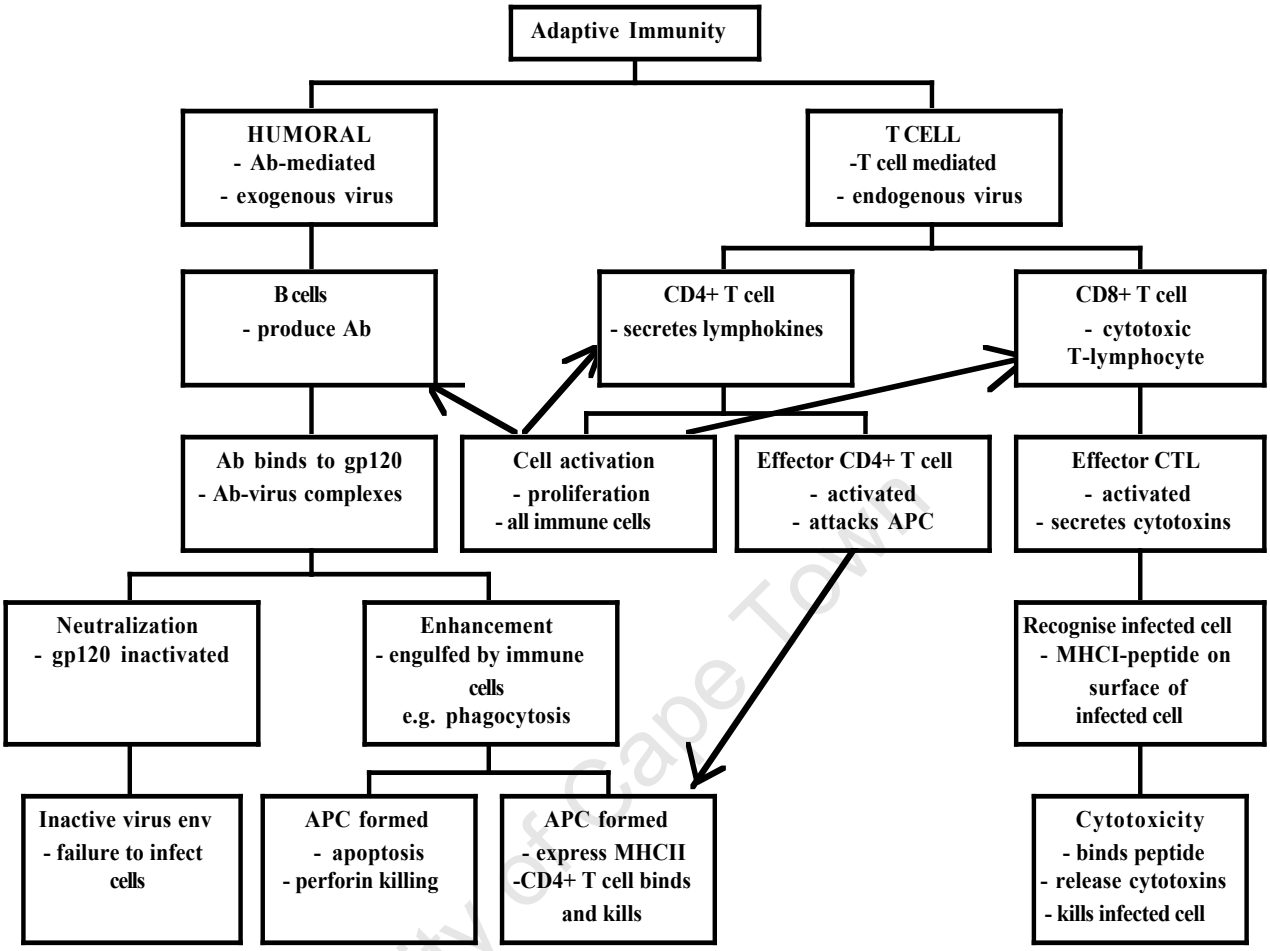


Figure 2.5.2: The adaptive immune response pathways (Flajnik *et al.,*2004; Levy,2007). Ab – antibody, APC – antigen-presenting cell, CTL – cytotoxic T-lymphocyte

The humoral, also known as antibody-mediated immunity is mainly directed by B-cells (developed in the bone-marrow). Activated B-cells produce antibodies which function to either neutralise exogenous virus or enhance recognition of the pathogen by the T cell immunity. A B-cell activated by the presence of a particular pathogen only produces antibodies specific to that pathogen. In neutralisation of HIV-1, antibodies bind to specific regions of the surfaces of gp120 molecules of the virus thereby neutralising them such that they are not able to infect host cells (Sagar *et al.*,2006; Yang *et al.*,2005a). The neutralisation of HIV-1 envelop proteins has been

found to be very high during early HIV-1 infection and correlated with an increase in immune escape mutations (Richman *et al.*,2003). Other evidences have also shown that neutralising antibodies play an important role in driving the evolution of HIV-1 envelop proteins (Frost *et al.*,2005; Wei *et al.*,2003). Enhancement involves binding of an antibody to the pathogen and thus enabling recognition of pathogen by macrophages or dendritic cells which engulf the antibody-pathogen complex (Levy,2007; Subbramanian *et al.,*2002). The macrophage bearing the pathogen, referred to as the antigen-presenting cell (APC), triggers its apoptosis or perforin killing (Yagita *et al.*,1992). APC presenting the antibody-bound pathogen on its surface can also be targeted by the T cell immunity depending on the presence of appropriate receptors on the APC surfaces, which allow for the T cells to bind. APC containing molecules known as major histocompatibility complex type 2 (MHCII), in particular, are bound by T cells that bear CD4 molecules on their membranes (Lanzavecchia,1998).

The T cell immune pathway, also known as cell mediated immunity (CMI), is mediated by white blood cells known as lymphocytes that develop in the Thymus glad hence known as T cells. There are two types, the CD4+ cells mentioned in the previous paragraph and the CD8+ T cells which bear CD8 molecules on the surfaces of their membranes. The CD4+ T cells also contain chemokine receptors, the main receptors enabling infection by HIV-1. They play an important role in adaptive immunity by mainly producing chemicals known as lymphokines which activate the proliferation and activation of all the other cells e.g. CD8+, macrophages and other CD4+ T cells. As mentioned in the previous paragraph, CD4+ T cells also attack antibody-bound pathogens presented on surfaces of APC by simultaneously binding to MHCII molecules (Lanzavecchia,1998). The CD8+ T cells carry out the cytotoxic T-lymphocyte (CTL) immune response, which plays a major role in anti-HIV-1 immune responses. Details of the CTL immune pathway are discussed in the next section.

2.6 The Cytotoxic T-lymphocyte Immune Response

The CTLs (CD8+ T cells) are activated to produce chemicals that destroy infected cells. The secretion of chemicals is activated by the interaction of the CTL with short

protein peptides of the pathogen and MHC type 1 molecules on the surface of the infected cell. The CTLs only bind to peptides bound by MHCI molecules (Ohno,1992). Therefore, proteins of the endogenous pathogen within the infected cell need to be cut into short peptides that can be bound by MHCs and presented to the CTLs on the cell surface (Ohno,1992). MHC molecules specific to humans are known as human leukocyte antigens (HLA). Throughout the rest of this thesis, the term human leukocyte antigen or HLA will be used to refer to the human MHC type I molecules.

2.6.1 Stages involved in the CTL immune response

Proteins of the pathogen are cut down into short peptides by the proteolytic enzymes in the cytoplasm during the degradation of some host cell proteins, a process that is part of the cell cycle (Ohno,1992). The resulting peptides are engulfed by golgi vesicles and transported to the smooth endoplasmic reticulum (ER). Non-self peptides are recognised by HLA molecules within the ER and bound in a specific manner (Ohno,1992). Each HLA molecule can only bind peptides with specific sequence motif patterns that complement the binding groove of the HLA. The HLA-peptide complexes within the lumen of the ER are transported to the surface of the infected cell. The CTLs, which have been activated by lymphokines produced by CD4+ T cells, recognise and bind only to HLA-presented peptides. The interaction of the CTL cell receptors with the HLA and peptide activates the CTL to secrete chemicals that destroy the infected cell. HIV-1 predominantly infects CD4+ T cells hence increased viral replication and cell infection results in the reduction of CD4+ T cell population (Boritz *et al.*,2004). The main steps involved in mounting a CTL immune response are presented in Figure 2.6.1. The HLA class I alleles form a major component of the cytotoxic T-lymphocyte immune response since the pathogenic peptides can only be recognised by the CTL in the context of HLAs (Frahm *et al.,*2007). No CTL immune response is elicited if peptides are not presented bound to HLA molecules.

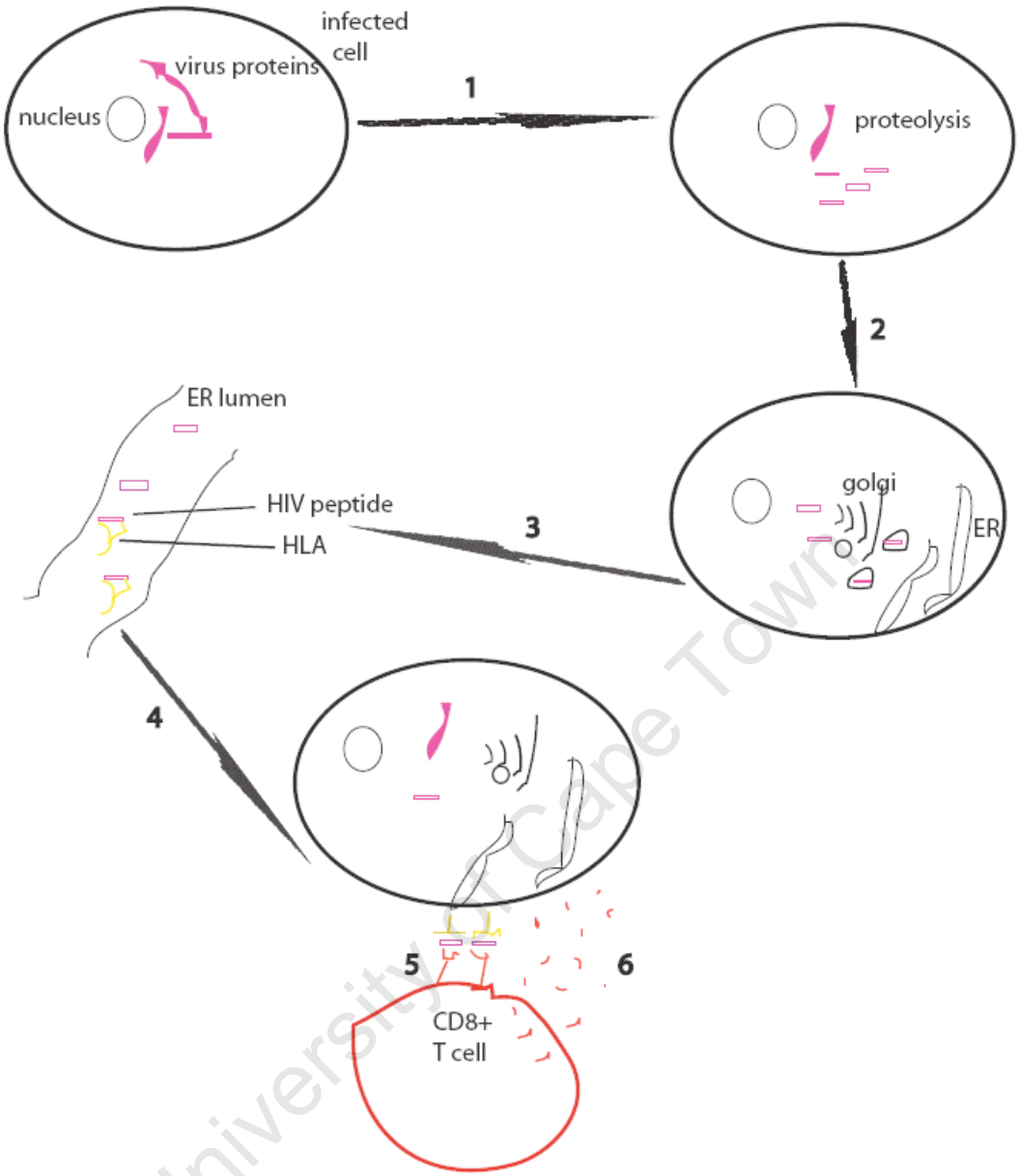


Figure 2.6.1: Schematic representation of the main steps involved in a CTL immune response adapted from various sources (Janeway *et al.*,1997; Levy,2007; Yang *et al.*,1997), 1; proteolysis of self and pathogen proteins, 2; transport to ER via golgi vesicles, 3; within ER, HLA-peptide binding, 4; HLA-peptide complexes presented on cell surface, 5; recognition and binding by CTLs, 6; production of cytotoxic chemicals and destruction of infected cell

2.6.2 HLA polymorphism and binding

The HLA gene is found within the short arm of human chromosome 6 (Marsh *et al.*,2000). Three loci within the gene region for HLA class 1 encode the HLA-A, -B and -C allotypes, which are the most active in mediating CTL immune responses. These three loci are highly polymorphic such that many alleles exist within each HLA allotype with specific sequence differences that affect their binding preferences (Williams,2001). The HLA-B locus is the most polymorphic comprising the highest number of individual alleles, followed by HLA-A and then C (Marsh *et al.*,2000). One of the most accurate methods for HLA typing is use of serological assays where antibody-HLA complexes are isolated from serum of multiparous women in which the mother's antibodies are produced against the fetus' HLA molecules (Marsh *et al.*,2000). The variation of HLA molecules can also be identified using the specific different CD8+ T cells that are targeted against HLA-peptide complexes on the surfaces of infected cells (Marsh *et al.*,2000). The antibody-HLA or peptide-HLA complexes are then degraded *in vitro* and sequenced in order to be classified and named. The HLA allele nomenclature is commonly represented by the locus (i.e. A, B or C), serological type (first two digits, eg. HLA B58) followed by another two digits which represent non-synonymous nucleotide substitution differences between alleles of the same serological type, e.g. HLA*B5801 and B5802. In some cases, closely related HLA alleles which only differ by synonymous nucleotide differences observable under high resolution are assigned one more digit (Marsh *et al.*,2000).

The basic structure of an HLA molecule consists of two identical heavy protein chains of about 45kd each (blue and purple beta sheets in Figure 2.6.2) and two identical light chains of about 25kd each (light blue and light purple alpha chains in Figure 2.6.2) (Marsh *et al.*,2000). The N-terminal half of each heavy chain is bound to one light chain via a single disulphide bond at the C-terminal end of the light chain. After this bond, towards the C-terminal of the heavy chain, there is a hinge that joins the two heavy chains together through another disulphide bond. The C-terminal regions of both the heavy and light chains are conserved. The N-terminal region of each heavy chain, situated close to the N-terminal of the light chain and the corresponding N-terminal half of the light chain, is variable. Within these variable regions are found hyper-variable regions. Therefore the HLA polymorphisms result from sequence

differences mostly found at the hyper-variable regions (Figure 2.6.2). The hyper variable sites are within the binding pockets of the HLA binding groove, the groove being formed by the variable regions of both the light and heavy chains.

The binding pockets known as the B and F pockets (shown by spherical representation of amino acid residues in Figure 2.6.2) form the strongest contact and binding to the antigenic peptides in a highly specific manner. It has been shown that the B and F pockets usually bind to the second and C-terminal positions of an antigenic peptide (Madden *et al.*,1992). Minimal amino acid variations are allowed at these major 'anchor residue' sites of peptides and less stringency is observed at the other positions within the peptide. However, successful binding is also dependant on the overall binding energy and morphology of the HLA-peptide complex. The extensive diversification of the binding groove region as a result of varying amino acid side changes enables the HLA molecules to recognize and present a diverse population of antigens thus enabling the immune system to combat various infecting pathogens. HLA molecules with closely related sequences and that share some common anchor residue preferences have been grouped under the same supertypes (Sette *et al.*,1999). For example, currently, alleles of the HLA-A and B loci, which have been extensively characterized have been classified into nine major supertypes, namely A1, A2, A24, A3, B27, B44, B58, B62, and B7 (Sette *et al.*,1999).

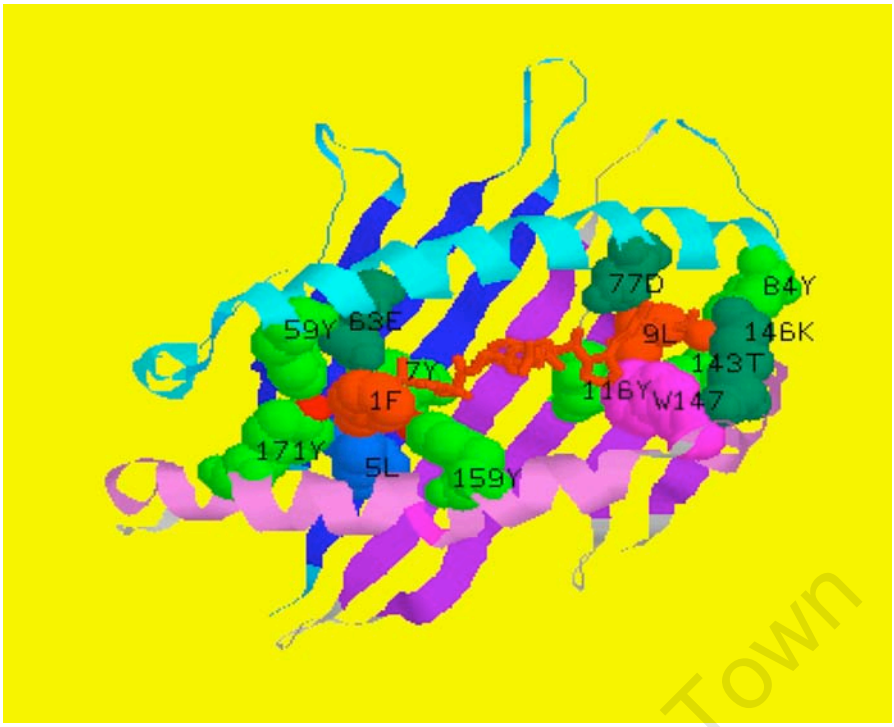


Figure 2.6.2: HLA A*0201 secondary structure obtained from *http://www.cryst.bbk.ac.uk/PPS2/projects/vun/H2kb_P1P9.gif*

Beta sheets represent heavy chains and alpha chains are the light chains. The residues of the B and F pockets of the HLA binding groove are represented in spheres, and the positions with respect to the amino acid sequence of the molecule are given for those that contact the peptide. The peptide bound in the groove is shown as a red chain with the N- and C-terminal regions of the peptide in close contact with HLA residues.

2.6.3 Detection of CTL epitopes

Various methods have been developed to determine peptides that are bound by HLA molecules. Both biochemical and computational approaches have been used. The central point in understanding the host-viral relationships is the evaluation of immune responses against HIV-1 and their impact on both the virus and the host. The primary step is the determination of specific peptides on the virus sequence that are targeted by specific HLA molecules. The most accurate methods for determination of CTL epitopes is the use of biochemical methods such as the frequently used interferon-γ ELISpot assays. In these assays, antigenic peptides that bind to peripheral blood mononucleocytes (PBMCs) isolated from infected individuals are determined

(Schmittel *et al.*,1997). The isolation of HLA-peptide complexes from cultured B-cells is also common (Marsh *et al.*,2000). The HLA-peptide complexes are denatured and peptides screened through High Performance Liquid Chromatography or the Edman Degradation technique (Marsh *et al.*,2000).

Although biochemical methods are very accurate and provide complete epitope sequences, they tend to be slow and expensive. Computational methods can be useful in speeding up the determination of HLA binding regions of protein sequences. Anchor residues as well as binding affinity are the key components of computational methods for predicting CTL epitopes. The anchor residue motifs themselves aid in identifying potential target sites of HLA alleles thereby reducing candidates that can be further analysed for optimal binding to HLAs either biochemically or computationally. Methods that determine the binding energy between HLA and candidate peptides can further reduce the quantity of biochemical tests by identifying good binders. Epitopes that bind with highest affinity are likely to be successfully transported from the ER to the cell surface of infected cell and elicit strong immune responses *in vivo*. A positive correlation between binding affinity and immunogenicity of a peptide has been observed in previous studies confirming the importance of binding affinity in predicting HLA binding epitopes (Sette *et al.*,1994).

The presence of anchor residues in query peptides is also used in combination with the determination of binding affinity between a HLA molecule and a peptide (Altfeld *et al.*,2001; Altuvia *et al.*,2004; Schueler-Furman *et al.*,2000). In this case, query peptides containing known anchor residues recognised by a HLA are analysed for the affinity of binding to the HLA molecule. There are a number of methods used to determine the HLA-peptide binding affinity. In one example, the half time for the dissociation of a HLA-peptide complex is used to measure the strength of binding and the longer the time, the higher the affinity (Parker *et al.*,1994). IC_{50} values, i.e., the concentration of a peptide that inhibits the binding of a standard peptide to an HLA molecule by up to 50%, from competitive binding assays, are also used (Altfeld *et al.*,2001). Yet other methods rely on the pairwise binding potentials of amino acids or binding free energy between a HLA crystal structure and a peptide (Altuvia *et al.*,2004; Logean *et al.*,2002).

Some computational approaches require the knowledge of known HLA-binding peptides as well as those that do not bind in order to classify candidates. One such example is the use of machine learning where algorithms learn the binding characteristics of known HLA binders and non-binders before testing a query peptide (Jojic *et al.,*2006; Lata *et al.*,2007). Query peptides are characterized by how best their binding characteristics resemble that of binders and non-binders. Another example of a frequently used computational approach is the use of position specific scoring matrices (PSSM) (Flower,2007). In PSSM, query peptides are compared against an alignment of epitopes known to bind to a HLA and each position of the query peptide scored based on its similarity to the alignment. Peptides that meet the required threshold scores are assigned as epitopes. A similar approach but based mostly on the presence and high scoring of anchor residues that were derived from natural HLA ligands is also used for epitope predictions in the SYFPEITHI database (Rammensee *et al.*,1995) .

2.6.4 Prediction of HLA anchor residue motifs

Amino acid residues located at the second and C-terminus of epitopes that bind to the B and F pockets of HLA molecules are known as anchor residues. They form the strongest binding points between the HLA and the peptide. As described in section 2.6.2, each HLA has high selectivity at the B and F pockets when binding to epitopes. Therefore, anchor residues are the most important components of peptides and first determinants of HLA-peptide binding. The sequence patterns that contain only the anchor residues are known as 'anchor residue motifs'. An example of an anchor residue motif is x-[P]-x-x-x-x-x-x-[WFL] for the B*5301 allele (Yusim *et al.,*2003). The 'x' represents any amino acid at that position of the 9-mer peptide. Phenylalanine is the anchor residue at position 2 of the peptide. For this HLA allele, each of the three residues given in square brackets at the C-terminus position of the peptide is preferred.

Anchor residue motifs are important in preliminary predictions of HLA binding peptides and have become important in studies of immune escape mutations at the anchor sites of peptides. Various methods are being used to predict anchor residue motifs of HLA alleles. The main approaches that are used to predict anchor residue

motifs are (i) the identification of residues that frequently occur at anchor sites of known epitopes found bound to HLA molecules endogenously e.g. (Boisgerault *et al.,*1996; Falk *et al.,*1991; Marsh *et al.*,2000; Seeger *et al.,*1999b)**.** Another prediction approach is (ii) the inference of a motif to a new HLA based on the similarity of residues and side chains at the binding groove to that of a HLA allele with a characterized motif e.g. (Honeyborne *et al.*,2006; Seeger *et al.*,1999a). With development in computer application in biology, (iii) structure-based methods for which the binding energy between a peptide and the HLA binding pockets have also been developed e.g. (Altuvia *et al.*,2004).

In the latter approach, the amino acids that confer strongest binding energy at the anchor positions are assigned as the optimal anchor residues for the specific allele. Anchor residue motifs that are predicted using the first two approaches can however, be biased by the residues that happen to occur in the viral sequences in which the motif was inferred. For instance, the HIV-1 subtypes have an amino acid sequence difference of up to 35% in the envelope proteins (Levy,2007; Wain *et al.,*2007). If such regions where the greatest sequence diversity exists are used to predict a motif, it is highly likely that the corresponding motif targeted by the same HLA allele at a sequence of another subtype could be different. When such motifs are used to predict binding regions on a sequence from a different source, optimal epitopes can be missed and be assumed to have been lost due to the difference in an anchor residue, which in fact does not hinder HLA binding. The third approach however is more universal in that the binding is predicted based on first principles of amino acid and protein interaction, i.e., the binding affinity between the HLA crystal structure and a peptide sequence.

2.7 CTL Immune Response and HIV Disease Progression

The CTL immune response is known to be important in anti-HIV-1 immune responses as well as in determining disease progression in infected individuals. In most individuals, viral load decreases drastically during early infection after the first activation and mounting of CTL immune responses. This can be followed by a latent stage in some individuals with protective immune responses where the viral load

decreases to very low levels. Final stages in HIV-1 infection have increasing viral load and low levels of CD4+ T-cells, which are the main targets of HIV-1 infection (McMichael *et al.*,2001). This marks the development of AIDS and typically death, in the absence of antiretroviral treatment. Host immune responses mediated through CTL and HLA alleles may varyingly exert selection pressure on the targeted regions of the HIV-1 sequences during infection (Carlson *et al.*,2008; Carrington *et al.*,2003; Rousseau *et al.*,2009). Most of the HLA alleles that exert strong selection pressure are from the B locus. The HLA-B alleles have also been found to be frequently involved in mediating CTL immune responses against HIV-1 and in determining disease progression (Bihl *et al.*,2006; Kiepiela *et al.*,2004). Individual HLA alleles within a single supertype can lead to different effects on the HIV-1 sequence and disease progression. Some HLA alleles are associated with either delayed or rapid progression to AIDS in infected individuals.

2.7.1 Long term non-progression in HIV-1 infection

Some HLA alleles have been found to mediate CTL responses that result in protection and delayed disease progression to AIDS. HLA B*5801 is one of the alleles first to be recognized for its association with long-term non-progression (LTNP) (Goulder *et al.*,2004). It is frequent in African populations (Marsh *et al.*,2000) and targets peptides within the Gag protein sequence (Brumme *et al.,*2008a; Goulder *et al.,*2004). HLA-B2705 targets a peptide within conserved regions of the Gag protein and has been found to be associated with a reduction in viral load and protective immunity (Kelleher *et al.*,2001). HLA B4 allele was also previously found to be associated with a prolonged disease-free period in HIV infected individuals bearing homozygosity for this genotype (Flores-Villanueva *et al.*,2001).

2.7.2 Rapid progression to AIDS

On the other hand, some HLA alleles tend to be associated with rapid progression to AIDS. The specific mechanisms by which the presence of these genotypes causes rapid progression are not yet well understood. Examples of such alleles include HLA A*2301 which was found to be frequent in individuals who progressed rapidly to AIDS (Chen *et al.*,1997; MacDonald *et al.*,2000). HLA A2301 is very frequent in

Africans and African-American populations and targets peptides located in the Nef protein (Marsh *et al.*,2000). B*5802, with the same serological group as B*5801 discussed in the previous section, is on the contrary associated with rapid disease progression in HIV-1 infections (Ngumbela *et al.*,2008). The allele is also frequently present in persons of African origins (Marsh *et al.*,2000). The B*35 genotype, frequent in Caucasoid and Amerindian ethnic groups, has been found to cause rapid progression to AIDS in Caucasians (Carrington *et al.*,1999).

2.8 Challenges in Designing anti-HIV Vaccines

Despite numerous studies that have been done on the HIV-1 pathogenesis in humans as well as development of a number of anti-HIV-1 therapeutics, there is still no cure for AIDS. The current hope lies in vaccines. However, vaccines that completely destroy the virus in humans have not been successfully developed yet (Singh,2006). Ideally, such a vaccine should boost immune responses that target regions of the sequence that are highly conserved and cannot easily mutate, and also prevent further replication of the virus.

2.8.1 Host and pathogen diversity

It is more likely to find similar immune responses targeting conserved regions within a single population of individuals of the same ethnic group who share common HLA genotypes. However, individuals of different ethnic origins usually have large genetic differences (Cao *et al.*,2001) such that the frequency of specific immune responses against HIV-1 may differ between the groups (Kawashima *et al.*,2009; Moore *et al.*,2002; Piontkivska *et al.*,2004; Travers *et al.*,2005). This implies that the selection pressure exerted on the viral sequences can differ remarkably between the different ethnic populations. The varying selection pressures therefore cause variations in immune escape mutations and the evolution of the virus sequences thus minimising the chances of finding identical and conserved HIV-1 regions across different populations. Even though there are highly conserved immunogenic regions in Nef and Gag proteins (da Silva *et al.,*1998; Masemola *et al.,*2004b) of different subtypes, the specific immune responses that target these regions in the different populations may

not all have a strong protective effect. Even individuals of the same ethnic group have different genotypic composition except for identical twins (Draenert *et al.*,2006), hence not all individuals within an ethnic or population group can have similar protective immune responses. Designing a vaccine for each individual or the few with similar desirable immune responses AIDS would be ideal but is very expensive and not feasible for the whole HIV-1 infected population worldwide. Therefore, an ideal vaccine reagent should comprise virus peptides that are conserved across different HIV-1 subtypes and elicit protective immune responses across at least most individuals in the different ethnic and population groups. The possibility of such a vaccine lies in the cooperation of researchers across the world and sharing of information on HIV-1 pathogenesis and anti-HIV immune responses.

2.8.2 Research limitations

Indeed, sharing of information has been made possible by the development of public databases for data storage. A good example is the Los Alamos HIV databases (http://www.hiv.lanl.gov/content/), where information such as HIV-1 sequences, epitopes, HIV specific anchor residue motifs for HLA allele binding, HLA supertypes and binding specificities as well as related analysis tools are provided. A lot of the data in the Los Alamos databases have been collected from various published studies. The curators of the database are therefore faced with a challenge of linking data isolated from the same infected individual (Learn, Jr. *et al.*,1996). Obviously, this cannot be possible if the source authors use different patient codes in subsequent studies of the same individual. This obviously causes limited amounts of longitudinal data that is publicly available thus limiting research studies that can monitor the evolution of specific sequences overtime.

Problems that may be faced by researchers include the inadequacy of the data in equally representing all infected populations and HIV-1 subtypes across the world. For example most studies that predict CTL epitopes have been carried out on HIV-1 subtype B which is most frequent in the developed countries which are better equipped to carry out more research analysis. Such data may not provide accurate information for some studies that focus on other subtypes. For example, epitope sequences located in regions that vary between subtypes may not be applicable to

studies of other subtypes if the provided sequences were obtained from a single subtype.

An example of another very useful database is dbMHC which provides the HLA anthropology data (Meyer *et al.*,2007). Estimated HLA frequencies from major population groups across the globe are provided. This resource however has its limitations too. In the first, the HLA allele frequencies are based on small population samples e.g. the Zulu tribe, a majority population group in South Africa is represented by a sample size of 290 individuals (Hammond *et al.*,2007). Also, not all HLA alleles that are found in these populations are provided in the database hence absence of an HLA in the database under a particular group does not necessarily mean that the HLA genotype is absent in the population. In some cases, the allele frequency is provided as belonging to a country as a whole even though people of different ethnic groups co-habit in the particular country.

Another major problem with the currently available data is that the HLA genotypes of individuals are provided separately from the HIV-1 autologous sequence data isolated from the same individual. Datasets comprising of both HLA background and autologous sequences of the virus are rarely found in public databases and only in small sample sizes. Large datasets of HIV sequences providing the corresponding genotypes of the corresponding patients would aid in more accurate analysis of HIV-1 evolution in relation to immune responses at larger population levels. In addition to the need for more studies to characterize host genotypes, there is a need for accuracy, consistency, cooperation and rapid depositing of data in order to speed up studies directed towards designing protective vaccines and other anti-HIV-1 therapeutics.

Chapter 3

The extent of purifying selection pressure acting on synonymous sites of HIV-1 sequences

Foreword

This chapter describes an analysis that was published in an international journal and is listed at the beginning of the thesis (Ngandu et al.,2008).

3.1 Summary

The ratio of nonsynonymous to synonymous substitution rates (dN/dS) is used to infer selection pressure acting on protein-coding sequences. A dN/dS >1 resulting from dN being significantly higher than dS indicates positive selection. Such an inference of selection pressure assumes that dS is a measure of the neutral rate of evolution against which dN, a measure of selection pressure acting at the amino acid level, is compared. However, purifying selection acting directly on the nucleotide sequence can lower dS, resulting in an underestimate of the neutral rate of evolution. This can cause dN/dS to be greater than 1 when there is actually no diversifying selection acting on the amino acid level and thus false inference of positive selection. Despite the fact that HIV-1 has a number of sequence motifs that function at the nucleotide level, the extent of purifying selection pressure acting on synonymous sites of the nucleotide sequence across the genome has not been evaluated to date. Yet dS is continually being used as a measure of neutral evolution in positive selection analysis of HIV-1 sequences. Here, site-to-site variation in dS across coding regions of the HIV-1 genome was modeled and found to vary significantly within and between genes. Fourteen regions of the nucleotide sequence with known functions appeared to be under strong purifying selection pressure with significantly lowered synonymous mutations. These included an exonic splicing enhancer, the rev-responsive element, the poly-purine tract and a transcription factor-binding site. A further five highly conserved regions were located within known functional domains. An additional four regions with un-characterized functions, possibly novel, located in *env* and *vpu* were also conserved. The coordinates of genomic regions with significantly lower synonymous substitution rates, which are putatively under the influence of strong purifying selection pressure

at the nucleotide sequence level are provided for consideration or exclusion from studies of positive selection acting on HIV-1 coding regions.

3.2 Background

Biological sequences that code for proteins are known to evolve either neutrally or in response to selection pressure exerted upon them. Several statistical models of codon evolution have been developed for the purposes of analysing the evolution rate of codons in protein-coding sequences of different species and pathogens such as viruses (Goldman *et al.*,1994; Muse *et al.*,1994; Nielsen *et al.*,1998; Yang *et al.*,2000). The primary application of these models has been the detection of evidence of diversifying selection acting on protein coding DNA sequences. Within maximum likelihood or Bayesian frameworks these models can be used to identify specific sites at which adaptive mutations have occurred. In the context of virus infections this information can be especially useful for identifying immune escape and drug resistance mutations (Lemey *et al.*,2007; Nielsen *et al.*,1998; Seoighe *et al.*,2007).

Selection pressure acting on protein-coding sequences is frequently inferred by comparing the rate of non-synonymous substitutions per non-synonymous site (dN) to the rate of synonymous substitutions per synonymous site (dS). The dN/dS ratio is often represented by the symbol ω. Diversifying selection can be inferred when ω is greater than one. However, this approach assumes that synonymous substitutions are neutral and that the synonymous substitution rate therefore approximates the neutral rate of evolution. Several methods exist to determine whether there is evidence that ω is greater than one i.e., the gene is evolving under diversifying selection at a subset of sites in a protein-coding gene and to identify the sites within the gene at which diversifying selection occurs (Choisy *et al.*,2004; de Oliveira *et al.*,2004; Nielsen *et al.*,1998; Yang *et al.*,2000; Zanotto *et al.*,1999).

Kosakovsky Pond & Muse reported that coding sequences from a wide range of taxa, including HIV-1, show strong evidence of variation in the rate of synonymous substitution across coding regions (Kosakovsky Pond *et al.*,2005a). Therefore, the assumption that synonymous substitutions are fixed at a constant rate and provide a good estimate of the neutral rate of evolution may not always hold true. There are two

possible causes of synonymous rate variation. Firstly, if synonymous substitutions are indeed neutral, variation in the mutation rate can cause the synonymous substitution rate to vary. In such a case, it is possible to include a varying synonymous substitution rate in the codon models of evolution and inference of positive selection from comparison of the local synonymous and nonsynonymous substitution rates remains feasible. Secondly, selection pressure acting to preserve functions that are encoded at the nucleotide level can cause a variation in synonymous substitution rate, which deviates from the neutral rate. In such a scenario, even a comparison of local nonsynonymous and synonymous substitution rates cannot be used to infer positive selection because the synonymous substitution rate has deviated from the neutral rate of evolution and the standard approach of inferring the action of diversifying selection when $\omega > 1$ becomes invalid.

In the second scenario, failure to model variation in synonymous substitution rate will result in an overall underestimate of the neutral rate of evolution. This undermines the validity of the inference of selection, because nonsynonymous substitution rates are compared against a rate, which is no longer a good estimate of the neutral rate, and this is likely to result in inference of diversifying selection at a proportion of the sites that are actually evolving neutrally. Indeed, as the number of taxa increases, we expect a greater proportion of amino acid sites that are evolving neutrally to be classified as diversifying selection sites in this scenario. Alternatively, if the synonymous substitution rate variation is modeled and selection inferred when dN is greater than the local dS rate then we expect a very high probability of false inference of selection at codons where the synonymous positions happen to be functionally important and conserved, and the nonsynonymous positions are neutral. Thus, in general, in a codon-based method, analysis of selection is unreliable when there is purifying selection acting to preserve functions at the nucleotide level. An example where an elevated ω was attributable to purifying selection acting on synonymous sites was reported by Hurst & Pal (Hurst *et al.*,2001).

Several examples of sequence motifs within protein-coding sequences that are expected to be under purifying selection at the nucleotide level are known in HIV-1, many of which are involved in regulating gene expression. In addition to functions of the nucleotide sequence given in section 2.2.5, some regions such as the LTR have

conserved RNA secondary structures (Pereira *et al.,*2000; Wilkinson *et al.,*2008). Some functionally important regions of the rev-responsive element (RRE) have also previously been found to be conserved at the nucleotide sequence level, presumably the result of purifying selection pressure to preserve this function (Phuphuakrat *et al.*,2003). If the functional sites of the nucleotide sequence are important for viral viability, then they are expected to be preserved by purifying selection resulting in dS rates that are lower than the neutral rate of evolution.

While it represents a significant challenge for studies of selection acting on the HIV-1 amino acid sequence, the variability in the synonymous substitution rate may also provide useful information about previously unknown sequence motifs within the coding fraction of the HIV-1 genome that function at the nucleotide level. Although some variability can be explained by a variable mutation rate, the identification of regions of very high conservation that cannot be explained by selection acting on the amino acid sequence or by known motifs that function at the nucleotide level has the potential to highlight novel functions encoded in the HIV-1 genome.

Here, an existing model of codon sequence evolution (Kosakovsky Pond *et al.*,2005a) was used to provide the first complete overview of site-to-site variation in synonymous substitution rate across all coding regions of the HIV-1 genome and identify selection pressures likely to be driving this variation. This model allows dN and dS to vary independently across sites, ensuring that the estimated dS values reflect selection pressure acting upon the nucleotide sequence and not at the amino acid level. It is worthwhile to distinguish between selective pressure acting at the nucleotide level, affecting both synonymous and nonsynonymous changes, and at the amino acid level, affecting nonsynonymous changes only. Unfortunately, quantifying the relative contributions of nucleotide and amino acid level effects on nonsynonymous changes is highly sensitive to model assumptions. Therefore, this study was restricted to the analysis of synonymous changes at the synonymous sites of the nucleotide sequence only. Recombination breakpoints were taken into account, as explained in the methods section, in order to avoid biased estimates that can result from fitting phylogenetic models that do not take recombination into account (Anisimova *et al.*,2003; Shriner *et al.*,2003). This chapter provides all sites of the HIV-1 genome where purifying selection was inferred directly on the nucleotide

sequence and which are likely to cause a substantial reduction in the synonymous substitution rate. These are provided with respect to the HXB2 reference strain, to enable other researchers to mask these regions from their analyses of positive selection acting on HIV-1 genes.

3.3 Materials and Methods

3.3.1 Sequence data

Nucleotide sequence alignments consisting of HIV-1 Group M subtype reference sequences were downloaded from the Los Alamos database for each gene of the HIV-1 genome (Leitner *et al.*,2005). Each alignment had at least one sequence from each of the 11 non-recombinant HIV-1 group M subtypes A1, A2, B, C, D, F1, F2, G, H, J and K [Genbank: AB253421, AB253429, AF004885, AF005494, AF005496, AF061641, AF061642, AF067155, AF069670, AF075703, AF077336, AF082394, AF082395, AF084936, AF190127, AF190128, AF286237, AF286238, AF377956, AF484509, AJ249235, AJ249236, AJ249237, AJ249238, AJ249239, AY173951, AY253311, AY331295, AY371157, AY371158, AY423387, AY612637, AY772699, DQ676872, DQ853463, K03454, K03455, U46016, U51190, U52953, U88824, U88826]. The total number of sequences per gene alignment ranged from 32 to 37. All regions encoding amino acids in more than one frame were identified and regions judged by eye to be unreliably aligned, i.e., positions 6544-6595, 6700-6715, 7318-7375 of the *env* gene region, were excluded from the analysis. The HIV-1 genome map and sequence annotations available from the Los Alamos database were used to identify the regions of the genome that encode proteins in a single reading frame (see Table 3.3.1) (Leitner *et al.*,2005).

Gene	Non-overlapping region used	Position on genome	Number of sequences
Env	88 – 2154	6313 – 8379	37
Gag	1 – 1295	790 – 2084	37
Nef	1 – 621	8797 – 9417	32
Pol	211 – 2955	2296 – 5040	37
Tat	22 – 138	5851 – 5967	37
Vif	58 – 519	5098 – 5559	37
Vpr	61 – 273	5620 – 5832	37
Vpu	1 – 162	6061 – 6222	37
Rev	total overlap		
Genome	Includes overlapping sites	790 – 9417	36

Table 3.3.1: Summary of HIV-1 reference sequence data used**, t**he HXB2 genome numbering is used.

Recombination breakpoints in each alignment were identified using the GARD (Genetic Algorithm for Recombination Detection) algorithm implemented in the HyPhy (Hypothesis testing using Phylogenies) package (Kosakovsky Pond *et al.*,2005d; Kosakovsky Pond *et al.*,2006). Evidence of recombination was detected in all genes except *tat*, *vpr* and *vpu*. GARD outputs both an alignment showing the positions of recombination breakpoints and separate tree topologies for each of the sequence alignment segments bounded by these breakpoints.

3.3.2 Synonymous substitution rate estimation

Synonymous substitution rates were estimated using a version of the MG94 codon substitution model (Kosakovsky Pond *et al.*,2005a; Muse *et al.*,1994). In the version used here, synonymous substitution rates were allowed to vary between sites. The Dual Model which allows dS to vary independently of dN was used with three

discrete categories for each rate, i.e., three dN/dS classes are estimated based on the data. The selected models were ran using a HyPhy batch script for analysis of selection acting on recombining sequences, which was developed by fellow researchers (Scheffler *et al.*,2006). This method uses separate tree topologies for each partition (between recombination breakpoints) of the sequence alignment while keeping the rest of the model parameters fixed across all partitions. Mean dS values over sliding windows was calculated over three neighboring codons and plotted to identify regions with low synonymous substitution rates.

An alignment of HIV-1 subtype C *gag* sequences from recent HIV-1 infections [Genbank: DQ792982-DQ793045] described in one of the publications listed in this thesis (Ngandu *et al.*,2007) was used to further assess the impact of conservation acting on synonymous substitutions on inference of positive selection. Positive selection was inferred using model M2a of Yang and colleagues (Yang *et al.*,2005b), taking recombination into account (Scheffler *et al.*,2006).

3.3.3 Simulations

Simulations were caried out in order to validate whether the observed dS rates per site were significantly different from what would be expected by chance. HyPhy was used to generate simulated data under a neutral model with trees generated from the original alignments (or the tree from the largest un-recombined region for alignments where recombination was detected). The same sequence alignments used as input in the initial analysis were used and one hundred simulated datasets were generated for each alignment. Each simulated dataset was then analyzed using the Dual Model as described above. For each gene the minimum value of mean dS across all sliding windows of three neighboring codons, in all of the one hundred simulated datasets, was used as a conservative threshold to identify windows of reduced dS in the observed data. This stringent threshold and a less stringent one that included 95% of the values inferred from the simulated data are shown in the sliding window plots.

3.4 Results

3.4.1 Evidence for site-to-site synonymous rate heterogeneity

For all genes the Dual Model (Kosakovsky Pond *et al.*,2005a), which allows independent variation of dS and dN had a much better fit to the data than a model with constant dN and dS (referred to as the Constant model in Table 3.4.1a) or than a model in which only dN varied across sites (the Nonsynonymous model in Table 3.4.1a). The Akaike Index Criterion is a score that gives a measure of how a model fits to the data in comparison to other models being tested. The score is calculated using a statistical measure of the maximum likelihood of the model given the data and taking into account the number of parameters to be fitted (Akaike, 1987). Too many parameters can cause overfitting thus a model with a low AIC, i.e., a high goodness of fit yet using the least number of parameters is preferred. Consistent with previous reports (Kosakovsky Pond *et al.*,2005a; Lemey *et al.*,2007), evidence of variation in synonymous substitution rates was observed within and across HIV-1 genes (Figures 3.4.1a and b). There was significant variation between genes (p value = 2×10^{-7}, from Levene's test) with the least site-to-site variation in dS observed in *vpu* and the most in *vpr* followed by *nef* and *env* (Figure 3.4.1b). The variance of dS gives an indication of the extent of site-to-site synonymous rate heterogeneity within the different genomic regions (Table 3.4.1b). Evidence of strong purifying selection acting directly on the nucleotide sequence was found at twenty-three nucleotide motifs across the HIV-1 genome, based on pvalues obtained using the simulations. The nucleotide motifs found to be conserved (all listed per gene in Table 3.4.1b) either had previously reported known functions, were within regions with known functions or novel with no record of known function in the literature. In addition to motifs that are highly conserved i.e., with mean dS lower than that obtained from all simmulated datasets, thus with a pvalue <0.0001, sites that were observed to be conserved at the 95% significance level are also given in Table 3.4.1b.

Gene regions	Akaike Information Criterion index (AIC) per Model		
	Dual	**Nonsynonymous**	**Constant**
Gag	23862.95	23963.83	25889.10
Pol	41504.20	41668.99	45642.96
Vif	9794.65	9843.41	10502.37
Vpr	9581.57	9692.29	10869.78
Tat	2834.49	2878.99	3110.23
Vpu	11701.57	11832.41	12938.54
Env	45201.62	45422.72	48658.16
Nef	13984.66	14061.46	14986.75

Table 3.4.1a: AIC model selection index to show how different models fit to the data, The Dual Model, which allows dS to vary across sites has the lowest AIC (best fit to the data) for all 8 genes.

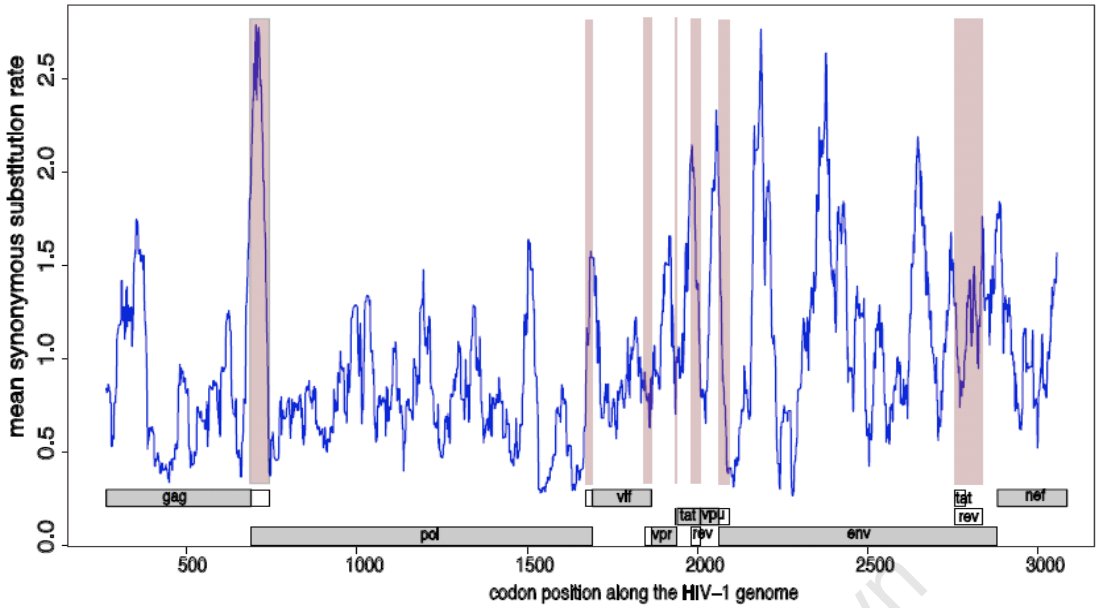


Figure 3.4.1a; HIV-1 genome-wide plot of mean nonsynonymous substitution rates. A 30 codon sliding window was used. Regions coding for proteins in more than one frame are shaded in pink. The frames that were used in each region are shown in grey rectangles, with frame 1 at the top and frame 3 at the bottom.

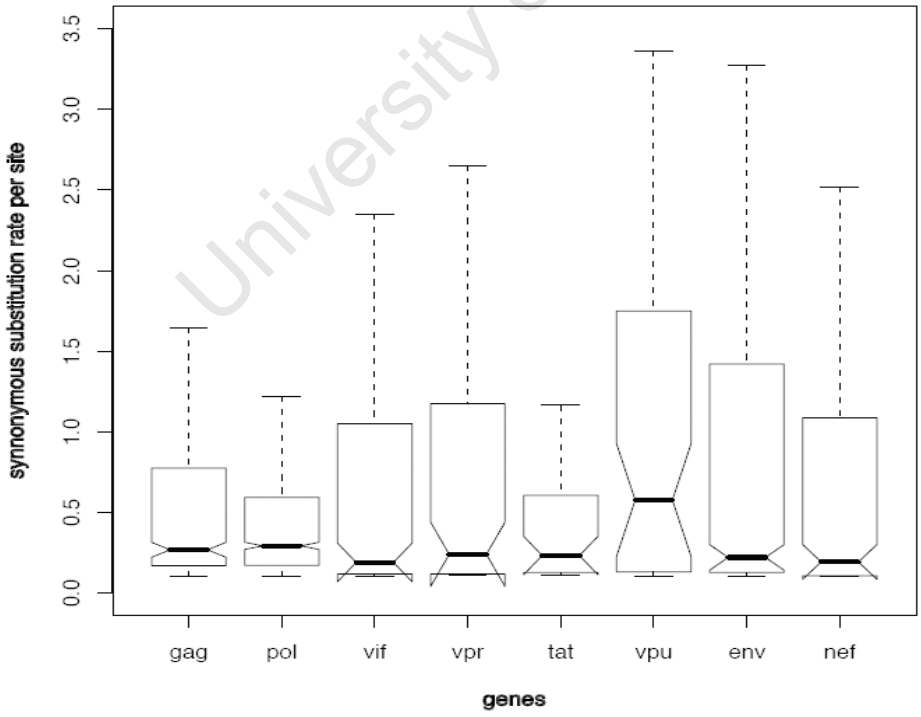


Figure 3.4.1b: Box-and-whisker plot showing variation of dS values per gene. The values are from non-overlapping regions of HIV-1 genes.

Gene	dS Variance	Overlapping regions	Highly conserved (most stringent cutoff)	Other sites conserved at 0.05 significance
gag	2.1	2086 – 2295 (gag/pol)	793 – 807 898 – 903 985 – 996 1309 – 1314	821-823 1036-1038 1810-1812 1831-1833 1969 – 1974 2002-2004 2023 – 2025
pol	1.8	5041 – 5097 (pol/vif)	4092 – 4094 4764 – 4790 4864 – 4866 4926 – 4937	2490-2495 2850 – 2858 3252 – 3257 3304 – 3312 3867 – 3881 3966 – 3971
vif	2.7	5560 – 5619 (vif/vpr)	-	-
vpr	3.9	5833 – 5850 (vpr/tat)	5769 – 5777 5794 - 5805	-
tat	2.9	5968 – 6060 (tat/rev)	5855 – 5863 5957 – 5968	-
vpu	1.4	6223 – 6312 (vpu/env)	6101 – 6106 6143 – 6151 6167 – 6178	-
env	3. 2	8380 – 8796 (env/rev)	7656 – 7667 7834 – 7842 8349 – 8354 8376 – 8378	7077 – 7082 7125 – 7130 7629 – 7634
nef	3.6	-	9067 – 9086 9087 – 9093 (nef/LTR) 9183 – 9192 (nef/LTR) 9391 – 9399 (nef/LTR)	8869 – 8874 8887 – 8892 9121 – 9126 9235 – 9237

Table 3.4.1b Regions of the HIV-1 sequence that should be considered for exclusion in positive selection analysis studies, Co-ordinates are adapted from the HXB2 numbering system. Dashes indicate that there are no sites within that category for a particular gene.

3.4.2 Purifying selection pressure in known functional motifs

Of the twenty-three highly conserved regions, fourteen (marked in black in Figures 3.4.2a through to 3.4.2g) coincided exactly with well-characterized functional motifs. Sequence logos illustrating the conservation in each of these fourteen significantly conserved regions with known specific function are shown in Table 3.4.2. Four of the significantly conserved regions with known function were in the *nef* gene and coincided with the poly-purine tract (PPT), integrase attachment site, negative regulatory element and Ets-1 transcription factor binding site (Figure 3.4.2a) (with HXB2 coordinates 9066-9083, 9084-9091, 9183-9192 and 9391-9399 respectively). The PPT precedes the start of the LTR and is known to associate with the 3' LTR, serving as a primer for the initiation of HIV-1 plus strand DNA replication (Luo *et al.*,1990; Miles *et al.*,2005; Quinones-Mateu *et al.*,1998; Rausch *et al.*,2007). A previous detailed RT RNase-H binding analysis revealed that priming of the plus strand occurs specifically at the 3' end of the PPT, at the "GGGGGG" motif (Powell *et al.*,1996; Pullen *et al.*,1993; Rausch *et al.*,2004). The region adjacent to the 3' end of the PPT was also highly conserved. This region corresponds to the start of the 3'LTR and contains the cleavage site of the PPT by RNAse H as well as the start of integrase attachment region for the integration of the viral genome into the genome of the host (Brown *et al.*,1999; Masuda *et al.*,1998; Miles *et al.*,2005; Rausch *et al.*,2004). Two codons in the central region of *nef* were highly variable, one dominated by G-to-A mutations in a sequence context consistent with APOBEC-induced hypermutations when compared to the Group M ancestral sequence (Rose *et al.*,2000) (see Figure A3.4.2a in the Appendix) and the second had a high rate of synonymous and nonsynonymous substitutions (labeled "G-A" and "var" in Figure 3.4.2a respectively).

Two regions in *vpr*, a 3' splice acceptor site (Kammler *et al.*,2006) and an RNase-V1 cleavage site (Jacquenet *et al.*,2001a; Jacquenet *et al.*,2001b) were also conserved (labeled ssa3 and rnase respectively in Figure 3.4.2b) at positions 5759-5777 and 5794-5805 respectively. An additional two highly conserved regions were observed within the *tat* gene, one containing an exonic splicing silencer, ESS2 between nucleotide positions 5855 and 5863 and the other at the 3' splice acceptor site A4b

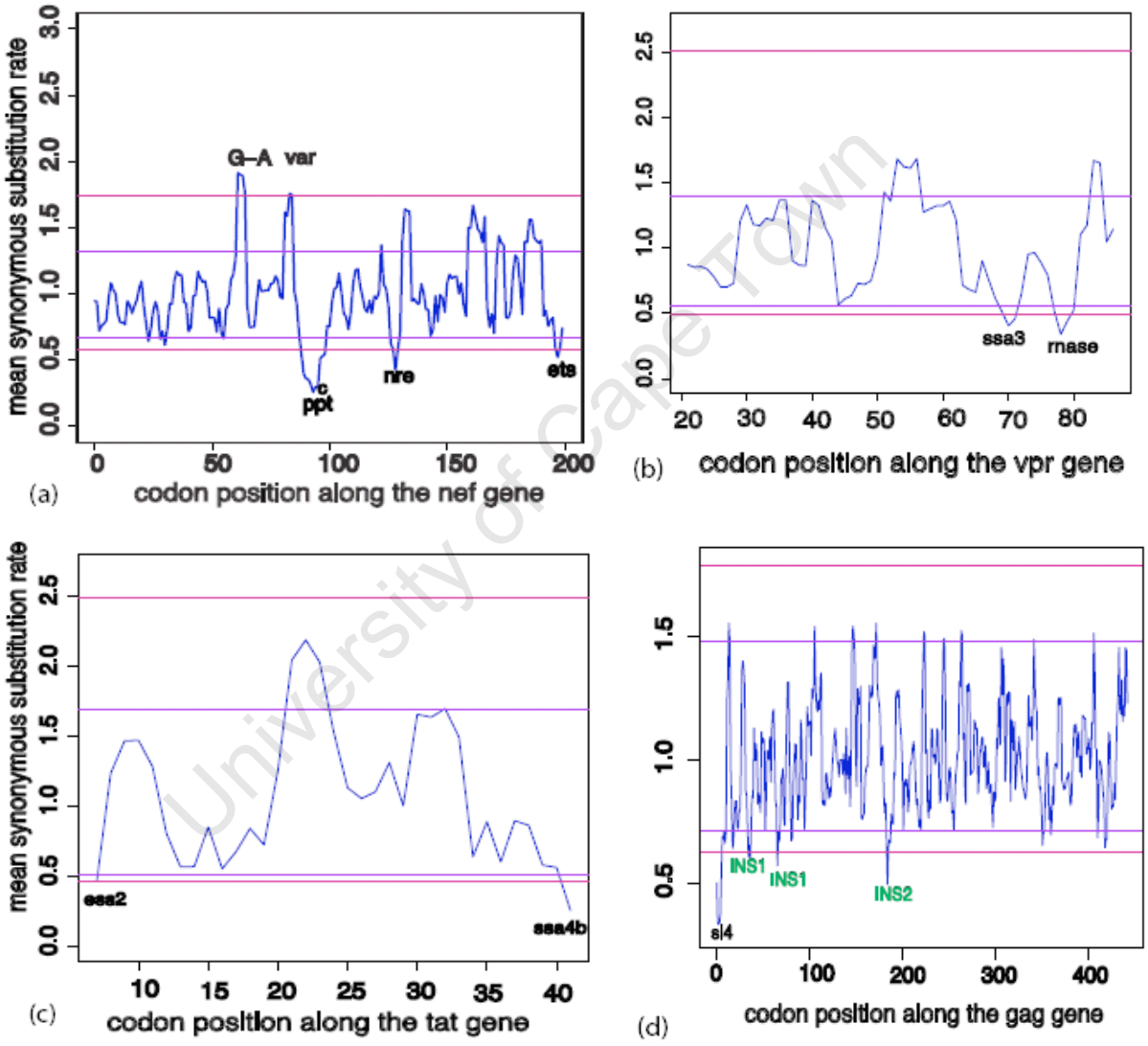
located at positions 5957 to 5968 (ess2 and ssa4b respectively in Figure 3.4.2c) (Kammler *et al.*,2006).

These fourteen regions with known functions included one region consisting of fifteen nucleotides following the *gag* start codon, positions 793-807 of HXB2. This forms the fourth stem loop (sl4, Figure 3.4.2d) of the dimerization/encapsidation signal. The encapsidation signal is a four stem-loop structure which stretches from the 5' LTR and interacts with the nucleo-capsid protein, promoting formation of genomic RNA and blocking the initiation of transcription (Clever *et al.*,1995; Huthoff *et al.*,2004; Wilkinson *et al.*,2008).

In the *pol* gene, two highly conserved regions with known functions were found. One corresponded to the first three nucleotides (HXB2 coordinate positions 4092-4094) of the 260 nucleotide long cis-repressive sequence (CRS; Figure 3.4.2e). The cis-repressive sequence inhibits expression of structural protein mRNAs by preventing their transportation from the nucleus – a process that is reversed by the RRE (Cochrane *et al.,*1991; Schneider *et al.,*1997; Wolff *et al.,*2003). The other was found at the 3' end of the intragenic nuclease hypersensitive domain (hs7 in Figure 3.4.2e). The latter motif, located between positions 4926 and 4937 and labeled “ese” in Figure 3.4.2e, is also located within HIV-1 exon 2 and the last six nucleotides, TGGAAA, of this conserved region form a known exonic splicing enhancer (ESE) of HIV mRNAs (Exline *et al.*,2008; Kammler *et al.*,2006; Krummheuer *et al.*,2001; Madsen *et al.*,2006; Schwartz *et al.*,1990a).

Three of the highly conserved regions with known functions were in the *env* gene and included a nine-nucleotide long motif from position 7834 to 7842 within the RRE. The approximately two hundred nucleotides long RRE element within *env*, is known to interact with the Rev protein and facilitates the transport of late un-spliced and partially-spliced RNAs from the nucleus to the cytoplasm (Hadzopoulou-Cladaras *et al.*,1989; Peterson *et al.*,1996; Renwick *et al.*,1995). Although the RRE is associated with a long stretch of sequence that forms a well characterized secondary structure with various conserved domains (Phuphuakrat *et al.*,2003), only the nine nucleotides that bind Rev with highest affinity (Hung *et al.*,2000; Peterson *et al.*,1996) were sufficiently conserved to be detected using our conservative threshold (Figure 3.4.2f).

Also conserved within the *env* gene were the two splice site regions at the end of the *tat*/*rev* exon, positions 8349-8354 and 8376-8378, usually referred to as 7a/7b, and 7 ("ss" in Figure 3.4.2f; (Kammler *et al.*,2006; Schwartz *et al.*,1990a; Schwartz *et al.*,1990b; Swanson *et al.*,1998)



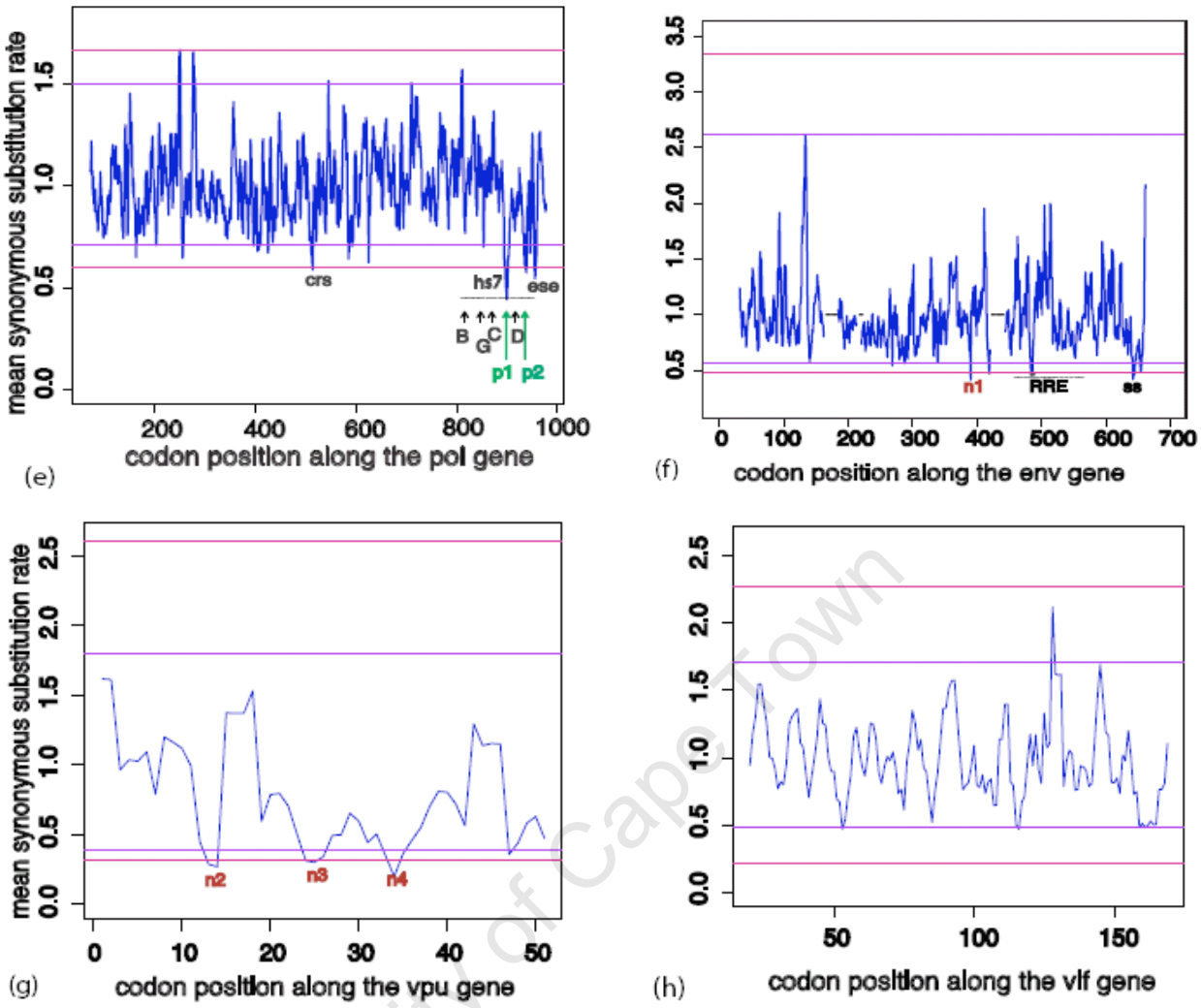


Figure 3.4.2: Mean (blue) synonymous substitution rates observed across each HIV-1 gene. Mean dS was calculated over sliding windows of three codons. Horizontal lines mark the most stringent (red) and less stringent (purple) significance thresholds. (a) dS across the *nef* gene. "G-A"; G-to-A hypermutations (see Figure A3.4.2a in the Appendix), 'var'; highly variable region, 'ppt'; poly-purine tract, "c"; PPT integrase attachment site, 'nre'; start of the negative repressive sequence, 'ets'; Ets-1 transcription factor binding site. (b) dS across the *vpr* gene. 'ssa3'; 3' splice acceptor site A3, 'rnase'; RNase-V1 cleavage site. (c) dS across the *tat* gene. 'ess2'; exonic splicing silencer ESS2, 'ssa4b'; 3' splice acceptor site A4b. (d) dS across the *gag* gene, 'sl4'; the fourth stem loop of the encapsidation signal, 'INS1'; a motif within the first inhibitory sequence region, 'INS2'; a motif within the second inhibitory sequence region, (e) dS across the *pol* gene. 'crs'; start of the cis-repressive sequence, horizontal dotted line is the nuclease hypersensitive region and sites 'B', 'G', 'C' and 'D' are confirmed transcription factor binding sites known as site-B, GC-box, site-C

and site-D respectively. ‘p1’and ‘p2’; conserved sites within nuclease hypersensitive region. “ese”; exonic splicing enhancer, (f) dS across the *env* gene. “n1” is the novel conserved site. The black dotted horizontal lines indicate poorly aligned regions that were excluded from the analysis. “rre”; rev-responsive element, *; the 9 nucleotides (5' GACGGUACA 3') which bind to the Rev protein with highest affinity, “ss”; splice site region for the *tat* and *rev* 3' exons, (g) dS across the *vpu* gene. ‘n2’, ‘n3’ and ‘n4’; novel conserved sites, (h) dS across the *vif* gene. No highly conserved sites were observed

Gene	HXB2 coordinates	Sequence Logo and degeneracy	Function
gag Figure 3a 'sl4'	793 - 807		Fourth stem loop of encapsidation signal
pol Figure 3b 'crs'	4092 - 4094		Cis-repressive sequence start site
Figure 3b 'ese'	4926 - 4937		"TGGAAA" is an exonic splicing enhancer
vpr Figure 3d 'ssa3'	5759 - 5777		3' splice acceptor site A3
Figure 3d 'rnase'	5794 - 5805		RNase-V1 cleavage site
tat Figure 3e 'ess2'	5855 - 5863		Exonic Splicing Silencer 2
env Figure 3g '*'	7834 - 7842		Rev binding loop in rev-responsive element
Figure 3g 'ss'	8349 - 8354		Tat/rev 3' exons splice sites 7a, 7b
Figure 3g 'ss'	8376 - 8378		Tat/rev 3' exons splice sites 7
nef Figure 3h 'ppt'	9066 - 9076		PPT (polypurine tract)
Figure 3h 'ppt'	9077 - 9086		PPT, Priming of transcription
Figure 3h 'c'	9084 - 9091		PPT cleavage region
Figure 3h 'nre'	9183 - 9192		3'LTR negative responsive element start sites
Figure 3h 'ets'	9391 - 9399		TF binding region for Ets-1 proteins

Table 3.4.2: Sequence motifs for the highly conserved regions with known function, the fourteen regions with known specific functions found to be under strong purifying

selection in HIV-1 genes. The range of coordinates on the HIV-1 genome for each motif is given in column 2. Numbers above each logo represent the degeneracy at each nucleotide site.

3.4.3 Purifying selection pressure within sites with unknown functions

The sequence logos for the remaining nine highly conserved regions are given in Table 3.4.3**.** Possible functions for another five were predicted based on the known functions of the sequence domains in which they were situated. In addition to the fourteen conserved regions with known functions, five motifs (marked in green in Figures 3.4.2d and e) without previously reported specific functions occurred within known functional domains. These include three short (3-6bp) motifs in the inhibitory (INS) sequence regions of gag (HXB2 positions 898-903, 985-996 and 1309-1314, Figure 3.4.2d). Previous *in-vitro* analyses have shown that short motifs within the approximately two hundred bp long INS regions, are responsible for the actual inhibition of mRNA expression (Schneider *et al.,*1997; Schwartz *et al.,*1992; Wolff *et al.,*2003). Although the three motifs we find within this region have not been specifically identified *in-vitro*, a computational study by Wolff et al (2003) showed that the INS sequences have several short functional motifs within them (Wolff *et al.*,2003). The conserved sites we identified within INS1 and INS2 could serve the same inhibitory function. Functional assays elucidating the role of these sites in the inhibition of mRNA expression could help to determine the precise mechanisms by which inhibition occurs and whether these sites also play a role. In another previous study which analyzed the RNA secondary structure of the 5' region of HIV-1, these two regions, labelled 'INS1' in Figure 3.4.2d, were found to be involved in conserved Watson-Crick base-pairing (Paillart *et al.*,2002).

The last two of the five conserved regions with unidentified specific functions were within the *pol* HS7. HS7 spans five hundred nucleotides, between positions 4481 and 4982 of the HIV-1 genome, and has an LTR-like regulatory function (Goffin *et al.*,2005; Van Lint *et al.*,1994). Previous studies revealed four domains towards the 3' end of this region (PU box, GC-box, site-C and site-D) that bind to specific

transcription factors (TFs) and are also important for viral infectivity (Goffin *et al.*,2005; Van Lint *et al.*,1994). In these studies, the Oct-1, Oct-2, PU.1, Sp1 and Sp3 transcription factors were found to bind to at least one of the four identified sites. The two conserved regions we identified are outside these specific identified functional domains, but one of them at positions 4767-4790 (labeled "p1" in Figure 3.4.2e) showed potential binding to Oct-1 using the MATCH tool from the TRANSFAC database (Kel *et al.*,2003; Matys *et al.*,2003). Potential association with a transcription factor was not observed for the three nucleotides (positions 4864-4866) labelled 'p2' and its adjacent sites.

The function of a novel twelve-nucleotide region with high degree of sequence conservation observed in the *env* gene and three other regions in *vpu* (marked in red in Figures 3.4.2f and g) has not been previously reported. One of these novel regions, the twelve nucleotide long motif in *env,* upstream of the RRE showed the highest degree of conservation of any region in the *env* gene ('n1' in Figure 3.4.2f; positions 7656-7667 of HXB2) and therefore it would be interesting to determine its function in the context of the extensively analysed functions of the envelop protein. This sequence was sent to a collaborating laboratory at the National Institute of Infectious Diseases (NICD) (http://www.nicd.ac.za), for fitness assay analysis. In brief, this nucleotide sequence, located at positions 7656-7667 of the HXB2 reference HIV-1 genome was mutated at 3 different synonymous sites (positions 7662, 7664 and 7667). Each mutant virus was tested for infectivity and the rate of replication in comparison to a wildtype virus that was not subjected to any synonymous mutations (Figure 3.4.3). The details of the experiments performed at NICD are described in the publication that discusses the findings of this project (Ngandu *et al.,*2008). Preliminary analysis showed that synonymous mutations within this region had no significant effect on the fitness of the virus and only showed a slight decrease in the replication rate, but can further be analysed in more sensitive replication assays.

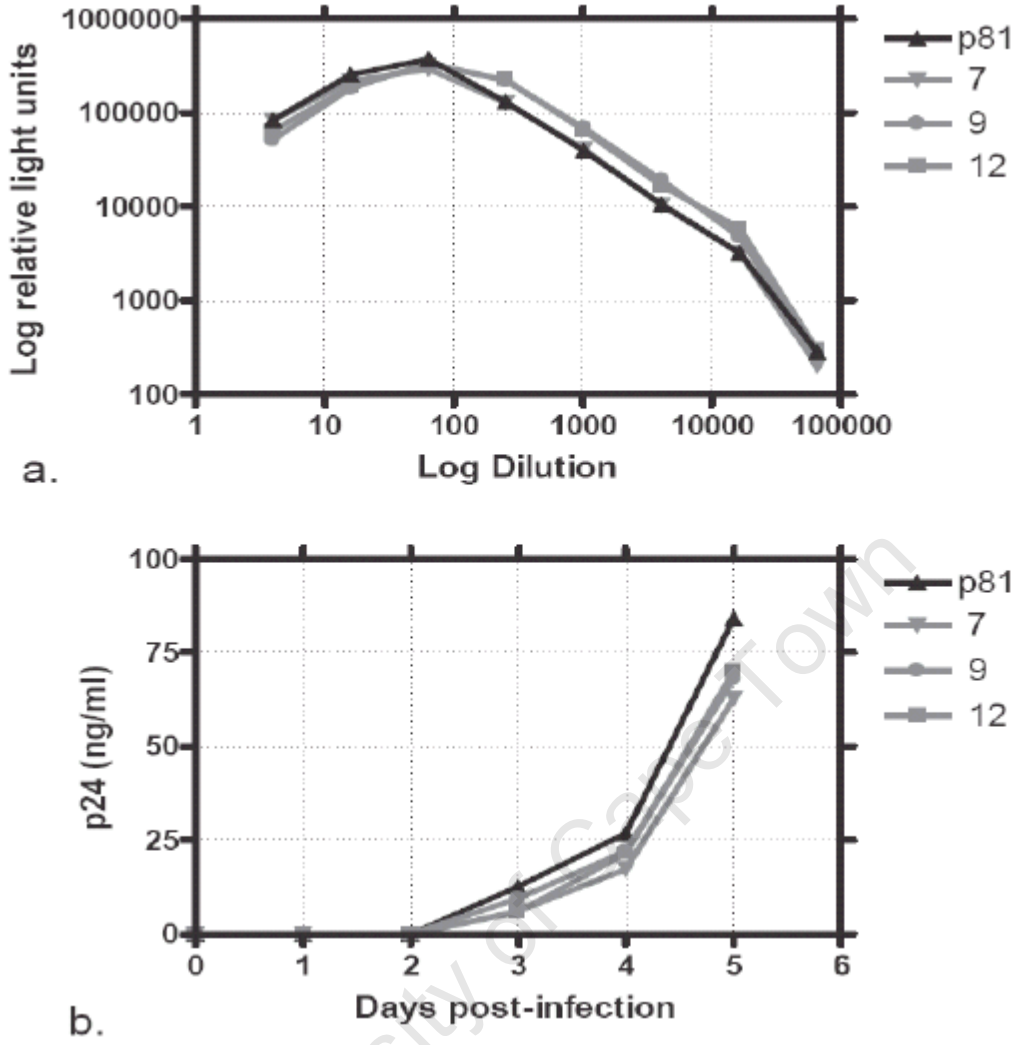


Figure 3.4.3 Functional analysis of a novel region found in the *env* gene. The synonymous mutations were introduced at positions 7662 (7), 7664 (9) and 7667 (12) each time and compared against a wildtype strain p81. (a) Comparison of infectivity between wildtype (p81) and the mutants (b) rate of replication as measured by the production of p24 protein from transfected wildtype virus and mutants (Ngandu *et al.,*2008).

Functional analyses of the three novel conserved regions in *vpu*, with HXB2 coordinates 6101-6106, 6143-6151 and 6167-6178 (n2, n3, n4 in Figure 3.4.2g) are being considered. The protein products of *vpu* and *env* are known to be produced from a bicistronic transcript (Schwartz *et al.*,1990b) and the conserved regions in *vpu* may be involved in the control of translation.

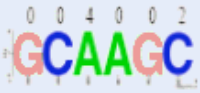
Gene	HXB2 coordinates	Sequence Logo and degeneracy	Function
		Regions with predicted functions	
gag Figure 3a 'INS1'	898 - 903		Within (INS1) first Inhibitory sequence region
Figure3a 'INS1'	985 - 996		Within INS1
Figure 3a 'INS2'	1309 - 1314		Within INS2
pol Figure 3b 'p1'	4764 - 4776		Within nuclease-hypersensitive region
Figure 3b: 'p1'	4777 - 4790		potential binding motif for Oct-1
Figure 3b 'p2'	4864 - 4866		In nuclease-hypersensitive region
		Novel regions	
vpu Figure 3f 'n2'	6101 - 6106		novel conserved region
Figure 3f 'n3'	6143 - 6151		novel conserved region
Figure 3f 'n4'	6167 - 6178		novel conserved region
env Figure 3g 'n1'	7656 - 7667		novel region, analyzed in fitness assays

Table 3.4.3: Sequence motifs for the highly conserved regions with unknown specific function, the five regions with predicted functions and four novel with unknown functions found to be under strong purifying selection in HIV-1 genes. The range of coordinates on the HIV-1 genome for each motif is given in column 2. Numbers above each logo represent the degeneracy at each nucleotide site.

3.4.4: Evidence of false inference of positive selection in the presence of purifying selection pressure

As mentioned earlier, purifying selection is likely to cause false inference of positive selection (Hurst *et al.*, 2001). To further assess this, the MG94 codon model (Muse *et al*., 1994) was used to detect positive selection in a subtype C *gag* coding sequence alignment described in one of the thesis publications and in Chapter 4 (Ngandu *et al.*,2007). Sites with ω significantly greater than one, implying positive selection, overlapped significantly with sites that had lower than average dS values (Fisher's exact test odds ratio = 4.4; p value = 0.006; Figure 3.4.4). This is consistent with a substantial proportion of the positive selection signals resulting from conservation of the synonymous sites rather than diversifying selection acting on the nonsynonymous sites.

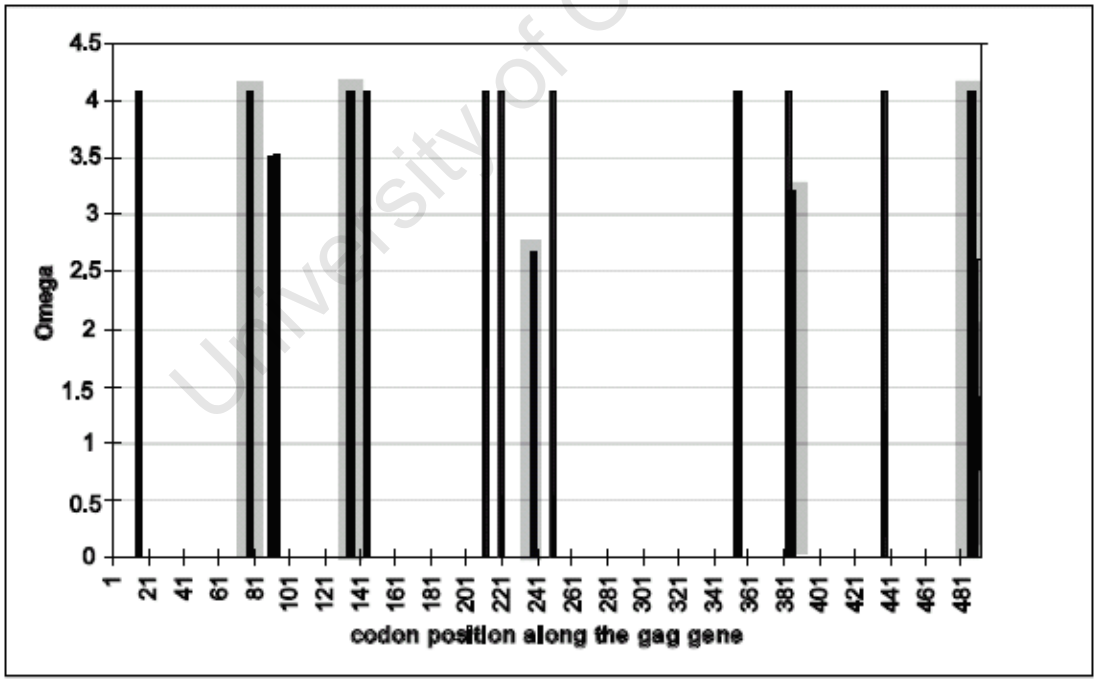


Figure 3.4.4: Evidence of overlap between high omega at a codon and low dS at the synonymous sites, positively selected sites at which a significantly low dS was observed at the synonymous sites. Positively selected sites are shown in black vertical lines and sites with low dS are shaded in light grey.

3.5 Discussion

This is the first study to provide a detailed analysis of site-to-site variation in the rate of synonymous substitutions across the HIV-1 genome. In the past, site-to-site variation in dS in HIV-1 has been investigated in a single gene (Kosakovsky Pond *et al.*,2005a; Lemey *et al.*,2007). In another study, a single overall synonymous substitution rate for the entire genome was determined for comparison to other viral lineages (Hanada *et al.*,2004). The method for determining site-to-site synonymous rate variation used in this study is similar to an approach from a previous study (Kosakovsky Pond *et al.*,2005a), in that case only one HIV-1 gene, *vif,* was considered and sites that encode proteins in multiple reading frames were included. As a consequence, it was not clear whether the observed site-to-site rate variation resulted from variation in the synonymous rate or from selection acting on nonsynonymous substitutions in another reading frame.

Here, all regions of the HIV genome that encode proteins in a single reading frame were analysed as well as the functions of nucleotide sequences that have the largest influence on synonymous substitution rate variation. Previous studies have demonstrated that recombination causes false inference of positive selection. Since recombination affects tree topologies used in fitting phylogenetic models, it is also likely to cause biased estimates of dS. The recent development of methods to account for recombination in selection analyses (Scheffler *et al.*,2006) enabled the removal of recombination as a source of bias in the estimates of synonymous substitution rates in this study.

All coding sites of the HIV-1 genome under the influence of purifying selection pressure were identified in this study. Substantial site-to-site variation in the rate of synonymous substitution with evidence of purifying selection pressure was observed within functional domains such as the Rev-responsive element. In addition to identifying putatively functional sites under purifying selection, these results contribute to the robustness of analyses of positive selection by identifying conserved synonymous sites that can cause false positive inference of selection. Conserved synonymous sites that can cause false detection of positive selection need to be either excluded from analyses or modeled appropriately. The danger is that some sites may

be assumed to be evolving adaptively simply as a result of the purifying selection acting directly on the nucleotide sequence. Evidence of the tendency for sites with significantly low dS to appear to be positively selected was observed in subtype C *gag*. In addition, a study by Hurst & Pal also showed false detection of positive selection caused by purifying selection pressure acting on synonymous sites (Hurst *et al.*,2001).

In many selection studies the synonymous rate is assumed to be constant. However, negative selection acting on synonymous sites can potentially reduce gene-wide estimates of the synonymous rates below the neutral evolution rate. Comparison of site-specific nonsynonymous substitution rates against this underestimate of the neutral rate is likely to cause a proportion of the selectively neutral nonsynonymous sites to seem as though they are evolving adaptively. A very stringent significance level cutoff was used here to identify the twenty-three regions within the HIV genome that have obviously reduced synonymous substitution rates. The rate of synonymous substitution was higher on average, in overlapping gene regions that encode proteins in more than one frame ($p = 6 \times 10^{-7}$ from Wilcoxon rank sum test; Figure 3.4a); however, lower dS was observed within some genome regions that are translated in multiple reading frames. Analysis of the most diverse sequence strains within subtype B and C revealed more highly conserved sites across the RRE and INS1 regions at the subtype level (Figure A3.5 in the Appendix). This indicates that false detection of positive selection is likely to increase in analysis carried out within a single subtype. For a more conservative analysis of selection, all the significantly conserved sites, i.e., including those conserved at 95% confidence (p value <0.05, listed in Table 3.4.1b) should be excluded from positive selection analyses along with sites that encode proteins in multiple frames. Sequence conservation around nucleotide sequence motifs with known functions and newly identified additional conserved nucleotide elements that do not fall within any currently characterized functional motifs are presented in Tables 3.4.1 and 3.4.2 and thus form a resource for future studies of selection pressures acting on HIV-1 genes.

Chapter 4

CTL response to HIV type 1 subtype C is poorly predicted by known epitope motifs

Foreword

The analysis discussed in this chapter has been published in one of the papers listed at the beginning of the thesis (Ngandu et al.,2007). The HIV-1 subtype C sequence data as well as the immunological responses data used in this study were provided by Dr. Clive Gray from the National Institute of Communicable Diseases in Johannesburg and Dr. Carolyn Williamson from the Institute of Infectious Diseases at the University of Cape Town. Therefore sections 4.3.1 through to 4.3.4 summarise the methods that were used by these collaborating laboratories to screen and isolate these data.

4.1 Summary

The cytotoxic T-lymphocyte immune response is important in controlling HIV-1 replication *in vivo*. Protective immune responses directed against conserved and immunodominant peptides are ideal targets for anti-HIV vaccine design. This chapter assesses the relationship between the frequency of targeting of HIV-1 Nef and Gag proteins and sequence evolution patterns in order to identify conserved immunodominant regions. In addition, the relationship between observed and predicted CTL responses given the HLA genotypes of the infected individuals was investigated. Amino acid sequence variability was measured by calculating sequence entropy in sequence alignments and by estimating the rate of evolution at each codon site using maximum likelihood. Using a robust statistical approach, sites in the central part of the *nef* gene and *gag* p24 that were frequently targeted by the CTL response were shown to be evolutionarily conserved. By contrast, the immunodominant peptides within the *gag* p17 region tended to be highly variable. CTL responses observed in the subtype C infected cohort poorly conformed to predicted sequence motifs required for binding by the HLA A or B alleles found in the corresponding patient. Only 52% of the Nef peptides and 64% of the Gag peptides that elicited a

CTL response contained predicted sequence motifs for the patient HLA. A comparable subtype B dataset showed a higher consistence between observed peptides that elicited a CTL response and patient HLA genotype (96% and 83% for Nef and Gag, respectively). This difference between subtypes C and B was demonstrated to be due to poor characterisation of HLA alleles common in Southern African populations and the tendency for sequence motifs associated with HLA recognition to be over-specified for sequence variation found in the B clade. The HLA binding motifs are therefore likely to be biased towards certain populations and subtypes and this can have important implications for understanding immune escape and predicting vaccine efficacy in the context of populations primarily infected with non-B subtype of HIV-1.

4.2 Background

The cytotoxic T-lymphocyte (CTL) immune pathway plays an important role in the control of viral replication in HIV-1 infected individuals (Betts *et al.,*1999; Letvin,1998; Moore *et al.,*2002). These immune responses are activated by antigenic peptides presented by human leukocyte antigens (HLAs) on the surface of infected cells. The HLA molecules are highly polymorphic and each molecule recognises peptides with specific sequence patterns known as anchor residue motifs (Madden *et al.,*1992). Some HLA molecules mediate CTL responses that confer resistance to disease progression yet others are associated with rapid progression to AIDS (Chen *et al.,*1997; Ross *et al.,*2002; Trachtenberg *et al.,*2003).

The disease outcome, however, has been found not only to be a function of the HLA allele and immune response itself but also a characteristic of the specific targeted region (Maurer *et al.,*2008; Miura *et al.,*2008a; Pereyra *et al.,*2008). The proteins encoded by the different genes of the viral genome vary in immunogenicity with some failing to elicit measurable immune responses (Kiepiela *et al.,*2007). Certain regions of the viral sequence are able to mutate and escape the immune response. The effect of the resulting mutant on the fitness of the virus determines the rate of viral replication and hence viral load. Persistent immune responses against regions of the viral sequence that mutate to weaker strains and fail to revert to the fit wild type virus typically delay the progression to AIDS (Ross *et al.,*2002). Short-lived immune responses that allow the virus to revert

back to the wild type state can be disadvantageous to the host (Ross *et al.,*2002). Highly conserved regions that are targeted by immunodominant and protective immune responses across many individuals are ideal targets for vaccine design.

The Nef and Gag proteins have the most immunodominant peptides across populations infected with different HIV-1 subtypes (Frahm *et al.,*2006; Masemola *et al.,*2004a). A negative relationship has been observed between amino acid sequence variation and the frequency of CTL recognition across HIV-1 proteins (Frahm *et al.,*2006; Yusim *et al.,*2002). The amino acid sequence variation in most of these studies has been inferred by calculating the entropy at each site of an amino acid sequence alignment. Entropy is calculated such that each site in a sequence is assumed to be independent. However, adjacent sites within a peptide that is recognised by an HLA molecule are not functionally independent. Therefore a direct correlation between the entropy values per site and the measure of the frequency of recognition cannot be justified. To compare the entropy of frequently targeted sites to less frequently targeted sites, the length of the targeted peptides needs to be taken into account. In some studies, the mean of the entropy over nine amino acid positions (Yusim *et al.,*2002) has been used, thus assuming that all targeted peptides have the same length yet in reality the length ranges from 8 to 11 amino acids in length (Frahm *et al.,*2007).

Even though most HIV-1 studies have focused on subtype B, subtype C poses the strongest challenge as it accounts for almost 50 percent of worldwide infections (Buonaguro *et al.,*2007). It is therefore important for vaccine design studies to pay more attention to the understanding of HIV-host viral relationship in subtype C infected populations. This chapter provides a detailed and more accurate analysis of the relationship between the CTL immune response and HIV-1 sequence variation using data consisting of CTL responses against the Nef and Gag proteins observed in a Southern Africa cohort, provided by collaborators at the South African National Institute of Communicable Disease (NICD). The extent to which predicted HLA binding motif data available in public databases is biased towards the better characterized HIV-1 B subtype was also investigated. This was done by determining the frequency with which an anchor residue motif associated with at least one of the HLA alleles of the host appear within peptides recognized by that host's CTL response. The occurrence of an anchor residue motif for one of the HLA alleles of the infected

individual does not prove that the observed immune response was mediated by that HLA allele. However, the absence from the targeted peptide of any sequence motif associated with the anchor residues of any of the HLA alleles of the host does imply missing information in the databases. Specifically, it implies that the anchor residues that have been bound by the HLA allele do not conform to the motif present in the database for that HLA allele. In order to investigate the extent to which known HLA anchor residue motifs are representative of all possible immune responses, we compared the fraction of CTL responses that can be accounted for by HLA anchor residue motifs between comparable subtype C and subtype B datasets.

4.3 Materials and Methods

4.3.1 Study subjects

The data were isolated from sixty-four HIV-1 infected individuals enrolled in a study of HIV-1 infection from four southern African countries: Malawi, Zimbabwe, Zambia and South Africa. The data consists of cytotoxic T-lymphocyte immune responses observed against the Gag and Nef proteins (see Table 4.3.5) as well as autologous viral sequences. The methods of data collection and the determination of immunological responses have been published in a previous study from Dr. Gray's laboratory as well as in the publications that report the findings of this study (Masemola *et al.,*2004a; Ngandu *et al.,*2007).

4.3.2 Summary of screening immune response data

Peptide sequences from regions observed to be targeted by the CTL response in HIV-1 subtype C infected individuals were provided by Dr Clive Gray. The methods used for identifying and screening the peptides were described in detail in a publication from his research group (Masemola *et al.,*2004a). In summary, responses against Gag were screened using sixty-six synthetic peptides based on a consensus subtype C sequence from Zambia. Nef-specific responses were confirmed using fifty peptides based on the viral strain Du151. The peptides used were synthesized using Fmoc-based solid phase chemistry (Natural and Medical Sciences Institute, University of Tuebingen, Germany). The peptide sets consisted of 15-18mers overlapping by 10

amino acids (for Gag) and 15 mers overlapping by 11 amino acids (for Nef). The immune responses to each peptide were confirmed in interferon-γ ELISpot assays. In ELISpot assays, PBMCs were plated on 96-well polyvinylidene difluoride plates that were coated with anti-IFNγ monoclonal antibody. Peptides were added directly to the wells at various final concentrations, incubated under specified conditions and subsequently washed to remove unbound peptides and displaced antibody. Streptavidin horseradish peroxidase (Pharmingen, Cupertino) was added to the wells and the plates incubated to develop spots on bound peptides. Both negative and positive controls were used for comparison. The number of spots per well was counted using the Immunospot (Cellular Technology Ltd., Cleveland, USA) automated cell counter.

4.3.3 Patient HLA allele data

HLA genotypes from the subtype C cohort were provided by Dr Clive Gray (National Institute of Communicable Diseases). High resolution typing of HLA Class I A and B loci was also performed using sequencing-based typing kits and the resulting sequences and subsequent allele assignment performed using the MatchMaker™ Allele Identification Software (Applied Biosystems). Two HLA alleles from each of A and B loci were provided from each patient. Similarly, each patient from the subtype B infected cohort had two alleles for each of A and B HLA loci. The HLA allele frequencies in both datasets are listed in the appendix Table A4.3.2.

4.3.4 Sequence data

A total of one hundred and seven full-length *nef* and *gag* sequences were amplified and sequenced as described in a previous study from Prof. Williamson's group (Bredell *et al.,*2007). In-frame alignments for each gene were generated using ClustalW with default parameter settings (Thompson *et al.,*2002). The GenBank accession numbers of the sequences used are DQ792982 to DQ793089.

4.3.5 Immune response data

Sixty-eight and eighty responses were observed against Nef and Gag peptides respectively. The average length of the targeted Gag peptides was 20 amino acids long and that for Nef was 15 amino acids. A comparative HIV-1 subtype B dataset from the REACH (Reaching for Excellence in Adolescent Care and Health) study was obtained online for comparative purposes (Bansal *et al.,*2003; Sabbaj *et al.,*2003; Wilson *et al.,*2001) (Table 4.3.5). The forty-five patients infected with subtype B were of African American, Hispanic and Caucasian origins. Thirty-one and thirty-nine Nef and Gag responses respectively, were observed from the subtype B cohort and the mean length of the targeted peptides was 20 amino acids in both genes.

	Subtype C		**Subtype B**	
Number of patients	64		45	
Gene	Gag	Nef	Gag	Nef
Number of responders	40	45	39	31
Number of peptide responses	80	68	113	70
Mean number of peptide responses per patient	2	1.5	2.9	2.3
Average peptide length	22	15	20	20

Table 4.3.5: CTL responses observed against the Nef and Gag peptides in a Southern African subtye C infected cohort and a comparative subtype B infected cohort from a previously published study of CTL responses (Bansal *et al.,*2003; Sabbaj *et al.,*2003).

4.3.6 HLA anchor residue motif data

HLA anchor residue motifs were obtained from the Los Alamos database (Korber *et al.,*2006), the SYFPEITHI database (Rammensee *et al.,*1999) and the HLA Ligand/Motif database (Sathiamurthy *et al.,*2003). We used only motifs with at least two anchor residues. Motifs from other databases were presented within these three databases. All anchor residue motifs with lengths 8,9,10 and 11 and with at least two anchor residues (at the second and C-terminal positions) were used.

4.3.7 Determination of amino acid variation and rate of sequence evolution

Amino acid sequence entropy was calculated at each site in the alignment. The mutation rates at a site across each gene were determined by estimating the dN/dS ratio (nonsynonymous to synonymous substitution rate ratio). The dN/dS ratio was calculated over the phylogenetic tree relating the sequences using the MG94 codon model (Muse *et al.,*1994) implemented in the HyPhy program (Kosakovsky Pond *et al.,*2005d).

4.3.8 Statistical analysis

A robust statistical method was developed to investigate whether the relationship between sequence variation and the frequency of epitope recognition could be the result of chance. For each peptide to which a CTL response was detected, the mean entropy of the sequence alignment within the region spanned by the peptide was calculated. To summarize the entropy of all of the regions targeted by CTL responses across all patients a statistic, referred to here as 'weighted-mean entropy' was calculated. The weighted mean entropy was the average entropy of regions spanned by targeted peptides weighted by the number of patients that responded to the peptide. The observed weighted-mean of the entropy was then compared to the same quantity calculated from randomized datasets. For the randomized datasets, a random position on the sequence alignment was assigned to each of the peptides 1000 times and the weighted-mean entropy recalculated for each replicate. This approach was similarly used for the dN/dS rates.

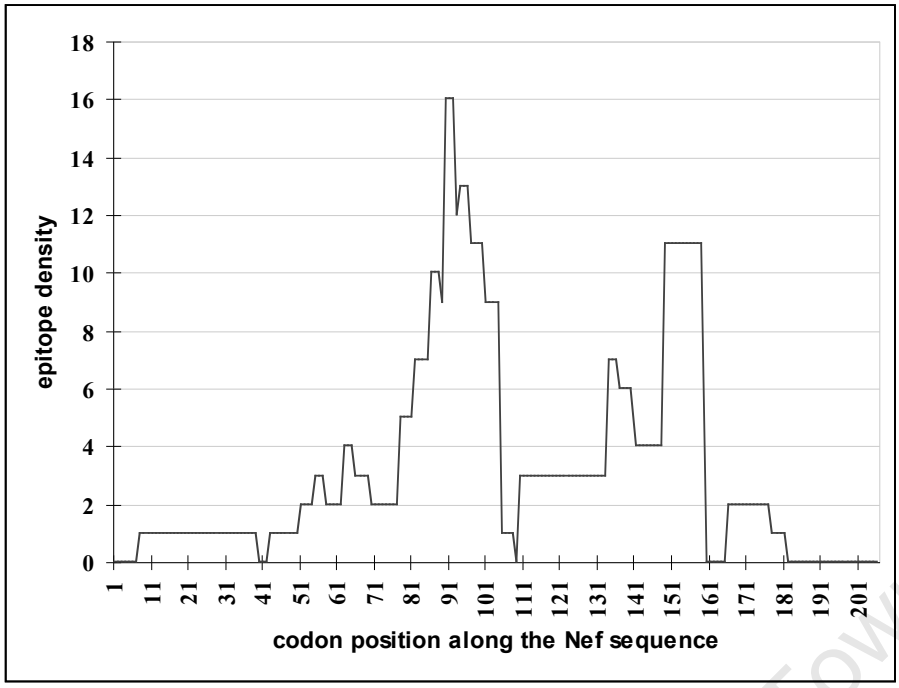
4.4 Results

The number of individuals that showed a response to a peptide, here referred to as epitope density, spanning each site across a gene was determined. High epitope density indicates high frequency of CTL recognition of a peptide across different individuals in the cohort. Low epitope density shows that few individuals had a CTL

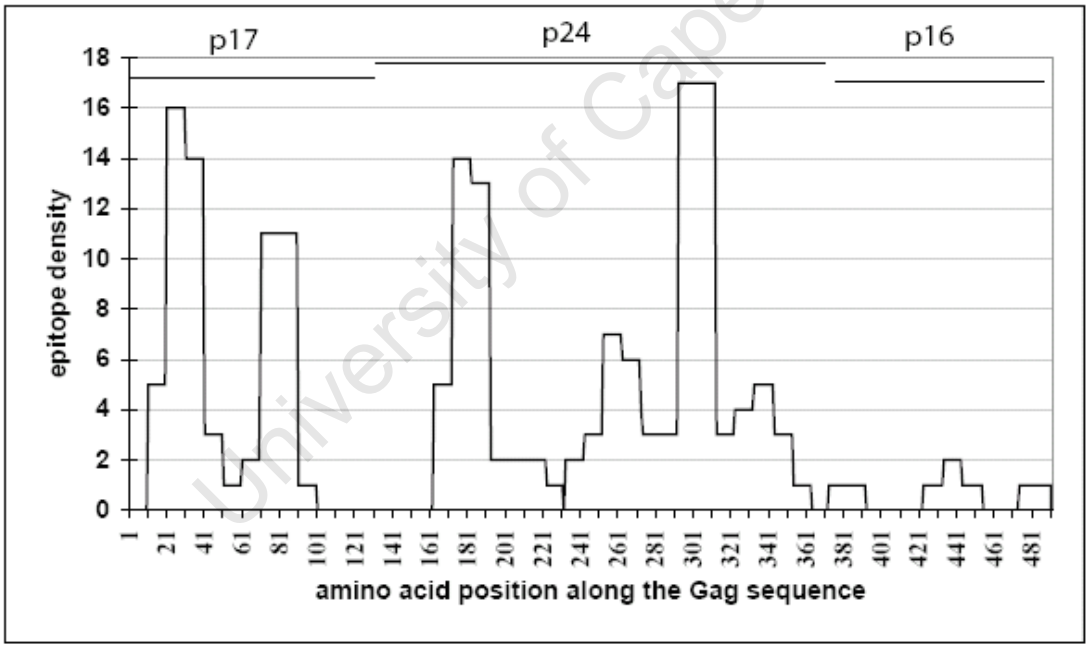
response against a peptide spanning a site. The observed epitope density is plotted against observed entropy values per site as well as against positive selection i.e. against dN/dS values greater than 1.

4.4.1 Observed epitope density

The most frequently targeted peptides were within the central regions of the Nef, Gag p17 and p24 protein regions. The terminal regions of Nef and the p16 region of Gag were poorly targeted. The frequencies with which a peptide was recognised by an immune response from different individuals are shown in Figure 4.4.1.



(a)

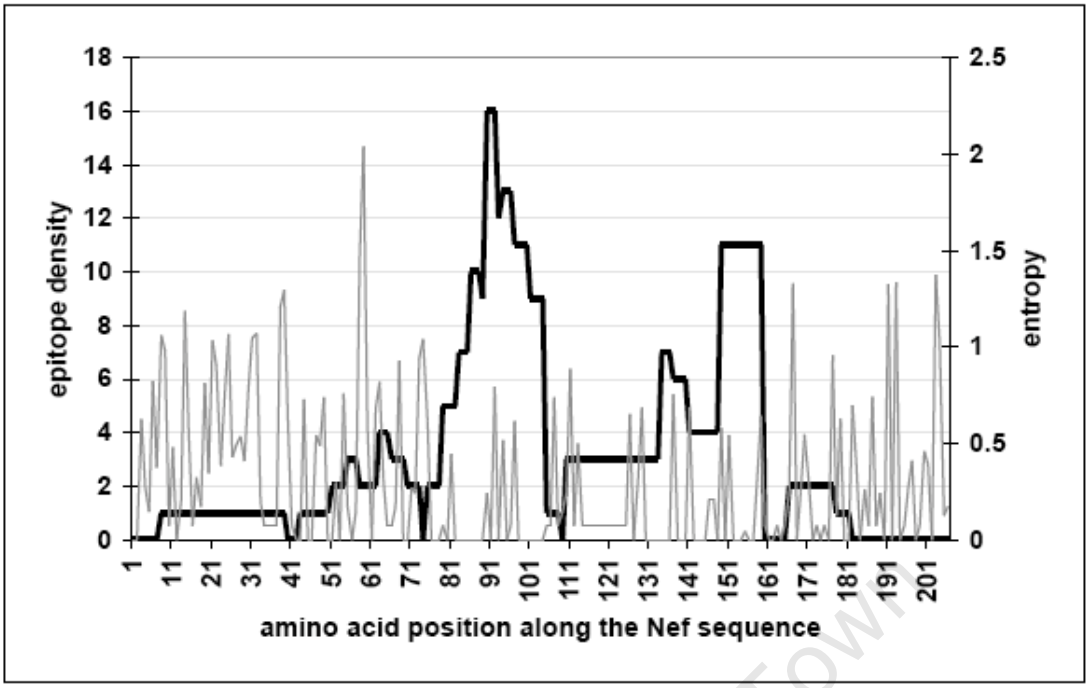


(b)

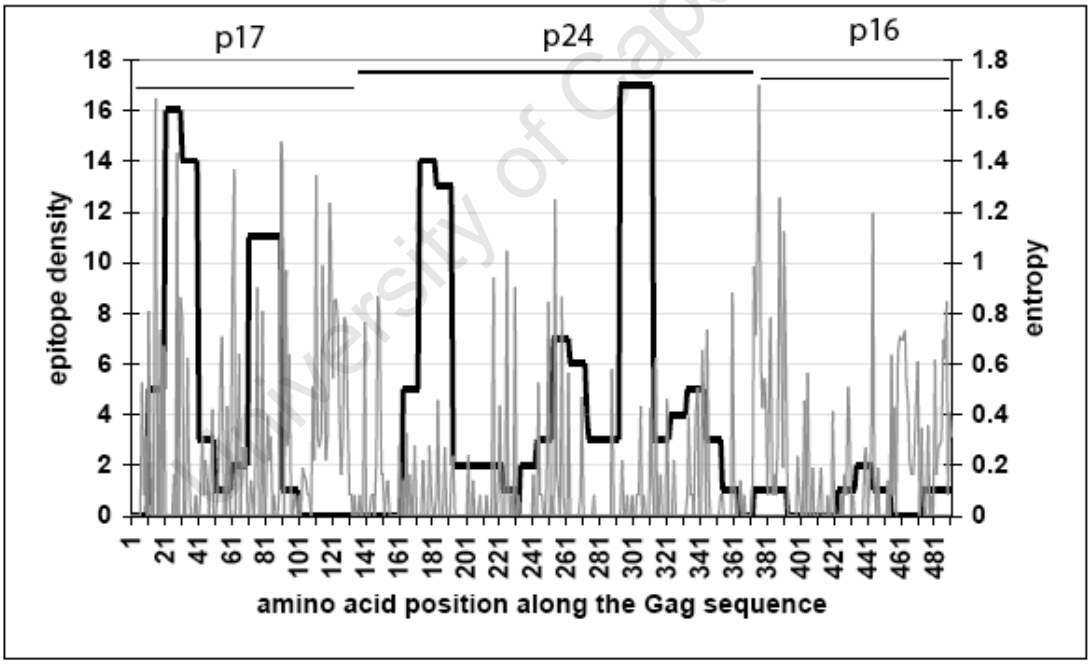
Figure 4.4.1: Epitope density (black) plotted across the HIV-1 Nef (a) and Gag (b) amino acid sequence position. (b) The p17, p24 and p16 regions of the *gag* gene are indicated by horizontal black bars.

4.4.2 Epitope density vs entropy

Higher entropy values were generally observed at the terminal regions of Nef which had very low epitope density (Figure 4.4.2a). In gag, high entropy values which indicate high amino acid variation were observed in p16 and p17 despite the latter showing the presence of highly immunodominant peptides (Figure 4.4.2b). The frequently targeted regions in p24 showed reduced sequence variation and a tendency to be conserved.



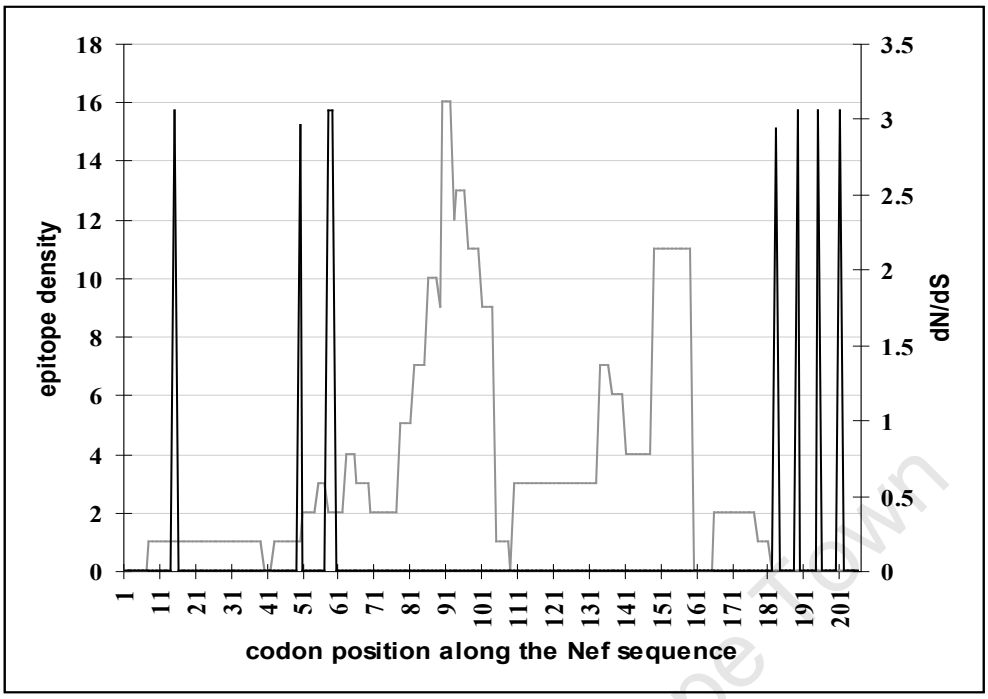
(a)



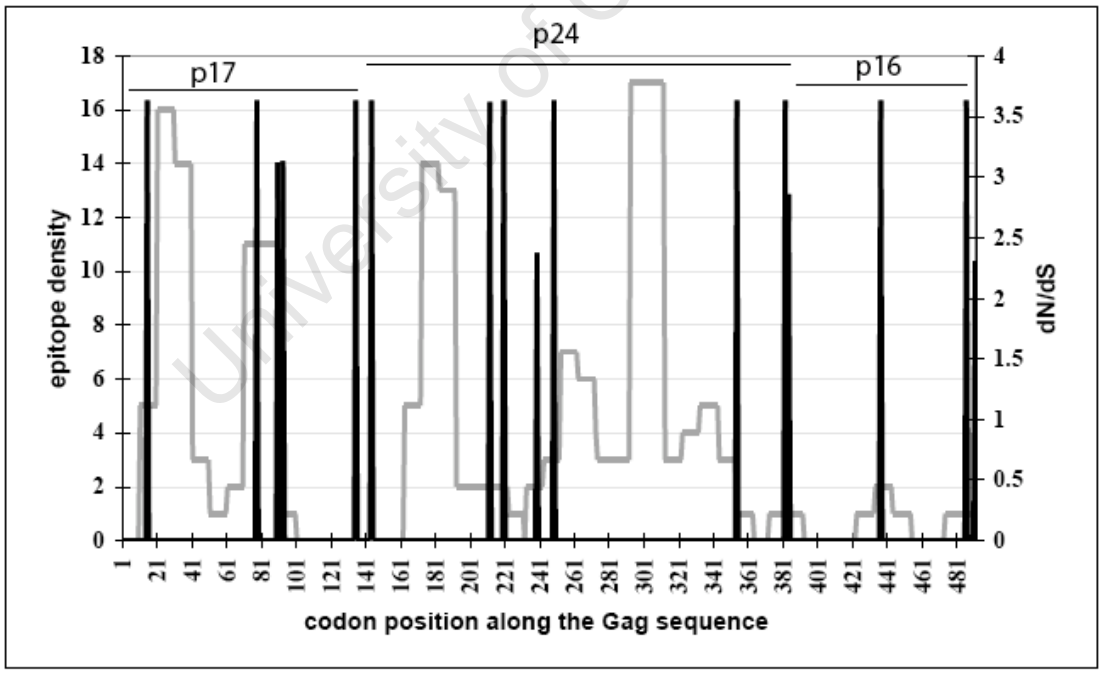
(b)

Figure 4.4.2: Epitope density (black) plotted alongside entropy (grey) for (a) Nef and (b) Gag

4.4.3 Epitope density plotted against omega



(a)



(b)

Figure 4.4.3: Positively selected sites (black vertical lines) vs epitope density (grey) along the *nef* (a) and *gag* (b) genes.

The dN/dS ratios were determined across both the *nef* and *gag* genes and positively selected sites, i.e., with dN/dS>1, an indication of adaptive evolution, were plotted against epitope density (Figure 4.4.3). A generally similar pattern was observed to that from the calculation of entropy. In *nef*, positively selected sites were only found at the poorly targeted terminal regions of the gene. Positively selected sites were scattered across the three different *gag* regions. However, in p24, the most immunodominant peptide regions were not under positive selection unlike in p17 where positively selected sites overlapped with sites frequently targeted by the CTL immune response (Figure 4.4.3(b)).

4.4.4 Statistical analysis using randomized datasets

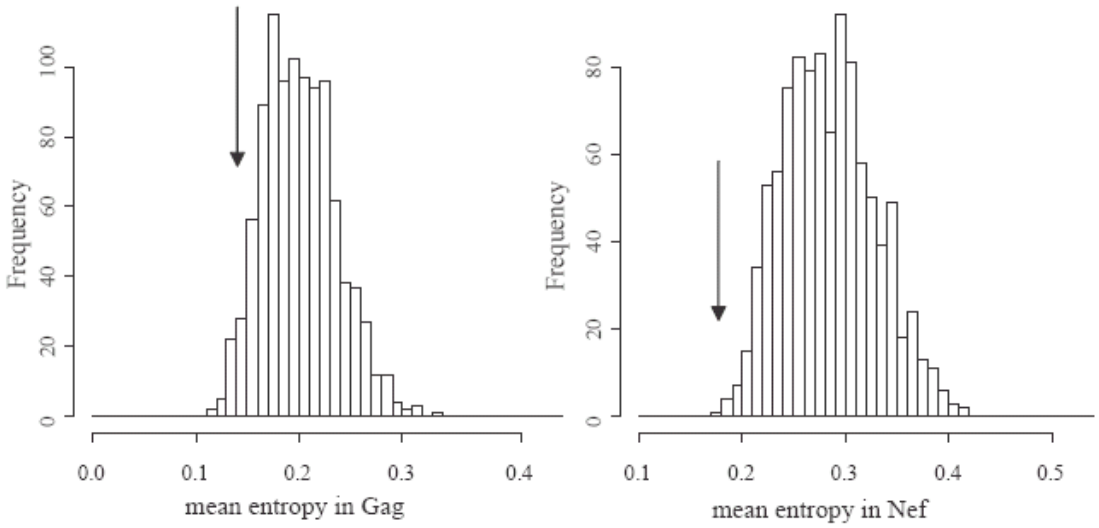
The weighted mean entropy values from randomized datasets carried out seperately for each of Nef, Gag, gag p17, gag p24 and gag p16, used to assess whether the observed entropy and dN/dS values could be achieved by chance, are shown in Tables 4.4.4(a) and 4.4.4(b) and Figure 4.4.4. The different regions of the *gag* gene coding for different proteins differed in sequence variation patterns. In both entropy and the dN/dS rates, the values for the Nef and the Gag p24 were significantly lower than what would be expected by chance indicating that the frequently targeted peptides encoded by these gene regions are generally conserved (entropy: Nef=0.015, Gag p24 = 0.001; dN/dS: Nef = 0.016, Gag p24=0.002). The rest of the targeted regions were not significantly conserved.

Gene region	Observed weighted mean entropy	P value
Nef	0.198	**0.015**
Gag	0.158	0.09
Gag p24	0.0988	**0.001**
Gag p17	0.242	0.753
Gag p16	0.334	0.80

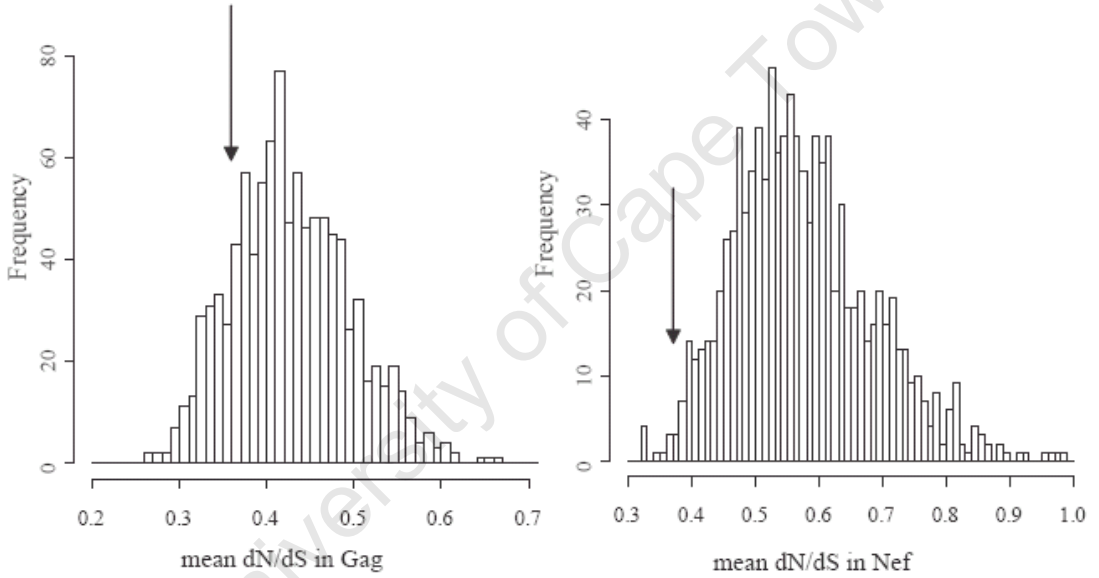
Table 4.4.4 (a): Observed weighted mean entropy**,** P value is the proportion of weighted mean entropy values from the randomized datasets that were lower than the mean values (column 2) observed from targeted regions. Significant p values are in bold.

Gene region	Observed weighted mean omega	P value
Nef	0.382	**0.016**
Gag	0.361	0.14
Gag p24	0.1552	**0.002**
Gag p17	0.179	0.76
Gag p16	0.56	0.94

Table 4.4.4 (b): Observed weighted mean dN/dS, P value is the proportion of weighted mean dN/dS rates from randomized datasets that were less than the observed weighted mean rate. Significant p values are in bold.



(a)



(b)

Figure 4.4.4 Histograms showing the distribution of weighted mean entropy and weighted mean dN/dS values obtained from the randomized datasets for both Nef and Gag whole proteins. (a) The frequency of entropy values from randomized datasets. (b) The frequency of dN/dS values from randomized datasets. In each case arrows show the position of the observed weighted mean entropy values.

4.4.5 Relationship between observed CTL responses and patient HLA genotype

CTL responses observed in the subtype C cohort poorly conformed to the patient HLA genotype as compared to the subtype B responses when considering predicted anchor residue motifs from the databases. Only 52.2% and 63.8% of the Nef and Gag responses respectively could potentially be linked to the patient HLA genotypes of classes A and B in this subtype cohort. In contrast, for the subtype B dataset 95.7% and 83.2% of the peptides recognized by CTL responses in *nef* and *gag* respectively, contained at least one of the anchor residue motifs restricted by HLA A and B alleles of the patient (Table 4.4.5).

	Subtype C	**Subtype B**
HLA A		
Nef peptides	17.4%	84.3%
Gag peptides	45%	52.2%
HLA B		
Nef peptides	39.1%	78.6%
Gag peptides	42.5%	66.4%

Table 4.4.5: Percentage of patient CTL responses to Nef and Gag peptides that contained at least one anchor residue motif for the HLA A or B Alleles of the patient

Two possible reasons for the poor association observed between the anchor residue motifs corresponding to the patient HLA alleles and the peptides recognized by the CTL response in the subtype C in comparison to the subtype B dataset were explored. Firstly, the binding motifs restricted by HLA alleles that are at high frequency in the Southern African population could be poorly characterized. The weak conformity between the predicted anchor residue motifs associated with patient HLA genotype and the peptides recognised by the CTL response could be because the HLA alleles found at high frequencies in the subtype C infected cohort are less likely to be associated with predicted HIV binding motifs in the databases. Secondly, the reason could be a result of the predicted binding motifs, which are possibly too specific for some HIV subtype sequences. Some of the anchor residue motifs could be too specific

for sequences observed in subtype B viruses even though the differences between the subtype B and C sequences do not cause failure to bind to HLA molecules. These two possible alternative explanations were explored to determine whether they are not mutually exclusive and the observed effect have involved a combination of the two.

4.4.5.1 Investigating the proportion of HLA alleles with characterized anchor residue motifs in the databases

The proportion of patient HLA alleles that have predicted anchor residue motifs available in the databases was determined for each of the subtype B and C cohorts (Table 4.4.5.1). Alleles from the HLA-A loci that were frequent in the subtype B cohort were mostly present in the database with characterized binding motifs compared to those frequent in the subtype C infected cohort. The HLA-B alleles frequent in the two cohorts were equally represented in the database. From a Fisher's exact test, all HLA alleles from the subtype B infected cohort appeared to be significantly associated with presence of an anchor residue motif in the database compared to alleles from the subtype C cohort (p value = 0.03). The statistic remained significant when looking at HLA-A alleles alone but not for the HLA-B alleles (p values = 0.01 and 0.75 respectively).

HLA locus	**Subtype C**	**Subtype B**	**P value (*Fisher's exact test*)**
HLA-A	59.7	76.7	0.01
HLA-B	76.9	77.2	0.75
HLA-A and HLA-B	66.8	76.4	0.03

Table 4.4.5.1: Percentages of individual occurrences of HLA alleles within a cohort for which anchor residue motifs were available in the databases.

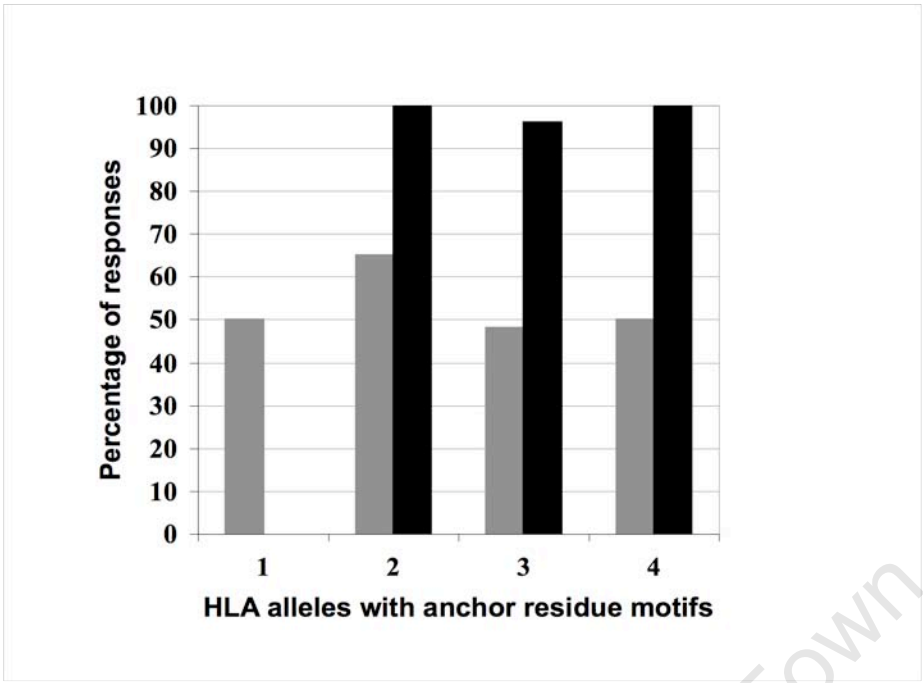
4.4.5.2 Investigating whether the anchor residue motif data is subtype-specific

The predicted HLA anchor residue motif data from the databases was used to determine patient HLA genotypes that conform to each observed CTL immune response. The motif predicted to be bound by each HLA allele from a patient was searched from the databases. For each patient, the HIV peptides observed to be targeted by the CTL immune response were then searched for the presence of the predicted anchor residue motifs restricted by each HLA allele. The distribution of HLA alleles with characterized anchor residue motifs available in the database as well as the proportion of the motifs that conformed to the observed CTL responses were compared between the subtypes B and C cohorts.

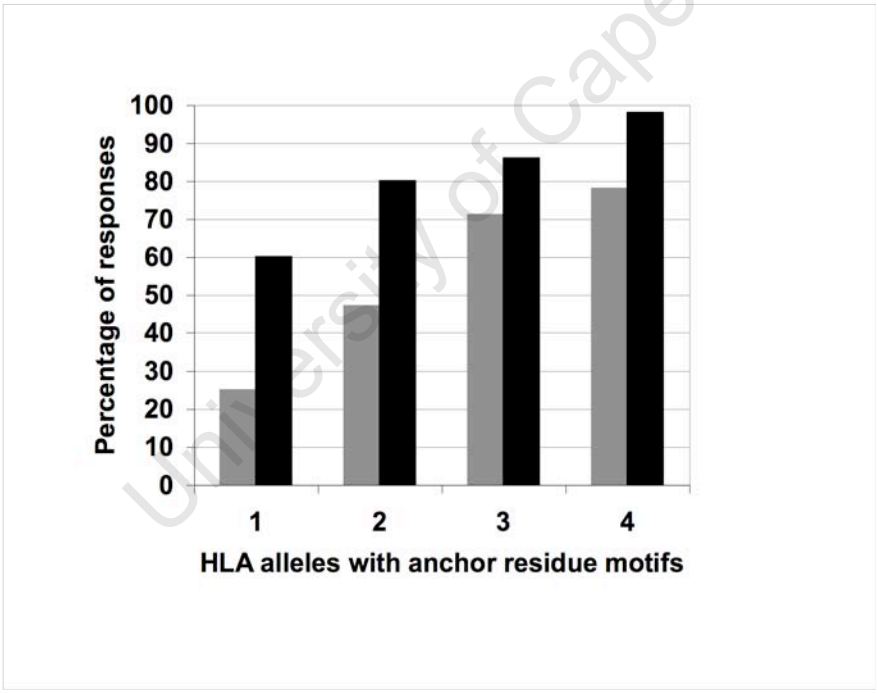
The possible bias of anchor residue motifs towards subtype B sequences was determined taking into account the differences in the proportions of alleles with characterized motifs. If most of the anchor residue motifs in the database were predicted from conserved patterns observed in HIV-1 subtype B epitopes, then it is possible that the anchor residue motifs reflect sequence diversity observed in subtype B to a greater extent than they reflect subtype C diversity. In order to determine if this holds true, inter-subtype comparison was made between patients with the same number of HLA alleles with characterized anchor residue motifs from the database. Patients were therefore classified as having 4, 3, 2, 1 or no alleles for which there was an anchor residue motif in the databases in both the C and B subtype cohorts. The proportions of observed CTL responses that could potentially be linked to the patient HLA alleles through the presence of an anchor residue motif in the targeted peptide were compared between groups of patients with the same numbers of characterized HLA alleles. In overall, there was a positive correlation between the proportion of CTL responses that could be linked to patient HLA genotype and the number of the patient's HLA alleles for which anchor residue motifs were available in the database in the case of Gag responses (one-way anova p values: subtype C = 0.002; subtype B = 0.0112). The same significant difference was not observed between the different groups for Nef responses.

The average proportion of responses that could be linked to patient HLA alleles within each group using available anchor residue motifs was consistently lower for

the subtype C data than for the subtype B data when patients were stratified according to the number of their alleles represented in the binding motif databases (p value = 0.02225, paired Wilcoxon rank-sum test). A two way ANOVA was carried out to test for an effect of subtype on the proportion of CTL responses that could be related to the patient HLA alleles when differences in the number of alleles that are in the databases were taken into account. Since patients with the same number of characterized HLA alleles were compared, the proportion of responses associated with patient HLA were expect to be the same between the two subtypes if the motifs are not biased towards one of the subtypes. In both the Nef and Gag responses, higher fractions of the targeted peptides contained anchor residue motifs restricted by at least one of the patient's HLA alleles compared to responses to subtype C peptides in groups of patients with same number of characterized HLA alleles (p values; Nef = 0.00004 and Gag = 0.002; Figure 4.4.5.2).



(a)



(b)

Figure 4.4.5.2: Proportion of CTL responses to peptides that contained an anchor residue motif for at least one of the patient's HLA A or B alleles in subtype C (grey) and subtype B (black), (a) responses observed against Nef peptides, (b) responses observed against Gag peptides. Patients were classified by the number of their HLA A or B alleles for which there were motifs available in the database (x-axis).

4.5 Discussion

The frequently targeted peptides in the middle region of the Nef protein sequence and the Gag p24 region appeared to be highly conserved, which is consistent with observations from previous studies (Gilbert *et al.,*2005; Masemola *et al.,*2004a; Mashishi *et al.,*2001; Yusim *et al.,*2002). However, the relationship between sequence diversity and the frequency of CTL recognition was not straightforward for Gag as *p17*, with a much greater number of variable sites, was also highly immunogenic. The Gag *p16* region and the terminal regions of the Nef protein sequence were highly variable and poorly targeted by the immune response. In future, autologous peptides could be considered for a similar analysis since some autologous viral sequences from the patients show a possibility of immune escape from the CTL response. In this study, only a proportion of the reactive peptides (20.6% of *nef* responses and 32.5% of *gag* responses) showed 100% sequence similarity with the autologous virus in 28.9% and 55% of Nef and Gag responders respectively.

The presence of anchor residue motifs restricted by a patient's HLA genotype in HIV-1 peptides observed to be targeted by the CTL response was determined in order to relate patient HLA genotype to the observed immune responses. Presence of anchor residue motifs known to be targeted by an HLA allele within a peptide does not necessarily imply that the HLA allele directed that particular immune response. However, a targeted peptide lacking at least one motif associated with the HLA alleles of the host shows inadequacy of the HLA anchor residue motif data. It was observed that HLA-A alleles common in Southern Africa were not well recorded in the databases and most of the anchor residue motifs were reflective of subtype B sequence peptides and failed to fully account for inter-subtype sequence diversity. However there were anchor residue motifs in the database for the majority of HLA-B alleles that were at high frequency in both the B and C cohorts. This could be because the HLA-B alleles are more extensively involved in presenting HIV peptides (Yusim *et al.,*2003) and are therefore better studied than other alleles.

This difference in the length of the peptides could introduce a bias in the comparison of the proportion of peptides that contained anchor residue motifs for a patient HLA

allele, since some of the matches between the peptides and the HLA anchor residue motifs could be the result of chance. The Nef peptides from the subtype C dataset were shorter (mean length = 15 amino acids) than the Nef peptides used in the subtype B study (mean length = 20 amino acids). However, the difference in mean length of the peptides alone is unlikely to explain the difference between the subtypes as the Gag peptides from the subtype C study were on average longer than the Gag peptides from subtype B (mean length 22 and 20 amino acids, respectively) and the difference between subtypes in the number of peptides containing the expected anchor residue motifs is consistent across the two genes.

Most of the anchor residue motifs used in this study were obtained from the Los Alamos database. These include motifs inferred using a variety of techniques. Some were determined from antigenic peptides endogenously bound to HLA molecule (Marsh *et al.*,2000). In some cases experimentally derived motifs for a given HLA allele were extended to HLA alleles with identical side chains at the binding pocket residues (Yusim *et al.,*2003). The database also includes predicted HLA motifs based on the conserved patterns found in at least two optimal HIV epitopes presented by an HLA allele (Yusim *et al.,*2003). In this latter case, the HIV optimal epitopes have been obtained experimentally using interferon-γ ELISpot, Cromium release or intracellular cytokine staining assays (Frahm *et al.,*2004). Binding motifs determined in this way are likely to be biased towards the more studied subtype-B than subtype-C virus sequence diversity, and this could explain why the anchor residue motifs appear to be more consistent with CTL responses to subtype B than to subtype C.

A more complete collection of HLA binding motifs would be useful for investigating viral escape from CD8+ T cell responses especially for studies focussing on developing effective anti-subtype C virus vaccines. Knowledge of the sequence motifs required for HLA binding can also be useful for monitoring vaccine efficacy. Breakthrough infections with mutations that are known to cause failure of patient HLA alleles to bind would suggest an effective immune response exerting selective pressure on the virus and could signal the spread of viral variants that have escaped the vaccine. Our results suggest that further efforts are required to elucidate the binding requirements of HLA alleles that are at relatively high frequency in South African populations and that efforts to determine the sequence binding requirements

of HLA alleles should take advantage of the global HIV-1 sequence diversity in order to determine the full range of sequence motifs that can be bound by a given HLA allele.

Chapter 5

Investigating HIV adaptation to host HLA background in global human populations

5.1 Summary

Human leukocyte antigens (HLA) present antigenic peptides that are recognized by cytotoxic T–lymphocytes during T-cell immune responses. Each HLA molecule requires the presence of a specific anchor residue motif in order to bind to a peptide. Immune escape mutations that occur at the anchor sites of peptides can hinder HLA recognition and binding. A mutation that allows escape from an immune response and which occurs at a site that is conserved due to functional or structural constraints is expected to revert to wildtype upon transmission to a host lacking the HLA allele required for the immune response. It has been suggested that HLA alleles that are frequent in a population are likely to force fixation of viral immune escape mutations and adaptation of the virus to the HLA allele. Such fixed escape mutations are expected to occur more often at variable sites of the HIV-1 sequence, which are not functionally conserved, because escape mutations at these sites do not involve a fitness cost to the virus. This hypothesis has been investigated using smaller study cohorts and limited to certain HIV-1 subtypes and proteins. HLA-directed immune escape mutations across all HIV-1 proteins isolated from six geographically distinct populations infected with one of the 3 most prevalent HIV-1 subtypes C, B and A1 were analysed in this study. The objective of this study was to test whether there was a lack of anchor residue motifs for common HLA alleles within the variable sites of HIV-1, consistent with the fixation of escape mutations in non-conserved sites. Even though some HLA genotypes were common in more than one population, they were associated with fewer anchor residues in only some of these populations and not in all of them. Further analysis of conserved regions and rare HLAs showed that both common and rare HLA genotypes appeared to be associated with lack of anchor residues in conserved regions. Therefore, the evidence for fixation of escape mutations at functionally unconstrained regions of HIV-1 may have been masked by

the opposing tendency for some alleles to cause functionally constrained regions to evolve rapidly through immune escape and reversion.

5.2 Background

Human Leukocyte antigens (HLAs) of class I direct cytotoxic T-lymphocyte (CTL) immune responses by recognising, binding to and presenting antigenic peptides to the CTL cells (Madden *et al.*,1992). Each HLA molecule binds to antigenic peptides which contain a specific sequence pattern known as an 'anchor residue motif' (Rammensee *et al.*,1995; Saper *et al.*,1991). Both the presence of an anchor residue motif as well as the overall binding energy between an HLA molecule and a peptide determine whether the peptide can successfully be transported and presented to the CTL (Bihl *et al.*,2006; Sette *et al.*,1994). HLA molecules are therefore key components of the CTL immune response and have been found to be associated with either rapid progression or long-term non-progression in HIV-1 infected individuals (Kiepiela *et al.*,2007; Ngumbela *et al.*,2008). The HLA gene itself is highly polymorphic and encodes a diverse set of HLA alleles, hence enabling the CTL immune response to target many types of antigenic proteins (Marsh *et al.*,2000). Closely related HLA alleles that share overlapping anchor residue motif requirements for binding have been grouped under the same supertype (Sette *et al.*,1999). The frequencies of some HLA alleles and supertypes differ between ethnic groups resulting in some inter-population differences in the frequency of targeting specific peptides and motifs (Cao *et al.,*2001; Hammond *et al.*,2007; Hollenbach *et al.,*2001).

The anchor residues given in anchor residue motifs are usually located at the second and C-terminal positions of peptides and directly contact the B and F binding pockets of the HLA binding groove respectively (Madden *et al.*,1992). It is at the binding pockets of the HLA groove that the highest binding affinity between the HLA and the peptide are obtained. The anchor residue motifs specify the most preferred binding residues at the anchor positions of peptides. These motifs can be used to predict CTL epitopes and to analyse immune escape patterns on HIV sequences. The presence of an anchor residue motif has been found to be associated with the presence of optimal CTL epitopes (Honeyborne *et al.*,2006; Nelson *et al.*,1997). It is not practical to use full length sequences of all peptides targeted by an individual HLA molecule since

amino acid variation is less restricted at the non-anchor sites such that peptides bound by a single HLA can vary. Anchor residue motifs are therefore the main determinants of HLA binding. Anchor residue motifs restricted by different HLA alleles have been deposited in various databases. The Los Alamos database (Korber *et al.,*2006) in particular accumulates anchor residue motifs from various sources, mainly from the SYFPEITHI database (Rammensee *et al.,*1999) and the HLA facts Book (Marsh *et al.*,2000).

Loss of optimally binding residues at the anchor sites of a peptide can cause an overall low HLA-peptide binding affinity and failure to elicit an immune response. In HIV-1 infection, mutations away from the wild-type residues located at important binding positions of viral peptides have been observed and mostly result in poor or loss of immune responses (McMichael *et al.*,2002). Such immune escape mutations either become fixed with time or revert to wild type depending on the presence or absence of the causative HLA molecule in the infected host and/or functional constraints of the target peptide region (Carlson *et al.*,2008; Kelleher *et al.*,2001; Moore *et al.*,2002; Peyerl *et al.*,2004). Fixation of a mutant can occur if the HLA exerting the strong selection pressure is frequent in a population such that the virus is exposed to the same immune pressure after transmission to new hosts which prevents its reversion (Kawashima *et al.*,2009; Trachtenberg *et al.*,2003). In some cases, compensatory mutations have been found to occur at adjacent sites or regions that restore the fitness of the virus (Crawford *et al.*,2007). The mutant residue can also become fixed if it enables the virus to adapt well to the host. Otherwise if the HLA allele causing an escape mutation is rare and the virus is transmitted to an individual lacking the HLA genotype, reversion to the wild-type residue is most likely to occur if the mutant compromises the fitness of the virus (Friedrich *et al.*,2004). Therefore, not only the frequency of an allele determines the fate of an immune escape mutation but the functional importance of the targeted sequence region in maintaining the fitness of the virus.

It has been suggested that HIV-1 can easily adapt to common HLA alleles which presumably target variable sites of the virus sequences and that individuals with rare HLA genotypes have an advantage (McMichael *et al.*,2002; Trachtenberg *et al.*,2003). Since sites that are conserved and under functional constraints are likely to

be under pressure to revert to wildtype in the absence of the causative HLA allele, it is expected that most mutations that become fixed occur at variable sites of the sequence which have no functional constraints. Studies on which these assumptions are based have been carried out on smaller datasets but the overall effect of all common HLA alleles in larger and different population across the most prevalent HIV-1 subtypes has not been assessed. This study set out to investigate HLA-driven immune escape mutations that cause loss of anchor residues in variable sites of HIV-1 sequences across the three most prevalent HIV-1 subtype sequences, C, B and A1 (Buonaguro *et al.*,2007). The specific objectives were to determine whether common HLA alleles within different populations are associated with fewer anchor residues than expected in variable sites of the different HIV-1 proteins. In addition, to evaluate whether loss of anchor residues frequently occurs at variable sites that are not likely to affect the fitness of the virus than at conserved regions. Evidence of positive selection pressure exerted by HLA alleles on known epitope regions is first evaluated in HIV-1 subtype C sequences from recently infected individuals from Southern Africa. The Nef and Gag proteins which contain peptides frequently targeted by the CTL response were used (Masemola *et al.*,2004a). Secondly, and on a larger scale, six major population groups infected with one of the three most prevalent HIV-1 subtypes were used to investigate HLA-directed loss of anchor residue motifs in variable sites. These populations include the developed countries infected with subtype B, i.e. America (US) and Australia (AU), Subtype C infected Brazilian (BR), Indian (IN) and South African (ZA) populations as well as the Kenyan (KE) population, infected with subtype A1.

5.3 Materials and Methods

5.3.1 Sequence data

Subtype C *nef* and *gag* codon sequences from a recently infected cohort from Southern Africa were provided by a collaborating laboratory headed by Prof Carolyn Williamson. The sequence data is described in Chapter 4 and in a publication listed at the beginning of this thesis (Ngandu *et al.*,2007). Pre-aligned HIV-1 amino acid gene sequence alignments for each of the six larger population groups infected with either subtype C, B or A1 were downloaded from the Los Alamos database (Leitner *et*

*al.,*2005) (Table 5.3.1). Regions with gaps in majority of the sequences and those judged by eye to be poorly aligned were manually removed from the alignment. Individual sequences with large stretches of gaps were excluded.

5.3.2 Determination of variable sites in HIV-1 sequences

For the subtype C sequences isolated from the Southern African cohort, the mutation rates at a site across each of the *nef* and *gag* genes were determined by estimating the nonsynonymous to synonymous substitution rate ratio (ω). The ω ratio was estimated over the phylogenetic tree relating the sequences using the MG94 codon model (Muse *et al.*,1994), implemented in the HyPhy package (Kosakovsky Pond *et al.,*2005d). The rate4site program (Mayrose *et al.,*2004; Pupko *et al.,*2002) was used to determine the rate of evolution at amino acid sites for each HIV-1 gene sequence alignment for each of the population regions. This analysis was carried out separately for each gene within each of the six population groups, where sequences from a single population belonged to the same HIV-1 subtype. The rate4site program outputs a score per site as a measure of conservation. Low scores, i.e., less than zero, indicate evolutionary conserved sites and scores above zero indicate variable sites. Positive selection was not analysed in the global population datasets but only in the local Southern African cohort. In all the individual gene alignments within each population group, a consensus sequence was generated and analysed for the distribution of HLA anchor residue motifs,

5.3.3 HLA anchor residue motif data

HLA anchor residue motifs were obtained from the Los Alamos database (Korber *et al.,*2006) which collects motifs from the literature and from databases such as SYFPEITHI (Rammensee *et al.,*1999) and the HLA Ligand/Motif database (Sathiamurthy *et al.,*2003). Motifs from other databases were present within these databases. All anchor residue motifs with lengths 8,9,10 and 11 and with at least two anchor residues (at the second and C-terminal positions) were used.

5.3.4 HLA allele frequency in different populations

HLA allele frequencies in the six major population groups used were obtained from the dbMHC HLA anthropology database (Meyer *et al.*,2007). The database provides good estimates of HLA allele frequencies in different populations worldwide. The supertypes for each allele were obtained from the Los Alamos immunology database (Korber *et al.,*2006) and the HLA Facts Book (Marsh *et al.*,2000). An allele was regarded as common in a population if its predicted frequency is greater than 0.05. The population group data used is shown in Table 5.3.1. The HLA allele frequencies do not link directly to the individuals from whom the sequences were obtained, they are representative of the frequencies at the population level. The data for individual-specific HLA genotypes is imiting for large analyses such as this and this issue has been discussed in 2.8.2.

Population Region	**HIV-1 sequence subtype**	**Mean number of sequences per gene**
Kenya (KE)	A1	15
America (US)	B	18
Australia (AU)	B	17
India (IN)	C	16
Brazil (BR)	C	15
South Africa (ZA)	C	17

Table 5.3.1: Populations for which the HIV-1 sequence data from the Los Alamos sequence database (Leitner *et al.,*2005) was analysed and for which the HLA frequency data was obtained from the dbMHC anthropology database (Meyer *et al.*,2007). The abbreviation for each population region is given in brackets.

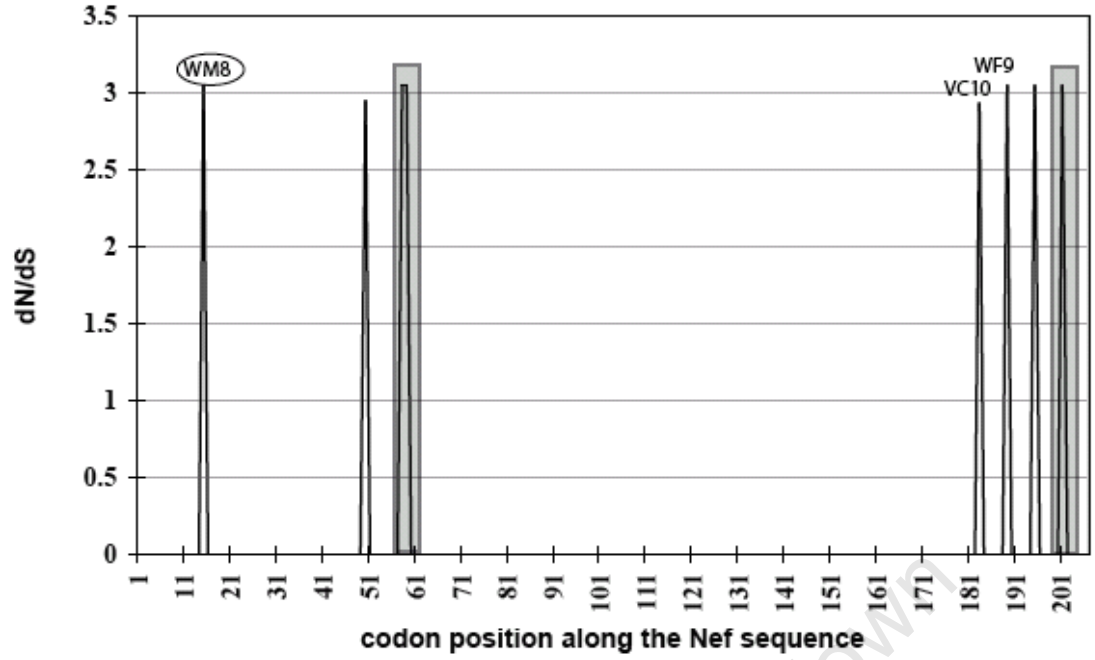
5.4 Results

The HLA-directed immune escape was first investigated using sequences from recent infections in a subtype C infected Southern African cohort. The presence of positive

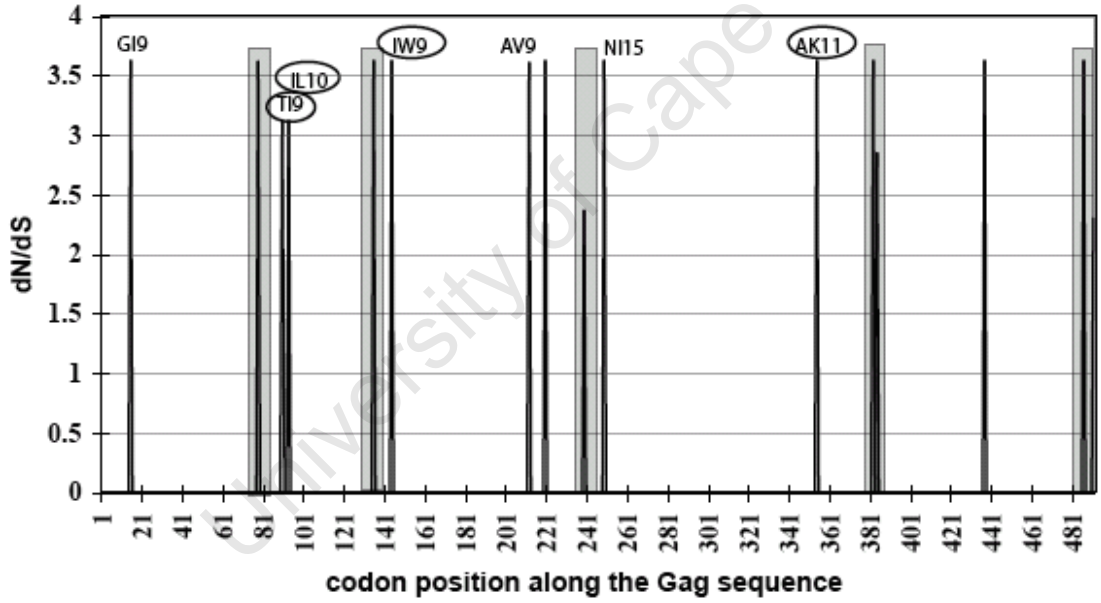
selection, i.e. ω>1, within known epitope regions listed in the Los Alamos database (Frahm *et al.,*2008) was investigated (section 5.4.1). Secondly in section 5.4.2, the loss of anchor motifs from variable sites associated with common alleles (in this chapter alleles that occur at a frequency greater than 0.05 in a population are considered to be common in that population) as well as rare HLA alleles (frequency of 0.05 or less) and supertypes across different populations was investigated. However, due to the observation in our previous publication, also discussed in Chapter 4, the anchor residue motifs in the databases appear to be biased towards certain populations and HIV-1 subtypes. If the predicted data equally represents responses observed against all subtypes and populations, then a direct inter-population comparison of the distribution of HLA anchor residue motifs can be made to determine whether anchor residues are less frequent in sequences from populations where a HLA allele is common. However, a direct comparison of the frequency of a motif between sequences from a population where the allele is rare and one where it is common can produce biased results. This is possible when motifs in one sequence are missed as a result of the difference in the anchor residues recognised by the same allele between different subtype sequences when the motif was only predicted using one of the subtypes. Therefore the motif can appear to be lost in one sequence yet the epitope is recognized and the predicted motif used in the analysis is too specific for the other sequence from a different HIV-1 subtype. In order to avoid this bias, the direct inter-population comparison is avoided and rather the association of a common/rare allele with the frequency of motifs in variable compared to conserved sites within a single subtype/population is investigated.

5.4.1 Positive selection within known epitope regions

Three of the five positively selected sites in *nef* and seven out of nine in *gag* were found within some known CTL targeted epitope regions listed in the Los Alamos database (Frahm *et al.,*2008) (Figure 5.4.1). Of these, one of the three Nef epitopes and four out of the seven Gag epitopes were previously observed to have sites exhibiting amino acid polymorphisms associated with the restricting HLA alleles (Brumme *et al.,*2007; Brumme *et al.,*2008b). Sites that were found to be positively selected but overlap with regions that are prone to false detection of positive selection (described in detail in chapter 3) were not considered.



(a)



(b)

Figure 5.4.1: Positively selected sites (black vertical lines) with ω >1 plotted across (a) Nef and (b) Gag amino acid sequence positions. Known optimal epitopes that overlap with positively selected sites are indicated above each corresponding site. Epitopes circled in black were also found to contain HLA-associated amino acid polymorphisms in previous studies by (Brumme *et al.,*2007; Brumme *et al.,*2008b). Sites with ω > 1 but which overlap with regions prone to false detection of positive selection (described in detail in Chapter 3) are shaded in grey and were not considered as positively selected in this study.

5.4.2 Determination of HLA-driven loss of anchor residue motifs in variable sites of HIV-1 sequences at the population level

The adaptation of virus sequences to CTL immune responses directed by common HLA alleles was investigated by determining the loss of HLA anchor residue motifs from variable sites of the HIV-1 sequence. A consensus sequence was generated from each HIV-1 sequence set isolated from the corresponding (to the HLA frequency data) geographical regions. A fisher's exact test was used to test whether a HLA allele or supertype is significantly associated with fewer anchor residue motifs in variable sites compared to conserved sites of a sequence. The test determines whether the HLA anchor residues within a motif (which is at the second and C-terminal positions for most motifs) are significantly associated with conserved sites as compared to variable sites. When the statistic shows a significant association of anchor residues with conserved sites, i.e. giving an odds ratio greater than 1 then it indicates that the variable sites significantly lack binding motifs for the particular HLA. The fewer anchor residue motifs than expected in variable sites of each HIV-1 gene was tested for all individual HLA alleles with anchor residue motifs available from the public databases. Each allele was tested against nine genes for each of the six population regions and the significant tests are given in Table 5.4.2.1. The analysis was also done by grouping all alleles of the same supertype since alleles within the same supertype share overlapping anchor residue motifs, hence in this way a single motif is evaluated once. The effect of common alleles (frequency greater than 0.05) within a supertype and that of rare alleles (frequency of 0.05 or less) was investigated separately (Figure 5.4.2.2 and Table 5.4.2.2). The effect of rare HLA genotypes was also tested to determine whether immune escape at the population level results only from common HLA alleles. Q-values (Storey *et al.,*2004) were determined for all p values obtained from the fisher's exact test to account for multiple hypothesis testing. A p value less than or equal to 0.05 with a q-value less than or equal to 0.2 was considered significant.

5.4.2.1 Loss of anchor residue motifs per HLA allele

Of 2,140 tests carried out for individual HLA alleles, 8 showed significant differences in the frequency of anchor residue motifs between conserved and variable sites (after correction for multiple hypothesis testing). The difference in frequency of anchor residue motifs between conserved and variable sites was observed in the Env, Gag and Pol protein sequences and associated with HLA s A*0201, A*0301, A*6801, B*2705, B*5301 and B*0801. However, a reduced frequency of anchor residue motifs was not only observed in variable sites but in conserved sites as well and was not only restricted to common HLA alleles (Table 5.4.2.1). Fewer motifs in variable sites (odds ratio significantly greater than one) was only observed in two cases, one in Env associated with a common HLA A*0201 in AU and the other in pol associated with a rare B*5301 allele in IN. Two of the three significant tests observed along the Pol sequence were associated with a lower frequency of motifs in conserved sites (odds ratio significantly less than one) restricted by a common B*2705 (US) and a rare B*0801 allele (AU). All four significant tests in Gag resulted from a reduced frequency of motifs in conserved sites only and were associated with either HLA A0301 or A6801 regardless of allele frequency.

HLA	[a]Type	[b]Gene	Frequency	Odds Ratio	P value	Q-value
A*0201	A2	ENV, B, AU	[c]0.127	[d]3.24	0.0004	0.16
A*0301	A3	GAG, B, AU	0.014	0.4	0.0001	0.11
A*0301	A3	GAG, C, ZA	[c]0.055	0.45	0.0006	0.16
A*6801	A3	GAG, B, AU	0.001	0.28	0.0002	0.14
A*6801	A3	GAG, C, IN	[c]0.08	0.31	0.0006	0.16
B*2705	B27	POL, B, US	[c]0.09	0.42	0.0005	0.16
B*5301	B7	POL, C, IN	0.01	[d] inf	0.0005	0.16
B*0801	B8	POL, B, AU	0.012	0.27	0.0001	0.11

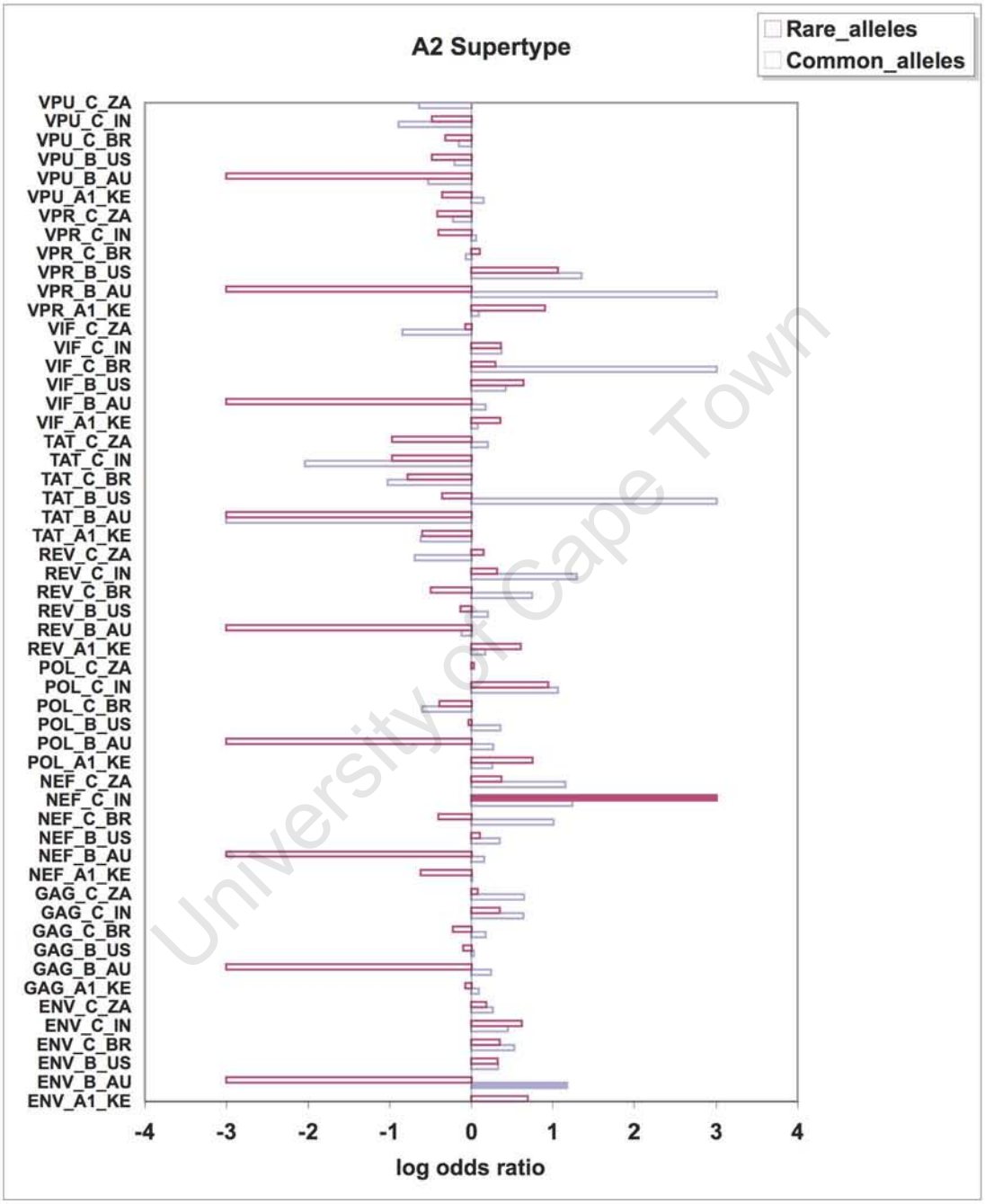
Table 5.4.2.1 HLA alleles and genes for which there was a significant difference in the frequency of anchor residue motifs between conserved and variable sites. [a] The corresponding supertype for each allele, [b] the gene, subtype and population region of the tested HIV-1 sequences, [c] the frequency of common alleles > 0.05), [d] Odds Ratio greater than 1 indicates a lack of motifs in variable sites.

5.4.2.2 Results for anchor residue motifs per supertype

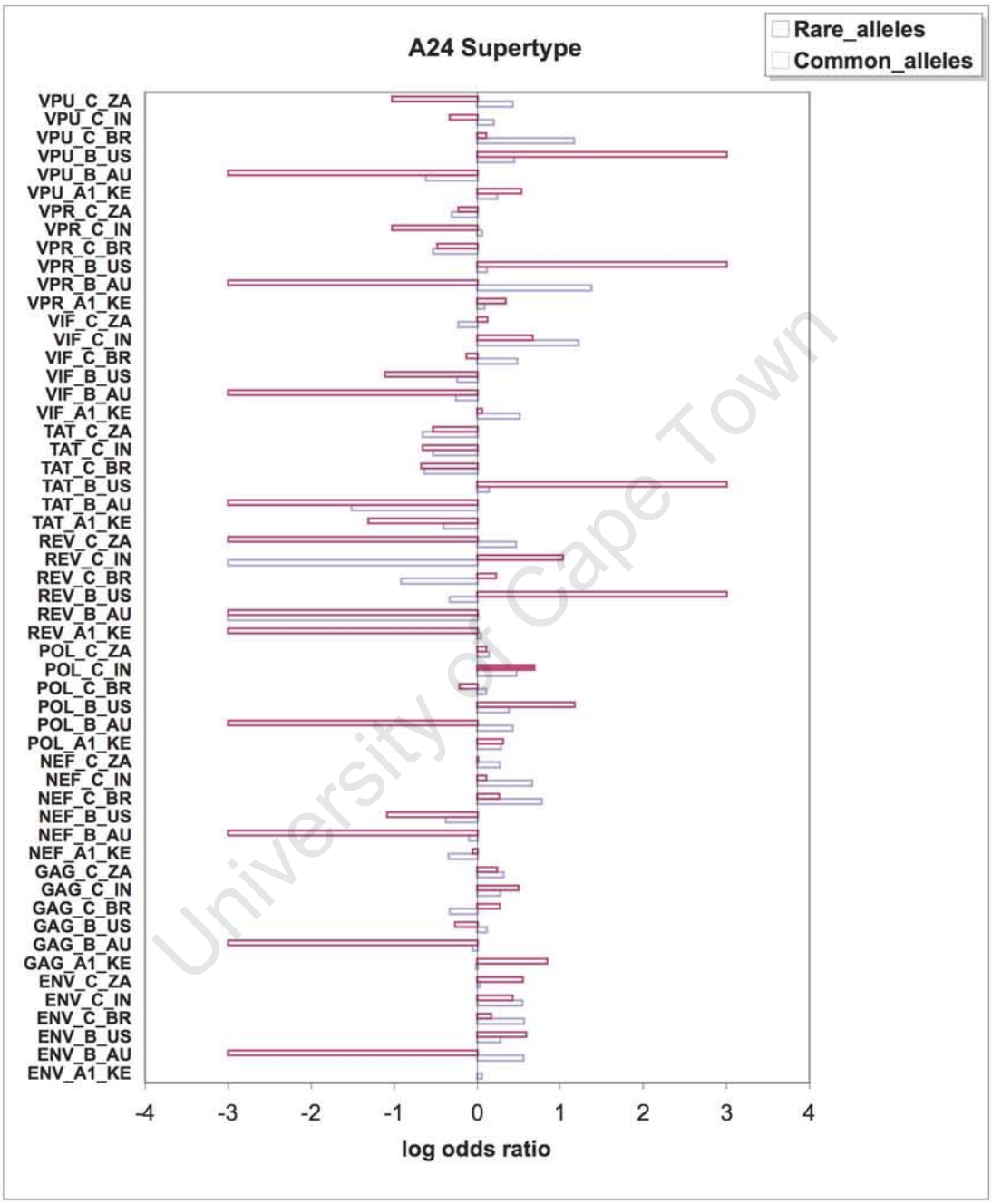
The HLA alleles were grouped into nine supertypes namely A1, A2, A3, A24, B7, B27, B44, B58 and B62 according to the classification in the Los Alamos database and the HLA Facts Book (Korber *et al.,*2006; Marsh *et al.*,2000; Sette *et al.,*1999). The frequency of overlapping motifs restricted by common and rare alleles within each supertype was investigated. All significant tests are presented in Table 5.4.2.2. There were no significant differences between conserved and variable sites in the frequency of consensus motifs associated with alleles of the A1, B62 and B7 supertypes (log odds ratios shown in the appendix). All tests for motifs restricted by rare and common alleles within each of the remaining six supertypes for each gene sequence per population group are presented in Figure 5.4.2.2. Similar to observations in individual alleles, significant tests were not restricted to fewer motifs in variable sites only nor associated with common supertypes only. No significant differences in the frequency of motifs between conserved and variable regions were observed in the Vpr and Vpu proteins. The Gag sequence had significantly fewer supertype anchor residue motifs in conserved sites as was the case when individual HLA alleles were considered separately. All but one of these significant tests were related to the HLA A3 supertype. The single significant test for Env was for the common A2 supertype alleles, which was associated with lack of motifs in variable sites (similarly to what was observed with the common A*0201 allele in the same AU dataset). The consensus motif restricted by rare alleles of the A2 supertype appeared to be less frequent in variable sites of the Nef sequences isolated from the subtype C infected Indian population. Significant differences in the frequency of motifs were observed in the Pol sequence. These included cases in which the frequency of motifs in variable sites was higher and cases in which it was lower than the frequency in conserved sites. The Rev, Tat and Vif proteins each showed a single significant test reflecting reduced frequencies in conserved sites of motifs associated with supertypes A3, B58 and B27, respectively.

HLA Supertype	Frequency	[a]Gene	Odds Ratio	P value	Q-value
A2	Common	ENV, B, AU	3.24	0.0004	0.02
A2	Rare	NEF, C, IN	5.0	0.004	0.16
A24	Rare	POL, C, IN	1.99	0.005	0.19
A3	Common	GAG, A1, KE	0.58	0.019	0.20
A3	Common	GAG, B, US	0.6	0.021	0.20
A3	Common	GAG, C, ZA	0.53	0.005	0.20
A3	Common	POL, B, AU	1.71	0.012	0.20
A3	Common	POL, C, IN	0.65	0.013	0.20
A3	Rare	GAG, B, AU	0.41	0.0001	0.005
A3	Rare	GAG, C, ZA	0.35	0.001	0.02
A3	Rare	POL, B, US	0.63	0.012	0.11
A3	Rare	POL, C, BR	0.52	0.011	0.11
A3	Rare	REV, A1, KE	0.08	0.009	0.11
B27	Common	POL, B, US	0.55	0.002	0.03
B27	Common	VIF, B, US	0.3	0.002	0.03
B44	Rare	POL, B, US	1.59	0.002	0.10
B58	Rare	GAG, B, US	0.5	0.008	0.20
B58	Rare	TAT, C, BR	0.15	0.002	0.09

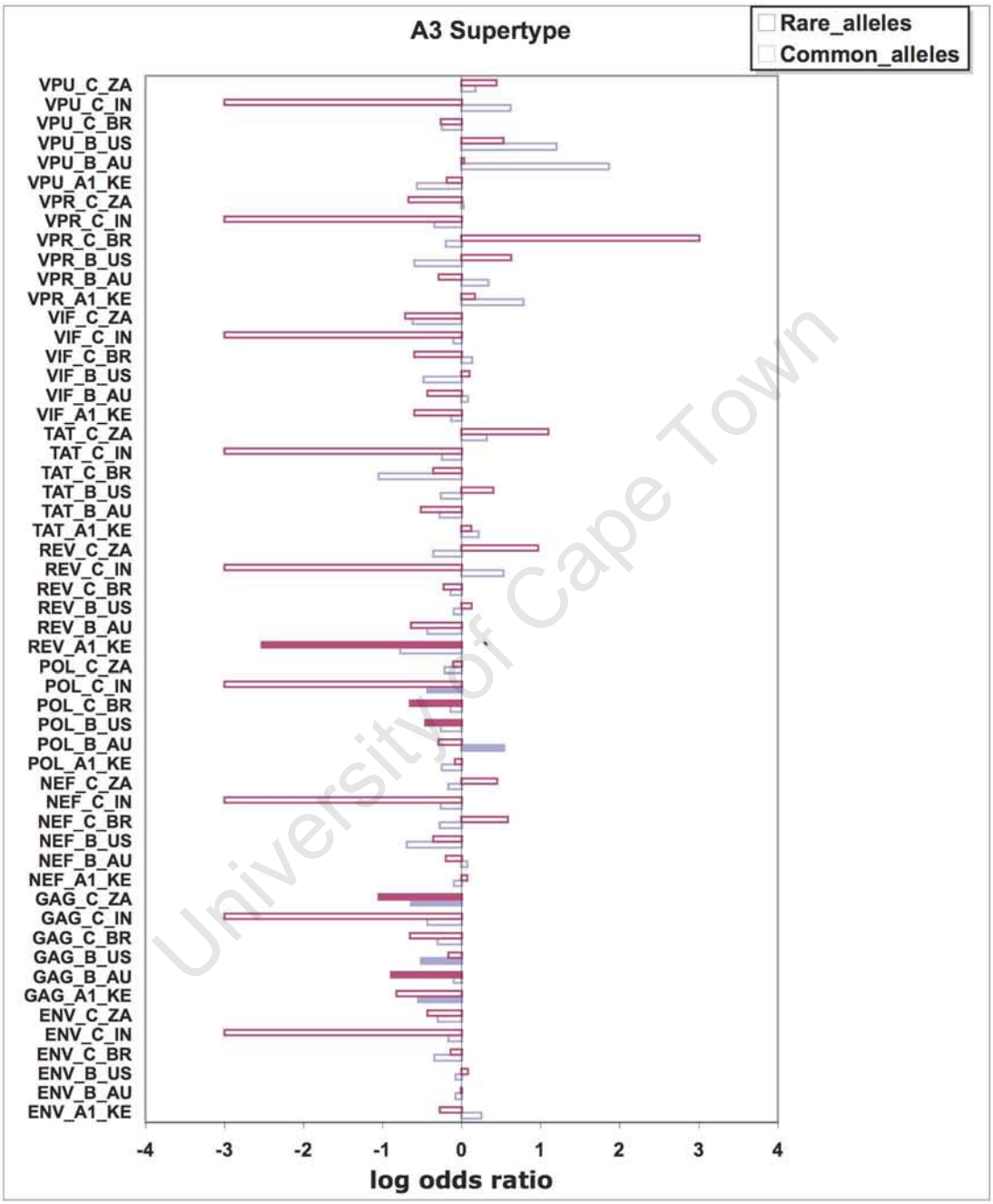
Table 5.4.2.2: HLA supertypes and genes for which there was a significant difference in the frequency of anchor residue motifs between conserved and variable sites, [a]Gene name, HIV-1 subtype, Country Code given in Table 3. A lower frequency of motifs in variable sites is indicated by an Odds Ratio>1.



(a)

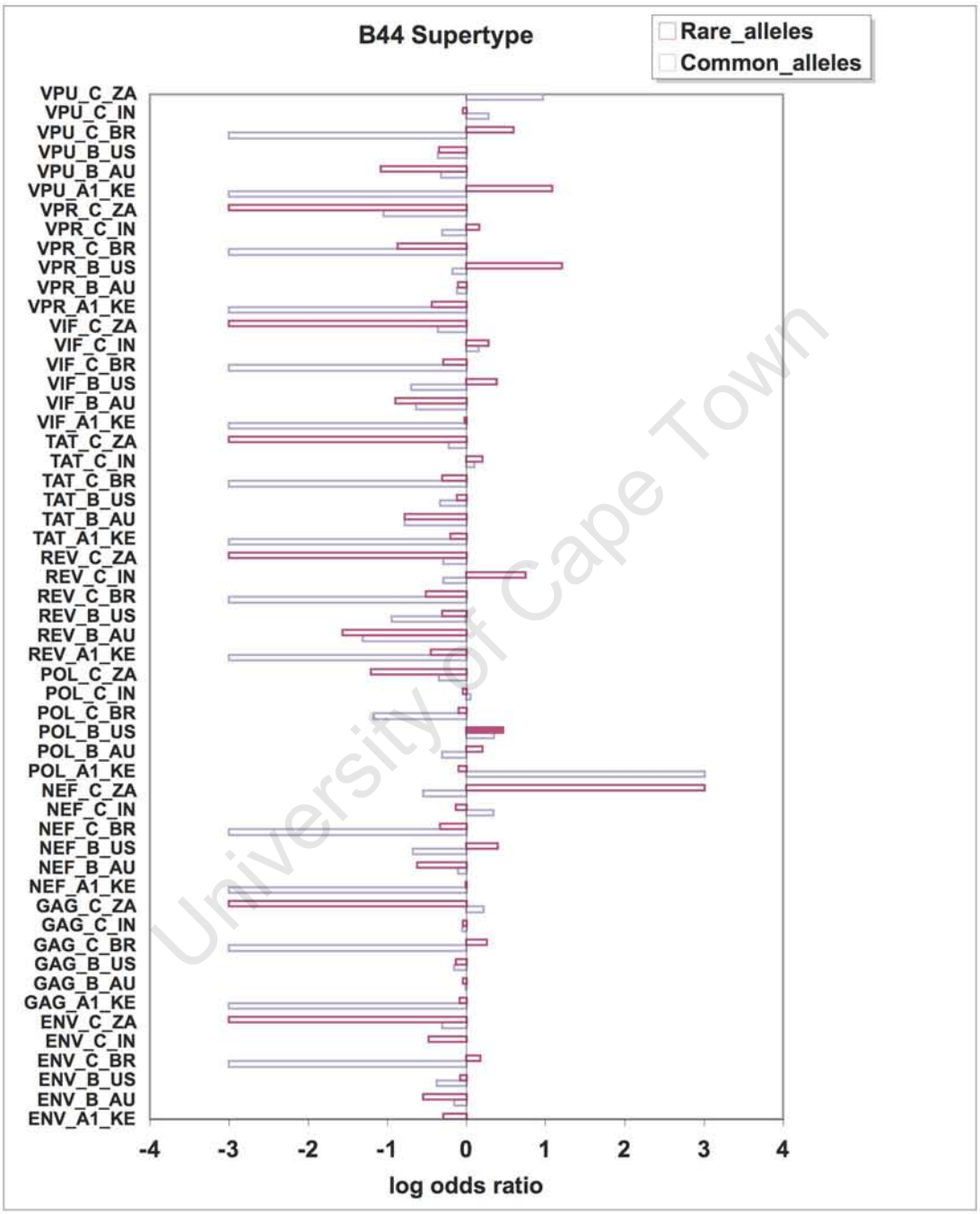


(b)



(c)

(d)



(e)

(f)

Figure 5.4.2.2: Log of the odds ratio values obtained from the association of fewer motifs with common alleles (blue) and rare alleles (red) within individual supertypes. Significant odds ratio values of association between common or rare alleles with fewer motifs are indicated by solid rectangles. Log odds ratios greater than 0 indicate

lack of motifs at variable sites and those less than 0 indicate lack of motifs at conserved sites. For the purposes of the plots, all infinite and zero odds ratio values were assigned the log values 3 and -3 respectively. All other log odds ratios were greater than -3 and less than 3.

5.4.2.3 Consistency between individual HLA alleles and corresponding supertypes

Results for five of the significant individual HLA allele tests were consistent with the corresponding supertypes that also showed significant results for the same protein and population region. Among the common HLA genotypes, the significant results from the A2 supertype could simply be a reflection of the effect of the common HLA-A0201 allele as this allele has been previously reported to be associated with immune escape that causes evolution of the virus (Edwards *et al.*,2005). Among the tests that showed lack of anchor residues in conserved sites, the common B2705 allele showed a similar pattern with the corresponding group of common alleles of the B27 supertype along the Pol gene from the subtype B-infected US population. The HLA-B27 supertype has been previously found to be associated with immune responses that result in reduced viral loads. The common A3 and A*0301 genotypes were also associated with lack of motifs in conserved sites of the Gag sequence from subtype C infected ZA population.

Two of the tests that were consistently significant between individual alleles and their supertypes were associated with rare HLA genotypes. The observation that the consensus motif shared by rare alleles of the A3 supertype was significantly less frequent in conserved sites of the Gag sequence from subtype B AU was consistent with results obtained from individual A3 rare alleles A0301 and A6801 in the same dataset. Although very few significant tests were observed, the *gag* gene consistently showed lower frequency of motifs in conserved sites in all six tests found to be significant regardless of population origin, subtype or HLA type. Gag is a conserved protein and most CTL epitopes are found in the highly conserved and immunogenic p24 region (Masemola *et al.*,2004b) and this could be the reason why in all cases the loss of motifs is found at conserved sites.

5.5 Discussion

In this study, some common HLA alleles were observed to be associated with fewer anchor residue motifs in variable sites of HIV-1 protein sequences, consistent with other previous observations that frequent HLA alleles cause fixation of immune escape mutations (Kawashima *et al.*,2009; Trachtenberg *et al.*,2003). However, this study was carried out on a larger scale, the data incorporating all HIV-1 genes and major pandemic subtypes from six distinct population regions. The large data size enabled the identification of rare patterns of HLA-directed immune escape mutations that have not been extensively reported in the past. One of these patterns is the lower frequency of anchor residue motifs than expected in conserved sites compared to variable sites. Generally, fixation of immune escape mutations is expected more frequently in variable sites that are not under functional constraints as compared to conserved sites. Therefore the frequency of anchor residues in conserved sites is not expected to be significantly lower than in variable sites. Another observation that has not been reported previously is that some rare HLA alleles were also associated with low frequencies of anchor residues. It can however be expected for rare HLA alleles, like common ones, to be associated with fewer motifs in variable sites since at such regions there is not always pressure to revert to wildtype in the absence of the causative HLA allele.

The association of either common or rare HLA alleles with significantly fewer anchor residue motifs in conserved sites is not expected to be frequent. This tendency for some genes to show fewer motifs in conserved sites than expected in association with some HLA alleles could be explained by CTL-driven evolution of epitopes in functionally constrained regions of the virus. The immune response exerting strong selection pressure on epitopes located in the conserved sites does to some extent cause these sites to evolve faster than they otherwise would. An interesting observation from this study was that all significant tests for the Gag protein (comprising 50% of significant tests from individual HLA alleles and 33% from supertypes) showed fewer motifs in conserved sites. This suggests that the immune response has a strong influence in the evolution of epitopes found in this protein, most of which are known to be in the functionally constrained p24 region.

As observed in the study, strong immune responses mediated by either common or rare HLA alleles can cause evolution of epitopes that are located in functionally constrained sites. Therefore, the effect of all individual HLA alleles on the evolution of the virus sequence needs to be determined regardless of their frequency in a population. The selection pressure exerted by the immune response causing conserved regions to evolve more rapidly can affect the way in which HLA anchor residues located at these sites are classified. Such functionally constrained epitope regions can appear as though they are not under functional constraints. In the analysis of sequence variation per site, these functionally constrained sites can produce high variation rates and the anchor residues can be classified as being in non-functionally important variable sites. This can cause the frequency of anchor residues in conserved sites to appear as though it is low and that there is higher tendency for fixation of immune escape mutations to occur in conserved sites compared to variable sites.

Some immune escape mutations either in conserved or variable sites can result in total loss of wild type anchor residue motifs and eventually failure of the specific immune response to target the region. This could have also affected this study because in such cases, epitope regions where fixation of mutant residues has occurred can appear as though they never contained CTL epitopes. Such regions can be mistakenly assumed to never been targeted by the immune response and not under immune selection pressure. These regions can be overlooked yet they contain important information with regard to the extent that the HIV-1 sequence evolves due to selection pressure exerted by the immune response. HIV-1 has been under selection from the human immune system since the early twentieth century when the virus was first transmitted from chimpanzees infected with simian immunodeficiency virus (SIVcpz) to humans. Therefore, there could be a number of regions across the HIV-1 genome especially in currently circulating viral sequences that have evolved over the years of HIV-1 infection in humans and appear to lack potential CTL epitopes. Information regarding the extent of HIV adaptation to the human immune response since the cross-species transmission event can be lost in such cases.

This study points to the complexity associated with analysis of HLA-related immune escape patterns in HIV-1 infection. It suggests that even though common HLA alleles tend to be associated with fewer motifs, the frequency of an HLA allele as well as

functional constraints in the HIV-1 sequence may not always determine the fate of an immune escape mutation. A further analysis is carried out in Chapter 6, in which the extent of selection pressure exerted by individual HLA alleles, regardless of HLA frequency and HIV-1 sequence constraints, is evaluated. The approach used is not affected by the potential misclassification of conserved regions and missing information discussed in the previous two paragraphs. It is carried out in such a way that the effect of the immune response on the evolution of all targeted sites in the HIV-1 sequence since the cross-species event can be determined. The SIVcpz sequence closely related to HIV-1 is used to infer sites that were targeted by the immune response since the first exposure of the virus to humans before any evolution due to human immune responses occurred. The HLA-directed selection pressure is determined across all regions of the HIV-1 sequence, inferred to have contained HLA epitopes before exposure to the human immune response.

Chapter 6

Evidence of HIV-1 adaptation to host HLA alleles following chimp-to-human transmission

A paper was recently published (Ngandu et al, 2009) reporting findings from this study and is listed at the beginning of the thesis.

6.1 Summary

Human leukocyte antigen (HLA) alleles can exert selection pressure on target peptide regions as they mediate cytotoxic T-lymphocyte immune responses against HIV-1 proteins. Immune escape mutations that occur at targeted sites can compromise virus fitness and simultaneously enable the virus to adapt to the immune response of the infected host. HIV-1 was transmitted to humans from chimpanzees infected with simian immunodeficiency virus (SIVcpz) in the early twentieth century; however, HIV-1 immune escape mutations that occurred during the zoonosis event and enabled the virus to adapt to the recipient human host HLA alleles have been poorly studied to date. Yet the adaptation of HIV-1 to human immune responses following zoonosis provides an opportunity to study how pathogens adapt to newly infected hosts. Here, the extent of adaptation of HIV-1 to human HLAs during the initial exposure of the virus to the human host immune responses following the zoonosis event was investigated. A SIVcpz consensus sequence was generated and used to approximate the virus sequence that was initially transmitted to the human host. A method based on HLA-peptide binding affinity was used to predict peptides that were potentially targeted by human HLA alleles on this sequence. Only HLA alleles with solved crystal structures could be used in this approach. The branch of the phylogenetic tree leading to the common ancestor of all of the HIV-1 sequences was used to investigate selection pressure that may have been exerted by individual HLA alleles following the transmission event. The average selection coefficient along this branch was estimated using codon-based phylogenetic models. Two methods were used, the branch *a priori* which allowed the nonsynonymous to synonymous substitution rate ratio (ω) to vary only along the HIV-1 ancestral branch and the GABranch algorithm which assigns

each branch of the phylogeny to the best fitting ω class. Evidence for adaptive evolution following the zoonosis event was observed at regions recognised by HLA A*6801 and A*0201. No evidence of adaptive evolution was observed along sites targeted by HLA-B*2705. Crystal structures of all HLA alleles involved in anti-HIV-1 immune responses need to be solved in order to fully investigate the extent of zoonotic adaptation of HIV-1 to human immune responses.

6.2 Background

Phylogenetic analysis indicates that the human immunodeficiency virus type 1 (HIV-1) originated from simian immunodeficiency virus infecting chimpanzees (SIVcpz) through a chimpanzee-to-human zoonotic transmission (Gao *et al.,*1999; Korber *et al.,*2000; Sharp *et al.,*1999; Zhu *et al.,*1998). The natural hosts of the virus, the chimpanzee, have been observed to remain asymptomatic throughout infection despite high viral loads (Pandrea *et al.,*2008; Silvestri *et al.,*2007; Silvestri,2008). It is only until recently (Keele, 2009) that AIDS-related symptoms have been reported in the chimpanzee. In humans however, an increase in viral load is usually associated with progression to the acquired immuno-deficiency syndrome (AIDS) and subsequently death (Lemey *et al.,*2007; Michael *et al.,*1995; Musey *et al.,*1997; Saksela *et al.,*1994; Shearer *et al.,*1997). The exact causes of these differences in host-viral relationships between the chimpanzee and humans are continuously being investigated by scientists worldwide. The causes of the difference in disease progression may involve either differences in the host and/or between the HIV-1 and the SIVcpz viruses.

A cross-species zoonotic event is expected to be accompanied by mutations that enable the pathogen to adapt to the new host environment. Indeed, sequence changes have been identified in HIV-1 that are evidence of selection pressure associated with the genetics of the human host (Brumme *et al.,*2007; Choisy *et al.,*2004; Soares *et al.,*2008; Wain *et al.,*2007). The cytotoxic T-lymphocyte (CTL) immune response directed against foreign antigens in particular plays a major role in exerting selective pressure on the antigenic proteins including those of HIV-1. The activation and characteristics of the immune responses against the virus have been found to differ remarkably between the two hosts (Bibollet-Ruche *et al.,*2008; Muller *et al.,*2006; Rutjens *et al.,*2003; Silvestri,2008). An elevated anti-HIV immune response upon

infection is characteristic in humans but the chimpanzee generally maintains a low level of immune activation (Bibollet-Ruche *et al.,*2008; Muller *et al.,*2006; Rutjens *et al.,*2003; Silvestri,2008). The human immune response therefore possibly exerts higher selection pressure on the virus sequence compared to immune responses of the natural host. However, the virus is still able to overcome the immune response and cause AIDS indicating that there are mutations which occurred immediately after the zoonotic transmission from chimpanzee to humans which enabled the virus to adapt to some extent to the human immune response.

Despite the efforts that have been made in analyzing selective pressures exerted by specific immune responses in infected populations within the HIV-1 group M subtypes, there has been limited analysis of the effect of the human CTL immune response on the viral sequence during the process of establishment of the virus in humans following transmission from chimpanzee. Yet a detailed understanding of the HIV sequence changes that occurred following the zoonosis process as a result of selective pressure exerted by the human immune response can broaden current knowledge of HIV-1 pathogenesis as well as the adaptation of pathogens to the human host. If the selection pressure exerted by the human immune response was strong enough to cause positive selection of the targeted sites and subsequent fixation of the mutations, it may be possible to identify such signatures of adaptation to new host following zoonosis by comparing the inferred ancestral sequence or the consensus sequence of HIV-1 group M subtypes to the consensus sequence of SIVcpz.

Anchor residue motifs can be used to predict epitopes bound by the various human leukocyte antigen (HLA) molecules in mediating CTL responses against HIV-1. However, successful binding, efficient transport and presentation of a peptide to a CTL depends on both the presence of the appropriate anchor residue motif and the overall affinity between the HLA binding groove and the peptide residues (Rovero *et al.,*1994; Sette *et al.,*1994). Therefore, the search for anchor residue motifs together with the prediction of the strength of binding between a HLA molecule can be used to best infer regions of the virus sequence that are likely targets of the CTL response *in vivo*. A variety of methods are currently used to predict anchor residue motifs as well

as the HLA-peptide binding affinity (DiBrino *et al.,*1994; Falk *et al.,*1991; Rammensee *et al.,*1999; Schiewe *et al.,*2007).

In this study, the HLA target regions were predicted using a structure-based method that determines the strength of binding between a sequence and HLA molecules using amino acid pair-wise potentials. Only regions with the highest binding affinity and anchor residue motifs are selected to ensure that most of the optimal epitopes recognised *in vivo* are detected with minimal false positives. This study limits the analysis of selection pressure only to potential HLA target sites across the HIV-1 genome so as to exclude sites that are not under selection pressure resulting from the CTL response. Models of codon sequence evolution were used to determine selection pressure in regions predicted to be targeted by specific HLA molecules. The methods employed for detecting selection pressure infer positive selection from the ratio of nonsynonymous substitution rates (dN) to synonymous substitution rates (dS) for individual branches in a phylogeny. Branch-specific analysis of selection pressure enabled the determination of whether there was evidence of strong selection pressure along the ancestral branch leading to the HIV sequences associated with each HLA allele. HIV-1 adaptation to HLA alleles with solved crystal structures is presented and this contributes to the understanding of how this pathogen adapted to the human host upon transmission.

6.3 Materials and Methods

6.3.1 Sequence data

An alignment of HIV-1 group M reference genome sequences and chimpanzee sequences was downloaded from the Los Alamos database (Leitner *et al.,*2005). Regions with gaps in majority of the sequences and those judged by eye to be poorly aligned were manually removed from the alignment. Individual sequences with large stretches of gaps were excluded. Regions that were found to be under strong purifying selection at synonymous sites (as described in chapter 3) were removed from the alignment. The resulting alignment consisted of 9 chimpanzee sequences and 32 HIV-1 sequences starting from codon 1 of the *gag* gene ending with the *nef* stop codon. The GenBank accession numbers of the sequences used are: 'DQ676872', 'AF004885',

'AB253421', 'AB253429', 'AF286238', 'AF286237', 'K03455', 'AY173951', 'AY331295', 'DQ853463', 'U52953', 'U46016', 'AY772699', 'K03454', 'AY253311', 'U88824', 'AF077336', 'AF005494', 'AF075703', 'AJ249238', 'AF377956', 'AF084936', 'AF061641', 'U88826', 'AY612637', 'AF190127', 'AF190128', 'AF005496', 'AF082394', 'AF082395', 'AJ249235', 'AJ249239', 'U42720', 'DQ373066', 'AF103818', 'AY169968', 'DQ373065', 'DQ373064', 'DQ373063', 'X52154' and 'AF447763'. We used HyPhy to build a phylogenetic tree from the alignment using a neighbour joining method and the pairwise distances calculated using maximum likelihood (Kosakovsky Pond *et al.,*2005d).

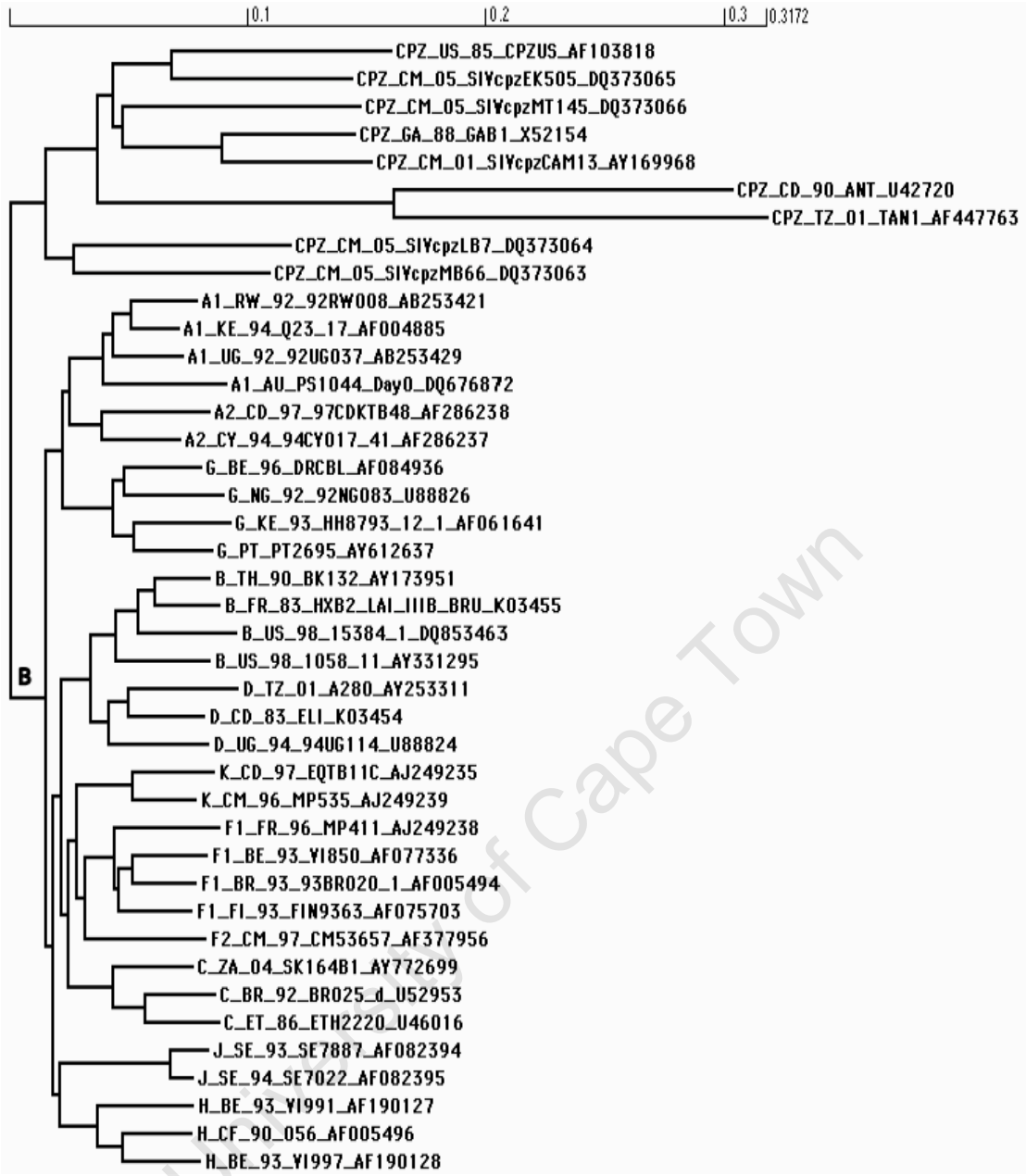


Figure 6.3.1: The phylogenetic tree of the 32 HIV-1 group M reference sequences and 9 SIVcpz sequences from the Los Alamos database comprising of all coding genes (Leitner *et al.,*2005). The chimpanzee sequence names start with 'CPZ' and the group M sequences, from branch "B", start with the subtype name. The branch lengths are scaled in reference to the scale given at the top of the tree.

6.3.2 Predicting HLA binding regions

PREDEP (Altuvia *et al.,*2004), a structure-based method for predicting HLA binding peptides was used to determine potential binding regions across the genome. The consensus chimpanzee sequence was used to predict the best HLA binding regions because it has not been exposed to the human HLA selection pressure. Consequently, it may be possible to detect epitopes in the chimpanzee sequence eliminated from HIV-1 shortly after transmission to humans. PREDEP does not require knowledge of known HLA-binding peptides hence is not biased towards any particular organism. The program requires solved crystal structures of the HLA molecules as well as knowledge of amino acid residues on the binding groove that interact with each position of the antigenic peptide sequence. Amino acid pair-wise potentials are used to determine the strength of binding between the peptide and the amino acids in the HLA binding groove based on backbone and side-chain interactions.

A score for each HLA-peptide interaction is calculated as the sum of amino acid pair-wise potentials between each peptide residue and the interacting residues of the HLA binding groove. The lower the score, the better the peptide binds to the HLA binding groove. Peptides with very low interaction scores have strong binding affinities to the HLA molecule hence are most likely to be targeted and successfully presented to CTLs *in-vivo*. A test of PREDEP performance showed that 80% of the top 15 percentile best binders were known optimal HLA binding peptides (Altuvia *et al.,*2004). In this study, only the top 5% of best binding regions containing the amino acid residues known to give optimal binding at the major binding pockets, i.e. peptides that matched the HLA anchor residue motifs, were taken for further analysis. All potential binding regions across the chimpanzee consensus sequence were predicted in this manner for each HLA allele. The downloaded HIV and chimpanzee reference sequence alignment was then edited for each HLA allele by taking only the sites in the alignment that fell within the potential binding regions predicted in the chimpanzee consensus sequence. This resulted in a new alignment for each HLA allele consisting of only the regions of the sequence likely to be bound by the HLA allele. In this way, the extent of selective pressure exclusively exerted by immune responses mediated by each individual HLA alleles could be evaluated. In particular, along the ancestral branch of HIV-1 sequences that is closely related to SIVcpz,

labelled 'B' in Figure 6.3.1, because it represents the sequence that encountered the first human immune responses immediately after the chimpanzee-to-human cross-species transmission event.

6.3.3 *A priori* analysis of selection pressure along the HIV branch compared to the SIV branch

The BranchAPriori model (Yang,1998) was used to determine whether there is high selection pressure on the branch ('B' in Figue 6.3.1), leading to the HIV sequences. For each HLA-related alignment described in the previous section, selection pressure along this branch was compared to the rest of the branches in the tree. Therefore, this particular branch was selected *a priori*. The program outputs a p value derived from the difference in the log likelihood between the null and the alternative models. The null model assumes a single global dN/dS ratio (ω) across the tree and ω for the selected branch leading to HIV-1 sequences is unconstrained in the alternative model. This *a priori* analysis however has the disadvantage that it assumes that all the other branches in the rest of the tree are under uniform selection pressure. The power of the analysis can be weakened for instance when there is actually strong among-branch heterogeneity in selection pressure between the branches in the rest of the tree (Yang,1998). It is therefore preferable to consider models that allow selection pressure along all branches in the tree to vary in order to construct a more realistic null model.

6.3.4 Branch-by-branch analysis of selection pressure

GABranch, a genetic algorithm (Kosakovsky Pond *et al.,*2005b) implemented in HyPhy was used to determine branch-specific selection pressure across the entire phylogeny of SIV and HIV-1 sequences. The genetic algorithm allows among branch variation in selection pressure across the entire tree. It calculates selection pressure along each branch of a tree and hence can be used to determine which branches of a tree are evolving under positive selection pressure. We therefore set out to determine whether there is positive selection pressure on the branch leading to HIV-1 sequences from the SIVcpz-HIV-1 ancestral node for each of the HLA alleles being analyzed. The best fitting nucleotide model for input in the genetic algorithm analysis was

determined using a maximum likelihood-based tool available in HyPhy (Kosakovsky Pond *et al.,*2005d). The GABranch algorithm first searches for the best fitting codon model for the phylogeny. All possible codon models are tested with varying ω rate classes starting with a model with a single rate class, i.e. a model that tests whether all the branches of the tree evolve uniformly. Models with more than 1 rate class are then tested whereby the evolutionary rate of each tree branch is tested for fit to each of the rate classes before assigning to the best fitting rate class. An Akaike Information Criterion weight (AIC) is calculated for each model based on its fit to the data in comparison to the model that assumes uniform selection across all the branches, i.e., the single rate model. The model with the best fit to the data as indicated by the lowest AIC is selected. The branches that fall under each ω rate class of the best fitting model are indicated on the phylogeny. Additionally for each branch, the averaged proportion of all tested models that showed support for dN > dS i.e., ω>1, for each branch i.e. all tested model that indicated that ω was significantly greater than 1 along a particular branch, is provided.

6.4 Results

6.4.1 Prediction of HLA binding regions in the chimp sequence

Of the six HLA alleles for which crystal structure and peptide binding preferences have been analyzed using Predep (http://bioinfo.md.huji.ac.il/marg/Teppred/mhc-bind/index.html), only HLAs A0201, A6801 and B2705 showed strong binding to regions of the SIVcpz consensus sequence which also contained the preferred anchor residue motifs. The total length of the overlapping 9mer peptide regions predicted to be the best binders, i.e. within the best five percent scores, for each HLA molecule across the chimp genome are given in Table 6.4.1. These predicted binding sites were distributed across different genes of the genome.

HLA allele	Anchor residue motif[1]	Binding regions for positive selection analysis
A0201	.[AILTVM]……[AILTVM]	395 codons
A6801	[AILTVM]……[RK].	345 codons
B2705	.RK……LFYRKHMI.	148 codons

Table 6.4.1 Sequence data for the regions predicted to have potential HLA binding peptides. [1] The anchor residue motifs were predicted from HLA-peptide structural conformations in Predep, residues in square brackets are the most preferred at the specific anchor site and the dots represent any other amino acid.

6.4.2 BranchApriori analysis of differential selection between the HIV and SIV lineages

The predicted binding regions for each HLA allele were tested for differential selection pressure between the HIV lineage and the SIV branches by estimating ω along the branch leading to the HIV-1 sequences. The ω was found to be higher in the branch leading to the HIV sequences compared to the SIVcpz branches although in both lineages it was greater than one (Table 6.4.2). Significant ω differences were observed only for HLA A6801 with a ω of 3.5 in the branch leading to the HIV-1 sequences and 1.6 in the rest of the tree (p value = 0.04). Selection pressure acting along the HIV-1 branch for sites associated with A0201 (ω = 2.5) was very high but failed to differ significantly from SIVcpz (ω = 1.2, p value = 0.08). For sites associated with B2705, there was no significant difference between HIV-1 and SIVcpz branches with no indication of positive selection pressure in HIV-1 and slightly low in SIVcpz (ω = 1.0 in HIV-1 and 1.1 in SIVcpz, p value = 0.65).

Allele	HIV branch ω	Rest of the tree ω	p value
A0201	2.5	1.2	0.08
A6801	3.5	1.6	0.04
B2705	1.0	1.1	0.65

Table 6.4.2 BranchApriori ω values for the HIV and SIVcpz lineage branches

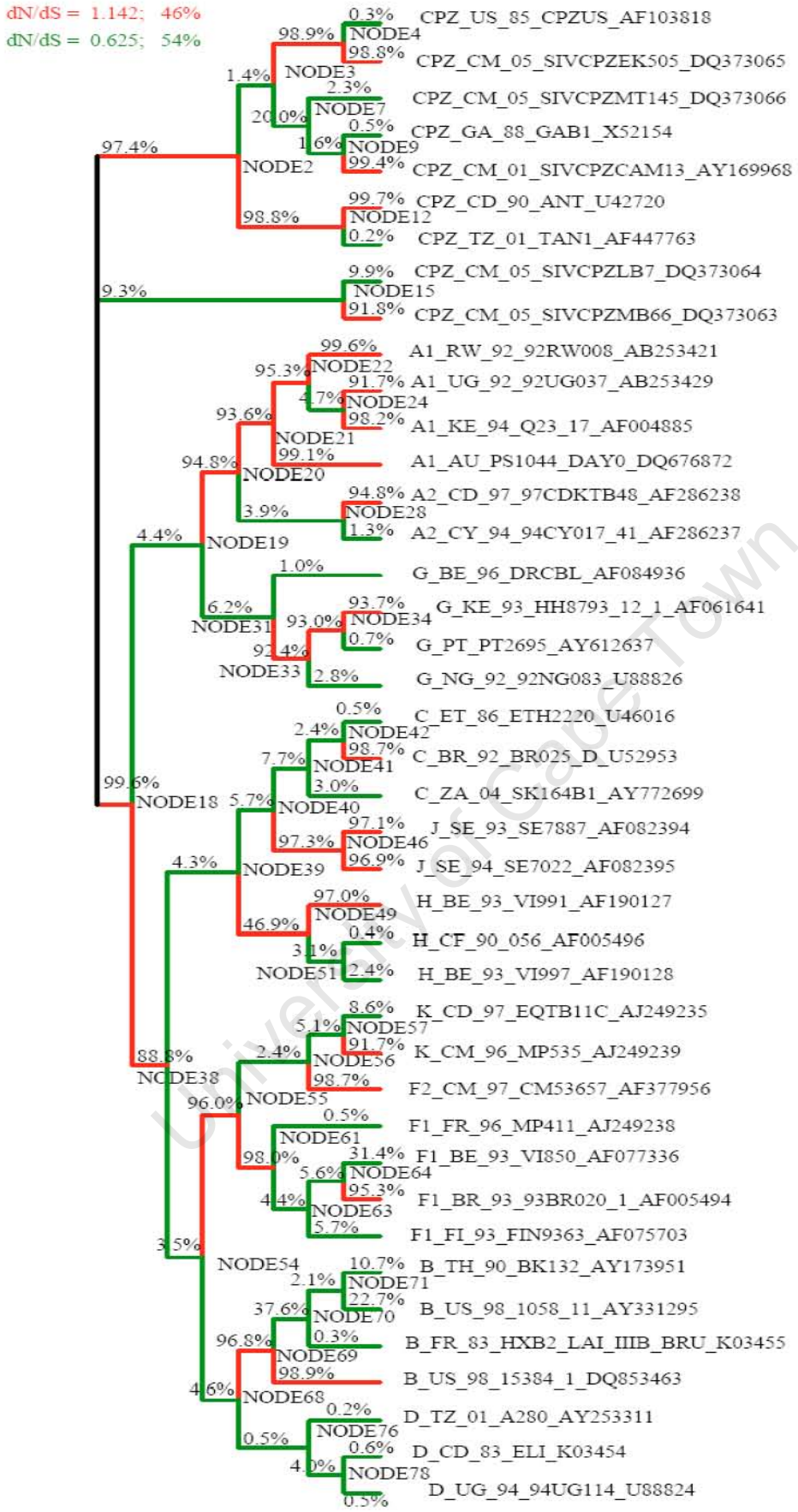
6.4.3 Branch-by-branch analysis of selection pressure using the GABranch algorithm

All possible codon models were run on the sequence alignments of predicted binding regions from each HLA allele and output given for the best fitting model. The ω rate classes for the best fitting model as well as the number of branches that fit to each class are given in Table 6.4.3. The proportion of models that have a high support for ω>1 for each branch is also given and highlighted in Figures 6.2a-c. No model support for ω>1 was observed on any branch in the phylogeny generated from the HLA B2705 (Figure 6.4.3c) predicted binding sites, suggesting that this allele may not have exerted strong selection pressure on the HIV-1 sequence. The values of ω along the branch leading to HIV-1 (from node 18 in Figure 6.2) as well as the proportion of models that showed support for ω>1 are shown in Table 6.4.3. The results strongly agreed with those observed from the *a priori* analysis. Selection pressure observed along the branch leading to HIV-1 sequences containing potential HLA-A6801 binding sites (from node 18) was high (ω = 1.15). A very high proportion of the tested models (0.996) showed a high support for ω>1 along this branch. There was also evidence of strong selection pressure along the HIV-1 branch for regions associated with targeting of HLA A0201 (ω = 1.14) and 0.996 of models showed support for ω>1. The mean omega values and model support data for all the 79 branches of the three trees are given in the appendix Tables A6.4.3a-c.

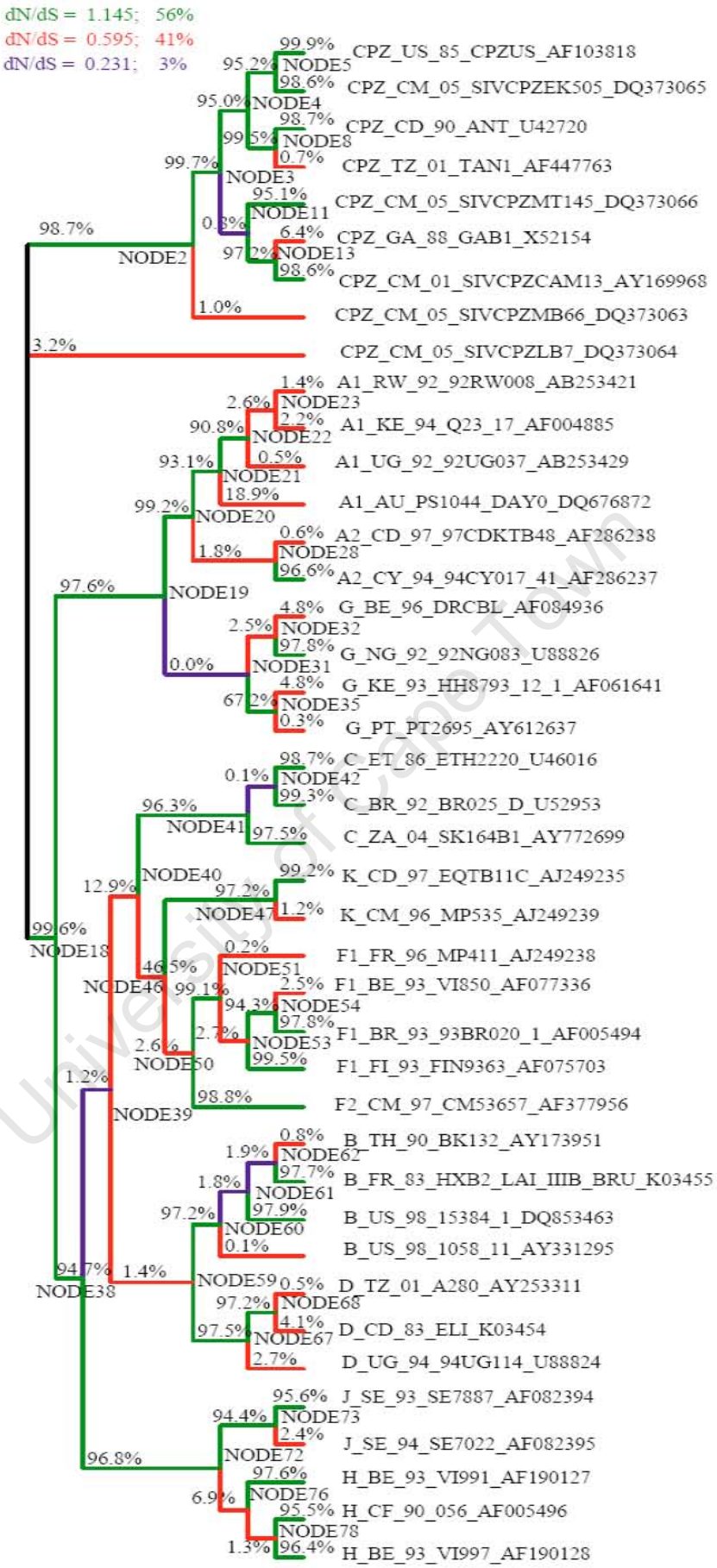
HLA allele	**ω rate classes (number of branches)[1]**	**ω HIV-1[2]** (from node 18)	**Prob(ω>1)[3] HIV-1**
A0201	1.14 (33), 0.62 (46)	1.14	0.996
A6801	1.15 (42), 0.60 (31), 0.23(6)	1.15	0.996
B2705	0.52 (79)	0.52	0.06

Table 6.4.3 The best fitting models determined from the Genetic algorithm analysis

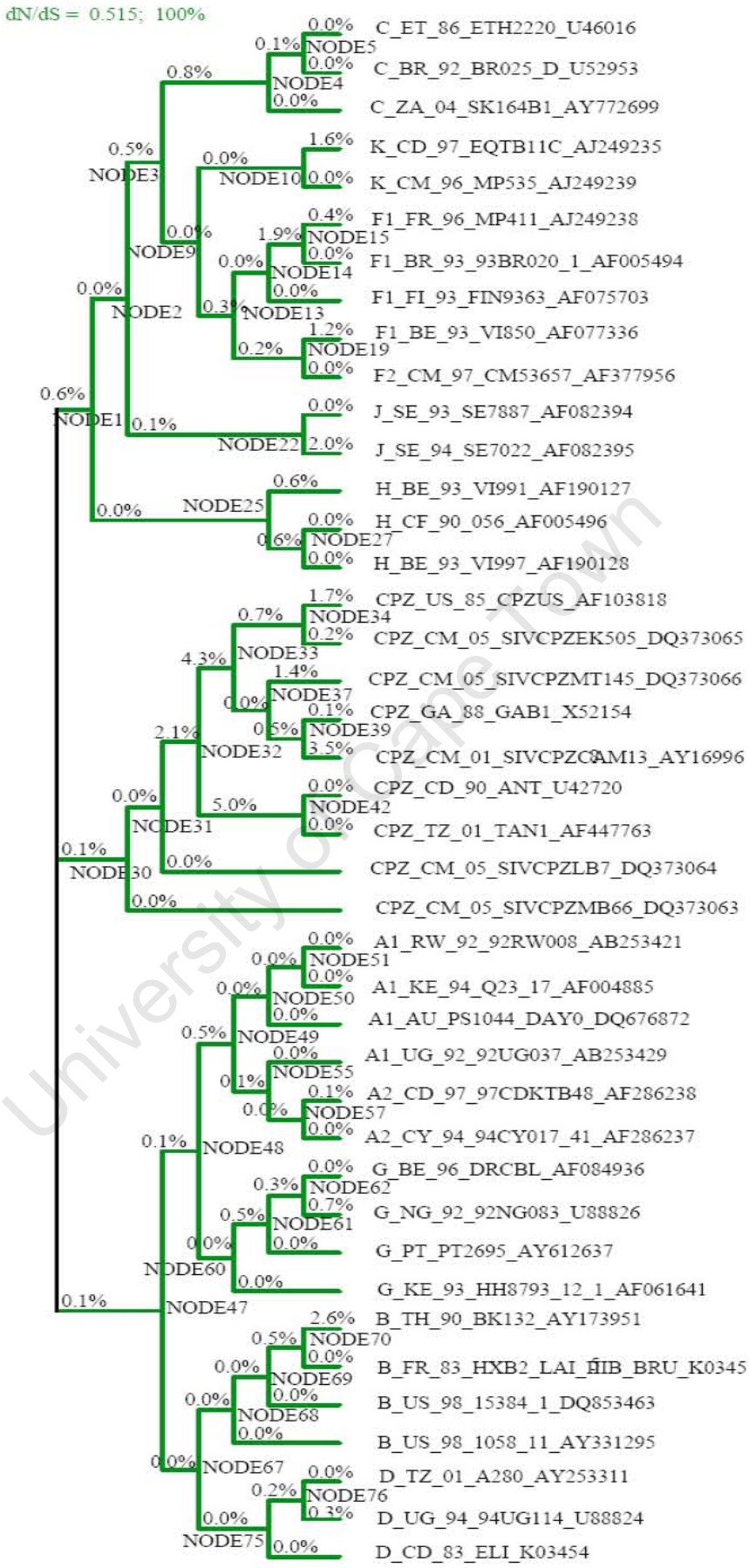
[1]Number of branches (out of a total of 79) that fall under each ω rate class are given in brackets. [2]The ω values are means over results from all the models. [3]Prob (ω>1) is the fraction of models that show support for ω>1



(a)



(b)



(c)

Figure 6.4.3: (a) Branch-by-branch selection pressure for regions predicted to be targeted by HLA-A0201. (b)Branch-by-branch selection pressure for regions predicted to be targeted by HLA-A6801. (c) Branch-by-branch selection pressure for regions predicted to be targeted by HLA-B2705. ω classes for each branch are shown in the colours given in the legends. The percentage of models that support ω>1 are written above each branch.

6.5 Discussion

The Predep program provides binding predictions for only a limited number of HLA molecules with solved crystal structures as well as preferred binding anchor residue motifs that were predicted from HLA-peptide structural conformations. Amongst these, only A0201, A6801 and B2705, were found to bind strongly to some regions of the consensus SIVcpz genome. Although higher ω was observed along the branch of the tree leading to the HIV sequences for HLAs A0201 and A6801 target sites using the BranchApriori analysis, strong selection pressure was still observed in the rest of the phylogeny. Selection pressure associated with A6801 was significantly higher in HIV-1 compared to the rest of the phylogeny (HIV ω = 3.5 versus 1.6, p value = 0.04). The difference observed for HLA A0201 was weaker (HIV ω = 2.5 versus 1.2, p value = 0.08). There was no significant evidence of differential selection pressure acting on the HIV-1 branch for the regions recognized by HLA B2705 with both ω values close to 1.

We ran the GABranch genetic algorithm to determine whether there is high selection pressure along the HIV-1 branch when ω is allowed to vary along each branch in the tree. This analysis does not assume that certain branches in the tree are under the same selection pressure as is the case in the *a priori* analysis, hence GABranch gives a more unbiased analysis of what is actually happening at each branch of the tree. The observations made in the branch-by-branch selection analysis were consistent with those obtained from the branch *a priori* analysis. A high mean ω was observed along the branch leading to the HIV-1 sequences predicted to bind to HLA-A6801 and in

those for A0201 with strong support for ω>1 from 99.6% of the tested models in both datasets. No evidence for strong selection pressure was observed at the HIV-1 branch or across the entire phylogeny in the case of HLA B2705 dataset.

It is interesting, however, that in both the *a priori* and GABranch analyses, there was no evidence for selection pressure for the HLA-B2705 alignment along the HIV-1 branch, nor any other branch. None of the tested models, including the best fitting one, showed any support for ω>1in all the branches of the phylogeny for the HLA B2705 dataset. The B27 alleles is associated with delayed progression to AIDS in HIV-1 infected individuals (McNeil *et al.,*1996). Delayed progression to AIDS has been found to be associated with persistent strong positive selection pressure at specific sites (Ross *et al.,*2002), as a result of reduced viral replication indicating that the positively selected sites are important for the fitness of the virus. One possible explanation to the observed B2705 result in this study is that the allele could have caused positive selection in only a few sites and this however became negligible since the given ω values are an average over all tested sites of the sequence.

Positive selection pressure was observed for the HIV-1 branch in the *a priori* analysis of the HLA-A0201 dataset and the GABranch analysis with very high support for ω>1. The HLA-A0201 allele is the most frequent in the Caucasian populations and many studies have been carried out to determine its effect on HIV disease progression (Marsh *et al.*,2000). Even though the allele recognizes immunodominant peptide regions of the HIV-1 sequence, it has previously been shown that it fails to exert strong selection pressure on some virus peptides during chronic infection (Brander *et al.,*1998). Some studies have also shown that the outcome of an immune response does not only depend on the HLA molecule but also on the specific peptide sequences that are targeted (Borghans *et al.,*2007; Maurer *et al.,*2008; Miura *et al.,*2008b; Nelson *et al.,*1997; Pereyra *et al.,*2008). In this study, all of the epitopes in the ancestral SIV genome that were mostly strongly predicted to have been bound by host HLA alleles, following transmission to human, were analysed. The results observed here suggest that immune escape mutations that occurred for HLA A0201 mediated CTL responses following the initial exposure of the virus to the human host had little

fitness cost to the virus. Rather they possibly became fixed, thus enabling the virus to adapt to the human immune response.

HLA A6801 appeared to exert strong selection pressure on the HIV-1 ancestral branch compared to the rest of the tree. High support (99.6% of the tested models) for $\omega > 1$ was observed at the ancestral HIV branch. This allele has anchor residue motif restrictions that are shared within the HLA A3 supertype, the second most frequent supertype in the human population (Sette *et al.,*1999). The HLA A6801 allele itself has been found to target the Tat protein which is expressed in the early stages of the HIV-1 lifecycle and CTL responses to this protein cause a significant reduction in disease progression rate (van Baalen *et al.,*1997). Escape mutations from the CTL immune response have also been identified within Tat at the population level causing reduced viral loads (Allen *et al.,*2000; Cao *et al.,*2003). The virus possibly adapted well to the A6801 responses early after the cross-species transmission event at sites that do not affect the replication of the virus and the recently observed association with a reduction in viral load indicates that there were also functionally important sites that contained A6801 epitopes that possibly failed to adapt to the immune response.

The results from this study suggest that HIV-1 adapted to CTL responses directed by HLAs A6801 and A0201, which are amongst the most common HLA genotypes in humans. It is therefore likely that the virus was frequently exposed to selection pressure exerted by common immune responses during initial exposure to the human host following transmission of the virus from chimpanzees. There was no evidence for strong selection pressure exerted by the HLA B2705 which is a generally rare genotype in the human population with extremely low frequencies in the African populations (Solberg *et al.,*2008; Trachtenberg *et al.,*2003). It is important to note that, not just the frequency of an allele could have had an effect but the functional importance of the targeted sites can also determine success or failure of fixation of an immune escape mutation that can allow the virus to adapt to its new host. This is the first study that analyses HLA-associated selection pressure following the transmission from chimpanzee to human across all potential target sites of the HIV-1 genome associated with alleles whose crystal structures have been solved. There are currently very few HLA alleles with solved structures. Crystal structures and peptide-binding

preferences of all human leukocyte antigens involved in CTL responses against HIV-1 need to be determined in order to expand analyses such as the one done here. The full analysis of the extent of HIV-1 adaptation to the human host following zoonosis will aid in better understanding of how pathogens adapt to the human host and can indirectly provide insight into designing effective therapeutics against pandemic diseases.

Chapter 7

Conclusion

Four related studies that contribute to the understanding of HIV-1 evolution and pathogenesis in humans are presented in this thesis. Each involved challenges and limitations, which are discussed in this chapter.

7.1 Quantification of functional sites that are under purifying selection pressure and functional analysis of conserved novel synonymous sites of the HIV-1 nucleotide sequence

In chapter 3, twenty-three motifs were found to be highly conserved showing evidence of purifying selection pressure acting on synonymous sites of the HIV-1 nucleotide sequence. Exhaustive searches of the literature were made in order to identify functions of all the regions that were highly conserved. The majority of the identified conserved sites were within functional regions that are well documented in the literature; however, the total number of sites that function at the nucleotide level is unknown. Even though many functional motifs are known in the HIV-1 nucleotide sequence, there is no database that collects and annotates all of them to provide a single consolidated resource. It was therefore not possible to quantify the proportion of known and novel functional sites that are under purifying selection pressure across the coding regions of the HIV-1 genome using the method applied here.

To date, preliminary functional analysis has been carried out on one of the four novel regions (i.e. one in the *env* gene and three along the *vpu* gene). The twelve-nucleotide region in the *env* gene was sent for functional analysis in a collaborating laboratory. In the preliminary results, synonymous mutations at these conserved sites appeared to slightly lower the replication rate of the mutant virus in comparison to the wildtype. There are plans to carry out replicated experiments to verify the accuracy of these preliminary results and determine whether there are significant differences between the mutants and the wildtype strain.

7.2 More studies are required to determine HLA alleles and their binding preferences from populations of the developing countries

As observed in Chapter 4 of this thesis, the HLA and anchor residue motif data associated with the Southern African population is limited in the database. Yet that from a subtype B infected cohort from North America is well characterized. HIV-1 subtype B is prevalent in the world's most developed countries, North America, Australia and parts of Europe (Peeters *et al.,*2000). Yet HIV prevalence is highest in Sub-Saharan Africa, a region mostly infected by subtype C. HIV-1 subtype C is responsible for almost half of the world's HIV-1 infections and deaths. Even though this analysis was based on two small cohorts, the differences between the two datasets does suggest that characterisation of HLA alleles involved in CTL immune responses against HIV-1 subtype C has not been completed comprehensively to date. Adequate HLA binding motif data for at least the most frequent immune responses against subtype C sequences is required in the databases. This will be useful for studies directed towards understanding immune responses in these populations and hence contribute towards the design of vaccines that are effective in the most affected populations. The curbing of the AIDS pandemic can be achieved much faster by focussing on populations where there is the highest prevalence of HIV-1 infections like the Southern African region. It is also possible that many studies are being carried out on HIV-1 subtype C sequences in this region but the results are not being deposited into public databases. There may, therefore, also be a need for researchers in this region to publish their research work and deposit more data into the publicly accessible databases.

7.3 Limitations of the HLA allele frequency data

As mentioned in section 2.8.2, there are no large publicly available datasets that provide both HIV-1 autologous sequence and HLA data of the same individual. In addition, the HLA allele frequencies in different populations are estimated from very small sample sizes. The ethnic groups of the individuals from whom the HIV-1 sequences were isolated are not given but instead the country, yet HLA allele frequency tends to vary by ethnicity (Marsh *et al.*,2000). In this particular project, the

allele frequency used was that estimated for a country as provided in the HLA allele frequency database, dbMHC (Meyer *et al.*,2007) in correspondence to the countries from which HIV-1 sequences available in the Los Alamos sequence database (Leitner *et al.,*2005) were isolated. Therefore the frequencies used in this study are estimates for a particular country and may not provide a good estimate of the HLA frequencies of the individuals from whom the viruses were isolated since different ethnic groups cohabit in a single country. Although the project managed to identify alleles with strong effects on the distribution of anchor residue motifs across HIV-1 sequences from different populations, the allele frequencies may not be accurate and this could have affected the classification of the HLA alleles as either common or rare in a population. This still however does not affect the presence or absence of an anchor residue motif in a sequence. However, the effect of the frequency of a HLA allele could have been clearer if the ethnic groups of the individuals from whom the HIV-1 sequences were obtained were identified in the sequence database.

7.4 There are currently very few HLA alleles with solved crystal structures

In chapter 6, regions that are bound by HLA alleles were identified using a method based on binding energy between the HLA molecule and the protein sequence. In this way, all the regions potentially subjected to selection pressure exerted by a specific HLA allele and the CTL immune response were analysed. This approach was not affected by poor characterisation of anchor residue motifs. However the PREDEP method (Altuvia *et al.,*2004) used is only applicable for a limited number of HLA alleles with solved crystal structures. There are currently few HLA alleles of class 1 with solved crystal structures. The method used here incorporated all the currently available HLA crystal structures. However, these structures and their features are embedded into the program code making it difficult to incorporate new HLA crystal structures as these become available. Programs using a similar approach as PREDEP but with flexible codes that allow the user to enter additional solved crystal structures for analysis would be useful when more crystal structures become available in future.

7.5 The novel achievements of this research work

Despite the challenges and limitations highlighted above, each of the projects presented in this thesis provide a novel finding. In the first a consolidated list of all synonymous sites of the HIV-1 coding regions which are under purifying selection pressure were identified. These regions are conserved across all the 11 subtypes of HIV-1 group M infecting humans. It is therefore evident that purifying selection pressure acting upon these sites is very strong since it causes conservation across all the subtypes. This shows that even though the virus has a high mutation rate, some regions are too critical for its viability to even accommodate many synonymous changes. The results of this work are also important for studies that analyse positive selection in coding regions of HIV-1 sequences by helping them identify and possibly avoid false positives in their analyses. When positive selection is inferred from a comparison of synonymous and non-synonymous substitution rates, purifying selection pressure acting at the synonymous sites can lower the synonymous substitution rate and cause the dN/dS ratio to be greater than 1, implying positive selection, even if there is no positive selection acting on the codon. The results of this research project were published and made available as a resource for other researchers.

In the second case, even though HIV-1 subtype C is the most prevalent world-wide, most scientific studies have been focused on subtype B. It is known that there is an average of 25% inter-subtype sequence difference (Buonaguro *et al.,*2007) but the accuracy of immune response data predicted using only subtype B in studying immune responses observed against other subtypes has not been extensively evaluated. This project managed to show that the HLA anchor residue motif data in the databases that is specific to HIV-1 is biased towards subtype B. In addition, the anchor residue motifs do not always accommodate the sequence variation that exists between different subtype sequences as observed by the fact that almost 50% of the subtype C peptides that were targeted by the CTL response did not contain the required HLA anchor residue motifs.

The third study was related to the effect of common HLA alleles in driving immune escape mutations in variable sites of HIV-1 sequences. The hypothesis here was that

regions of the virus that are not under strong functional constraint may have permanently lost anchor residue motifs required for presentation by common human HLA alleles. Many reports have been published on the effect of common HLA alleles in causing HIV-1 to adapt to the immune response. Here, the relationship between HLA allele frequency and the number of their anchor residue motifs across HIV-1 proteins in major infected global populations was evaluated. Some HLA alleles that were less frequent in a population were also found to be associated with fewer motifs in variable regions and contributed to the evolution of conserved regions of the HIV-1 sequence. Therefore, not only common HLA alleles but some rare ones involved in directing the less frequent immune responses are involved in shaping the evolution of conserved regions of the HIV-1 sequence.

Adaptation of the HIV-1 sequence to HLA alleles within infected populations through immune escape mutations and positive selection has been widely evaluated. However, sequence changes that occurred to enable HIV-1 to adapt to the recipient human host immune responses following the zoonotic transmission from chimpanzee has not been considered. The fourth study presented here set out to find evidence for the adaptation of HIV-1 to common human HLA alleles with solved crystal structures following the zoonosis event. Despite the limited number of HLA alleles with known crystal structures, the extent of selection pressure on the targeted regions of the alleles with solved structures was successfully evaluated. The results showed evidence that the virus did adapt to CTL immune responses directed by specific HLA alleles after initial transmission to the human host. In this study in particular, adaptation was observed towards the HLA alleles that are common in the human populations.

Reference List

Akaike H. (1987). **Factor Analysis and AIC**. *Psychometrika* 54(3):317-332

Alexander L, Weiskopf E, Greenough TC, Gaddis NC, Auerbach MR, Malim MH, O'Brien SJ, Walker BD, Sullivan JL and Desrosiers RC. (2000). **Unusual polymorphisms in human immunodeficiency virus type 1 associated with nonprogressive infection**. *J.Virol.* 74(9): 4361-4376

Allen TM, O'Connor DH, Jing P, Dzuris JL, Mothe BR, Vogel TU, Dunphy E, Liebl ME, Emerson C, Wilson N, Kunstman KJ, Wang X, Allison DB, Hughes AL, Desrosiers RC, Altman JD, Wolinsky SM, Sette A and Watkins DI. (2000). **Tat-specific cytotoxic T lymphocytes select for SIV escape variants during resolution of primary viraemia**. *Nature* 407(6802): 386-390

Altfeld MA, Livingston B, Reshamwala N, Nguyen PT, Addo MM, Shea A, Newman M, Fikes J, Sidney J, Wentworth P, Chesnut R, Eldridge RL, Rosenberg ES, Robbins GK, Brander C, Sax PE, Boswell S, Flynn T, Buchbinder S, Goulder PJ, Walker BD, Sette A and Kalams SA. (2001). **Identification of novel HLA-A2-restricted human immunodeficiency virus type 1-specific cytotoxic T-lymphocyte epitopes predicted by the HLA-A2 supertype peptide-binding motif**. *J Virol.* 75(3): 1301-1311

Altuvia Y and Margalit H. (2004). **A structure-based approach for prediction of MHC-binding peptides**. *Methods* 34(4): 454-459

Anisimova M, Nielsen R and Yang Z. (2003). **Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites**. *Genetics* 164(3): 1229-1236

Asang C, Hauber I and Schaal H. (2008). **Insights into the selective activation of alternatively used splice acceptors by the human immunodeficiency virus type-1 bidirectional splicing enhancer**. *Nucleic Acids Res.* 36(5): 1450-1463

Bansal A, Sabbaj S, Edwards BH, Ritter D, Perkins C, Tang J, Szinger JJ, Weiss H, Goepfert PA, Korber B, Wilson CM, Kaslow RA and Mulligan MJ. (2003). **T cell responses in HIV type 1-infected adolescent minorities share similar epitope specificities with whites despite significant differences in HLA class I alleles**. *AIDS Res.Hum.Retroviruses* 19(11): 1017-1026

Berkowitz RD, Hammarskjold ML, Helga-Maria C, Rekosh D and Goff SP. (1995). **5' regions of HIV-1 RNAs are not sufficient for encapsidation: implications for the HIV-1 packaging signal**. *Virology* 212(2): 718-723

Betts MR, Krowka JF, Kepler TB, Davidian M, Christopherson C, Kwok S, Louie L, Eron J, Sheppard H and Frelinger JA. (1999). **Human immunodeficiency virus type 1-specific cytotoxic T lymphocyte activity is inversely correlated with HIV type 1 viral load in HIV type 1-infected long-term survivors**. *AIDS Res.Hum.Retroviruses* 15(13): 1219-1228

Bibollet-Ruche F, McKinney BA, Duverger A, Wagner FH, Ansari AA and Kutsch O. (2008). **The quality of chimpanzee T-cell activation and simian immunodeficiency virus/human immunodeficiency virus susceptibility achieved via antibody-mediated T-cell receptor/CD3 stimulation is a function of the anti-CD3 antibody isotype**. *J.Virol.* 82(20): 10271-10278

Bihl F, Frahm N, Di Giammarino L, Sidney J, John M, Yusim K, Woodberry T, Sango K, Hewitt HS, Henry L, Linde CH, Chisholm JV, III, Zaman TM, Pae E, Mallal S, Walker BD, Sette A, Korber BT, Heckerman D and Brander C. (2006). **Impact of HLA-B alleles, epitope binding affinity, functional avidity, and viral coinfection on the immunodominance of virus-specific CTL responses**. *J.Immunol.* 176(7): 4094-4101

Boisgerault F, Khalil I, Tieng V, Connan F, Tabary T, Cohen JH, Choppin J, Charron D and Toubert A. (1996). **Definition of the HLA-A29 peptide ligand motif allows prediction of potential T-cell epitopes from the retinal soluble antigen, a candidate autoantigen in birdshot retinopathy**. *Proc.Natl.Acad.Sci.U.S.A* 93(8): 3466-3470

Borghans JA, Molgaard A, De Boer RJ and Kesmir C. (2007). **HLA alleles associated with slow progression to AIDS truly prefer to present HIV-1 p24**. *PLoS ONE.* 2(9): e920

Boritz E, Palmer BE and Wilson CC. (2004). **Human immunodeficiency virus type 1 (HIV-1)-specific CD4+ T cells that proliferate in vitro detected in samples from most viremic subjects and inversely associated with plasma HIV-1 levels**. *J Virol.* 78(22): 12638-12646

Brander C, Hartman KE, Trocha AK, Jones NG, Johnson RP, Korber B, Wentworth P, Buchbinder SP, Wolinsky S, Walker BD and Kalams SA. (1998). **Lack of strong immune selection pressure by the immunodominant, HLA-A*0201-restricted cytotoxic T lymphocyte response in chronic human immunodeficiency virus-1 infection**. *J.Clin.Invest* 101(11): 2559-2566

Bredell H, Martin DP, Van Harmelen J, Varsani A, Sheppard HW, Donovan R, Gray CM and Williamson C. (2007). **HIV type 1 subtype C gag and nef diversity in Southern Africa**. *AIDS Res.Hum.Retroviruses* 23(3): 477-481

Brown HE, Chen H and Engelman A. (1999). **Structure-based mutagenesis of the human immunodeficiency virus type 1 DNA attachment site: effects on integration and cDNA synthesis**. *J.Virol.* 73(11): 9011-9020

Brumme ZL, Brumme CJ, Carlson J, Streeck H, John M, Eichbaum Q, Block BL, Baker B, Kadie C, Markowitz M, Jessen H, Kelleher AD, Rosenberg E, Kaldor J, Yuki Y, Carrington M, Allen TM, Mallal S, Altfeld M, Heckerman D and Walker BD. (2008a). **Marked epitope and allele-specific differences in rates of mutation in HIV-1 Gag, Pol and Nef CTL epitopes in acute/early HIV-1 infection**. *J.Virol.*

Brumme ZL, Brumme CJ, Heckerman D, Korber BT, Daniels M, Carlson J, Kadie C, Bhattacharya T, Chui C, Szinger J, Mo T, Hogg RS, Montaner JS, Frahm N, Brander C, Walker BD and Harrigan PR. (2007). **Evidence of differential HLA class I-mediated viral evolution in functional and accessory/regulatory genes of HIV-1**. *PLoS Pathog.* 3(7): e94

Brumme ZL, Tao I, Szeto S, Brumme CJ, Carlson JM, Chan D, Kadie C, Frahm N, Brander C, Walker B, Heckerman D and Harrigan PR. (2008b). **Human leukocyte antigen-specific polymorphisms in HIV-1 Gag and their association with viral load in chronic untreated infection**. *AIDS* 22(11): 1277-1286

Buonaguro L, Tornesello ML and Buonaguro FM. (2007). **Human immunodeficiency virus type 1 subtype distribution in the worldwide epidemic: pathogenetic and therapeutic implications**. *J.Virol.* 81(19): 10209-10219

Cao J, McNevin J, Malhotra U and McElrath MJ. (2003). **Evolution of CD8+ T cell immunity and viral escape following acute HIV-1 infection**. *J.Immunol.* 171(7): 3837-3846

Cao K, Hollenbach J, Shi X, Shi W, Chopek M and Fernandez-Vina MA. (2001). **Analysis of the frequencies of HLA-A, B, and C alleles and haplotypes in the five major ethnic groups of the United States reveals high levels of diversity in these loci and contrasting distribution patterns in these populations**. *Hum.Immunol.* 62(9): 1009-1030

Carlson JM and Brumme ZL. (2008). **HIV evolution in response to HLA-restricted CTL selection pressures: a population-based perspective**. *Microbes.Infect.* 10(5): 455-461

Carrington M, Nelson GW, Martin MP, Kissner T, Vlahov D, Goedert JJ, Kaslow R, Buchbinder S, Hoots K and O'Brien SJ. (1999). **HLA and HIV-1: heterozygote advantage and B*35-Cw*04 disadvantage**. *Science* 283(5408): 1748-1752

Carrington M and O'Brien SJ. (2003). **The influence of HLA genotype on AIDS**. *Annu.Rev.Med.* 54535-551

Casado C, Garcia S, Rodriguez C, del Romero J, Bello G and Lopez-Galindez C. (2001). **Different evolutionary patterns are found within human immunodeficiency virus type 1-infected patients**. *J Gen.Virol.* 82(Pt 10): 2495-2508

Chan DC and Kim PS. (1998). **HIV entry and its inhibition**. *Cell* 93(5):681-684

Chen Y, Winchester R, Korber B, Gagliano J, Bryson Y, Hutto C, Martin N, McSherry G, Petru A, Wara D and Ammann A. (1997). **Influence of HLA alleles on the rate of progression of vertically transmitted HIV infection in children: association of several HLA-DR13 alleles with long-term survivorship and the potential association of HLA-A*2301 with rapid progression to AIDS. Long-Term Survivor Study**. *Hum.Immunol.* 55(2): 154-162

Choisy M, Woelk CH, Guegan JF and Robertson DL. (2004). **Comparative study of adaptive molecular evolution in different human immunodeficiency virus groups and subtypes**. *J.Virol.* 78(4): 1962-1970

Clever J, Sassetti C and Parslow TG. (1995). **RNA secondary structure and binding sites for gag gene products in the 5' packaging signal of human immunodeficiency virus type 1**. *J.Virol.* 69(4): 2101-2109

Cochrane A, Jones K, Beidas S, Dillon P, Skalka A and Rosen C. (1991). **Identification and characterization of intragenic sequences which repress human immunodeficiency virus structural gene expression**. *J.Virol.* 65(10): 5305-5313

Cormier EG and Dragic T. (2002). **The crown and stem of the V3 loop play distinct roles in human immunodeficiency virus type 1 envelope glycoprotein interactions with the CCR5 coreceptor**. *J.Virol.* 76(17): 8953-8957

Cormier EG, Tran DN, Yukhayeva L, Olson WC and Dragic T. (2001). **Mapping the determinants of the CCR5 amino-terminal sulfopeptide interaction with soluble human immunodeficiency virus type 1 gp120-CD4 complexes**. *J.Virol.* 75(12): 5541-5549

Crawford H, Prado JG, Leslie A, Hue S, Honeyborne I, Reddy S, van der SM, Mncube Z, Brander C, Rousseau C, Mullins JI, Kaslow R, Goepfert P, Allen S, Hunter E, Mulenga J, Kiepiela P, Walker BD and Goulder PJ. (2007). **Compensatory mutation partially restores fitness and delays reversion of escape mutation within the immunodominant HLA-B*5703-restricted Gag epitope in chronic human immunodeficiency virus type 1 infection**. *J.Virol.* 81(15): 8346-8351

da Silva J and Hughes AL. (1998). **Conservation of cytotoxic T lymphocyte (CTL) epitopes as a host strategy to constrain parasite adaptation: evidence from the nef gene of human immunodeficiency virus 1 (HIV-1)**. *Mol.Biol.Evol.* 15(10): 1259-1268

Das A, Klaver B and Berkhout B. (1998). **The 5' and 3' TAR elements of human immunodeficiency virus exert effects at several points in the virus life cycle**. *J.Virol.* 72(11): 9217-9223

De Leys R, Vanderborght B, Vanden Haesevelde M, Heyndrickx L, van Geel A, Wauters C, Bernaerts R, Saman E, Nijs P, Willems B and . (1990). **Isolation and partial characterization of an unusual human immunodeficiency retrovirus from two persons of west-central African origin**. *J.Virol.* 64(3): 1207-1216

de Oliveira T, Salemi M, Gordon M, Vandamme AM, van Rensburg EJ, Engelbrecht S, Coovadia HM and Cassol S. (2004). **Mapping sites of positive selection and amino acid diversification in the HIV genome: an alternative approach to vaccine design?** *Genetics* 167(3): 1047-1058

Delgado E, Thomson MM, Villahermosa ML, Sierra M, Ocampo A, Miralles C, Rodriguez-Perez R, Diz-Aren J, Ojea-de Castro R, Losada E, Cuevas MT, Vazquez-de Parga E, Carmona R, Perez-Alvarez L, Medrano L, Cuevas L, Taboada JA and Najera R. (2002). **Identification of a newly characterized HIV-1 BG intersubtype circulating recombinant form in Galicia, Spain, which exhibits a pseudotype-like virion structure**. *J.Acquir.Immune Defic.Syndr.* 29(5): 536-543

Delport W, Scheffler K and Seoighe C. (2009). **Models of coding sequence evolution**. *Brief.Bioinform.* 10(1): 97-109

DiBrino M, Parker KC, Shiloach J, Turner RV, Tsuchida T, Garfield M, Biddison WE and Coligan JE. (1994). **Endogenous peptides with distinct amino acid anchor residue motifs bind to HLA-A1 and HLA-B8**. *J.Immunol.* 152(2): 620-631

Dimitrov DS, Willey RL, Sato H, Chang LJ, Blumenthal R and Martin MA. (1993). **Quantitation of human immunodeficiency virus type 1 infection kinetics**. *J.Virol.* 67(4): 2182-2190

Doms RW and Trono D. (2000). **The plasma membrane as a combat zone in the HIV battlefield.** *Genes & Dev*. 14: 2677-2688

Draenert R, Allen TM, Liu Y, Wrin T, Chappey C, Verrill CL, Sirera G, Eldridge RL, Lahaie MP, Ruiz L, Clotet B, Petropoulos CJ, Walker BD and Martinez-Picado J. (2006). **Constraints on HIV-1 evolution and immunodominance revealed in monozygotic adult twins infected with the same virus**. *J Exp.Med.* 203(3): 529-539

Edwards CT, Pfafferott KJ, Goulder PJ, Phillips RE and Holmes EC. (2005). **Intrapatient escape in the A*0201-restricted epitope SLYNTVATL drives evolution of human immunodeficiency virus type 1 at the population level**. *J.Virol.* 79(14): 9363-9366

Exline CM, Feng Z and Stoltzfus CM. (2008). **Negative and positive mRNA splicing elements act competitively to regulate human immunodeficiency virus type 1 vif gene expression**. *J.Virol.* 82(8): 3921-3931

Falk K, Rotzschke O, Stevanovic S, Jung G and Rammensee HG. (1991). **Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules**. *Nature* 351(6324): 290-296

Flajnik MF and Du PL. (2004). **Evolution of innate and adaptive immunity: can we draw a line?** *Trends Immunol.* 25(12): 640-644

Flores-Villanueva PO, Yunis EJ, Delgado JC, Vittinghoff E, Buchbinder S, Leung JY, Uglialoro AM, Clavijo OP, Rosenberg ES, Kalams SA, Braun JD, Boswell SL, Walker BD and Goldfeld AE. (2001). **Control of HIV-1 viremia and protection from AIDS are associated with HLA-Bw4 homozygosity**. *Proc.Natl.Acad.Sci.U.S.A* 98(9): 5140-5145

Flower D. (2007). **Immunoinformatics *Predicting Immunogenicity In Silico***. *Humana Press Inc.*

Foster JL and Garcia JV. (2008). **HIV-1 Nef: at the crossroads**. *Retrovirology* 584

Frahm N, Baker B and Brander C. (2008). **Identification and Optimal Definition of HIV-Derived Cytotoxic T-Lymphocyte (CTL) Epitopes for the Study of CTL**

Escape, Functional Avidity and Viral Evolution. *HIV Molecular Immunology.Los Alamos National Laboratory, Theoretical Biology and Biophysics, Los Alamos, New Mexico* LA-UR 08-050963-24

Frahm N, Goulder P and Brander C. (2004). **Broad HIV-1 Specific CTL Responses Reveal Extensive HLA Class I Binding Promiscuity of HIV-Derived, Optimally Defined CTL Epitopes**. *Los Alamos HIV Sequence Databases: Review Articles* LA-UR04-24

Frahm N, Linde C and Brander C. (2007). **Identification of HIV-Derived, HLA Class I Restricted CTL Epitopes: Insights into TCR Repertoire, CTL Escape and Viral Fitness**. *HIV Molecular Immunology.Los Alamos National Laboratory, Theoretical Biology and Biophysics, Los Alamos, New Mexico* LA-UR 07-47523-28

Frahm N, Kiepiela P, Adams S, Linde CH, Hewitt HS, Sango K, Feeney ME, Addo MM, Lichterfeld M, Lahaie MP, Pae E, Wurcel AG, Roach T, St John MA, Altfeld M, Marincola FM, Moore C, Mallal S, Carrington M, Heckerman D, Allen TM, Mullins JI, Korber BT, Goulder PJ, Walker BD and Brander C. (2006). **Control of human immunodeficiency virus replication by cytotoxic T lymphocytes targeting subdominant epitopes**. *Nat.Immunol.* 7(2): 173-178

Frankel AD and Young JA. (1998). **HIV-1: fifteen proteins and an RNA**. *Annu.Rev.Biochem.* 671-25

Friedrich TC, Dodds EJ, Yant LJ, Vojnov L, Rudersdorf R, Cullen C, Evans DT, Desrosiers RC, Mothe BR, Sidney J, Sette A, Kunstman K, Wolinsky S, Piatak M, Lifson J, Hughes AL, Wilson N, O'Connor DH and Watkins DI. (2004). **Reversion of CTL escape-variant immunodeficiency viruses in vivo**. *Nat.Med.* 10(3): 275-281

Frost SD, Wrin T, Smith DM, Kosakovsky Pond SL, Liu Y, Paxinos E, Chappey C, Galovich J, Beauchaine J, Petropoulos CJ, Little SJ and Richman DD. (2005). **Neutralizing antibody responses drive the evolution of human immunodeficiency virus type 1 envelope during recent HIV infection**. *Proc.Natl.Acad.Sci.U.S.A* 102(51): 18514-18519

Gao F, Bailes E, Robertson DL, Chen Y, Rodenburg CM, Michael SF, Cummins LB, Arthur LO, Peeters M, Shaw GM, Sharp PM and Hahn BH. (1999). **Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes**. *Nature* 397(6718): 436-441

Geretti AM. (2006). **HIV-1 subtypes: epidemiology and significance for HIV management**. *Curr.Opin.Infect.Dis.* 19(1): 1-7

Gilbert PB, Novitsky V and Essex M. (2005). **Covariability of selected amino acid positions for HIV type 1 subtypes C and B**. *AIDS Res.Hum.Retroviruses* 21(12): 1016-1030

Goffin V, Demonte D, Vanhulle C, de Walque S, de Launoit Y, Burny A, Collette Y and Van Lint C. (2005). **Transcription factor binding sites in the pol gene intragenic regulatory region of HIV-1 are important for virus infectivity**. *Nucleic Acids Res.* 33(13): 4285-4310

Gog J, Afonso E, Dalton R, Leclercg I, Tiley L, Elton D, Von Kirchbach J, Naffakh N, Escriou N and Digard P. (2007). **Codon conservation in the influenza A virus genome defines RNA packaging signals**. *Nucleic Acids Research* 35(6): 1897-1907

Goldman N and Yang Z. (1994). **A codon-based model of nucleotide substitution for protein-coding DNA sequences**. *Mol.Biol.Evol.* 11(5): 725-736

Gottlinger HG. (2001). **The HIV-1 assembly machine**. *AIDS* 15 Suppl 5S13-S20

Gottlinger HG, Dorfman T, Sodroski JG and Haseltine WA. (1991). **Effect of mutations affecting the p6 gag protein on human immunodeficiency virus particle release**. *Proc.Natl.Acad.Sci.U.S.A* 88(8): 3195-3199

Goulder PJ and Watkins DI. (2004). **HIV and SIV CTL escape: implications for vaccine design**. *Nat.Rev.Immunol.* 4(8): 630-640

Greenberg M, DeTulleo L, Rapoport I, Skowronski J and Kirchhausen T. (1998). **A dileucine motif in HIV-1 Nef is essential for sorting into clathrin-coated pits and for downregulation of CD4**. *Curr.Biol.* 8(22): 1239-1242

Greene WC and Peterlin BM. (2002). **Charting HIV's remarkable voyage through the cell: Basic science as a passport to future therapy**. *Nat.Med.* 8(7): 673-680

Grigorov B, Arcanger F, Roingeard P, Darlix JL and Muriaux D. (2006). **Assembly of infectious HIV-1 in human epithelial and T-lymphoblastic cell lines**. *J.Mol.Biol.* 359(4): 848-862

Hadzopoulou-Cladaras M, Felber B, Cladaras C, Athanassopoulos A, Tse A and Pavlakis G. (1989). **The rev (trs/art) protein of human immunodeficiency virus type 1 affects viral mRNA and protein expression via a cis-acting sequence in the env region**. *J.Virol.* 63(3): 1265-1274

Hahn BH, Shaw GM, De Cock KM and Sharp PM. (2000). **AIDS as a zoonosis: scientific and public health implications**. *Science* 287(5453): 607-614

Hammond M, Middleton D and Anley D. (2007). **Zulu from Natal Province, South Africa: Anthropology/human genetic diversity population reports**. *IHWG Press* 1590-591

Hanada K, Suzuki Y and Gojobori T. (2004). **A large variation in the rates of synonymous substitution for RNA viruses and its relationship to a diversity of viral infection and transmission modes**. *Mol.Biol.Evol.* 21(6): 1074-1080

Helseth E, Olshevsky U, Furman C and Sodroski J. (1991). **Human immunodeficiency virus type 1 gp120 envelope glycoprotein regions important for association with the gp41 transmembrane glycoprotein**. *J.Virol.* 65(4): 2119-2123

Hollenbach JA, Thomson G, Cao K, Fernandez-Vina M, Erlich HA, Bugawan TL, Winkler C, Winter M and Klitz W. (2001). **HLA diversity, differentiation, and haplotype evolution in Mesoamerican Natives**. *Hum.Immunol.* 62(4): 378-390

Honeyborne I, Rathod A, Buchli R, Ramduth D, Moodley E, Rathnavalu P, Chetty S, Day C, Brander C, Hildebrand W, Walker BD, Kiepiela P and Goulder PJ. (2006). **Motif inference reveals optimal CTL epitopes presented by HLA class I alleles highly prevalent in southern Africa**. *J.Immunol.* 176(8): 4699-4705

Huet T, Cheynier R, Meyerhans A, Roelants G and Wain-Hobson S. (1990). **Genetic organization of a chimpanzee lentivirus related to HIV-1**. *Nature* 345(6273): 356-359

Hung LW, Holbrook EL and Holbrook SR. (2000). **The crystal structure of the Rev binding element of HIV-1 reveals novel base pairing and conformational variability**. *Proc.Natl.Acad.Sci.U.S.A* 97(10): 5107-5112

Hurst LD and Pal C. (2001). **Evidence for purifying selection acting on silent sites in *BRCA1***. *TRENDS in Genetics* 17(2): 62-65

Huthoff H, Das AT, Vink M, Klaver B, Zorgdrager F, Cornelissen M and Berkhout B. (2004). **A human immunodeficiency virus type 1-infected individual with low viral load harbors a virus variant that exhibits an in vitro RNA dimerization defect**. *J.Virol.* 78(9): 4907-4913

Jacquenet S, Mereau A, Bilodeau PS, Damier L, Stoltzfus CM and Branlant C. (2001a). **A second exon splicing silencer within human immunodeficiency virus type 1 tat exon 2 represses splicing of Tat mRNA and binds protein hnRNP H**. *J.Biol.Chem.* 276(44): 40464-40475

Jacquenet S, Ropers D, Bilodeau PS, Damier L, Mougin A, Stoltzfus CM and Branlant C. (2001b). **Conserved stem-loop structures in the HIV-1 RNA region containing the A3 3' splice site and its cis-regulatory element: possible involvement in RNA splicing**. *Nucleic Acids Res.* 29(2): 464-478

Janeway C, Travers P, Hunt S and Walport M. (1997). **Immunobiology : the immune system in health and disease**. *London, Current Biology Edinburgh, Churchill Livingstone*

Jojic N, Reyes-Gomez M, Heckerman D, Kadie C and Schueler-Furman O. (2006). **Learning MHC I--peptide binding**. *Bioinformatics* 22(14): e227-e235

Jonassen TO, Stene-Johansen K, Berg ES, Hungnes O, Lindboe CF, Froland SS and Grinde B. (1997). **Sequence analysis of HIV-1 group O from Norwegian patients infected in the 1960s**. *Virology* 231(1): 43-47

Kammler S, Otte M, Hauber I, Kjems J, Hauber J and Schaal H. (2006). **The strength of the HIV-1 3' splice sites affects Rev function**. *Retrovirology* 389

Kawashima Y, Pfafferott K, Frater J, Matthews P, Payne R, Addo M, Gatanaga H, Fujiwara M, Hachiya A, Koizumi H, Kuse N, Oka S, Duda A, Prendergast A, Crawford H, Leslie A, Brumme Z, Brumme C, Allen T, Brander C, Kaslow R, Tang J, Hunter E, Allen S, Mulenga J, Branch S, Roach T, John M, Mallal S, Ogwu A, Shapiro R, Prado JG, Fidler S, Weber J, Pybus OG, Klenerman P, Ndung'u T, Phillips R, Heckerman D, Harrigan PR, Walker BD, Takiguchi M and Goulder P. (2009). **Adaptation of HIV-1 to human leukocyte antigen class I**. *Nature* 458(7238): 641-645

Keele BF, Jones JH, Terio KA, Estes JD, Rudicell RS, Wilson ML, Li Y, Learn GH, Mark B, Schumacher-Stankey J, Wroblewski W, Mosser A, Raphael J, Kamenya S, Lonsdorf EV, Travis DA, Mlengenya T, Kinsel MJ, Else JG, Silvestri G, Goodall J, Sharp PM, Shaw GM, Pusey AE and Hahn BH. (2009). I**ncreased mortality and AIDS-like immunopathology in wild chimpanzees infected with SIVcpz**. *Nature* 460: 515-519

Keele BF, Van Heuverswyn F, Li Y, Bailes E, Takehisa J, Santiago ML, Bibollet-Ruche F, Chen Y, Wain LV, Liegeois F, Loul S, Ngole EM, Bienvenue Y, Delaporte E, Brookfield JF, Sharp PM, Shaw GM, Peeters M and Hahn BH. (2006). **Chimpanzee reservoirs of pandemic and nonpandemic HIV-1**. *Science* 313(5786): 523-526

Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV and Wingender E. (2003). **MATCH: A tool for searching transcription factor binding sites in DNA sequences**. *Nucleic Acids Res.* 31(13): 3576-3579

Kelleher AD, Long C, Holmes EC, Allen RL, Wilson J, Conlon C, Workman C, Shaunak S, Olson K, Goulder P, Brander C, Ogg G, Sullivan JS, Dyer W, Jones I, McMichael AJ, Rowland-Jones S and Phillips RE. (2001). **Clustered mutations in HIV-1 gag are consistently required for escape from HLA-B27-restricted cytotoxic T lymphocyte responses**. *J.Exp.Med.* 193(3): 375-386

Kiepiela P, Leslie AJ, Honeyborne I, Ramduth D, Thobakgale C, Chetty S, Rathnavalu P, Moore C, Pfafferott KJ, Hilton L, Zimbwa P, Moore S, Allen T, Brander C, Addo MM, Altfeld M, James I, Mallal S, Bunce M, Barber LD, Szinger J, Day C, Klenerman P, Mullins J, Korber B, Coovadia HM, Walker BD and Goulder

PJ. (2004). **Dominant influence of HLA-B in mediating the potential co-evolution of HIV and HLA**. *Nature* 432(7018): 769-775

Kiepiela P, Ngumbela K, Thobakgale C, Ramduth D, Honeyborne I, Moodley E, Reddy S, de Pierres C, Mncube Z, Mkhwanazi N, Bishop K, van der SM, Nair K, Khan N, Crawford H, Payne R, Leslie A, Prado J, Prendergast A, Frater J, McCarthy N, Brander C, Learn GH, Nickle D, Rousseau C, Coovadia H, Mullins JI, Heckerman D, Walker BD and Goulder P. (2007). **CD8+ T-cell responses to different HIV proteins have discordant associations with viral load**. *Nat.Med.* 13(1): 46-53

Kimura T, Hashimoto I, Yamamoto A, Nishikawa M and Fujisawa JI. (2000). **Rev-dependent association of the intron-containing HIV-1 gag mRNA with the nuclear actin bundles and the inhibition of its nucleocytoplasmic transport by latrunculin-B.** *GENES CELLS* **5:289-307**

Korber B, Muldoon M, Theiler J, Gao F, Gupta R, Lapedes A, Hahn BH, Wolinsky S and Bhattacharya T. (2000). **Timing the ancestor of the HIV-1 pandemic strains**. *Science* 288(5472): 1789-1796

Korber B, Theiler J and Wolinsky S. (1998). **Limitations of a molecular clock applied to considerations of the origin of HIV-1**. *Science* 280(5371): 1868-1871

Korber B, Brander C, Haynes B, Koup R, Moore J, Walker B and Watkins D. (2006). **HIV Molecular Immunology**. *Los Alamos National Laboratory, Theoretical Biology and Biophysics, Los Alamos, New Mexico* LA-UR07-4752

Kosakovsky Pond S and Muse SV. (2005a). **Site-to-site variation of synonymous substitution rates**. *Mol.Biol.Evol.* 22(12): 2375-2385

Kosakovsky Pond SL and Frost SD. (2005b). **A genetic algorithm approach to detecting lineage-specific variation in selection pressure**. *Mol.Biol.Evol.* 22(3): 478-485

Kosakovsky Pond SL and Frost SD. (2005c). **Datamonkey: rapid detection of selective pressure on individual sites of codon alignments**. *Bioinformatics* 21(10): 2531-2533

Kosakovsky Pond SL, Frost SD and Muse SV. (2005d). **HyPhy: hypothesis testing using phylogenies**. *Bioinformatics* 21(5): 676-679

Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH and Frost S. (2006). **Automated Phylogenetic Detection of Recombination Using a Genetic Algorithm**. *Mol.Biol.Evol.* 23(10): 1891-1901

Krummheuer J, Lenz C, Kammler S, Scheid A and Schaal H. (2001). **Influence of the small leader exons 2 and 3 on human immunodeficiency virus type 1 gene expression**. *Virology* 286(2): 276-289

Lama J, Mangasarian A and Trono D. (1999). **Cell-surface expression of CD4 reduces HIV-1 infectivity by blocking Env incorporation in a Nef- and Vpu-inhibitable manner**. *Curr.Biol.* 9(12): 622-631

Lanzavecchia A. (1998). **Immunology. Licence to kill**. *Nature* 393(6684): 413-414

Lapham CK, Ouyang J, Chandrasekhar B, Nuyen NY, Dimitrov DS and Golding H. (1996). **Evidence for cell-surface association between Fusin and the CD4-gp120 complex in human cell lines**. *Science* 274:62-605

Lassen KG, Ramyar KX, Bailey JR, Zhou Y and Siliciano RF. (2006). **Nuclear Retention of Multiply Spliced HIV-1 RNA in Resting CD4$^+$ T Cells**. *PLoS Pathog* 2(7): e68

Lata S, Bhasin M and Raghava F. (2007). **Application of Machine Learning Techniques in Predicting MHC binders**. *The Humana Press Inc.*(14):

Learn GH, Jr., Korber BT, Foley B, Hahn BH, Wolinsky SM and Mullins JI. (1996). **Maintaining the integrity of human immunodeficiency virus sequence databases**. *J Virol.* 70(8): 5720-5730

Leitner T, Korber B, Daniels M, Calef C and Foley B. (2005). **HIV-1 Subtype and Circulating Recombinant Form (CRF) Reference Sequences**. *Los Alamos National Laboratory Reviews* 200541-48

Leitner T, Foley B, Hahn B, Marx P, McCutchan F, Mellors J, Wolinsky S and Korber B. (2005). **HIV Sequence Compendium**. *Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, NM, LA-UR*06-0680

Lemey P, Kosakovsky Pond SL, Drummond AJ, Pybus OG, Shapiro B, Barroso H, Taveira N and Rambaut A. (2007). **Synonymous substitution rates predict HIV disease progression as a result of underlying replication dynamics**. *PLoS Comput.Biol.* 3(2): e29

Letvin NL. (1998). **Progress in the development of an HIV-1 vaccine**. *Science* 280(5371): 1875-1880

Levy J. (2007). **HIV and the Pathogenesis of AIDS**. *American Society for Microbiology Press* 3rd13-20

Li WH, Tanimura M and Sharp PM. (1988). **Rates and dates of divergence between AIDS virus nucleotide sequences**. *Mol.Biol.Evol.* 5(4): 313-330

Logean A and Rognan D. (2002). **Recovery of known T-cell epitopes by computational scanning of a viral genome**. *J Comput.Aided Mol.Des* 16(4): 229-243

Luo GX, Sharmeen L and Taylor J. (1990). **Specificities involved in the initiation of retroviral plus-strand DNA**. *J.Virol.* 64(2): 592-597

MacDonald KS, Fowke KR, Kimani J, Dunand VA, Nagelkerke NJ, Ball TB, Oyugi J, Njagi E, Gaur LK, Brunham RC, Wade J, Luscher MA, Krausa P, Rowland-Jones S, Ngugi E, Bwayo JJ and Plummer FA. (2000). **Influence of HLA supertypes on susceptibility and resistance to human immunodeficiency virus type 1 infection**. *J.Infect.Dis.* 181(5): 1581-1589

Madden DR, Gorga JC, Strominger JL and Wiley DC. (1992). **The three-dimensional structure of HLA-B27 at 2.1 A resolution suggests a general mechanism for tight peptide binding to MHC**. *Cell* 70(6): 1035-1048

Madsen JM and Stoltzfus CM. (2006). **A suboptimal 5' splice site downstream of HIV-1 splice site A1 is required for unspliced viral mRNA accumulation and efficient virus replication**. *Retrovirology* 310

Marsh S, Parham P and Barber L. (2000). **The HLA Facts Book**. *Academic Press* 1

Masemola A, Mashishi T, Khoury G, Mohube P, Mokgotho P, Vardas E, Colvin M, Zijenah L, Katzenstein D, Musonda R, Allen S, Kumwenda N, Taha T, Gray G, McIntyre J, Karim S, Sheppard H and Gray C. (2004a). **Hierarchical targeting of subtype C human immunodeficiency virus type 1 proteins by CD8+ T cells: correlation with viral load**. *J.Virol.* 78(7): 3233-3243

Masemola A, Mashishi T, Khoury G, Bredell H, Paximadis M, Mathebula T, Barkhan D, Puren A, Vardas E, Colvin M, Zijenah L, Katzenstein D, Musonda R, Allen S, Kumwenda N, Taha T, Gray G, McIntyre J, Karim S, Sheppard H and Gray C. (2004b). **Novel and promiscuous CTL epitopes in conserved regions of Gag targeted by individuals with early subtype C HIV type 1 infection from southern Africa**. *J.Immunol.* 173(7): 4607-4617

Mashishi T, Loubser S, Hide W, Hunt G, Morris L, Ramjee G, Abdool-Karim S, Williamson C and Gray CM. (2001). **Conserved domains of subtype C nef from South African HIV type 1-infected individuals include cytotoxic T lymphocyte epitope-rich regions**. *AIDS Res.Hum.Retroviruses* 17(17): 1681-1687

Masuda T, Kuroda MJ and Harada S. (1998). **Specific and independent recognition of U3 and U5 att sites by human immunodeficiency virus type 1 integrase in vivo**. *J.Virol.* 72(10): 8396-8402

Masur H, Michelis MA, Greene JB, Onorato I, Stouwe RA, Holzman RS, Wormser G, Brettman L, Lange M, Murray HW and Cunningham-Rundles S. (1981). **An outbreak of community-acquired Pneumocystis carinii pneumonia: initial manifestation of cellular immune dysfunction**. *N.Engl.J Med.* 305(24): 1431-1438

Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Munch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S and Wingender E. (2003).

TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.* 31(1): 374-378

Maurer K, Harrer EG, Goldwich A, Eismann K, Bergmann S, Schmitt-Haendle M, Spriewald B, Mueller SM and Harrer T. (2008). **Role of cytotoxic T-lymphocyte-mediated immune selection in a dominant human leukocyte antigen-B8-restricted cytotoxic T-lymphocyte epitope in Nef**. *J.Acquir.Immune Defic.Syndr.* 48(2): 133-141

Mayrose I, Graur D, Ben-Tal N and Pupko T. (2004). **Comparison of Site-Specific Rate-Inference Methods for Protein Sequences: Emperical Bayesian Methods Are Superior**. *Molecular Biology and Evolution* 21(9): 1781-1791

McMichael A and Klenerman P. (2002). **HIV/AIDS. HLA leaves its footprints on HIV**. *Science* 296(5572): 1410-1411

McMichael AJ and Rowland-Jones SL. (2001). **Cellular immune responses to HIV**. *Nature* 410(6831): 980-987

McNeil AJ, Yap PL, Gore SM, Brettle RP, McColl M, Wyld R, Davidson S, Weightman R, Richardson AM and Robertson JR. (1996). **Association of HLA types A1-B8-DR3 and B27 with rapid and slow progression of HIV disease**. *QJM.* 89(3): 177-185

Meyer D, Singe R, Mack S, Lancaster A, Nelson M, Erlich H, Fernandez-Vina M and Thomson G. (2007). **Single Locus Polymorphism of Classical HLA Genes.** *IHWG Press* 1

Michael NL, Mo T, Merzouki A, O'Shaughnessy M, Oster C, Burke DS, Redfield RR, Birx DL and Cassol SA. (1995). **Human immunodeficiency virus type 1 cellular RNA load and splicing patterns predict disease progression in a longitudinally studied cohort**. *J.Virol.* 69(3): 1868-1877

Miles LR, Agresta BE, Khan MB, Tang S, Levin JG and Powell MD. (2005). **Effect of polypurine tract (PPT) mutations on human immunodeficiency virus type 1 replication: a virus with a completely randomized PPT retains low infectivity**. *J.Virol.* 79(11): 6859-6867

Miura T, Brockman MA, Brumme CJ, Brumme ZL, Carlson JM, Pereyra F, Trocha A, Addo MM, Block BL, Rothchild AC, Baker BM, Flynn T, Schneidewind A, Li B, Wang YE, Heckerman D, Allen TM and Walker BD. (2008a). **Genetic Characterization of Human Immunodeficiency Virus type 1 in Elite Controllers: Lack of gross genetic defects or common amino acid changes**. *J.Virol.*

Miura T, Brockman MA, Schneidewind A, Lobritz M, Pereyra F, Rathod A, Block BL, Brumme ZL, Brumme CJ, Baker B, Rothchild AC, Li B, Trocha A, Cutrell E, Frahm N, Brander C, Toth I, Arts EJ, Allen TM and Walker BD. (2008b). **HLA-B57/B*5801 HIV-1 ELITE CONTROLLERS SELECT FOR RARE GAG VARIANTS ASSOCIATED WITH REDUCED VIRAL REPLICATION CAPACITY AND STRONG CTL RECOGNITION**. *J.Virol.*

Moore CB, John M, James IR, Christiansen FT, Witt CS and Mallal SA. (2002). **Evidence of HIV-1 adaptation to HLA-restricted immune responses at a population level**. *Science* 296(5572): 1439-1443

Muller V and De Boer RJ. (2006). **The integration hypothesis: an evolutionary pathway to benign SIV infection**. *PLoS Pathog.* 2(3): e15

Murakami T and Freed EO. (2000). **The long cytoplasmic tail of gp41 is required in a cell type-dependent manner for HIV-1 envelope glycoprotein incorporation into virions**. *Proc.Natl.Acad.Sci.U.S.A* 97(1): 343-348

Muse SV and Gaut BS. (1994). **A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome**. *Mol.Biol.Evol.* 11(5): 715-724

Musey L, Hughes J, Schacker T, Shea T, Corey L and McElrath MJ. (1997). **Cytotoxic-T-cell responses, viral load, and disease progression in early human immunodeficiency virus type 1 infection**. *N.Engl.J.Med.* 337(18): 1267-1274

Negroni M and Bruc H. (2000). **Copy-choice recombination by reverse transcriptases: Reshuffling of genetic markers mediated by RNA chaperones**. *PNAS* 97(12): 6385-6390

Nelson GW, Kaslow R and Mann DL. (1997). **Frequency of HLA allele-specific peptide motifs in HIV-1 proteins correlates with the allele's association with relative rates of disease progression after HIV-1 infection**. *Proc.Natl.Acad.Sci.U.S.A* 94(18): 9802-9807

Ngandu N, Bredell H, Gray CM, Williamson C and Seoighe C. (2007). **CTL response to HIV type 1 subtype C is poorly predicted by known epitope motifs**. *AIDS Res.Hum.Retroviruses* 23(8): 1033-1041

Ngandu NK, Seoighe C, Scheffler K. (2009). **Evidence of HIV-1 adaptation to host HLA alleles following chimp-to-human transmission**. *Virology Journal* 6:164

Ngandu NK, Scheffler K, Moore P, Woodman Z, Martin D and Seoighe C. (2008). **Extensive purifying selection acting on synonymous sites in HIV-1 Group M sequences**. *Virol.J.* 5(1): 160

Ngumbela KC, Day CL, Mncube Z, Nair K, Ramduth D, Thobakgale C, Moodley E, Reddy S, de Pierres C, Mkhwanazi N, Bishop K, van der SM, Ismail N, Honeyborne I, Crawford H, Kavanagh DG, Rousseau C, Nickle D, Mullins J, Heckerman D, Korber B, Coovadia H, Kiepiela P, Goulder PJ and Walker BD. (2008). **Targeting of a CD8 T cell env epitope presented by HLA-B*5802 is associated with markers of HIV disease progression and lack of selection pressure**. *AIDS Res.Hum.Retroviruses* 24(1): 72-82

Nielsen R and Yang Z. (1998). **Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene**. *Genetics* 148(3): 929-936

Nisole S and Saib A. (2004). **Early steps of retrovirus replicative cycle**. *Retrovirology* 19

Ohno S. (1992). **How cytotoxic T cells manage to discriminate nonself from self at the nonapeptide level**. *Proc.Natl.Acad.Sci.U.S.A* 89(10): 4643-4647

Paillart JC, Skripkin E, Ehresmann B, Ehresmann C and Marquet R. (2002). **In vitro evidence for a long range pseudoknot in the 5'-untranslated and matrix coding regions of HIV-1 genomic RNA**. *J.Biol.Chem.* 277(8): 5995-6004

Pandrea I, Ribeiro RM, Gautam R, Gaufin T, Pattison M, Barnes M, Monjure C, Stoulig C, Dufour J, Cyprian W, Silvestri G, Miller MD, Perelson AS and Apetrei C. (2008). **Simian immunodeficiency virus SIVagm dynamics in African green monkeys**. *J.Virol.* 82(7): 3713-3724

Parker KC, Bednarek MA and Coligan JE. (1994). **Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains**. *J.Immunol.* 152(1): 163-175

Pascual R, Moreno MR and Villalain J. (2005). **A peptide pertaining to the loop segment of human immunodeficiency virus gp41 binds and interacts with model biomembranes: implications for the fusion mechanism**. *J.Virol.* 79(8): 5142-5152

Pavesi A, De Iaco B, Granero MI and Porati A. (1997). **On the informational content of overlapping genes in prokaryotic and eukaryotic viruses**. *J Mol.Evol.* 44(6): 625-631

Peeters M and Sharp PM. (2000). **Genetic diversity of HIV-1: the moving target**. *AIDS* 14 Suppl 3S129-S140

Pereira LA, Bentley K, Peeters A, Churchill MJ and Deacon NJ. (2000). **A compilation of cellular transcription factor interactions with the HIV-1 LTR promoter**. *Nucleic Acids Res.* 28(3): 663-668

Perelson AS, Neumann AU, Markowitz M, Leonard JM and Ho DD. (1996). **HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time**. *Science* 271(5255): 1582-1586

Pereyra F, Addo MM, Kaufmann DE, Liu Y, Miura T, Rathod A, Baker B, Trocha A, Rosenberg R, Mackey E, Ueda P, Lu Z, Cohen D, Wrin T, Petropoulos CJ, Rosenberg ES and Walker BD. (2008). **Genetic and immunologic heterogeneity among persons who control HIV infection in the absence of therapy**. *J.Infect.Dis.* 197(4): 563-571

Peterson RD and Feigon J. (1996). **Structural change in Rev responsive element RNA of HIV-1 on binding Rev peptide**. *J.Mol.Biol.* 264(5): 863-877

Peyerl FW, Bazick HS, Newberg MH, Barouch DH, Sodroski J and Letvin NL. (2004). **Fitness costs limit viral escape from cytotoxic T lymphocytes at a structurally constrained epitope**. *J Virol.* 78(24): 13901-13910

Phuphuakrat A and Auewarakul P. (2003). **Heterogeneity of HIV-1 Rev response element**. *AIDS Res.Hum.Retroviruses* 19(7): 569-574

Piontkivska H and Hughes AL. (2004). **Between-host evolution of cytotoxic T-lymphocyte epitopes in human immunodeficiency virus type 1: an approach based on phylogenetically independent comparisons**. *J Virol.* 78(21): 11758-11765

Powell MD and Levin JG. (1996). **Sequence and structural determinants required for priming of plus-strand DNA synthesis by the human immunodeficiency virus type 1 polypurine tract**. *J.Virol.* 70(8): 5288-5296

Preston B, Poiesz B and Loeb L. (1988). **Fidelity of HIV-1 reverse transcriptase**. *Science* 242(4882): 1168-1171

Pullen KA, Rattray AJ and Champoux JJ. (1993). **The sequence features important for plus strand priming by human immunodeficiency virus type 1 reverse transcriptase**. *J.Biol.Chem.* 268(9): 6221-6227

Pupko T, Bell R, Mayrose I, Glaser F and Ben-Tal N. (2002). **Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues**. *Bioinformatics* 18(1): S71-S77

Quinones-Mateu M, Albright J, Torre V, Reinis M, Vandasova J, Bruckova M and Arts E. (1999). **Molecular epidemiology of HIV type 1 isolates from the Czech Republic: identification of an env E subtype case**. *AIDS Res.Hum.Retroviruses* 15(1): 85-89

Quinones-Mateu M, Mas A, Lain dL, Soriano V, Alcami J, Lederman M and Domingo E. (1998). **LTR and tat variability of HIV-1 isolates from patients with divergent rates of disease progression**. *Virus Res.* 57(1): 11-20

Rammensee H, Bachmann J, Emmerich N, Bachor O and Stevanovic S. (1999). **SYFPEITHI: database for MHC ligands and peptide motifs**. *Immunogenetics* 50(3-4): 213-219

Rammensee H, Friede T and Stevanoviic S. (1995). **MHC ligands and peptide motifs: first listing**. *Immunogenetics* 41(4): 178-228

Rausch JW and Le Grice SF. (2004). **'Binding, bending and bonding': polypurine tract-primed initiation of plus-strand DNA synthesis in human immunodeficiency virus**. *Int.J.Biochem.Cell Biol.* 36(9): 1752-1766

Rausch JW and Le Grice SF. (2007). **Purine analog substitution of the HIV-1 polypurine tract primer defines regions controlling initiation of plus-strand DNA synthesis**. *Nucleic Acids Res.* 35(1): 256-268

Renwick S, Critchley A, Adams C, Kelly S, Price N and Stockley P. (1995). **Probing the details of the HIV-1 Rev-Rev-responsive element interaction: effects of modified nucleotides on protein affinity and conformational changes during complex formation**. *Biochem.J.* 308 (Pt 2)447-453

Richman DD, Wrin T, Little SJ and Petropoulos CJ. (2003). **Rapid evolution of the neutralizing antibody response to HIV type 1 infection**. *Proc.Natl.Acad.Sci.U.S.A* 100(7): 4144-4149

Robertson DL, Anderson JP, Bradac JA, Carr JK, Foley B, Funkhouser RK, Gao F, Hahn BH, Kalish ML, Kuiken C, Learn GH, Leitner T, McCutchan F, Osmanov S, Peeters M, Pieniazek D, Salminen M, Sharp PM, Wolinsky S and Korber B. (2000). **HIV-1 nomenclature proposal**. *Science* 288(5463): 55-56

Rose P and Korber B. (2000). **Detecting hypermutations in viral sequences with an emphasis on G --> A hypermutation**. *Bioinformatics* 16(4): 400-401

Ross HA and Rodrigo AG. (2002). **Immune-mediated positive selection drives human immunodeficiency virus type 1 molecular variation and predicts disease duration**. *J.Virol.* 76(22): 11715-11720

Rousseau CM, Lockhart DW, Listgarten J, Maley SN, Kadie C, Learn GH, Nickle DC, Heckerman DE, Deng W, Brander C, Ndung'u T, Coovadia H, Goulder PJ, Korber BT, Walker BD and Mullins JI. (2009). **Rare HLA drive additional HIV evolution compared to more frequent alleles**. *AIDS Res.Hum.Retroviruses* 25(3): 297-303

Rovero P, Riganelli D, Fruci D, Vigano S, Pegoraro S, Revoltella R, Greco G, Butler R, Clementi S and Tanigaki N. (1994). **The importance of secondary anchor residue motifs of HLA class I proteins: a chemometric approach**. *Mol.Immunol.* 31(7): 549-554

Rutjens E, Balla-Jhagjhoorsingh S, Verschoor E, Bogers W, Koopman G and Heeney J. (2003). **Lentivirus infections and mechanisms of disease resistance in chimpanzees**. *Front Biosci.* 8d1134-d1145

Sabbaj S, Bansal A, Ritter GD, Perkins C, Edwards BH, Gough E, Tang J, Szinger JJ, Korber B, Wilson CM, Kaslow RA, Mulligan MJ and Goepfert PA. (2003). **Cross-reactive CD8+ T cell epitopes identified in US adolescent minorities**. *J Acquir.Immune Defic.Syndr.* 33(4): 426-438

Sagar M, Wu X, Lee S and 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**. *J Virol.* 80(19): 9586-9598

Saksela K, Stevens C, Rubinstein P and Baltimore D. (1994). **Human immunodeficiency virus type 1 mRNA expression in peripheral blood cells predicts disease progression independently of the numbers of CD4+ lymphocytes**. *Proc.Natl.Acad.Sci.U.S.A* 91(3): 1104-1108

Saper MA, Bjorkman PJ and Wiley DC. (1991). **Refined structure of the human histocompatibility antigen HLA-A2 at 2.6 A resolution**. *J.Mol.Biol.* 219(2): 277-319

Sathiamurthy M, Hickman HD, Cavett JW, Zahoor A, Prilliman K, Metcalf S, Fernandez VM and Hildebrand WH. (2003). **Population of the HLA ligand database**. *Tissue Antigens* 61(1): 12-19

Sato K, Fushimi N and Ohya M. (2007). **The Code Structure of HIV-1**. *Open Systems & Information Dynamics* 14(3): 295-306

Scheffler K, Martin D and Seoighe C. (2006). **Robust inference of positive selection from recombining coding sequences**. *Bioinformatics* 22(20): 2493-2499

Schiewe AJ and Haworth IS. (2007). **Structure-based prediction of MHC-peptide association: algorithm comparison and application to cancer vaccine design**. *J.Mol.Graph.Model.* 26(3): 667-675

Schmittel A, Keilholz U and Scheibenbogen C. (1997). **Evaluation of the interferon-gamma ELISPOT-assay for quantification of peptide specific T lymphocytes from peripheral blood**. *J Immunol.Methods* 210(2): 167-174

Schneider R, Campbell M, Nasioulas G, Felber B and Pavlakis G. (1997). **Inactivation of the human immunodeficiency virus type 1 inhibitory elements allows Rev-independent expression of Gag and Gag/protease and particle formation**. *J.Virol.* 71(7): 4892-4903

Schubert U, Ferrer-Montiel AV, Oblatt-Montal M, Henklein P, Strebel K and Montal M. (1996). **Identification of an ion channel activity of the Vpu transmembrane domain and its involvement in the regulation of virus release from HIV-1-infected cells**. *FEBS Lett.* 398(1): 12-18

Schueler-Furman O, Altuvia Y, Sette A and Margalit H. (2000). **Structure-based prediction of binding peptides to MHC class I molecules: application to a broad range of MHC alleles**. *Protein Sci.* 9(9): 1838-1846

Schwartz S, Campbell M, Nasioulas G, Harrison J, Felber B and Pavlakis G. (1992). **Mutational inactivation of an inhibitory sequence in human immunodeficiency virus type 1 results in Rev-independent gag expression**. *J.Virol.* 66(12): 7176-7182

Schwartz S, Felber B, Benko D, Fenyo E and Pavlakis G. (1990a). **Cloning and functional analysis of multiply spliced mRNA species of human immunodeficiency virus type 1**. *J.Virol.* 64(6): 2519-2529

Schwartz S, Felber B, Fenyo E and Pavlakis G. (1990b). **Env and Vpu proteins of human immunodeficiency virus type 1 are produced from multiple bicistronic mRNAs**. *J.Virol.* 64(11): 5448-5456

Seeger FH, Schirle M, Gatfield J, Arnold D, Keilholz W, Nickolaus P, Rammensee HG and Stevanovic S. (1999a). **The HLA-A*6601 peptide motif: prediction by pocket structure and verification by peptide analysis**. *Immunogenetics* 49(6): 571-576

Seeger FH, Schirle M, Keilholz W, Rammensee HG and Stevanovic S. (1999b). **Peptide motif of HLA-B*1510**. *Immunogenetics* 49(11-12): 996-999

Seoighe C, Ketwaroo F, Pillay V, Scheffler K, Wood N, Duffet R, Zvelebil M, Martinson N, McIntyre J, Morris L and Hide W. (2007). **A model of directional selection applied to the evolution of drug resistance in HIV-1**. *Mol.Biol.Evol.* 24(4): 1025-1031

Sette A and Sidney J. (1999). **Nine major HLA class I supertypes account for the vast preponderance of HLA-A and -B polymorphism**. *Immunogenetics* 50(3-4): 201-212

Sette A, Vitiello A, Reherman B, Fowler P, Nayersina R, Kast WM, Melief CJ, Oseroff C, Yuan L, Ruppert J, Sidney J, del Guercio MF, Southwood S, Kubo RT, Chesnut RW, Grey HM and Chisari FV. (1994). **The relationship between class I binding affinity and immunogenicity of potential cytotoxic T cell epitopes**. *J.Immunol.* 153(12): 5586-5592

Sharp P, Bailes E, Chaudhuri R, Rodenburg C, Santiago M and Hahn B. (2001). **The origins of acquired immune deficiency syndrome viruses: where and when?** *Philos.Trans.R.Soc.Lond B Biol.Sci.* 356(1410): 867-876

Sharp P, Bailes E, Robertson D, Gao F and Hahn B. (1999). **Origins and evolution of AIDS viruses**. *Biol.Bull.* 196(3): 338-342

Sharp P, Robertson D and Hahn B. (1995). **Cross-species transmission and recombination of 'AIDS' viruses**. *Philos.Trans.R.Soc.Lond B Biol.Sci.* 349(1327): 41-47

Shearer WT, Quinn TC, LaRussa P, Lew JF, Mofenson L, Almy S, Rich K, Handelsman E, Diaz C, Pagano M, Smeriglio V and Kalish LA. (1997). **Viral load and disease progression in infants infected with human immunodeficiency virus type 1. Women and Infants Transmission Study Group**. *N.Engl.J.Med.* 336(19): 1337-1342

Shpaer E and Mullins J. (1993). **Rates of amino acid change in the envelope protein correlate with pathogenicity of primate lentiviruses**. *J.Mol.Evol.* 37(1): 57-65

Shriner D, Nickle DC, Jensen MA and Mullins JI. (2003). **Potential impact of recombination on sitewise approaches for detecting positive natural selection**. *Genet.Res.* 81(2): 115-121

Silvestri G. (2008). **Immunity in natural SIV infections**. *Journal of INTERNAL MEDICINE* 101365-2796

Silvestri G, Paiardini M, Pandrea I, Lederman M and Sodora D. (2007). **Understanding the benign nature of SIV infection in natural hosts**. *J.Clin.Invest* 117(11): 3148-3154

Simon JH, Gaddis NC, Fouchier RA and Malim MH. (1998). **Evidence for a newly discovered cellular anti-HIV-1 phenotype**. *Nat.Med.* 4(12): 1397-1400

Singh M. (2006). **No vaccine against HIV yet--are we not perfectly equipped?** *Virol.J* 360

Soares AE, Soares MA and Schrago CG. (2008). **Positive selection on HIV accessory proteins and the analysis of molecular adaptation after interspecies transmission**. *J.Mol.Evol.* 66(6): 598-604

Solberg OD, Mack SJ, Lancaster AK, Single RM, Tsai Y, Sanchez-Mazas A and Thomson G. (2008). **Balancing selection and heterogeneity across the classical**

human leukocyte antigen loci: a meta-analytic review of 497 population studies. *Hum.Immunol* 69(7): 443-464

Storey J, Taylor J and Siegmund D. (2004). **Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach.** *Journal of the Royal Statistical Society* 66187-205

Subbramanian RA, Xu J, Toma E, Morisset R, Cohen EA, Menezes J and Ahmad A. (2002). **Comparison of human immunodeficiency virus (HIV)-specific infection-enhancing and -inhibiting antibodies in AIDS patients**. *J Clin.Microbiol.* 40(6): 2141-2146

Swanson AK and Stoltzfus CM. (1998). **Overlapping cis sites used for splicing of HIV-1 env/nef and rev mRNAs**. *J.Biol.Chem.* 273(51): 34551-34557

Thompson JD, Gibson TJ and Higgins DG. (2002). **Multiple sequence alignment using ClustalW and ClustalX**. *Curr.Protoc.Bioinformatics* Chapter 2Unit

Todar K (2009). Todar's Online Textbook of Bacteriology. http://www.textbookofbacteriology.net/index.html

Trachtenberg E, Korber B, Sollars C, Kepler TB, Hraber PT, Hayes E, Funkhouser R, Fugate M, Theiler J, Hsu YS, Kunstman K, Wu S, Phair J, Erlich H and Wolinsky S. (2003). **Advantage of rare HLA supertype in HIV disease progression**. *Nat.Med.* 9(7): 928-935

Travers SA, O'Connell MJ, McCormack GP and McInerney JO. (2005). **Evidence for heterogeneous selective pressures in the evolution of the env gene in different human immunodeficiency virus type 1 subtypes**. *J Virol.* 79(3): 1836-1841

van Baalen CA, Pontesilli O, Huisman RC, Geretti AM, Klein MR, de Wolf F, Miedema F, Gruters RA and Osterhaus AD. (1997). **Human immunodeficiency virus type 1 Rev- and Tat-specific cytotoxic T lymphocyte frequencies inversely correlate with rapid progression to AIDS**. *J.Gen.Virol.* 78 (Pt 8)1913-1918

Van Lint C, Ghysdael J, Paras P, Jr., Burny A and Verdin E. (1994). **A transcriptional regulatory element is associated with a nuclease-hypersensitive**

site in the pol gene of human immunodeficiency virus type 1. *J.Virol.* 68(4): 2632-2648

Vartanian JP, Meyerhans A, Asjo B and Wain-Hobson S. (1991). **Selection, recombination, and G----A hypermutation of human immunodeficiency virus type 1 genomes**. *J Virol.* 65(4): 1779-1788

Vartanian J, Meyerhans A, Asjo B and Wain-Hobson S. (1991). **Selection, recombination, and G----A hypermutation of human immunodeficiency virus type 1 genomes**. *J Virol.* 65(4): 1779-1788

Verdin E, Becker N, Bex F, Droogmans L and Burny A. (1990). **Identification and characterization of an enhancer in the coding region of the genome of human immunodeficiency virus type 1**. *Proc.Natl.Acad.Sci.* 874874-4878

Wain LV, Bailes E, Bibollet-Ruche F, Decker JM, Keele BF, Van Heuverswyn F, Li Y, Takehisa J, Ngole EM, Shaw GM, Peeters M, Hahn BH and Sharp PM. (2007). **Adaptation of HIV-1 to its human host**. *Mol.Biol.Evol.* 24(8): 1853-1860

Wei X, Decker JM, Wang S, Hui H, Kappes JC, Wu X, Salazar-Gonzalez JF, Salazar MG, Kilby JM, Saag MS, Komarova NL, Nowak MA, Hahn BH, Kwong PD and Shaw GM. (2003). **Antibody neutralization and escape by HIV-1**. *Nature* 422(6929): 307-312

Wilkinson KA, Gorelick RJ, Vasa SM, Guex N, Rein A, Mathews DH, Giddings MC and Weeks KM. (2008). **High-throughput SHAPE analysis reveals structures in HIV-1 genomic RNA strongly conserved across distinct biological states**. *PLoS Biol.* 6(4): e96

Williams T. (2001). **Human leukocyte antigen gene polymorphism and the histocompatibility laboratory**. *J Mol.Diagn.* 3(3): 98-104

Wilson CM, Houser J, Partlow C, Rudy BJ, Futterman DC and Friedman LB. (2001). **The REACH (Reaching for Excellence in Adolescent Care and Health) project: study design, methods, and population profile**. *J Adolesc.Health* 29(3 Suppl): 8-18

Wolff H, Brack-Werner R, Neumann M, Werner T and Schneider R. (2003). **Intergrated functional and bioinformatics approach for the identification and experimental verification of RNA signals: application to HIV-1 INS**. *Nucleic Acids Research* 31(11): 2839-2851

Wong WS and Nielsen R. (2004). **Detecting selection in noncoding regions of nucleotide sequences**. *Genetics* 167(2): 949-958

Wu Y. (2004). **HIV-1 gene expression: lessons from provirus and non-integrated DNA**. *Retrovirology* 113

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

Yagita H, Nakata M, Kawasaki A, Shinkai Y and Okumura K. (1992). **Role of perforin in lymphocyte-mediated cytolysis**. *Adv.Immunol.* 51215-242

Yang OO, Kalams SA, Trocha A, Cao H, Luster A, Johnson RP and Walker BD. (1997). **Suppression of human immunodeficiency virus type 1 replication by CD8+ cells: evidence for HLA class I-restricted triggering of cytolytic and noncytolytic mechanisms**. *J Virol.* 71(4): 3120-3128

Yang QE, Stephen AG, Adelsberger JW, Roberts PE, Zhu W, Currens MJ, Feng Y, Crise BJ, Gorelick RJ, Rein AR, Fisher RJ, Shoemaker RH and Sei S. (2005a). **Discovery of small-molecule human immunodeficiency virus type 1 entry inhibitors that target the gp120-binding domain of CD4**. *J Virol.* 79(10): 6122-6133

Yang Z. (1998). **Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution**. *Mol.Biol.Evol.* 15(5): 568-573

Yang Z. (2007). **PAML 4: phylogenetic analysis by maximum likelihood**. *Mol.Biol.Evol.* 24(8): 1586-1591

Yang Z, Nielsen R, Goldman N and Pedersen AM. (2000). **Codon-substitution models for heterogeneous selection pressure at amino acid sites**. *Genetics* 155(1): 431-449

Yang Z, Wong WS and Nielsen R. (2005b). **Bayes empirical bayes inference of amino acid sites under positive selection**. *Mol.Biol.Evol.* 22(4): 1107-1118

Yusim K, Kesmir C, Gaschen B, Addo M, Altfeld M, Brunak S, Chigaev A, Detours V and Korber B. (2002). **Clustering patterns of cytotoxic T-lymphocyte epitopes in human immunodeficiency virus type 1 (HIV-1) proteins reveal imprints of immune evasion on HIV-1 global variation**. *J.Virol.* 76(17): 8757-8768

Yusim K, Szinger J, Honeyborne I, Calef C, Goulder P and Korber B. (2003). **Enhanced Motif Scan: A Tool to Scan for HLA Anchor Residues in Proteins**. *HIV Immunol HIV/SIV Vaccine Databases*25-36

Zanotto PM, Kallas EG, de Souza RF and Holmes EC. (1999). **Genealogical evidence for positive selection in the nef gene of HIV-1**. *Genetics* 153(3): 1077-1089

Zhu T, Korber BT, Nahmias AJ, Hooper E, Sharp PM and Ho DD. (1998). **An African HIV-1 sequence from 1959 and implications for the origin of the epidemic**. *Nature* 391(6667): 594-597

APPENDIX

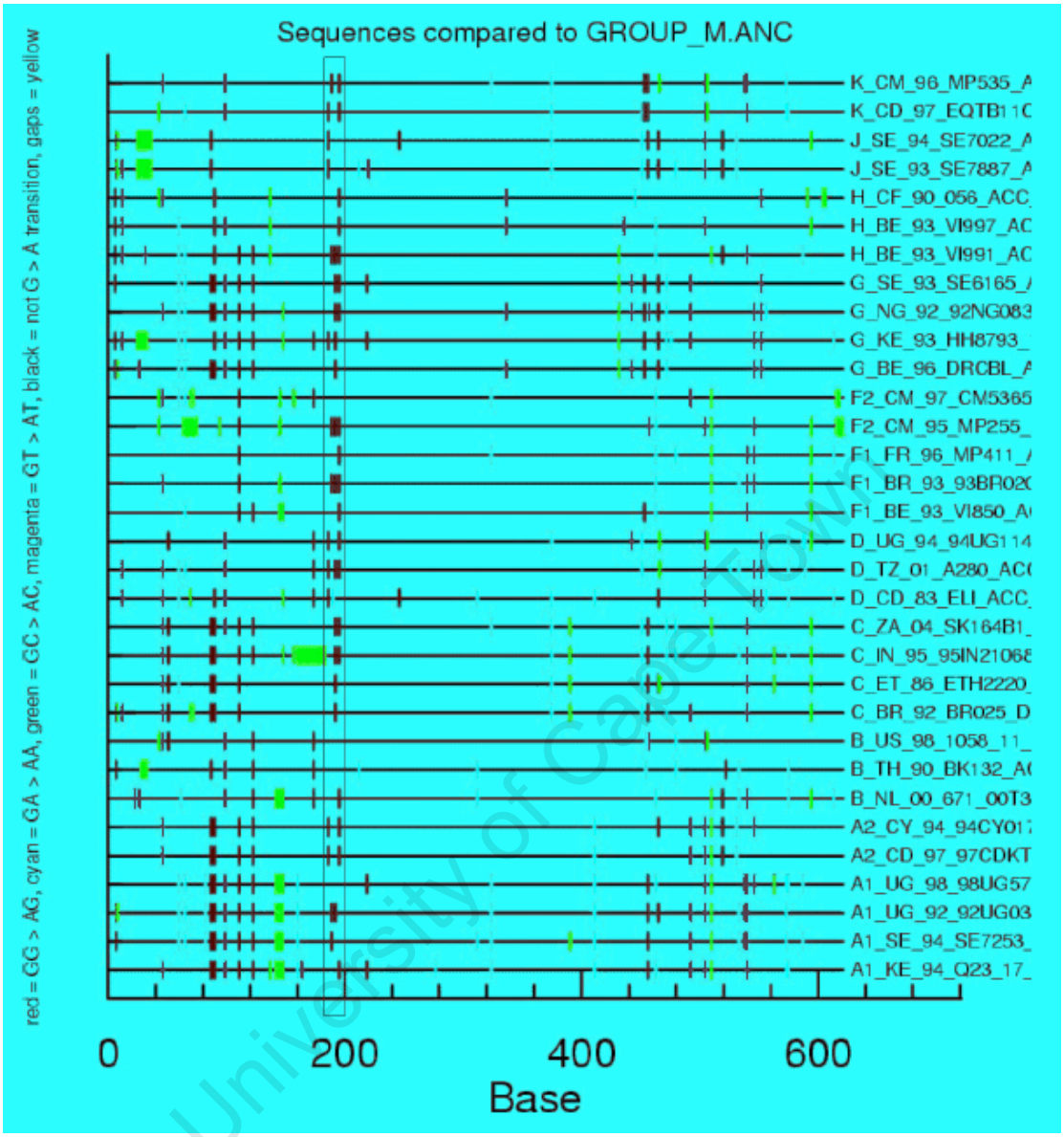


Figure A3.4.2a: G-A mutations in a variable region in the *nef* gene, Mutations observed in reference sequences in comparison to Group M ancestral sequence identified using the hypermut tool available in the Los Alamos database. The highly variable region (labeled "G-A" in Figure 3d) showed G-A mutations and is boxed in red.

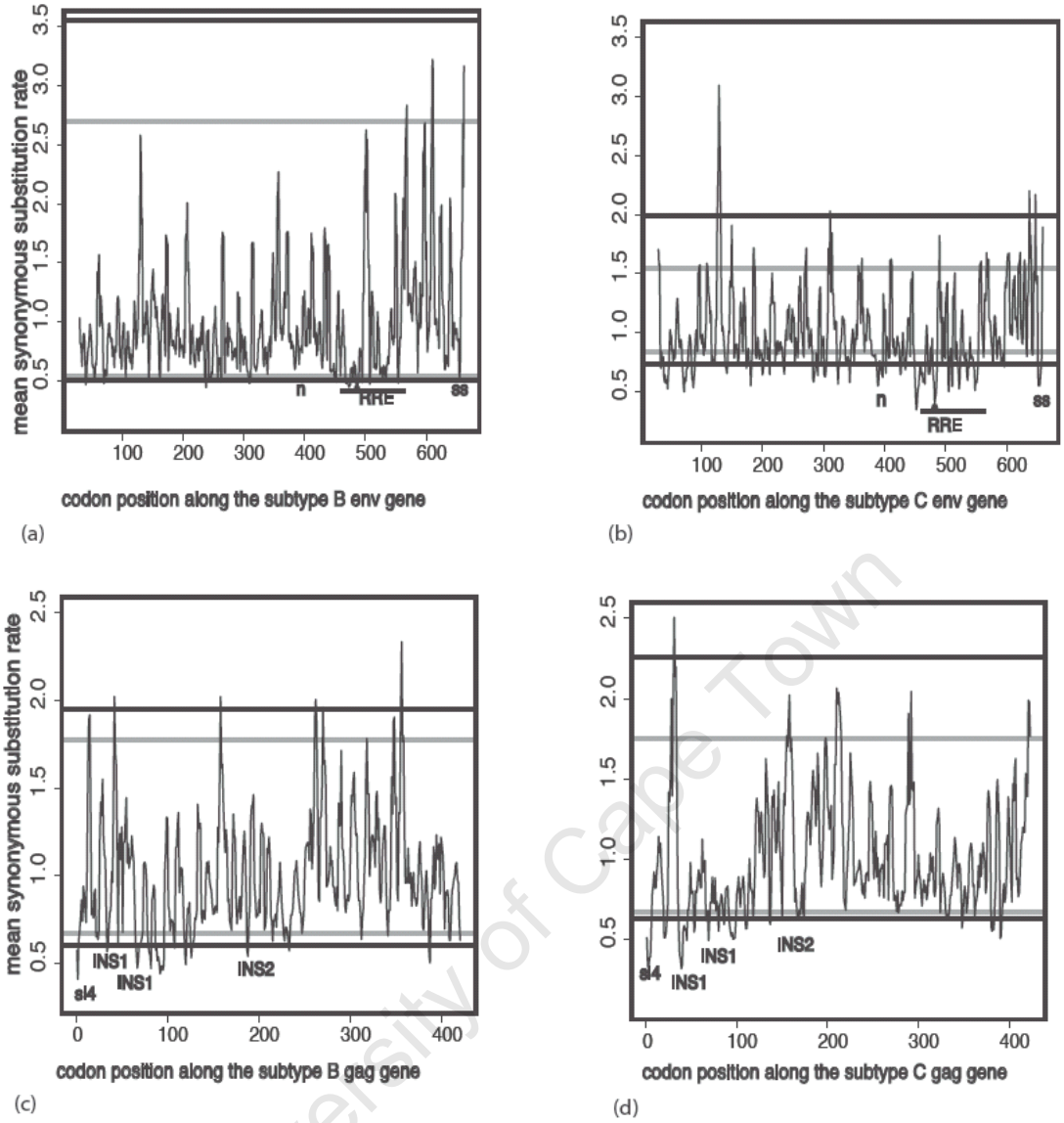


Figure A3.5: Highly conserved regions observed at the subtype level, dS across subtypes B and C *gag* and *env* genes showing more conserved sites at the subtype sequence level within the INS regions in *gag* and RRE in *env*. Black horizontal lines indicate the most stringent cuttoff and grey ones indicate the 5 percentile significance level.

Table A4.3.5 HLA allele frequency in Subtype B and C cohorts used in this study

HLA Allele	Subtype B	Subtype C
A*0101	0.044444	0.007813
A*0201	0.166667	0.039063
A*0202	0.022222	0.015625
A*0205	0.011111	0.046875
A*0207	0.011111	Nil
A*0217	0.011111	Nil
A*0209	Nil	0.007813
A*03	Nil	0.007813
A*0301	0.1	0.007813
A*1101	0.022222	Nil
A*2301	0.1	0.109375
A*2402	0.044444	0.015625
A*2601	0.033333	Nil
A*29	Nil	0.03125
A*2902	0.011111	0.140625
A*2911	Nil	0.007813
A*30	Nil	0.007813
A*3001	0.044444	0.0625
A*3002	0.088889	0.148438
A*3004	0.011111	0.03125
A*3201	0.022222	0.007813
A*3301	0.033333	Nil
A*3303	0.011111	0.007813
A*3304	0.011111	Nil
A*3402	0.022222	0.054688
A*36	Nil	0.007813
A*3601	Nil	0.039063
A*4301	Nil	0.023438
A*6601	0.022222	0.023438
A*6801	0.022222	0.023438
A*6802	0.011111	0.070313
A*74	Nil	0.007813
A*7401	0.055556	0.0625
A*7402	Nil	0.007813
A*7403	0.011111	Nil
A*8001	0.011111	0.023438
B*0702	0.044444	0.03125
B*0801	0.055556	0.0625
B*13	Nil	0.070313
B*14	0.022222	0.023438
B*1402	0.022222	0.007813
B*15	Nil	0.007813
B*1501	0.033333	0.015625
B*1503	0.011111	0.109375
B*1510	Nil	0.046875
B*1516	0.011111	0.007813

B*1530	0.011111	Nil
B*18	Nil	0.007813
B*1801	0.011111	0.03125
B*2705	0.022222	0.007813
B*35	0.011111	Nil
B*3501	0.066667	0.023438
B*3503	0.011111	Nil
B*3801	0.011111	Nil
B*3902	0.011111	Nil
B*3910	Nil	0.007813
B*40	0.011111	Nil
B*4001	0.011111	Nil
B*4002	0.022222	Nil
B*4101	Nil	0.015625
B*4102	0.011111	Nil
B*4201	0.033333	0.085938
B*4202	Nil	0.007813
B*44	Nil	0.007813
B*4402	0.022222	Nil
B*4403	0.033333	0.117188
B*4501	0.088889	Nil
B*45	Nil	0.0625
B*4901	0.033333	0.007813
B*5101	0.033333	Nil
B*5301	0.1	0.03125
B*5308	Nil	0.007813
B*5601	0.011111	Nil
B*57	Nil	0.046875
B*5701	0.022222	Nil
B*5703	0.055556	Nil
B*58	Nil	0.03125
B*5801	0.022222	0.046875
B*5802	0.022222	0.078125
B*7020	Nil	0.007813
B*81	Nil	0.007813
B*8101	0.066667	0.023438
B*8201	0.022222	Nil

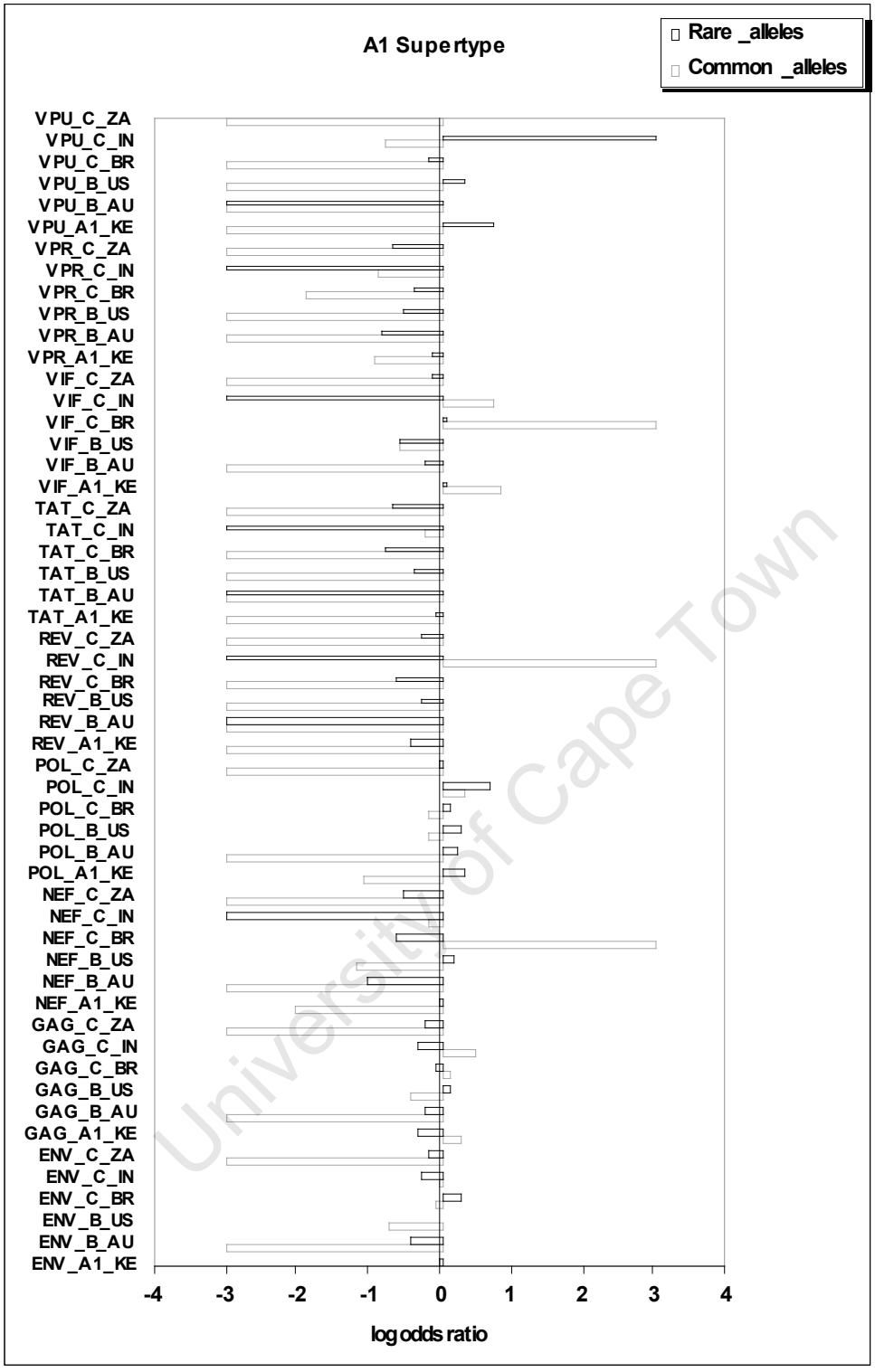


Figure A5.4.2.2a

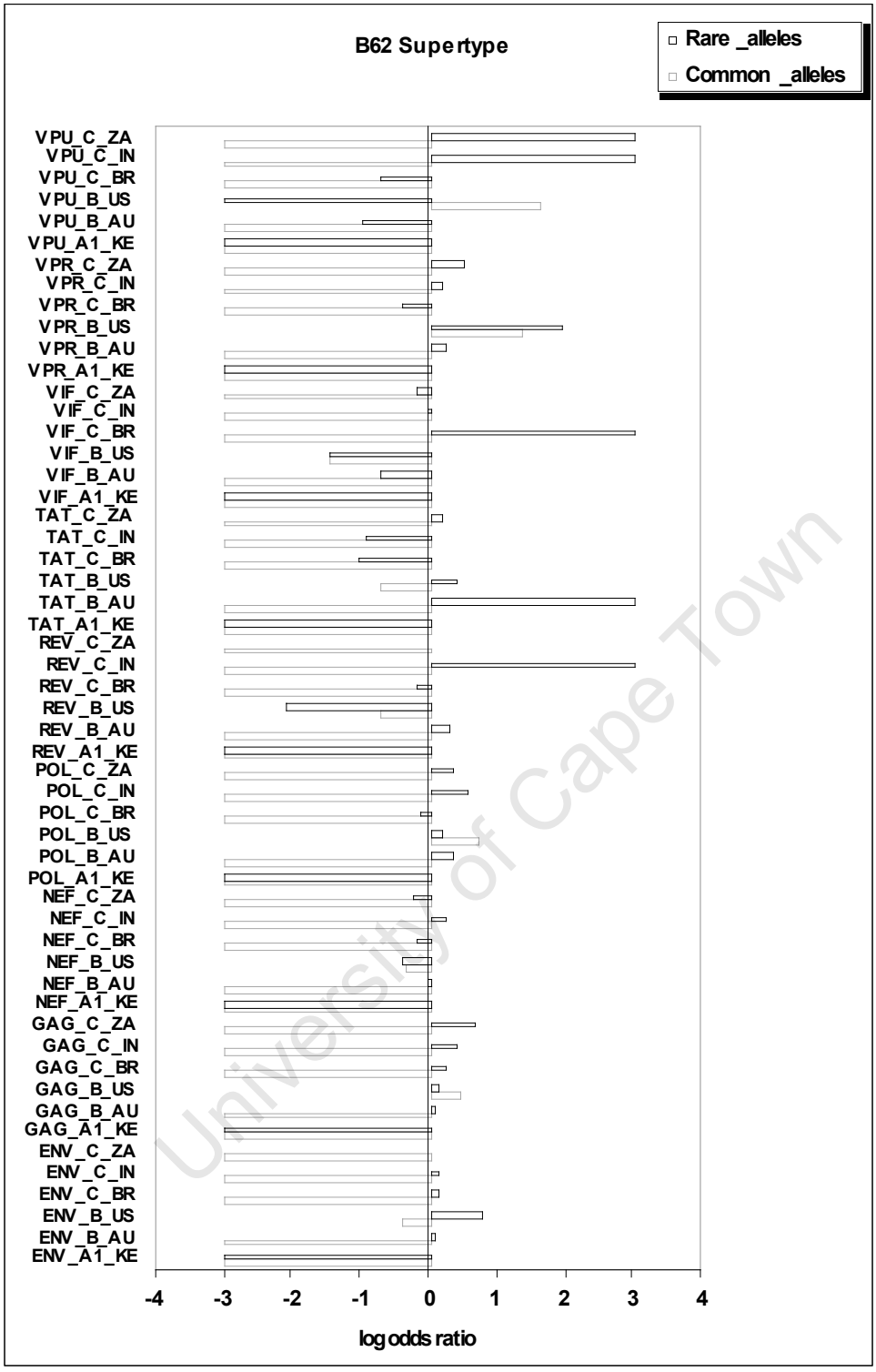


Figure A5.4.2.2b

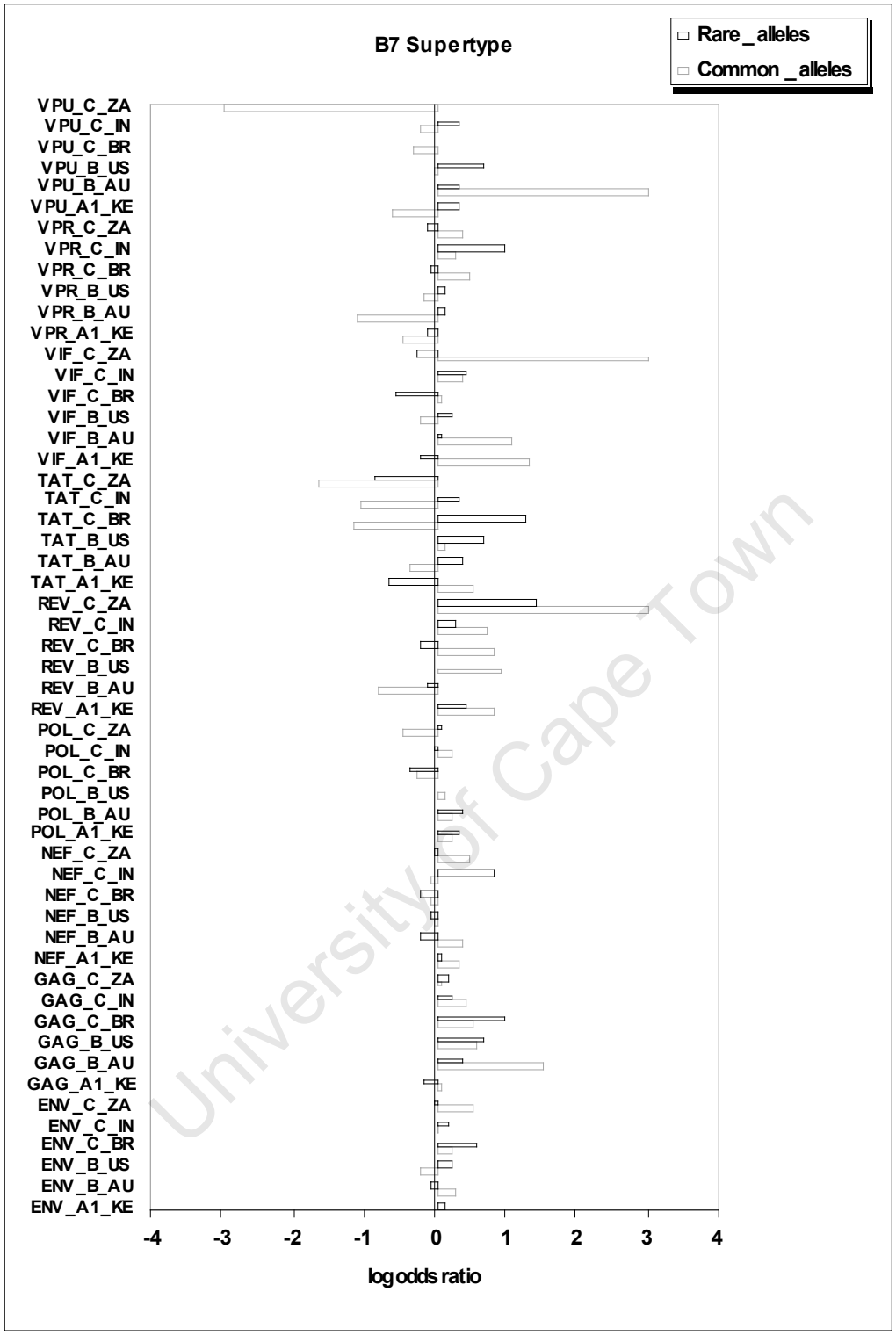


Figure A5.4.2.2c

Figure A5.4.2.2 Log of the odds ratio values obtained from the association of fewer motifs with common alleles (grey) and rare alleles (black) within individual supertypes where no significant associations were found, a; A1 supertype, b; B62 supertype and c; B7 supertype. Log odds ratios greater than 0 indicate lack of motifs

at variable sites and those less than 0 indicate lack of motifs at conserved sites. For the purposes of the plots, all infinite and zero odds ratio values were assigned the log values 3 and -3 respectively. All other log odds ratios were greater than -3 and less than 3.

Table A6.4.3a: Model Averaged Branch dN/Ds for HLA A0201: branches with high model-averaged support for dN>dS are shown in bold

Branch Name	Mean	Std.Dev.	2.5%	Median	97.5%	Prob{dN>dS}†
CPZ_US_85_CPZUS_AF103818	0.630	0.027	0.612	0.628	0.648	0.003
CPZ_CM_05_SIVCPZEK505_DQ373065	1.113	0.055	1.046	1.120	1.162	**0.988**
Node4	1.113	0.054	1.047	1.120	1.163	**0.989**
CPZ_CM_05_SIVCPZMT145_DQ373066	0.639	0.070	0.612	0.628	0.685	0.023
CPZ_GA_88_GAB1_X52154	0.630	0.032	0.612	0.628	0.649	0.005
CPZ_CM_01_SIVCPZCAM13_AY169968	1.116	0.044	1.051	1.120	1.163	**0.994**
Node9	0.637	0.059	0.612	0.628	0.681	0.016
Node7	0.724	0.186	0.616	0.632	1.119	0.200
Node3	0.634	0.055	0.611	0.628	0.652	0.014
CPZ_CD_90_ANT_U42720	1.117	0.034	1.053	1.120	1.163	**0.997**
CPZ_TZ_01_TAN1_AF447763	0.629	0.023	0.611	0.628	0.649	0.002
Node12	1.113	0.052	1.046	1.120	1.161	**0.988**
Node2	1.107	0.081	0.984	1.119	1.163	**0.974**
CPZ_CM_05_SIVCPZLB7_DQ373064	0.679	0.139	0.614	0.630	1.101	0.099
CPZ_CM_05_SIVCPZMB66_DQ373063	1.078	0.133	0.632	1.116	1.159	0.918
Node15	0.673	0.141	0.612	0.629	1.127	0.093
A1_RW_92_92RW008_AB253421	1.117	0.050	1.052	1.120	1.163	**0.996**
A1_UG_92_92UG037_AB253429	1.078	0.135	0.628	1.116	1.161	0.917
A1_KE_94_Q23_17_AF004885	1.109	0.068	1.036	1.119	1.163	**0.982**
Node24	0.651	0.104	0.612	0.628	1.114	0.047
Node22	1.099	0.144	0.629	1.119	1.163	**0.953**
A1_AU_PS1044_DAY0_DQ676872	1.115	0.055	1.048	1.120	1.163	**0.991**
Node21	1.087	0.121	0.628	1.118	1.162	0.936
A2_CD_97_97CDKTB48_AF286238	1.093	0.105	0.635	1.117	1.159	0.948
A2_CY_94_94CY017_41_AF286237	0.634	0.054	0.611	0.628	0.652	0.013
Node28	0.647	0.091	0.612	0.629	1.090	0.039
Node20	1.093	0.110	0.631	1.118	1.162	0.948
G_BE_96_DRCBL_AF084936	0.633	0.046	0.612	0.628	0.651	0.010
G_KE_93_HH8793_12_1_AF061641	1.087	0.120	0.631	1.117	1.160	0.937
G_PT_PT2695_AY612637	0.631	0.041	0.611	0.628	0.650	0.007
Node34	9.241	285.455	0.628	1.117	1.162	0.930
G_NG_92_92NG083_U88826	0.641	0.078	0.612	0.628	1.060	0.028
Node33	1.084	0.133	0.624	1.119	1.163	0.924

Node31	0.658	0.117	0.612	0.629	1.120	0.062
Node19	0.649	0.097	0.612	0.629	1.091	0.044
C_ET_86_ETH2220_U46016	0.630	0.033	0.611	0.628	0.649	0.005
C_BR_92_BR025_D_U52953	1.112	0.059	1.045	1.120	1.163	**0.987**
Node42	0.639	0.075	0.611	0.628	0.675	0.024
C_ZA_04_SK164B1_AY772699	0.642	0.079	0.612	0.629	1.059	0.030
Node41	0.666	0.128	0.612	0.629	1.121	0.077
J_SE_93_SE7887_AF082394	17.020	398.613	0.645	1.119	1.163	**0.971**
J_SE_94_SE7022_AF082395	16.396	390.728	0.642	1.119	1.163	**0.969**
Node46	1.106	0.077	0.960	1.119	1.162	**0.973**
Node40	0.656	0.112	0.612	0.629	1.121	0.057
H_BE_93_VI991_AF190127	1.105	0.077	0.895	1.118	1.161	**0.970**
H_CF_90_056_AF005496	0.630	0.030	0.612	0.628	0.649	0.004
H_BE_93_VI997_AF190128	0.640	0.073	0.612	0.628	0.677	0.024
Node51	0.643	0.084	0.612	0.628	1.081	0.031
Node49	0.855	0.236	0.618	0.645	1.142	0.469
Node39	0.647	0.102	0.611	0.628	1.098	0.043
K_CD_97_EQTB11C_AJ249235	0.669	0.130	0.612	0.629	1.100	0.086
K_CM_96_MP535_AJ249239	1.079	0.134	0.630	1.118	1.162	0.917
Node57	0.653	0.108	0.612	0.629	1.114	0.051
F2_CM_97_CM53657_AF377956	1.112	0.057	1.044	1.120	1.162	**0.987**
Node56	0.640	0.074	0.612	0.628	0.682	0.024
F1_FR_96_MP411_AJ249238	0.630	0.036	0.611	0.628	0.649	0.005
F1_BE_93_VI850_AF077336	0.778	0.217	0.616	0.634	1.128	0.314
F1_BR_93_93BR020_1_AF005494	1.096	0.104	0.636	1.119	1.162	**0.953**
Node64	0.655	0.112	0.612	0.628	1.121	0.056
F1_FI_93_FIN9363_AF075703	0.656	0.109	0.612	0.629	1.101	0.057
Node63	0.649	0.100	0.611	0.628	1.110	0.044
Node61	1.109	0.071	1.032	1.120	1.163	**0.980**
Node55	1.100	0.097	0.638	1.119	1.163	**0.960**
B_TH_90_BK132_AY173951	0.678	0.143	0.612	0.630	1.107	0.107
B_US_98_1058_11_AY331295	0.736	0.196	0.614	0.632	1.127	0.227
Node71	0.638	0.069	0.611	0.628	0.661	0.021
B_FR_83_HXB2_LAI_IIIB_BRU_K03455	0.629	0.028	0.611	0.628	0.648	0.003
Node70	0.808	0.230	0.614	0.636	1.138	0.376
B_US_98_15384_1_DQ853463	1.113	0.050	1.047	1.120	1.163	**0.989**
Node69	1.103	0.088	0.639	1.119	1.163	**0.968**
D_TZ_01_A280_AY253311	0.627	0.027	0.611	0.628	0.647	0.002
D_CD_83_ELI_K03454	0.631	0.036	0.611	0.628	0.649	0.006
D_UG_94_94UG114_U88824	0.631	0.035	0.611	0.628	0.649	0.005
Node78	0.647	0.092	0.612	0.629	1.088	0.040
Node76	0.630	0.032	0.612	0.628	0.649	0.005
Node68	0.650	0.099	0.612	0.629	1.092	0.046

Node54	0.645	0.089	0.612	0.628	1.092	0.035
Node38	1.065	0.155	0.625	1.116	1.163	0.888
Node18	1.117	0.037	1.052	1.120	1.163	**0.996**

Table A6.4.3b: Model Averaged Branch dN/Ds for HLA A6801: branches with high model-averaged support for dN>dS are shown in bold

Branch Name	Mean	Std.Dev.	2.5%	Median	97.5%	Prob{dN>dS}†
CPZ_US_85_CPZUS_AF103818	1.129	0.032	1.066	1.134	1.171	**0.999**
CPZ_CM_05_SIVCPZEK505_DQ373065	1.122	0.063	1.052	1.133	1.168	**0.986**
Node5	1.103	0.115	0.607	1.132	1.168	**0.952**
CPZ_CD_90_ANT_U42720	1.122	0.061	1.053	1.133	1.166	**0.987**
CPZ_TZ_01_TAN1_AF447763	0.598	0.043	0.543	0.598	0.625	0.007
Node8	1.127	0.042	1.062	1.134	1.170	**0.995**
Node4	1.102	0.120	0.602	1.132	1.169	0.950
CPZ_CM_05_SIVCPZMT145_DQ373066	1.103	0.111	0.617	1.132	1.157	**0.951**
CPZ_GA_88_GAB1_X52154	0.628	0.120	0.549	0.600	1.087	0.064
CPZ_CM_01_SIVCPZCAM13_AY169968	1.122	0.065	1.052	1.134	1.170	**0.986**
Node13	1.114	0.090	0.619	1.133	1.170	**0.972**
Node11	0.328	0.152	0.218	0.242	0.602	0.008
Node3	1.128	0.042	1.065	1.134	1.171	**0.997**
CPZ_CM_05_SIVCPZMB66_DQ373063	0.600	0.056	0.541	0.598	0.628	0.010
Node2	1.124	0.077	1.054	1.134	1.171	**0.987**
CPZ_CM_05_SIVCPZLB7_DQ373064	0.611	0.087	0.545	0.599	1.049	0.032
A1_RW_92_92RW008_AB253421	0.601	0.064	0.542	0.598	0.630	0.014
A1_KE_94_Q23_17_AF004885	0.605	0.082	0.541	0.599	0.646	0.022
Node23	0.593	0.110	0.262	0.598	1.045	0.026
A1_UG_92_92UG037_AB253429	0.591	0.061	0.532	0.598	0.625	0.005
Node22	3.610	159.511	0.241	1.131	1.169	0.908
A1_AU_PS1044_DAY0_DQ676872	0.695	0.194	0.588	0.604	1.119	0.189
Node21	1.092	0.141	0.597	1.132	1.168	0.931
A2_CD_97_97CDKTB48_AF286238	0.596	0.048	0.540	0.598	0.625	0.006
A2_CY_94_94CY017_41_AF286237	1.111	0.100	0.612	1.133	1.170	**0.966**
Node28	0.600	0.078	0.538	0.598	0.633	0.018
Node20	7.889	259.986	1.059	1.134	1.171	**0.992**
G_BE_96_DRCBL_AF084936	0.620	0.107	0.547	0.600	1.080	0.048
G_NG_92_92NG083_U88826	1.117	0.089	1.033	1.133	1.170	**0.978**
Node32	0.599	0.100	0.334	0.598	0.656	0.025
G_KE_93_HH8793_12_1_AF061641	0.619	0.116	0.541	0.599	1.127	0.048

G_PT_PT2695_AY612637	0.594	0.041	0.540	0.598	0.623	0.003
Node35	0.951	0.266	0.539	1.125	1.169	0.672
Node31	0.285	0.106	0.208	0.236	0.560	0.000
Node19	1.117	0.089	1.011	1.133	1.170	**0.976**
C_ET_86_ETH2220_U46016	1.123	0.068	1.053	1.134	1.171	**0.987**
C_BR_92_BR025_D_U52953	1.126	0.052	1.059	1.134	1.171	**0.993**
Node42	0.291	0.113	0.211	0.236	0.567	0.001
C_ZA_04_SK164B1_AY772699	1.116	0.084	0.999	1.133	1.170	**0.975**
Node41	1.109	0.107	0.605	1.133	1.169	**0.963**
K_CD_97_EQTB11C_AJ249235	1.127	0.072	1.059	1.134	1.171	**0.992**
K_CM_96_MP535_AJ249239	0.595	0.072	0.533	0.598	0.628	0.012
Node47	1.114	0.095	0.615	1.133	1.170	**0.972**
F1_FR_96_MP411_AJ249238	0.594	0.034	0.541	0.598	0.622	0.002
F1_BE_93_VI850_AF077336	0.604	0.090	0.538	0.599	0.873	0.025
F1_BR_93_93BR020_1_AF005494	1.117	0.083	1.028	1.133	1.169	**0.978**
Node54	1.097	0.138	0.561	1.133	1.169	0.943
F1_FI_93_FIN9363_AF075703	1.146	0.474	1.063	1.134	1.171	**0.995**
Node53	0.607	0.090	0.540	0.599	1.083	0.027
Node51	1.125	0.061	1.059	1.134	1.171	**0.991**
F2_CM_97_CM53657_AF377956	1.123	0.061	1.055	1.134	1.171	**0.988**
Node50	0.604	0.093	0.537	0.598	1.063	0.026
Node46	0.839	0.266	0.549	0.616	1.155	0.465
Node40	0.655	0.190	0.501	0.600	1.142	0.129
B_TH_90_BK132_AY173951	0.597	0.053	0.541	0.598	0.626	0.008
B_FR_83_HXB2_LAI_IIIB_BRU_K03455	1.117	0.083	1.021	1.134	1.171	**0.977**
Node62	0.316	0.163	0.218	0.238	0.601	0.019
B_US_98_15384_1_DQ853463	1.118	0.079	1.034	1.134	1.170	**0.979**
Node61	0.309	0.156	0.212	0.237	0.599	0.018
B_US_98_1058_11_AY331295	0.588	0.049	0.507	0.598	0.622	0.001
Node60	1.114	0.094	0.625	1.133	1.169	**0.972**
D_TZ_01_A280_AY253311	0.596	0.045	0.540	0.598	0.624	0.005
D_CD_83_ELI_K03454	0.615	0.106	0.541	0.599	1.113	0.041
Node68	1.115	0.095	0.617	1.133	1.171	**0.972**
D_UG_94_94UG114_U88824	0.608	0.086	0.541	0.599	1.056	0.027
Node67	1.116	0.088	0.976	1.133	1.170	**0.975**
Node59	0.602	0.064	0.542	0.599	0.631	0.014
Node39	0.315	0.151	0.213	0.238	0.598	0.012
J_SE_93_SE7887_AF082394	1.104	0.124	0.600	1.133	1.171	**0.956**
J_SE_94_SE7022_AF082395	0.603	0.090	0.537	0.599	0.948	0.024
Node73	1.099	0.121	0.605	1.132	1.167	0.944
H_BE_93_VI991_AF190127	1.117	0.084	1.016	1.133	1.170	**0.976**
H_CF_90_056_AF005496	1.104	0.115	0.605	1.132	1.168	**0.955**
H_BE_93_VI997_AF190128	1.110	0.101	0.609	1.133	1.169	**0.964**

Node78	0.595	0.077	0.527	0.598	0.629	0.013
Node76	0.630	0.131	0.546	0.600	1.127	0.069
Node72	1.111	0.106	0.607	1.133	1.170	**0.968**
Node38	1.100	0.123	0.597	1.132	1.169	0.947
Node18	1.128	0.040	1.064	1.134	1.171	**0.996**

Table A6.4.3c: Model Averaged Branch dN/Ds for HLA B2705: branches with high model-averaged support for dN>dS are shown in bold

Branch Name	Mean	Std.Dev.	2.5%	Median	97.5%	Prob{dN>dS}†
C_ET_86_ETH2220_U46016	0.506	0.020	0.468	0.515	0.529	0.000
C_BR_92_BR025_D_U52953	0.506	0.025	0.468	0.515	0.529	0.000
Node5	0.507	0.046	0.466	0.515	0.532	0.001
C_ZA_04_SK164B1_AY772699	0.508	0.026	0.473	0.515	0.534	0.000
Node4	0.516	0.060	0.475	0.515	0.652	0.008
K_CD_97_EQTB11C_AJ249235	17.789	414.835	0.476	0.515	0.971	0.016
K_CM_96_MP535_AJ249239	0.506	0.020	0.468	0.515	0.529	0.000
Node10	0.504	0.041	0.466	0.515	0.531	0.000
F1_FR_96_MP411_AJ249238	0.508	0.044	0.468	0.515	0.531	0.004
F1_BR_93_93BR020_1_AF005494	0.508	0.032	0.468	0.515	0.533	0.000
Node15	74.891	859.119	0.473	0.515	0.688	0.019
F1_FI_93_FIN9363_AF075703	0.509	0.035	0.474	0.515	0.544	0.000
Node14	0.505	0.027	0.468	0.515	0.529	0.000
F1_BE_93_VI850_AF077336	0.539	0.115	0.477	0.515	0.936	0.012
F2_CM_97_CM53657_AF377956	0.509	0.030	0.473	0.515	0.545	0.000
Node19	9.339	297.041	0.468	0.515	0.529	0.002
Node13	0.513	0.062	0.468	0.515	0.546	0.003
Node9	0.524	0.070	0.475	0.515	0.800	0.000
Node3	2.385	136.929	0.468	0.515	0.531	0.005
J_SE_93_SE7887_AF082394	0.507	0.049	0.468	0.515	0.545	0.000
J_SE_94_SE7022_AF082395	62.748	786.222	0.474	0.515	0.829	0.020
Node22	0.516	0.053	0.475	0.515	0.657	0.001
Node2	0.505	0.023	0.468	0.515	0.529	0.000
H_BE_93_VI991_AF190127	0.512	0.069	0.468	0.515	0.534	0.006
H_CF_90_056_AF005496	0.503	0.032	0.466	0.515	0.529	0.000
H_BE_93_VI997_AF190128	0.508	0.024	0.475	0.515	0.542	0.000
Node27	0.515	0.061	0.472	0.515	0.617	0.006
Node25	0.505	0.029	0.468	0.515	0.529	0.000
Node1	0.514	0.065	0.475	0.515	0.607	0.006
CPZ_US_85_CPZUS_AF103818	0.566	0.135	0.481	0.515	0.979	0.017
CPZ_CM_05_SIVCPZEK505_DQ373065	0.552	0.113	0.481	0.515	0.901	0.002

Node34	0.520	0.138	0.474	0.515	0.607	0.007
CPZ_CM_05_SIVCPZMT145_DQ373066	0.532	0.105	0.475	0.515	0.875	0.014
CPZ_GA_88_GAB1_X52154	0.548	0.110	0.481	0.515	0.909	0.001
CPZ_CM_01_SIVCPZCAM13_AY169968	0.547	0.138	0.475	0.515	1.036	0.035
Node39	0.521	0.071	0.475	0.515	0.792	0.005
Node37	0.507	0.029	0.473	0.515	0.532	0.000
Node33	77.942	876.219	0.475	0.515	1.232	0.043
CPZ_CD_90_ANT_U42720	0.514	0.039	0.475	0.515	0.648	0.000
CPZ_TZ_01_TAN1_AF447763	0.512	0.043	0.468	0.515	0.565	0.000
Node42	0.604	0.177	0.488	0.516	1.086	0.050
Node32	38.390	614.142	0.473	0.515	0.916	0.021
CPZ_CM_05_SIVCPZLB7_DQ373064	0.508	0.025	0.473	0.515	0.538	0.000
Node31	0.514	0.059	0.468	0.515	0.714	0.000
CPZ_CM_05_SIVCPZMB66_DQ373063	0.506	0.023	0.468	0.515	0.531	0.000
Node30	0.512	0.049	0.472	0.515	0.546	0.001
A1_RW_92_92RW008_AB253421	0.509	0.030	0.475	0.515	0.546	0.000
A1_KE_94_Q23_17_AF004885	0.504	0.049	0.421	0.515	0.536	0.000
Node51	0.505	0.029	0.468	0.515	0.529	0.000
A1_AU_PS1044_DAY0_DQ676872	0.512	0.038	0.475	0.515	0.607	0.000
Node50	0.506	0.026	0.468	0.515	0.531	0.000
A1_UG_92_92UG037_AB253429	0.506	0.022	0.468	0.515	0.529	0.000
A2_CD_97_97CDKTB48_AF286238	0.531	0.096	0.473	0.515	0.904	0.001
A2_CY_94_94CY017_41_AF286237	0.504	0.032	0.461	0.515	0.528	0.000
Node57	0.506	0.025	0.468	0.515	0.529	0.000
Node55	8.063	274.786	0.468	0.515	0.529	0.001
Node49	0.524	0.282	0.468	0.515	0.536	0.005
G_BE_96_DRCBL_AF084936	0.505	0.025	0.468	0.515	0.529	0.000
G_NG_92_92NG083_U88826	0.513	0.068	0.468	0.515	0.555	0.007
Node62	18.052	418.486	0.468	0.515	0.531	0.003
G_PT_PT2695_AY612637	0.506	0.022	0.468	0.515	0.529	0.000
Node61	13.033	353.648	0.468	0.515	0.532	0.005
G_KE_93_HH8793_12_1_AF061641	0.506	0.027	0.468	0.515	0.529	0.000
Node60	0.506	0.024	0.468	0.515	0.531	0.000
Node48	0.509	0.036	0.473	0.515	0.543	0.001
B_TH_90_BK132_AY173951	0.561	0.225	0.480	0.515	1.006	0.026
B_FR_83_HXB2_LAI_IIIB_BRU_K03455	0.502	0.032	0.444	0.515	0.527	0.000
Node70	4.174	191.289	0.475	0.515	0.607	0.005
B_US_98_15384_1_DQ853463	0.507	0.031	0.468	0.515	0.531	0.000
Node69	0.490	0.075	0.188	0.514	0.528	0.000
B_US_98_1058_11_AY331295	0.506	0.028	0.468	0.515	0.529	0.000
Node68	0.509	0.031	0.474	0.515	0.549	0.000
D_TZ_01_A280_AY253311	0.497	0.053	0.280	0.515	0.528	0.000
D_UG_94_94UG114_U88824	0.510	0.054	0.468	0.515	0.532	0.003

Node76	0.508	0.044	0.468	0.515	0.531	0.002
D_CD_83_ELI_K03454	0.502	0.032	0.441	0.515	0.527	0.000
Node75	0.509	0.029	0.474	0.515	0.544	0.000
Node67	0.505	0.029	0.468	0.515	0.529	0.000
Node47	0.513	0.050	0.475	0.515	0.633	0.001