Molecular Characterisation of the HIV-1 Epidemic in Karonga District, Malawi, with Particular Emphasis on Long-Term Survivors

A thesis submitted for the degree of Ph.D. to the National University of Ireland, Galway by:

Ishla Seager

Supervisors:

Dr. Grace P McCormack, B.Sc., Ph.D. (NUI)

Dr. Simon A Travers, B.Sc., Ph.D. (NUI)

School of Natural Sciences,

National University of Ireland,

Galway

September 2011

Acknowledgments

Firstly I would like to thank my supervisor Dr Grace McCormack. Through the highs and lows of the past few years you have been a constant source of inspiration and will be forever a women I both admire and look up to. Your faith and encouragement in me will stay with me forever. Thank you is completely inadequate. You are nothing less than amazing.

And Dr Simon Travers, whether here or oceans away your insight and ideas have always been able to lift even the heaviest clouds of confusion. Thank you for patience and everything you have taught me over the years.

To Mom and Dad, I have no idea how to express how much I appreciate EVERYTHING you have ever done for me, so I dedicate this thesis to you. I love you both. A big Thanks to Zak for your encouragement and faith.

Thank you to all my partners in crime in the zoology department over the last four years. The mammal ecology lab rocks! Special thanks to Cat, I can't wait to celebrate your success soon too, and Peter, thanks for all the book recommendations, we all missed you loads when you left.

Thank you Kelly, Conor and Vijay for making me laugh, and supporting me through it all, I was very lucky to be able to share a lab and office with such great people.

To Niamh Quinn, thank you for catching my tears in the middle and despite being in a different country, holding me up at the end. There's no doubt that I would not have finished if it weren't for you your friendship (and your amazing dinners, a nice fire, and Enchanted).

To Niamh Redmond, I really wish you could be here to celebrate with me, you made the first few years of my PhD so much fun, but I may never forgive you for abandoning me! I miss you lots and thank you for all your words of encouragement over the years.

To the girls at home, Janette, Katherine and Sarah Jane (my own personal cheerleader), thank you for all your encouragement and understanding. I'm lucky to have you guys. And Ciara, What can I say other than I think your fantastic and thank you for all the phone calls, they kept me going through it all.

There are some many others over the years who have been there for me, Anne for all the tea and adventures, Katrina, Jenny, Lorraine Clancy (who is too beautiful for words). Thank you all.

A big "That's what she said" to all the people in the Eric Arts lab. I remember my time there to be so full of laughter in spite of the science! Thank you Dr Eric Arts for letting come to his lab and teaching me so much. Thank you to Ashley, Mel and Kendall and a special thank you to Susannah for taking me in and treating me like a true friend. Most of all thank you to Rick for his immense patience and support, and for so graciously imparting his knowledge to me.

Thank you to everyone out in KPS who have allowed me be part of the amazing work they do.

Thank you to all the staff and technicians in the Zoology Department for teaching me everything I know from my first year as an undergrad to today. You are all Awesome and I am so glad I got to be part of the zoology Department.

Table of Contents

Abbreviations

°C	Degrees Celsius
A	Adenine
aa	Amino Acid
AIDS	Acquired Immune Defiency Syndrome
ART	Anti Retroviral Therapy
bp	Base Pairs
C	Cytosine
CCR5	C-C Chemokine Receptor 5
CD4	Cluster of Differentiation 4
CRFs	Circulation Recombinant Forms
CXCR4	C-X-C Receptor 4
DBS	Dried Blood Spots
DNA	Deoxyribonucleic Acid
dNTPs	Deoxynucleoside Triphosphates
FPR	False Positive Ratio
G	Guanine
GTR	General Time Reversible
HARRT	Highly Active Anti retroviral Therapy
HIV	Human Immunodeficiency Virus
HSS	HIV-1 Sero Survery
HTLV	Human T-cell Leukemia Virus
indels	Insertions and Deletions
KPS	Karonga Prevention Study
LTNPs	Long-Term Non-Progressors
LTS	Long-Term Survivors
Ma	Malawi
$MgCl_2$	Magnesium Chloride
ML	Maximum Likelihood
ml	Milliliters
mM	Millimoles
nfl	Near Full Lentgh
PCR	Polymerase Chain Reaction
pmol	Picomole
RNA	Ribonucleic Acid
RT	Reverse Trancriptase
SIV	Simian Immunodeficiency Virus
T	Thymine
TB	Tuburculosis
URFs	Unique Recombinant Forms
WHO	World Health Organisation
µl	Microliters
µM	Micromoles

Chapter 1: General Introduction

1 Chapter 1

1.1 The history of discovery of HIV

HIV-1 is the causative agent of AIDS and is responsible for the deaths of over 25 million people since the beginning of the pandemic almost 30 years ago (UNAIDS, 2009b). The first cases of AIDS in America occurred in 1981, though at that time the cause of AIDS was unknown. Kaposi's Sarcoma, a rare form of a relatively benign cancer that was normally seen in older people was found, in a much more aggressive form, in eight young gay men in New York in 1981 (Hymes et al., 1981). At the same time there was an increase in the number of cases of a rare lung infection, *Pneumocystis carinii* pneumonia, in both California and New York (MMWR Weekly, 1981), which was soon noticed by the Centres of Disease Control (CDC). A number of theories emerged about the possible causes of these opportunistic infections and cancers, some of which included infection with cytomegalovirus (a herpes virus), the use of the recreational drug amyl nitrate or butyl nitrate (poppers), a fungus and 'immune overload' (Goedert et al., 1982; Gottlieb et al., 1981). Identifying the cause of AIDS presented a unique challenge, because unlike other viral diseases responsible for past epidemics, AIDS was characterised by clinical signs that developed years after infection had occurred, by which point patients usually had numerous other infections. The various manifestations of AIDS were however, unified by one biologic marker, a decrease in the levels of $CD4^+$ cells (Gottlieb et al., 1981; Masur et al., 1981).

It soon became clear that the disease was not limited to homosexual men as the disease was soon found in other population groups such as injecting drug users (Masur et al., 1981) and haemophiliacs (MMWR Weekly, 1982). In 1983 it was thought that the disease might be caused by an infectious agent that was both sexually transmitted and transmitted through blood products (Francis, Curran, and Essex, 1983). At the time of the emergence of AIDS, it was known that a retrovirus, HTLV-1, was also transmitted through sexual and blood exposure and also targeted $CD4^+$ T cells. This led a number of scientists to propose that AIDS might be caused by a new retrovirus of the HTLV family (Gallo, 2002). In 1983 a group in France successfully cultured a retrovirus from the peripheral blood cells of a patient with AIDS, but found that it was not related to HTLV and appeared to be a new retrovirus (Barre-Sinoussi et al., 1983). It took only two more years for the HIV-1 genome to be cloned and

fully sequenced (Wain-Hobson et al., 1985). In the following two years blood tests became commercially available, which reduced to almost to zero the transmission of AIDS through blood transfusion in developed countries (Barre-Sinoussi et al., 1983). Since its first discovery a vast amount of knowledge has been amassed on HIV, and with the advent of new molecular and cellular techniques research on HIV continues to be extensive.

1.2 The replication cycle of HIV-1

Infection begins with virus entry into the host cell. This is an intricate and complex multistep process that involves both viral proteins (envelope proteins gp120 and gp41), and the corresponding receptor proteins present on the host cells. Early in the AIDS epidemic it was observed that HIV specifically targets and depletes the $CD4^+$ subset of T-lymphocytes in the peripheral blood of infected individuals (Bour, Geleziunas, and Wainberg, 1995; Harper et al., 1986). The primary receptor for HIV is this glycoprotein CD4, which is expressed on the surface of T helper cells, regulatory T cells, monocytes, macrophages and dendritic cells (Harper et al., 1986). CD4 normally functions as a coreceptor (with the T cell receptor) to activate T-cells following interaction with antigen- presenting cells during an immune response. The gp41 envelope proteins are anchored to, and span the viral membrane with the extracellular domain binding (non- covalently) to the envelope protein gp120 (Tavassoli, 2010). These viral proteins associate into trimers (Zhu et al., 2006) and mediate binding to the CD4 receptor and entry into the target cells (Figure 1.1: step 1).

The binding of gp41 to CD4 causes gp120 to undergo conformational changes (while still maintaining its association with gp41) (Pancera et al., 2010), which subsequently results in the exposure of a second binding site to a chemokine host cell coreceptor (Kwong et al., 1998). Twelve different chemokine receptors have been identified experimentally as being able to function as coreceptors for HIV-1, but only two have been observed *in vivo* (Doms and Trono, 2000). CCR5 tropic non-syncytium inducing viruses bind to macrophages with the CCR5 cell surface protein, while CXCR4 tropic syncytium inducing viruses bind to the CXCR4 cell surface protein on T cells (Greene and Peterlin, 2002). In subtype B infections, the early phase of HIV-1 infection is generally associated with CCR5 tropic viruses with CXCR4 tropic viruses being generally dominant during the later stage of the disease (Scarlatti et al., 1997). It had been previously reported that subtype C isolates almost

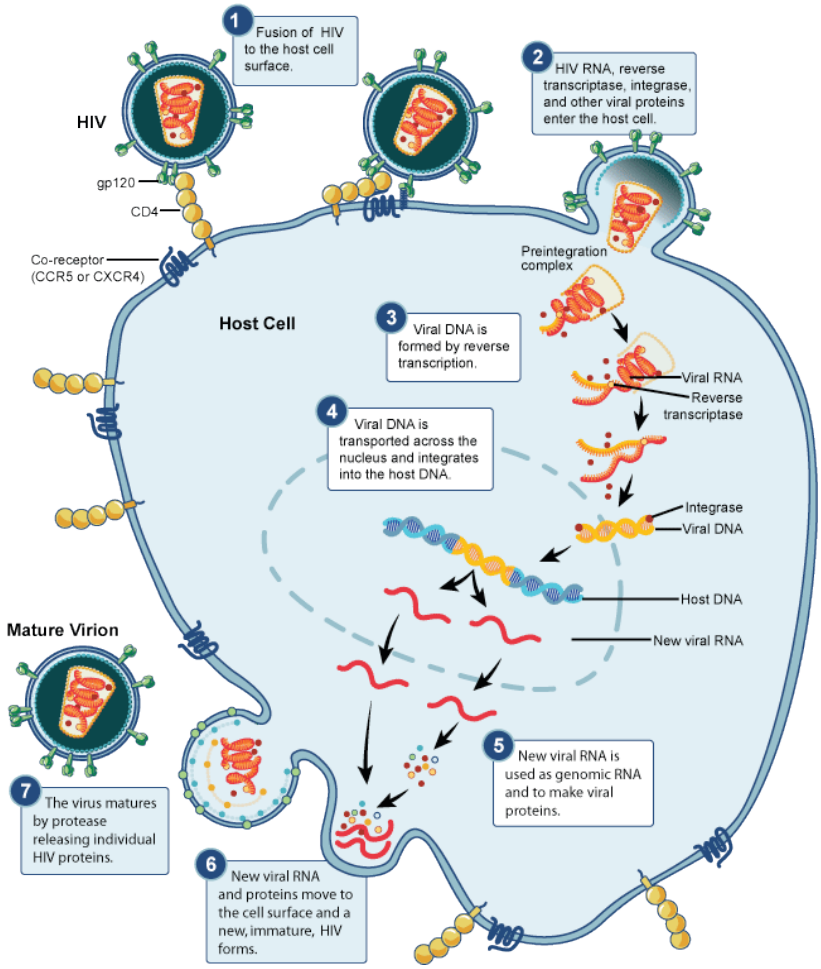


Figure 1.1 The HIV-1 Replication cycle from binding to the host cell, entry into the host cell, un-coating of the virus, reverse transcription into the host genome, viron assembly and virion budding (Source www.niaid.nih.gov).

exclusively use the CCR5 coreceptor (Michler et al., 2008). However a number of more recent studies have have indicated that CXCR4 tropic subtype C viruses may be more wide spread than initially thought with prevelance levels of CXCR4 viruses of approximatly 30 % reported in two subtype C studies (Connell et al., 2008; Raymond et al., 2010).

Following binding to the coreceptor (CCR5 or CXCR4), gp120 (now anchored to the host by two protein-protein interactions) undergoes a further conformational shift, which brings a hydrophobic region in gp41 close to the host cell resulting in its insertion into the host cell's membrane. A fusion pore is then formed enabling the viral core to enter the cell cytoplasm (Chan and Kim, 1998) (Figure 1.1: step 2). The events that follow remain the least understood stages of the virus life cycle whereby the "uncoating" of the virus occurs during which the core of the virus is converted to the preintegration complex (PIC) (Freed, 2001; Nisole and Saib, 2004). Upon entry of the viral capsid into the cell cytoplasm, reverse transcription begins and leads to the generation of double-stranded DNA (cDNA) from the single-stranded RNA genome (Miller, Farnet, and Bushman, 1997; Negroni and Galetto, 2009) (Figure 1.1:step 3). The PIC is composed of both viral and cellular proteins. Known cellular proteins associated with the PIC include BAF, HRP-2, and LEDGF/p75 (Al-Mawsawi and Neamati, 2007; Friedrich et al., 2011; Vandegraaff et al., 2006). Viral elements include the reverse transcribed linear blunt-ended viral DNA, integrase, matrix, and Vpr (Tavassoli, 2010). The PIC is driven towards the nucleus by the nuclear targeting action of Vpr (Fouchier et al., 1998; Popov et al., 1998). It travels down cellular microtubules to the nucleus using cytoplasmic dynein motors where it then must cross the nuclear membrane (McDonald et al., 2002). This is achieved using nuclear localization signals and a number of host cell nuclear import factors including transportin-SR2 (TNO3/TRN-SR2) and importin (Christ et al., 2008; Goff, 2008; Lee et al., 2010; Zaitseva et al., 2009).

Following the nuclear import of the viral PIC the viral protein integrase catalyzes the insertion of the viral DNA into the host cell chromosome where it is known as a provirus (Brown, 1997) (Figure 1.1: step 4). Once the provirus is established, the *cis*-acting elements that regulate viral transcription in the U3 region of the LTR in the provirus, lead to the synthesis of full length genomic RNA from the 5' LTR region. The U3 region within the LTR contains the core promoter as well as enhancer sequences. The promoter harbours a

TATA box and three Sp1 transcription factor sites (Negroni and Galetto, 2009; Taube et al., 1999). The 5'LTR also contains sites for several host transcription factors. Among these are the binding site(s) for NF-kB (a host complex that controls DNA transcription) which plays a central role in mediating and inducing HIV gene expression (Freed, 2001; Tavassoli, 2010). The proviral insert is subsequently transcribed into RNA using the viral enzyme RNA polymerase and cellular host apparatus, a process that is promoted by the viral protein Tat, though interactions with cellular transcription machinery (Zhou et al., 1998; Zhou and Sharp, 1995).

Transcription from the HIV-1 LTR leads to the generation of a large number of viral mRNAs (Purcell and Martin, 1993). These fall into three major classes: 1) unspliced RNAs, which function as the mRNAs for the Gag and GagPol polyprotein precursors and are packaged into progeny virions as genomic RNA, 1) partially spliced mRNAs (around 5kb in size) which encode the Env, Vif, Vpu and Vpr proteins and 3) small (1.7 to 2.0kb), multiply spliced mRNAs which are translated into Rev, Tat and Nef (Freed, 2001). The viral mRNA is initially fully spliced by the host's splicing factors. The viral protein Rev and a *cis*-acting RNA element, the Rev responsive element (RRE) located in the *env* gene is responsible for producing partially spliced and full length viral mRNA (Malim et al., 1989). Rev binds to the RRE and enhances the export of unspliced mRNAs from the nuclease (Strebel, 2003).

Following the synthesis of the full complement of viral proteins, the assembly process begins (Figure 1.1: step 5). The Gag precursor polyprotein ($Pr55^{gag}$) is responsible for viral assembly. This protein contains determinants that target it to the plasma cell membrane, then bind the membrane itself where subsequent Gag-Gag interactions form the structural shell of the budding virual particle (Bieniasz, 2009; Ganser-Pornillos, Yeager, and Sundquist, 2008). The matrix (MA) domain of Gag also recruits Env glycoprotein's to the plasma membrane and the nucleocapsid (NC) region of Gag recognises and binds the dimeric full length viral RNA transcript. Gag also binds to and actively directs the packaging of other viral proteins to the assembling particle (Freed, 1998; Freed, 2001) (Figure 1.1: step 6). Host proteins are also recruited by Gag to aid in the budding and release of the particle from the cell (Goff et al., 2003; Strack et al., 2003). In particular HIV-1 uses the host cell's endosomal sorting complex required for transport (ESCRT) pathway which promotes membrane budding within

the host cell and in turn the HIV-1 virion (Weiss and Gottlinger, 2011). The resulting virion buds through the cell membrane and is released. Cells can become infected with more than one virus. At the virion packaging stage two heterogeneous strands of RNA may be packaged into the one virion. This can subsequently result in both inter or intra-subtype recombinant genomes. Once budding has occurred, viral protease is used within the particle to cleave protein precursors into functioning proteins resulting in an infectious particle that can go on to infect a new cell (Kohl et al., 1988) (Figure 1.1: step 7).

1.3 The structure of the HIV-1 genome

The HIV-1 genome is over 9000 nucleotides long and is flanked by long terminal repeats at both ends (Figure 1.2). The full HIV-1 genome codes for nine genes. These genes are split into three main categories based on the function of the resulting proteins: regulatory genes (*tat* and *rev*), accessory genes (*vif, vpr, vpu* and *nef*) and structural genes (*gag, pol* and *env*).

1.3.1 Structural genes

Env

The *env* gene encodes for a protein, gp160, that is later cleaved into two proteins, gp120 and gp41 (Montagnier et al., 1985). These two proteins are found on the cell surface in a trimeric complex called the envelope spike, which consists of three gp120 monomers and three gp41 monomers (Farzan et al., 1998; Lu, Blacklow, and Kim, 1995). The gp120 is located on the surface of the viron and its function is the binding of the host receptor CD4 found on the cell surface. This is followed by interaction with a cell surface co receptor (CCR5 or CXCR4) (Dragic et al., 1996; Feng et al., 1996; Landau, Warton, and Littman, 1988). The gp120 gene region is further subdivided into five variable regions (V1-V5) and five constant regions (C1-C5) (Leonard et al., 1990; Pollard et al., 1992). This highly variable nature of *env* has been used to subdivide the HIV-1 Group M epidemic into nine different subtypes using phylogenetic analysis (Robertson et al., 2000). The average genetic variability between subtypes is approximately 25 % in *env* (Buonaguro, Tornesello, and Buonaguro, 2007). The V3 loop is composed of 35 amino acids with a conserved disulfide bridge at the base (De Jong et al., 1992; Fouchier et al., 1992). The V3 loop plays a vital role in HIV-1 tropism by determining which co receptor the virus can use to enter the cell, either CCR5 or CXCR4 (Hwang et al., 1991). The gp41 protein is a trans-membrane protein that contains an N-

terminal ectodomain, a trans-membrane domain and an intra-viral domain, the latter of which interacts with the Gag matrix protein. The function of gp41 is the creation of a pore between the viral and cellular membrane allowing the virion contents to enter the host cell (Gallaher, 1987). Certain sequence mutations and insertions within the *env* gene have been associated with long-term non-progression in certain infected individuals (Alexander et al., 2000; Wang et al., 2000).

Gag

The *gag* gene is approximately 1500 nucleotides in length. The initial precursor protein, p55, becomes associated with the host cell membrane after translation, and recruits other viral proteins and RNA to the site before virion assembly. During, or shortly after the virus buds from the cell, HIV-1 protease cleaves the p55 into the four proteins, the matrix (MA, p17), the capsid (CA, p24), the nucleocapsid (NC, p9) and p6 (Gottlinger, 2001). Cleavage of p55 also produces two small spacer peptides, p1 and p2. Although the precise function for these domains are not known, they are thought to be involved in regulating the rate of cleavage of p55 (Krausslich et al., 1995; Pettit et al., 1994). The MA forms a protective shell attached to the inner surface of the plasma membrane of the virus (Gottlinger, Sodroski, and Haseltine, 1989). It has also been implicated in targeting of Gag to the plasma membrane and Env glycoprotein incorporation into virions (Ehrlich et al., 1996; Hermida-Matsumoto and Resh, 2000; Zhou et al., 1994). The MA is also essential in early post entry events of the virus life cycle and is involved in recruiting the host cellular nuclear import machinery to the viral pre-integration complex (PIC) and is required for the efficient nuclear import of the PIC (Reil et al., 1998). The CA forms the conical shell around the core of the HIV-1 virion (Gelderblom, 1991). It has also been suggested that the CA plays an important role in virus assembly and early post entry steps (Reviewed in (Freed, 1998). The NC is essential for the specific packaging of two copies of the genomic viral RNA into the assembling viral particle (Berkowitz, Fisher, and Goff, 1996). It is a highly basic protein and has a non-specific nucleic acid binding activity and assists in various annealing reactions during reverse transcription (Berkowitz, Fisher, and Goff, 1996). The p6 protein is unique to lentiviruses and is known to be involved in the release of the virus from the host cell (Gottlinger et al., 1991). It is also essential for the incorporation of the Vpr viral protein in to the viral particle (Paxton, Connor, and Landau, 1993). Phylogenetic analysis of the diversity in *gag* has been

used to divide HIV-1 group M into nine of subtypes with an average 15 % diversity noted between different subtypes (Buonaguro, Tornesello, and Buonaguro, 2007). Small deletions in the C terminus of MA and p6, and mutations within *gag* have been identified within HIV-1 long-term survivors. Direct links, however, between the mutations and the reduced pathogenesis of the virus have not yet been established and require further investigation (Alexander et al., 2000; Huang, Zhang, and Ho, 1998; Miura et al., 2008).

Pol

The *pol* gene lacks an initiation codon, partially overlaps, and is in the -1 reading frame with respect to *gag*. As a result Pol is only synthesized as part of a Gag-Pol fusion protein (Pr160-Gag-Pol)(Hill, Tachedjian, and Mak, 2005). A small proportion of ribosomes undergo a -1 ribosomal frame shift during translation of the Gag polyprotein $Pr55^{gag}$ to facilitate Pol protein expression (Jacks et al., 1988; Parkin, Chamorro, and Varmus, 1992). Further processing results in four viral enzymes, protease (PR, p10), reverse transcriptase (RT, p51), RNaseH (p15) and integrase (IN, p31). The PR enzyme is involved in the cleaving of the *gag-pol* precursor protein into the structural products of *gag* and the enzymatic products of *pol* (Ross et al., 1991). RT is responsible for transcribing RNA into cDNA during the HIV-1 life cycle (Miller, Farnet, and Bushman, 1997). This allows for the viral genome to be subsequently incorporated into the DNA of the host cell, an essential part in the life cycle of a retrovirus. The lack of fidelity and of proof reading ability by RT during this process is vital in the rapid evolution of HIV-1 resulting in highly diverse viral population within an infected individual and across populations of hosts (Rambaut et al., 2004). RNaseH removes the RNA template after RT has finished synthesizing the DNA to allow for synthesis of complementary DNA (Wintersberger, 1990). The IN enzyme is involved in the insertion of the newly created DNA into the host genome where it is now known as proviral DNA (Andrake and Skalka, 1996). This proviral DNA can be easily extracted using molecular techniques and sequenced enabling detailed studies on the molecular epidemiology of HIV-1.

1.3.2 Regulatory genes

Tat

The Tat protein is a transcriptional transactivator found in the nucleus of infected cells and is involved in the promotion of transcription of the integrated viral DNA (Feinberg, Baltimore, and Frankel, 1991). Unlike other transactivators which only bind to DNA, HIV-Tat can bind to HIV proviral DNA (El Kharroubi et al., 1998), host cell DNA (Pessler and Cron, 2004), and viral RNA. Tat allows for a several fold increase in the rate of HIV transcription after binding to the TAR element found in the LTR promoter (Bres et al., 2002). Tat is not only required for viral transcription but also ensures that full-length viral RNA transcripts are produced through the commandeering of the elongation factor P-TEFb, found in human host cells (Zhou et al., 1998).

Rev

The Rev protein is involved in the nuclear export of the viral mRNA in the correct unspliced form (Malim et al., 1989). An RNA binding motif in Rev interacts with the Rev responsive element (RRE) found in *env* to facilitate the export of unspliced or partially spliced HIV-1 RNA genomes from the nucleus (Malim et al., 1989; Truant and Cullen, 1999). It contains both a nuclear localization sequence and nuclear export sequence which allows it to shuttle between the nucleus and the cytoplasm of the cell in order to accomplish its function (Meyer, Meinkoth, and Malim, 1996; Nekhai and Jeang, 2006). A number of host proteins such as Crm1 (Fukuda et al., 1997) and importin-ß (Truant and Cullen, 1999) are also utilised in this process. Genomic defects described in *rev* have been associated with long-term non-progression within the individuals infected with the *rev* attenuated virus (Papathanasopoulos et al., 2003) (Iversen et al., 1995; Oelrichs et al., 1998).

1.3.3 Accessory genes

Vif

The function of Vif (Viral Infectivity Factor) remained unclear until a number of years ago. Vif was found to interact with an element of the innate immune system of the host cell, APOBEC3G, which is found in peripheral blood lymphocytes and macrophages (Navarro and Landau, 2004; Sheehy et al., 2002). The antiviral activity of APOBEC3G is due

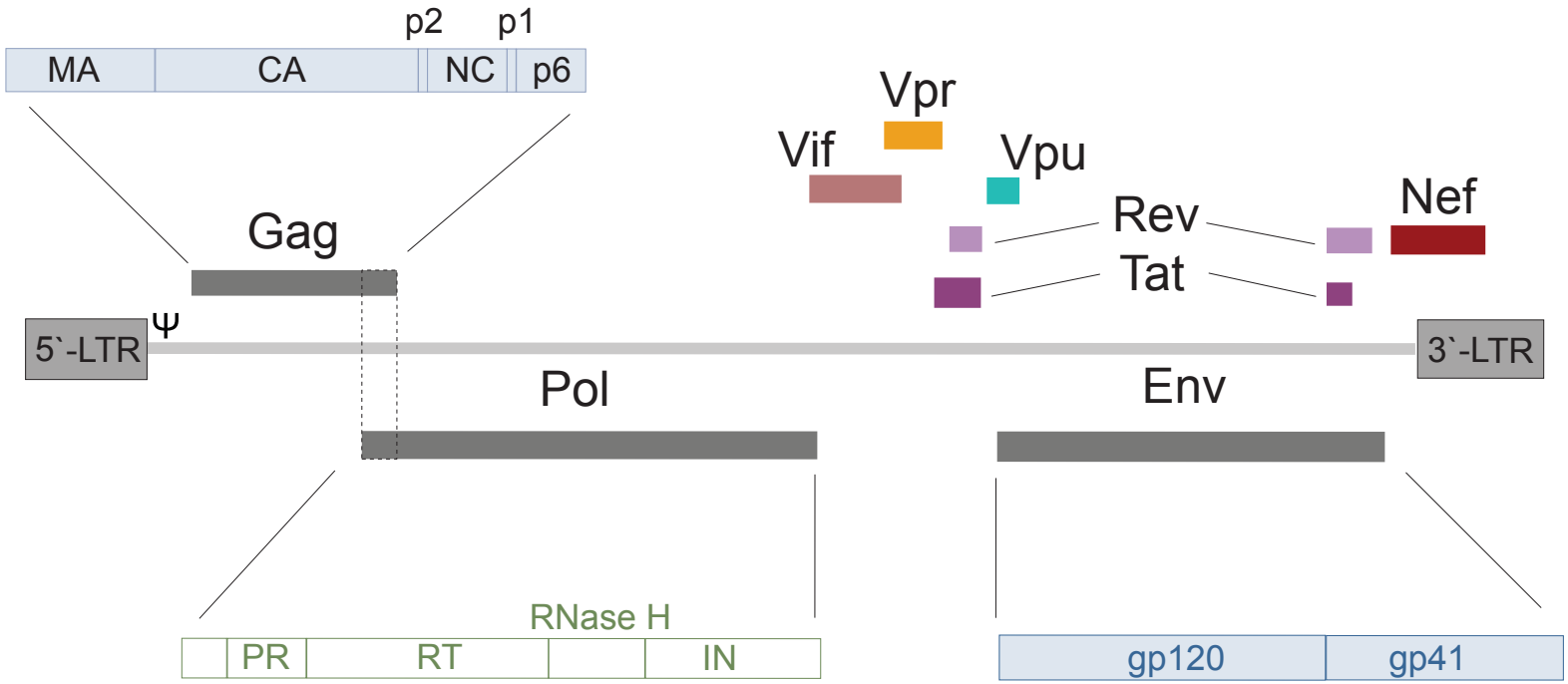


Figure 1.2 Genetic organization of HIV-1. HIV-1 contains nine genes, three structural genes, two regulatory genes and four accessory genes. The location of each gene in the viral mRNA is shown here with proteins created for each gene. The open reading frameshift region associated with the creation of Gag or Pol is marked with a dashed outline. Abbreviations: LTS = Long Terminal Repeat; MA = Matrix; CA = Capsid; NC = Nucleocapsid; = PR = Protease; RT = Reverse Transcriptase; IN = Integrase. (Source Dr Conor Meehan).

primarily to its deamination of the viral cDNA cytidines to uridines resulting in mutations within the virus, which inhibit viral reverse transcription and integration (Chiu and Greene, 2008; Harris and Liddament, 2004). Vif interacts with APOBEC3G reducing its ability to prevent successful HIV-1 replication.

Vpr

Vpr is highly conserved among HIV-1 and HIV-2 and SIV. Approximately 100 copies of the protein are incorporated into the HIV viron indicating its importance in the viral life cycle (Sato et al., 1990). HIV-1 Vpr enhances the ability of HIV-1 to replicate in terminally differentiated macrophages by active nuclear import of the viral PIC (Heinzinger et al., 1994). Vpr also causes G2 cell cycle arrest in infected cells (Jowett et al., 1995; Rogel, Wu, and Emerman, 1995), which allows for a two to three fold enhancement of virus production (Goh et al., 1998). Expression of the Vpr protein has also been linked to cell death in the infected cell (Casey, Wen, and de Noronha, 2010). It has also been suggested that apoptosis in uninfected bystander cells is also a result of Vpr expression (Romani and Engelbrecht, 2009). However, it is unclear whether cell death is an advantageous function of Vpr for the virus or merely a consequence of another process; for example does holding cells in G2 cell arrest for extended periods contribute to cell death? (Casey, Wen, and de Noronha, 2010). Vpr has also been shown to affect not only virus replication, but also cellular gene expression, proliferation and differentiation and it is thought that as Vpr circulates in the blood it may also affect the gene expression of non-infected cells (Balasubramanyam et al., 2007; Xiao et al., 2008)

Vpu

Vpu is a small integral trans membrane protein, which is found exclusively in HIV-1 (Strebel et al., 1989; Strebel, Klimkait, and Martin, 1988). Vpu is expressed from the same mRNA as *env* but is translated at much lower rates than *env* due to an inefficient translation initiation codon (Sato et al., 1990). The protein is involved in the degradation of CD4 at the endoplasmic reticulum. This prevents the formation of stable complexes between cellular CD4 and the HIV-1 Env protein, which essentially traps the Env protein and prevents it from being incorporated into a new virion (Willey et al., 1992). An important clue to the function of Vpu was the discovery of the host restriction factor Tetherin. Tetherin (BST-2, CD317) is

an interferon inducible protein that restricts the release of enveloped viruses by tethering the newly budded virions to the plasma membrane and causing their endocytosis back into the cell for destruction (Neil, Zang, and Bieniasz, 2008). Vpu has been shown to remove Tetherin from the cell surface allowing for efficient release of viral particles (Skasko et al., 2011; Van Damme et al., 2008). It is also thought that Vpu interacts with the host cells MHC decreasing its presentation on the cell surface in order to avoid the host's immune system(Hussain et al., 2008).

Nef

Nef is the first detectable viral protein following HIV-1 infection (Kim et al., 1989) and plays a role in several aspects of the HIV-1 lifecycle. Firstly, Nef is known to reduce the level of CD4 on the infected cell surface by binding the cytoplasmic tail of the CD4 molecule and trapping it in intracellular vesicles (daSilva et al., 2009; Garcia, Alfano, and Miller, 1993; Schwartz et al., 1995). Nef is also involved in the down regulation of cell surface expression of MHC-1 which decreases the ability of T-cells to kill infected cells (Schwartz et al., 1996). However, the mechanism by which Nef achieves this is still not understood (Foster et al., 2011). Interactions between Nef and host kinases such as PAK2 (Arora et al., 2000) alter the cell-signalling pathways through up-regulation of the kinase activity resulting in arrested cell growth and increased viral infectivity (Briggs et al., 1997). The presence of Nef is also known to enhance viral infection by acting within the infected cell to alter the virions prior to release in such a manner that subsequent infection of a new cell is more efficient (Goldsmith et al., 1995; Pizzato et al., 2007).

A number of studies have looked at the correlation between mutations within accessory genes and long-term non-progression. Perhaps the most well known is the attenuation of a virus due to a large deletion in *nef* that was found in a blood donor in Australia who subsequently infected eight others through blood transfusions. Six of these individuals went on to be LTNPs (three died due to causes unrelated to HIV-1 infection) (Learmont et al., 1992; Learmont et al., 1999). Since then a number of other studies have reported deletions in *nef* HIV-1, which are thought to be associated with non-progression (Kirchhoff et al., 1995; Rhodes et al., 2000; Salvi et al., 1998). Other mutations within *vif* and *vpr* have also been

linked to non-progression (Mologni et al., 2006; Wang et al., 2003; Yamada and Iwamoto, 2000).

1.4 The pathogenesis of HIV-1

Sexual transmission of HIV-1 is associated with large amounts of virus in the genital fluids (Phillips and Bourinbaiar, 1992). It has been suggested however, that the initial infection by HIV-1 occurs in most patients by a single virus (Keele et al., 2008). Large amounts of virus in the genital fluids are most commonly associated with advanced disease progression and during acute infection when the risk of transmission can be increased over 20 fold compared to the asymptomatic period (Hollingsworth, Anderson, and Fraser, 2008). HIV-1 infection is characterized by three stages: (i) primary infection phase, which is associated with a massive increase in viral load and a drop in $CD4^+$ T cell count followed by a decrease to a viral load set-point after the initiation of an immune response; (ii) the asymptomatic or chronic phase, which is associated with a gradual increase in viral load from primary infection set-point concurrent with a gradual, but irreversible decrease in $CD4^+$ T cell numbers and (iii) the symptomatic phase of HIV-1, AIDS, associated with the terminal failure of the immune system (Pierson, McArthur, and Siliciano, 2000).

During primary infection the gut-associated lymphoid tissue is profoundly depleted of $CD4^+$ T cells as 60 % of total $CD4^+$ T cells reside there (Cheroutre and Madakamutil, 2004). During this phase the level of $CD4^+$ T cell depletion is inversely correlated with the increase in the observed number of $CD8^+$ T cells, which includes HIV-1 specific $CD8^+$ T cells (Smit-McBride et al., 1998). Levels of viremia within the blood start to decline, coinciding with the development to an immune response to HIV-1 and the levels of $CD4^+$ T cells begin to recover (Figure 1.3). Despite a widespread immune response to HIV-1 the virus is not completely cleared and ongoing replication can readily be detected during the asymptomatic phase (Pierson, McArthur, and Siliciano, 2000). During this time $CD4^+$ T cells continue to be depleted for a number of reasons. Firstly, because the cells are destroyed by the HIV-1 infection and secondly, because the production of new $CD4^+$ T cells is impaired. $CD4^+$ T cells are destroyed due to HIV-1 mediated apoptosis, disruption of the cell membrane due to the HIV-1 life cycle, the accumulation of un-integrated viral DNA, destruction of infected

cells by the immune system and incorporation of cells into syncytia by neighbouring infected cells. The production of $CD4^+$ T cells is impaired by infection of progenitor cells, destruction of supporting stromal networks, cytokine dysfunction, HIV-1 induced apoptosis, opportunistic infections of the bone marrow, and suppressed thymic functions (Review by (McCune, 2001). The final phase of infection is characterised by the emergence of clinical immunodeficiency. In the year or two before AIDS develops, there is often a more rapid decline in $CD4^+$ T cells. This decline may be preceded by an increase in viral load with viral replication occurring in many sites of the body in addition to occurring in the lymphoid tissue (Connor et al., 1993; Reinhart et al., 1997). Opportunistic infections begin to occur when $CD4^+$ T cells drop below 200 cells/μl and the already badly damaged immune system is unable to fight the infections off and death in inevitable.

The length of time from infection to the onset of AIDS is approximately 10 years in normal progressors, however drug therapy can increase this time dramatically (Langford, Ananworanich, and Cooper, 2007). In a minority of individuals (long-term non-progressors), gut-associated lymphoid tissue $CD4^+$ T cell depletion is not visible and the $CD4^+$ T cell numbers are close to normal (Guadalupe et al., 2003) (Figure 1.3). These individuals can remain asymptomatic for much longer than 10 years without drug therapy. This could reflect the ability of the host anti viral immune response to control viral infection and to maintain the integrity of the immune system preserving $CD4^+$ T cell maturation and their role in co-ordinating HIV-1 specific immune responses (Sankaran et al., 2005). It has also been suggested that these individuals can be infected with an attenuated viral strain resulting in reduced viral fitness and non-progression (Reviewed in (Poropatich and Sullivan, 2011). Other individuals show rapid disease progression and can show signs of AIDS as early as three months following infection (Pantaleo and Fauci, 1996).

1.5 The origins of HIV

Lentiviruses cause chronic persistent infections in various mammalian species including bovines, horses, sheep, felines and primates (Sharp and Hahn, 2010). Ever since HIV was first discovered, the reasons for its sudden emergence, pandemic spread and unique pathogenicity have been the subject of intense study. Improved molecular techniques and

.

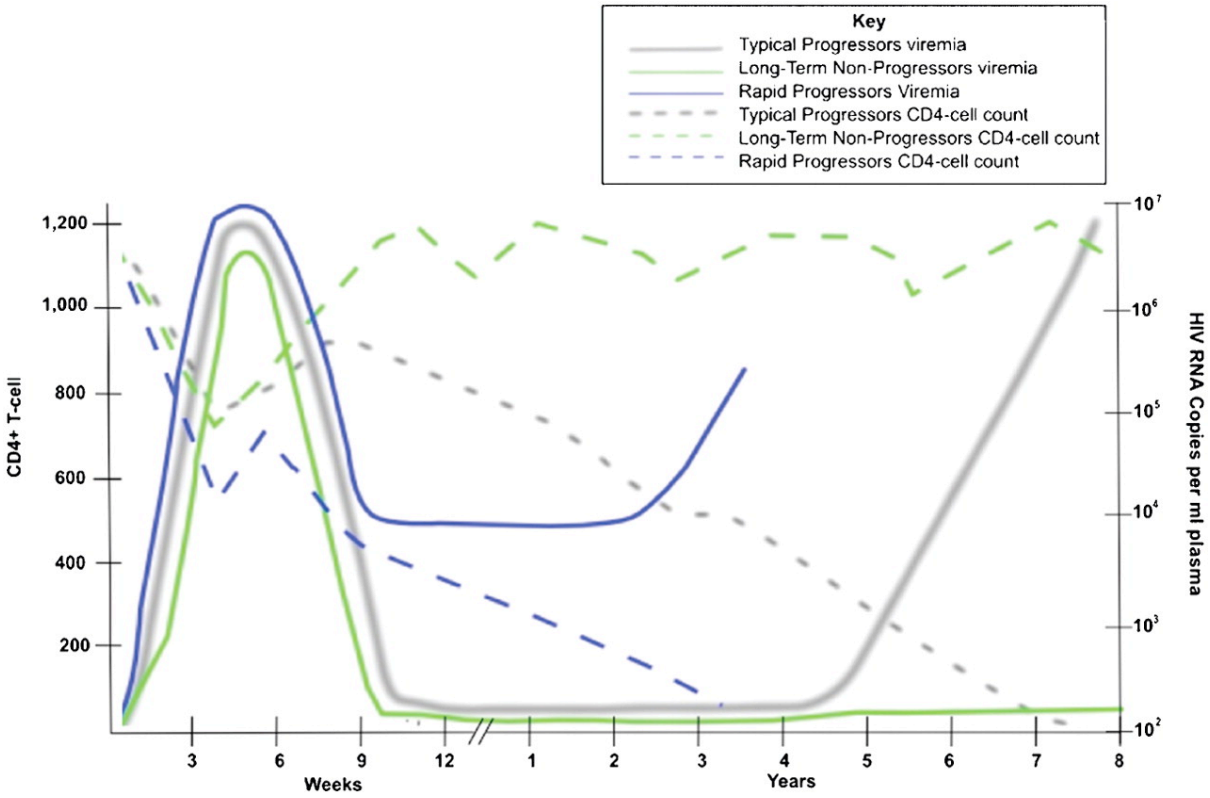


Figure 1.3 Disease progression in HIV-1-infected typical progressors, LTNPs and rapid progressors according to $CD4^+$ T-cell counts and viremia. Length of infection is on the X-axis while $CD4^+$ T cell counts and viral count are on the Y-axis (Poropatich and Sullivan, 2011).

phylogenetic analysis have been able to answer a number of unanswered questions such as, where did HIV come from? Two distinct retroviruses, HIV-1 and HIV-2, are known to cause AIDS. Together with related SIVs (Simian Immunodeficiency Viruses) found in 20 other African primate species, these comprise the primate lentiviruses (Rambaut et al., 2004) (Figure 1.4) . Interestingly (other than laboratory associated infections of Asian macaque monkeys) SIV infections are thought not to cause disease in their hosts, although only a few studies of their natural history in wild populations has been undertaken (Rambaut et al., 2004).

1.5.1 HIV-1

HIV first arose in humans through zoonotic transmission from SIV infected primates (Sharp et al., 2001). Phylogenetic analysis showed close simian relatives of HIV-1 and HIV-2 were found in chimpanzees (Huet et al., 1990) and sooty mangabeys (Hirsch et al., 1989) respectively (Figure 1.4). The first isolates of SIVcpz (Chimpanzee SIV) were all derived from animals housed in primate centres or sanctuaries, although infection was very rare in these populations (Sharp, Shaw, and Hahn, 2005). This raised doubts about whether chimpanzees represented a true SIV reservoir. The endangered status of these animals made in depth studies of SIVcpz infection in the wild difficult. Non-invasive diagnostic methods were developed that detected SIVcpz specific antibodies and nucleic acids in chimpanzee faecal and urine samples with high specificity and sensitivity (Keele et al., 2006; Santiago et al., 2003). These technical innovations permitted a comprehensive analysis of chimpanzee populations throughout central Africa. These studies identified the common chimpanzee as a natural SIVcpz reservoir. Chimpanzees can be divided into four subspecies using mitochondrial DNA sequence diversity (Gagneux et al., 1999). Only two of the four subspecies, located in central and east Africa, harboured SIVcpz, and within those two subspecies prevalence varied widely from 50 % in some communities to rare or absent in others. SIVcpz was not found in the two subspecies found in the remaining habitat of chimpanzees (mainly west Africa) (Keele et al., 2006; Santiago et al., 2003). The reason therefore for the scarcity of SIV in captive chimpanzees was that most were imported from west Africa and were of a subspecies that was shown not to harbour SIVcpz (Prince et al., 2002; Switzer et al., 2005). Initially it was thought SIVcpz did not cause disease in chimpanzees, however recent evidence suggests the contrary. It is now thought that SIVcpz

has a substantial negative impact on the health, reproduction and life span of all chimpanzees that harbour SIV in the wild (Etienne et al., 2011; Keele et al., 2009; Soto et al.).

Molecular analysis of the faecal and urine samples above, also discovered a new lineage of SIV found in Gorillas (SIVgor) (Van Heuverswyn et al., 2006). This lineage clusters with SIV strains from chimpanzees (Figure 1.4) and it is thought the SIVgor was the result of a single chimpanzee to gorilla transmission that occurred at least 100 to 200 years ago (Takehisa et al., 2009). SIVgor has only been found in a few gorilla communities and within these, prevalence was low (approximately 5 %) (Neel et al., 2010). Presently, it is unclear whether SIVgor is pathogenic to its host or not.

1.5.2 HIV-2

The origin of HIV-2 is a much simpler story. HIV-2 was thought to have originated as a result of cross species transmission between humans and sooty mangabeys (SIVsmm), a hypothesis confirmed by phylogenetic analysis (Gao et al., 1992)(Figure 1.4). SIVsmm is highly prevalent both in animals in captivity and in the wild, and is non-pathogenic to its host (Silvestri, 2005). There are at least eight different lineages (A-H) of HIV-2, each of which appears to be the result of independent cross species transmissions (Sharp and Hahn, 2010). HIV-2 is largely restricted to west Africa (de Silva, Cotten, and Rowland-Jones, 2008). The prevalence rates are currently in decline and in most West African countries HIV-2 has increasingly been replaced by HIV-1 (Hamel et al., 2007). Viral loads tend to be lower in HIV-2 than in HIV-1, and most individuals infected with HIV-2 do not progress to AIDS. Those individuals who do progresses to AIDS, display clinical symptoms that are indistinguishable from symptoms associated with AIDS caused by HIV-1 (Rowland-Jones and Whittle, 2007).

1.6 HIV-1 subtypes

HIV-1 is made up of four distinct lineages, M, N, O and P, each of which resulted from an independent cross species transmission. Group M was the first to be discovered and represents the pandemic form of HIV-1. Group O was discovered later and is much less prevalent than group M (De Leys et al., 1990). It represents less than 1 % of global HIV-1

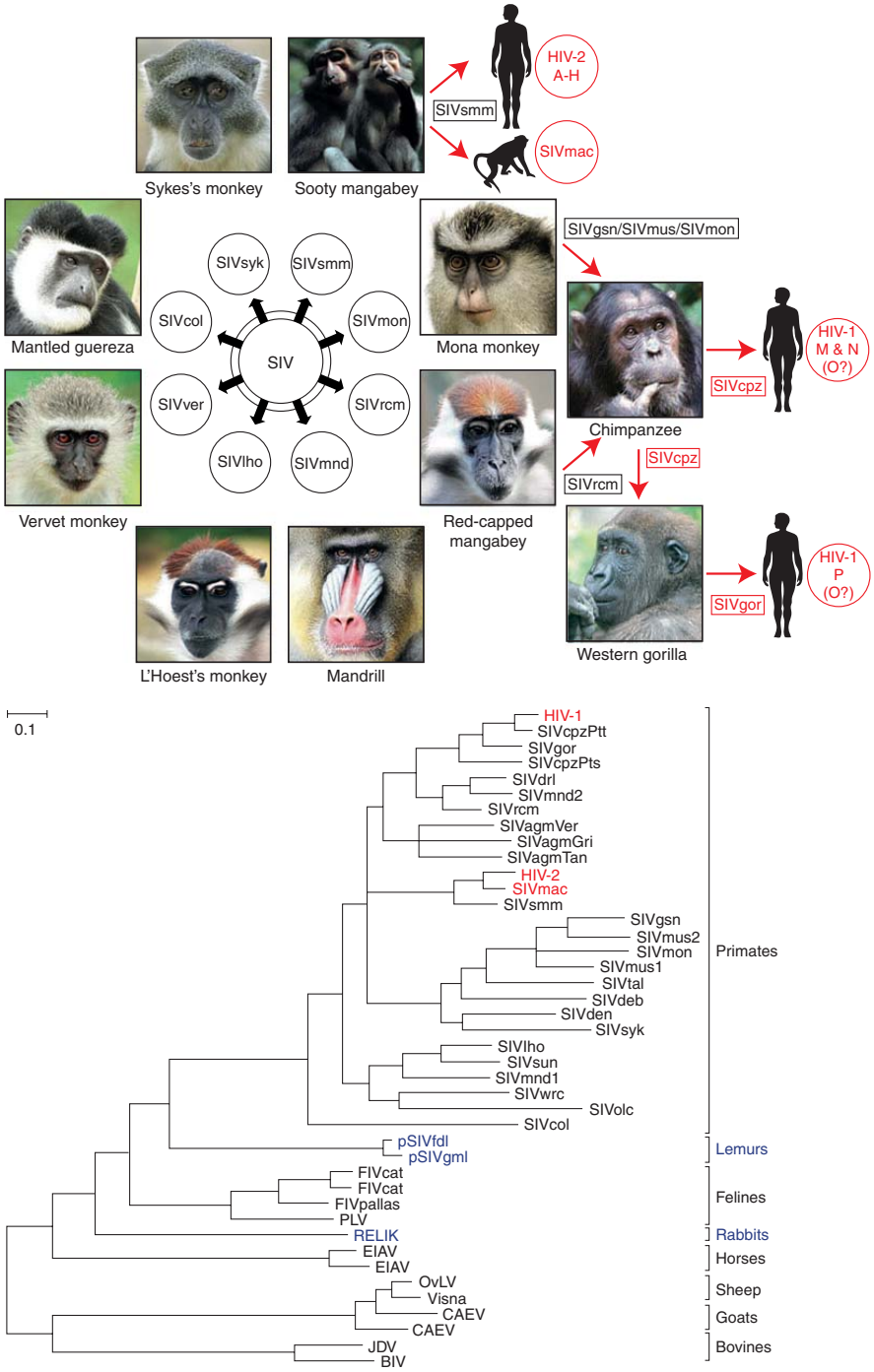


Figure 1.4 Origins of HIV. Old world monkeys are naturally infected with SIVs with a suffix to donate their primate species of origin (e.g. SIVsmm from sooty mangabeys). Phylogenetic reconstruction of the evolutionary relationships among *pol* sequences show HIV-1 to be more closely related to SIVcpz and HIV-2 to be more closely related to SIVsmm (Sharp et al., 2010).

infections and is largely found in Cameroon, Gabon and neighbouring countries (Mauclere et al., 1997; Peeters et al., 1997). Group N was discovered in 1998 and is even less prevalent than group O with only 13 cases in Cameroon documented to date (Vallari et al., 2010a). Finally, group P was recently discovered in a Cameroonian woman living in France (Plantier et al., 2009) and has been found only in one other individual also from Cameroon (Vallari et al., 2010b).

Phylogenetic and statistical analyses have dated the last common ancestor of HIV-1 group M to around 1910 - 1930 with narrow confidence levels (Korber et al., 2000; Worobey et al., 2008). It is thought that the pandemic originated in the Democratic Republic of Congo. One line of evidence for this comes from samples collected in Kinshasa between 1959-1960. Proviral DNA extraction from these samples and PCR amplification of small fragments of *gag* and *pol* and *env* revealed that HIV-1 had already diversified into different subtypes by that time (Worobey et al., 2008). Group M has been divided into nine different subtypes (A-D, F-H, J, K) (Robertson et al., 2000), which are approximately equidistant from one another. The average inter-subtype genetic variability is 15 % for *gag* and 25 % for *env*, however this difference between subtypes has been shown to be as high as 35 % for *env* (Gaschen et al., 2002). Within a number of subtypes, it is possible to identify groups of viral isolates from genetically related sister clades (sub subtypes) which appear to be phylogenetically more closely related to each other than to other subtypes. This is seen in the case of subtype A and F, which have both been sub divided into A1, A2, F1 and F2 respectively (Gao et al., 2001; Triques et al., 2000). Subtypes B and D are more closely related to each other than to other subtypes and clade D is considered to be an early clade B African variant. Their original designation as subtypes is retained by authors in order to keep consistency with earlier published works (Gao et al., 1996; Louwagie et al., 1993).

Sequencing of multiple genomic regions from the same viral isolate in conjunction with improved methods of classification has identified inter-subtype recombinant forms of HIV-1. These inter-subtype recombinants have originated in individuals multiply infected with viruses of two or more subtypes (Robertson et al., 1995). Full genome sequencing has further identified isolates with the same pattern of recombination. Currently 49 Circulating Recombinant Forms (CRFs) have been identified (www.hiv.lanl.gov) from the sequencing of

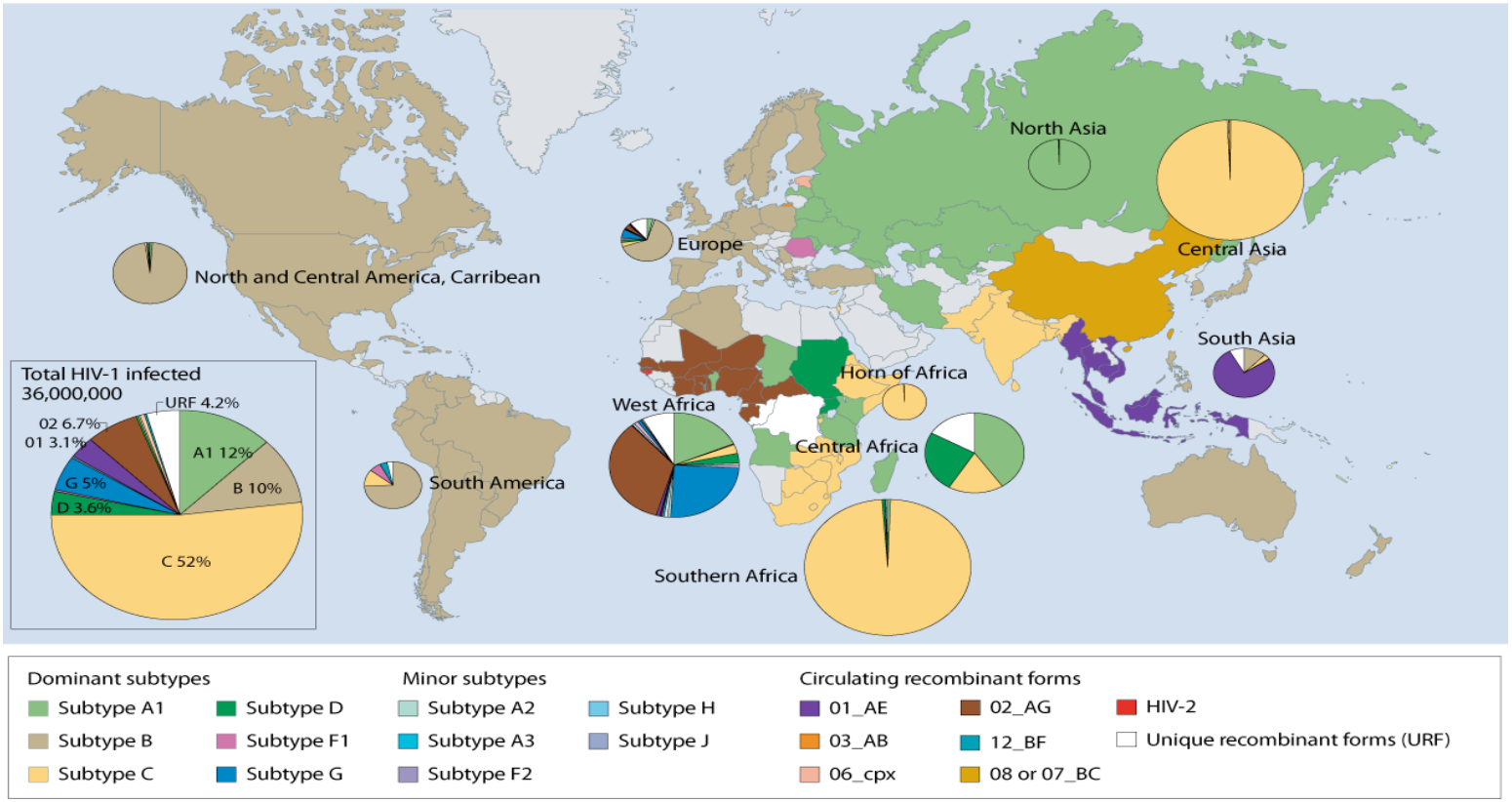


Figure 1.5 The Global subtype Distribution for HIV (Arien, Vanham, and Arts, 2007).

three or more genomes from epidemiologically unlinked individuals with identical patterns of recombination.

The classification of HIV-1 strains has helped in tracking the course of the HIV-1 pandemic. Subtype C infections represent approximately 50 % of infections world wide (Buonaguro, Tornesello, and Buonaguro, 2007) and is the dominant viral form in Ethiopia and all countries in southern Africa including Malawi, and India (Figure 1.5). The prevalence of subtype C is also increasing in Brazil and China (Rodenburg et al., 2001; Soares et al., 2005). Subtype A is predominant in areas of central and eastern Africa, and CRF02_AG is found in western Africa. Subtype B is the main genetic form in western and central Europe, the Americas and Australia. It is also common in several countries of South East Asia, northern Africa and the Middle East and among South African and Russian homosexual men (Buonaguro, Tornesello, and Buonaguro, 2007). Because of its prevalence in developed countries of the world, scientific research has focused on subtype B. This has lead to critical gaps in our knowledge of non-subtype B variants, including subtype C, which is not only the most prevalent subtype worldwide but is found in some of the poorest regions of the world where the HIV-1 epidemic adds to the burden of poverty and high mortality. Studies on the molecular epidemiology of subtype C are essential for future vaccine and therapy strategies.

1.7 HIV in Africa

Experts studying the spread of the epidemic suggest that HIV-1 was present and circulating in the population of the Democratic Republic of Congo as early as the 1960s (Vidal et al., 2000; Worobey et al., 2008). Stored blood samples from an American malaria research project carried out in the Congo in 1959 prove one such example of early HIV-1 infection (Motulsky, Vandepitte, and Fraser, 1966; Nahmias et al., 1986). The first epidemic of HIV/AIDS was thought to have occurred in Kinshasa in the 1970s where in the capital a surge in opportunistic infections such as crytococcal meningitis, Kaposi's sarcoma, and tuberculosis and specific forms of pneumonia was seen. The description of 'slim disease' was coined to describe the severe wasting of the patient (Denolf et al., 2001). From here, HIV-1 spread into eastern Africa in the 1970s but it did not reach epidemic levels until the 1980s (Serwadda et al., 1985). Within east Africa it is thought that sex workers played a part in the accelerated transmission rate, for example, in Nairobi 85 % of sex workers were HIV-1

positive in 1986 (Piot et al., 1987). Truck drivers and migrants such as soldiers, traders and miners, facilitated the initial rapid spread as they engaged with sex workers and spread HIV-1 via the transport routes (Iliffe, 2006). For example, one study identified 35 % of Ugandan truck drivers as HIV-1 positive in the 1980s (Carswell, Lloyd, and Howells, 1989). Also contributing to the spread of HIV-1 was the high ratio of men in urban populations, the low status of women, lack of circumcision and the prevalence of sexually transmitted diseases (Iliffe, 2006). Uganda in particular was affected by high prevalence levels and by the end of the 1980s this had peaked at over 30 % prevalence among pregnant women attending antenatal care clinics (Stoneburner and Low-Beer, 2004). Later in the 1980s the epidemic progressed and moved south, and by the end of the 1980s Malawi, Zambia, Zimbabwe and Botswana were on the verge of overtaking east Africa as the focus of the global HIV-1 epidemic (Buve, Bishikwabo-Nsarhaza, and Mutangadura, 2002) (Figure 1.6). By 2001, of the 40 million people living with HIV-1, 70 % were from sub-Saharan Africa (Buve, Bishikwabo-Nsarhaza, and Mutangadura, 2002).

In 2002 Botswana became first African country to launch a national anti retroviral treatment project and by 2007 approximately 95 % of HIV-1 positive people in the country, who required treatment, were being treated with ART (UNAIDS/WHO, 2007). While the ART program in Botswana is thought to be the most successful, other countries are not far behind. In 2006 Namibia was treating 71 % of those in need of ART, Rwanda 72 %, Malawi 43 % and Uganda 41 % (UNAIDS/WHO, 2007). Recently however, WHO revised the previous recommendation that individuals with a $CD4^+$ T cell count of 350 cells/mm^3 rather than 200 cells/mm^3 should be placed on ART suggesting many people in need of ART are yet to be treated.

In 2009 in sub Saharan Africa, 22.5 million people were living with AIDS (UNAIDS, 2009a). Despite the overwhelming numbers of people infected with HIV-1 the overall growth of the global epidemic appears to have stabilized. The annual number of HIV-1 infections has been steadily declining since the 1990s and there are fewer AIDS related deaths due to the significant scale up of anti retroviral drug therapy over the past few years (UNAIDS, 2009b). In sub Saharan Africa the number of newly infected individuals fell from 2.2 million in 2001 to 1.8 million in 2009 (UNAIDS, 2009a). Despite this fall in the number

of new infections, the levels of infection are still high and HIV-1 is still one of the most significant causes of mortality and morbidity in sub Saharan Africa. In South Africa alone 5.6 million people are living with HIV-1. Swaziland in southern Africa has the world's highest prevalence levels with 25.9 % of the adult population infected with HIV-1 (UNAIDS, 2009a). Within Malawi the prevalence also remains high with 11 % of adults infected with HIV-1 (UNAIDS, 2010) (Figure 1.6). Since the start of the epidemic, more than 15 million African people had died of AIDS (UNAIDS, 2009b). It is almost impossible to comprehend the impact the epidemic has had on families, communities, work places, and national and regional development and with no cure or vaccine expected in the near future, HIV-1 will continue to have devastating effects in Africa.

1.8 HIV-1 in Malawi

The first epidemiologically identified case of HIV-1 in Malawi comes from a sample collected in 1982 (Glynn et al., 2001). After a period of rapid increase in prevalence in the 1980s and 1990s, HIV-1 prevalence has stabilised at around 11 % in Malawi. This had been observed in both national estimates and in epidemiological studies (Crampin et al., 2003; UNAIDS, 2010; UNAIDS, 2007; White et al., 2007). Within Malawi HIV-1 prevalence varies by age, gender and other socio-economic characteristics. According to the 2004 demographic and health survey, prevalence in the 15-49 age-group, is higher among women (13.3 %) than in men (10.2 %) and higher in urban (17 %) than in rural areas (10.8 %).

The difference between males and females is particularly seen in adolescents (15 -19 year olds) where 3.7 % of females are infected with HIV-1 when compared to 0.4 % of their male counterparts (UNGASS, 2010). Prevalence is also higher in the south of the country than in the north (Bello, Chipeta, and Aberle-Grasse, 2006). The relatively high countrywide prevalence of 12 % is thought to be related to multiple and concurrent sexual partnerships, discordance in long-term couples where protection is not being used, low prevalence of male circumcision, low and inconsistent condom use, sub-optimal implementation of HIV-1 prevention interventions, late initiation of HIV-1 treatment and TB/HIV co infection (UNGASS, 2010). In 2009, a total of 840, 156 adults and 111, 510 children were estimated to be living with HIV-1 in Malawi.

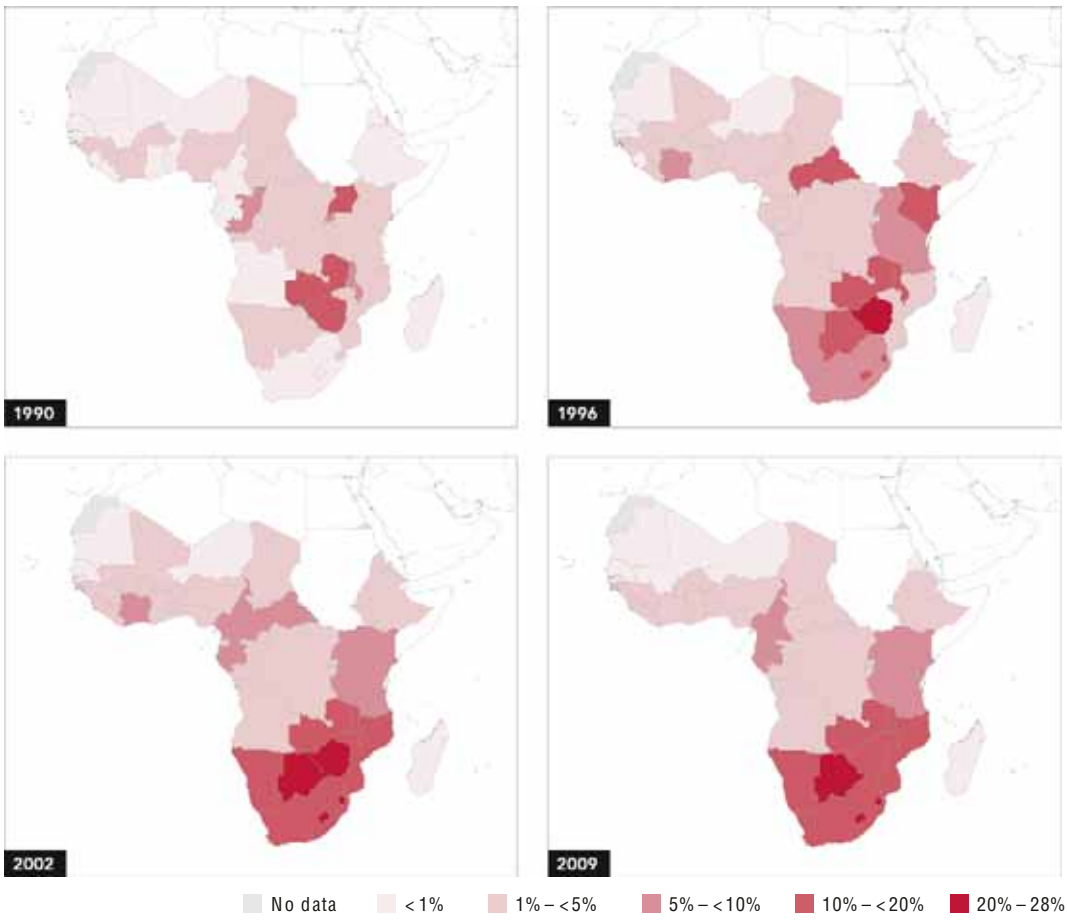


Figure 1.6 The prevalence of HIV-1 infections in 1990, 1996, 2002 and finally in 2009. (Source (UNAIDS, 2009a))

An ART rollout program, under a Malawian national scheme, was started in 2004. The Malawi national ART programme follows a public health model focusing on 'service delivery to all who need it' (Harries et al., 2008). Individuals are eligible for ART in Malawi if, upon physical examination and history, during a screening clinic visit, they are assessed to be in stage III/IV or stage II with a CD4 count < 250 cell/mm^3. In 2006 the estimated number of individuals needing ART in Malawi was 190, 000 and RT coverage was estimated to be 43 % (UNAIDS/WHO, 2007).

Karonga is a rural region located in the north of Malawi (Figure 1.7). It is bordered by Tanzania to the north, Zambia in the west and Lake Malawi in the east. The majority of individuals are subsistence farmers and fishermen. The Karonga Prevention Study (KPS) was established in Northern Malawi in the late 1970s for population based epidemiological studies of leprosy. As part of this study, a house to house total population survey was conducted in 1981-984 and 1986-1989 and 44 150 dried blood spot (DBS) samples were collected to examine the prevalence incidence and risk factors for TB and Leprosy (Ponnighaus et al., 1993; Ponninghaus et al., 1987). These samples were later tested for the presence of HIV-1 antibodies. The prevalence of HIV-1 was 0.1 % between 1981-1984 and was 2 % between 1987-1989 (Glynn et al., 2001). To measure the possible effect of HIV-1 on survival in rural Africa, and to aid planning and counselling, a retrospective cohort study was carried out in the late 1990s. Individuals identified as HIV-1 positive in the first study, their spouses and offspring, were traced as well as some of the individuals who were HIV-1 negative during the 1980s survey (Crampin et al., 2002). At this time point the prevalence had risen to approximately 10-15 % (Glynn et al., 2001). DNA was later extracted from these samples and regions of *env* and *gag* were PCR amplified. Between 1982-1984 subtype C viruses represented 55 % of infections. By 1987-1989 the prevalence of subtype C infections had increased to 90 % (McCormack et al., 2002). Thirty-eight individuals who had been seen in the 1980s survey were found to be still alive in the 1990s. These same individuals were sought again in 2004 at which time 17 were still alive and were characterised as long-term survivors (LTS) (McCormack et al., 2006).

Figure 1.7 Map of Malawi. Karonga is a rural region found in the north of the country on west shore of Lake Malawi. Malawi borders Tanzania to the north, Zambia is to the west and Mozambique is to the south. A national highway runs through Karonga from Tanzania.

As part of the Anti Retroviral Therapy (ART) rollout scheme the first ART clinic opened mid 2005 in Karonga District (White et al., 2007). To facilitate a rapid scale up in a health care system where even basic laboratory services are often not available, such as in the Karonga District Hospital, ART eligibility is primarily determined by clinical staging (Libamba et al., 2005; McGrath et al., 2007). A recent study has suggested that focusing on clinical staging alone failed to identify two thirds of those in Karonga who would have potentially benefited from receiving ART treatment (McGrath et al., 2007). Improvements in ART delivery are clearly needed and the most recent study in KPS focused on the effect of ART on the HIV-1 epidemic in Karonga. The study aimed to measure the impact of increasing access to ART on the spread, socio-demographic and health consequences of the HIV-1 epidemic in Karonga, quantifying the benefit and defining the limitations of ART roll out. These findings would subsequently be used to support the development of solutions and to improve outcomes. KPS is also involved in studying strains of TB present in Karonga, and the affect of HIV-1 infection on TB. The transmission of pneumococci to infants in HIV-1 affected and unaffected households and the impact of lower acute respiratory tract disease on young children together with studies on eliminating lymphatic filariasis infections are also carried out by KPS. Clinical services in Karonga district, predominantly diagnosis and care of TB patients, support of ART services, provision of paediatric care and provision or radiography services are also supported by KPS.

The studies carried out by KPS have provided an expansive dataset mapping the HIV-1 epidemic in Karonga from 1982, when the first HIV-1 positive sample was collected, to the present day. Through this dataset, the distribution of subtypes and molecular characterises of the viral population in Karonga have been explored. Subtype C viruses represented 90 % of HIV-1 infections in Karonga in the late 1980s having increased from 55 % in the early 1980s (McCormack et al., 2002). Using samples collected from Karonga in the 1990s and 2000s I hoped to answer the question of whether the subtype distribution had changed within Karonga? What other subtypes or recombinant subtypes were also present with subtype C and, how had the molecular evolution of the epidemic within Karonga changed since the first survey in the 1980s? (Chapters 4 and 5).

These studies also identified a number of particular cohorts in Karonga; in particular a cohort of long-term survivors was apparent. Chapters 2 and 3 investigate the possible viral elements associated with non-progression in the 17 LTS still alive in 2004. A three amino acid deletion in *gag,* found in sequences from the 1990s but not the 1980s in some of the LTS and two thirds of the HIV-1 positive population in Karonga, had been previously thought to be associated with non-progression and increased onward transmission (McCormack et al., 2006). With this work I hoped to answer a number of questions related to non-progression in Karonga. Firstly, was the three amino acid deletion in *gag* present as a minor variant in the 1980s and simply not detected via the methods employed at the time? Does the three amino acid deletion have an effect on the replicative fitness of the virus that subsequently results in non-progression in those infected with this viral variant? Are there any other mutations in *env* and *gag* in the LTS that may be related to their survival? What viral elements, such as diversity and co receptor tropism have changed during the twenty years that the LTS in Karonga have been followed?

Chapter 2: Mutational Patterns and Replicative Fitness in HIV-1 Subtype C Infected Long-Term Survivors in Karonga District Malawi

2 Chapter 2

2.1 Introduction

The natural history of HIV-1 infection and disease progression has been well established in adults with 10 years being the average time from the initial infection to the development of AIDS in normal progressors (Crampin et al., 2002; Langford, Ananworanich, and Cooper, 2007). Rapid progressors (RP) are those individuals who develop AIDS within three to five years after infection, and in some cases in less than 6 months after contracting the HIV-1 virus (Pantaleo and Fauci, 1996). Other individuals progress to AIDS over a much longer period of time. Long-term non-progressors (LTNPs) are those individuals characterised as having remained asymptomatic for >10 years, without antiretroviral therapy and who maintain an elevated $CD4^+$ count (>500 cells μl^{-1}). Long-Term Survivors (LTS) are those individuals who remain asymptomatic for over 10 years, without antiretroviral therapy, but show a steady decline in the number of circulating $CD4^+$ cells (<500 cells μl^{-1})(Learmont et al., 1992). A third group known as elite suppressors (ES) maintain almost undetectable levels of viremia (>50 copies of HIV-1 RNA per millilitre of blood) while also maintaining a normal $CD4^+$ count, without antiretroviral therapy (Walker, 2007).

HIV-1 has a highly diverse viral population, both within an infected individual and across populations of hosts. This diversity derives from two sources: (i) the rapidly evolving nature of the virus, which continually generates new mutant forms due to mutation caused by the error prone nature of HIV-1 reverse transcriptase and the absence of a 3'-exonuclease proofreading activity, its propensity to recombine and its high rate of replication, and (ii) the active immune response that promotes diversifying selection (Choisy et al., 2004; Wolinsky et al., 1996; Yang, Bielawski, and Yang, 2003). Any region of the viral genome may have many roles. Individual component proteins within the virus are required to perform regulatory, functional and structural roles in the course of its life cycle. Both the structure and the function of the protein, therefore restricts the amount of change that can occur within the viral genome (Woo, Robertson, and Lovell, 2010). Certain mutations may lead to an in-active virus or a virus with reduced fitness. Studies focused on the HIV-1 viruses found in non-progressing patients have shown alterations in one or more of the genes, which seem to correlate with disease progression.

One of the first studies to associate gross genetic defects with long-term survivorship described a large deletion in *nef*/LTR that was transmitted from an asymptomatic blood donor to multiple recipients. The donor and all of the recipients became LTNPs (Churchill et al., 2004). Since then a number of studies have associated deletions, insertions and polymorphisms in nearly all gene regions of the HIV-1 genome (*env, gag, nef, vpu, vif, rev* and *tat*) with viral attenuation and non-progression. These defects do not result in a non-functioning virus but can have a wide range of repercussions on the viral life cycle such as affecting the transmission ability of the virus, the ability to infect a cell or the replicative ability of the virus. A number of different methods have been created to measure the replicative fitness of HIV-1 variants. One such method utilises the naturally occurring homologous recombination (Gap Repair System) seen in yeast, to facilitate recombination between a PCR product and a neutral vector backbone via short sequences of comparable homology in both DNA fragments (Marozsan and Arts, 2003). These newly created chimeric viruses can then be competed in dual infection assays to ascertain the replicative fitness of viruses containing a particular mutation in a gene of interest. This method was recently used in a study by Lassen et al. (2009), which reported that the *env* gene from a number of elite suppressors supported less efficient replication than that seen in typical progressors.

Currently there is very little information for LTS found in sub-Saharan Africa (Laeyendecker et al., 2009). McCormack et al. (2006) described a three amino acid deletion at the end of *gag* p17 found in 15 LTS from Karonga District, Malawi. In each LTS, the deletion was observed in sequences dating from the late 1990s but was not present in any sequences dating from the 1980s. It was also described in two thirds of the other HIV-1 positive individuals from Karonga District included in that study from the late 1990s and it was suggested that the deletion could be associated with longer survival and onwards transmission (McCormack et al., 2006). The p17 matrix protein plays a key role in several steps during viral replication both in early and late stages of the virus life cycle, and is involved in directing precursor poly proteins to the plasma membrane, incorporation of the envelope into the virion and particle assembly. The p17 protein is also required for efficient transport of the pre-integration complex into the nucleus of the cell (Cannon et al., 1997; Fiorentini et al., 2006). It is possible that indels or polymorphisms may disrupt the ability of p17 to function and subsequently have an impact on the replicative fitness of the virus.

In this work I sought to further explore the viral factors involved in long-term survival in Karonga District, Malawi including;

a) the pattern of emergence, and spread of the three amino acid deletion in the LTS and in the Karonga HIV-1 positive population,
b) other mutations in *gag* and *env* that might be associated with survival,
c) the replicative fitness of the infecting viruses in the LTS using the yeast homologous recombination system described by Marozsan and Arts, (2003).

2.2 Materials and methods

2.2.1 Materials

Thirty-eight HIV positive individuals were seen in Karonga District Malawi in both the late 1980s and late 1990s and Dried Bloods Spot (DBS) samples were collected (Crampin et al., 2002). Seventeen of the individuals were still alive in 2004 (McCormack et al., 2006) and ten provided DBS samples. Fourteen of these were sought again in 2010 (three were not sought as they had refused to participate the previous two times) when it was found that three had died and one had left the region. Nine individuals agreed to provide whole blood samples. Of the nine, five had begun antiretroviral therapy (ART) (one in 2005, two in 2006, one in 2008 and one in 2009). Four individuals had not begun ART and have been HIV-1 positive, without treatment for a minimum of 21 years (although one of these has now been referred) (Table 2.1).

Plasma or cell pellets were collected from HIV-1 positive individuals from retrospective studies in the Karonga District. Approximately 100 of these, chosen randomly, were utilized in this study to explore the frequency of the three amino acid deletion in the 2007-2008 time periods. Nine of these, categorised as Normal Progressors, were also used in the replicative fitness assays (Table 2.2). Four HIV-1 subtype C control viruses (where the replicative fitness was already known (Abraha et al., 2009)) propagated on U87 cells were also used in this project (Table 2.2). The cell supernatant was collected previously by lab personal and stored at -80 °C.

2.2.2 Nucleic acid extraction and storage

DNA was extracted from available dried blood spots (DBS) collected from LTS between 1986-1989 and 2004 using the QIAamp DNA Micro Kit (Qiagen) according to the manufacturer's instructions (Table 2.1). Whole blood samples collected from both the LTS and other HIV-1 positive individuals in Karonga (including nine normal progressors (Table 2.2)) were separated into cell pellet and plasma by centrifugation. Proviral DNA was extracted from 200 μl of cell pellet or plasma using the QIAamp DNA Blood Mini Kit (Qiagen).

RNA was extracted in the KPS, Chilumba (Karonga District) from 200 μl of plasma collected from nine LTS in 2010 using the QIAamp Viral RNA Mini Kit (Qiagen). The extracted RNA was stored using the stabilising product, RNAstable (Biomatrica), to allow for transport from the laboratory in Karonga to the laboratory in Galway, Ireland. Once in Galway the RNA samples were re eluted in nuclease free water and stored at -80 °C. They were again placed in the stabilising agent to allow transport to the laboratory of Dr. Eric Arts in Cleveland, USA. On arrival in Cleveland, the RNA samples were re-eluted in 20 μl of nuclease free water and stored at -80 °C. RNA was isolated from the four subtype C HIV-1 control samples listed in Table 2.2 using the MagMAX-96 viral RNA Isolation Kit (Applied Biosystems) according to the manufacturer's instructions.

Ethanol precipitation was used to purify and remove any possible PCR inhibitory compounds introduced by the RNAstable preservation process carried out on the RNA extracted from plasma collected from the nine LTS in 2010. The procedure was carried out on ice and in a nuclease free environment. The RNA samples were first thawed gently on ice, after which one tenth the volume of Ammonium Acetate was added, followed by 3 times the volume of cold 100 % ethanol. The samples were then vortexed and placed at -80 °C for over 2 hours. The rotor of the centrifuge was cooled in the -20 °C freezer during this time. After 2 hours the mixture was centrifuged at full speed for 10 minutes. The supernatant was removed and 500 μl of cold 70 % ethanol was added to wash the pellet. After the sample was spun at full speed for 5 minutes the supernatant was again removed and the wash step was repeated. Following the final wash step the supernatant was removed by pipetting. Any remaining ethanol was removed using a speed vacuum. The dried RNA was then re-eluted in 20 μl of nuclease free water. Five microliters of this was then used in the RT PCR (this amount was later increased to 15 μl) in section 2.2.4.1.

Table 2. 1 A list of the type of sample collected from the LTS in Karonga. The samples are labelled with the LTS number and the year of collection. The type of nucleic acid available is also listed. –ve = had not started ART, +ve = had started ART, D* = had died in 2008, D** = had died in 2005, L = Left the region after 2004.

Samples collected in 2004				Samples collected in 2010		
LTS number	**Type of sample**	**ART**	**Nucleic acid available**	**LTS number**	**Type of sample**	**Nucleic acid available**
LTS1_2004	DBS	-ve	Proviral DNA	LTS1_2010	Cell pellet, plasma	Proviral DNA, RNA
LTS2_2004		+ve	No Sample	LTS2_2010	Cell pellet, plasma	Proviral DNA, RNA
LTS5_2004	DBS	+ve	Proviral DNA	LTS5_2010	Cell pellet, plasma	Proviral DNA, RNA
LTS8_2004	DBS	+ve	Proviral DNA	LTS8_2010	Cell pellet, plasma	Proviral DNA, RNA
LTS9_2004	DBS	D*	Proviral DNA	LTS9_2010		No sample
LTS10_2004	DBS	-ve	Proviral DNA	LTS10_2010	Cell pellet, plasma	Proviral DNA, RNA
LTS12_2004	DBS	+ve	Proviral DNA	LTS12_2010	Cell pellet, plasma	Proviral DNA, RNA
LTS17_2004	DBS	L	Proviral DNA	LTS17_2010		No sample
LTS20_2004	DBS	+ve	Proviral DNA	LTS20_2010	Cell pellet, plasma	Proviral DNA, RNA
LTS21_2004		-ve	No sample	LTS21_2010	Cell pellet, plasma	Proviral DNA, RNA
LTS22_2004	DBS	D**	Proviral DNA	LTS22_2010		No sample
LTS30_2004	DBS	-ve	Proviral DNA	LTS30_2010	Cell pellet, plasma	Proviral DNA, RNA

Table 2.2 A list of the type of sample collected from the normal progressors in Karonga. The samples are labelled with a unique number. The four control viruses (C3, C5, C8 and C9) and their country of origin are also listed along with the type of nucleic acid available.

Sample number	**Place of Origin**	**Type of sample**	**Type of nucleic acid available**
57968	Karonga	Cell pellet	Proviral DNA
58022	Karonga	Cell pellet	Proviral DNA
60560	Karonga	Cell pellet	Proviral DNA
61355	Karonga	Cell pellet	Proviral DNA
61591	Karonga	Cell pellet	Proviral DNA
61788	Karonga	Cell pellet	Proviral DNA
63921	Karonga	Cell pellet	Proviral DNA
65096	Karonga	Cell pellet	Proviral DNA
65867	Karonga	Cell pellet	Proviral DNA
C3	South Africa	Propagated cell supernatant	Viral RNA
C5	South Africa	Propagated cell supernatant	Viral RNA
C8	Nigeria	Propagated cell supernatant	Viral RNA
C9	Malawi	Propagated cell supernatant	Viral RNA

2.2.3 Section 1. Molecular characterisation of HIV-1 in LTS from Karonga

2.2.3.1 PCR amplification and cloning

Nested PCR and sequencing of a 750 bp region of *gag* p17p24 and a 500 bp region of *env* C2V3 was carried out as previously described (McCormack et al., 2002) on the 10 samples collected from the LTS in 2004 and the nine samples collected in 2010. Nested *gag* PCRs were applied to the 100 samples collected from 100 randomly chosen individuals in Karonga to search for the presence of the deletion in the current population. Briefly, both the primary and the secondary PCR amplifications were carried out in 100 µl reactions using the Expand High Fidelity PCR system (Roche). The PCR reaction consisted of 10 µl of 10X PCR buffer, 1.5 µM of $MgCl_2$, 2 µl of 10 mM dNTPs, 6 µl of each primer (Table 2.3) (5 pmol/µl) and 0.75 µl (2.6 units) of Expand *Taq* DNA Polymerase. The amount of DNA added to the primary PCR varied from 5 µl to 10 µl and later 5 µl of the primary reaction was added to the secondary nested reaction. The thermocycling conditions for both *gag* and *env* for the primary round of PCR were 94 °C for 2 minutes, followed by 15 cycles of 94 °C for 15 seconds, 50 °C for 30 seconds and 68 °C for 1 minute, followed by 25 cycles of the same with 5 second increment to each cycle during the extension stage followed by a final 72 °C for 10 minutes. The secondary PCR thermocycling conditions for *env* were 94 °C for 2 minutes, followed by 15 cycles of 94 °C for 15 seconds, 55 °C for 30 seconds and 72 °C for 1 minutes, followed by 25 cycles of the same with a 5 second increment to each cycle during the extension stage, followed by a final 72 °C for 10 minutes. The *gag* secondary PCR thermocycling conditions were the same except for an annealing temperature of 56 °C.

Table 2.3 List of the primers used in the *env* and *gag* nested PCR.

env primary PCR primers	ED5	5' ATGGGATCAAAGCCTAAAGCCATGTG 3'
	ED12	5' AGTGCTTCCTGCTGCTCCCAAGAACCCAAG 3'
env secondary PCR primers	ED31	5' CCTCAGCCATTACACAGGCCTGTCCAAAG 3'
	ED33	5' TTACAGTAGAAAAATTCCCCTC 3'
gag primary PCR primers	DT1	5' ATGGGTGCGAGAGCGTCAGTATT 3'
	DT7	5' CCCTGACATGCTGTCATCATTTCTTCT 3
gag secondary PCR primers	DT3	5' CATCTAGTATGGGCAAGCAGGGA 3'
	DT6	5' ATGCTGACAGGGCTATACATTCTTAC 3'

Successfully PCR amplified samples were quantified using a nanodrop spectrophotometer (Thermo Scientific) and were PCR purified using the HiYield Gel/PCR DNA Fragments Extraction Kit (Real Genomics) according to the manufacturer's instructions. The resulting purified PCR products were sequenced in both directions by LGC Genomics (Germany). Sequence chromatographs were examined and sequence contigs assembled in Seqman (DNAStar Inc.). Cloning of *gag* PCR products from one LTS (LTS2) who contained the deletion in the 1990s was carried out using the Stratagene PCR Cloning Kit (Agilent Technologies).

2.2.3.2 Sequence analysis

Multiple alignments of all *env* (74) and *gag* (65) sequences from subtype C infected LTS along with 40 control sequences were assembled and optimised in MacClade 4 (Sinauer Associates). Phylogenetic trees were reconstructed under the GTR + gamma model of DNA substitution implemented by RAxML 7.0.3 (Stamatakis, 2006) with all parameters optimised by RAxML. Confidence levels in the groupings in the phylogeny were assessed using 1000 bootstrap replicates as part of the RAxML phylogeny reconstruction. The subtype C ancestral sequence derived in previous work (Travers et al., 2004) was employed as the out-group for both *gag* and *env* trees. One LTS (LTS30) showed divergent sequence in the subtype C phylogeny and was subsequently subtyped by aligning the *env* and *gag* sequences to downloaded LANL subtype reference sequences and drawing a phylogenetic tree as described above.

Pairwise genetic distances from nucleotide sequences were computed by PAUP* 4.0 (D.L. Swofford, Sinauer Associates, Inc. Publishers) using models of evolution described by McCormack et al. (2002). For *gag,* the Kimura three-parameter model of evolution was used, with base frequencies A; 0.42, C; 0.21, G; 0.21, T; 0.16, Rate matrix A-C; 1.00, A-G; 3.13, A-T; 0.64, C-G; 0.64, C-T; 3.13, G-T; 1.00, I = 0.53 and G = 0.93. For *env,* the general time reversible model was used, with base frequencies A; 0.41, C; 0.19, G; 0.19, T; 0.22, Rate matrix A-C; 1.31, A-G; 4.51, A-T; 0.45, C-G; 0.65, C-T; 3.55, G-T; 1.00, I = 0.20 and G= 0.67. Intra patient genetic divergence at each time point was calculated by comparing the genetic distance between the earliest sequence available and all subsequent sequences from later time points. The Mann-Whitney U test carried out in SPSS (SPSS Statistics 17.0.1 -

December 2008) was used to look at the differences in pairwise genetic distances between LTS21 and all other LTS.

The BLOSUM62 matrix was used to score the likelihood of individual substitutions with indels, shared deletions relative to the other sequences and ambiguities also identified. The sequence available from the earliest time point for an individual was compared to all the sequences available from each subsequent sampling time point. A graphical representation of these observed amino acid substitutions that have occurred within the LTS was constructed.

2.2.4 Section 2. Replicative fitness

2.2.4.1 RT PCR and nested PCR amplification

Reverse Transcriptase PCR –(a)

The SuperScript III Reverse Transcriptase system (Invitrogen) was used to create cDNA using the RNA from the LTS. In a clean PCR tube 5 µl of RNA was added to a master mix comprising of 1.6 µl of 3' LTR B CAT primer (Appendix 1) (25 pmol/µl), 1 µl of 10 mM dNTPs and 13 µl of water. This was placed in a thermocycler for 1 cycle of 88 °C for 2 minutes, 70 °C for 10 minutes, 55 °C for 10 minutes with a hold then of 42 °C. Four µl of 5X First strand Buffer, 2 µl of 0.1 M DTT and 1 µl of SuperScript III was individually added to each sample (a master mix was not used as high concentrations of DTT can disable the SuperScript III enzyme). The samples were then held at 55 °C for 60 minutes followed by one cycle of 70 °C for 15 minutes. Five microliters of cDNA was then used in a primary nested PCR described.

Reverse Transcriptase PCR –(b)

RT PCR was used to create cDNA from the RNA extracted from the four subtype C control samples. Four microliters of the extracted viral RNA was added to 2 µl AccuScript RT Buffer, 1 µl of primer Env End (Appendix 1) (20 pmol/µl), 2 µl of 10 mM dNTPs and 7.25 µl of nuclease free water. The samples were then placed in a thermocycler at 88 °C for 2 minutes, 70 °C for 10 minutes, 55 °C for 10 minutes and finishing at 42 °C. Two microliters of 100 mM DTT, .25 µl of RNase Inhibitor and 1 µl of AccuScript RT (2.5 U/µl) (Agilent

Technologies) were then added to each sample individually followed by one cycle of 42 °C for 90 minutes and ending with 15 minutes at 70 °C.

Full Genome nested PCR:

Efforts were made to amplify the full genome (approximately 10, 000 bp) in two overlapping fragments in order to recombine it with a neutral backbone using yeast homologues recombination in order to create chimeric viruses. Initially, proviral DNA was used in a nested PCR with a set of primers for both the 5' half and the 3' half of the genome (Appendix 1). Both the primary and the secondary PCR amplifications were carried out in 50 µl reactions, to which 5 µl of PCR buffer, 1.5 µM of $MgCl_2$, 1 µl of 10 mM dNTPs, 0.5 µl of each primer (20 pmol/µl) and 2.0 units of Platinum *Taq* DNA Polymerase (Invitrogen) were added. The amount of DNA added to the primary PCR varied from 3 µl to 5 µl and later 5 µl of the primary reaction was added to the secondary nested reaction. The thermocycling conditions for both rounds of the nested PCR were 94 °C for 2 minutes, followed by 35 cycles of 94 °C for 30 seconds, 55 °C for 30 seconds and 72 °C for 3.5 minutes followed by 72 °C for 10 minutes.

The Expand Long Template PCR system (Roche) was also employed to amplify the 3' end of the HIV-1 genome. Five microliters of cDNA generated from RT PCR (a) described above, was used in a nested PCR. Both of the rounds of the nested PCR reactions were carried out in 50 µl, to which 5 µl of cDNA, 5 µl of 10X Expand Long Template Buffer, 1.75 µl of 10 mM dNTPs, 0.75 µl of each primer (20 pmol/µl) (Appendix 1) and 0.75 µl of Expand Long Template enzyme mix (5 U/µl) were added. The temperature regime was an initial denaturation of 94 °C for 1 minute followed by 10 cycles of 94 °C for 15 seconds, 55 °C for 30 seconds and 72 °C for 4 minutes. This was followed by 25 cycles of 94 °C for 15 seconds, 55 °C for 30 seconds and 72 °C for 4 minutes with an additional 20 seconds added per cycle.

Full env and gag nested PCR:

Four sets of primer pairs (Appendix 1) were used in efforts to amplify the whole *env* gene (2500 bp). Both primary and secondary PCRs were carried out in 50 µl reactions, to which 5

μl of PCR buffer, 1.5 μM of $MgCl_2$, 1 μl of 10 mM dNTPs, 0.5 μl of each primer (20 pmol/μl) and 2.0 units of Platinum *Taq* DNA Polymerase (Invitrogen) were added. Three microliters of the DNA template was used in the primary PCR and later 5 μl of this was added to the secondary nested PCR. After an initial denaturation of 94 °C for 2 minutes the reactions were exposed to cycles of 94 °C for 30 seconds, 55 °C for 30 seconds and 72 °C for 3.5 minutes followed by 72 °C for 10 minutes. The extension time was later extended to 6 minutes in some reactions. Modifications were also made to the reaction volume (100 μl) and the annealing temperature was lowered. A second polymerase was also employed in a number of reactions, *PfuTurbo* DNA Polymerase (Stratagene), in an attempt to achieve positive amplification. Primers designed to amplify full *gag* (1500 bp) (Appendix 1) were applied to a number of samples using the same PCR master mix and thermocycling conditions as used for full *env.*

env C2-V3 nested PCR:

Due to the immense difficulty encountered attempting to amplify the larger full genome and full *env,* primers designed to amplify a much smaller fragment of *env* that could then be incorporated into the vector, were used (480 bp)(Table 2.4). The primers and the vector contained regions of homology, which allowed yeast homologous recombination to incorporate the resulting PCR product into the vector. The nested PCR was carried out using 50 μl reactions and the amount of template DNA added was varied from 1 μl to 9 μl. Five microliters of the primary reaction was subsequently added to the secondary reaction. The PCR master mix consisted of 5 μl of PCR buffer, 1.5 μM of $MgCl_2$, 1 μl of 10 mM dNTPs, 0.5 μl of each primer (20 pmol/μl) and 2.0 units of Platinum *Taq* DNA Polymerase (Invitrogen). The thermocycling conditions included an initial denaturation of 94 °C for 2 minutes followed by 35 cycles of 94 °C for 30 seconds, 47 °C for 30 seconds and 72 °C for 1 minute followed by 72 °C for 10 minutes. Two microliters of cDNA created in RT PCR (b) was added to the C2-V3 nested PCR described above for the four subtype C control viruses. All successfully amplified PCR products were purified using the PureLink PCR Purification Kit (Invitrogen) according to the manufacturer's instructions.

Table 2.4 A list of the primers designed to amplify *env* C2V3.

Outside primary PCR primers	EnvB	3'AGAAAGAGCAGAAGACAGTGGCAATGA5'
	ED14	3'TCTTGCCTGGAGCTGTTTGATGCCCCAGAC5'
Inside secondary PCR primers	E80	3'CCAATTCCCATACATTATTGTG5'
	E125	3'CAATTTCTGGGTCCCCTCCTGAGG5'

2.2.4.2 Bacterial transformation with the vector pREC$_{nfl\Delta V3}$

The cloning vector, pREC$_{nfl\Delta V3,}$ a modification of pREC$_{nfl\ HIV-1}$ (Marozsan and Arts, 2003) with orotidine-5'-phosphate decarboxylase gene (*URA3*) in place of the C2-V3 region (480 bp) was utilised in this study as a neutral backbone, into which PCR products would be inserted using the yeast homologues recombination system in order to created chimeric viruses. Bacterial transformation using a stock of pREC$_{nfl\Delta V3}$ stored at -80 °C was carried out in order to create more of the vector. A 1.5 ml eppendorf tube and a 0.1 cm cuvette were placed on ice. Electrocompetent STBL4 cells (Invitrogen) were gently thawed on ice. Twenty microliters of cells were transferred to the cold 1.5 ml eppendorf and 80 µl of cold sterile water was added along with 1 µl of the vector, pREC$_{nfl\Delta V3}$. The mixture was incubated on ice for 1 minute before being transferred to the cold cuvette. The cells were electroporated at capacitance 25 µF, resistance 200 Ω and voltage 1.2 kV, after which 500 µl of SOC was immediately added to the cuvette. The cells were transferred to a 1.5 ml eppendorf and incubated on a shaker at 30 ºC for 1 hour. A 1:100 dilution of the culture was made using SOC and 100 µl of this was plated on pre warmed LB+Ampicilin plates. The plates were incubated at 30 ºC for 24 hours.

2.2.4.3 Bacterial miniprep to extract the vector

The vector DNA, pREC$_{nfl\Delta V3,}$ created above was extracted from the bacteria using the PureLink Quick Plasmid Miniprep Kit (Invitrogen). One colony was inoculated into 5 ml of LB+Ampiciln broth and incubated overnight on a shaker at 30 ºC. The next day the culture was centrifuged at 4000 rpm for 10 minutes. The supernatant was poured off and the cells were resuspended with 250 µl of Resuspension Buffer R3 (with RNase A) and transferred to a 1.5ml eppendorf tube. Following this 250 µl Lysis Buffer L7 was added and the tube was inverted gently 6 times and before incubation at room temperature for no more than 5 minutes. Three hundred and fifty microliters of Precipitation Buffer N4 was added to the

mixture and the tube was immediately inverted 6 times. The mixture was centrifuged for 10 minutes at 14000 rpm, after which supernatant from this was added to a spin column and centrifuged for 1 minute at 14000 rpm. After the flow though was discarded, 500 µl of Wash Buffer W10 was added to the column was incubated for 1 minute. The column was centrifuged for 1 minute at 14000 rpm and the flow through discarded. Seven hundred microliters of Wash Buffer W9 was then added and the column was centrifuged for 1 minute at 14000 rpm. The flow through was again discarded and the column was centrifuged for 1 minute at 14000 rpm to remove any residual wash buffer. The column was placed in a 1.5 ml eppendorf tube and 75 µl of sterile water was added to the centre. After a 1 minute incubation at room temperature the sample was centrifuged for 2 minutes at full speed. The concentration of plasmid DNA extracted was then measured using a spectrophotometer.

2.2.4.4 Homologues yeast recombination

The cloning vector, $pREC_{nfl\Delta V3,}$ was linearised using the restriction enzyme *Sac*II. Four micrograms of the vector, along with 2 µl of *Sac*II, 3 µl of Buffer and x µl of nuclease free water to bring the final volume to 30 µl were added to a 1.5 ml eppendorf tube and mixed before incubation over night at 37 °C. The enzyme was denatured the following day by incubation at 65 °C for 15 minutes. Successful digestion was confirmed by running 5 µl of the digested plasmid on 1 % agarose gel. This would then be recombined using the homologues yeast recombination, to create a chimeric virus containing the PCR.

A single yeast colony was inoculated into 50 ml of YPD medium and was grown up overnight by shaking at 30 ºC. The cells were pelleted in a centrifuge for 5 minutes at 4000 x g and then resuspended in 1 ml of sterile water and transferred to a 1.5 ml eppendorf tube. Again the cells were centrifuged for 5 minutes at 5000 x g, the supernatant poured off and the cells resuspended in 1 ml of freshly prepared TE/LiAc solution (1 volume of 10X TE Buffer, 1volume of 10X LiAc, and 8 volumes of sterile water). Fifty micrograms of salmon sperm carrier DNA was denatured by heating for 95 ºC for 8 minutes and was then cooled on ice until the temperature reached approximately 4 ºC. In a 1.5 ml eppendorf tube, 4 µg of the digested plasmid was mixed with the denatured sperm carrier DNA and 50 µl of the purified PCR product. To this, 300 µl or freshly made PEG solution (8 volumes of 50 % PEG 4000, 1

volume of 10X TE Buffer, and 1 volume of LiAc) was added along with 50 µl of the prepared yeast cells. The reaction was incubated on a shaker at 30 ºC for 30 minutes. The cells were heat shocked for 15 minutes in a water bath set to 42 ºC, after which they were gently pelleted by centrifugation. The supernatant was removed and the cells were resuspended in 250 µl of 1X TE Buffer. All 250 µl was plated on to a pre warmed C-Leu + 5 FOA plate and incubated at 30 ºC for 2 to 5 days.

2.2.4.5 Yeast miniprep to extract the recombined vector

In order to retrieve the successfully recombined vectors containing the PCR product and the neutral $pREC_{nfl\ HIV\text{-}1}$ backbone all yeast colonies that grew were inoculated into 2 ml of C-Leu and grown up overnight on a shaker at 30 ºC. The yeast cells were centrifuged at 4000 x g for 10 minutes and the cell pellet resuspended in 200 µl of breaking buffer (2 % v/v Trition X0100, 1 % sodium dodecyl sulfate, 100 mM NaCl, 10 mM Tis-HCL, and 1 mM EDTA) and transferred to a 1.5 ml eppendorf. Glass beads (0.3 g) along with 200 µl phenol/chloroform/isoamylalcohol were added to the mixture and vortexed for 2 minutes. The mixture was then centrifuged for 5 minutes at 14000 rpm. The aqueous top layer (approximately 100 µl) was removed and added to 170 µl of sterile water and 30 µl of 3 M sodium acetate. To this 700 µl of 100 % ethanol was added and the mixture was vortexed for 5 seconds before the mixture was centrifuged for 10 minutes at 14000 rpm. The supernatant was aspirated off and the pellet was washed with 700 µl of 70 % ethanol and then centrifuged for 5 minutes at 14000 rpm. The supernatant was pipetted off and the pellet was air dried for 5 minutes before being resuspended in 20 µl sterile water and stored at -20 ºC.

2.2.4.6 Bacterial transformation and miniprep

The vector DNA extracted from the yeast was then transformed into Electrocompetent STBL4 bacterial cells (Invitrogen) in order to create more of the recombined vector as described above with a few minor changes. Two microliters of plasmid DNA was used in the electroporation step and all 250µl SOC medium and bacterial cells were plated. The plasmid DNA was then extracted from the bacteria as described above, however 15 bacterial colonies were inoculated into 5 ml of LB+ampicillin broth and incubated overnight on a shaker at 30 ºC.

2.2.4.7 Transfection of 293T cells to produce HIV-1 chimeric virus

Recombinant chimeric virus was produced using 293T packing cells and the Effectene lipid system (Qiagen). The 293T cells were transfected with the C2-V3 recombinant HIV-1 vector created in the previous section along with the complementing vector pCMV_cplt, which contained the 5' LTR missing from the $pREC_{nfl\Delta V3}$ vector used as the neutral backbone. Five hundred nanaograms of the C2-V3 plasmid DNA along with 500 ng of the pCMV_cplt, plasmid and 3 μl of Fugene 6 were added to x μl of DMEM Media to obtain a final volume of 100 μl. The mixture was gently flicked and left to incubate for 15 minutes at room temperature. Each sample was then added drop by drop to individual 293T cells and swirled gently. The cells were then incubated at 37 ºC with 5 % CO_2 for 24 hours. After 24 hours the media containing all of the transfection reagents was removed and replaced with 2 ml DMEM Media with 10 % FBS and penicillin/streptomycin. The cells were incubated at 37 ºC with 5 % CO_2 for 48 hours.

2.2.4.8 Infection of U87.CD4.CXCR4 and U87.CD4.CCR5 cells with the chimeric virus

The cell free supernatant containing the HIV-1 virons was collected from the 293T cells and was used to infect U87.CD4.CCR5/CXCR4 cells, a human glioma cell line that expresses the receptor CD4 and can express either of the co receptors CCR5 and CXCR4. The cell free supernatant from the 293T cells containing the newly constructed virus composed of the PCR product and the neutral backbone $pREC_{nfl\Delta V3,}$ was removed and centrifuged at 1500 rpm for 5 minutes. Ten microliters of supernatant was added to a 96 well plate and later used to test for RT activity. The remaining media was equally divided and placed on top of U87.CD4.CXCR4 and U87.CD4.CCR5 cells in order to infect them with the newly constructed virus. The cells were then incubated at 37 ºC with 5 % CO_2. Ten microliters of supernatant was collected and added to a 96 well plate every 2 days starting 3 days post infection and was this used to monitor for RT activity using a radioactive assay described below. After 9 days all supernatant was removed and stored at -80 °C.

2.2.4.9 RT assay

A Radioactive RT assay was used to measure the amount Reverse Transcriptase activity present in the infected 293T cells, U87.CD4.CXCR4 and U87.CD4.CCR cells from the chimeric virus created above, thus indicating the presence and amount of replicating virus.

Twenty five microliters of RT master mix (1 mM Tris-HCL (pH 7.8), 2 mM KCl, 1 mM dithiothreitol, 200 mM $MgCl_2$, 1 U/ml, 100 μg/ml of poly (rA) · poly $(dT)_{12-18}$, 0. 5 % (vol/vol) NP-40, and 1 μl of fresh 10-mCi/ml [α-32P]-dTTP per ml) was added to the 96 well containing 10 μl of cell free supernatant collected from both the 293Ts and U87s. After incubation at 37 ºC for 2.5 hours, 10 μl of the RT reaction mixtures were blotted onto a Whatman DE81 filtermat with 96 wells (Whatman), and allowed to dry for 10 minutes at room temperature. The filtermat was washed 5 times with 1X SCC (0.15 M NaCl, 0.015 M Sodium Citrate) for 5 minutes in a shaker platform. It was then washed twice with 85 % ethanol for 5 minutes each also in a shaker platform. After the filtermat was allowed to dry, the CPM (counts/Minute) was measured using a 96 direct Beta Counter.

2.2.4.10 VERITROP Assay

VERITROP is a fusion based assay that allows you not only to determine if the HIV-1 virus being studied contains a functional V3 loop and is therefore able to successfully fuse with and infect cells presenting the CD4 receptor and the CCR5/CXCR4 co-receptor on the cell surface. It can also be used to determine, which co receptor the virus in question utilises. It utilises a firefly luciferase reporter vector that is activated in the presence of HIV-1 rev and tat.

Day 1: U87.CD4.CCR5/CXCR4 cells were transfected with a pDM128fluc reporter vector using the Effectene lipid system (Qiagen). Six micrograms of the pDM128fluc reporter vector and 3 μl of Fugene 6 were added to x μl DMEM Media to obtain a final volume of 100 μl. The tube was gently flicked and left to incubate for 15 minutes at room temperature. The mixture was then added drop by drop to the cells and the plate was swirled gently. The cells were incubated at 37 ºC with 5 % CO_2 for 24 hours. On the same day 300 000 293T cells were plated into 6 well plates using DMEM (10 % FBS/Penicillin/Streptomycin) Media to a total volume of 2 mls.

Day 2: 65 000 of the transfected U87.CD4.CCR5/CXCR4 transfected with a pDM128fluc reporter vector from day, were plated using DMEM Media (10 % FBS/Penicillin/Streptomycin) to a total volume of 500 μl. The 293T cells from day 1 were

transfected with the vector containing the PCR product in the neutral $pREC_{nfl\Delta V3}$ backbone using the Effectene lipid system (Qiagen) as described above in section 2.2.4.7.

Day 3: The supernatant from the transfected 293Ts from day 2, was gently removed and 1 ml of DMEM (15 % FBS/Penicillin/Streptomycin) Media was added to the cells and the cells were gently resuspended by pipetting up and down 5 times. Sixty five thousand of the cells were then transferred to the plated U87.CD4.CCR5/CXCR4 cells from day 2. These cells were incubated at 37 °C with 5 % CO_2 for 18 hours. On the fourth day the supernatant was removed from the U87.CD4.CCR5/CXCR4 cells and the cells were lysed with 100 μl of Glo Lysis Buffer (Promega) for 15 minutes. Fifty microliters of the cell lysate was combined with 50 μl of Bright-Glo (Promega). The amount of luciferase expression was measured by reading the amount of luminescence in the cell lysate on a VICTOR plate reader.

2.2.4.11 Sequencing and phylogenetic analysis of the vectors

All successfully recombined yeast vectors were partially sequenced (to cover the region of interest containing the PCR product) in both directions using the E80 and E125 primer set. Sequence chromatographs were manually edited in SeqMan (DNA Star Inc). The sequences were then assembled into a multiple alignment in MacClade 4.0 (Sinauer Assoc) and were aligned to all previously sequenced LTS *env* samples from the 1980s, 1990s, 2004 and 2010. Phylogenetic trees were reconstructed under the GTR + gamma model of DNA substitution implemented by RAxML 7.0.3 (Stamatakis, 2006) with all parameters optimised by RAxML. jpHMM (jumping profile Hidden Markov Model) (Zhang et al., 2006) was used to identify possible genomic recombinant events within the C2-V3 region of *env*.

2.3 Results

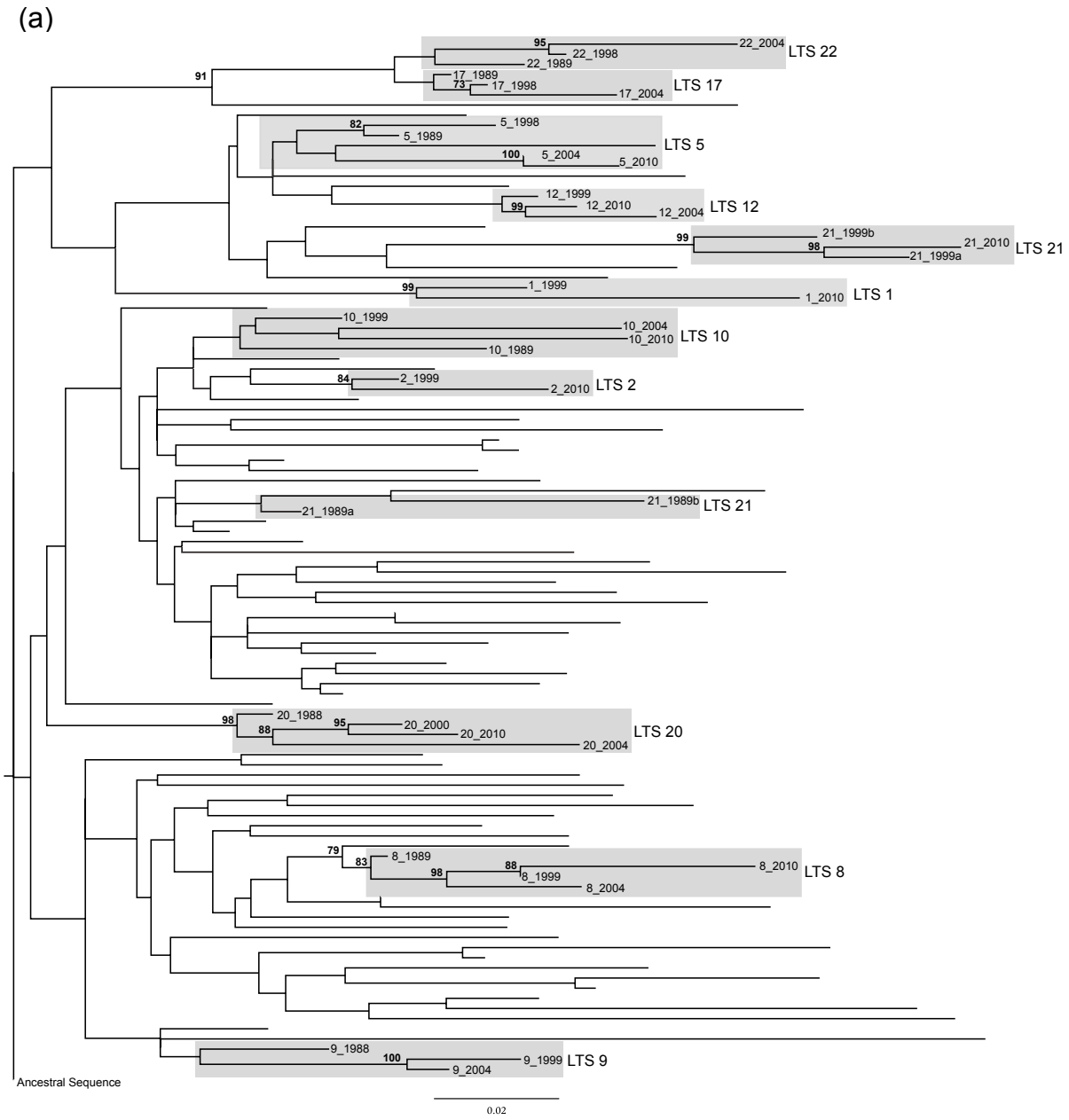
2.3.1 Section 1. Molecular characterisation of HIV-1 in LTS from Karonga

2.3.1.1 Three amino acid deletion

Amplification from the dried blood spots from the 1980s to explore evidence of the three amino acid deletion in that time period was largely unsuccessful. The three amino acid deletion was not found in any *gag* sequences produced from LTS in 2004 and 2010 or in over 100 sequences produced from blood samples collected in 2007-2008. Furthermore, none of 50 sequences from 50 clones produced from the 1998 sample from LTS2 (which had previously showed the deletion) contained the deletion. All of the raw data from the sequences used in McCormack et al. (2006) was re-examined and the deletion was not found in any of the sequences.

2.3.1.2 Phylogeny and genetic divergence of the LTS

No PCR amplification was achieved from DBS collected from LTS30 in 1989 or 1999 (McCormack et al., 2006). Both *gag* and *env* sequences were retrieved from DNA extracted from the cell pellet retrieved in 2010. Phylogenetic reconstruction identified LTS30 as an unclassifiable subtype grouping with other unclassifiable subtype described by McCormack et al. (2002) (Appendix 2 and 3). Both subtype C gene trees (*gag* and *env*) showed that for most individuals the sequences from the different time points grouped together (10/11 for *gag* and 8/10 for *env*) but only half grouped with significant bootstrap support (Figure 2.1 a and 1b). Sequences from LTS21 formed multiple clusters on both gene trees, which is consistent with the pattern seen in McCormack et al. (2006). In order to further explore this, additional *gag* and *env* consensus sequences were produced from DNA re-extracted from the DBS collected in 1989 and 1999 for this individual with additional *env* sequences also produced from the 2010 DNA sample. In *gag,* the 1990s sequences were ancestral to the 2010 sequences and the two 1989 sequences grouped distantly from them (Figure 2.1 a). In *env* the 2010 sequences showed further variation with two sequences grouping away from the 1999 sequences (Figure 2.1 b). The average genetic distance between the eight *env* sequences (across all time points) was 12 %, which was significantly higher than the 8.8 % genetic distance between all other LTS sequences from all individuals at all the different time


(a)
95
22_2004
22_1998
22_1989
LTS 22
91
17_1989
73
17_1998
17_2004
LTS 17
82
5_1998
5_1989
LTS 5
100
5_2004
5_2010
12_1999
99
12_2010
12_2004
LTS 12
99
21_1999b
98
21_2010
21_1999a
LTS 21
99
1_1999
1_2010
LTS 1
10_1999
10_2004
10_2010
10_1989
LTS 10
84
2_1999
2_2010
LTS 2
21_1989b
21_1989a
LTS 21
98
20_1988
95
20_2000
88
20_2010
20_2004
LTS 20
79
83
8_1989
88
8_2010
98
8_1999
8_2004
LTS 8
9_1988
100
9_1999
9_2004
LTS 9
Ancestral Sequence
0.02

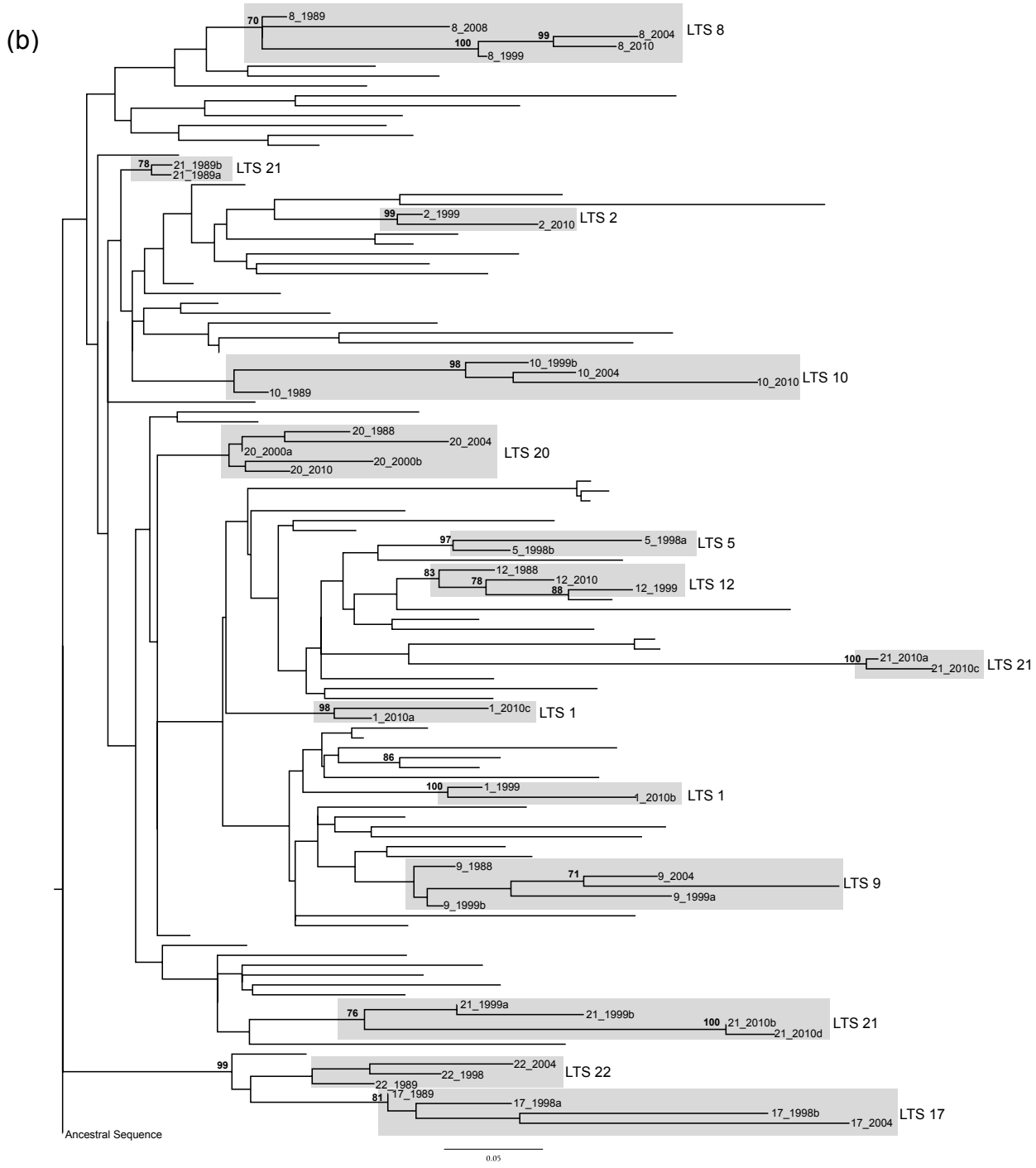


Figure 2.1 Maximum Likelihood trees generated from (a) *gag* and (b) *env* gene sequences from LTS dating from 2004 and 2010 along with local control sequences. Only the LTS sequences are labelled and are named by the LTS number (see Table 1) and the year the sample was collected. Sequences labelled a, b, c and d refer to multiple consensus sequences generated from the same time point. Bootstrap values of over 70 are marked on the relevant branches.

points (Mann Whitney U=432, P<0.001). This level of genetic variation within an individual may indicate that there may be unrelated strains present in this individual.

The genetic distance between two of the sequences collected from 2010 was higher (17.5 %) than between the sequences collected in the 1980s and 1990s (7.8 %) from the same individual. The sequences came from a female who was 21 when she was first identified as being HIV-1 positive. She has maintained a very low $CD4^+$ count for the last six years, 47 cells/mm^3 in 2004 and 32 cells/mm^3 in 2010. At both visits she was described as being healthy and showing no signs of AIDS and refused ART on both occasions. Divergent *env* sequences were also found amongst sequences from the 2010 sample of LTS1, a female who was 37 when she was first identified as being HIV-1 positive. Her CD4 count was 586 cells/mm^3 in 2004 and had fallen slightly to 449 cells/mm^3 in 2010. At that time she had not been referred for ART.

2.3.1.3 BLOSUM62 matrix

The BLOSUM62 matrix was used to assess the likelihood of amino acid substitutions between sequential sequences of *env* and *gag* for each LTS, with indels and shared deletions relative to the other sequences also noted. A graphical representation of these observed amino acid substitutions within each LTS sample from one time period to another shows the large amount of change that was apparent across most of *env* C2-V3 in all individuals (Figure 2.2 a). There are a number of positions that changed within nearly all of the LTS, e.g. HXB2 positions 268-269 showed substitutions in eight LTS (* Figure 2.2 a). Some of these changes are less likely mutations, e.g. in six individuals (LTS8, 9, 10, 12, 20 and 22) there was a change from glycine to glutamic acid or *vice versa*. In one individual (LTS17) there was deletion at position 269 within the 1980s sequence and one of the sequences from the 1990s. This deletion was not present in the second sequence retrieved from the 1990s or 2004.

Mutations at positions 11, 24 and 25 within the V3 loop have been associated with a change in co-receptor usage with a shift from negatively charged amino acids to positively charge amino acids being suggested to result in a switch from the use of CCR5 to CXCR4-usage. Only LTS17 showed evidence of a change to a positive charge in this region by 2004 but had left the district by 2010 and so no additional information is available on this person. A large

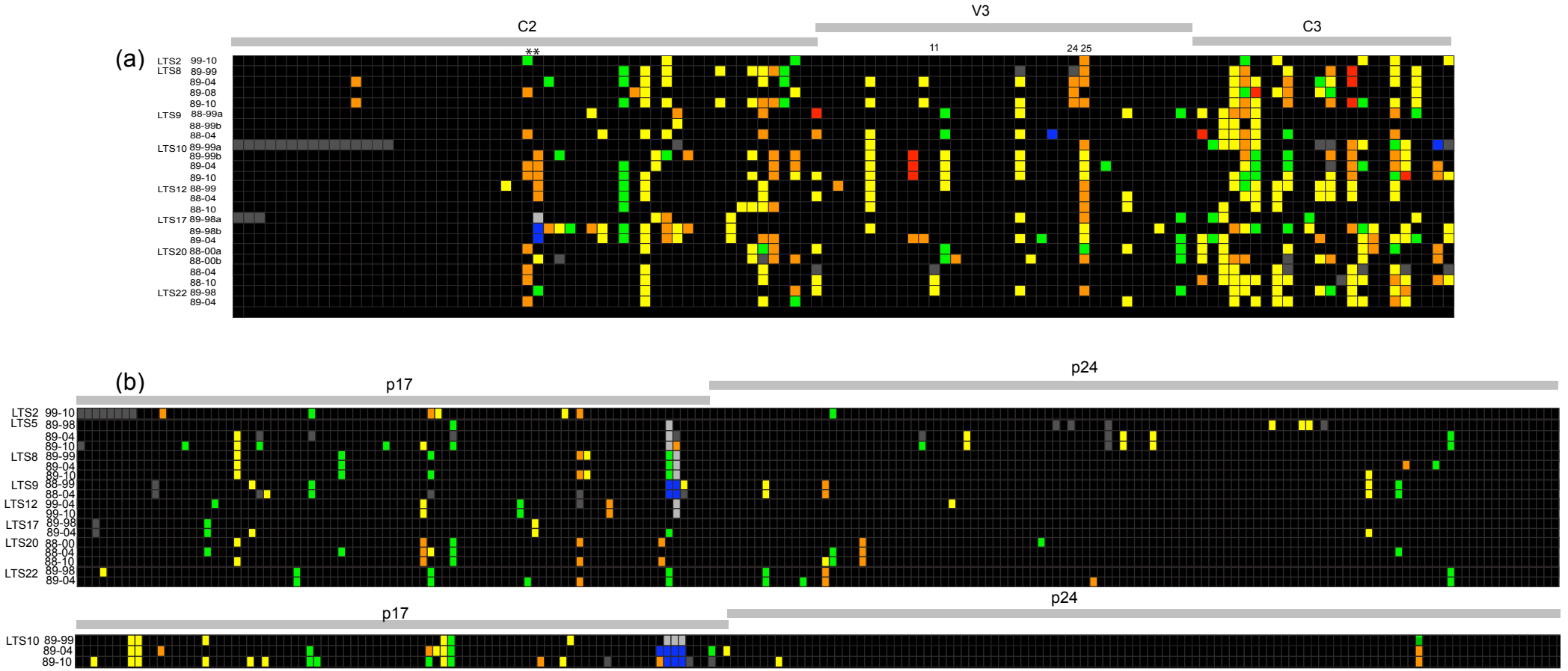


Figure 2.2 A graphical representation of observed amino acid substitutions that have occurred (a) *env* and (b) *gag* within survivors over time. Substitutions are colour coded by likelihood according to the BLOSUM62 matrix with green being the most likely and red the least (green>yellow>orange> red), blue indicating an insertion in one of the sequences relative to the other, pale grey a shared deletion and dark grey and ambiguous site e.g. a shot codon. The sequence collected at the earliest available time point was compared to all available *env* sequences from subsequent time points. The comparisons are labelled by the LTS number and the years being compared. Those labelled with an 'a' or 'b' refer to multiple sequences generated from the same time point. ** mark positions 268 and 269 in *env* using HXB2 numbering. LTS10 contains a two amino acid insertion in *gag* and is therefore has a longer *gag* than the other LTS sequences.

amount of change at the amino acid level was seen just after the V3 loop in the C3 region. Indeed the high degree of genetic divergence between multiple *env* sequences from the same time point in some LTS rendered comparisons of pairwise genetic distances between time points meaningless (even when we excluded the individuals with possible superinfection mentioned above). For example, two consensus sequences from LTS17 from 1999 showed a genetic distance of 4 %.

Comparing amino acid substitutions in *gag* sequences showed higher numbers of substitutions within the *gag* p17 domain when compared to p24 as might be expected as the p17 region is more variable and, using the BLOSUM62 matrix as a reference, most of these substitutions were changes that were more likely to occur (Figure 2.2 b). Different patterns of genetic divergence in *gag*, calculated using pair wise distances, were apparent amongst the LTS over time. Two LTS (LTS5 and LTS8) showed an overall trend of an increase in genetic divergence in *gag* over time similar to the trend described by Shankarappa et al. (1999) (Figure 2.3). Both LTS10 and LTS20 showed a general increase in genetic divergence from 1989 to 2004, however in 2010 the amount of genetic divergence from 1989 decreased for LTS20 and plateaued in LTS10. Three of these individuals had begun ART between 2004 and 2010 and one has since been referred. Two individuals (LTS9 and LTS22) who were alive in 2004 but had died by 2010 showed very different patterns of divergence. LTS9 showed an increase in divergence from 1989 to 1999, which decreased in 2004, while in LTS22 there was a linear increase in divergence from 1989 to 2004 (Figure 2.3).

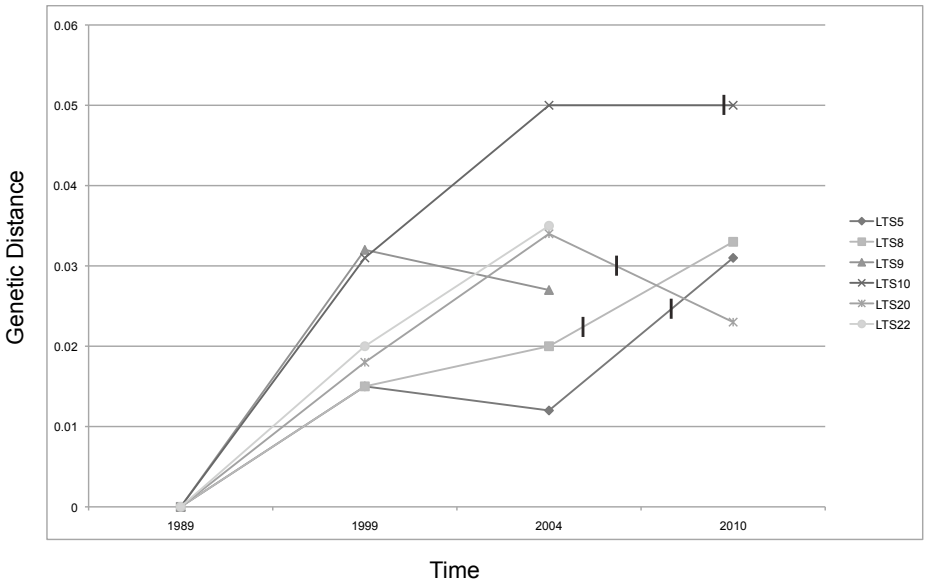


Figure 2.3 The genetic divergence seen in *gag* over time in LTS5, 8, 9, 10, and 20. All sequences from one individual from different time points were compared to the sequence generated from the earliest time point available. LTS9 and LTS22 had died before 2010. LTS5 had begun ART in 2008, LTS8 in 2005, LTS20 in 2006 and LTS10 had been referred for ART in 2010 as indicated by the vertical lines.

2.3.2 Section 2. Replicative fitness

2.3.2.1 PCR amplification

A significant amount of time was spent trying to amplify large fragments of the HIV-1 genome from the proviral DNA found in the LTS and normal progressors to utilize these fragments in the yeast based recombination cloning system to ascertain the replicative fitness of the viruses present. Attempts to amplify the full HIV-1 genome in two over lapping fragments from extracted proviral DNA proved unsuccessful despite employing numerous primer combinations and extensive DNA concentration optimizations. Either PCR amplification failed completely or amplification of multiple smaller products of the wrong size occurred. RT PCR RNA extracted from the LTS plasma was also unsuccessful with no amplification products of any kind. Ethanol precipitation to purify RNA extracted from the LTS did not result in amplification of the HIV-1 genome yielding smearing on the agarose gel due to spurious amplification.

The focus was shifted to amplification of smaller regions of the genome due to the unsuccessful efforts in amplifying the full genome. A large array of different primers designed to amplify the full *env* were applied to all the LTS and normal progressor proviral samples, however all attempts proved unsuccessful. Amplification of multiple or single products smaller than the expected 2500 bp was seen or no amplification was observed. Primers designed for full *gag* also failed to generate PCR products. A smaller section (C2-V3, 480 bp) of *env,* which contains the V3 loop became the main focus. Primers designed to amplify the C2-V3 region were applied to the extracted proviral DNA from all 28 Karonga samples and the cDNA from the four subtype C control viruses (Table 2.1 and 2.2) and proved to be more effective than the attempts to amplify the larger fragments of the HIV-1 genome. Thirteen LTS and five NP and the four subtype C control viruses were successfully PCR amplified and PCR purified as can be seen in Table 2.5.

An almost constant problem encountered during PCR amplification was contamination of PCR reagents or master mix. Contamination was thought to have been caused by the vector $pREC_{nfl,}$ which was used ubiquitously in the lab. All PCR reagents were added in a PCR hood following UV light sterilisation for 15 minutes. This did not prevent the DNA contamination in spite of increasing the UV light sterilisation times and wiping of all surfaces

within the PCR hood and gloves with DNase wipes along with using fresh reagents. Subsequently, PCRs were carried out in a separate laboratory within the building and using pipettes, which had in the past never been used in PCR or any other DNA or RNA reactions. This minimised the problem however contaminated PCR products led to long time delays during the project.

2.3.2.2 Homologous yeast recombination

The homologous yeast recombination approach (Marozsan and Arts, 2003) using the vector $pREC_{nfl\Delta V3}$ was applied to all successfully amplified C2-V3 products (LTS1, 2, 5, 10, 12 (2004 and 2010 sample), 21, and 30). Between 3 and 30 yeast colonies were produced as a result of recombination between the PCR products and the vector $pREC_{nfl\Delta V3}$ (Table 2.5). Approximately 100 colonies are thought to be indicative of a successful recombination however, the amount of colonies for each sample was comparable to the positive control, which produced 17 colonies. The recombined vector DNA was extracted from the yeast and was transformed into *E.coli* bacteria in order to create more copies of the vector. The amount of plasmid DNA successfully extracted from each of the samples ranged from 31-151 ng/µl (Table 2.5). The samples collected from LTS9, 10, 17, 20 22 and 30 as well as C3, C5, C8 and C9 all failed to recombine with the $pREC_{nfl\Delta V3}$ vector and produced no colonies when grown on the C-Leu + 5 FOA plates.

2.3.2.3 Cell infection

A vector containing the 5' LTR (pCMV_cplt) and the yeast recombined vectors containing the $pREC_{nfl\Delta V3}$ neutral backbone and the PCR products were transfected into 293T cells. Virus containing cell-free culture medium was collected 3 days post infection. A radioactive reverse transcriptase assay was used to measure the amount of RT activity for each sample. The presence of RT is indicative of successful complementation of the two vectors into a budding viron. All 8 LTS and 5 NP samples showed positive signs of RT activity with the counts per minute range from 322 to 1766. The results of the RT assay can be seen in Table 2.6. Full $pREC_{nfl}$ was used as a control and recorded 1299 counts per minute.

Table 2.5 PCR amplification and yeast recombination results. The number of colonies following yeast recombination is reported along with the amount of extracted plasmid DNA from the bacterial transformation process.

Samples collected in 2004				Samples collected in 2010			
LTS number	**PCR Result**	**No. of colonies**	**DNA concentration**	**LTS number**	**PCR Result**	**No. of colonies**	**DNA concentration**
LTS1_2004	Negative			LTS1_2010	Amplified	9	117 ng/μl
LTS2	No sample			LTS2_2010	Amplified	5	42 ng/μl
LTS5_2004	Negative			LTS5_2010	Amplified	30	31 ng/μl
LTS8_2004	Negative			LTS8_2010	Multiple bands		
LTS9_2004	Amplified			LTS9	No sample		
LTS10_2004	Amplified			LTS10_2010	Amplified	3	65 ng/μl
LTS12_2004	Amplified	12	151 ng/μl	LTS12_2010	Amplified	8	108 ng/μl
LTS17_2004	Amplified			LTS17	No sample		
LTS20_2004	Amplified			LTS20_2010	Negative		
LTS21	No sample			LTS21_2010	Amplified	6	139 ng/μl
LTS22_2004	Amplified			LTS22	No sample		
LTS30_2004	Amplified			LTS30_2010	Amplified	9	113 ng/μl
57968	Amplified	3	119 ng/μl				
58022	Negative						
60560	Negative						
61355	Amplified	5	32 ng/μl				
61591	Amplified	14	65 ng/μl				
61788	Amplified	9	69 ng/μl				
63921	Negative						
65096	Negative						
65867	Amplified	13	120 ng/μl				
C3	Amplified						
C5	Amplified						
C8	Amplified						
C9	Amplified						

The RT assay showed that the HIV-1 chimeric viruses failed to infect both cell types expressing the two different co receptors, CCR5 and CXCR4. The results of the RT assay can be seen in Table 2.6. If cells were successfully infected, an increase in RT activity as infection time increased would be expected until the level of activity levelled off due to cell death and therefore lack of new cells to infect. Three days post infection the CPM ranged from 21.3 to 137.0 in the U87.CD4.CCR5, which was below the 193 CPM seen in the control sample. The CPM reading within U87.CD4.CXCR4 cells ranged from 23.3 to 116.0, which was lower than 150 the CPM seen in the positive control sample. An increase in RT activity was not detected in either cell lines or in the positive control samples five or seven days post infection. By the ninth day post infection the CPM reading ranged from 21.0 to 74.7 CPM in the U87.CD4.CCR5 cells, which was lower than the positive control (192.0 CPM). Within the U87.CD4.CXCR4 cells the range went from 4.7 to 57.7 CPM with a positive control reading of 18 CPM, all of which were unacceptably low readings and indicative of a negative infection. If the infection had taken hold, CPM readings comparable to the readings in the 293T RT assay would have been expected.

2.3.2.4 VERITROP assay

Despite the fact the RT assay from above showed no presence of replicating viruses in the U87.CD4.CXCR4/CCR5 cell lines the functionality of the V3 loop and the co-receptor usage was assayed for the eight long term survivor and five normal progressor chimeric viruses containing the C2V3 PCR product and the $pREC_{nfl\Delta V3}$ neutral backbone using the VERITROP Assay. All the 13 chimeric viruses failed to infect the U87.CD4.CXCR4 cells with readings similar to the negative control as can be seen in Figure 2.4. Four chimeric viruses (LTS2, LTS5, 61591, and 61788) successfully fused with the U87.CD4.CCR5 cells. Three other samples (LTS1, LTS10, and 65887) showed luciferase activity slightly higher than the negative control in the U87.CD4.CCR5 cells.

Table 2.6 Results of RT assays carried out on the supernatant collect from the 293T cells and the U87.CD4.CXCR4/CCR5 cells following infection with the reconstructed viruses.

Sample Number	293Ts CPM	U87.CD4. CCR5 CPM	U87.CD4. CXCR4 CPM	U87.CD4. CCR5 CPM	U87.CD4. CXCR4 CPM	U87.CD4. CCR5 CPM	U87.CD4. CXCR4 CPM	U87.CD4. CCR5 CPM	U87.CD4. CXCR4 CPM
		3 Days Post infection		5 Days Post Infection		7 Days Post Infection		9 Days Post Infection	
LTS1_2010	734.0	57.7	48.0	58.7	41.7	27.7	38.3	36.7	9.7
LTS2_2010	1032.0	71.0	65.3	76.7	62.0	22.7	48.3	43.0	12.0
LTS5_2010	1531.0	99.0	79.7	109.0	59.3	25.7	61.3	62.0	9.7
LTS10_2010	1087.0	70.7	94.7	76.7	88.7	30.3	101.0	38.0	57.7
LTS12_2010	829.0	77.7	48.7	69.0	49.7	28.3	45.3	38.3	12.0
LTS12_2004	800.0	35.3	52.0	18.7	49.7	7.0	45.3	28.7	23.0
LTS21_2010	322.0	21.3	23.3	23.0	18.0	14.0	24.0	21.0	10.7
LTS30_2010	1144.0	95.3	108.0	109.0	105.0	36.7	111.0	68.0	52.3
57968	1042.0	128.0	49.7	104.0	47.7	28.3	49.3	55.3	4.7
61355	931.0	74.0	85.3	45.0	57.7	14.0	43.7	31.7	8.7
61591	1288.0	123.0	73.3	89.0	58.3	20.7	45.3	67.7	9.0
61788	1776.0	137.0	116.0	88.3	79.3	24.3	77.3	74.7	11.7
65867	425.0	38.3	35.0	28.0	28.0	4.7	29.7	25.7	16.0
Control	1299.0	193.0	150.0	120.0	126.0	30.0	129.0	192.0	18.0
Negative	55.7	14.3	12.3	10.0	7.3	4.0	9.7	7.0	5.0

2.3.2.5 Phylogeny of the recombinant vectors

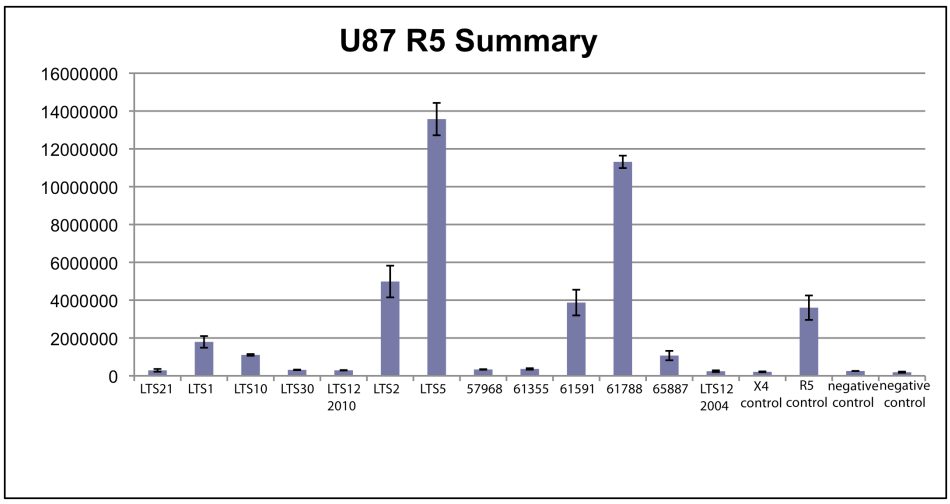
All 13 yeast recombined vector samples were sent for sequencing using the E80 and E125 primers, however only four of the LTS vectors (LTS2, 5, 10 and 12) and three of the normal progressors vectors (57968, 61788, and 65867) were successfully sequenced in one direction with the E80 primer. Phylogenetic analysis of the four LTS sequences obtained from the vectors showed only one sequenced vector (LTS10_vector) to group with its parallel sequence obtained directly from extracted DNA from the same time period as can be seen in Figure 2.5. The remaining six sequences formed a unique cluster together, which was highly supported by bootstrapping. No other LTS sequences fell within this grouping.

jpHMM was used to analyze the seven vector sequences further in order to identify any possible recombinant events. jpHMM identified six samples, LTS2, 5, 12, 57968, 61788, and 65867 as B/C recombinants within the C2-V3 region. The break point varied between positions 7084 and 7148 (HXB2 numbering) within the samples, which is located in the middle of the V3 loop (HXB2 positions 7109 to 7217). LTS10 was not a recombinant and was identified as subtype C within the sequenced region (Table 2.7).

Table 2.7 jpHMM results for the sequenced vectors. HXB2 numbering is used to identify the break point between the two subtypes.

Vector	HXB2 numbering	Subtype	HXB2 numbering	Subtype
LTS2vector	6945 – 7123	Subtype B	7124 – 7289	Subtype C
LTS5vector	6945 – 7148	Subtype B	7149 – 7289	Subtype C
LTS10vector	6945 – 7289	Subtype C		
LTS12vector	6945 – 7084	Subtype B	7085 – 7289	Subtype C
57968vector	6945 – 7100	Subtype B	7101 – 7289	Subtype C
61788vector	6945 – 7103	Subtype B	7104 – 7289	Subtype C
65867vector	6945 – 7084	Subtype B	7085 – 7289	Subtype C

(a)



(b)

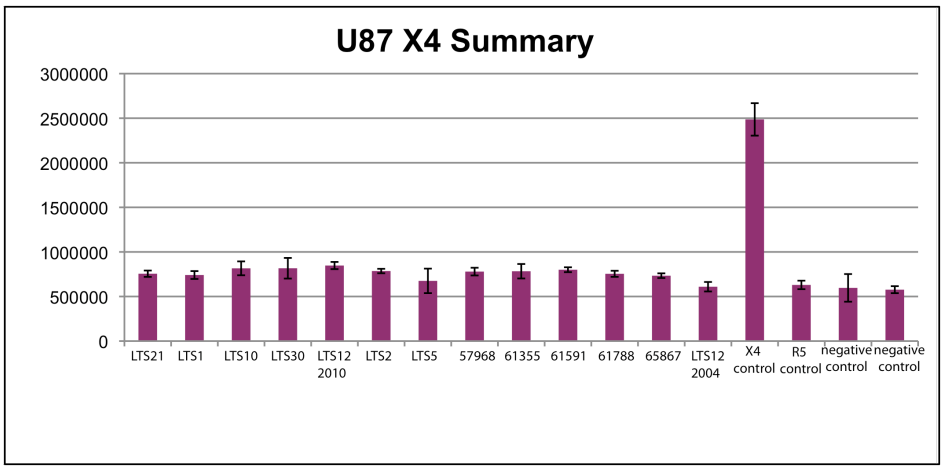


Figure 2.4 Results of the Veritrop assay for the CCR5 co receptor (a) and the CXCR4 co receptor (b). The sample labels are on the X-axis and the Reflective Light Units are on the Y-axis. The positive and negative controls are the last 4 columns.

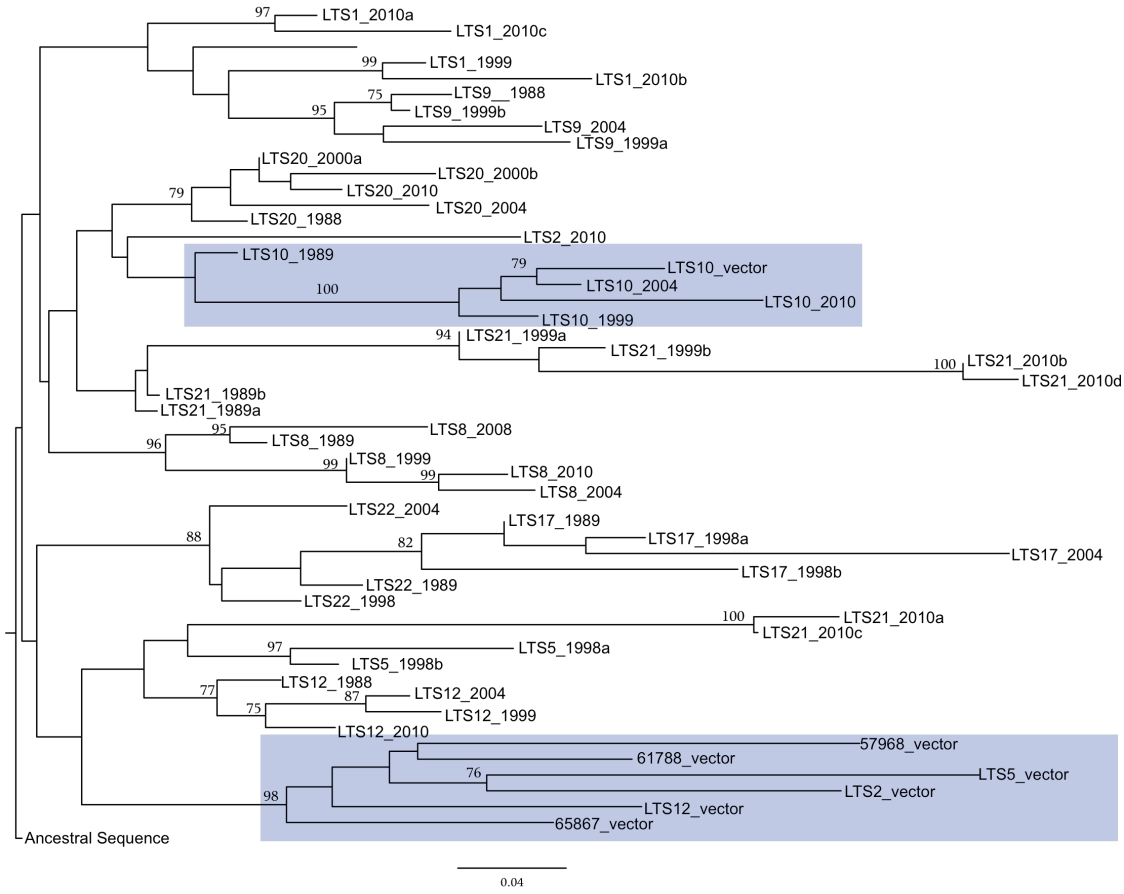


Figure 2.5 Maximum Likelihood tree featuring the sequenced vectors and all other *env* sequences from the LTS at all time points. The clades containing the vectors are boxed. Bootstrap values of above 70 are marked.

2.4 Discussion

As part of an investigation to study viral elements enabling disease non-progression within HIV-1 infected long-term survivors from Karonga District, Malawi, I set out to further characterise the emergence of a three amino acid deletion in the p17 matrix of *gag* previously described within these LTS (McCormack et al., 2006). Proviral DNA extracted from Dried Blood Spots collected in the 1980s proved difficult to amplify using PCR. The DBS had been frozen at -20 °C for over 20 years and had been exposed to a number of freeze thaw events and it is highly probable that this had led to fragmentation of DNA present. Previous studies found –20 °C suitable for long term storage of DBS (six years) (Cassol et al., 1991; Cassol et al., 1992; McNulty et al., 2007; Steegen et al., 2007) but this work suggests an upper limit to the length of time such samples can be stored successfully in this way for DNA-based studies.

Significant difficulties were also encountered when trying to amplify larger fragments of the HIV-1 genome within the LTS. The whole blood and DBS samples were collected in Karonga, a resource poor setting, that often suffers from electricity failures resulting in sub optimal storage for the collected cell pellet and plasma samples. Another difficulty associated with working in these limited settings is the transport of the extracted DNA or RNA to the research laboratory, which is often in another city or country as in the case here where the samples had to travel to Ireland and then to America. The financial cost of transporting HIV positive body fluids is very high due to the infectious nature of the material and the need to transport the samples on dry ice. Another option is to extract the DNA or RNA at the point of sample collection. DNA is more stable than RNA and can travel at room temperature, but RNA is highly susceptible to fragmentation and so was placed in a stabilizing agent twice to allow for transport from Karonga to the research laboratory in Ireland and then to Dr Eric Arts laboratory in America. It is possible that contaminants or PCR inhibitory products might have been introduced to the samples due to the preservation and stabilisation process.

Amplifying large fragments of the HIV-1 genome from proviral DNA is also affected by the interference of the human DNA extracted along with the viral DNA, which can interfere with PCR (Weidner et al., 2011). A combination of poor template quality, and limited time

available to fully optimize the PCRs contributed to the unsuccessful attempts to amplify large fragments of the HIV-1 genome. The viral load was not available for any of the LTS thus the starting copy number of template was unknown and a low amount of template may also have hampered PCR amplifications of large fragments of the HIV-1 genome. This problem was also encountered in Chapter 5 when attempting to amplify the full HIV-1 genome from three inter-subtype recombinant viruses from Karonga. Problems of PCR contamination within Dr Eric Art's laboratory also hindered progress. Contamination was almost certainly due to the ubiquitous use of vector $pREC_{nfl}$ which is constructed from the nearly full length HIV-1 genome, and which is used as a neutral backbone in the yeast recombination process to create chimeric virus. This highlights the disadvantage of multiple people performing similar procedures and using the same workstations and equipment (e.g. PCR hoods and pipettes).

While amplification of larger fragments of the HIV-1 genome to subsequently use in the yeast recombination process to create chimeric viruses was unsuccessful, the smaller *env* C2-V3 fragment was successfully amplified from most of the LTS and normal progressors. However, homologous yeast recombination of the PCR products with the $pREC_{nfl\Delta V3}$ vector produced only a maximum of 30 yeast colonies (Range: 2-30). Approximately 100 colonies is indicative of successful recombination, subsequently little confidence can be placed in any the results obtained from the resulting chimeric viruses. This was also highlighted by the fact that most of the vectors sequenced were inter-subtype recombinants (subtypes B and C) in the C2-V3 region. The subtype within the C2-V3 region should have matched that of the sample, from which the PCR product was generated, in the case here, subtype C. Unfortunately, due to time constraints the experiments were not repeated. More time spent both optimizing the PCRs and the homologous recombination procedure would have been needed to obtain meaningful results.

When *gag* was amplified from DBS and cell pellets collected from the LTS in 2004 and 2010 the three amino acid deletion described by McCormack et al. (2006) was not present. Although it is impossible to determine how or when an error occurred it is very likely that an alignment error was made at an early stage of multiple alignment assembly of the relevant sequences in the original study and the three amino acid deletion in *gag* p17 was the result this error. This work highlights some of the pitfalls associated with sequence analysis and

serves as a stark reminder of the dangers of such errors when handling large numbers of sequences. All affected sequences from McCormack et al. (2006) that were submitted to GenBank have been re-examined and the correct sequences re-deposited. Furthermore, many HIV-1 sequences within the LANL sequence database contain single nucleotide insertions or deletions that are most likely errors due to poor quality control during the sequencing and a chromatograph editing process. Therefore care must be taken when relying on sequences from databases as it is unknown how many other possible errors may have been left undetected**.**

As the HIV-1 epidemic shows no sign of abating, both worldwide and in sub Saharan Africa, it is increasingly important to collect data on long-term survivors infected with subtype C virus. In depth studies of LTS are important to produce information regarding viral and host factors that correlate with control of infection. This has implications for the development of more effective vaccines and therapies. The long-term solution maybe to increase the amount of technology in Karonga and train local technicians in nucleic acid extraction procedures and PCR protocols. This investment would have a positive effect on the development of indigenous research and would benefit the community by up-skilling local laboratory technicians and researchers. It would also empower them to take responsibility with authority, for an epidemic that directly affects them and their local community.

Despite the problems discussed above, new sequence data was generated from HIV-1 subtype C infected long-term survivors from Malawi, Africa. Many amino acid changes were seen in *env* C2-V3 in LTS over time (1980s, 1990s, 2004 and 2010). A number of different patterns of divergence were observed in *gag* over time in the LTS, and within *env* a large amount of diversity was seen within a single time point for some of the LTS (e.g. LTS 17). In order to gain a more accurate and comprehensive picture of divergence and diversity within the LTS multiple clones from multiple time points were sequenced (Chapter 3). Two LTS who are possibly superinfected with two different viruses were also identified. The two 1989 sequences (*env* and *gag*) retrieved from LTS21 appeared very divergent from the sequences retrieved at later time points (1999 and 2010). While sample mislabelling of the DBS is very unlikely, as the individual's name was written on the filter paper as well as a unique identifier, the possibility cannot be excluded. The multiple *env* sequences from 2010 also

appeared to come from a highly divergent population of viruses and did not group together on the phylogenetic tree. LTS1 also appeared to have two divergent populations of viruses in 2010. PCR contamination is also a possibility; however the negative control in the PCRs showed no amplification products. The population of viruses within the two individuals is explored further by the sequencing multiple clones in Chapter 3.

While no clear association between mutations and survival could be shown, amino acid changes that are present in a number of LTS may in the future be shown to be important for survival but future work in this regard will require data from virus and host. Combining the sequence information with replicative fitness assays may have enabled us to draw connections between certain amino acid changes seen and the affect those changes may have had on the fitness of the virus as it is thought that the efficiency of HVI-1 replication may map to the *env* gene (Ball et al., 2003). Unfortunately, the work surrounding the replicative fitness within this project was affected by a number of setbacks.

Future work on the replicative fitness of the viruses within this cohort of subtype C LTS is of great interest. These individuals live in a resource poor area where HIV-1 is the principal cause of premature adult death and the most important co-factor for the development of several diseases such as TB and pneumococcal pneumonia (Floyd et al., 2008). With limited access to health care and poverty these individuals showed no signs of disease progression for at least 16 years without ART and some continue to show no progression over twenty years after initial infection. Understanding the factors behind their survival will no doubt shed some light on the complicated interactions between the virus and host, which can then be used to develop an effective vaccine and therapeutic interventions.

Chapter 3: Molecular Characterisation of HIV-1 Subtype C infected LTS in Karonga District Malawi

3 Chapter 3

3.1 Introduction

HIV/AIDS continues to be one of the most significant infectious diseases globally. The clinical course of infection is variable between different human hosts. Some patients such as long-term survivors can remain asymptomatic for over 15 years without anti retroviral therapy(Learmont et al., 1992). The pathogenic mechanisms that underlie HIV-1 infection are complex and highly variable, and depend on the interplay between numerous viral and host factors. A study by Shankarappa et al. (1999) followed nine HIV-1 positive men for between six to twelve years of infection. Sampling began eight months or less between their last seronegative and first seropositive visit and samples were collected from each individual over an average of 12 time points with approximately 12 *env* sequences generated from each time point. This detailed study identified a number of patterns of diversity and divergence during HIV-1 disease progression. After infection with a single virus (Keele et al., 2008), HIV-1 begins to immediately gather mutations during viral replication due to the lack of proof reading by reverse transcriptase and a high rate of viral turn over (Williams et al., 2011). Divergence within the nine individuals increased linearly for several years after seroconversion, but then appeared to slow or stabilize late in infection. The breadth of viral population diversity at a given time point increased in parallel with divergence for a few years after seroconversion, before reaching a peak and then levelling off or decreasing prior to point of divergence stabilization (Shankarappa et al., 1999). Little is known about the viral diversity and divergence within in individuals who remain asymptomatic for longer periods of infection such as the LTS in Chapter 2 who have been alive and HIV-1 positive for over 20 years. A number of studies have recorded higher levels of diversity within non-progressors when compared to normal progressors indicative of immunocompetence within the LTNPs with the resulting effect of rapid evolution by the virus to avoid the immune system (Delwart et al., 1994; Wang et al., 1997; Wang et al., 2000). Contrary to this, other studies have found much lower levels of diversity in non-progressors suggesting minimal replication and evolution of the virus. This is perhaps indicative of infection with a virus with low replicative fitness resulting in disease non-progression in the infected individual (Bailey et al., 2006; Joos et al., 2005).

A study by McCormack et al. (2006) reported the presence of a three amino acid deletion within the p17 region of *gag* in 15 LTS in Karonga and speculated that this deletion may be associated with non-progression. While the presence of this deletion later proved to be inaccurate (Chapter 2) this cohort of LTS in still of huge importance and interest because currently there is still no consensus on the most relevant mechanism involved in non-progression, in part because LTNPs form a very heterogeneous population (Saez-Cirion et al., 2007). This cohort is unique in that it is the only group of Subtype C survivors in sub Saharan Africa who have been followed for over 20 years (McCormack et al., 2006; Chapter 2). These individuals were first identified during a survey in Karonga in the 1980s where approximately 200 people were identified as being HIV-1 positive (Glynn et al., 2001). In the 1990s, the 200 HIV-1 positive individuals were followed up at which point 38 were still alive. These 38 were followed up again in 2004 and 17 were found to still be alive and then again in 2010 when nine were still alive (McCormack et al., 2006). These individuals were only discovered due to the longitudinal nature of the studies being carried out by KPS in Karonga, and were identified as LTS retrospectively. Therefore samples are available for only four time points for most individuals. As it is, there are only a few cohorts of survivors who have been followed for 20 years and those that are, are found in Europe, America and Australia with the majority focused on individuals infected with subtype B (Birch et al., 2001; Kloosterboer et al., 2005; Lambotte et al., 2005; Lopez et al., 2008; Migueles et al., 2008). Very little information is available in the literature on non-progression in Africa and even less information on subtype C non-progression in Africa. Tzitzivacos et al. (2009) studied non-progression in a subtype C cohort of children and Archary et al. (2010) explored the diversity in adult subtype C infected slow progressors; however, the average follow up was less than two years. Laeyendecker et al. (2009) and Fang et al. (2004) studied long-term survivors in Uganda and Nairobi respectively; however neither of these studies were based on individuals infected with subtype C. The extensive period of follow up on the LTS in Karonga has allowed us to examine changes in the subtype C viral population within each individual over 20 years. LTNPs serve as important models for effective immunologic control of HIV-1, and can provide clues to natural therapeutics and possible clues to therapeutic vaccines for HIV-1. For these reasons, it is of great importance to continue to study both the host factors and the complex viral populations involved in non-progression. Within this chapter, I aim to look at the viral diversity and divergence over time in the LTS

by sequencing multiple clones of *env* and *gag*. I also will investigate further the possibility of superinfection within two of the LTS (Chapter 2) by sequencing multiple clones.

Co receptor tropism in HIV-1 has been extensively studied due to its apparent link to pathogenesis and more recently due to the introduction of CCR5 antagonists as a form of anti retroviral therapy (Dorr et al., 2005). Most of these studies have been carried out on subtype B isolates. In the majority of HIV-1 infections, CCR5 tropic viruses are found to predominate during early infection (van't Wout et al., 1994; Xiao et al., 1998b; Zhang et al., 1993). Subsequently in approximately 50 % of subtype B infections CXCR4 emerges and is often associated with accelerated loss in $CD4^+$ T cells and progression to clinical AIDS (Jekle et al., 2003; Richman and Bozzette, 1994; Schuitemaker et al., 1992; Xiao et al., 1998b). When compared to CCR5 viruses, CXCR4 viruses show increased cytopathicity *in vitro* (Glushakova et al., 1998), which may account for the link between co receptor switching and increased pathogenicity of HIV-1 *in vivo* (Connor et al., 1997; Fauci, 1996).

Studies on subtype C have reported that viral isolates almost exclusively use the CCR5 co receptor with CXCR4 usage being only very rarely observed even within individuals with more advanced disease progression. By 2008, less than 30 CXCR4-utilizing subtype C viruses had been isolated despite the fact subtype C represents over 50 % of HIV-1 infections worldwide (Cecilia et al., 2000; Choge et al., 2006; Cilliers et al., 2003; Engelbrecht et al., 2001; Michler et al., 2008; Morris et al., 2001; Papathanasopoulos et al., 2003; Ping et al., 1999). More recently a study by Connel et al. (2008) found that 30 % of the subtype C viral isolates retrieved from 19 individuals with advanced AIDS could efficiently utilize CXCR4 and exhibited the syncytium-inducing phenotype *in vitro*. The V3 loops of the CXCR4 tropic isolates had at least one basic amino acid residue at positions 11 or 25 and were all predicted by C-PSSM to use CXCR4 (Jensen et al., 2006). A second study by Raymond et al. (2010) found that 29 % of the subtype C viruses from 52 HIV-1 infected individuals on HAART were able to use CXCR4. Both phenotypic and a number of genotypic assays were used to identify the CXCR4 tropic viruses.

Co receptor tropism can be established by functional assays, or infection of cell indicator lines (Jensen et al., 2003), however these methods are both labour intensive and expensive.

Subtype C infects some of the poorest regions of the world making it in many cases not economically or practically possible to obtain an active virus sample to test *in vitro*. Alternatively, bioinformatic methods can be utilised to analyze the amino acid make up of the V3 loop as it is widely acknowledged to play a major role in the co receptor switch (Cann et al., 1992; Jensen et al., 2006; Milich, Margolin, and Swanstrom, 1997). To date the most used genotypic predictor is the 11/25 rule, which looks for the presence of positively charged amino acids, arginine or lysine at positions 11 and /or 25 of the V3 region of the *env*elope protein to distinguish between syncytium-inducing viruses (CXCR4) or non-syncytium inducing viruses (CCR5) (de Jong et al., 1992; Fouchier et al., 1992). More recently the 11/25 rule has been replaced by more informed tools such as SVM_{wetcat} (Pillai et al., 2003), PSSM (Jensen et al., 2003), Geno2pheno (Sing et al., 2007), the Briggs method (Briggs et al., 2000) and R5X4-red (Xu et al., 2007), however these are generally based on subtype B viruses. Jensen et al (2006) has recently introduced C-PSSM, which has been trained on subtype C sequences and has been shown to be more accurate at predicting the presence of CXCR4 viruses in subtype C isolates and is thought to perform comparably well on subtype C V3 tropic prediction as existing subtype B specific methods do on subtype B V3 tropic predication (Jensen et al., 2006). Using some of these bioinformatic tools, my aim was to look at the co receptor tropism within the LTS in Karonga over multiple time periods, which will allow me to identify any switch in co receptor usage as some individuals begin to progress.

This chapter focuses on the cohort of LTS from the Karonga district of northern Malawi previously described in Chapter 2. Phylogenetic analysis of *env* and *gag* has identified 11 of the 12 LTS as subtype C and one individual as an Unclassifiable subtype (McCormack et al., 2002; Seager et al., 2011). Different amounts of genetic divergence were seen among the LTS within *gag*. A number of individuals showed an increase in viral divergence between 1989 and 2004-2010 while two individuals showed a decrease in viral divergence over time. Within *env* a comparison of genetic divergence among the LTS was not explored in chapter 2 due to the large amount of diversity seen within consensus sequences retrieved from the 1990s. Further more, two individuals were suspected to be superinfected due to the amount of divergence seen within each individual resulting in different consensus sequences from

different time points and within time points grouping separately on a phylogenetic reconstruction of *env* and *gag*.

Within this chapter I will further explore,

(a) the evolution of the viral population within the Karonga LTS, including both viral genetic diversity and divergence through the sequencing of multiple clones over a number of time points,

(b) the possibility of superinfection within two LTS and,

(c) the co receptor tropism of the viral populations using new bioinformatic tools to ascertain if any major changes within the V3 section of *env* indicate a co receptor switch and if so at what time point.

3.2 Materials and methods

3.2.1 Patients and samples

Dried blood spot (DBS) samples were collected from 10 of the 17 long-term survivors in 2004 and stored at -20 °C. Fourteen of these were sought again in 2010 (three were not sought as they had refused to participate on the previous two occasions), at which point it was found that three had died (LTS9 – 2008, LTS22 – 2005, LTS25 – 2004), and one had left the region (LTS17). By this time five LTS had been placed on ART (LTS2, 5, 8, 12 and 20) and two other individuals, LTS10 and LTS21 had been referred for ART due to low $CD4^+$ cell counts of 138 and 36 cells/mm^3 respectively. Two LTS had $CD4^+$ cell counts greater than 200 cells/mm^3 (LTS1 and LTS30) and have not been referred for ART. Whole blood cell pellet samples were collected from nine of the remaining individuals (LTS29 refused to provide a sample) and were stored at -80 °C. Two individuals (LTS2 and LTS8) were seen in unrelated studies in 2008 and whole blood cell pellet were available from that time also.

3.2.2 DNA extraction, PCR amplification, and cloning

Proviral DNA was extracted from the DBS collected in 2004 using a QIAamp DNA Micro Kit (Qiagen) or from 200 µl of cell pellets collected in 2009 and 2010 using a QIAamp DNA Blood Mini Kit (Qiagen). Nested PCR of a 750 bp region of *gag* p17p24 and a 549 bp region of *env* C2V3 was carried out as previously described (McCormack et al., 2002). Three secondary PCR products were pooled and TA cloned using the pCR2.1-TOPO Cloning Kit (Invitrogen) or using the StrataClone PCR Cloning Kit (Agilent Technologies). Approximately 20 individual clones were sequenced in one direction. Primer DT6 was used in the sequencing of *gag* and ES8 was used to generate the *env* cloned sequences (McCormack et al., 2002). Sequence chromatographs were examined for quality in Seqman 8.0.2 (DNASTAR).

3.2.3 Sequence alignment and phylogenetic analysis

Consensus sequences were available for partial *env* (C2V3 – 549 bp) and *gag* (p17-p24 750 bp) from the 1980s (McCormack et al., 2002), 1990s (unpublished data), 2004, and 2010 (Seager et al., 2011 and Chapter 2) as can be seen in Table 3.1. Multiple alignments of all LTS consensus sequences from the four time points and cloned sequences from 2004, 2008

and 2010 along with 40 other sequences randomly chosen from Karonga (20 sequences retrieved in 1990s and 20 retrieved in 2008) were assembled for *env* and *gag*. Phylogenetic trees were reconstructed under the GTR + gamma model of DNA substitution implemented by RAxML 7.0.3(Stamatakis, 2006) with all parameters optimised by RAxML. Confidence levels in the groupings in the phylogeny were assessed using 1000 bootstrap replicates as part of the RAxML phylogeny reconstruction. The subtype C ancestral sequences derived by (Travers et al., 2004) was employed as the out-group for the *env* and *gag* trees. In addition, for each LTS, multiple alignments of each gene region (*env* and *gag*) were assembled and optimised by eye in MacClade 4 (Sinauer Associates). Each alignment contained all cloned and consensus sequences from the different time points. Phylogenetic trees were reconstructed in a similar manner to that described above for that individual.

Pairwise evolutionary nucleotide distances from nucleotide sequences were computed by PAUP* 4.0 (D.L. Swofford, Sinauer Associates, Inc. Publishers) under Kimura's two-parameter model of evolution (Kimura, 1980). Intra patient genetic divergence for each time point was examined by comparing the pairwise genetic distance between the earliest sequence available and all subsequent sequences from later time points. When both cloned and consensus sequences were available from the same time point, the average pairwise genetic distance between all the sequences was calculated and used as an estimate of the amount of genetic divergence seen at that time point. The genetic divergence over time was plotted on a graph for each LTS with time on the X-axis and pairwise genetic distance on the Y-axis. The intra patient genetic diversity present at a particular time point was also plotted on the graph using error bars to mark the maximum and minimum pairwise genetic distance values seen at that time point. LTS21 and LTS1 were not included due to the possibility of super infection (see later). All averages and standard deviations were calculated in Excel (Microsoft). A Z test was calculated by hand to test the difference between the average genetic diversities.

3.2.4 Co receptor tropism prediction

Co receptor usage was determined for all *env* sequences for all of the LTS using two genotypic predictor tools. The PSSM (Jensen et al., 2003) (http://indra.mullins.microbiol.washington.edu/webpssm/) is a bioinformatic tool for predicting HIV-1 co receptor usage based

on the amino acid sequence of the V3 loop of the *env* gene, using position-specific scoring matrices (PSSM), which uses background genetic variation as a baseline comparison to facilitate comparison of the residues of a sequence fragment to those of a group of aligned sequences known to have a desired property. The C-PSSM, the subtype C specific SINSI matrix, was used to predict co receptor tropism for all the *env* sequences of the subtype C LTS, as this PSSM predictor was trained on 279 HIV-1 subtype C V3 sequences known to possess the syncytium -inducing phenotype (Jensen et al., 2006). Gene2pheno (5 % FPR) (Lengauer et al., 2007) (http://coreceptor.bioinf.mpi-inf.mpg.de/), a separate web based bioinformatic tool to predict co receptor tropism was used on any sequences, which C-PSSM failed to predict with confidence. Geno2pheno is based on a statistical learning method called a support vector machine and was trained on 1100 sequences from 332 patients. However, it is not subtype specific (Sing et al., 2007). LTS30 is infected with an unclassifiable subtype (Chapter 2) and due to this both PSSM and geno2pheno were used.

3.2.5 Constraint analysis

LTS1 and LTS21 were indicated in Chapter 2 as possible super infections. To explore this further, phylogenies were produced in MacClade 4 where relationships were constrained such that sequences from LTS1 or LTS21 were monophyletic in various arrangements. The likelihood of each phylogeny was then calculated in PAUP* 4.0. The Shimodaira-Hasegawa test was employed to ascertain if any of the reconstructed phylogenies were significantly worse than the original maximum likelihood tree.

3.3 Results

3.3.1 Sequences generated

To investigate the properties of the viral populations present within the LTS, multiple cloned sequences were generated of the *env* and *gag* fragments from samples dating from 2004, 2008 and 2010. Six samples were successfully cloned from the 10 DBS collected in 2004, and approximately 20 sequences were generated for both *gag* and *env* from each. Twenty *env* sequences were generated from the whole blood cell pellet sample collected from LTS2 in 2008 and 39 *env* sequences from LTS8 in 2008. Approximately 18-23 *env* clones were generated from nine and between 17-20 *gag* clones from eight, of the LTS cell pellet samples from 2010 (Table 3.1). Phylogenetic reconstruction of all sequences (consensus and clonal) from each LTS and 40 control sequences, for both *env* and *gag* showed all sequences from each individual to form monophyletic clusters with the exceptions of LTS1 and LTS21, which are discussed in section 3.3.5 (Appendix 4 and 5).

3.3.2 Diversity

Two different patterns of clonal diversity were identified in the *gag* clonal sequences generated from the LTS. The *gag* clonal sequences of five LTS (LTS2, 5, 9, 12 and 20) were homogenous with a pairwise genetic difference of less than 0.5 % (Table 3.2 and Figure 3.1 a-e). For each of these individuals the clonal sequences and the consensus sequences from the corresponding time period were found to group within the same clade or the consensus sequence was found as a sister clade to the clade containing the clonal sequences (Figure 3.1 a-e). The *gag* clonal sequence diversity for the remaining LTS (LTS8, 10, 17 and 22) was higher than 1 % at a particular time period (Table 3.2 and Figure 3.1 f-i). The clonal sequences generated from LTS8 and LTS10 in 2004 were less heterogeneous than what was then observed in 2010. A pairwise genetic diversity of 0.2 % in 2004 for LTS8 rose significantly to 1.2 % in 2010 ($Z=18.41$, $P<0.0001$). In LTS10, a pairwise genetic diversity of 0.6 % in 2004 rose significantly to 2.6 % in 2010 ($Z=10.75$, $P<0.0001$). The phylogenetic relationships between the clones and the corresponding consensus sequences from the same time period showed less temporal identity within these four LTS. Within LTS17 (Figure 3.1 f) the 2004 consensus sequence was found as a sister to a clade containing the 2004 clones

Table 3.1 A summary of the clonal sequence data, co receptor tropism, $CD4^+$ cell counts and ART information for each LTS.

LTS Number	Year	Consensus Sequence Co Receptor Information	Number of Cloned sequences	Cloned Sequence Co Receptor Information	$CD4^+$ T-cell Count and ART Status
LTS1	1999	R5			2004 - 586
	2010	R5	18 *env,* 17 *gag*	18 R5	2010 - 449
					no ART
LTS2	1999	X4			2010 - 244
	2009	no consensus	20 *env*	20 X4	ART - 2009
	2010	R5	22 *env,* 20 *gag*	2 R5, 20 X4	
LTS5	1998	X4 and R5			2010 -789
	2010	no consensus	22 *env*, 19 *gag*	21 R5, 1 X4	ART - 2008
LTS8	1989	R5			
	1999	R5			2004 - 265
	2004	R5	20 *env,* 30 *gag*	20 R5	2010 - 734
	2009	no consensus	39 *env*	39 R5	ART - 2005
	2010	R5	23 *env*, 19 *gag*	23 R5	
LTS9	1988	R5			2004 - 56
	1999	R5			Died in 2008 of
	2004	R5	20 *env*, 20 *gag*	20 R5	AIDS
LTS10	1989	R5			2004 - 328
	1999	R5 and X4			2010 - 138
	2004	X4	20 *env*, 20 *gag*	1 R5, 19 X4	no ART
	2010	X4	23 *env*, 20 *gag*	23 X4	
LTS12	1988	R5			2004 - 390
	1999	R5			2010 - 439
	2004	R5	20 *env*, 20 *gag*	20 R5	ART 2006
	2010	R5	20 *env*, 19 *gag*	20 R5	
LTS17	1989	R5			2004 - 452
	1998	R5			Left the region
	2004	X4	20 *env,* 20 *gag*	20 X4	
LTS20	1988	R5			
	2000	R5			2010 - 139
	2004	R5			ART - 2006
	2010	R5	18 *env*, 19 *gag*	18 R5	
LTS21	1989	R5			2004 - 47
	1999	R5			2010 - 36
	2010	R5	21 *env*, 18 *gag*	14 R5, 7 X4	no ART
LTS22	1989	R5			2004 - 475
	1998	R5			Died in 2005 of
	2004	R5	19 *env*, 20 *gag*	18 R5, 1 X4	AIDS
LTS30	2010	R5	19 *env*	19 R5	2004 - 675
					2010 - 362
					no ART

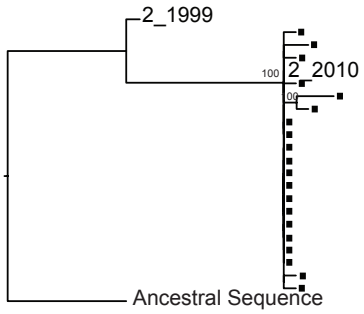
and the consensus sequence generated from the 1998 sample. The 2004 consensus sequence generated from LTS10 (Figure 3.1 i) was found within the clade containing the consensus sequences generated from the 1999 sample and all sequences generated from the 2010 sample. The 2004 clonal sequences were found in a sister clade to this larger clade.

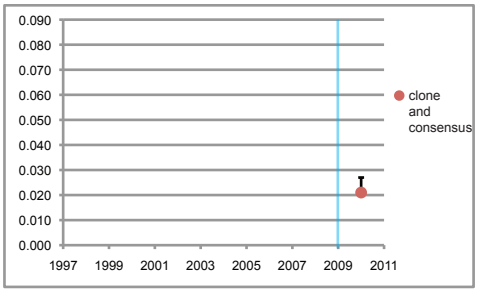
The clonal sequences of LTS2, 5, 9, and 20 were also homogenous in *env* (as well as LTS10 and LTS17) with a pairwise genetic diversity of less than 1.5 % with four of those (LTS5, 9, 17, and 20) showing a genetic diversity of 0.2 % or less within a particular time point (Table 3.2 and Figure 3.2 a-f). The consensus sequences generated from 2004 for LTS9, 10, 17 and from 2010 for LTS20 appeared almost identical to the clonal sequences generated from the corresponding time points (Figure 3.2 c-f). The consensus sequence generated from LTS10 in 2010 however, was found on a sister branch to a clade that was composed of the clonal sequences generated from 2004, 2010 and the consensus sequences from 2004 (Figure 3.2 e). A larger amount of diversity was seen in the clonal *env* sequences generated from LTS8, 12 and 22 (Table 3.2 and Figure g-i) with pairwise genetic distance of 1.8 % up to 4 % within a time period. The amount of diversity seen in LTS 12 in 2004 reached 1.8 %, but by 2010 the quantity of diversity detected was only 0.1 % ($Z=18.41$, $P<0.0001$) (Figure 3.2 g). Sequences generated from LTS8 in 2004 and 2008 had a pairwise genetic distance of 2.8 % and 4 % respectively, however the clonal sequences generated in 2010 were identical (Figure 3.2 h). These clones generated from 2010 were found in a single clade, however the 2010 consensus sequence was more closely related to a clonal sequence from 2008.

3.3.3.2 Divergence

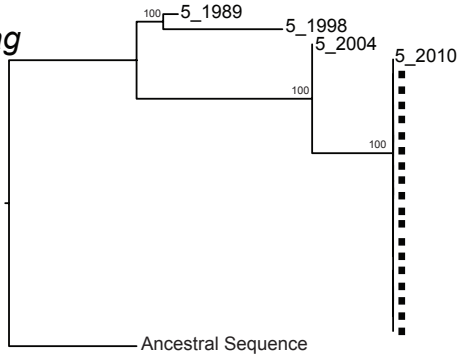
Genetic divergence in *gag* and *env* was calculated by comparing the pairwise genetic distance between the earliest consensus sequence for each LTS to each subsequent sequence from the 1990s, 2004 and 2010. Where more than one sequence for a time point was available the average was used (Table 3.3). Seven of the nine individuals showed a overall increase in genetic divergence between the first sampling time point in 1988/89 and the last sampling time point of either 2004 (LTS9, 17 and 22) or 2010 (LTS2, 5, 8, and 10) as is consistent with previous reports on HIV-1 infection (Shankarappa et al., 1999) (Figure 3.1, Table 3.2). However, due to the limited number of sequences generated in the 1990s and in 2004,

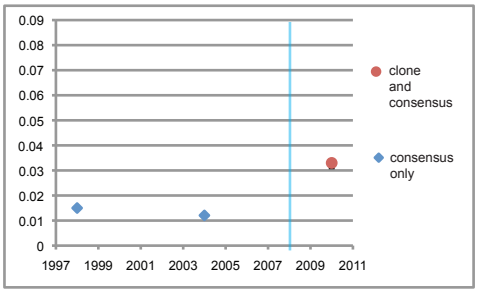
(a) LTS2 *gag*



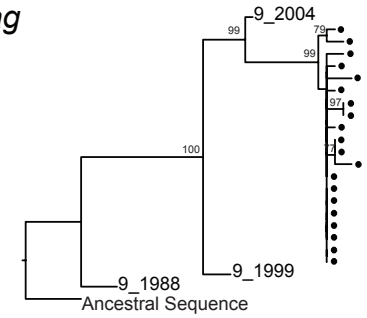


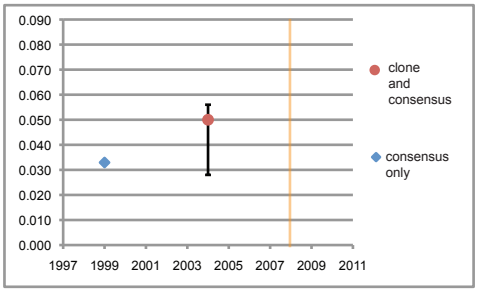
(b) LTS5 *gag*



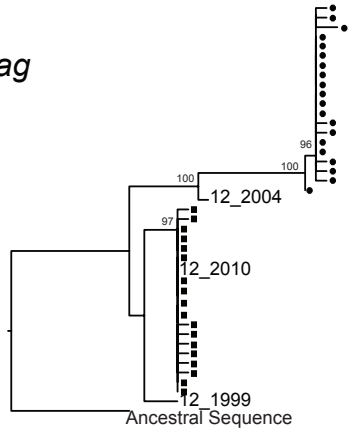


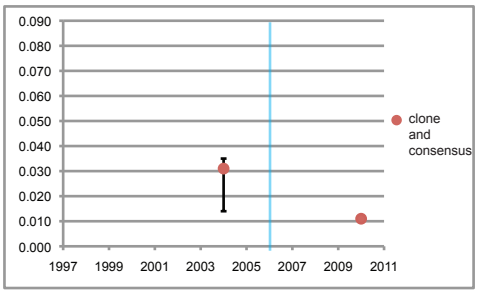
(c) LTS9 *gag*



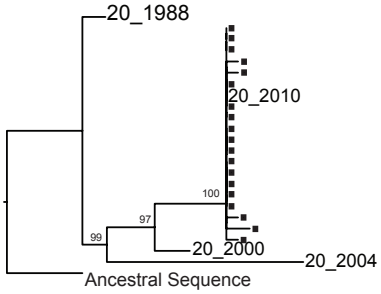


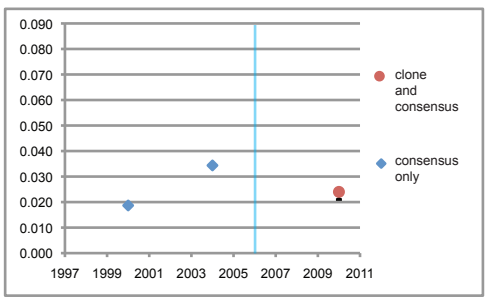
(d) LTS12 *gag*





(e) LTS20 *gag*





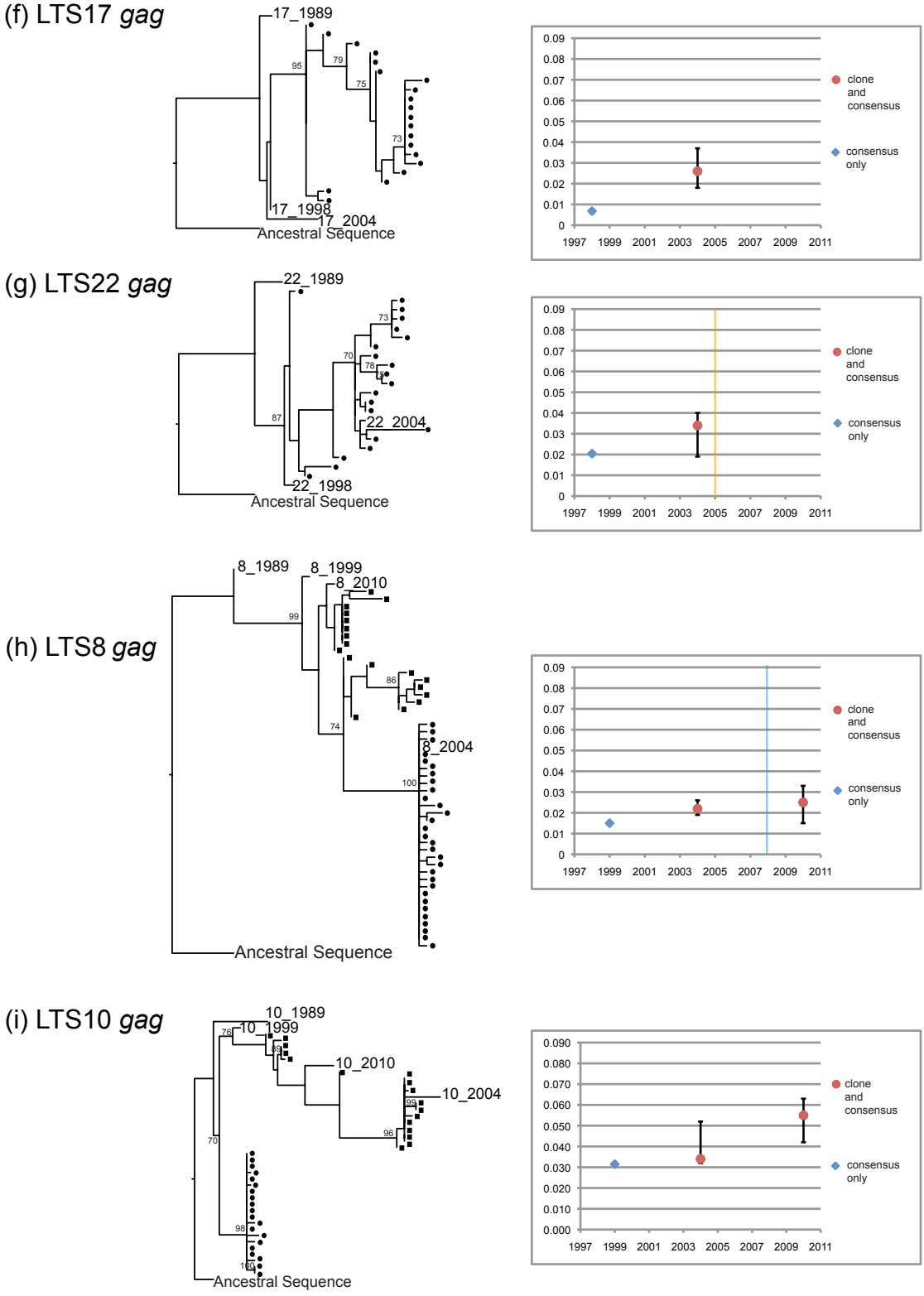
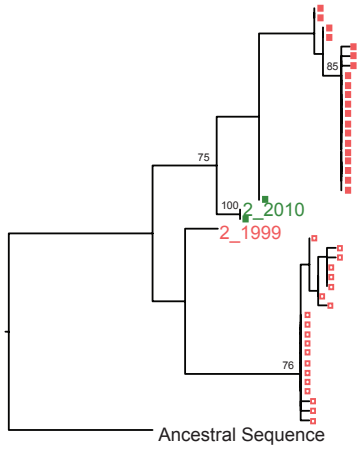
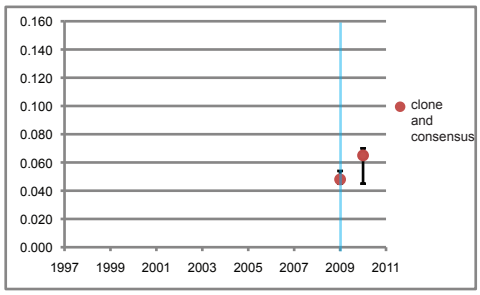


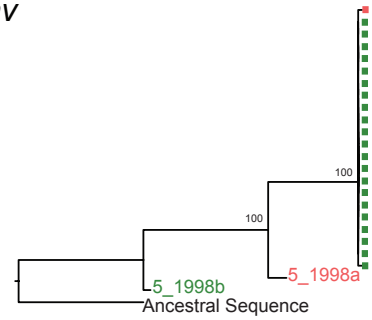
Figure 3.1 Individual Maximum Likelihood trees generated from all *gag* sequences for each LTS. Branches containing consensus sequences are labelled with the LTS number followed by the year of sampling. ● represents clones dating from 2004, ■ represent clones dating from 2010. Bootstrap values of over 70 are marked on relevant branches. The graphs show the genetic distance between each time point and the sequence generated from the earliest time point available. The error bars mark the max and min genetic distance at that time point. Time in years is on the X-axis and the genetic distance is on the Y-axis. Blue vertical line = ART start date. Orange vertical line = Date of Death

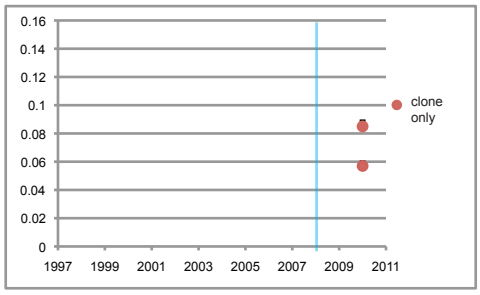
(a) LTS2 *env*



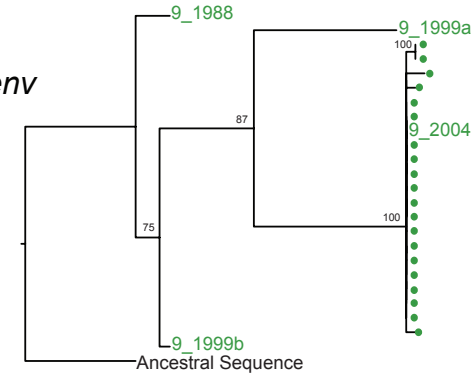


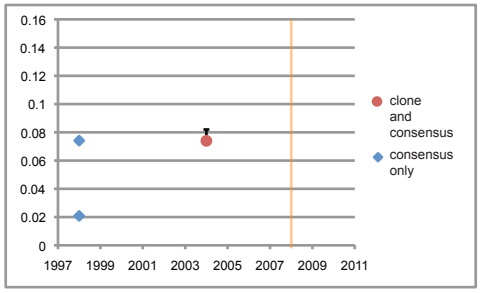
(b) LTS5 *env*



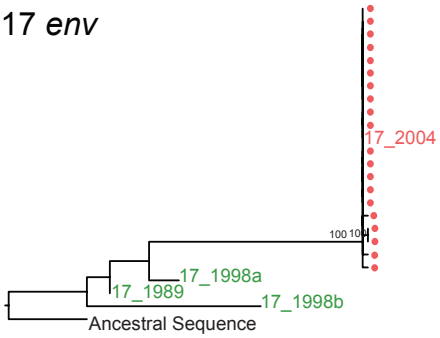


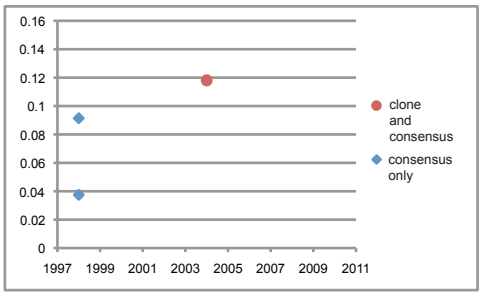
(c) LTS9 *env*



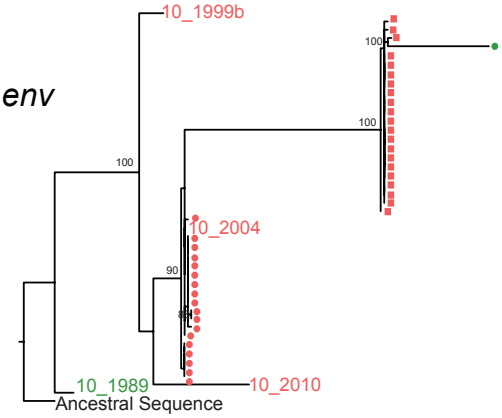


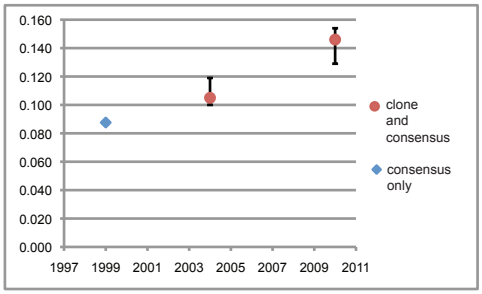
(d) LTS17 *env*



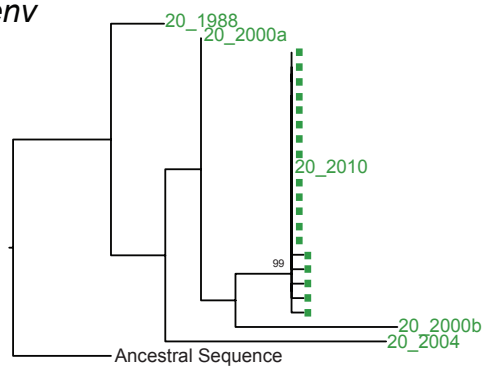


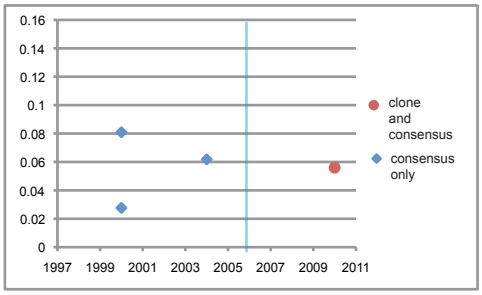
(e) LTS10 *env*





(f) LTS20 *env*





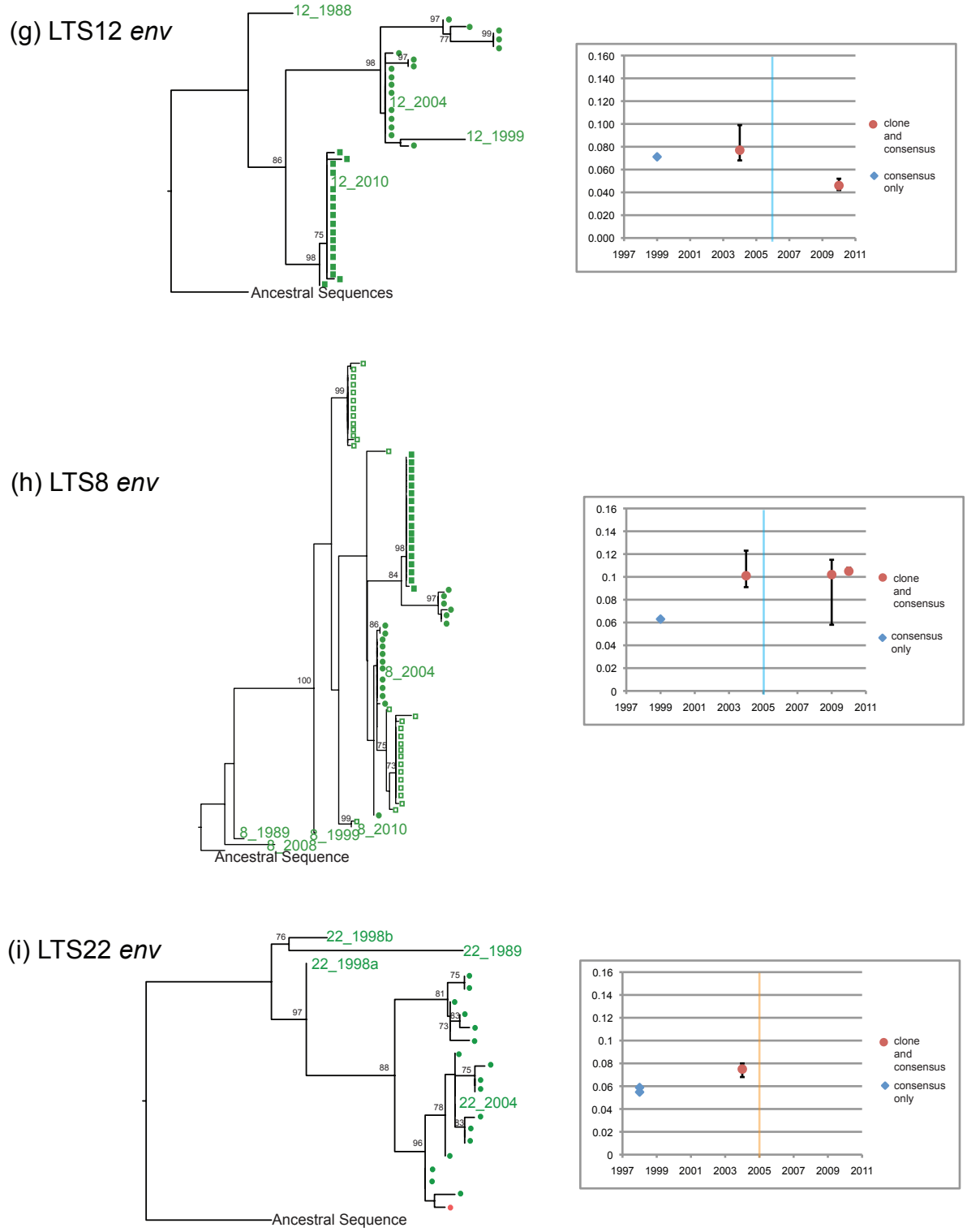


Figure 3.2 Individual Maximum Likelihood trees generated from all sequences for each LTS. Branches containing consensus sequences are labelled with the LTS number followed by the year of sampling. ● represent clones dating from 2004, □ represent clones dating from the 2008, ■ represent clones dating from 2010. Green = sequences predicted to use CCR5, red = sequences predicted to use CXCR4. Bootstrap values of over 70 are marked on the relevant branches. The graphs show the pairwise genetic distance between each time point and the sequence generated from the earliest time point available. The error bars mark the max and min pairwise genetic distance at that time point. Time in years is on the X-axis and the pairwise genetic distance is on the Y-axis. Blue vertical line = ART start date. Orange vertical line = Date of Death

Table 3.2 Average pairwise genetic diversity for all LTS sequences generated for *gag* and *env*. The standard deviation is in brackets after the average.

LTS Number	*env*			*gag*	
	2004	2008	2010	2004	2010
LTS2		0.6% (SD=0.4%)	1.2% (SD=1.7%)		0.2% (SD=0.2%)
LTS5			0.1% (SD=0.1%)		0.1% (SD=0.2%)
LTS8	2.8% (SD=3%)	4% (SD=3.7%)	0.6% (SD=1.9%)	0.2% (SD=0.2%)	1.2% (SD=0.7%)
LTS9	0.2% (SD=0.2%)			0.5% (SD=0.4%)	
LTS10	2.1% (SD=5%)		1.4% (SD=4.3%)	0.6% (SD=1%)	2.6% (SD=2.1%)
LTS12	1.8% (SD=1.8%)		0.1% (SD=0.2%)	0.3% (SD=0.5%)	0.1% (SD=0.1%)
LTS17	0.1% (SD=0.2%)			1.1% (SD=0.8%)	
LTS20			0.2% (SD=0.2%)		0.1% (SD=0.1%)
LTS22	1.9% (SD=1.3%)			1.5% (SD=0.9%)	

Table 3.3 Average pairwise genetic divergence in *gag* and *env* between the sequence from the earliest time point and all sequences (cloned and consensus) available from each subsequent time point. The two genetic divergences for LTS8, 17, 20 and 22 in the 1990s represent two different consensus sequences for that time point. The two genetic divergences for LTS5 in 2010 represent two consensus sequences available from the 1990s and as this was the earliest time point with sequence information this was used to calculate the amount of genetic divergence by 2010.

LTS Number	*env*				*gag*		
	1990s	2004	2008	2010	1990s	2004	2010
LTS2			4.8%	6.5%			2.1%
LTS5				5.7%, 8.5%	1.5%	1.2%	3.3%
LTS8	6.3%	10.1%	10.2%	10.5%	1.5%	2.2%	2.5%
LTS9	7.4%, 2.1%	7.4%			3.3%	5.0%	
LTS10	8.8%	10.5%		14.6%	3.1%	3.4%	5.5%
LTS12	7.1%	7.7%		4.6%		3.1%	1.1%
LTS17	3.7%, 9.1%	11.9%			0.7%	2.6%	
LTS20	2.8%, 8.1%	6.2%		5.6%	1.9%	3.4%	2.4%
LTS22	5.9%, 5.5%	7.5%			2.0%	3.4%	

statistical significance could not be calculated. In LTS 20 an overall increase in genetic divergence in *gag* was seen in the first 16 years of infection (1988-2004). This was subsequently followed by a drop in divergence between 2004 and 2010 (Figure 3.1 e) resulting in a cumulative overall increase in divergence of just 0.5 % between 2000 and 2010. This measure, and thus any trend seen, could possibly be altered if a different consensus sequence had been sampled at each time. LTS12 displayed a decrease in the amount of divergence over time in both genes. A small rise of divergence in *env* of 0.06 % was seen between 1999 and 2004 before a fall of 3.1 % between the years 2004 and 2010 (Table 3.3, Figure 3.2 g).

3.3.4 Prediction of co receptor tropism

LTS30, an unclassifiable subtype (Chapter 4), was predicted as being R5 tropic by PSSM and by Geno2pheno. All *env* sequences from five of the subtype C LTS (LTS1, 8, 9, 12 and 20) were predicted to be CCR5 tropic at all time points where sequence information was available (Table 3.1). Six individuals (LTS2, 5, 10, 17, 21 and 22) were found to show a switch from using CCR5 to CXC4 at some point in time. CXCR4 constituted more than 90 % of the sequenced viral population within three individuals (LTS2, 10, 17) at the time of the last sampling time point (Figure 3.2 a, d, and e).

All sequences collected in 1988/89 were predicted to be CCR5 tropic in all patients. By the next time point of 1998-2000, CXCR4 tropic sequences was detected within three individuals (LTS2, 5, and 10) (Figure 3.2 a, b, and e). The earliest sequence for LTS2 (1999) was predicted to be CXCR4 tropic (Figure 3.2 a). By the next sampling time point (2008), 11 of 20 individual clones were predicted to use CXCR4. While the consensus sequences from 2010 and two of the cloned sequences from the same time point were predicted to use CCR5, the remaining clones from this time point were predicted to use CXCR4 (but without confidence) in conjunction with Geno2pheno. These latter clones, along with all clones from 2008, had an arginine at position 11 of the V3 loop, a basic amino acid often associated with a switch in co receptor usage (Fouchier et al., 1992). The phylogenetic reconstruction (Figure 3.2 a) of the *env* C2-V3 fragment from LTS2 showed two viral lineages, one containing all CXCR4 tropic viruses from 1999 and 2008, and the other showed a clade of CXCR4 sequences from 2010 having emerged from a CCR5 ancestral strain. Both CCR5 and

CXCR4 strains were present in 2010, at which point LTS2 had a $CD4^+$ cell count of 244 and had been on ART since 2008.

Viral sequences from LTS10 were predicted to be CCR5 tropic in 1989 (Figure 3.2 e). The *env* V3 sequences retrieved from the next time point in 1999, was predicted to be CXCR4 tropic. All sequences from the following time points, 2004 and 2010, were predicted to use CXCR4 with the exception of one clone retrieved from the 2004 sample, which was predicted to be CCR5 tropic. This single 2004 CCR5 sequence shared a most recent common ancestry with CXCR4 clones from 2010. The emergence of CXCR4 dominance was not accompanied by positively charged amino acid mutations at position 11, 24 or 25.

The V3 sequences obtained from LTS17 in 1989 and in 1998 were both predicted to be CCR5 tropic (Figure 3.2 d). By 2004 both the consensus sequence and 20 cloned sequences were predicted to use CXCR4. All V3 sequences from 2004 had acquired the positively charged amino acid Lysine at position 25, which is often associated with a co receptor switch to CXCR4 usage (Xiao et al., 1998a). No further follow up information was available for LTS17, as the individual had left the Karonga region by 2010.

One sequence from 1998 was predicted to use CCR5 in LTS5 (Figure 3.2 b). This sequence was a sister sequence to a cluster, which diverged into two lineages emerging from a common node, one lineage containing a second sequence dating from 1998 predicted to use CXCR4. The second lineage was composed of cloned sequences dating from 2010, of which the majority were CCR5 tropic; however, one clone was predicted to use CXCR4. It is likely that the common ancestor of these two lineages was CCR5 tropic and CXCR4 emerged independently within the two lineages.

V3 sequence data was available from LTS22 (Figure 3.2 i) from 1989, 1998 and 2004, all of which were predicted to utilise CCR5 except one of 19 clones from 2004. The one clone that was predicted to use CXCR4 contained a positively charged arginine at position 11. No sequence data was available for 2010, as the patient had died of AIDS in 2005.

3.3.5 Possible superinfections

Previous phylogenetic analysis of consensus sequences of the twelve LTS still alive in 2004 and or 2010 revealed that all LTS formed monophyletic groups with the exception of LTS1 and LTS21, which separated into more than one group situated on different branches on the *env* phylogenetic tree (Chapter 2) (Figure 3.3). This was also seen for LTS21 on the *gag* tree (Figure 3.4). This suggests that a possible HIV-1 superinfection has occurred within these two patients.

Constraint analyses of LTS1 showed that neither of the two *env* trees, where all sequences generated from LTS1 were forced to group together, were significantly worse than the original tree as seen in Table 3.5. All V3 sequences from LTS1 were predicted to use CCR5. Four different *env* phylogenies were created for the constraint analysis on LTS21, of which only one proved to be significantly worse than the original maximum likelihood tree. This tree involved grouping two of the sequences generated in 2010 (Clade 2, Figure 3.3) with the two sequences generated from 1989 (Clade 1, Figure 3.3). All three alternative *gag* topologies, which placed all sequences from LTS21 in a monophyletic grouping, were found to be significantly poorer representations of the data than the original maximum likelihood tree (Table 3.5) making it likely that this individual is infected by more than one strain.

The majority of the *env* clones from LTS21 generated from the 2010 sample grouped within a clade with two consensus sequences from 2010 (Figure 3.5 b). This clade is equivalent to Clade 2 in Figure 3.3. One clone and the remaining two consensus sequences from 2010 are found in a sister clade to this latter grouping. These two consensus sequences are seen to group together also in a Clade in Figure 3.3 (Clade 3). A single clone is found in-between these two groupings on the tree. No clones are found to group with the 1989 or 1999 consensus sequences. All the V3 consensus sequences available from 1988, 1999 and 2010 were predicted to be CCR5 tropic, however seven of the 20 clones from 2010 were predicted as CXCR4 tropic while the rest were CCR5 tropic. The majority of the clones predicted to use CXCR4 grouped together with a CCR5 tropic consensus sequence from 2010. One CXCR4 tropic clone grouped with one consensus sequence from 2010 and within the clones predicted to use CCR5 and (21_2010c). It is likely that the common ancestor of these two lineages was CCR5 tropic and CXCR4 emerged twice within the population (Figure 3.5 b).

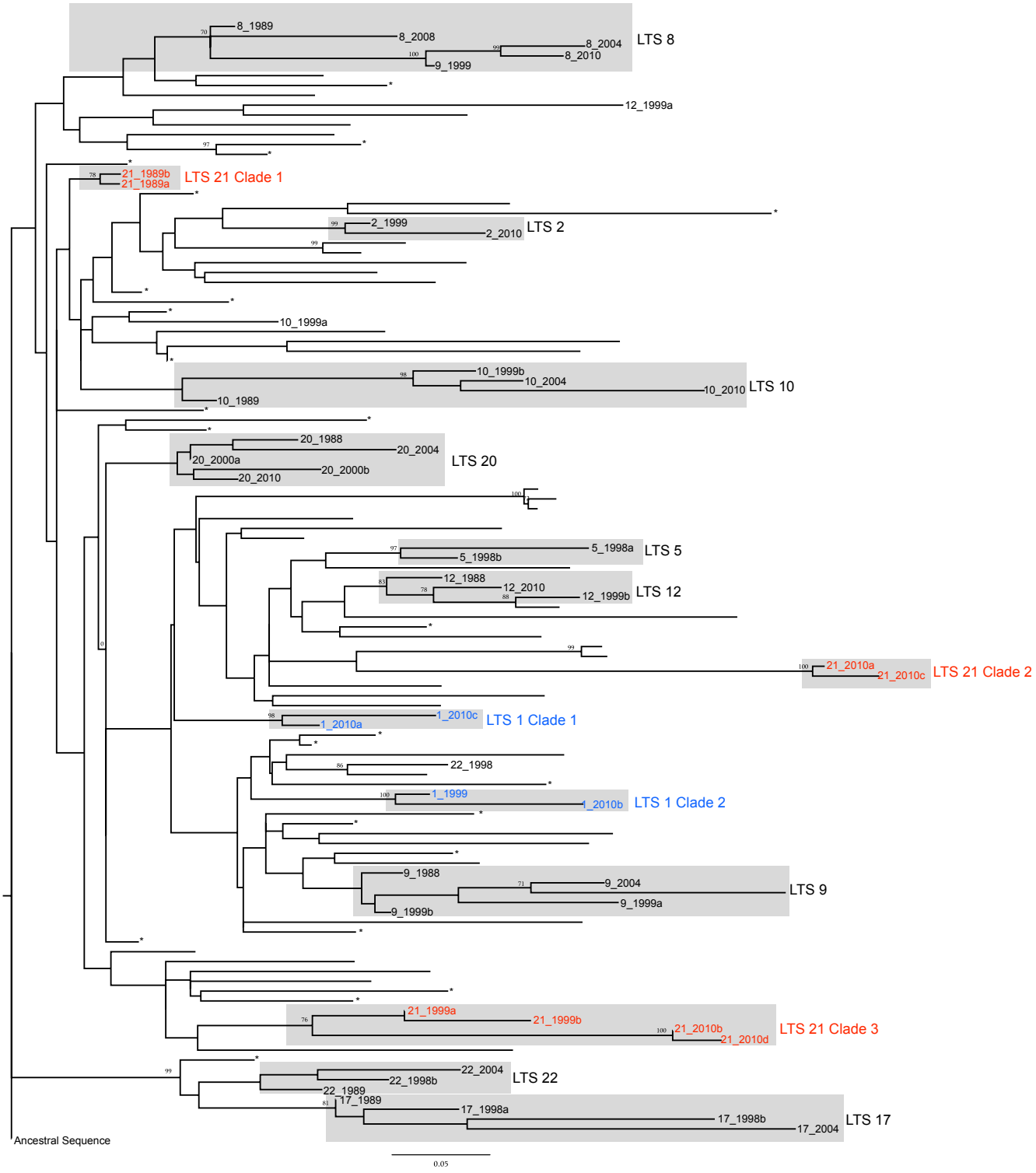


Figure 3.3 Maximum Likelihood tree constructed from *env* sequences altered from chapter 2. LTS1 is labelled blue and LTS21 is labelled red with each separate clade numbered.

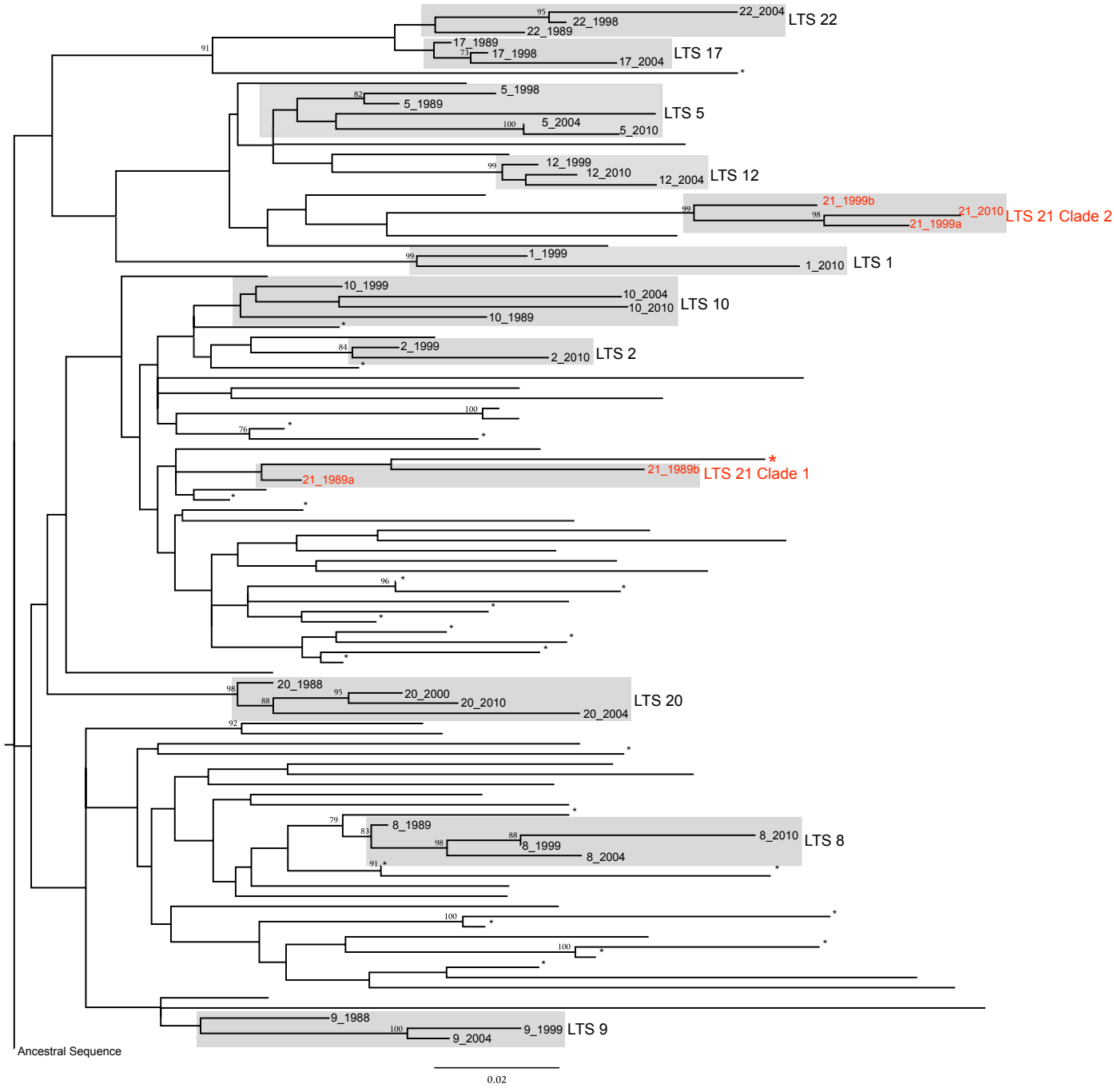


Figure 3.4 Maximum Likelihood tree constructed from *gag* sequences altered from chapter 2. LTS21 is labelled red with each separate clade numbered. The non-related sequence grouped in Clade 2 with LTS21 is labelled with an *.

Table 3.5 Constraint analysis results carried out on the maximum likelihood trees from Figure 3.3 and Figure 3.4 for LTS1 and LTS21.

Tree	Likelihood score	p-value	Description of tree
LTS1 *env*			
Original tree	8414.17	Best	Original ML tree, LTS1 forms two separate clusters (Figure 3)
Tree 1	8425.47	0.087	Clade 1 grouped with Clade 2
Tree 2	8426.99	0.070	Clade 2 grouped with Clade 1
LTS21 *env*			
Original tree	8414.19	Best	Original ML tree, LTS21 forms three separate clades (Figure 4)
Tree 1	8422.66	0.466	Clade 2 grouped with Clade 3
Tree 2	8436.83	0.094	All three Clades grouped together
Tree 3	8428.66	0.228	Clade 1 grouped with Clade 3
Tree 4	8448.39	0.009*	Clade 1 grouped with Clade 2
LTS21 *gag*			
Original tree	8256.43	Best	Original ML tree, LTS21 forms two separate clades (Figure 3)
Tree 1	8296.69	0.006*	Clade 1 and a third epidemiologically unlinked sequence within the clade grouped with Clade 2
Tree 2	8320.85	0.007*	The two 1989 sequences in Clade1 moved one at a time to Clade 2
Tree 3	8296.73	0.015*	The two 1989 sequences were separated from the third unrelated sequence grouped with Clade 2

*p<0.05

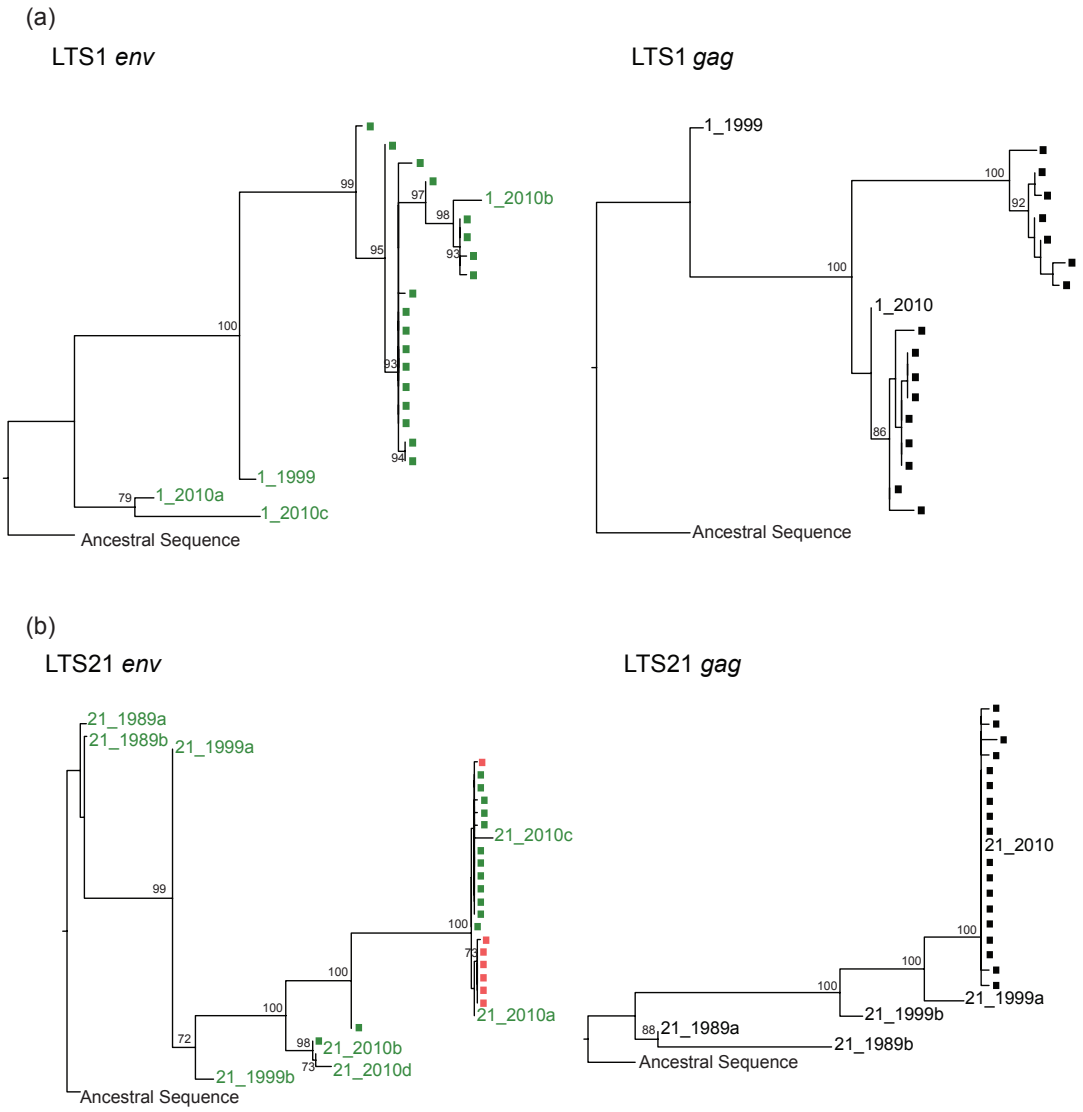


Figure 3.5 Individual Maximum Likelihood trees generated from all sequences for each LTS. Branches containing consensus sequences are labelled with the LTS number followed by the year of sampling. ● represent clones dating from 2004, ■ represent clones dating from 2010. Green = sequences predicted to use CCR5, red = sequences predicted to use CXCR4. Bootstrap values of over 70 are marked on the relevant branches.

A phylogenetic tree was reconstructed for *env* and *gag* with all sequences (clonal and consensus) from the LTS and 40 subtype C control sequences from Karonga (Appendix 4 and 5). In this *env* tree (Appendix 4), the viral sequences generated from LTS1 separated into two clades. One clade contained the sequence dating from 1999, one consensus sequence from the 2010 and all of the cloned sequences from 2010. The second contained two consensus sequences generated from the 2010 sample. In the *gag* phylogenetic tree (Appendix 5) all available nucleotide sequences from LTS1 were found in one clade. In both the *env* and *gag* phylogenetic reconstructions all the cloned and the consensus sequences from 2010, and the two consensus sequences dating from 1999 formed a monophyletic group that was highly supported for LTS21. However, in both trees the two sequences generated from the 1989 sample grouped on a different branch of the tree (Appendix 4 and 5).

3.3.6 Amino acid translation

No large indels were seen in either gene region. Premature stop codons were found within two individuals and were the result of a mutation of TGG (tryptophan) to TAG. The first was found in LTS5 within the C2 region of *env* in one of the consensus sequences recovered from the 1998 sample and all the cloned sequences from 2010. Two stop codons were found in one consensus sequence retrieved from LTS21 in 1989 within the p24 region of *gag*. The GPGQ motif typically found in subtype C (Ou et al., 1992) at the crown of the V3 loop in *env* was conserved in seven of the LTS. Variations of the tetrapeptide motif were seen in LTS1 where an H (histidine) was in place of the Q (glutamine), and LTS20 had an R (arginine) in place of the Q in a single clone from 2010. LTS5 had an R in place of the first G (glutamine) in one of the consensus sequences retrieved from 1998. The cloned sequences retrieved from LTS2 in 2008 showed two different variations of the GPGQ motif, the first was found in 12 of the 20 clones and contained one mutation, GPGR. The second was found in the remaining 8 clones and contained two mutations, GTGR. By 2010 GTGR dominated the cloned sequences population and was found in 20 of the 22 sequences, the remaining two sequences were the characteristic subtype C motif of GPGQ.

3.4 Discussion

Patterns of viral genetic diversity and divergence has previously thought to be important in the monitoring of HIV-1 disease progression (Shankarappa et al., 1999). In this work, 11 Subtype C LTS from Karonga have been followed over a 20 year period providing a unique opportunity to explore the longitudinal evolution of the virus. Previous work on non-progressors has mainly focused on non-subtype C infected individuals thus this study also provides a unique insight to behaviour of subtype C in LTS. In this work I have shown the presence of CXCR4 tropic viruses in subtype C infected LTS in the Karonga district individuals who show no signs of disease progression and in those who have been placed on ART due to disease progression. I have shown different patterns of diversity and divergence within the LTS from Karonga, however it is difficult to draw many conclusions from the diversity data due to the uncertainty in how well the cloning approach sampled the viral population across individuals.

Of the nine individuals infected with subtype C, that were not suspected to be dual infected, clonal sequences generated of both *env* and *gag* from four (LTS2, 5, 9 and 20) of the LTS appeared highly homogenous with a diversity of below 1.5 % in *env* and 0.5 % for *gag*. Low levels of viral diversity have previously been recorded within LTNPS (Bello et al., 2005; Bello et al., 2007; Braibant et al., 2008; Wang et al., 2003). Bello et al. (2005) reported a mean heterogeneity in *env* of <1 % in seven LTNPS who were in the 8th to 15th year of infection. Braibant et al. (2008) noted that viral diversity did not exceed 1 % in three LTNPS who had been HIV positive for over 8 years with stable $CD4^+$ cell counts. Low levels of diversity have been suggested to be indicative of variants of lower fitness, which in turn has been associated with non progression (Joos et al., 2005; Mens et al., 2010). However, the four homogenous individuals studied here have all shown signs of progression with low $CD4^+$ cell counts (Table 3.1) and placement on ART so the low diversity detected may instead be a result of homogenisation of the viral population associated with advanced disease. Within normal progressing subtype B infected individuals, homogenisation of viral diversity occurred after nine years of infection (Shankarappa et al., 1999), much sooner than what was observed in the LTS here.

It is also possible that the PCR and cloning approach may have contributed to the limited viral diversity above. This method may be influenced by reduced template detection in the early stages of PCR. Currently it is impossible to tell to what extent experimental design may have impacted on the results. Single genome analysis has been used in a number of studies looking at viral diversity (Archary et al., 2010; Fernandez et al., 2006; Mens et al., 2010) and has been cited as a more efficient method in the detection of viral diversity (Mens et al., 2010). However a study by (Jordan et al., 2010) comparing standard PCR/Cloning to single genome sequencing determined that both methods are likely to provide a similar measure of viral population diversity within a given sample. Clonal sequences generated for LTS5 were almost identical in *env* and *gag* indicating a possible lack of viral evolution and restricted viral diversity rather than PCR bottlenecking. Twenty identical clonal sequences were identified in a LTNPS nine years after infection in a study by Bello et al. (2007) with a second LTNPS producing 39 almost identical clones 17 years after infection in the same study (Bello et al., 2007). In LTS 8, the high levels of viral diversity in *env* in 2008 (4 %) were followed by a sharp decline in diversity in 2010 (0.6 %) with identical clonal sequences generated. This restricted clonal diversity in LTS8 2010 was not mirrored in *gag* clonal sequences (1.2`%), which points to possible PCR bottlenecking in the 2010 *env* sample. The inability to differentiate between PCR bottlenecking and detection of highly homogenous populations makes it difficult to draw definitive conclusions from the data here.

Higher levels of genetic diversity in *env* and *gag* were seen in LTS8, 10, and 22, and in *env* for LTS12 and *gag* for LTS17 (*gag*: Mean 1.4 %, Range 0.6 – 2.6 %, *env:* Mean 2.6 %, Range 1.8 – 4 %) which are also comparable to diversity values described in the literature. Very few viral diversity studies have focused on *gag*. In one study by Huang et al. (1998), an average of 1.7 % diversity in full *gag* (Range 0.03-3 %) was described in eight LTNPS 12-15 years after infection, while Braibant et al. (2008) described a range of viral diversity from <1 % to 5.6 % in *env* V1-V5 in nine LTNPs both similar to what was recorded here. Wang et al. (2000) reported a range of 2.8 – 11.3 % *env* V2-V3 diversity in 15 LTS and 3 LTNPs 8 to 10 years after infection.

Due to the nature of the sampling for this project, unfortunately normal progressors were not sampled at multiple time points to enable comparison with the patterns found in the LTS.

However, (Yoshimura et al., 1996) showed that four normal progressors who had been infected for an average of 7 years had a higher average diversity of 2.86 % (Range 0.55-2.86 %) in full *gag*. A comparison between seven normal progressors and nine elite progressors was carried out by Bailey et al. (2006) an average of four years and nine years after diagnosis respectively. A significantly lower diversity of full *env* was found in the elite suppressors when compared to the normal progressors (0.4 % and 2.1 % respectively) (Bailey et al., 2006). Specific studies on viral diversity are often based on more regular sampling intervals but over shorter periods of time compared to what was carried out here (Shankarappa et al., 1999; Wang et al., 2000). This makes it is very difficult to draw any direct comparisons or detect patterns between the data described here and what is described in the literature. Studies are generally focused on small subsets of patients using different gene regions, sampling times, frequency of sampling and number of sequences. However this work is consistent with both types of patterns described in the literature for diversity within LTS, i.e. some individuals showing very reduced diversity (Bello et al., 2005; Bello et al., 2007; Braibant et al., 2008; Wang et al., 2003), while others show a greater amount of diversity (e.g. Huang et al. 1998; Braibant et al. 2008) although apparently still a lower diversity than what is found in normal progressors(Yoshimura et al., 1996). Only studies employing ultradeep sequencing with the primer ID approach (Jabara et al., 2011) will determine the extent to which both the cloning and SGA approaches have affected measures of diversity within hosts.

In normal progressors, as HIV-1 infection progresses, the virus accumulates new mutations resulting in an increase in genetic divergence over time from the original infecting strain (Shankarappa et al., 1999). This pattern was also observed here in eight of the LTS from Karonga (LTS2, 5, 8, 9, 10, 17, 20 and 22). The observed increase in divergence within these individuals indicates that evolution and replication has not been arrested in spite of the observed slow disease progression. Different rates of divergence were seen between the LTS and may be the result of the different rates of evolution of the HIV-1 virus within each individual making it impossible to identify a pattern unique to LTS. Within two individuals (LTS12 and LTS20) however, a decrease in viral divergence was seen after an initial increase in divergence in 2004. LTS12 began ART in 2005 and LTS20 began ART in 2006 thus altering the environment, within which the HIV-1 virus replicates and evolves. This decrease

in divergence maybe due to the sampling of an ancestral lineage that had existed previously as a minor variant and had remained latent, as has been described in other individuals (Chun et al., 1997; Finzi et al., 1997; Metzner et al., 2003). A similar pattern was shown by Bello et al, (2005) where 7 of 16 LTNPs were seen to show slow or arrested viral divergence within the C2-V5 *env* region in conjunction with reduced viral diversity and viral loads (Bello et al., 2005).

C-PSSM identified a co receptor switch from CCR5 to CXCR4 usage in 55 % (n=11) of the subtype C LTS in this study. Early studies of subtype C tropism reported little if any usage of the CXCR4 co receptor (Bjorndal et al., 1999; Cecilia et al., 2000; Cilliers et al., 2003; Ping et al., 1999). The amount of CXCR4 tropism reported here is comparable to what is observed in subtype B progressing individuals (Jekle et al., 2003; Richman and Bozzette, 1994; Schuitemaker et al., 1992; Xiao et al., 1998b). A recent study by Connell et al. (2008) found that 30 % (n=20) of primary isolates from subtype C infected individuals from South Africa were able to use CXCR4 for cell entry (Connell et al., 2008). Raymond et al, (2010) also detected CXCR4 tropic virus in 29 % (n=52) subtype C infected individuals in Malawi. Both of these studies were focused on individuals who had progressed to AIDS or had started on HAART respectively while in this study CXCR4 tropic viruses were seen in non-progressing individuals in Karonga. The methods employed here were genotypic and have not been confirmed using phenotypic approaches. However, studies on subtype C samples where both phenotypic methods (i.e. virus grown in cell lines which have the CXCR4 co receptor only) and C-PSSM were used showed good concordance between the results (C-PSSM displayed a specificity of over 93 % in detecting CXCR4 tropism)(Connell et al., 2008; Jensen et al., 2006; Raymond et al., 2010). CXCR4 usage was detected in LTS2, 5 and 10 as early as 1998/99. LTS2 transitioned from non-progression to progression in 2009, and was placed on ART, 10 years after CXCR4 tropism was first detected. LTS5 was placed on ART in 2008, at which point a reversion back to CCR5 was seen, which may have occurred in response to treatment as has been reported in other studies (Ida et al., 1997; Kitchen et al., 2004; McDonald et al., 1997). LTS10 was referred for ART in 2010 as their $CD4^+$ cell count had dropped to 138 cells/mm^3 indicating progression, which was 11 years after viruses able to use CXCR4 were detected. This may suggest that a switch to CXCR4 tropism within subtype C is not as strongly correlated with disease progression as has been seen in subtype B

which would be consistent with the suggestion by (Meehan et al., 2010). Of the six who have been placed on ART, suggesting disease progression, four were CCR5 tropic, and of the two who had died (LTS9 and LTS22) both were also CCR5 tropic with the exception of one cloned sequence found in LTS22. However as all of these individuals were classified as LTS, further exploration of the emergence of CXCR4 usage in non-progressors as well as normal progressors is required to fully understand the implications within subtype C infected individuals.

C-PSSM provides a probable indication of co receptor tropism, however it only looks at the V3 loop of *env* and many studies have alluded to the fact that areas outside of this region were also involved in co receptor tropism. (Huang et al., 2008) identified patient *env* sequences with identical V3 loops but different co receptor tropism, and mapped possible determents of CXCR4 tropism to the GP 41 trans membrane region of *env*. The gp120 V1, V2 and C4 regions of *env* have also been implicated in co receptor binding (Carrillo and Ratner, 1996; Koito, Stamatatos, and Cheng-Mayer, 1995; Labrosse et al., 2001; Pastore et al., 2006; Suphaphiphat, Essex, and Lee, 2007). Also proviral DNA rather than circulating virus was utilized in this study thus may not represent circulating virus. The transition from CCR5 to CXCR4 is most likely a gradual multistep process and in order to get a clearer understanding of how this occurs in subtype C infected individuals more frequent sampling in longitudinal studies is required. The advent of next generation sequencing coupled with more detailed longitudinal studies will be essential in exploring the emergence of CXCR4 tropism and the implications this might have. Abbate et al, (2011) identified the presence of CXCR4 tropic variants in half of a cohort of 20 newly infected subtype B infected patients using next generation sequencing exposing the many gaps in our understanding of the HIV-1 tropism as it is.

Out of the 12 LTS studied two individuals (LTS1 and LTS21), i.e. 16.6 % (n=12), were identified as having possible dual infections. Obtaining systematic data on the frequency of dual infections is difficult due to the difficulty in detecting them. Here, further analyses using clones suggests that one of the LTS (LTS21) is likely to be dually infected but data from the other is also consistent with the sampling of a single diverse population. Ultra deep pyrosequencing is likely to be the best way forward for detecting multiple infections. By

2002 only 16 cases of superinfection had been reported in the literature (Smith, Richman, and Little, 2005). However today 49 circulating recombinant forms (CRFs) have been identified (http://www.hiv.lanl.gov) and up to 10 % of infections occur with a unique recombinant forms (URFs) (Kosakovsky Pond and Smith, 2009) indicating that dual infections occur more frequently than previously thought. Braibant et al. (2008) reported that in a cohort of nine LTNPs three individuals i.e. 30 % (n=9) were identified as being either coinfected or superinfected with two different strains of HIV-1 (Braibant et al., 2008). Bello et al. (2005) reported that one individual within a cohort of 15 LTNPs i.e. 6.25 % (n=15) was infected with two separate viral strains (Bello et al., 2005). The high level of superinfection within non-progressors may be due the extended period of infection, increasing the possibility of reinfection occurring. It is known that safe sex practices were often not adhered to in rural Malawi where the Karonga district is located (Boileau et al., 2009; Chimbiri, 2007) increasing the possibility of superinfection.

Phylogenetic analysis of LTS21 showed that the viral strains isolated from 1989 were very different to the subsequent strains obtained at all the later time points (1999 and 2010). This earlier variant was not isolated at any of the later time points. It is possible that it was simply not detected using the methods employed. A second possibility is that when LTS21 was infected with the second strain, this superinfecting strain subsequently replaced the earlier original infecting variant. The *gag* sequences obtained from 1989 both contained premature stop codons that would truncate the *gag* protein resulting in non-functional virus making replacement by a functional virus possible. The replacement of one viral population with a second population from a superinfection has been previously reported where a woman infected with subtype A was then superinfected with a subtype C virus, which subsequently recombined with the previous subtype A. On re sampling the individual 10 years after initial infection the subtype A strain was no longer detected (Fang et al., 2004).

It has been documented that superinfection results in lower $CD4^+$ cell counts, higher viral loads and disease progression (Gottlieb et al., 2004; Grobler et al., 2004) and has been noted to rapidly increase disease progression in two elite controllers (Clerc et al., 2010). However a second study found that two LTNPs showed no clinical consequences in spite of dual infection (Casado et al., 2007). LTS21 has maintained an extremely low $CD4^+$ cell count

since 2004, which may be the result of the possible dual infection. Unfortunately no viral loads were available preventing an accurate account of disease progression but the individual refuses ART and was reported as otherwise healthy at the last sampling point.

While much research has been carried out on non-progressing individuals, studies have focused on a small subsets of patients and different gene regions, sampling times, frequency of sampling and the amount of sequences created making it difficult to identify any possible links between non-progressors worldwide. A number of reasons have been suggested for non-progression encompassing host, viral and environmental factors however it is most likely to be interplay between these elements. More comprehensive and collaborative studies will need to be carried out incorporating all these elements to understand how long term control of HIV-1 is achieved. In resource poor areas such as Malawi, more elaborate studies may not be currently financially or physically practical. Here we present a preliminary view of subtype C infected long-term survivors. One of the advantages of this study was the length of time the individuals have been followed allowing us to observe changes that have occurred as many of them made the transition from a state of non-progression to progression. CXCR4 was found in a higher percentage of individuals than expected when compared to previous studies. Another interesting point was the emergence of CXCR4 tropism in individuals before the onset of disease progression. The recent introduction of ART to Karonga has involved only RT inhibitors, however in the future the recently discovered CCR5 antagonistic anti-HIV drugs may also be dispensed thus making a more informed and detailed understating of co receptor usage in the area vital. Also understanding the interplay of subtype C and co receptor tropism is important for any future preventative methods as subtype C remains the most prolific subtype of HIV-1. The wide range of pattern of divergence and diversity seen was to be expected when one considers the complicated interactions of unique quasi-species interacting with a unique immune system. The information collected here will aid in the design of future studies, which are vitally important as HIV-1 is still one leading causes of death in Malawi in the economically productive age group. This places a huge burden on the economy and the social structure (Bello, Chipeta, and Aberle-Grasse, 2006) of an already resource poor country. More frequent sampling and exploration of larger regions of the HIV-1 genome will be needed to be explored in future longitudinal studies within in the region.

Chapter 4: Molecular Evolution and Genetic Diversity of HIV-1 Karonga District Malawi

4 Chapter 4

4.1 Introduction

4.1.1 HIV-1 subtype C

HIV-1 group M viruses can be subdivided into nine (A-D, F-H, J and K) recognised phylogenetic subtypes or clades, which are all phylogenetically distinct from each other. Within group M, the average inter subtype genetic variability is 15 % for the *gag* gene and 25 % for the *env* gene (Buonaguro, Tornesello, and Buonaguro, 2007; Robertson et al., 2000). Subtype C accounts for almost 50 % of all HIV-1 infections world wide and predominates in the countries that together constitute over 80 % of all global infections such as the southern Africa and India (Buonaguro, Tornesello, and Buonaguro, 2007). Work by Travers et al. (2004) suggests that the most recent common ancestor of subtype C appeared in the late 1960s with the first evidence of subtype C virus seen in 1983 in Malawi (McCormack et al., 2002; Travers et al., 2004). By the late 1980s subtype C began a devastating spread across southern Africa and major outbreaks have now occurred in every country of southern Africa with some regions reporting the prevalence of HIV-1 infections in adults to be as high as 40 % by the year 2000 (Novitsky et al., 2002). However, it should be noted that there are other subtypes within southern Africa (e.g. Subtype A and D) (Arroyo et al., 2004; Hoelscher et al., 2001; McCormack et al., 2002; Van Harmelen et al., 1999).

One trend in the subtype distribution throughout the world is certain; subtype C infections have risen in prevalence during the past 10 years of the AIDS epidemic. This is due in part to the raging epidemics in southern Africa and India and increasing subtype C infections in South America, China and South East Asia (Cassol et al., 1996; Essex, 1999; Luo et al., 1995; Piyasirisilp et al., 2000). The introduction of subtype C into these regions may have been due to a founder event. Other subtypes did pre-exist in some populations before subtype C (Kuiken, Korber, and Shafer, 2003) and the introduction of subtype C into many regions has resulted in an apparent displacement of existing HIV-1 subtypes (such as subtypes B and CRF01_AE in south China, many subtypes in Kinshasa, Democratic Republic of Congo and subtype B in southern Brazil (Soares et al., 2005). Within Rio do Sul (Brazil) the prevalence of subtype C increased from 35 % in 1996 to 52 % in 2002, while in the Yunnan province

(China) subtype C rose in prevalence from 5.1 % in 1992 to 90 % by 2002 (Soares et al., 2005; Luo et al., 1995; Piyasirisilp et al., 2000).

HIV has been studied in Karonga District, Malawi, for over 20 years and a similar pattern of the displacement of other subtypes in favour of subtype C has been observed there. The prevalence of people infected with subtype D and A was 9 % each in 1982-1984. An unclassifiable subtype was found to have infected 27 % (n=11) of individuals during the same time period, while subtype C represented 55% (n=11) of infections. By the late 1980s the prevalence of people infected with subtype C had risen substantially to 90 % (n=168) (McCormack et al., 2002). The monitoring of HIV in Karonga from the 1980s to the present day provides a unique opportunity to follow the evolutionary changes occurring within a growing subtype C epidemic. The rapid emergence of specific subtypes in a particular region has been attributed in part to specific modes and routes of transmission. For example, intravenous drug use in south East Asia in the 1980s let to the rapid spread of CRF01_AE (Lau, Wang, and Saksena, 2007). A similar expansion of subtype B HIV-1 transmission occurred among men who have sex with men in North America and Europe in the 1980s (Foley et al., 2000). However, HIV-1 subtype C appears to have emerged slowly through the world over the past 10 to 15 years following multiple introductions (Arien, Vanham, and Arts, 2007) and without any apparent unique route of transmission.

Subtype C does however; appear to be spreading more rapidly than other subtypes. The disproportionate increase in C viruses relative to other HIV-1 strains suggests that subtype C may be more easily transmitted or that it has a higher level of "fitness" at the population level. At the viral level, it has been suggested that an extra NF-κB binding site in the LTR may enhance gene expression, altering the transmissibility and pathogenesis of C viruses (Shankarappa et al., 2001). Additional features of subtype C include a prematurely truncated Rev protein (Gao et al., 1998; Shankarappa et al., 2001), a five amino acid insertion in *Vpu* (McCormick-Davis et al., 2000), which may influence viral gene expression and alter the virulence of C viruses (Tatt et al., 2001). A more active protease with increased catalytic activity has also been reported in subtype C isolates (Velazquez-Campoy et al., 2001).

Subtype C infected individuals have been shown to have higher viral loads than that observed for other subtypes, which may in turn cause increased levels of viral transmission (Neilson et al., 1999). The transmission efficiency of the different subtypes may have an effect on subtype prevalence distributions also. A study in Tanzania has suggested that the maternal subtype could play a part in the rates of vertical transmission, with subtypes A and C and recombinant viruses being more likely to be perinatally transmitted than subtype D (Renjifo et al., 2001). HIV-1 subtype specific differences in disease progression appear to be conflicting. For example, no difference in disease progression was documented between patients infected with subtypes B and C in Israel (Weisman et al., 1999), or among patients infected with subtypes A, B, C and D in Sweden (Alaeus et al., 1999). In a alternate study by Neilson et al. (1999) a cohort of pregnant women in Kenya subtype C infected patients, displayed higher viral loads and lower $CD4^+$ cell counts when compared to those infected with subtypes A and D (Neilson et al., 1999). In a detailed *in vitro* study by Ball et al. (2003), subtype C HIV-1 isolates were outcompeted by subtype B HIV-1 isolates in PBMCs, CD4 T cell lymphocytes and blood derived macrophages in pairwise competition experiments. However, the subtype C isolates were equally fit on Langerhans cells. Since epidermal Langerhans cells provide a model for primary HIV-1 infection in the genital tract, it is conceivable that subtype C HIV-1 isolates may be efficiently transmitted but less fit during disease progression (Ball et al., 2003). A number of *in vivo* studies suggest subtype C is transmitted as efficiently as other HIV-1 group M isolates in human cohorts (Renjifo et al., 2001; Walker et al., 2005). If subtype C viruses are less virulent than other subtypes, subtype C infections might result in a slower disease progression, longer periods of asymptomatic infection and more opportunities for transmission (Arien, Vanham, and Arts, 2007).

4.1.2 HIV-1 in Karonga

The first evidence of HIV-1 in Malawi dates back to 1982 (Glynn et al., 2001; McCormack et al., 2002), while the first case of AIDS was recorded in 1985 (Kachapila, 1998). According to epidemiological data, HIV-1 spread rapidly in Malawi in the 1980s and infection rates among pregnant women continued to increase during the 1990s and rose to 19.8 % in 2003 (The Global Fund, 2010). The prevalence in adults has since stabilised at about 14 % (UNAIDS, 2006). There appear to be regional differences in the spread of HIV-1, with infection rates higher in the South than in the North of the country and evidence from antenatal surveys

suggest that infection levels are higher in urban than in rural areas (UNAIDS, 2007). In the rural northern Karonga District prevalence has been relatively stable at around 12 % since the early 1990s (White et al., 2007). In this District, two total population surveys were carried out in the early and late 1980s as part of a large study of mycobacterial diseases (Ponninghaus et al., 1987). During these studies over 44,000 blood samples were collected and 198 individuals were found to be HIV-1 positive (Glynn et al., 2001; McCormack et al., 2002). Phylogenetic analysis of HIV-1 showed the presence of three subtypes (A, C and D) and an 'unclassifiable' HIV strain (McCormack et al., 2002) as early as 1984 and by the late 1980s three recombinant forms were present in the population (AC, AD and DC) based on partial *gag* and *env* gene sequences.

Phylogenetic trees reconstructed by McCormack et al. (2002) using subtype C sequences dating form the 1980s based on both *gag* and *env* were largely without bootstrap support, yet a certain amount of structure was evident and a number of clusters were identified. The majority of Karonga subtype C sequences from the 1980s were shown by McCormack et al. (2002) to belong to one main clade that was characterised by very short internal branches and included subtype C sequences from the LANL database derived from other African countries such as Botswana, Zambia and South Africa, as well as India. Within the C clade a number of clusters composed of primarily Karonga sequences were also identified, one of which represented almost 40 % of Karonga sequences from the 1980s and was thought to have been introduced by a single individual around 1980 (McCormack et al., 2002; Travers et al., 2004).

To measure the possible effect of HIV on survival in rural Africa and to aid planning and counseling, a retrospective cohort study was carried out (Crampin et al., 2002). In this study individuals whose HIV status was known from the 1980s surveys were followed up in the period 1998-2000. One hundred and ninety seven known HIV-1 positive individuals and 397 HIV-1 negative individuals and their families were interviewed, examined and HIV tested after counseling and if consent was given (Crampin et al., 2002). Sequence data for *gag* and *env* were retrieved from individuals who were alive between the two surveys (1980s and the 1990s) and their spouses and off spring as well as from unrelated individuals that were HIV negative in the first survey (in the 1980s). Control sequences from additional HIV-1 positive adults in Karonga collected between 1997 and 2000 were also available.

More recently between 2006 and 2011 a number of studies (HIV-1 Sero Survey (HSS) and ART study) have been carried out in Karonga to study the patterns, risk factors, trends and dynamics of HIV by monitoring changes in sexual behavior and HIV incidence as ART becomes more available, to identify factors affecting adherence to ART and to asses the role of ART in mitigating the socio-demographic impact of HIV/AIDS. Blood samples collected from HIV-1 positive individuals in 2007 to 2009 as part of these studies have been utilised in this project

The availability of HIV-1 positive blood samples collected over the past 20 years provided an opportunity to;

a) explore the evolution of the epidemic within Karonga from 1982, when the first HIV-1 positive sample was identified, to 2010 including patterns of diversity and divergence over time and possible geographic associations within Karonga,

b) explore the HIV-1 genetic diversity seen in spouse pairs, and patterns associated with transmission and long-term infection and,

c) utilise new bioinformatics tools to monitor the emergence of CXCR4 usage within the population.

4.2 Materials and Methods

4.2.1 Samples

In 2008/09, 151 whole blood samples were collected from HIV-1 positive individuals in Karonga District, Malawi as part of a much larger observational cohort study screening for ART suitability, side affects of ART and the general health of the individual (Bansode et al., 2010). The 151 samples were chosen randomly for this study in order to further explore the molecular diversity of HIV-1 in Karonga. A further 136 plasma samples collected between 2007 and 2008 as part of a larger HIV-1 Sero Survey (HSS) were also utilized within this project.

4.2.2 DNA extraction

All samples collected between 2007 and 2009 as part of both the ART and HSS studies were separated into plasma and cell pellet by local laboratory technicians using centrifugation and were stored at – 20 °C. DNA was extracted in Karonga from cell pellets (ART samples) and from plasma (HSS samples) using the QIAamp DNA Blood Mini Kit (Qiagen) according to the manufactures instructions. The extracted proviral DNA was transported to the Molecular Evolution and Systematic Laboratory in NUI Galway.

4.2.3 PCR amplification

Fragments of *env* (C2V3: 549 bp) and of *gag* (p17/p24: 750 bp) were amplified using a nested PCR as previously described McCormack et al. (2002). Outer primers ED5 (forward) (5' ATGGGATCAAAGCCTAAAGCCATGTG 3') and ED12 (reverse) (5' AGT GCTTCCTGCTGCTCCCAAGAACCCAAG 3') and inner primers ED31 (forward) (5' CCTCAGCCATTACACAGGCCTGTCCAAAG 3') and ED33 (reverse) (5' TTACAGTAGAAAAATTCCCCTC) (Delwart et al., 1995; Delwart et al., 1994) were employed in a nested PCR for amplification of the *env* C2-V3 region. Primers for the *gag* p17/p24 region included outer primers DT1 (forward) (5' ATGGGTGCGAGAGCGTCAG TATT 3') and DT7 (reverse) (5' CCCTGACATGCTGTCATCATTTCTTCT 3") and inner primers DT3 (forward) (5' CATCTAGTATGGGCAAGCAGGGA 3') and DT6 (reverse) (5' ATGCTGACAGGGCTATACATTCTTAC 3') (Tatt, Barlow, and Clewley, 2000). Successfully PCR amplified samples were purified using the HiYield Gel/PCR DNA Fragments Extraction Kit (Real Genomics) according to the manufacturers instructions. The

resulting purified PCR products were sequenced in both directions by LGC Genomics (Germany). Sequence chromatographs were examined and sequence contigs assembled in Seqman (DNAStar Inc.).

4.2.4 Alignments and phylogenetic analysis

Subtyping analysis; For both *env* and *gag* gene fragments, all newly generated sequences from 2007-2009 were aligned to unpublished sequences from Karonga retrieved from the 1990s and representative sequences of all the other subtypes (reference alignments downloaded from the HIV database, (www.hiv.lanl.gov). The sequences were manually aligned in MacClade 4 (Sinauer Associates). Phylogenetic trees were reconstructed under the GTR + gamma model of DNA substitution implemented by RAxML 7.0.3 (Stamatakis, 2006) with all parameters optimised by RAxML. Confidence levels in the groupings in the phylogeny were assessed using 1000 bootstrap replicates as part of the RAxML phylogeny reconstruction. A bootstrap value equal to or greater than 70 % was considered significant. Recombination between different subtypes within the partial *env* and *gag* sequences was identified using the jumping profile Hidden Markov Model internet application (jpHMM) (http://jphmm.gobics.de/) (Zhang et al., 2006).

Subtype C phylogeny; Newly generated subtype C sequences were added to alignments of subtype C sequences retrieved from dried blood spots collected in the Karonga District in the time periods 1983-1984, 1986-1989 (McCormack et al., 2002) and 1997-2000. HIV-1 subtype C sequences from a number of different countries were also downloaded from the HIV-1 database (http://www.hiv.lanl.gov/) and were also included as reference sequences within the alignments. All known epidemiologically linked sequences were removed. The multiple alignments were assembled and optimized in MacClade 4 (Sinauer Assoc.). The *gag* alignment contained 560 sequences; 119 of which were collected in the Karonga District in the 1980s, 118 collected in Karonga between 1997 and 2000, 93 collected in Karonga between 2007 and 2008 and 230 were from other locations. The *env* alignment contained 563 sequences, 110 of which were collected in the Karonga District in the 1980s, 100 collected in Karonga between 1997 and 2000, 101 collected in Karonga between 2008 and 2009, and 252 were from other locations. Phylogenetic trees were reconstructed by RAxML

as described above. The subtype C ancestral sequences derived by (Travers et al., 2004) was employed as the out-group.

Spouse phylogeny; Multiple alignments were assembled with all sequences generated from Karonga between the 1980s and the 2000s including all epidemiologically linked sequences from spouses and their offspring for both *gag* and *env*. Phylogenetic trees were reconstructed by RAxML as described above. The subtype C ancestral sequences derived by (Travers et al., 2004) was employed as the out-group.

4.2.5 Pairwise genetic distances

jModelTest 0.1.1 (Guindon and Gascuel, 2003; Posada, 2008) (using the Akaike information criterion) was employed to select the optimal substitution model for both *env* and *gag*. The substitution model suggested by jModelTest 0.1 for the *env* Karonga data set was TIM2+G with base frequencies A; 0.46, C; 0.19, G; 0.21, T; 0.15, Rate matrix A-C; 0.99, A-G; 3.96, A-T; 0.99, C-G; 1.00, C-T; 6.28, G-T; 1.00, and G= 0.37. The substitution model suggested by jModelTest 0.1 for the *gag* Karonga data set was TPMuf+I+G with base frequencies A; 0.37, C; 0.23, G; 0.23, T; 0.17, Rate matrix A-C; 1.59, A-G; 6.00, A-T; 1.59, C-G; 1.00, C-T; 6.00, G-T; 1.00, I= 0.34, and G= 0.65. Using this model of substitution, pairwise evolutionary nucleotide distances were computed by PAUP* 4.0 (D.L. Swofford, Sinauer Associates, Inc. Publishers) for both the spouse pairs and the remaining Karonga dataset containing no epidemiologically linked individuals. The amount of genetic divergence seen in the 1980s, 1990s and 2000s was calculated using the average genetic distance between all sequences from that time point to the subtype C ancestral sequences previously derived by (Travers et al., 2004). This average was then plotted on a graph in Microsoft Excel with the time points on the X-axis and the genetic distance on the Y-axis. Error bars were used to mark the minimum and maximum genetic distance between the sequences from each time point and the subtype C ancestral sequence. Sequences from a number of other African countries, India and Brazil of similar size to the partial *env* and *gag* amplified here, were downloaded from the LANL sequence database (http://www.hiv.lanl.gov/). Multiple sequence alignments were created for each country and the genetic distances were computed by PAUP* 4.0 using the model of substitution obtained from jModelTest for the Karonga *env* and *gag* sequence alignments. All averages and standard deviations were calculated in Excel

(Microsoft). A Z test was calculated by hand to test the difference between the average genetic diversity in Karonga compared to a number of other African countries, India and Brazil, and between the average genetic divergences in Karonga in the 1980s, 1990s and 2000s.

4.2.6 Co receptor tropism prediction

The PSSM (Jensen et al., 2003) (http://indra.mullins.microbiol.washington.edu/webpssm/) subtype C specific SINSI matrix was used to predict the co receptor tropism for all *env* sequences retrieved from Karonga. A multiple alignment of all the *env* sequences retrieved from Karonga was constructed in MacClade 4 (Sinauer Assoc.) and was trimmed to include only the V3 loop. V3 loop sequences containing deletions or ambiguous sites were not analysed.

4.3 Results

4.3.1 Subtyping of *gag* and *env* sequences

Plasma samples had originally been collected as part of the HSS study. Of the 136 plasma samples used in this study, *gag* was successfully sequenced from only 36 samples (26 %). Subsequently DNA extracted from 151 cell pellets collected as part of the ART study were used for remaining *gag* and *env* PCR's. *Gag* was sequenced from 55 of 71 cell pellets and *env* was sequenced from 107 of 151 cell pellets. DNA sequences were retrieved for both gene regions from 44 samples.

Of the 167 sequences retrieved from samples collected in 1997-2000 15 were not subtype C (11 %) in one or both of the *env* and *gag* gene regions (Table 4.1). Two samples were subtype A (3070Ma99, 3173Ma00) and three were subtype D (314Ma99, 218 Ma99, 259 Ma99) in both gene regions. One sample was unclassifiable (223Ma99) in both gene regions. Recombinants (*env/gag*), found were as follows; AC (2427Ma98), CA (3311Ma99, 3274Ma00) DC (2439Ma98) (Table 4.2). The *env* C2-V3 region did not amplify from one sample (340Ma99), which was subtype A in *gag* (p17/p24). No *gag* sequence was retrieved from four samples, two of which were subtype A (307Ma99, 2981Ma99) and two of which were subtype D (262Ma99, 346Ma99) in *env*.

Of the 154 sequences retrieved from samples collected in 2007-2009 where either an *env* or *gag* sequence was available, eight were not subtype C (5.2 %) (Table 4.1). Three samples were different subtypes in the two genes as follows (*env/gag*): AC (53288, 54781) and CA (54332) (Table 4.2). No *gag* sequence was retrieved from one sample (47024) that was subtype A1 in *env*. jpHMM identified recombination within a gene fragment for four individuals. Three individuals (54653, 54754 and 54702) were subtype A1/C within *gag* and subtype C within *env*. These three sequences were found to have an identical breakpoint and so were explored further in Chapter 5. One individual (48734) was subtype A1/D in *env* but no *gag* sequence was retrieved (Table 4.2).

Table 4.1 Numbers of sequences produced in Karonga in the 1990s and 2000s that were subtype A, D or C or were inter subtypes recombinants.

		Number of sequences that were:				
Sample Period	**Region(s) sequenced**	**Subtype A**	**Subtype D**	**Subtype C**	**Un-classifiable**	**Recominant**
1998-2000	*gag* only	1		34		
	env only	2	2	17		
	gag and env	2	3	101	1	4
	Total	5	5	152	1	4
2007-2009	*gag* only	1		45		3
	env only	2		61		1
	gag and env			38		3
	Total	3		146		7

Table 4.2 A list of the inter subtype recombinants identified from 1998-2009. The sequences are labeled by the sample number, Ma for Malawi, and the year of collection. HXB2 numbering is used to identify the breakpoints when recombination was seen to occur within the *env* or *gag*.

Sequence name	**Subtype within *env***	**Subtype within *gag***	**Inter subtype Recombinant Break Points**	
2427Ma98	A1	C		
53288Ma08	A1	C		
54781Ma08	A1	C		
3311Ma99	C	A1		
3274Ma00	C	A1		
54332Ma08	C	A1		
2439Ma98	D	C		
54653Ma08	C	A1/C	925-1202 A1	1203-1503 C
54574Ma08	C	A1/C	925-1202 A1	1203-1503 C
54702Ma08	C	A1/C	925-1202 A1	1203-1503 C
48734Ma08	A1/D	n/a*	6870-7186 A1	7187-7284 D

**gag* sequence not retrieved.

4.3.2 Subtype C phylogeny

Phylogenetic reconstructions for both *gag* and *env* showed sequences from Karonga to be spread throughout the entire subtype C phylogenies (Figure 4.1 and 4.2). Sequences from the three different time points did not form distinct groupings and all three time points were found mixed together. Both trees were characterised by short internal branch lengths and long terminal branches. A number of large clusters were seen in both trees but the majority received no support from bootstrapping. However, as might be expected from the large number of sequences included a number of small clusters composed of sequences from individuals with unknown epidemiological links did group together with high bootstrap support (Appendix 6 and 7).

Cluster 1 originally, described by McCormack et al. (2002) as the Malawi Cluster, was still present within the *gag* tree and was expanded with sequences from both the 1990s and from 2000s (Figure 4.1). The cluster was now made up of 112 sequences of which only five were not of Karonga origin (one from Brazil and Botswana and three from South Africa) and represented 32.3 % of all Karonga sequences. The sequences from a small cluster identified previously as Cluster 3 by McCormack et al. (2002) were no longer present as a discrete cluster but were part of a much larger cluster made up of an additional 26 sequences from Karonga from both the 1990s and 2000s as well as four sequences from Botswana and one sequence from South Africa. A second small cluster identified previously as Cluster 4, was originally composed of sequences from a spouse pair and a third unrelated sequence. These three sequences were now found within a much larger, highly supported cluster that was composed of a number of sequences from Brazil as well as two other sequences retrieved from Karonga in 1998. Another cluster, not previously identified by McCormack et al. (2002) was present and was comprised of 44 sequences from Karonga across all three-time points as well as three additional sequences from Botswana and one sequence from Tanzania. However, the cluster was not seen in the *env* ML tree. The remaining sequences were intermixed with sequences from other countries in addition to a number of small sub clusters consisting of Karonga sequences though the majority of these were not supported by bootstrapping. The majority of sequences from India and Bangladesh formed a separate cluster with no sequences from any other location, as did the majority of sequences from Brazil (along with the five sequences from Karonga). The sequences from the remaining

countries were not grouped in distinct geographical clusters and were distributed across the whole tree.

Within the *env* phylogenetic tree, Cluster 1 had separated into two clusters and was expanded upon by a number of sequences from 1990s and 2000s but to a lesser extant than that seen within *gag* (Figure 4.2). One cluster contained 11 sequences from Karonga and one contained 44 sequences from Karonga as well as an additional two sequences from South Africa. Cluster 1 only accounted for 17.6 % of the sequences from Karonga compared to 40 % as had been described previously by McCormack et al. (2002). As in *gag*, the previously described Cluster 3 was now part of a much larger cluster composed of 26 new sequences from the 1990s and 2000s and six sequences from the 1980s in addition to three sequences from Botswana and one sequence retrieved from Israel. Cluster 4 was still present within *env* and was found as a sister cluster to a cluster containing sequences from Brazil, Burundi, Botswana, Ethiopia, Djibouti, Uganda, Tanzania and Kenya as well as one sequence from Karonga retrieved in the 1990s. A separate cluster not previously described by McCormack et al., (2002) comprising 28 sequences from Karonga and two other Africa sequences was identified, however, this grouping was not observed in the *gag* tree. The remaining Karonga sequences were spread across the tree, intermixed with the sequences downloaded from LANL. The majority of sequences retrieved from India were found within one cluster with the addition of three sequences retrieved from Karonga, two from the 1990s and one from the 2000s. Sequences retrieved from the remaining countries were intermixed with each other and Karonga sequences with the exception of Brazil where the majority of sequences were found within one cluster.

4.3.3 Genetic diversity and divergence

The mean pairwise genetic diversity within subtype C in Karonga across all time points was 11.3 % with a range of 0.0 % to 23.6 % in *env* and 6.2 % with a range of 0.0 % to 14.97 % in *gag*. The level of diversity seen in Karonga was significantly higher than what was observed between sequences from Brazil and India in *env* who have newer subtype C epidemics (Bredell et al., 2007; Soares et al., 2005) but was significantly lower than that seen in the subtype C epidemics in other African countries as can be seen in Table 4.3. This situation

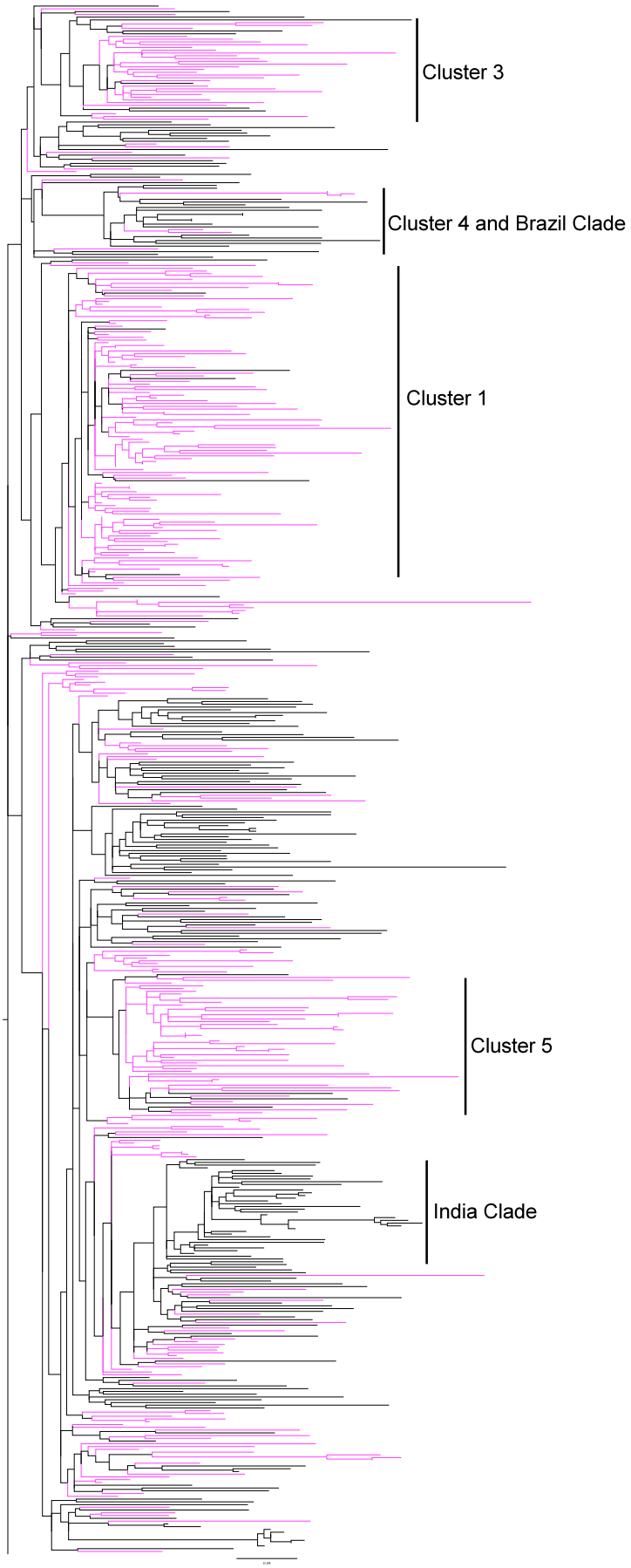


Figure 4.1 Maximum likelihood *gag* tree for Subtype C sequences from Karonga District and from other countries. Karonga sequences are in pink and sequences from other countries are in black. Clusters previously descried by McCormack et al. (2002) are marked (Cluster 1, 3 and 4) in addition to a new Karonga cluster and the Indian clade and the Brazilian clade.

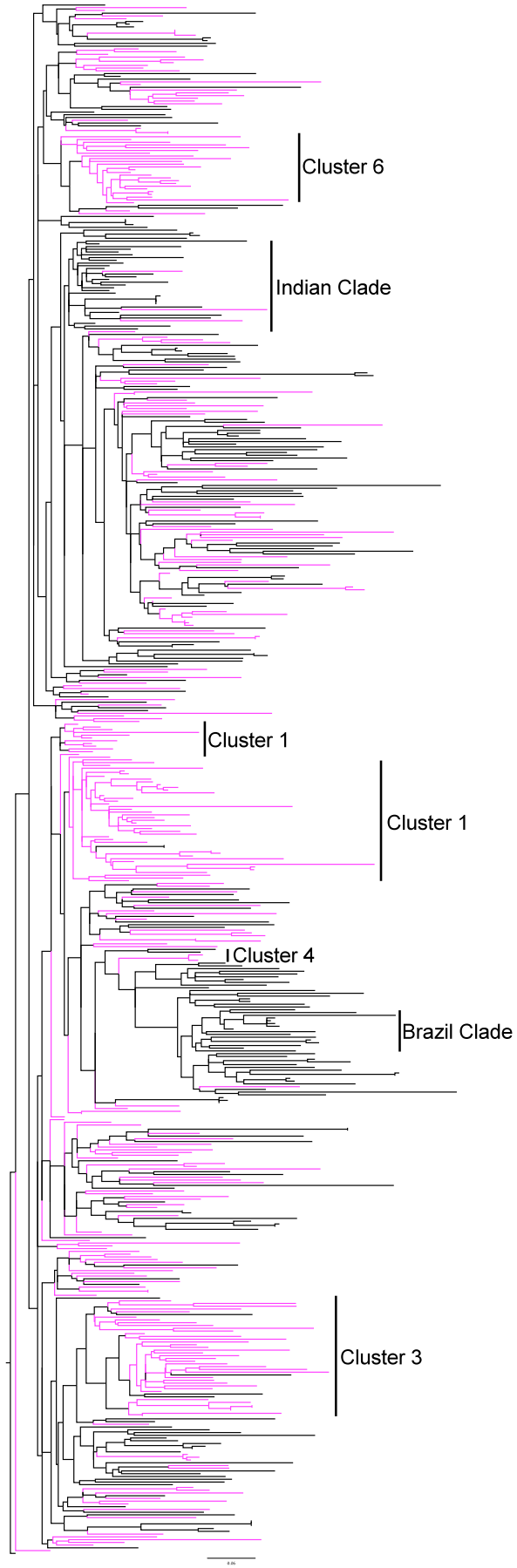


Figure 4.2 Maximum likelihood *env* tree for Subtype C sequences from Karonga District and from other countries. Karonga sequences are in pink and sequences from other countries are in black. Clusters previously described by McCormack et al. (2002) are marked (cluster 1, 3 and 4) in addition to a new Karonga cluster and the Indian clade and the Brazilian clade.

was mirrored in *gag* with the exception of Brazil where the average diversity was not significantly different to that seen in Karonga.

The pairwise genetic distances were calculated for all the sequences retrieved from 1989 to 2000 and then for all the sequences retrieved from 2007 to 2009. As might be expected, all of the countries, showed a significant increase in diversity between the first time period and the second time period with the exception of Botswana (Table 4.4). Within each of the two time periods Karonga showed a significantly lower average diversity than the other African countries but showed significantly higher diversity than India. The ranges of diversities were similar across the countries. In Karonga the mean genetic diversity in *env* (C2/V3) increased from 10.05 % to 13.60 % ($Z=94.53$ $P<0.0001$) between 1989-2000 and 2007-2009 (Table 4.4). Within *gag* (p17p24) the pairwise genetic diversity also increased from 5.57 % to 7.84 % ($Z=65.63$ $P<0.0001$) between sequences collected in 1989-2000 and 2007-2009 respectively. Mean divergence from the reconstructed ancestral sequence (Travers et al., 2004) increased significantly over the three time periods (1980s, 1990s and 2000s) for both *gag* and *env* (Figure 4.3 a and b). Between the 1980s and the 1990s divergence increased by 2.5 % and 1.5 % in *env* and *gag* respectively (*env*: $Z=10.07$, *gag*: $Z=10.92$ $P<0.0001$). Between the 1990s and the 2000s divergence increased by a further 1.1 % in *env* and 0.9 % in *gag* (*env*: $Z=4.10$, *gag*: $Z=4.36$ $P<0.0001$).

4.3.4 Spouses

As part of the retrospective cohort study (Crampin et al., 2002) the families of HIV-1 positive and HIV-1 negative individuals from the 1986-1989 survey were followed up between 1998-2000. This provided sequence data from pairs of individuals that were married or co-habiting as spouses. Analysis of such data provided an opportunity to explore the extent of evolution in viruses in related hosts and over different time periods. For the five pairs of individuals who were HIV-1 positive and sampled in the 1990s (i.e. HIV-1 negative in the 1980s) (Table 4.5)(Appendix 6 and 7) all five spouse pair sequences grouped close to each other. The average pairwise genetic distance between spouse pairs for *gag* was 1.2 % (SD ± 1.1) and for *env* was 5.1 % (SD ± 1.2). For HIV-1 positive individuals seen in the 1980s but who had died by the late 1990s and their spouses who had only become HIV-1 positive by the 1990s (Table 4.5) sequence information for *gag* or *env* or both, was available for 14 spouse pairs.

Table 4.3 The average genetic distance for *env* and *gag* in Karonga and other countries from the early 1980s to the 2000s. The Standard Deviation is included within the brackets and n=the number of sequences. The Z test was used to examine the differences between the mean genetic distance in Karonga compared to all other countries.

env

Country		Mean Genetic Distance (Standard Deviation)	Range	Z score
Karonga	n=313	11.3% (±2.8%)	0.00 - 23.6%	
Botswana	n=113	13.44% (±2.1%)	0.29 - 22.79%	Z=71.72*
Brazil	n=28	10.24% (±2.1%)	1.18 - 14.75%	Z=10.11*
India	n=152	9.69% (±2.7%)	0.29 - 24.94%	Z=56.71*
Mozambique	n=103	13.25% (±2.5%)	0.58 - 26.86%	Z=51.68*
South Africa	n=300	14.42% (±2.5%)	2.92 - 26.91%	Z=189.45*
Tanzania	n=64	14.03% (±2.4%)	1.76 - 20.84%	Z=49.15*
Zambia	n=285	14.12% (±2.5%)	0.29 -24.93%	Z=157.76*

* P<0.0001

gag

Country		Mean Genetic Distance (Standard Deviation)	Range	Z score
Karonga	n=332	6.24% (±1.9%)	0.00 - 14.97%	
Botswana	n=76	7.84% (±1.3%)	3.35 - 12.09%	Z=62.33*
Brazil	n=18	6.27% (±1.7%)	0.67 - 9.77%	Z=0.73"
India	n=47	5.25% (±1.7%)	1.01 - 11.27%	Z=19.10*
South Africa	n=138	7.22% (±1.6%)	0.67 - 14.02%	Z=54.51*
Zambia	n=139	6.98% (±1.6%)	0.17 - 13.56%	Z=43.86*

*P<0.0001

"P 0.4676

Table 4.4 The mean genetic distance for *env* and *gag* seen from 1988 to 2000 and from 2007 to 2008 for Karonga and a number of other countries. The Standard Deviation is included in the brackets and n=the number of sequences. The Z test was used to look at the difference between the mean genetic distance between the two time periods in Karonga and the corresponding time period in the other country.

env

Country	Years		Mean Genetic Distance (Standard Deviation)	Range	Z Score
Karonga	1988-2000	n=211	10.05% (±2.7%)	0.29 -23.56%	
	2008-2009	n=103	13.60% (±2.4%)	0.00 - 21.21%	Z=94.53*
South Africa	1990-2000	n=123	13.07% (±2.2%)	2.92 - 22.51%	Z=99.16*
	2001-2009	n=159	15.11% (±2.5%)	6.73 - 26.91%	Z=40.38*
Zambia	1989-2000	n=57	12.41% (±2.6%)	2.90 - 20.94%	Z=35.51*
	2001-2005	n=147	14.51% (±2.4%)	2.90 - 24.93%	Z=23.78*
India	1991-2000	n=53	7.44% (±2.3%)	0.58 - 17.25%	Z=39.73*
	2003-2008	n=98	10.67% (±2.7%)	0.29 - 24.94%	Z=56.31*

* P <0.0001

gag

Country	Years		Mean Genetic Distance (Standard Deviation)	Range	Z Score
Karonga	1988-2000	n=239	5.57% (±1.8%)	0.00 - 15.02%	
	2008-2009	n=99	7.84% (±2.1%)	0.36 - 16.48%	Z=65.63*
South Africa	1997-2000	n=62	6.78%(±1.4%)	0.67 - 11.44%	Z=35.37*
	2001-2009	n=76	7.51% (±1.6%)	2.23 - 12.86%	Z=6.87*
Zambia	1988-2000	n=40	6.04% (±1.5%)	0.17 - 16.40%	Z=7.30*
	2001-2006	n=99	7.31% (±1.6%)	0.17 - 13.56%	Z=12.75*
Botswanna	1996-2000	n=49	7.80% (±1.3%)	3.70 - 11.67%	Z=55.84*
	2007	n=27	7.90% (±1.3%)	4.19 - 12.09%	Z=1.31"

* P <0.0001

" P 0.1906

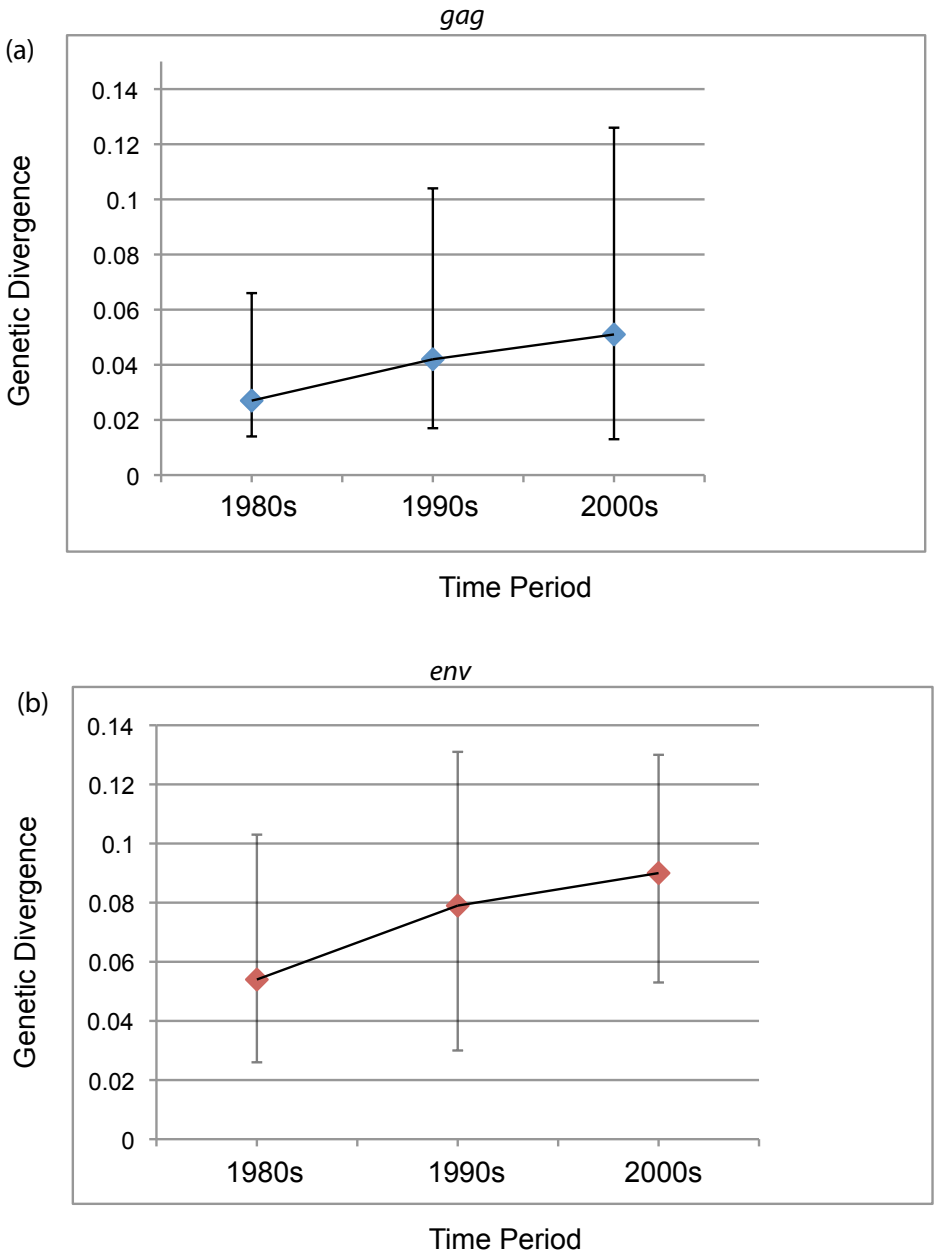


Figure 4.3 Genetic divergence seen in (a) *gag* and (b) *env* over the three time periods. The average genetic distance between all the sequences from that time period and the reconstructed MRCA of subtype C (Travers et al., 2004) is represented by the diamond. The error bars mark the range of diversity.

The average pairwise genetic distance between spouse pairs for *gag* was 3.9 % (SD ± 1.5) and for *env* was 7.0 % (SD ± 2.2). Only five spouse pair sequences grouped close to each other on the tree.

4.3.4 Prediction of co receptor tropism in Karonga

The web based tool C-PSSM was used to predict CXCR4 co receptor tropism in the whole Karonga data set. C-PSSM predicted a total of 31 *env* sequences from Karonga to be CXCR4 tropic. The number of sequences predicted to be CXCR4 tropic increased over the three sampling time points. In the 1980s five of 127 sequences (3.9 %, n=127) were predicted to be CXCR4 tropic, in the 1990s eight of 100 sequences (8 %, n=100) were predicted to be CXCR4 tropic and in the 2000s 21 of 84 sequences (24.7 %, n=84) were predicted to be CXCR4 tropic (Table 4.6). Of all 31 CXCR4 sequences, only one sequence retrieved in 1999 contained a positively charged amino acid (arginine - R) at position 11, this sequence also had a positively charged amino acid (R) at position 25. The remaining CXCR4 predicted sequences contained a Serine at position 11. Four CXCR4 tropic sequences had a positive amino acid (lysine - K) at position 24 and five CXCR4 tropic viruses had a positive amino acid (two had a K and three had an R) at position 25. Fifteen sequences had a basic amino acid at position 24 (either and R or a K) but were not predicted as CXCR4 tropic by C-PSSM. Five sequences were missing the amino acid at position 24 of the V3 loop and one sequence was missing the amino acid at potion 25. Two sequences collected in 1989 from a husband and his wife were found previously to have an insertion of a serine in the V3 loop, the sequence retrieved from the wife who was still alive in 1999 (sequence 205eMa99) no longer had this insertion.

Table 4.5 Summary of results for subtype C spouse data to include sequencing success for *env* and *gag* gene fragments and the percentage genetic distance between sequences of individuals that were HIV positive in the 1980s (most whom had died by the 1990s) and their spouses subsequently seen in the 1990s (Spouse pair 1 to 14), and individuals who were initially HIV negative in the 1980s but who were positive in the 1990s and their spouses (Spouse pair 15 to 19). x indicates missing sequence data, * indicates where sequence pairs clustered directly with each other on the ML trees, ** indicates those spouse pairs that clustered with an additional 1990s sequence and *** indicates those spouse pairs that clustered with additional 2000s sequences.

	Pairwise Genetic Distance	
	***env*%**	***gag*%**
Spouse pair 1	7	2.8
Spouse pair 2	6.7	2.9
Spouse pair 3	4.2*	2.9*
Spouse pair 4	4.2*	3.2*
Spouse pair 5	4.1*	3.1
Spouse pair 6	x	3.1*
Spouse pair 7	7.0**	2.6**
Spouse pair 8	8.5	4.6
Spouse pair 9	8.8	4.8
Spouse pair 10	x	4.9
Spouse pair 11	10.5	7.9
Spouse pair 12	6.8	4.5
Spouse pair 13	5.8	3.7
Spouse pair 14	10.2	x
Spouse pair 15	x	2.8**
Spouse pair 16	4.2*	0.7*
Spouse pair 17	5.9*	x
Spouse pair 18	x	0.3*
Spouse pair 19	x	1.1***

Table 4.6 Sequences predicted to use CXCR4 by C-PSSM. Any result below the 0.95 percentile is predicted to be CXCR4 tropic. * = a positive amino acid at position 11; ** = a positive amino acid at position 11 and 25; $^{+}$ = a positive amino acid at position 25.

Time Period	Sequence Name	Predicted Co Receptor Tropism	Percentile
1980s:	10eMa89	CXCR4	0.75 *
	48eMa89	CXCR4	0.75
	77eMa88	CXCR4	0.67 *
	105eMa88	CXCR4	0.87
	126eMa88	CXCR4	0.68 $^{+}$
1990s:	245eMa99	CXCR4	0.78
	256eMa99	CXCR4	0.69 $^{+}$
	278eMa00	CXCR4	0.74 *
	301eMa99	CXCR4	0.94
	312eMa99	CXCR4	0.93 **
	1981eMa97	CXCR4	0.91
	3274eMa00	CXCR4	0.84
	3310eMa00	CXCR4	0.86
2000s:	53231eMa08	CXCR4	0.82
	53274eMa08	CXCR4	0.86
	53354eMa08	CXCR4	0.77
	53394eMa08	CXCR4	0.77 $^{+}$
	53443eMa08	CXCR4	0.68
	53900eMa08	CXCR4	0.79
	54138eMa08	CXCR4	0.68
	54348eMa08	CXCR4	0.86
	54434eMa08	CXCR4	0.68
	54535eMa08	CXCR4	0.82
	54618eMa08	CXCR4	0.74 *
	54653eMa08	CXCR4	0.61
	54702eMa08	CXCR4	0.61
	54784eMa08	CXCR4	0.69
	54451eMa08	CXCR4	0.61
	62597eMa09	CXCR4	0.84 $^{+}$
	62601eMa09	CXCR4	0.82 $^{+}$
	63400eMa09	CXCR4	0.95
	64417eMa09	CXCR4	0.92
	66517eMa09	CXCR4	0.83
	67776eMa09	CXCR4	0.81

4.4 Discussion

As part of the Karonga prevention study, HIV has been studied in the Karonga District Malawi for over twenty years. This has provided a rare opportunity to study the evolution of HIV within a population. HIV-1 subtype C emerged as the most successful strain of the virus in the Karonga population with the percentage of people infected with subtype C increasing from 55% (n=11) in the early 1980s to 90% (n=168) by the late 1980s. By the late 1990s, subtype C was still the most prevalent subtype having risen slightly to the prevalence of 91 % (n=167) of HIV-1 infections. While the rate of increase is much reduced from the early to late 1980s, the prevalence of subtype C continues to rise and represented 94 % (n=156) of all HIV-1 infections between 2007 and 2009. Subtype D was only found as part of one recombinant strain in the 2000s and the presence of subtype A had fallen to below 2 %. The number of recombinant viruses had risen from 2.4 % to 4.5 % between the 1990s and 2000s with A/C recombinants representing over 80 % of the recombinant viruses. This overwhelming dominance of subtype C is also found in most of the countries surrounding Malawi (Zambia, Zimbabwe and Mozambique) (Abreu et al., 2008; Dalai et al., 2009; Handema et al., 2003).

The success of subtype C in establishing itself in the Karonga district is almost certainly due to a combination of factors. Unprotected heterosexual sex is the main transmission route in Karonga and an initial number of subtype C viruses introduced to the susceptible population could have resulted in a strong founder effect (Foley et al., 2000). The rapid spread of subtype B in the developed world as opposed to other subtypes (Hu et al., 1999) is similarly thought to be the result of a strong founder effect into the male homosexual and intravenous drug user population. Within Karonga, the success of subtype C is additionally compounded by the fact that subtype C is the most prevalent subtype found in most of the countries surrounding Malawi (Batra et al., 2000; Handema et al., 2003; Lahuerta et al., 2008) thus limiting the possible introduction of new subtypes. There is also evidence to suggest that subtype C viruses may be less virulent than other subtypes. This may result in longer periods of asymptomatic infections increasing the opportunity for onward transmissions (Arien, Vanham, and Arts, 2007; Ball et al., 2003). Further evidence of higher transmissibility rates in subtype C provides ideal conditions for subtype C to become the dominant subtype (Renjifo et al., 2001).

The disappearance of subtype D in Karonga may be due to its limited introduction into the region. This factor, in combination with the overwhelming success of subtype C, has resulted in an evolutionary 'dead end' in Karonga for subtype D. The continued presence of subtype A (although its prevalence continues to fall) and the increased prevalence of A/C recombinants was probably affected by introductions from the south of Tanzania. In the northern parts of Tanzania, which are geographically close to the Ugandan subtype D and Kenyan subtype A epidemics, subtypes A and D are more prominent (Arroyo et al., 2004). However, in the Mbeya, situated in the South West of Tanzania, subtype C represents approximately 40 % of infections, subtype A represents approximately 22 % of infections and A/C recombinants represent approximately 25 % of infections with subtype D still present as a minor variant (Herbinger et al., 2006; Arroyo, (2004). Traffic between Karonga and Tanzania is regular, allowing for the introduction of subtype A and A/C recombinants into Karonga. The introduction of different subtypes into an area has been previously attributed to commercial sex workers and the resulting sexual networks (e.g. in Thailand) (Ou et al., 1993). It is probable that subtype C will continue to dominate the epidemic within Karonga in the future. If so, this will have implications for drug and vaccine therapies in the future.

Short internal branches and lack of bootstrap support are features associated with the subtype C phylogeny, a pattern consistent with large numbers of short sequences that are closely related. There was also some structure evident within the phylogenetic trees. In earlier work by McCormack et al. (2002) four clusters were identified in the subtype C gene trees drawn using 1980s data from *env* and *gag* gene fragments. It was hypothesized that one cluster may have been introduced to Karonga from a single event sometime in the early 1980s. This subsequently spread and accounts for over 40 % of infections seen in the district in the later 1980s (McCormack et al., 2002; Travers et al., 2004). This group appeared to be largely restricted to sequences from Malawi (Malawi cluster). The Malawian cluster (Cluster 1) was still present in the *gag* phylogenetic tree here and had been expanded by sequences from 1990s and 2000s. Within the *env* phylogenetic reconstruction the cluster had separated into two and only approx. 17% of sequences fell into this group. In both gene trees very few sequences from different countries were found within the Malawi cluster. It is likely, therefore, that this cluster had its origins in Malawi, is restricted to the region and is

preserved by sexual networks within the region. A second smaller cluster (cluster 3) also previously described by McCormack et al. (2002) had similarly increased in size, with newly generated sequences from the 1990s and 2000s. It also contained very few sequences from other African countries indicating the presence of a possible second Malawi lineage circulating within Karonga

Other clusters (Cluster 2 and 4) described in McCormack et al were more radically altered with the addition of new sequences from 1990s and 2000s. The large cluster (cluster 2), composed of almost half of sequences from Malawi from the 1980s and most of the sequences from Africa and India, was no longer present in either gene tree when the new data from Karonga (from the 1990s and 2000s) and further sequences from India, Africa and Brazil were added. The new sequences from Karonga were interspersed among other subtype C sequences from other countries in Africa across the whole span of the tree. This observation supports the concept of the largely unrestricted movement of subtype C genotypes across the entire sub-continent (Bredell et al., 2007). This, in conjunction with a longstanding epidemic in Malawi and multiple introductions into the region from different geographic areas, results in multiple lineages within Karonga. A similar pattern of multiple lineages has been described in most of the Sub Saharan HV-1 subtype C epidemics (Batra et al., 2000; Novitsky et al., 1999; van Harmelen et al., 2001; Van Harmelen et al., 1999).

Within Karonga, two epidemic patterns appear to be present; a number of related geographically restricted circulating strains such as those in cluster 1 (Malawian cluster) and cluster 3 and a constant introduction of new viral strains from the surrounding regions resulting in multiple diverse lineages circulating within Karonga. This situation is somewhat similar to what has been reported in Ethiopia, where the HIV epidemic is thought to have begun around the same time (the 1960s) as it did in Malawi (Travers et al., 2004; Tully and Wood, 2010). Within Ethiopia, a geographically restricted lineage (subtype C') has been identified. Subtype C' is thought to be the result of a single introduction into Ethiopia and now represents approximately 50 % of infections there (Abebe et al., 2000). This is also supported by Thomson et al. (2011) who suggested a more well defined geographic structure to the subtype C epidemic within Africa that is masked by the presence of recombinant sequences. The study of the HIV-1 epidemic within Karonga over the past 20 years has

allowed us to track the history of the epidemic and has identified the emergence of two patterns for viral circulation within the region,.

The average genetic diversity seen in Karonga across all time frames was significantly lower than what was seen in subtype C epidemics in the surrounding counties such as Botswana, Mozambique, Tanzania, Zambia and South Africa. When compared to India and Brazil, two countries with newer subtype C epidemics (Bredell et al., 2007; Soares et al., 2005), the level of diversity in *env* is significantly higher in Karonga. The range of diversities was similar across the countries in Africa. The wide range of diversity in Karonga, in conjunction with a lower average diversity than in other African countries may reflect the two patterns of epidemiology present in rural Karonga discussed above That is, a number of related viruses circulating within Karonga resulting in a lower average diversity, while at the same time, a number of divergent sequences are leading to high maximum diversity.

Genetic distances measured between spouses from data collected in the late 1980s (McCormack et al., 2002) suggested a highest value of 2.3 % and 5 % for *gag* and *env* respectively, with most measures between pairs being less than these values (McCormack et al., 2002). Where data exists for spouse pairs of whom both were identified as HIV-1 positive only in the 1990s survey, the genetic distances were consistent with the values above, i.e. below 2.8% in gag and 5.9% in env above. Furthermore all clustered near to each other on the tree which is consistent with transmission from one spouse to another and less time for intra individual evolution to have occurred. Unsurprisingly, comparisons of sequences retrieved from HIV-1 positive individuals in the 1980s, and ten years later from their spouses in the 1990s, showed *gag* genetic distances to reach up to 4.9% and 8.8 % in *env* in most cases. In all these cases the genetic distances likely represent transmissions between the two individuals and that the higher genetic distance seen when compared to the 1980s data is due to the length of time that the virus may have evolved within the spouse since transmission. However, over half of sequences from theses spouse pairs did not group together on the trees highlighting some limitations of using phylogenetic methods to indicate transmission. The genetic distance between spouse pair 11 was 7.9 % in *gag*. The average genetic distance between all unrelated viruses amongst individuals seen in the 1990s was 6.7 % (range 0.2 % to 15 %). The virus was also very distantly related in *env* for this pair (10.5 %). In this case,

it is possible that the sequences present in the pair were from unrelated viruses and do not represent transmission. This is also a possibility for spouse pair 14 where the average genetic distance in *env* was 10.2 %.

Previous studies on co receptor tropism in subtype C isolates reported a dominance of CCR5 tropic isolates in all stages of disease progression, including advanced disease with minimal switch to CXCR4 tropism in the 1990s (Choge et al., 2006; Ndung'u et al., 2006), By 2008 only 30 CXC4 utilizing subtype C isolates had been reported in the literature (Cecilia et al., 2000; Choge et al., 2006; Cilliers et al., 2003; Engelbrecht et al., 2001; Michler et al., 2008; Morris et al., 2001; Papathanasopoulos et al., 2003; Ping et al., 1999). Within Karonga, five CXCR4 tropic viruses were predicted in the 1980s and a further eight were predicted in sequences from the 1990s and by the 2000s a further 21 had been identified. As the co receptor tropism in Karonga was predicted using bioinformatic tools (PSSM (Jensen et al., 2006)) based on proviral DNA the results should be treated with some caution. The subtype C epidemic is found to exist mostly in the poorest regions of the world making it diffucult in the past to carry out the costly and time consuming phenotypic tropism assays. This has almost certainly impacted on the low levels of CXCR4 tropism detection within the subtype C epidemic. The C-PSSM web based tool (Jensen et al., 2006) is a bioinformatic tool which has been trained on subtype C viruses. This tool could be used on all V3 sequences retrieved from the subtype C epidemic since the first detection of subtype C viruses to reevaluate the presence of subtype C viruses that use CXCR4 and could potentially reveal a higher prevalence of CXCR4 tropism than previously documented.

What was most striking in the Karonga data set was the rise in the prevalence of CXC4 tropic viruses from 3.9 % to 8 % to 24.7 % (n= 127, 100, and 84) in 1980s, 1990s and 2000s. This increase may be linked to the increase in diversity over the same time period also observed in Karonga. The stage of disease progression, however, is not known for these individuals so it is not clear if this increase in CXCR4 prevalence is associated with rapid $CD4^+$ T cell decline and progression to AIDS as is observed in subtype B (Connor et al., 1997; Scarlatti et al., 1997). Increased prevalence of subtype C CXCR4 tropic viruses has recently been reported in two cohorts of individuals who had progressed to AIDS or were on ART in Malawi and South Africa.

These two studies reported that approximately 30 % (n=20 and 52) of subtype C viruses isolated were CXCR4 tropic (Connell et al., 2008; Raymond et al., 2010). These studies focused on individuals showing disease progression however we have found evidence of CXCR4 tropism in non progressing individuals (Chapter 3). Further studies are needed to explore the expansion of CXCR4 tropic viruses within the subtype C epidemic.

The subtype C epidemic has shown that the virus is continuing to evolve and change over time in Karonga. The implications of increased diversity and increased CXCR4 detection are not yet known. The accumulation of diversity will certainly make it more difficult for any vaccine to be effective within the worst affected countries and worldwide. The emergence of CXCR4 will need to be monitored closely as this may affect future use of CCR5 antagonists, which have recently been introduced as part of HIV drug therapy (Dorr et al., 2005). ART has recently been introduced into Karonga (Bansode et al., 2010), but only a limited number of drugs are available. Diversity is continuing to increase, making the accumulation of drug mutations highly likely, and may impact of the efficacy of the current ART program.

Chapter 5: HIV-1 A1/C Recombinant viruses in Karonga District Malawi

5 Chapter 5

5.1 Introduction

Three phylogenetic groups of HIV-1 are currently recognised, groups M (main), O (outlier), and N (non-M, non-O) (Korber and Daniels, 1997; Leitner et al., 1996; Simon et al., 1998). Within group M there are nine subtypes identified by the letters A-D, F-H, J and K, all of which are thought to have originated in central Africa (Buonaguro, Tornesello, and Buonaguro, 2007). The co-circulation of different subtypes in the same geographic regions followed by the dual infection of individuals with viruses from different subtypes has led to the generation of inter-subtype recombinant forms that are characterised by having distinct parts of their genome matching the consensus sequences of different subtypes. The region within the recombinant genome where one sequence switches to another sequence is defined as a "breakpoint" (Ramirez et al., 2008). In some cases, the isolates from epidemiologically unlinked individuals have identical mosaic genomes with identical breakpoints, and are believed to have come about due to recombination event. If such an isolate plays an important role in the HIV pandemic they are identified as circulating recombinant forms (CRFs). In order to identify a CRF, three full genomes or two full genomes and a third partial genome is required to be sequenced from individuals who are not epidemiologically linked (Robertson et al., 2000). CRFs are identified by numbers correlating to the order of their discovery, followed by the letters of the parental subtypes (U indicating unclassified segments), or cpx (complex) for recombinant viruses made up of three or more parental subtypes. Currently 49 CRFs have been described (Table 5.1) (http://www.hiv.lanl.gov/). The importance of CRFs in the global HIV-1 pandemic is increasingly recognised. In a recent survey they were found to make up 18 % of infections worldwide (Buonaguro, Tornesello, and Buonaguro, 2007). Unique recombinant forms (URFs) are seen in only one person or in a few epidemiologically linked persons. CRFs are considered to be more significant than unique recombinant forms (URFs) as some CRFs are responsible for disproportionate numbers of infections globally (Fan, Negroni, and Robertson, 2007). On the other hand, extensive numbers of URFs have been isolated from individual patients and are recognised to be abundant throughout the world. A reliable estimate of their global incidence in the pandemic has still to be established (Ramirez et al., 2008).

Table 5.1 A list of off all 49 CRFs described to date. They are listed in the order by which they were originally described. Modified from LANL Sequence Database (http://www.hiv.lanl.gov/).

Name	Reference Strain	Subtypes
CRF01_AE	CM240	A, E
CRF02_AG	IbNG	A, G
CRF03_AB	Kal153	A, B
CRF04_cpx	94CY032	A, G, H, K, U
CRF05_DF	VI1310	D, F
CRF06_cpx	BFP90	A, G, J, K
CRF07_BC	CN54	B, C
CRF08_BC	CX-6F	B, C
CRF09_cpx	96GH2911	A, G, U
CRF10_CD	TZBF061	C, D
CRF11_cpx	GR17	A, CRF01, G, J
CRF12_BF	ARMA159	B, F
CRF13_cpx	96CM-1849	A, CRF01, G, J, U
CRF14_BG	X397	B, G
CRF15_01B	99TH.MU2079	CRF01, B
CRF16_A2D	97DR004	A2, D
CRF17_BF	ARMA038	B, F
CRF18_cpx	CU76	A1, F, G, H, K, U
CRF19_cpx	CU7	A1, D, G
CRF20_BG	Cu103	B, G
CRF21_A2D	99KE_KER2003	A2, D
CRF22_01A1	CM001BBY	CRF01, A1
CRF23_BG	CB118	B, G
CRF24_BG	CB378	B, G
CRF25_cpx	02CM_1918LE	A, G, U
CRF26_AU	MBTB047	A ,U
CRF27_cpx	04FR-KZS	A, E, G, H, J, K, U
CRF28_BF	BREPM12609	B, F
CRF29_BF	BREPM16074	B, F
CRF30_0206	NE36	CRF02, CRF06
CRF31_BC	04BR142	B, C
CRF32_06A1	EE0369	CRF06, A1
CRF33_01B	05MYKL007	CRF01, B
CRF34_01B	OUR2275P	CRF01, B
CRF35_AD	AF095	A, D
CRF36_ cpx	NYU830	A, G, CRF01, CRF02
CRF37_cpx	NYU926	A, G, CRF01, CRF02, U
CRF38_BF	UY03_3389	B, F1
CRF39_BF	03BRRJ103	B, F1
CRF40_BF	05BRRJ055	B, F
CRF41_CD	CO650V1	C, D
CRF42_BF	luBF_13_05	B, F1
CRF43_02G	J11223	CRF02, B
CRF44_BF	CH80	B, F1
CRF45_cpx	04FR.AKU	A, K, U
CRF46_BF	01BR087	B, F1
CRF47_FB	P1942	B, C
CRF48_01B	07MYKT014	CRF01, B
CRF49_cpx	N28353	A1, C, J, K, U

5.1.1 The recombination process in HIV

Recombination is an important property of retroviruses. During every round of proviral DNA synthesis, reverse transcriptase is required to carry out two template switches. These template jumps by reverse transcriptase are thought to make retroviruses prone to recombination (Coffin, 1979; Onafuwa-Nuga and Telesnitsky, 2009; Temin, 1993). In order for recombination to take place, it is first necessary for two divergent viruses to infect the same cell. Originally it was proposed that a HIV infected cell would become resistant to dual infection due to the down regulation of the CD4 receptor required for viral entry (Piguet et al., 1999). However, the widespread occurrences of inter-subtype recombinant viruses has provided indirect evidence for the frequent occurrences of dual infection in a single cell. Further evidence was provided by a study using fluorescence *in situ* hybridization, which identified up to eight different proviruses per cell in the spleen (Jung et al., 2002). Dual infection can lead to the integration of two variant viruses into the genome of the infected cell. The resulting viral progeny will be composed of both heterozygous and homozygous virions. Infection of a second target cell by a heterozygous virion will result in the generation of a chimeric provirus during reverse transcription which can subsequently lead to the production of recombinant HIV viral particles (Galetto and Negroni, 2005) (Figure 5.1).

Recombination is thought to be the result of template switching by reverse transcriptase during synthesis of the first DNA strand (Vogt, 1973). The basic process described below is known as “copy choice recombination” (Figure 5.2). After the RNA genome (the donor RNA) is copied, the RNaseH activity of reverse transcriptase degrades the RNA genome creating a region of single stranded (-) DNA (ssDNA) that is complementary to the second copy of the RNA genome (acceptor RNA) in the viral particle (Krug and Berger, 1989; Oyama et al., 1989; Schultz, Zhang, and Champoux, 2004; Wisniewski et al., 2000). The 3’ end of the growing ssDNA is transferred and annealed to the acceptor RNA strand in a process known as “the docking step” for which a template switch by reverse transcriptase is required (Ramirez et al., 2008). There are a number of proposed hypothesises to explain why this template switch occurs (Figure 5.2).

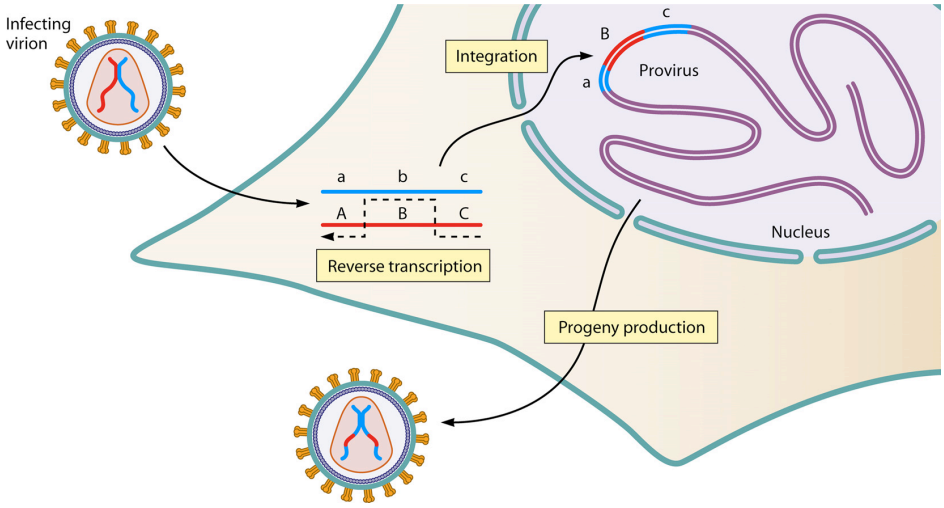


Figure 5.1 Outline of the recombination process. Viral particles package two genetically distinct copies of the viral RNA (red and blue lines). If a template switch occurs during reverse transcription a recombinant virus can be created (Onafuwa-Nuga and Telesnitsky, 2009).

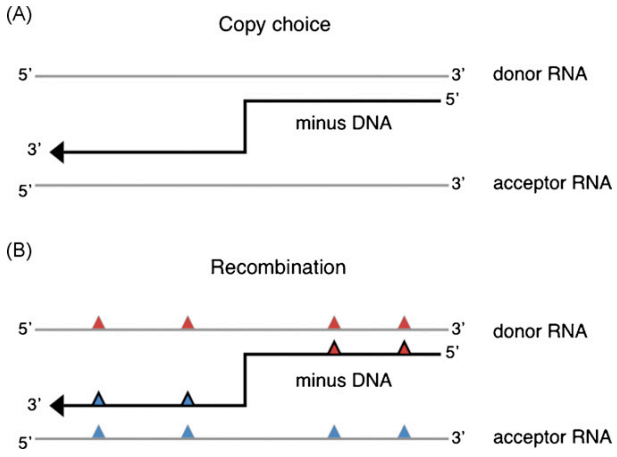


Figure 5.2 Copy choice recombination (panel A) and Recombination generated by copy choice (panel B). The black arrow gives the direction of reverse transcriptase switching strands. The blue and red triangles represent different amino acids (Ramirez et al., 2008).

The presence of breaks in the genomic RNA was the earliest hypotheses proposed to account for the template switch in copy-choice recombination (Coffin, 1979). Alternatively, the presence of strong natural pause sites on the RNA genome led to the suggestion that pausing of the reverse transcriptase enzyme constituted a trigger for template switching (DeStefano, Bambara, and Fay, 1994; DeStefano et al., 1992). Another possible factor responsible for the generation of preferential sites for recombination is the presence of secondary structured regions on the RNA template such as dimmer initiation sequences (Balakrishnan, Fay, and Bambara, 2001).

5.1.2 Effects of recombination in HIV-1

Recombination contributes to the evolution of HIV-1 on several levels, from the origins of the virus to its adaptability in individual patients (Telesnitsky and Goff, 1993). The chimpanzee virus that gave rise to HIV-1 is thought to be a recombinant form of other SIVs (Heeney, Dalgleish, and Weiss, 2006) and phylogenetic analysis has suggested that HIV-1 group N arose from recombination between an SIV and an early form of group M (Gao et al., 1999). Within an individual patient, recombination is responsible for increased viral diversity by rearranging different viral variants into new distinct viral variants (Charpentier et al., 2006; Moutouh, Corbeil, and Richman, 1996; Rambaut et al., 2004). It is thought that the ability to maintain extensive diversity is important for viral pathogenesis as it allows the virus to evade the host immune responses or to overcome anti-retroviral therapy (Charpentier et al., 2006). Recombination is therefore thought to be an advantage to changing evolutionary pressures within the host while also being able to repair deleterious deletions or mutations (Charpentier et al., 2006).

Recombination can also link together mutations that can then provide resistance to anti-retroviral drugs (Fraser, 2005). *In vitro* experiments have provided evidence that recombination can speed up the emergence of drug resistant viral strains (Gu et al., 1995; Kellam and Larder, 1995; Moutouh, Corbeil, and Richman, 1996) and more recently this has been observed *in vivo.* The evolution of a viral quasi species population was followed within a patient both before and after anti retroviral treatment (Nora et al., 2007). The emergence of resistance to the antiviral treatment directed against reverse transcriptase and protease were identified as the consequence of recombination involving three different

sequences that had pre existed in the population before treatment, plus an additional mutation that was thought to be generated by base substitution (Nora et al., 2007). For most communicable diseases, vaccines are often the most cost-effective and viable control strategy as in the case of polio, influenza and hepatitis. The extraordinary worldwide diversity of HIV-1 presents the greatest hurdle to vaccine development and as it stands, the breadth of cross subtype immune protection provided by vaccine strategies focused on current subtypes is still unclear. The continuous emergence of new recombinant strains potentially threatens strategies aimed at providing protection against "non recombinant" viral subtypes (Barouch, 2008). It is vital therefore, to be aware of the viral variants being transmitted within a target population, be they pure subtypes or recombinant subtypes, for the development of a HIV-1 vaccine.

5.1.3 Global distribution of CRFs

CRFs have now been found in almost all regions where HIV-1 infection occurs (Figure 5.3). The first CRFs were characterised by full-length sequence analysis of HIV from Thailand (CRF01_AE) and central Africa (CRF02_AG). CRF01_AE, which currently circulates mostly in South East Asia, is one of the most epidemiologically abundant CRFs, and is responsible for almost 5 % of infections worldwide. Originally, it was designated as subtype E based on partial *env* sequences. Further phylogenetic analysis based on full genomic sequences identified large portions of the genome (*gag* and *pol*) as subtype A. The accessory genes and *env* were hybrids between subtypes A and E (Carr et al., 1996; Gao et al., 1996). The origin of subtype E is as of yet unknown, however, it is has been proposed that a single recombination event in Africa between a subtype A and E virus preceded the introduction of the CRF01_AE recombinant virus in Thailand (Gao et al., 1996).

The second CRF discovered, CRF02_AG, makes up 5 % of infections world wide and is the most prevalent subtype in West and parts of Central Africa (Hemelaar et al., 2006). CRF02_AG was already widespread early on in the HIV-1 pandemic. By 1999 it was already more prevalent than the parental strain subtype G in West Central Africa where subtype G was thought to have originated (Abecasis et al., 2007; Carr et al., 1999; Carr et al., 1998) (Montavon C, 1999). More recent analysis has established that subtype G is a

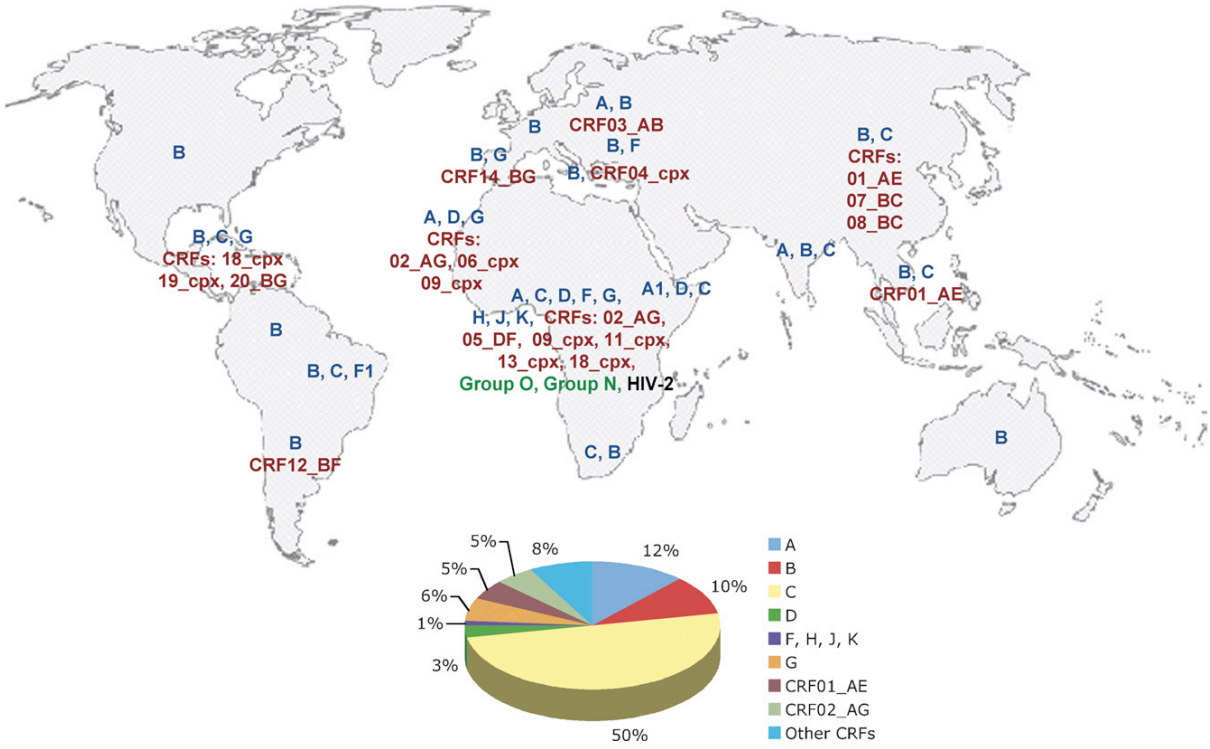


Figure 5.3 Geographic distribution of HIV genetic forms. The approximate location of the different HIV subtypes is indicated. HIV-1 group M pure subtypes are indicated in blue, while CRFs are in red. The other HIV-1 groups O and N and HIV-2 are indicated in green and black, respectively. The pie chart gives the prevalence of HIV-1 group M genetic forms. The global prevalence of each subtype and CRFs is expressed as the percentage of the total number of group M HIV-1 isolates identified worldwide (data from (Buonaguro, Tornesello, and Buonaguro, 2007) and Figure from (Ramirez et al., 2008)).

recombinant virus itself made up of subtypes A and J and a putative subtype G parent (Abecasis et al., 2007). Consequently it has been proposed that subtype G is in fact a recombinant form of the parental lineage CFR02_AG (Abecasis et al., 2007). The remaining CRFs have a lesser epidemic relevance on a global scale.

No CRFs comprising solely of subtypes A1 and C have yet been described, which requires the sequencing of three full length genomes containing the same recombinant pattern from three epidemiologically unlinked individuals (Robertson et al., 1995). This is in spite of the fact that subtype C is the most prevalent subtype globally and is responsible for 50 % of the worldwide infections. Subtype A1 is the second most prevalent subtype and is responsible for 12 % of worldwide infections. Both subtypes are also found to co exist in a number of large geographic regions such as central Africa and India (Arien, Vanham, and Arts, 2007; Buonaguro, Tornesello, and Buonaguro, 2007). Recently however de Silva et al. (2010) described a URF composed of subtypes A1 and C together with subtypes J, K, and U (CRF49_cpx) in Gambia. The Los Alamos sequence database lists 355 A1/C recombinant HIV-1 sequences (URFs) from both partial and complete *gag, pol* and *env.* Currently there are eight full A1/C recombinant genome sequences found within one individual in the LANL database each with a unique mosaic pattern (Figure 5.4). To date these viruses have only been found in the single individual they were isolated from as follows, five in Tanzania, one in Kenya, one in India and one in Ethiopia.

5.1.4 Recombinant viruses in Karonga

Between 1982-1984 the percentage of individuals in Karonga District, Malawi, infected with subtype C was 55 %, the percentage of those infected with subtype A and D was 9 % each, and the percentage of those infected with an unclassifiable subtype was 27 %. By 1989 the percentage of individuals infected with subtype C had risen to 90 %. Also by this time three recombinant forms were identified (*env/gag:* D/A, two C/A and C/D) (McCormack et al., 2002). Four more recombinant viruses were recorded between 1997-2001 (Chapter 4) (*env/gag:* A/C, two C/A, and C/D) and the prevalence of those infected with subtype C had risen to 91 %. By 2008, nearly all *env* and *gag* sequences were phylogenetically identified as subtype C with subtype D having disappeared. One *gag* sequence and three *env* sequences were identified as subtype A1 (Chapter 4). Three individuals harboured viruses that were

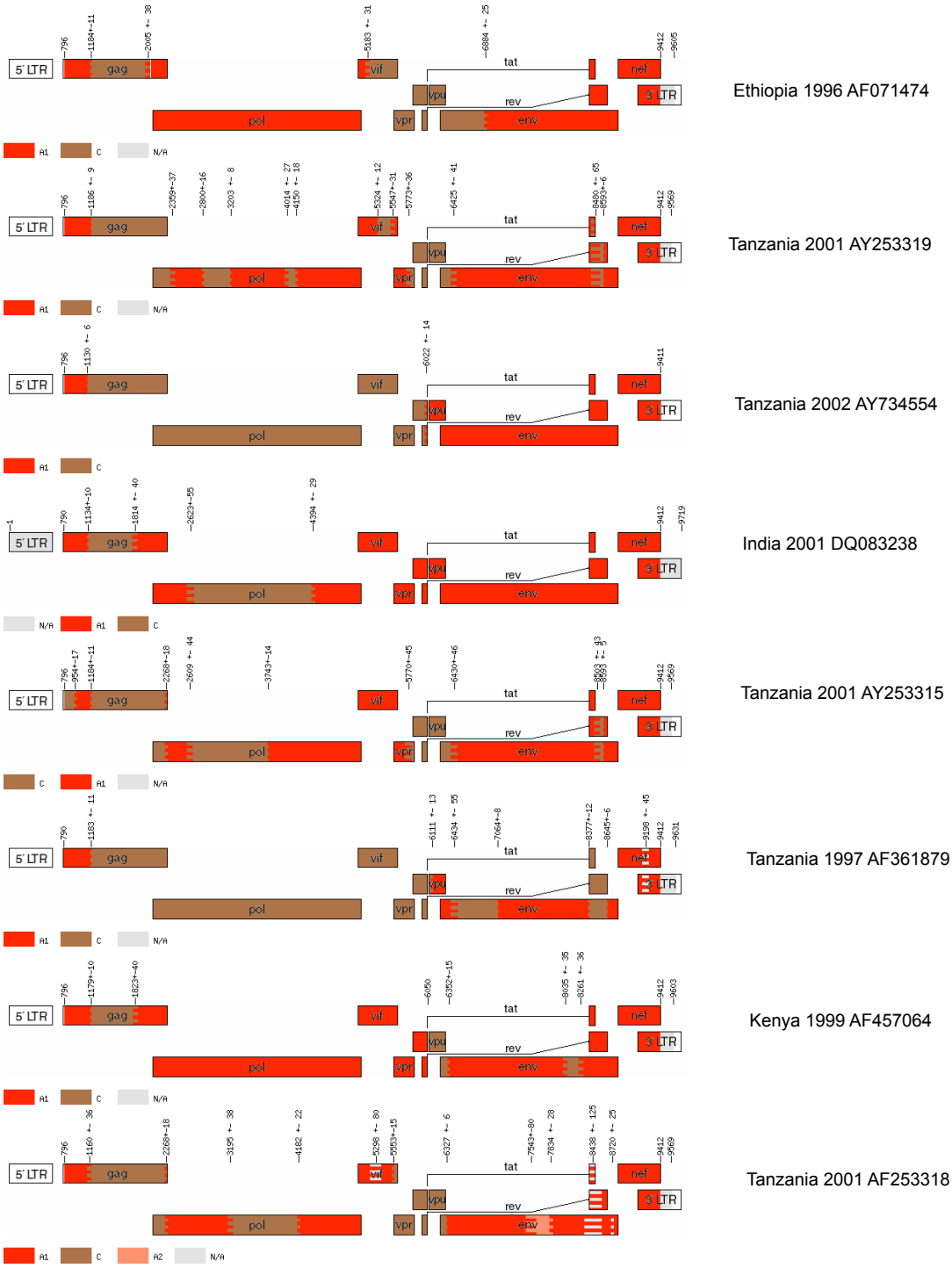


Figure 5.4 Mosaic structures of the HIV-1 URFs recorded in the LANL database made up of subtypes A1 and C. Subtype A1 is indicated by red and subtype C is indicated by brown. They are labelled with the country of origin, year, and accession number. The recombinant genome from Tanzania 2001 AF253318 is composed mostly of A1 and C however it does contain a small amount of A2 within *env*. (Reproduced with minor modifications from the LANL sequence database http://www.hiv.lanl.gov/ and (Schultz et al., 2009).

different subtypes within *env* and *gag* (*env/gag*: two A1/C and C/A1). A further three A1/C recombinants with identical breakpoints within *gag* were detected from three epidemiologically unlinked individuals. The identification of a unique identical breakpoint within these three individuals thus prompted further investigation. The aim of this chapter was to

a) characterise the A1/C recombinants further and,

b) attempt to produce full genome sequences of all three A1/C recombinants in order to identify and describe a new CRF in the Karonga district of Malawi.

5.2 Materials and methods

5.2.1 Sample information

In 2008, 71 whole blood samples were collected from HIV-1 positive individuals in Karonga District, Malawi as part of a much larger observational cohort study screening for ART suitability, side affects of ART and the general health of the individual (Bansode et al., 2010). The 71 samples were chosen randomly for this study to explore the molecular evolution of HIV-1 in Karonga. All samples were separated by local laboratory technicians by centrifugation into plasma and cell pellet and were stored at – 20 °C. DNA was extracted in Karonga from cell pellets using the QIAamp DNA Blood Mini Kit (Qiagen) according to the manufacture's instructions. The extracted proviral DNA was transported to the Molecular Evolution and Systematic Laboratory in NUI Galway.

5.2.2 PCR amplification and sequencing of *gag* (p17/p24)

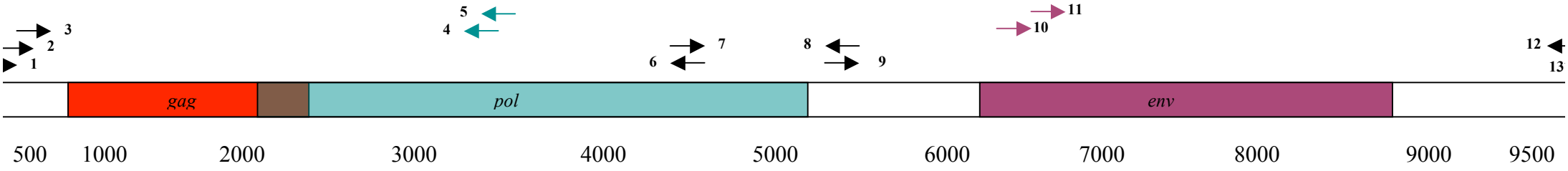
Nested PCR was used to amplify a fragment of *gag* spanning part of p17/p24 (750 bp) and a fragment of *env* encompassing the C2 to V3 region (549 bp) as described in McCormack et al, (2002). After verification of fragment size by gel electrophoresis, the PCR products were purified using HiYield Gel/PCR DNA Fragments Extraction Kit (Real Genomics) according to the manufacturer's instructions. The resulting purified PCR products were sequenced in both directions by LGC Genomics (Germany). DNA sequences were assembled and edited in Seqman 8.0.2 (DNASTAR). Any positions where sequencing ambiguities were present were assigned the appropriate IUPAC designation.

5.2.3 PCR amplification of the full genome

Five μl of extracted proviral DNA was visualised by gel electrophoresis to ascertain the presence of high molecular weight DNA. A nanodrop spectrophotometer (Thermo Scientific) was used to quantify the amount of DNA present in each sample. Amplification of virtually full-length HIV-1 genomes in one continuous segment was attempted using a well-described procedure (Carr et al., 1999; Papathanasopoulos et al., 2003; Piyasirisilp et al., 2000; Streeck et al., 2008; Tzitzivacos et al., 2009). Briefly, virtually full-length genomes were amplified in a nested PCR using the Expand Long Template PCR System (Roche

Table 5.2 A list of all the primer sequences used in the PCR for full genome amplification. HXB2 numbering is used to identify the position of the primers within the HIV-1 genome. Tm represents the melting point of each primer.

Primer name	Primer Composition	Tm (°C)	HXB2 position	Orientation	Reference
FG5F	GGTCTCTCTGGTTAGACCAG	59.4	455	Forward	Fang et al 2001
MSF12	AAATCTCTAGCAGTGGCGCCCGAACAG	68.0	623	Forward	Salminem et al 1995
F2nst	GCGGAGGCTAGAAGGAGAGAGATGG	67.9	768	Forward	Van Harmelen et al 2001
DRRT15	TCCCACTAACTTCTGTATATC	54.0	3320	Reverse	Handema et al 2003
DRRT4	TTCTGTTAGTGCTTTGGC	51.4	3420	Reverse	Handema et al 2003
FGF46for	GCATTCCCTACAATCCCCAAAG	60.3	4647	Forward	Fang et al 2001
FGF46rev	CTTTGGGGATTGTAGGGAATGC	60.3	4669	Reverse	Fang et al 2001
FGR53for	GGAGGAAAAAGAGATATAGCACACAATGAGAC	65.6	5305	Forward	Fang et al 2001
FGR53rev	GTCTACTTGTGTGCTATATCTCTTTTTCCTCC	65.6	5347	Reverse	Fang et al 2001
ED31	CCTCAGCCATTACACAGGCCTGTCCA AG	69.5	6840	Forward	McCormack et al 2002
ED5	ATGGGATCAAAGCCTAAAGCCATGTG	63.2	7380	Forward	McCormack et al 2002
MSR5	GCACTCAAGGCAAGCTTTATTGAGGCTTA	65.3	9603	Reverse	Salminem et al 1995
ofm-r1	TGAGGGATCTCTAGTTACCAGAGTC	63.0	9661	Reverse	Van Harmelen et al 2001



1	FG5F
2	MSF12
3	F2nst
4	DRRT15
5	DRT4
6	FGF46rev
7	FGF46for
8	FGR53rev
9	FGR53for
10	ED5
11	ED31
12	MSR5
13	ofm-r1

Figure 5.5 A schematic representation of the HIV-1 genome and approximate position of primers used in the PCR for the amplification of the full HIV-1 genome. The arrows represent the position of the primers. Forward facing arrows are forward primers and backward facing arrows are reverse primers. The primer names are listed below.

Applied Science). The outer nested primers were MSF12 and ofm-r1 and the inner nested primers were F2nst and MSR5 (Details in Table 5.2 and Figure 5.5). These primers were designed to amplify all but a small section of the 5' Long Terminal Repeat and have been used to successfully amplify a number of different HIV-1 subtypes including subtype C and A1. Cycling conditions for both the first and the second round of the nested PCR were the same and included a hot start (94 °C for 2 m) followed by 10 cycles of denaturation (94 °C) for 10 s, annealing (60 °C) for 30 s, and extension (68 °C) for 10 minutes. This was followed by 20 cycles with a lower annealing temperature (55 °C). A second programme including an 8 m extension period and a 15 s increment after every cycle was also attempted. Extracted proviral DNA (5 – 15 μl) was used as the template in a PCR of 50 μl with 5 μl of 10 X Buffer 2 (containing 17.5 mM $MgCl_2$), 500 μM dNTPs (each), 20 μM of each primer and 0.75 μl of Expand Long Template enzyme. For the second round of the nested PCR 5 μl of the primary product was used as a template.

Amplification of the full genome in two overlapping fragments was also attempted with a number of different primer sets (Table 5.2). This was first attempted with the primers FG5F and FGR53rev for the 5' end described by Fang et al. 2001. When this proved unsuccessful a number of primer combinations (seen in Table 5.2 and Figure 5.5) in both nested and hemi nested PCRs were attempted, including a number of in-house primers that had previously resulted in amplification of partial *env* and partial *pol* (ED5, ED31, DRRT15 and DRT4). The initial thermo cycling protocol was a hot start (94 °C for 2 m) followed by 10 cycles of denaturation (94 °C) for 10 s, annealing (60 °C) for 30 s, and extension (68 °C) for 4 minutes. This was followed by 20 cycles with the annealing temperature reduced (55 °C). A range of lower annealing temperatures was also employed. The amount of proviral DNA used in both rounds of the nested PCRs was varied from 1 μl up to 20 μl in the primary PCR and 1 μl to 10 μl in the secondary PCR, for each sample to account for different amounts of provirus within each individual. PCR products were then visualised on a 0.6 % agarose gel.

5.2.4 Phylogenetic analysis

Two multiple sequence alignments of all successfully amplified sequences from the 71 samples along with previously amplified *gag* and *env* sequences from Karonga (McCormack et al., 2002) were generated along with reference sequences of the main subtypes downloaded

from the LANL database (http://www.hiv.lanl.gov/). The sequences were manually aligned in MacClade 4 (Sinauer Associates) and gaps were removed along with areas of ambiguous alignment. Phylogenetic trees were reconstructed under the GTR + gamma model of DNA substitution implemented by RAxML 7.0.3 (Stamatakis, 2006) with all parameters optimised by RAxML. Confidence levels in the groupings in the phylogeny were assessed using 1000 bootstrap replicates as part of the RAxML phylogeny reconstruction. A bootstrap value equal to or greater than 70 % was considered significant.

5.2.5 Recombination analysis

For all putative recombinants identified from the phylogenetic analysis the Recombination Identification Program 3.0 (RIP) (Siepel et al., 1995) was used to further identify any potential mosaic structures using a window size of 200. Briefly, RIP identifies recombination in a query sequence by calculating its similarity to a background alignment of HIV-1 sequences of different subtypes in a sliding window that is moved in increments of one nucleotide residue from left to right in the alignment. In order to further determine the approximate breakpoints in the mosaic structures the jumping profile Hidden Markov Model application (jpHMM) (Zhang et al., 2006) was used. jpHMM predicts whether any HIV-1 genomic sequence is composed of different subtypes. It then estimates the recombination breakpoints and assigns a parental subtype among the major HIV-1 subtypes to each segment on either side of the breakpoint (Schultz et al., 2009). The LANL sequence database (http://www.hiv.lanl.gov/) was searched for any partial *gag* sequences (p17/p24) that are part of an A1/C recombinant genome. RIP was then applied to the downloaded sequences in order to identify the recombinant breakpoints in the *gag* sequences in order to compare them to the breakpoints found within the Karonga A1/C recombinants.

5.3 Results

5.3.1 Phylogeny of the recombinant viruses

Of the 71 samples, partial *gag* was successfully sequenced for 67 samples and partial *env* was sequenced for 55 samples. Phylogenetic analysis of *gag* showed three samples, 54702, 54574 and 54653, to group within subtype A as a monophyletic clade with high bootstrap (Figure 5.6). They did not group with any other subtype A sequences from Karonga and did not group within A1 but were seen to form a sister clade to the Subtype A2 reference sequences. This grouping was not supported by bootstrap. Phylogenetic analysis of partial *env* identified all three samples as subtype C. The *env* sequence from 54653 was seen to still group with 54702 with significant bootstrap support, however, sequence 54574 was seen to group within subtype C but on a different branch from the other two samples. Three other samples were seen to cluster with different subtypes depending on the gene region examined. One sample grouped with subtype C in *gag* but subtype A1 in *env* while two other samples grouped within subtype A1 in *env* and within subtype C in *gag* (chapter 4).

5.3.2 Identification of A1/C recombinant viruses

RIP identified 54574, 54653 and 54702 as being A1/C recombinants within *gag* (Figure 5.7), and identified *env* as being subtype C only within all three sequences. On the other hand three recombinants showed no evidence of recombination within the two gene fragments (*env* and *gag*) studied here. Eighty-three partial *gag* sequences found as part of an A1/C recombinant genome were located in LANL and were downloaded (Table 5.3). Further analysis using RIP identified 58 sequences as either pure subtype A or C with the p17/p24 fragment. The remaining 25 sequences were found to contain an A1/C recombinant breakpoint within the *gag* fragment. Of these 25 sequences, 18 were subtype A1 first followed by subtype C within the *gag* fragment similar to the sequences from Karonga and the other 17 sequences were the opposite, subtype C followed by subtype A1. One sequence (AY253315 from Tanzania) contained two breakpoints with the first third of the p17/p24 *gag* fragment identified as subtype C, the second third was identified as subtype A1 and the final section was identified subtype A1.

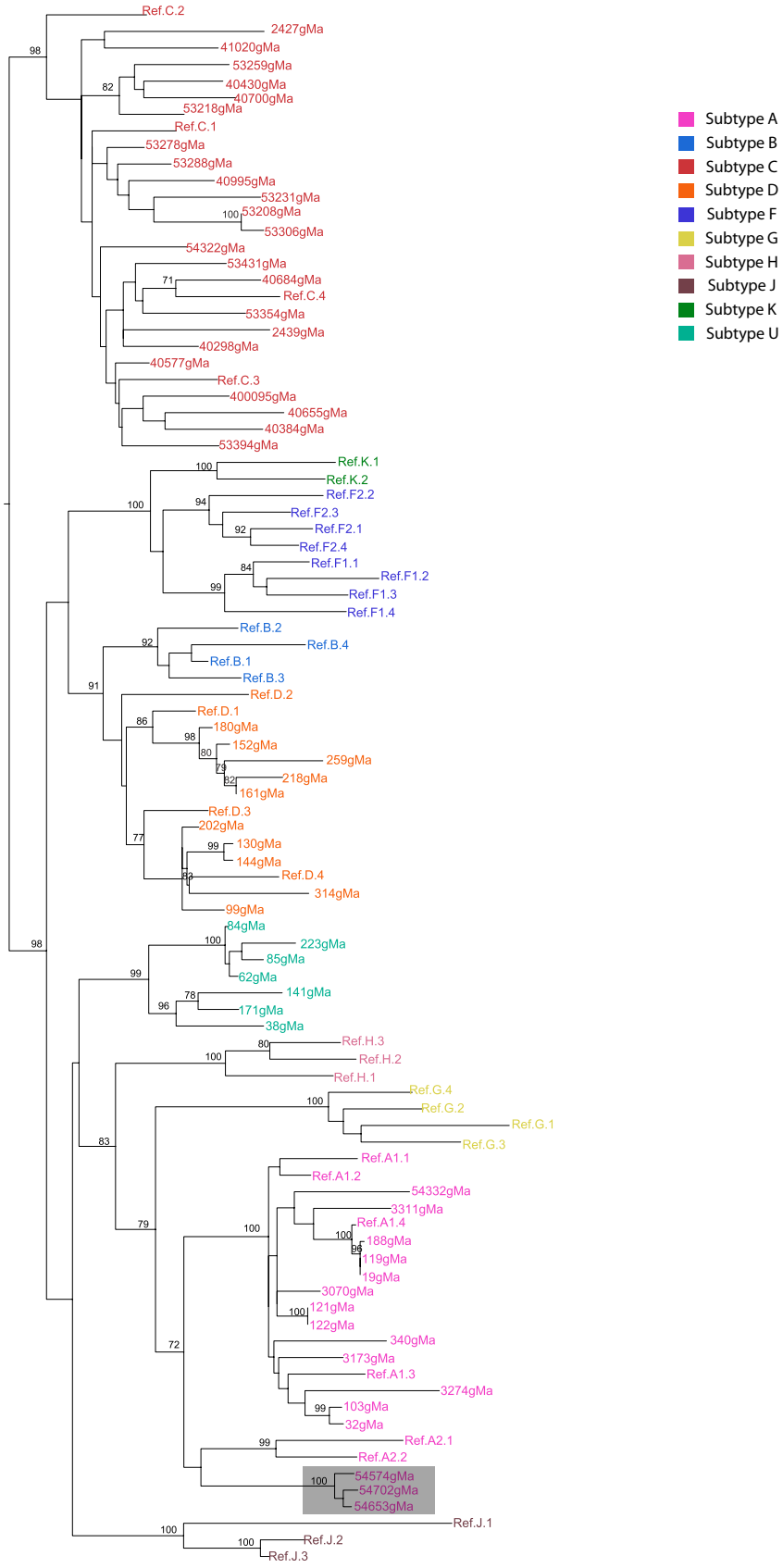


Figure 5.6 Maximum Likelihood tree generated from *gag* sequences with subtype reference sequences from the LANL sequence database (http://www.hiv.lanl.gov/), 20 subtype C sequences and all non subtype C sequences collected from Karonga. Sequences are either labelled as Ref (indicating a reference sequence from the LANL sequence database) and the subtype or with the sample name, g (*gag*) and Ma (Malawi). The different subtypes are individually coloured and the three A1/C recombinant sequences are enclosed in a grey box. Bootstrap values of over 70 are marked on the relevant branches.

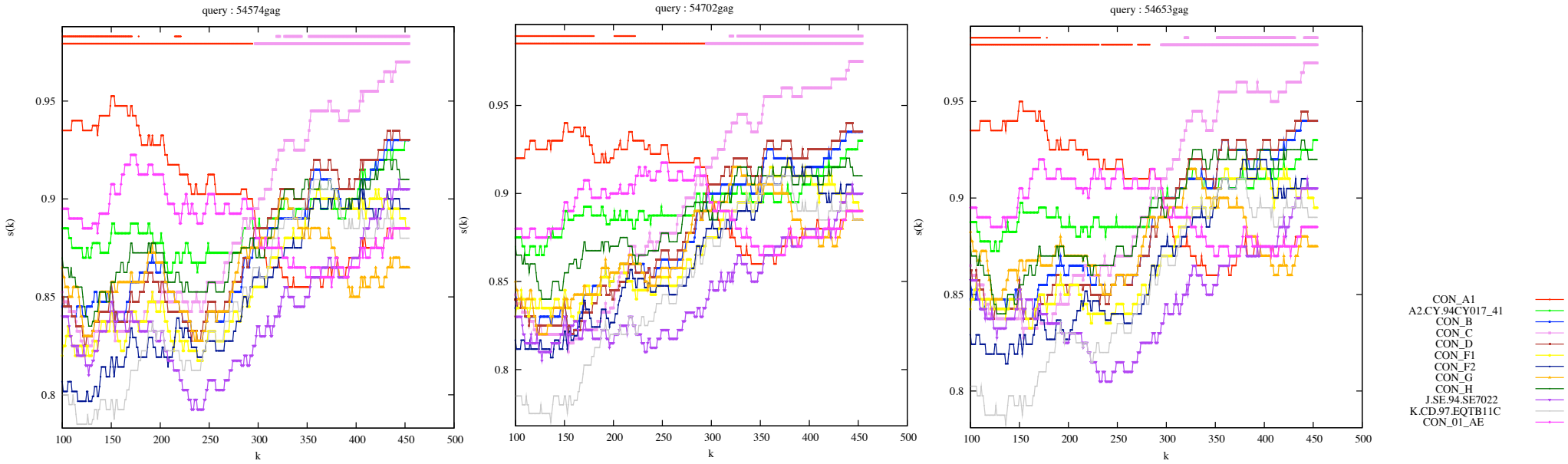


Figure 5.7 The RIP results for the three A1/C recombinants. The coloured curves trace the similarity between the query sequences and subtype consensus sequence. Consensus A1 is red and consensus C is pink. All three query sequences show the largest similarity to consensus A1 for approximately the first half of the sequence and for consensus C for the second half of the sequence.

Table 5.3. A list of all 83 sequences found in the LANL sequence database that are from an A1C recombinant genome. Subsequently the subtype found within the gag17/p24 fragment is also identified.

Sequence name	Accession number	Year	Country	Subtype within gag p17/p24 fragment
clone_1090	GQ430954	1990	Kenya	A1
clone_1089	GQ430953	1990	Kenya	A1
clone_1088	GQ430952	1990	Kenya	A1
clone_1086	GQ430950	1990	Kenya	A1
clone_1084	GQ430948	1990	Kenya	A1
clone_1083	GQ430947	1990	Kenya	A1
clone_1082	GQ430946	1990	Kenya	A1
clone_1081	GQ430945	1990	Kenya	A1
clone_1080	GQ430944	1990	Kenya	A1
clone_1079	GQ430943	1990	Kenya	A1
clone_1078	GQ430942	1990	Kenya	A1
clone_1076	GQ430940	1990	Kenya	A1
clone_1075	GQ430939	1990	Kenya	A1
clone_1074	GQ430938	1990	Kenya	A1
clone_1073	GQ430937	1990	Kenya	A1
clone_1072	GQ430936	1990	Kenya	A1
clone_1071	GQ430935	1990	Kenya	A1
clone_1068	GQ430932	1990	Kenya	A1
clone_1066	GQ430930	1990	Kenya	A1
clone_1064	GQ430928	1990	Kenya	A1
clone_1063	GQ430927	1990	Kenya	A1
clone_1062	GQ430926	1990	Kenya	A1
clone_1061	GQ430925	1990	Kenya	A1
clone_279	GQ430143	1990	Kenya	A1/C
clone_277	GQ430141	1990	Kenya	C
clone_270	GQ430134	1990	Kenya	C
clone_260	GQ430124	1990	Kenya	A1/C
clone_258	GQ430122	1990	Kenya	C
clone_256	GQ430120	1990	Kenya	A1/C
E3099G	U92049	1991	Ethiopia	A1
92009_06	U88823	1992	Rwanda	C
ML170_	AF539406	1995	Kenya	A1
9521301	AF067156	1995	India	C
clone_1930	GQ431794	1995	Kenya	A1/C
clone_1928	GQ431792	1995	Kenya	C
clone_1920	GQ431784	1995	Kenya	C
9488	AF071474	1996	Ethiopia	A1/C
clone_2390	GQ432254	1996	Kenya	A1/C
clone_2381	GQ432245	1996	Kenya	A1/C
1997ML170_1997	AF539404	1997	Kenya	A1
9709	AF361879	1997	Tanzania	A1/C
9708	AF361878	1997	Tanzania	C
9706	AF361876	1997	Tanzania	C
9701	AF361871	1997	Tanzania	C
KNH1314	DQ367293	1999	Kenya	A1
KNH1300	DQ367292	1999	Kenya	A1
99K30889	AF484501	1999	Uganda	A1
99F25926	AF484491	1999	Uganda	A1
KNH1097	AF457064	1999	Kenya	A1/C
MSA4080	AF457087	2000	Kenya	A1
KISII5011	AF457061	2000	Kenya	C
1579A	DQ083238	2001	India	A1/C
A359	AY253319	2001	Tanzania	A1/C
A306	AY253318	2001	Tanzania	A1/C

A355	AY253315	2001	Tanzania	A1/C
CO9	AY734563	2002	Tanzania	C
CO3720	AY734562	2002	Tanzania	C
CO6968	AY734555	2002	Tanzania	C
CO6770	AY734554	2002	Tanzania	A1/C
CO3710	AY734553	2002	Tanzania	C
CO346	AY734552	2002	Tanzania	C
clone_1132	GQ430996	2002	Kenya	A1
H587	FJ853587	2003	Tanzania	C
H580	FJ853585	2003	Tanzania	A1/C
H514	FJ853570	2003	Tanzania	A1
H456	FJ853565	2003	Tanzania	A1/C
H347	FJ853556	2003	Tanzania	A1/C
H312	FJ853552	2003	Tanzania	C
H274	FJ853547	2003	Tanzania	A1
H270	FJ853546	2003	Tanzania	A1/C
H247	FJ853543	2003	Tanzania	A1
H236	FJ853540	2003	Tanzania	A1/C
H216	FJ853536	2003	Tanzania	A1
H201	FJ853533	2003	Tanzania	A1/C
H189	FJ853531	2003	Tanzania	A1/C
H123	FJ853527	2003	Tanzania	A1/C
H048	FJ853513	2003	Tanzania	A1/C
H034	FJ853509	2003	Tanzania	A1/C
H018	FJ853506	2003	Tanzania	C
H014	FJ853505	2003	Tanzania	A1/C
47750	EU220698	2004	Canada	A1
200606Cst_004	FJ623489	2006	Kenya	A1
2008BBCR06	GU201611	2008	South Africa	A1

Of the 18 A1/C recombinants with the same pattern of recombination as was seen in the Karonga recombinant sequences, 10 were collected in Tanzania (Arroyo et al., 2004; Geldmacher et al., 2009; Hoelscher et al., 2001), six were collected from Kenya (Dowling et al., 2002) one from Ethiopia (Carr et al., 1999) and one from India (Rodriguez et al., 2006).

jpHMM was employed to map more precisely the recombination break point within the *gag* p17/p24 fragment of both the Karonga and all 25 of the *gag* sequences downloaded from LANL including both the A1/C and C/A1 *gag* sequences. Identical breakpoints were characterised by jpHMM for the three Karonga sequences (54574, 54653 and 54702). From HXB2 position 925 to 1202 all three are subtype A1, and from position 1203 to 1503 all three are subtype C (Table 5.4). The breakpoints from the 25 LANL sequences ranged from HXB2 position 1124 (p17) to 1331 (p24) (Table 5.4). None of the LANL sequences contained a breakpoint identical to the three Karonga sequences (HXB2 1202) distinguishing them as unique recombinant forms (URFs). However the majority of the sequence breakpoints (80 %) did cluster around the same area (between HXB2 position 1169 to position 1202) with the highest number of recombination breakpoints seen at position 1184 (Figure 5.9). Position 1184 marks the junction between p17 and p24 with p24 starting at position 1185. The breakpoint for the three Karonga sequences was 18 bp after this junction.

The region around a breakpoint where the posterior probabilities of two subtypes are lower than a certain threshold, but higher than the posterior probabilities of all other subtypes is known as the breakpoint interval (Zhang et al., 2006). The average interval calculated by jpHMM region for the 25 LANL sequences made up of subtypes A1 and C and the three Karonga sequences was 25 bp (Table 5.4) with the highest being 70 bp and the lowest being 11 bp. The Karonga sequences were each identical with an interval of 34 bp.

5.3.3 PCR amplification of the full genome

Full genome amplification from extracted proviral DNA was attempted on the three A1/C recombinant samples 54702, 54653, and 54654. The presence of high molecular weight DNA was confirmed by gel electrophoresis (Figure 5.11). Spectrophotometer results, however, revealed very low amounts of DNA present for each sample (54574: 11.2 ng/µl, 54653: 15.8 ng/µl, and 54702: 20 ng/µl).

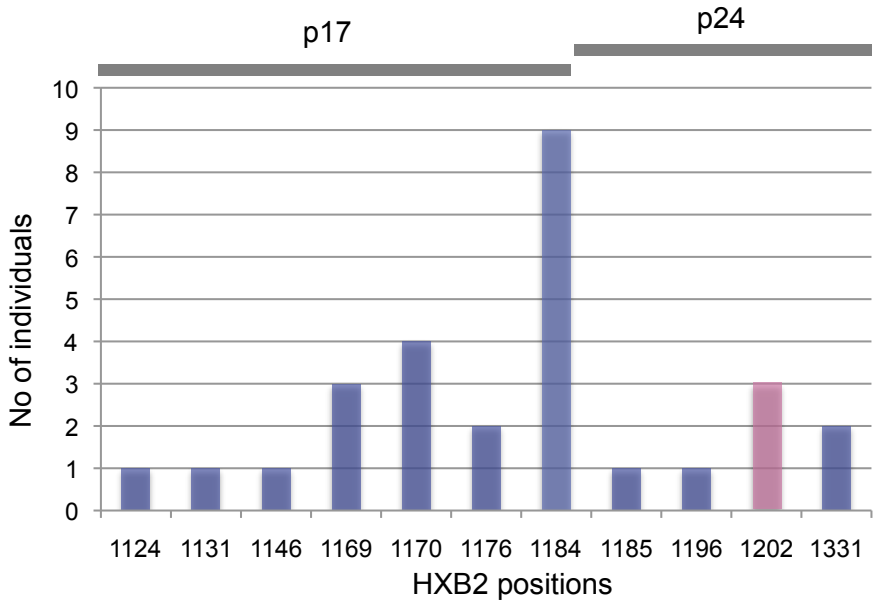


Figure 5.9 A graph representing the distribution of breakpoints found with in the p17/p24 *gag* fragment. The HXB2 numbering is on the X-axis, which the number of individuals with a breakpoint at a particular position is on the Y-axis. A pink bar represents the three Karonga A1/C recombinants.

Table 5.4 A list of the recombinant breakpoins predicted by jphHMM. Sequences are labelled with the subtypes, country of origin, year collected, sample name, and accession numbers. The sequences collected from the Karonga District are the first three listed. The length of each fragment is described using HXB2 numbering followed by what subtype is found within that fragment. The length of the breakpoint interval is described by the number of base pairs within the interaval.

Name of Sequence	First Fragment			Second Fragment			Breakpoint interval (bp)	Third Fragment			Breakpoint interval (bp)
A1C.MA.2008.54653.XXXXXX	925	1202	A1	1203	1503	C	34				
A1C.MA.2008.54702XXXXX	925	1202	A1	1203	1503	C	34				
A1C.MA.2008.54574XXXXXX	925	1202	A1	1203	1503	C	35				
A1C.KE.1999.KNH1097.AF457064	925	1169	A1	1170	1503	C	15				
A1C.KE.1996.clone_2390.GQ432254	925	1184	A1	1185	1503	C	24				
A1C.KE.1996.clone_2381.GQ432245	925	1184	A1	1185	1503	C	24				
A1C.KE.1996.clone_279.GQ430143	925	1185	A1	1186	1503	C	19				
A1C.KE.1996.clone_260.GQ430124	925	1176	A1	1177	1503	C	23				
A1C.KE.1996.clone_256.GQ430120	925	1176	A1	1177	1503	C	23				
A1C.TZ.2003.H048.FJ853513	925	1169	A1	1170	1503	C	16				
A1C.TZ.2003.H034.FJ853509	925	1184	A1	1185	1503	C	23				
A1C.TZ.2003.H236.FJ853540	925	1184	A1	1185	1503	C	25				
A1C.TZ.2003.H580.FJ853585	925	1169	A1	1170	1503	C	16				
A1C.TZ.2003.H347.FJ853556	925	1184	A1	1185	1503	C	24				
A1C.TZ.2003.H201.FJ853533	925	1184	A1	1185	1503	C	24				
A1C.TZ.2001.A359.AY253319	925	1184	A1	1185	1503	C	20				
A1C.TZ.2002.CO6770.AY734554	925	1124	A1	1125	1503	C	12				
A1C.TZ.2001.A306.AY253318	925	1196	A1	1197	1503	C	24				
A1C.TZ.1997.97TZ09	925	1184	A1	1185	1503	C	22				
A1C.SE.1996.SE9488.AF071474	925	1184	A1	1185	1503	C	24				
A1C.IN.2001.1579A.DQ083238	925	1131	A1	1132	1503	C	19				
A1C.TZ.2003.H014.FJ853505	925	1170	C	1171	1503	A1	26				
A1C.TZ.2003.H270.FJ853546	925	1170	C	1171	1503	A1	25				
A1C.TZ.2003.H189.FJ853531	925	1146	C	1147	1503	A1	36				
A1C.TZ.2003.H456.FJ853565	925	1170	C	1171	1503	A1	11				
A1C.TZ.2003.H123.FJ853527	925	1170	C	1171	1503	A1	11				
A1C.TZ.2001.A355.AY253315	925	960	C	961	1184	A1	70	1185	1503	C	23
A1C.KE.1995.clone_1930.GQ431794	925	1331	C	1332	1503	A1	36				

No amplification was seen for the three samples despite numerous optimization efforts including altering DNA concentration, lowering the annealing temperature and altering the extension times.

Attempts to amplify the full genome in smaller fragments also proved unsuccessful with no amplification observed from most primer combinations. Use of primers DRRT4 and MSF12 in a primary PCR followed by a secondary PCR with DRRT15 and F2nst (Table 5.2) resulted in amplification of two products (approximately 200 bp and 700 bp) for each sample (Figure 5.12). The expected amplification size was approximately 3300 bp. When primers ED5 and ofm-r1 for the primary and ED31 and MSR5 were combined the expected amplification size was approximately 2700 bp. Three amplification products were seen, around 1500 bp, 1000 bp and 800 bp, all smaller than the expected product. Changes in DNA concentrations and annealing temperatures did not result in amplification of the correct product for either primer combinations.

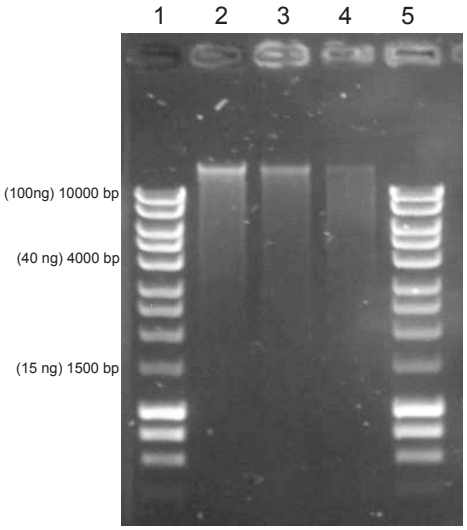


Figure 5.11 Agarose gel showing the quality of extracted DNA from the three samples. Each well was loaded with 5 µl of the three A1/C recombinant samples and 5 µl of the ladder was loaded and shows both size and quantity of DNA present.

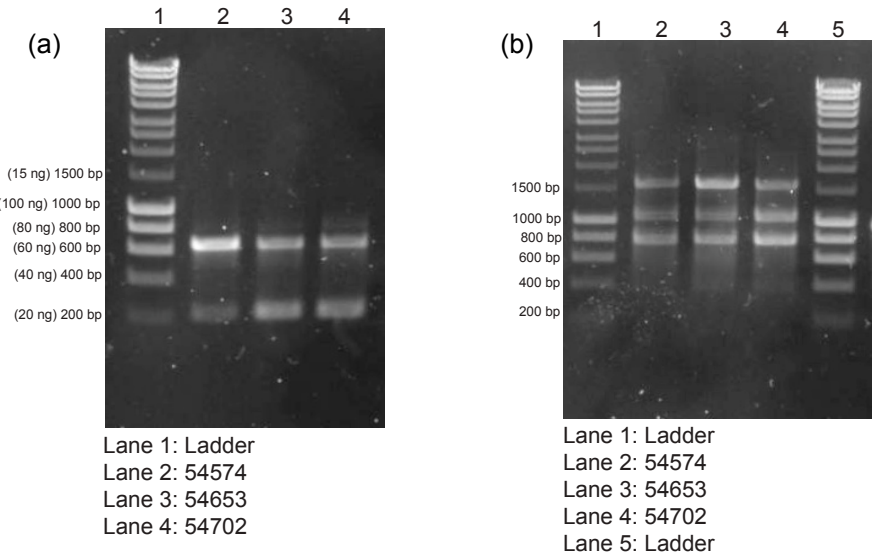


Figure 5.12 Agarose gels showing PCR amplification products from using primers (a) DRRT15 and F2nst and (b) ED31 and MSR5. Each well was loaded with 5 µl of the PCR product and 5 µl of the ladder was loaded and shows both size and quantity of DNA present.

5.4 Discussion

Fourteen viruses showing evidence of recombination have been identified within Karonga, 12 composed of subtypes A1 and C, one composed of subtypes A and D and one composed of subtypes D and C (McCormack et al., 2002) (and Chapter 4). Of the 13 A1/C recombinants two were classified as subtype A1 in *env* and subtype C in *gag* and four were subtype A1 in *gag* and subtype C in *env*. Three of the sequences (54702, 54653 and 54354) from the epidemiologically unlinked individuals were found to have a unique and identical recombination breakpoint within *gag* with the first half of the p17/p24 fragment classified as subtype A1 and the second half classified as subtype C. Identifying three sequences with an identical breakpoint, in conjunction with the fact all three are also subtype C in *env* implies a possible CRF within Karonga rather than the presence of URFs. This is also substantiated by the identical size of the breakpoint interval of 34 bp described by jpHMM, which was found to vary between 11 and 70 bp in the LANL A1/C sequences.

While subtype C is the most prevalent subtype within Karonga (McCormack et al., 2002), the prevalence of subtype A1 is less than 1 % making superinfection with these two different subtypes an unlikely event providing further evidence for a possible CRF. Full genome sequencing of the three A1/C samples is required to designate is as a CRF (Robertson et al., 2000) and to identify any other recombination points or any other subtypes present within the genome. However, three subtype A1/C sequences found in Mbeya, Tanzania were seen to have very similar breakpoints to each other within *gag,* but, full genome amplification revealed very different mosaic patterns for the rest of the genome (Arroyo et al., 2004). This may also be the case for the three A/C recombinants with identical breakpoints found in Karonga consequently full genome sequencing is essential.

Full genome amplification was attempted on the three A1/C recombinant samples with no success. HIV-1 full genome amplification is carried out regularly in a number of laboratories worldwide with over 2700 full genome sequences in the Los Alamos sequence database (http://www.hiv.lanl.gov/). Much of the reported full genome amplification involves extracting DNA from HIV-1 infected phytohemagglutinin-stimulated PBMCs followed by a single round of PCR or a further nested step when required (Gao et al., 1996; Laukkanen et al., 2000; Piyasirisilp et al., 2000; Salminen et al., 1997; Su et al., 2000). HIV co-culture

was not available in this case, none the less whole genome amplification from patient PMBC's has been shown to be possible with a nested PCR (Tzitzivacos et al., 2009). High molecular weight DNA was successfully extracted from at least two of the samples with little evidence of fragmentation. The primers (MSF12, F2nst, MSR5 and ofm-r1) and the PCR protocols used here have been used to amplify almost complete HIV-1 genomes from both subtype C and subtype A1 viruses (Carr et al., 1999; Papathanasopoulos et al., 2003; Piyasirisilp et al., 2000; van Harmelen et al., 2001) so theoretically should have successfully annealed to the extracted DNA from Karonga during the PCR process. It remains unclear why, despite numerous optimizations to the PCR protocol, amplification of full genome ultimately proved unsuccessful. Proviral DNA is a result of the integration of the HIV-1 genome into the infected individual's DNA. The DNA extraction process not only extracts the 9.7 kilo nucleotides of the HIV-1 genome but also the 6 billion nucleotides of the human genome, which could lead to false primer attachment and poor DNA synthesis in an already difficult PCR amplification procedure. Full genome amplification using proviral DNA has been described as both difficult and time consuming (Personal Communication Dr Jean K Carr, Dr Maria A Papathanasopoulos and Dr Eric Arts). Time constraints meant that only a few months could be spent optimising the PCR, which was most likely insufficient time to completely optimize the PCR to allow for successful amplification.

Some studies have successfully used viral RNA to amplify the complete HIV-1 genome (Fang et al., 2004; Fang et al., 1996; Gao, 2005; Rousseau et al., 2006), which removes the interference introduced by the presence of human DNA in the PCR. RNA was not available for the three Karonga samples but could potentially be collected from the three individuals in future projects. Amplification of a 750 bp fragment of *gag* proved successful in all three sequences presenting the possibility of amplifying the full genome in a number of overlapping smaller fragments. Unique primers could then be designed to amplify the regions next to this fragment and from those sequences new primers designed. The overlapping fragments could then be assembled into a contiguous sequence.

Within the *gag* fragment amplified and examined here the recombination break points of the three A1/C recombinants from Karonga and all 25 of the *gag* sequences downloaded from LANL database, including both the A1/C and C/A1 recombinants, seemed to cluster around

the p17/p24 interface. A number of studies have looked into the presence of recombination hot spots along the HIV-1 genome. A recent study (Fan, Negroni, and Robertson, 2007) analysed the position of all recombination breakpoints found in all full HIV-1 genome sequences listed in the Los Alamos sequence database. They found that the breakpoints were scattered at similar frequencies all along the genome with the exception of the borders of the *env* gene where breakpoints appeared to be more frequent and no recombination hot spots were identified within *gag*. Opposed to this a number of *in vitro* studies have identified a possible hot spot for recombination at the beginning of *gag* (Dykes et al., 2004; Zhuang et al., 2002) but no hotspots were identified surrounding p17/p24 junction. It has also been proposed that the RNA structure near gene borders promotes recombination locally to minimize deleterious interruptions to the gene due to recombination (Simon-Loriere et al., 2010). It is possible that the high frequency of breakpoints found here at the p17/p24 junction accommodates recombination without resulting in non-functional viruses.

In conclusion three A1/C recombinants have been identified in Karonga with a unique break point within *gag*. Attempts to amplify the full HIV-1 genome for all three samples failed. Understanding the full extent of diversity present within HIV-1 is critical for places such as Karonga where the impact of the epidemic is staggering. The expanding diversity will need to be monitored in conjunction with the implementation of ART. In the future, as vaccine studies continue, understanding the strains present within different regions will be crucial as it is still widely believed that vaccination will be vital to overcome HIV-1 in underdeveloped regions such as sub Saharan Africa. The potential for drug resistant strains to arise by recombination will also need to be monitored closely. This is particularly pertinent in places like Karonga where very limited drug choices are available.

Chapter 6: General Discussion

6 Chapter 6

HIV-1 has been studied in detail in Karonga District from the start of the epidemic in the 1980s, up until present day with the introduction of ART to the region (Glynn et al., 2001; White et al., 2007). This has provided a rare opportunity to map the history of HIV-1 in Karonga. Subtype C represented 91 % of infections in the 1990s and 93 % of infections by 2008/09. Subtypes A1 and D were only represented by a few sequences in 1990s and 2000s (eight and five sequences respectively). In the 1980s, four recombinant viruses were identified (McCormack et al., 2002), and four more were recorded in the 1990s. By the 2000s, seven recombinant viruses had been identified, three of which contained identical breakpoints within *gag*.

Recent discussions among scientists at the 18th International HIV Dynamics and Evolution in Galway questioned the authenticity of "pure" subtypes. With the advent of sub-subtypes as HIV-1 continues to diversify (e.g. A1 and A2, F1 and F2) and with the identification of inter-subtype recombinant viruses, the discrete identity of the nine subtypes described by Robertson et al. (2000) are less obvious. In spite of this accumulation of diversity in HIV-1, there are geographic elements to the epidemic as described here. In Karonga, two different patterns of circulation were observed; (i) a number of related, geographically restricted circulating strains and (ii) a constant flow of introductions of new viral strains from the surrounding regions. These geographically restricted clusters found within Karonga are not supported by bootstrap but are constantly present even after numerous phylogenetic reconstructions using two gene regions (Described here and by McCormack et al., 2002). A similar double pattern has also been seen in the Ethiopia epidemic (Abebe et al., 2000). A serious caveat to the work carried out here, and numerous other studies too many to list, is the use of proviral DNA which represents archived virus and not circulating virus. The rapid rate at which HIV-1 accumulates mutations results in the generation of sequences that represent evolutionary dead ends and make no contribution to the ongoing infection within an individual. Some of these sequences contain amino acids that may be detrimental to the protein structure, which can result in the viral fitness of the virus being significantly reduced, or, the virus being made non-functional (Williams et al.). A study looking at subtype B and C sequences of *gag* p17 from the LANL database and investigating the functionality p17

protein noted that many of the sequences in LANL represented non-functioning proteins (Williams et al., 2011).

In this work here, the range of viral diversity in subtype C within Karonga reached a high of 23.6 % in *env* and 16.48 % in *gag*. This level of diversity matches that seen between different subtypes in *gag* and the diversity in *env* is just below that which is seen between subtypes in *env*. The level of viral divergence from a subtype C common ancestor is continuing to rise within Karonga, although there was some evidence that the amount of increased divergence is beginning to level off. However, is the range of viral diversity recorded here in Karonga, a true representation of the functional viral diversity present? More thorough analysis is needed using bioinformatics and structural biology techniques to identify functioning viruses that have an actual impact on the epidemic. Replicative fitness assays can also be used to identify structurally viable viral variants. Identifying which variants represent replicating virus is also important for studies on the mutations in *pol* associated with drug resistance, as it is possible that the detection of these mutations using proviral DNA will not represent accurately the viral population within that individual resulting in erroneous judgements on patient care.

Ultimately, control of the HIV-1 pandemic is dependent on the development of an effective and preventative vaccine. The amount of diversity within HIV-1 is often cited as an almost insurmountable obstacle to vaccine designs (Woo, Robertson, and Lovell, 2010). Currently, only two vaccine concepts have completed clinical efficacy studies, both of which failed to provide detectable protective efficacy towards HIV-1 (Barouch, 2008). HIV-1 vaccine development, despite years of research, is still in its infancy (Barouch, 2008; Johnston and Fauci, 2008). Understanding the level of diversity within and between subtypes is still important for vaccine development, and therefore exploring the evolution of a HIV-1 within a population such as Karonga provides valuable information. More research is required to map the diversity observed through sequencing to the biology of the virus, in order to identify which variants are actually contributing to the spread of infections. This may reduce the amount of diversity and provide clearer separation of the nine subtypes described for HIV-1, information that can then be applied to future vaccine designs.

CXCR4 tropic viruses were seen within the Karonga subtype C population as early as the 1980s where they represented 3.9 % of the viral population. The prevalence of CXCR4 viruses continued to increase and in 2010 they represented 24.7 % of the viral population. It has been previously reported that subtype C viruses rarely use CXCR4 as a co receptor (Ping et al., 1999; van Harmelen et al., 2001) even during advanced disease progression when it has been previously shown that subtype B viruses show a greater propensity to switch to using CXCR4 (Jekle et al., 2003; Richman and Bozzette, 1994). CXCR4 tropic viral variants were found in over 50 % of the subtype C LTS in Karonga. The emergence of CXCR4 tropic viruses appeared not to be linked with disease progression within the LTS and was detected 10 years before some LTS were placed on ART, which is indicative of disease progression. Co receptor phenotypic assays are both expensive and time consuming and so are often not carried out in resource poor areas such as Karonga and other African districts. Lack of detection of CXCR4 viruses in the subtype C epidemic may have been hampered by this limitation. New bioinformatic tools such as C-PSSM and Geno2pheno could be used to investigate the prevalence of CXCR4 tropic viruses in the LANL database. The implications of increased CXCR4 tropism within subtype C are unknown but may potentially have an affect on the drug treatments available in the future in Karonga such as the CCR5 antagonist Maraviroc (Dorr et al., 2005) which is only effective on CCR5 tropic viruses.

Sequences generated from samples collected at multiple time points from LTS between 1988 and 2010 identified a number of mutations in *env* and *gag*. Some of these mutations were highly unlikely to occur according to the Blossum62 matrix. Viral attenuation has long been associated with disease non-progression and it is possible that some of the mutations seen in the LTS here have potentially negative implications for the virus. These implications could include a fitness cost to the virus or in extreme cases, the mutations could result in a non-functioning protein (and in turn non-function virus) that can then result in non-progression in the infected individual. Understanding the interaction between the virus and the immune system of non-progressors will provide information on the biology of the virus, which can then be incorporated into vaccines and drug therapies. In order for this to occur, more work on the affects of sequence mutations to the structure and function of the viral proteins and the resulting affect on the fitness of the virus is required.

A number of difficulties were encountered during the course of this project. Attempts to amplify large fragments of the HIV-1 genome from the LTS and from three individuals harbouring a recombinant viral genome proved unsuccessful. Efforts were most likely hampered by reliance on proviral genome amplification and due to the lack of high quality PCR template. Consensus sequencing is limited in the information it can provide as it gives little indication about the viral diversity present within an infected individual. This impacts not only on LTS studies, but also on other studies such drug resistant studies and CTL escape mutation studies. Sequencing of clones provides more detail on diversity however; it is both cost and time inefficient. Next generation sequencing can provide a much more accurate picture of the diversity within an individual. More recently population level next generation sequencing has been developed, this, in conjunction with an increase in sequence depth, a reduction in the error rate, and a reduction in cost makes the application of this new technology much more accessible and will provide a much more in depth understanding of HIV-1

It was also difficult to compare the results here with results from other studies due to the individual nature of research whereby different groups follow non-progressors identified with different criteria, for different time periods, using different genes and different methods. A common element to all long-term survivor studies is the low number of individuals within the studies. In the future more collaborative efforts globally would allow for more consistent collection of information and data. Within Africa, non-progressors have been identified in South Africa, Uganda and Nairobi (Archary et al., 2010; Fang et al., 2004; Laeyendecker et al., 2009; Tzitzivacos et al., 2009). Longitudinal, collaborative studies within Africa are needed to both identify new non-progressors and to provide a more in-depth detailed study of survivorship in Africa. Future work must also focus on host genomic elements that may be contributing to survival, such as HLA and CCR5 genotype to gain a more inclusive picture of non-progression.

In order for research on the molecular characterisation of LTS in Karonga to move forward it is vital that improved nucleic acid collection methods are employed. A long-term solution to this problem would be more investment of technology and training in Karonga with more sample processing and research carried out at the point of sample collection. This investment

would have a positive effect on the development of indigenous research and would benefit the community by up-skilling local laboratory technicians and researchers. Also as currently a large proportion of research about HIV 1 is undertaken outside of Africa, for long-term behavioral change to take root, local collaboration is essential. Increasing local skills and professional capacity will enable those who are most affected by the epidemic to influence change and grow in understanding at a community level.

References

Abebe, A., Pollakis, G., Fontanet, A. L., Fisseha, B., Tegbaru, B., Kliphuis, A., Tesfaye, G., Negassa, H., Cornelissen, M., Goudsmit, J., and Rinke de Wit, T. F. (2000). Identification of a genetic subcluster of HIV type 1 subtype C (C') widespread in Ethiopia. *AIDS Res Hum Retroviruses* **16**(17)**,** 1909-14.

Abecasis, A. B., Lemey, P., Vidal, N., de Oliveira, T., Peeters, M., Camacho, R., Shapiro, B., Rambaut, A., and Vandamme, A. M. (2007). Recombination confounds the early evolutionary history of human immunodeficiency virus type 1: subtype G is a circulating recombinant form. *J Virol* **81**(16)**,** 8543-51.

Abraha, A., Nankya, I. L., Gibson, R., Demers, K., Tebit, D. M., Johnston, E., Katzenstein, D., Siddiqui, A., Herrera, C., Fischetti, L., Shattock, R. J., and Arts, E. J. (2009). CCR5- and CXCR4-tropic subtype C human immunodeficiency virus type 1 isolates have a lower level of pathogenic fitness than other dominant group M subtypes: implications for the epidemic. *J Virol* **83**(11)**,** 5592-605.

Abreu, C. M., Brindeiro, P. A., Martins, A. N., Arruda, M. B., Bule, E., Stakteas, S., Tanuri, A., and de Moraes Brindeiro, R. (2008). Genotypic and phenotypic characterization of human immunodeficiency virus type 1 isolates circulating in pregnant women from Mozambique. *Arch Virol* **153**(11)**,** 2013-7.

Al-Mawsawi, L. Q., and Neamati, N. (2007). Blocking interactions between HIV-1 integrase and cellular cofactors: an emerging anti-retroviral strategy. *Trends Pharmacol Sci* **28**(10)**,** 526-35.

Alaeus, A., Lidman, K., Bjorkman, A., Giesecke, J., and Albert, J. (1999). Similar rate of disease progression among individuals infected with HIV-1 genetic subtypes A-D. *Aids* **13**(8)**,** 901-7.

Alexander, L., Weiskopf, E., Greenough, T. C., Gaddis, N. C., Auerbach, M. R., Malim, M. H., O'Brien, S. J., Walker, B. D., Sullivan, J. L., and Desrosiers, R. C. (2000). Unusual polymorphisms in human immunodeficiency virus type 1 associated with nonprogressive infection. *J Virol* **74**(9)**,** 4361-76.

Andrake, M. D., and Skalka, A. M. (1996). Retroviral integrase, putting the pieces together. *J Biol Chem* **271**(33)**,** 19633-6.

Archary, D., Gordon, M. L., Green, T. N., Coovadia, H. M., Goulder, P. J., and Ndung'u, T. (2010). HIV-1 subtype C envelope characteristics associated with divergent rates of chronic disease progression. *Retrovirology* **7,** 92.

Arien, K. K., Vanham, G., and Arts, E. J. (2007). Is HIV-1 evolving to a less virulent form in humans? *Nat Rev Microbiol* **5**(2)**,** 141-51.

Arora, V. K., Molina, R. P., Foster, J. L., Blakemore, J. L., Chernoff, J., Fredericksen, B. L., and Garcia, J. V. (2000). Lentivirus Nef specifically activates Pak2. *J Virol* **74**(23)**,** 11081-7.

Arroyo, M. A., Hoelscher, M., Sanders-Buell, E., Herbinger, K. H., Samky, E., Maboko, L., Hoffmann, O., Robb, M. R., Birx, D. L., and McCutchan, F. E. (2004). HIV type 1 subtypes among blood donors in the Mbeya region of southwest Tanzania. *AIDS Res Hum Retroviruses* **20**(8)**,** 895-901.

Bailey, J. R., Lassen, K. G., Yang, H. C., Quinn, T. C., Ray, S. C., Blankson, J. N., and Siliciano, R. F. (2006). Neutralizing antibodies do not mediate suppression of human immunodeficiency virus type 1 in elite suppressors or selection of plasma virus variants in patients on highly active antiretroviral therapy. *J Virol* **80**(10)**,** 4758-70.

Balakrishnan, M., Fay, P. J., and Bambara, R. A. (2001). The kissing hairpin sequence promotes recombination within the HIV-I 5' leader region. *J Biol Chem* **276**(39)**,** 36482-92.

Balasubramanyam, A., Mersmann, H., Jahoor, F., Phillips, T. M., Sekhar, R. V., Schubert, U., Brar, B., Iyer, D., Smith, E. O., Takahashi, H., Lu, H., Anderson, P., Kino, T., Henklein, P., and Kopp, J. B. (2007). Effects of transgenic expression of HIV-1 Vpr on lipid and energy metabolism in mice. *Am J Physiol Endocrinol Metab* **292**(1)**,** E40-8.

Ball, S. C., Abraha, A., Collins, K. R., Marozsan, A. J., Baird, H., Quinones-Mateu, M. E., Penn-Nicholson, A., Murray, M., Richard, N., Lobritz, M., Zimmerman, P. A., Kawamura, T., Blauvelt, A., and Arts, E. J. (2003). Comparing the ex vivo fitness of CCR5-tropic human immunodeficiency virus type 1 isolates of subtypes B and C. *J Virol* **77**(2)**,** 1021-38.

Bansode, V., Drebert, Z. J., Travers, S. A., Banda, E., Molesworth, A., Crampin, A., Ngwira, B., French, N., Glynn, J. R., and McCormack, G. P. (2010). Drug resistance mutations in drug-naive HIV type 1 subtype C-infected individuals from rural Malawi. *AIDS Res Hum Retroviruses* **27**(4)**,** 439-44.

Barouch, D. H. (2008). Challenges in the development of an HIV-1 vaccine. *Nature* **455**(7213)**,** 613-9.

Barre-Sinoussi, F., Chermann, J. C., Rey, F., Nugeyre, M. T., Chamaret, S., Gruest, J., Dauguet, C., Axler-Blin, C., Vezinet-Brun, F., Rouzioux, C., Rozenbaum, W., and Montagnier, L. (1983). Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science* **220**(4599)**,** 868-71.

Batra, M., Tien, P. C., Shafer, R. W., Contag, C. H., and Katzenstein, D. A. (2000). HIV type 1 envelope subtype C sequences from recent seroconverters in Zimbabwe. *AIDS Res Hum Retroviruses* **16**(10)**,** 973-9.

Bello, G., Casado, C., Sandonis, V., Alonso-Nieto, M., Vicario, J. L., Garcia, S., Hernando, V., Rodriguez, C., del Romero, J., and Lopez-Galindez, C. (2005). A subset of human immunodeficiency virus type 1 long-term non-progressors is characterized by the unique presence of ancestral sequences in the viral population. *J Gen Virol* **86**(Pt 2)**,** 355-64.

Bello, G., Casado, C., Sandonis, V., Alvaro-Cifuentes, T., Dos Santos, C. A., Garcia, S., Rodriguez, C., Del Romero, J., Pilotto, J. H., Grinsztejn, B., Veloso, V. G., Morgado, M. G., and Lopez-Galindez, C. (2007). Plasma viral load threshold for sustaining intrahost HIV type 1 evolution. *AIDS Res Hum Retroviruses* **23**(10)**,** 1242-50.

Bello, G. A., Chipeta, J., and Aberle-Grasse, J. (2006). Assessment of trends in biological and behavioural surveillance data: is there any evidence of declining HIV prevalence or incidence in Malawi? *Sex Transm Infect* **82 Suppl 1,** i9-13.

Berkowitz, R., Fisher, J., and Goff, S. P. (1996). RNA packaging. *Curr Top Microbiol Immunol* **214,** 177-218.

Bieniasz, P. D. (2009). The cell biology of HIV-1 virion genesis. *Cell Host Microbe* **5**(6)**,** 550-8.

Birch, M. R., Learmont, J. C., Dyer, W. B., Deacon, N. J., Zaunders, J. J., Saksena, N., Cunningham, A. L., Mills, J., and Sullivan, J. S. (2001). An examination of signs of disease progression in survivors of the Sydney Blood Bank Cohort (SBBC). *J Clin Virol* **22**(3)**,** 263-70.

Bjorndal, A., Sonnerborg, A., Tscherning, C., Albert, J., and Fenyo, E. M. (1999). Phenotypic characteristics of human immunodeficiency virus type 1 subtype C isolates of Ethiopian AIDS patients. *AIDS Res Hum Retroviruses* **15**(7)**,** 647-53.

Boileau, C., Clark, S., Bignami-Van Assche, S., Poulin, M., Reniers, G., Watkins, S. C., Kohler, H. P., and Heymann, S. J. (2009). Sexual and marital trajectories and HIV

infection among ever-married women in rural Malawi. *Sex Transm Infect* **85 Suppl 1,** i27-33.

Bour, S., Geleziunas, R., and Wainberg, M. A. (1995). The human immunodeficiency virus type 1 (HIV-1) CD4 receptor and its central role in promotion of HIV-1 infection. *Microbiol Rev* **59**(1)**,** 63-93.

Braibant, M., Agut, H., Rouzioux, C., Costagliola, D., Autran, B., and Barin, F. (2008). Characteristics of the env genes of HIV type 1 quasispecies in long-term nonprogressors with broadly neutralizing antibodies. *J Acquir Immune Defic Syndr* **47**(3)**,** 274-84.

Bredell, H., Martin, D. P., Van Harmelen, J., Varsani, A., Sheppard, H. W., Donovan, R., Gray, C. M., and Williamson, C. (2007). HIV type 1 subtype C gag and nef diversity in Southern Africa. *AIDS Res Hum Retroviruses* **23**(3)**,** 477-81.

Bres, V., Tagami, H., Peloponese, J. M., Loret, E., Jeang, K. T., Nakatani, Y., Emiliani, S., Benkirane, M., and Kiernan, R. E. (2002). Differential acetylation of Tat coordinates its interaction with the co-activators cyclin T1 and PCAF. *EMBO J* **21**(24)**,** 6811-9.

Briggs, D. R., Tuttle, D. L., Sleasman, J. W., and Goodenow, M. M. (2000). Envelope V3 amino acid sequence predicts HIV-1 phenotype (co-receptor usage and tropism for macrophages). *Aids* **14**(18)**,** 2937-9.

Briggs, S. D., Sharkey, M., Stevenson, M., and Smithgall, T. E. (1997). SH3-mediated Hck tyrosine kinase activation and fibroblast transformation by the Nef protein of HIV-1. *J Biol Chem* **272**(29)**,** 17899-902.

Brown, P. O. (1997). Integration.

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

Buve, A., Bishikwabo-Nsarhaza, K., and Mutangadura, G. (2002). The spread and effect of HIV-1 infection in sub-Saharan Africa. *Lancet* **359**(9322)**,** 2011-7.

Cann, A. J., Churcher, M. J., Boyd, M., O'Brien, W., Zhao, J. Q., Zack, J., and Chen, I. S. (1992). The region of the envelope gene of human immunodeficiency virus type 1 responsible for determination of cell tropism. *J Virol* **66**(1)**,** 305-9.

Cannon, P. M., Matthews, S., Clark, N., Byles, E. D., Iourin, O., Hockley, D. J., Kingsman, S. M., and Kingsman, A. J. (1997). Structure-function studies of the human immunodeficiency virus type 1 matrix protein, p17. *J Virol* **71**(5)**,** 3474-83.

Carr, J. K., Laukkanen, T., Salminen, M. O., Albert, J., Alaeus, A., Kim, B., Sanders-Buell, E., Birx, D. L., and McCutchan, F. E. (1999). Characterization of subtype A HIV-1 from Africa by full genome sequencing. *Aids* **13**(14)**,** 1819-26.

Carr, J. K., Salminen, M. O., Albert, J., Sanders-Buell, E., Gotte, D., Birx, D. L., and McCutchan, F. E. (1998). Full genome sequences of human immunodeficiency virus type 1 subtypes G and A/G intersubtype recombinants. *Virology* **247**(1)**,** 22-31.

Carr, J. K., Salminen, M. O., Koch, C., Gotte, D., Artenstein, A. W., Hegerich, P. A., St Louis, D., Burke, D. S., and McCutchan, F. E. (1996). Full-length sequence and mosaic structure of a human immunodeficiency virus type 1 isolate from Thailand. *J Virol* **70**(9)**,** 5935-43.

Carrillo, A., and Ratner, L. (1996). Cooperative effects of the human immunodeficiency virus type 1 envelope variable loops V1 and V3 in mediating infectivity for T cells. *J Virol* **70**(2)**,** 1310-6.

Carswell, J. W., Lloyd, G., and Howells, J. (1989). Prevalence of HIV-1 in east African lorry drivers. *Aids* **3**(11)**,** 759-61.

Casado, C., Pernas, M., Alvaro, T., Sandonis, V., Garcia, S., Rodriguez, C., del Romero, J., Grau, E., Ruiz, L., and Lopez-Galindez, C. (2007). Coinfection and superinfection in patients with long-term, nonprogressive HIV-1 disease. *J Infect Dis* **196**(6)**,** 895-9.

Casey, L., Wen, X., and de Noronha, C. M. (2010). The functions of the HIV1 protein Vpr and its action through the DCAF1.DDB1.Cullin4 ubiquitin ligase. *Cytokine* **51**(1)**,** 1-9.

Cassol, S., Salas, T., Arella, M., Neumann, P., Schechter, M. T., and O'Shaughnessy, M. (1991). Use of dried blood spot specimens in the detection of human immunodeficiency virus type 1 by the polymerase chain reaction. *J Clin Microbiol* **29**(4)**,** 667-71.

Cassol, S., Salas, T., Gill, M. J., Montpetit, M., Rudnik, J., Sy, C. T., and O'Shaughnessy, M. V. (1992). Stability of dried blood spot specimens for detection of human immunodeficiency virus DNA by polymerase chain reaction. *J Clin Microbiol* **30**(12)**,** 3039-42.

Cassol, S., Weniger, B. G., Babu, P. G., Salminen, M. O., Zheng, X., Htoon, M. T., Delaney, A., O'Shaughnessy, M., and Ou, C. Y. (1996). Detection of HIV type 1 env subtypes A, B, C, and E in Asia using dried blood spots: a new surveillance tool for molecular epidemiology. *AIDS Res Hum Retroviruses* **12**(15)**,** 1435-41.

Cecilia, D., Kulkarni, S. S., Tripathy, S. P., Gangakhedkar, R. R., Paranjape, R. S., and Gadkari, D. A. (2000). Absence of coreceptor switch with disease progression in human immunodeficiency virus infections in India. *Virology* **271**(2)**,** 253-8.

Chan, D. C., and Kim, P. S. (1998). HIV entry and its inhibition. *Cell* **93**(5)**,** 681-4.

Charpentier, C., Nora, T., Tenaillon, O., Clavel, F., and Hance, A. J. (2006). Extensive recombination among human immunodeficiency virus type 1 quasispecies makes an important contribution to viral diversity in individual patients. *J Virol* **80**(5)**,** 2472-82.

Cheroutre, H., and Madakamutil, L. (2004). Acquired and natural memory T cells join forces at the mucosal front line. *Nat Rev Immunol* **4**(4)**,** 290-300.

Chimbiri, A. M. (2007). The condom is an 'intruder' in marriage: evidence from rural Malawi. *Soc Sci Med* **64**(5)**,** 1102-15.

Chiu, Y. L., and Greene, W. C. (2008). The APOBEC3 cytidine deaminases: an innate defensive network opposing exogenous retroviruses and endogenous retroelements. *Annu Rev Immunol* **26,** 317-53.

Choge, I., Cilliers, T., Walker, P., Taylor, N., Phoswa, M., Meyers, T., Viljoen, J., Violari, A., Gray, G., Moore, P. L., Papathanosopoulos, M., and Morris, L. (2006). Genotypic and phenotypic characterization of viral isolates from HIV-1 subtype C-infected children with slow and rapid disease progression. *AIDS Res Hum Retroviruses* **22**(5)**,** 458-65.

Choisy, M., Woelk, C. H., Guegan, J. F., and Robertson, D. L. (2004). Comparative study of adaptive molecular evolution in different human immunodeficiency virus groups and subtypes. *J Virol* **78**(4)**,** 1962-70.

Christ, F., Thys, W., De Rijck, J., Gijsbers, R., Albanese, A., Arosio, D., Emiliani, S., Rain, J. C., Benarous, R., Cereseto, A., and Debyser, Z. (2008). Transportin-SR2 imports HIV into the nucleus. *Curr Biol* **18**(16)**,** 1192-202.

Chun, T. W., Carruth, L., Finzi, D., Shen, X., DiGiuseppe, J. A., Taylor, H., Hermankova, M., Chadwick, K., Margolick, J., Quinn, T. C., Kuo, Y. H., Brookmeyer, R., Zeiger, M. A., Barditch-Crovo, P., and Siliciano, R. F. (1997). Quantification of latent tissue reservoirs and total body viral load in HIV-1 infection. *Nature* **387**(6629)**,** 183-8.

Churchill, M., Sterjovski, J., Gray, L., Cowley, D., Chatfield, C., Learmont, J., Sullivan, J. S., Crowe, S. M., Mills, J., Brew, B. J., Wesselingh, S. L., McPhee, D. A., and Gorry, P. R. (2004). Longitudinal analysis of nef/long terminal repeat-deleted HIV-1 in blood and cerebrospinal fluid of a long-term survivor who developed HIV-associated dementia. *J Infect Dis* **190**(12)**,** 2181-6.

Cilliers, T., Nhlapo, J., Coetzer, M., Orlovic, D., Ketas, T., Olson, W. C., Moore, J. P., Trkola, A., and Morris, L. (2003). The CCR5 and CXCR4 coreceptors are both used by human immunodeficiency virus type 1 primary isolates from subtype C. *J Virol* **77**(7)**,** 4449-56.

Coffin, J. M. (1979). Structure, replication, and recombination of retrovirus genomes: some unifying hypotheses. *J Gen Virol* **42**(1)**,** 1-26.

Connell, B. J., Michler, K., Capovilla, A., Venter, W. D., Stevens, W. S., and Papathanasopoulos, M. A. (2008). Emergence of X4 usage among HIV-1 subtype C: evidence for an evolving epidemic in South Africa. *Aids* **22**(7)**,** 896-9.

Connor, R. I., Mohri, H., Cao, Y., and Ho, D. D. (1993). Increased viral burden and cytopathicity correlate temporally with CD4+ T-lymphocyte decline and clinical progression in human immunodeficiency virus type 1-infected individuals. *J Virol* **67**(4)**,** 1772-7.

Connor, R. I., Sheridan, K. E., Ceradini, D., Choe, S., and Landau, N. R. (1997). Change in coreceptor use correlates with disease progression in HIV-1--infected individuals. *J Exp Med* **185**(4)**,** 621-8.

Crampin, A. C., Floyd, S., Glynn, J. R., Sibande, F., Mulawa, D., Nyondo, A., Broadbent, P., Bliss, L., Ngwira, B., and Fine, P. E. (2002). Long-term follow-up of HIV-positive and HIV-negative individuals in rural Malawi. *Aids* **16**(11)**,** 1545-50.

Crampin, A. C., Glynn, J. R., Ngwira, B. M., Mwaungulu, F. D., Ponnighaus, J. M., Warndorff, D. K., and Fine, P. E. (2003). Trends and measurement of HIV prevalence in northern Malawi. *Aids* **17**(12)**,** 1817-25.

Dalai, S. C., de Oliveira, T., Harkins, G. W., Kassaye, S. G., Lint, J., Manasa, J., Johnston, E., and Katzenstein, D. (2009). Evolution and molecular epidemiology of subtype C HIV-1 in Zimbabwe. *Aids* **23**(18)**,** 2523-32.

daSilva, L. L., Sougrat, R., Burgos, P. V., Janvier, K., Mattera, R., and Bonifacino, J. S. (2009). Human immunodeficiency virus type 1 Nef protein targets CD4 to the multivesicular body pathway. *J Virol* **83**(13)**,** 6578-90.

De Jong, J. J., De Ronde, A., Keulen, W., Tersmette, M., and Goudsmit, J. (1992). Minimal requirements for the human immunodeficiency virus type 1 V3 domain to support the syncytium-inducing phenotype: analysis by single amino acid substitution. *J Virol* **66**(11)**,** 6777-80.

de Jong, J. J., Goudsmit, J., Keulen, W., Klaver, B., Krone, W., Tersmette, M., and de Ronde, A. (1992). Human immunodeficiency virus type 1 clones chimeric for the envelope V3 domain differ in syncytium formation and replication capacity. *J Virol* **66**(2)**,** 757-65.

De Leys, R., Vanderborght, B., Vanden Haesevelde, M., Heyndrickx, L., van Geel, A., Wauters, C., Bernaerts, R., Saman, E., Nijs, P., Willems, B., and et al. (1990). Isolation and partial characterization of an unusual human immunodeficiency retrovirus from two persons of west-central African origin. *J Virol* **64**(3)**,** 1207-16.

de Silva, T. I., Cotten, M., and Rowland-Jones, S. L. (2008). HIV-2: the forgotten AIDS virus. *Trends Microbiol* **16**(12)**,** 588-95.

Delwart, E. L., Herring, B., Rodrigo, A. G., and Mullins, J. I. (1995). Genetic subtyping of human immunodeficiency virus using a heteroduplex mobility assay. *PCR Methods Appl* **4**(5)**,** S202-16.

Delwart, E. L., Sheppard, H. W., Walker, B. D., Goudsmit, J., and Mullins, J. I. (1994). Human immunodeficiency virus type 1 evolution in vivo tracked by DNA heteroduplex mobility assays. *J Virol* **68**(10)**,** 6672-83.

Denolf, D., Musongela, J. P., Nzila, N., Tahiri, M., and Colebunders, R. (2001). The HIV epidemic in Kinshasa, Democratic Republic of Congo. *Int J STD AIDS* **12**(12)**,** 832.

DeStefano, J. J., Bambara, R. A., and Fay, P. J. (1994). The mechanism of human immunodeficiency virus reverse transcriptase-catalyzed strand transfer from internal regions of heteropolymeric RNA templates. *J Biol Chem* **269**(1)**,** 161-8.

DeStefano, J. J., Mallaber, L. M., Rodriguez-Rodriguez, L., Fay, P. J., and Bambara, R. A. (1992). Requirements for strand transfer between internal regions of heteropolymer templates by human immunodeficiency virus reverse transcriptase. *J Virol* **66**(11)**,** 6370-8.

Doms, R. W., and Trono, D. (2000). The plasma membrane as a combat zone in the HIV battlefield. *Genes Dev* **14**(21)**,** 2677-88.

Dorr, P., Westby, M., Dobbs, S., Griffin, P., Irvine, B., Macartney, M., Mori, J., Rickett, G., Smith-Burchnell, C., Napier, C., Webster, R., Armour, D., Price, D., Stammen, B., Wood, A., and Perros, M. (2005). Maraviroc (UK-427,857), a potent, orally bioavailable, and selective small-molecule inhibitor of chemokine receptor CCR5 with broad-spectrum anti-human immunodeficiency virus type 1 activity. *Antimicrob Agents Chemother* **49**(11)**,** 4721-32.

Dowling, W. E., Kim, B., Mason, C. J., Wasunna, K. M., Alam, U., Elson, L., Birx, D. L., Robb, M. L., McCutchan, F. E., and Carr, J. K. (2002). Forty-one near full-length HIV-1 sequences from Kenya reveal an epidemic of subtype A and A-containing recombinants. *Aids* **16**(13)**,** 1809-20.

Dragic, T., Litwin, V., Allaway, G. P., Martin, S. R., Huang, Y., Nagashima, K. A., Cayanan, C., Maddon, P. J., Koup, R. A., Moore, J. P., and Paxton, W. A. (1996). HIV-1 entry into CD4+ cells is mediated by the chemokine receptor CC-CKR-5. *Nature* **381**(6584)**,** 667-73.

Dykes, C., Balakrishnan, M., Planelles, V., Zhu, Y., Bambara, R. A., and Demeter, L. M. (2004). Identification of a preferred region for recombination and mutation in HIV-1 gag. *Virology* **326**(2)**,** 262-79.

Ehrlich, L. S., Fong, S., Scarlata, S., Zybarth, G., and Carter, C. (1996). Partitioning of HIV-1 Gag and Gag-related proteins to membranes. *Biochemistry* **35**(13)**,** 3933-43.

El Kharroubi, A., Piras, G., Zensen, R., and Martin, M. A. (1998). Transcriptional activation of the integrated chromatin-associated human immunodeficiency virus type 1 promoter. *Mol Cell Biol* **18**(5)**,** 2535-44.

Engelbrecht, S., de Villiers, T., Sampson, C. C., zur Megede, J., Barnett, S. W., and van Rensburg, E. J. (2001). Genetic analysis of the complete gag and env genes of HIV type 1 subtype C primary isolates from South Africa. *AIDS Res Hum Retroviruses* **17**(16)**,** 1533-47.

Essex, M. (1999). Human immunodeficiency viruses in the developing world. *Adv Virus Res* **53,** 71-88.

Etienne, L., Nerrienet, E., LeBreton, M., Bibila, G. T., Foupouapouognigni, Y., Rousset, D., Nana, A., Djoko, C. F., Tamoufe, U., Aghokeng, A. F., Mpoudi-Ngole, E., Delaporte, E., Peeters, M., Wolfe, N. D., and Ayouba, A. (2011). Characterization of a new

simian immunodeficiency virus strain in a naturally infected Pan troglodytes troglodytes chimpanzee with AIDS related symptoms. *Retrovirology* **8,** 4.

Fan, J., Negroni, M., and Robertson, D. L. (2007). The distribution of HIV-1 recombination breakpoints. *Infect Genet Evol* **7**(6)**,** 717-23.

Fang, G., Kuiken, C., Weiser, B., Rowland-Jones, S., Plummer, F., Chen, C. H., Kaul, R., Anzala, A. O., Bwayo, J., Kimani, J., Philpott, S. M., Kitchen, C., Sinsheimer, J. S., Gaschen, B., Lang, D., Shi, B., Kemal, K. S., Rostron, T., Brunner, C., Beddows, S., Sattenau, Q., Paxinos, E., Oyugi, J., and Burger, H. (2004a). Long-term survivors in Nairobi: complete HIV-1 RNA sequences and immunogenetic associations. *J Infect Dis* **190**(4)**,** 697-701.

Fang, G., Weiser, B., Kuiken, C., Philpott, S. M., Rowland-Jones, S., Plummer, F., Kimani, J., Shi, B., Kaul, R., Bwayo, J., Anzala, O., and Burger, H. (2004b). Recombination following superinfection by HIV-1. *Aids* **18**(2)**,** 153-9.

Fang, G., Weiser, B., Visosky, A. A., Townsend, L., and Burger, H. (1996). Molecular cloning of full-length HIV-1 genomes directly from plasma viral RNA. *J Acquir Immune Defic Syndr Hum Retrovirol* **12**(4)**,** 352-7.

Farzan, M., Choe, H., Desjardins, E., Sun, Y., Kuhn, J., Cao, J., Archambault, D., Kolchinsky, P., Koch, M., Wyatt, R., and Sodroski, J. (1998). Stabilization of human immunodeficiency virus type 1 envelope glycoprotein trimers by disulfide bonds introduced into the gp41 glycoprotein ectodomain. *J Virol* **72**(9)**,** 7620-5.

Fauci, A. S. (1996). Host factors and the pathogenesis of HIV-induced disease. *Nature* **384**(6609)**,** 529-34.

Feinberg, M. B., Baltimore, D., and Frankel, A. D. (1991). The role of Tat in the human immunodeficiency virus life cycle indicates a primary effect on transcriptional elongation. *Proc Natl Acad Sci U S A* **88**(9)**,** 4045-9.

Feng, Y., Broder, C. C., Kennedy, P. E., and Berger, E. A. (1996). HIV-1 entry cofactor: functional cDNA cloning of a seven-transmembrane, G protein-coupled receptor. *Science* **272**(5263)**,** 872-7.

Fernandez, G., Llano, A., Esgleas, M., Clotet, B., Este, J. A., and Martinez, M. A. (2006). Purifying selection of CCR5-tropic human immunodeficiency virus type 1 variants in AIDS subjects that have developed syncytium-inducing, CXCR4-tropic viruses. *J Gen Virol* **87**(Pt 5)**,** 1285-94.

Finzi, D., Hermankova, M., Pierson, T., Carruth, L. M., Buck, C., Chaisson, R. E., Quinn, T. C., Chadwick, K., Margolick, J., Brookmeyer, R., Gallant, J., Markowitz, M., Ho, D. D., Richman, D. D., and Siliciano, R. F. (1997). Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy. *Science* **278**(5341)**,** 1295-300.

Fiorentini, S., Marini, E., Caracciolo, S., and Caruso, A. (2006). Functions of the HIV-1 matrix protein p17. *New Microbiol* **29**(1)**,** 1-10.

Floyd, S., Crampin, A. C., Glynn, J. R., Mwenebabu, M., Mnkhondia, S., Ngwira, B., Zaba, B., and Fine, P. E. (2008). The long-term social and economic impact of HIV on the spouses of infected individuals in northern Malawi. *Trop Med Int Health* **13**(4)**,** 520-31.

Foley, B., Pan, H., Buchbinder, S., and Delwart, E. L. (2000). Apparent founder effect during the early years of the San Francisco HIV type 1 epidemic (1978-1979). *AIDS Res Hum Retroviruses* **16**(15)**,** 1463-9.

Foster, J. L., Denial, S. J., Temple, B. R., and Garcia, J. V. (2011). Mechanisms of HIV-1 Nef function and intracellular signaling. *J Neuroimmune Pharmacol* **6**(2)**,** 230-46.

Fouchier, R. A., Groenink, M., Kootstra, N. A., Tersmette, M., Huisman, H. G., Miedema, F., and Schuitemaker, H. (1992). Phenotype-associated sequence variation in the third variable domain of the human immunodeficiency virus type 1 gp120 molecule. *J Virol* **66**(5)**,** 3183-7.

Fouchier, R. A., Meyer, B. E., Simon, J. H., Fischer, U., Albright, A. V., Gonzalez-Scarano, F., and Malim, M. H. (1998). Interaction of the human immunodeficiency virus type 1 Vpr protein with the nuclear pore complex. *J Virol* **72**(7)**,** 6004-13.

Francis, D. P., Curran, J. W., and Essex, M. (1983). Epidemic acquired immune deficiency syndrome: epidemiologic evidence for a transmissible agent. *J Natl Cancer Inst* **71**(1)**,** 1-4.

Fraser, C. (2005). HIV recombination: what is the impact on antiretroviral therapy? *J R Soc Interface* **2**(5)**,** 489-503.

Freed, E. O. (1998). HIV-1 gag proteins: diverse functions in the virus life cycle. *Virology* **251**(1)**,** 1-15.

Freed, E. O. (2001). HIV-1 replication. *Somat Cell Mol Genet* **26**(1-6)**,** 13-33.

Friedrich, B. M., Dziuba, N., Li, G., Endsley, M. A., Murray, J. L., and Ferguson, M. R. (2011). Host factors mediating HIV-1 replication. *Virus Res* **161**(2)**,** 101-14.

Fukuda, M., Asano, S., Nakamura, T., Adachi, M., Yoshida, M., Yanagida, M., and Nishida, E. (1997). CRM1 is responsible for intracellular transport mediated by the nuclear export signal. *Nature* **390**(6657)**,** 308-11.

Gagneux, P., Wills, C., Gerloff, U., Tautz, D., Morin, P. A., Boesch, C., Fruth, B., Hohmann, G., Ryder, O. A., and Woodruff, D. S. (1999). Mitochondrial sequences show diverse evolutionary histories of African hominoids. *Proc Natl Acad Sci U S A* **96**(9)**,** 5077-82.

Galetto, R., and Negroni, M. (2005). Mechanistic features of recombination in HIV. *AIDS Rev* **7**(2)**,** 92-102.

Gallaher, W. R. (1987). Detection of a fusion peptide sequence in the transmembrane protein of human immunodeficiency virus. *Cell* **50**(3)**,** 327-8.

Gallo, R. C. (2002). Historical essay. The early years of HIV/AIDS. *Science* **298**(5599)**,** 1728-30.

Ganser-Pornillos, B. K., Yeager, M., and Sundquist, W. I. (2008). The structural biology of HIV assembly. *Curr Opin Struct Biol* **18**(2)**,** 203-17.

Gao, F. (2005). Amplification and cloning of near full-length HIV-2 genomes. *Methods Mol Biol* **304,** 399-407.

Gao, F., Bailes, E., Robertson, D. L., Chen, Y., Rodenburg, C. M., Michael, S. F., Cummins, L. B., Arthur, L. O., Peeters, M., Shaw, G. M., Sharp, P. M., and Hahn, B. H. (1999). Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. *Nature* **397**(6718)**,** 436-41.

Gao, F., Morrison, S. G., Robertson, D. L., Thornton, C. L., Craig, S., Karlsson, G., Sodroski, J., Morgado, M., Galvao-Castro, B., von Briesen, H., Beddows, S., Weber, J., Sharp, P. M., Shaw, G. M., and Hahn, B. H. (1996a). Molecular cloning and analysis of functional envelope genes from human immunodeficiency virus type 1 sequence subtypes A through G. The WHO and NIAID Networks for HIV Isolation and Characterization. *J Virol* **70**(3)**,** 1651-67.

Gao, F., Robertson, D. L., Carruthers, C. D., Morrison, S. G., Jian, B., Chen, Y., Barre-Sinoussi, F., Girard, M., Srinivasan, A., Abimiku, A. G., Shaw, G. M., Sharp, P. M., and Hahn, B. H. (1998). A comprehensive panel of near-full-length clones and

reference sequences for non-subtype B isolates of human immunodeficiency virus type 1. *J Virol* **72**(7)**,** 5680-98.

Gao, F., Robertson, D. L., Morrison, S. G., Hui, H., Craig, S., Decker, J., Fultz, P. N., Girard, M., Shaw, G. M., Hahn, B. H., and Sharp, P. M. (1996b). The heterosexual human immunodeficiency virus type 1 epidemic in Thailand is caused by an intersubtype (A/E) recombinant of African origin. *J Virol* **70**(10)**,** 7013-29.

Gao, F., Vidal, N., Li, Y., Trask, S. A., Chen, Y., Kostrikis, L. G., Ho, D. D., Kim, J., Oh, M. D., Choe, K., Salminen, M., Robertson, D. L., Shaw, G. M., Hahn, B. H., and Peeters, M. (2001). Evidence of two distinct subsubtypes within the HIV-1 subtype A radiation. *AIDS Res Hum Retroviruses* **17**(8)**,** 675-88.

Gao, F., Yue, L., White, A. T., Pappas, P. G., Barchue, J., Hanson, A. P., Greene, B. M., Sharp, P. M., Shaw, G. M., and Hahn, B. H. (1992). Human infection by genetically diverse SIVSM-related HIV-2 in west Africa. *Nature* **358**(6386)**,** 495-9.

Garcia, J. V., Alfano, J., and Miller, A. D. (1993). The negative effect of human immunodeficiency virus type 1 Nef on cell surface CD4 expression is not species specific and requires the cytoplasmic domain of CD4. *J Virol* **67**(3)**,** 1511-6.

Gaschen, B., Taylor, J., Yusim, K., Foley, B., Gao, F., Lang, D., Novitsky, V., Haynes, B., Hahn, B. H., Bhattacharya, T., and Korber, B. (2002). Diversity considerations in HIV-1 vaccine selection. *Science* **296**(5577)**,** 2354-60.

Gelderblom, H. R. (1991). Assembly and morphology of HIV: potential effect of structure on viral function. *Aids* **5**(6)**,** 617-37.

Geldmacher, C., Metzler, I. S., Tovanabutra, S., Asher, T. E., Gostick, E., Ambrozak, D. R., Petrovas, C., Schuetz, A., Ngwenyama, N., Kijak, G., Maboko, L., Hoelscher, M., McCutchan, F., Price, D. A., Douek, D. C., and Koup, R. A. (2009). Minor viral and host genetic polymorphisms can dramatically impact the biologic outcome of an epitope-specific CD8 T-cell response. *Blood* **114**(8)**,** 1553-62.

Glushakova, S., Grivel, J. C., Fitzgerald, W., Sylwester, A., Zimmerberg, J., and Margolis, L. B. (1998). Evidence for the HIV-1 phenotype switch as a causal factor in acquired immunodeficiency. *Nat Med* **4**(3)**,** 346-9.

Glynn, J. R., Ponnighaus, J., Crampin, A. C., Sibande, F., Sichali, L., Nkhosa, P., Broadbent, P., and Fine, P. E. (2001). The development of the HIV epidemic in Karonga District, Malawi. *Aids* **15**(15)**,** 2025-9.

Goedert, J. J., Neuland, C. Y., Wallen, W. C., Greene, M. H., Mann, D. L., Murray, C., Strong, D. M., Fraumeni, J. F., Jr., and Blattner, W. A. (1982). Amyl nitrite may alter T lymphocytes in homosexual men. *Lancet* **1**(8269)**,** 412-6.

Goff, A., Ehrlich, L. S., Cohen, S. N., and Carter, C. A. (2003). Tsg101 control of human immunodeficiency virus type 1 Gag trafficking and release. *J Virol* **77**(17)**,** 9173-82.

Goff, S. P. (2008). Knockdown screens to knockout HIV-1. *Cell* **135**(3)**,** 417-20.

Goh, W. C., Rogel, M. E., Kinsey, C. M., Michael, S. F., Fultz, P. N., Nowak, M. A., Hahn, B. H., and Emerman, M. (1998). HIV-1 Vpr increases viral expression by manipulation of the cell cycle: a mechanism for selection of Vpr in vivo. *Nat Med* **4**(1)**,** 65-71.

Goldsmith, M. A., Warmerdam, M. T., Atchison, R. E., Miller, M. D., and Greene, W. C. (1995). Dissociation of the CD4 downregulation and viral infectivity enhancement functions of human immunodeficiency virus type 1 Nef. *J Virol* **69**(7)**,** 4112-21.

Gottlieb, G. S., Nickle, D. C., Jensen, M. A., Wong, K. G., Grobler, J., Li, F., Liu, S. L., Rademeyer, C., Learn, G. H., Karim, S. S., Williamson, C., Corey, L., Margolick, J.

B., and Mullins, J. I. (2004). Dual HIV-1 infection associated with rapid disease progression. *Lancet* **363**(9409)**,** 619-22.

Gottlieb, M. S., Schroff, R., Schanker, H. M., Weisman, J. D., Fan, P. T., Wolf, R. A., and Saxon, A. (1981). Pneumocystis carinii pneumonia and mucosal candidiasis in previously healthy homosexual men: evidence of a new acquired cellular immunodeficiency. *N Engl J Med* **305**(24)**,** 1425-31.

Gottlinger, H. G. (2001). The HIV-1 assembly machine. *Aids* **15 Suppl 5,** S13-20.

Gottlinger, H. G., Dorfman, T., Sodroski, J. G., and Haseltine, W. A. (1991). Effect of mutations affecting the p6 gag protein on human immunodeficiency virus particle release. *Proc Natl Acad Sci U S A* **88**(8)**,** 3195-9.

Gottlinger, H. G., Sodroski, J. G., and Haseltine, W. A. (1989). Role of capsid precursor processing and myristoylation in morphogenesis and infectivity of human immunodeficiency virus type 1. *Proc Natl Acad Sci U S A* **86**(15)**,** 5781-5.

Greene, W. C., and Peterlin, B. M. (2002). Charting HIV's remarkable voyage through the cell: Basic science as a passport to future therapy. *Nat Med* **8**(7)**,** 673-80.

Grobler, J., Gray, C. M., Rademeyer, C., Seoighe, C., Ramjee, G., Karim, S. A., Morris, L., and Williamson, C. (2004). Incidence of HIV-1 dual infection and its association with increased viral load set point in a cohort of HIV-1 subtype C-infected female sex workers. *J Infect Dis* **190**(7)**,** 1355-9.

Gu, Z., Gao, Q., Faust, E. A., and Wainberg, M. A. (1995). Possible involvement of cell fusion and viral recombination in generation of human immunodeficiency virus variants that display dual resistance to AZT and 3TC. *J Gen Virol* **76 (Pt 10),** 2601-5.

Guadalupe, M., Reay, E., Sankaran, S., Prindiville, T., Flamm, J., McNeil, A., and Dandekar, S. (2003). Severe CD4+ T-cell depletion in gut lymphoid tissue during primary human immunodeficiency virus type 1 infection and substantial delay in restoration following highly active antiretroviral therapy. *J Virol* **77**(21)**,** 11708-17.

Guindon, S., and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52**(5)**,** 696-704.

Hamel, D. J., Sankale, J. L., Eisen, G., Meloni, S. T., Mullins, C., Gueye-Ndiaye, A., Mboup, S., and Kanki, P. J. (2007). Twenty years of prospective molecular epidemiology in Senegal: changes in HIV diversity. *AIDS Res Hum Retroviruses* **23**(10)**,** 1189-96.

Handema, R., Terunuma, H., Kasolo, F., Kasai, H., Sichone, M., Yamashita, A., Deng, X., Mulundu, G., Ichiyama, K., Munkanta, M., Yokota, T., Wakasugi, N., Tezuka, F., Yamamoto, N., and Ito, M. (2003). Prevalence of drug-resistance-associated mutations in antiretroviral drug-naive Zambians infected with subtype C HIV-1. *AIDS Res Hum Retroviruses* **19**(2)**,** 151-60.

Harper, M. E., Marselle, L. M., Gallo, R. C., and Wong-Staal, F. (1986). Detection of lymphocytes expressing human T-lymphotropic virus type III in lymph nodes and peripheral blood from infected individuals by in situ hybridization. *Proc Natl Acad Sci U S A* **83**(3)**,** 772-6.

Harries, A. D., Makombe, S. D., Schouten, E. J., Ben-Smith, A., and Jahn, A. (2008). Different delivery models for antiretroviral therapy in sub-Saharan Africa in the context of 'universal access'. *Trans R Soc Trop Med Hyg* **102**(4)**,** 310-1.

Harris, R. S., and Liddament, M. T. (2004). Retroviral restriction by APOBEC proteins. *Nat Rev Immunol* **4**(11)**,** 868-77.

Heeney, J. L., Dalgleish, A. G., and Weiss, R. A. (2006). Origins of HIV and the evolution of resistance to AIDS. *Science* **313**(5786)**,** 462-6.

Heinzinger, N. K., Bukinsky, M. I., Haggerty, S. A., Ragland, A. M., Kewalramani, V., Lee, M. A., Gendelman, H. E., Ratner, L., Stevenson, M., and Emerman, M. (1994). The Vpr protein of human immunodeficiency virus type 1 influences nuclear localization of viral nucleic acids in nondividing host cells. *Proc Natl Acad Sci U S A* **91**(15)**,** 7311-5.

Hemelaar, J., Gouws, E., Ghys, P. D., and Osmanov, S. (2006). Global and regional distribution of HIV-1 genetic subtypes and recombinants in 2004. *Aids* **20**(16)**,** W13-23.

Hermida-Matsumoto, L., and Resh, M. D. (2000). Localization of human immunodeficiency virus type 1 Gag and Env at the plasma membrane by confocal imaging. *J Virol* **74**(18)**,** 8670-9.

Hill, M., Tachedjian, G., and Mak, J. (2005). The packaging and maturation of the HIV-1 Pol proteins. *Curr HIV Res* **3**(1)**,** 73-85.

Hirsch, V. M., Olmsted, R. A., Murphey-Corb, M., Purcell, R. H., and Johnson, P. R. (1989). An African primate lentivirus (SIVsm) closely related to HIV-2. *Nature* **339**(6223)**,** 389-92.

Hoelscher, M., Kim, B., Maboko, L., Mhalu, F., von Sonnenburg, F., Birx, D. L., and McCutchan, F. E. (2001). High proportion of unrelated HIV-1 intersubtype recombinants in the Mbeya region of southwest Tanzania. *Aids* **15**(12)**,** 1461-70.

Hollingsworth, T. D., Anderson, R. M., and Fraser, C. (2008). HIV-1 transmission, by stage of infection. *J Infect Dis* **198**(5)**,** 687-93.

Hu, D. J., Buve, A., Baggs, J., van der Groen, G., and Dondero, T. J. (1999). What role does HIV-1 subtype play in transmission and pathogenesis? An epidemiological perspective. *Aids* **13**(8)**,** 873-81.

Huang, W., Toma, J., Fransen, S., Stawiski, E., Reeves, J. D., Whitcomb, J. M., Parkin, N., and Petropoulos, C. J. (2008). Coreceptor tropism can be influenced by amino acid substitutions in the gp41 transmembrane subunit of human immunodeficiency virus type 1 envelope protein. *J Virol* **82**(11)**,** 5584-93.

Huang, Y., Zhang, L., and Ho, D. D. (1998). Characterization of gag and pol sequences from long-term survivors of human immunodeficiency virus type 1 infection. *Virology* **240**(1)**,** 36-49.

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)**,** 356-9.

Hussain, A., Wesley, C., Khalid, M., Chaudhry, A., and Jameel, S. (2008). Human immunodeficiency virus type 1 Vpu protein interacts with CD74 and modulates major histocompatibility complex class II presentation. *J Virol* **82**(2)**,** 893-902.

Hwang, S. S., Boyle, T. J., Lyerly, H. K., and Cullen, B. R. (1991). Identification of the envelope V3 loop as the primary determinant of cell tropism in HIV-1. *Science* **253**(5015)**,** 71-4.

Hymes, K. B., Cheung, T., Greene, J. B., Prose, N. S., Marcus, A., Ballard, H., William, D. C., and Laubenstein, L. J. (1981). Kaposi's sarcoma in homosexual men-a report of eight cases. *Lancet* **2**(8247)**,** 598-600.

Ida, S., Gatanaga, H., Shioda, T., Nagai, Y., Kobayashi, N., Shimada, K., Kimura, S., Iwamoto, A., and Oka, S. (1997). HIV type 1 V3 variation dynamics in vivo: long-term persistence of non-syncytium-inducing genotypes and transient presence of syncytium-inducing genotypes during the course of progressive AIDS. *AIDS Res Hum Retroviruses* **13**(18)**,** 1597-609.

Iliffe, J. (2006). "The African AIDS epidemic: A History." James Currey Oxford.

Iversen, A. K., Shpaer, E. G., Rodrigo, A. G., Hirsch, M. S., Walker, B. D., Sheppard, H. W., Merigan, T. C., and Mullins, J. I. (1995). Persistence of attenuated rev genes in a human immunodeficiency virus type 1-infected asymptomatic individual. *J Virol* **69**(9)**,** 5743-53.

Jabara, C. B., Jones, C. D., Roach, J., Anderson, J. A., and Swanstrom, R. (2011). Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. *Proc Natl Acad Sci U S A* **108**(50)**,** 20166-71.

Jacks, T., Power, M. D., Masiarz, F. R., Luciw, P. A., Barr, P. J., and Varmus, H. E. (1988). Characterization of ribosomal frameshifting in HIV-1 gag-pol expression. *Nature* **331**(6153)**,** 280-3.

Jekle, A., Keppler, O. T., De Clercq, E., Schols, D., Weinstein, M., and Goldsmith, M. A. (2003). In vivo evolution of human immunodeficiency virus type 1 toward increased pathogenicity through CXCR4-mediated killing of uninfected CD4 T cells. *J Virol* **77**(10)**,** 5846-54.

Jensen, M. A., Coetzer, M., van 't Wout, A. B., Morris, L., and Mullins, J. I. (2006). A reliable phenotype predictor for human immunodeficiency virus type 1 subtype C based on envelope V3 sequences. *J Virol* **80**(10)**,** 4698-704.

Jensen, M. A., Li, F. S., van 't Wout, A. B., Nickle, D. C., Shriner, D., He, H. X., McLaughlin, S., Shankarappa, R., Margolick, J. B., and Mullins, J. I. (2003). Improved coreceptor usage prediction and genotypic monitoring of R5-to-X4 transition by motif analysis of human immunodeficiency virus type 1 env V3 loop sequences. *J Virol* **77**(24)**,** 13376-88.

Johnston, M. I., and Fauci, A. S. (2008). An HIV vaccine--challenges and prospects. *N Engl J Med* **359**(9)**,** 888-90.

Joos, B., Trkola, A., Fischer, M., Kuster, H., Rusert, P., Leemann, C., Boni, J., Oxenius, A., Price, D. A., Phillips, R. E., Wong, J. K., Hirschel, B., Weber, R., and Gunthard, H. F. (2005). Low human immunodeficiency virus envelope diversity correlates with low in vitro replication capacity and predicts spontaneous control of plasma viremia after treatment interruptions. *J Virol* **79**(14)**,** 9026-37.

Jordan, M. R., Kearney, M., Palmer, S., Shao, W., Maldarelli, F., Coakley, E. P., Chappey, C., Wanke, C., and Coffin, J. M. (2010). Comparison of standard PCR/cloning to single genome sequencing for analysis of HIV-1 populations. *J Virol Methods* **168**(1-2)**,** 114-20.

Jowett, J. B., Planelles, V., Poon, B., Shah, N. P., Chen, M. L., and Chen, I. S. (1995). The human immunodeficiency virus type 1 vpr gene arrests infected T cells in the G2 + M phase of the cell cycle. *J Virol* **69**(10)**,** 6304-13.

Jung, A., Maier, R., Vartanian, J. P., Bocharov, G., Jung, V., Fischer, U., Meese, E., Wain-Hobson, S., and Meyerhans, A. (2002). Recombination: Multiply infected spleen cells in HIV patients. *Nature* **418**(6894)**,** 144.

Kachapila, L. (1998). The HIV/AIDS epidemic in Malawi. *Int Nurs Rev* **45**(6)**,** 179-81.

Keele, B. F., Giorgi, E. E., Salazar-Gonzalez, J. F., Decker, J. M., Pham, K. T., Salazar, M. G., Sun, C., Grayson, T., Wang, S., Li, H., Wei, X., Jiang, C., Kirchherr, J. L., Gao, F., Anderson, J. A., Ping, L. H., Swanstrom, R., Tomaras, G. D., Blattner, W. A., Goepfert, P. A., Kilby, J. M., Saag, M. S., Delwart, E. L., Busch, M. P., Cohen, M. S., Montefiori, D. C., Haynes, B. F., Gaschen, B., Athreya, G. S., Lee, H. Y., Wood, N., Seoighe, C., Perelson, A. S., Bhattacharya, T., Korber, B. T., Hahn, B. H., and Shaw, G. M. (2008). Identification and characterization of transmitted and early founder

virus envelopes in primary HIV-1 infection. *Proc Natl Acad Sci U S A* **105**(21)**,** 7552-7.

Keele, B. F., Jones, J. H., Terio, K. A., Estes, J. D., Rudicell, R. S., Wilson, M. L., Li, Y., Learn, G. H., Beasley, T. M., Schumacher-Stankey, J., Wroblewski, E., Mosser, A., Raphael, J., Kamenya, S., Lonsdorf, E. V., Travis, D. A., Mlengeya, T., Kinsel, M. J., Else, J. G., Silvestri, G., Goodall, J., Sharp, P. M., Shaw, G. M., Pusey, A. E., and Hahn, B. H. (2009). Increased mortality and AIDS-like immunopathology in wild chimpanzees infected with SIVcpz. *Nature* **460**(7254)**,** 515-9.

Keele, B. F., Van Heuverswyn, F., Li, Y., Bailes, E., Takehisa, J., Santiago, M. L., Bibollet-Ruche, F., Chen, Y., Wain, L. V., Liegeois, F., Loul, S., Ngole, E. M., Bienvenue, Y., Delaporte, E., Brookfield, J. F., Sharp, P. M., Shaw, G. M., Peeters, M., and Hahn, B. H. (2006). Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. *Science* **313**(5786)**,** 523-6.

Kellam, P., and Larder, B. A. (1995). Retroviral recombination can lead to linkage of reverse transcriptase mutations that confer increased zidovudine resistance. *J Virol* **69**(2)**,** 669-74.

Kim, S., Ikeuchi, K., Byrn, R., Groopman, J., and Baltimore, D. (1989). Lack of a negative influence on viral growth by the nef gene of human immunodeficiency virus type 1. *Proc Natl Acad Sci U S A* **86**(23)**,** 9544-8.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* **16**(2)**,** 111-20.

Kirchhoff, F., Greenough, T. C., Brettler, D. B., Sullivan, J. L., and Desrosiers, R. C. (1995). Brief report: absence of intact nef sequences in a long-term survivor with nonprogressive HIV-1 infection. *N Engl J Med* **332**(4)**,** 228-32.

Kitchen, C. M., Philpott, S., Burger, H., Weiser, B., Anastos, K., and Suchard, M. A. (2004). Evolution of human immunodeficiency virus type 1 coreceptor usage during antiretroviral Therapy: a Bayesian approach. *J Virol* **78**(20)**,** 11296-302.

Kloosterboer, N., Groeneveld, P. H., Jansen, C. A., van der Vorst, T. J., Koning, F., Winkel, C. N., Duits, A. J., Miedema, F., van Baarle, D., van Rij, R. P., Brinkman, K., and Schuitemaker, H. (2005). Natural controlled HIV infection: preserved HIV-specific immunity despite undetectable replication competent virus. *Virology* **339**(1)**,** 70-80.

Kohl, N. E., Emini, E. A., Schleif, W. A., Davis, L. J., Heimbach, J. C., Dixon, R. A., Scolnick, E. M., and Sigal, I. S. (1988). Active human immunodeficiency virus protease is required for viral infectivity. *Proc Natl Acad Sci U S A* **85**(13)**,** 4686-90.

Koito, A., Stamatatos, L., and Cheng-Mayer, C. (1995). Small amino acid sequence changes within the V2 domain can affect the function of a T-cell line-tropic human immunodeficiency virus type 1 envelope gp120. *Virology* **206**(2)**,** 878-84.

Korber, B., Muldoon, M., Theiler, J., Gao, F., Gupta, R., Lapedes, A., Hahn, B. H., Wolinsky, S., and Bhattacharya, T. (2000). Timing the ancestor of the HIV-1 pandemic strains. *Science* **288**(5472)**,** 1789-96.

Korber, N., and Daniels, J. (1997). A Hydrogen Polyphosphide Salt of the First Homoleptic Strontium Ammine Complex: Synthesis and Structure Analysis of [Sr(NH(3))(8)]HP(11).NH(3). *Inorg Chem* **36**(21)**,** 4906-4908.

Kosakovsky Pond, S. L., and Smith, D. M. (2009). Are all subtypes created equal? The effectiveness of antiretroviral therapy against non-subtype B HIV-1. *Clin Infect Dis* **48**(9)**,** 1306-9.

Krausslich, H. G., Facke, M., Heuser, A. M., Konvalinka, J., and Zentgraf, H. (1995). The spacer peptide between human immunodeficiency virus capsid and nucleocapsid proteins is essential for ordered assembly and viral infectivity. *J Virol* **69**(6)**,** 3407-19.

Krug, M. S., and Berger, S. L. (1989). Ribonuclease H activities associated with viral reverse transcriptases are endonucleases. *Proc Natl Acad Sci U S A* **86**(10)**,** 3539-43.

Kuiken, C., Korber, B., and Shafer, R. W. (2003). HIV sequence databases. *AIDS Rev* **5**(1)**,** 52-61.

Kwong, P. D., Wyatt, R., Robinson, J., Sweet, R. W., Sodroski, J., and Hendrickson, W. A. (1998). Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. *Nature* **393**(6686)**,** 648-59.

Labrosse, B., Treboute, C., Brelot, A., and Alizon, M. (2001). Cooperation of the V1/V2 and V3 domains of human immunodeficiency virus type 1 gp120 for interaction with the CXCR4 receptor. *J Virol* **75**(12)**,** 5457-64.

Laeyendecker, O., Redd, A. D., Lutalo, T., Gray, R. H., Wawer, M., Ssempijja, V., Gamiel, J., Bwanika, J. B., Makumbi, F., Nalugoda, F., Opendi, P., Kigozi, G., Ndyanabo, A., Iga, B., Kiwanuka, N., Sewankambo, N., Reynolds, S. J., Serwadda, D., and Quinn, T. C. (2009). Frequency of long-term nonprogressors in HIV-1 seroconverters From Rakai Uganda. *J Acquir Immune Defic Syndr* **52**(3)**,** 316-9.

Lahuerta, M., Aparicio, E., Bardaji, A., Marco, S., Sacarlal, J., Mandomando, I., Alonso, P., Martinez, M. A., Menendez, C., and Naniche, D. (2008). Rapid spread and genetic diversification of HIV type 1 subtype C in a rural area of southern Mozambique. *AIDS Res Hum Retroviruses* **24**(2)**,** 327-35.

Lambotte, O., Boufassa, F., Madec, Y., Nguyen, A., Goujard, C., Meyer, L., Rouzioux, C., Venet, A., and Delfraissy, J. F. (2005). HIV controllers: a homogeneous group of HIV-1-infected patients with spontaneous control of viral replication. *Clin Infect Dis* **41**(7)**,** 1053-6.

Landau, N. R., Warton, M., and Littman, D. R. (1988). The envelope glycoprotein of the human immunodeficiency virus binds to the immunoglobulin-like domain of CD4. *Nature* **334**(6178)**,** 159-62.

Langford, S. E., Ananworanich, J., and Cooper, D. A. (2007). Predictors of disease progression in HIV infection: a review. *AIDS Res Ther* **4,** 11.

Lau, K. A., Wang, B., and Saksena, N. K. (2007). Emerging trends of HIV epidemiology in Asia. *AIDS Rev* **9**(4)**,** 218-29.

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

Learmont, J., Tindall, B., Evans, L., Cunningham, A., Cunningham, P., Wells, J., Penny, R., Kaldor, J., and Cooper, D. A. (1992). Long-term symptomless HIV-1 infection in recipients of blood products from a single donor. *Lancet* **340**(8824)**,** 863-7.

Learmont, J. C., Geczy, A. F., Mills, J., Ashton, L. J., Raynes-Greenow, C. H., Garsia, R. J., Dyer, W. B., McIntyre, L., Oelrichs, R. B., Rhodes, D. I., Deacon, N. J., and Sullivan, J. S. (1999). Immunologic and virologic status after 14 to 18 years of infection with an attenuated strain of HIV-1. A report from the Sydney Blood Bank Cohort. *N Engl J Med* **340**(22)**,** 1715-22.

Lee, K., Ambrose, Z., Martin, T. D., Oztop, I., Mulky, A., Julias, J. G., Vandegraaff, N., Baumann, J. G., Wang, R., Yuen, W., Takemura, T., Shelton, K., Taniuchi, I., Li, Y., Sodroski, J., Littman, D. R., Coffin, J. M., Hughes, S. H., Unutmaz, D., Engelman,

A., and KewalRamani, V. N. (2010). Flexible use of nuclear import pathways by HIV-1. *Cell Host Microbe* **7**(3)**,** 221-33.

Leitner, T., Escanilla, D., Franzen, C., Uhlen, M., and Albert, J. (1996). Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. *Proc Natl Acad Sci U S A* **93**(20)**,** 10864-9.

Lengauer, T., Sander, O., Sierra, S., Thielen, A., and Kaiser, R. (2007). Bioinformatics prediction of HIV coreceptor usage. *Nat Biotechnol* **25**(12)**,** 1407-10.

Leonard, C. K., Spellman, M. W., Riddle, L., Harris, R. J., Thomas, J. N., and Gregory, T. J. (1990). Assignment of intrachain disulfide bonds and characterization of potential glycosylation sites of the type 1 recombinant human immunodeficiency virus envelope glycoprotein (gp120) expressed in Chinese hamster ovary cells. *J Biol Chem* **265**(18)**,** 10373-82.

Libamba, E., Makombe, S., Harries, A. D., Chimzizi, R., Salaniponi, F. M., Schouten, E. J., and Mpazanje, R. (2005). Scaling up antiretroviral therapy in Africa: learning from tuberculosis control programmes--the case of Malawi. *Int J Tuberc Lung Dis* **9**(10)**,** 1062-71.

Lopez, M., Soriano, V., Lozano, S., Ballesteros, C., Cascajero, A., Rodes, B., De La Vega, E., Gonzalez-Lahoz, J., and Benito, J. M. (2008). No major differences in the functional profile of HIV Gag and Nef-specific CD8+ responses between long-term nonprogressors and typical progressors. *AIDS Res Hum Retroviruses* **24**(9)**,** 1185-95.

Louwagie, J., McCutchan, F. E., Peeters, M., Brennan, T. P., Sanders-Buell, E., Eddy, G. A., van der Groen, G., Fransen, K., Gershy-Damet, G. M., Deleys, R., and et al. (1993). Phylogenetic analysis of gag genes from 70 international HIV-1 isolates provides evidence for multiple genotypes. *Aids* **7**(6)**,** 769-80.

Lu, M., Blacklow, S. C., and Kim, P. S. (1995). A trimeric structural domain of the HIV-1 transmembrane glycoprotein. *Nat Struct Biol* **2**(12)**,** 1075-82.

Luo, C. C., Tian, C., Hu, D. J., Kai, M., Dondero, T., and Zheng, X. (1995). HIV-1 subtype C in China. *Lancet* **345**(8956)**,** 1051-2.

Malim, M. H., Hauber, J., Le, S. Y., Maizel, J. V., and Cullen, B. R. (1989). The HIV-1 rev trans-activator acts through a structured target sequence to activate nuclear export of unspliced viral mRNA. *Nature* **338**(6212)**,** 254-7.

Marozsan, A. J., and Arts, E. J. (2003). Development of a yeast-based recombination cloning/system for the analysis of gene products from diverse human immunodeficiency virus type 1 isolates. *J Virol Methods* **111**(2)**,** 111-20.

Masur, H., Michelis, M. A., Greene, J. B., Onorato, I., Stouwe, R. A., Holzman, R. S., Wormser, G., Brettman, L., Lange, M., Murray, H. W., 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)**,** 1431-8.

Mauclere, P., Loussert-Ajaka, I., Damond, F., Fagot, P., Souquieres, S., Monny Lobe, M., Mbopi Keou, F. X., Barre-Sinoussi, F., Saragosti, S., Brun-Vezinet, F., and Simon, F. (1997). Serological and virological characterization of HIV-1 group O infection in Cameroon. *Aids* **11**(4)**,** 445-53.

McCormack, G. P., Glynn, J. R., Clewley, J. P., Crampin, A. C., Travers, S. A., Redmond, N., Keane, T. M., Sibande, F., Mulawa, D., and Fine, P. E. (2006). Emergence of a three codon deletion in gag p17 in HIV type 1 subtype C long-term survivors, and general population spread. *AIDS Res Hum Retroviruses* **22**(2)**,** 195-201.

McCormack, G. P., Glynn, J. R., Crampin, A. C., Sibande, F., Mulawa, D., Bliss, L., Broadbent, P., Abarca, K., Ponnighaus, J. M., Fine, P. E., and Clewley, J. P. (2002).

Early evolution of the human immunodeficiency virus type 1 subtype C epidemic in rural Malawi. *J Virol* **76**(24)**,** 12890-9.

McCormick-Davis, C., Dalton, S. B., Singh, D. K., and Stephens, E. B. (2000). Comparison of Vpu sequences from diverse geographical isolates of HIV type 1 identifies the presence of highly variable domains, additional invariant amino acids, and a signature sequence motif common to subtype C isolates. *AIDS Res Hum Retroviruses* **16**(11)**,** 1089-95.

McCune, J. M. (2001). The dynamics of CD4+ T-cell depletion in HIV disease. *Nature* **410**(6831)**,** 974-9.

McDonald, D., Vodicka, M. A., Lucero, G., Svitkina, T. M., Borisy, G. G., Emerman, M., and Hope, T. J. (2002). Visualization of the intracellular behavior of HIV in living cells. *J Cell Biol* **159**(3)**,** 441-52.

McDonald, R. A., Mayers, D. L., Chung, R. C., Wagner, K. F., Ratto-Kim, S., Birx, D. L., and Michael, N. L. (1997). Evolution of human immunodeficiency virus type 1 env sequence variation in patients with diverse rates of disease progression and T-cell function. *J Virol* **71**(3)**,** 1871-9.

McGrath, N., Kranzer, K., Saul, J., Crampin, A. C., Malema, S., Kachiwanda, L., Zaba, B., Jahn, A., Fine, P. E., and Glynn, J. R. (2007). Estimating the need for antiretroviral treatment and an assessment of a simplified HIV/AIDS case definition in rural Malawi. *Aids* **21 Suppl 6,** S105-13.

McNulty, A., Jennings, C., Bennett, D., Fitzgibbon, J., Bremer, J. W., Ussery, M., Kalish, M. L., Heneine, W., and Garcia-Lerma, J. G. (2007). Evaluation of dried blood spots for human immunodeficiency virus type 1 drug resistance testing. *J Clin Microbiol* **45**(2)**,** 517-21.

Meehan, C. J., Hedge, J. A., Robertson, D. L., McCormack, G. P., and Travers, S. A. (2010). Emergence, dominance, and possible decline of CXCR4 chemokine receptor usage during the course of HIV infection. *J Med Virol* **82**(12)**,** 2004-12.

Mens, H., Kearney, M., Wiegand, A., Shao, W., Schonning, K., Gerstoft, J., Obel, N., Maldarelli, F., Mellors, J. W., Benfield, T., and Coffin, J. M. (2010). HIV-1 continues to replicate and evolve in patients with natural control of HIV infection. *J Virol* **84**(24)**,** 12971-81.

Metzner, K. J., Bonhoeffer, S., Fischer, M., Karanicolas, R., Allers, K., Joos, B., Weber, R., Hirschel, B., Kostrikis, L. G., and Gunthard, H. F. (2003). Emergence of minor populations of human immunodeficiency virus type 1 carrying the M184V and L90M mutations in subjects undergoing structured treatment interruptions. *J Infect Dis* **188**(10)**,** 1433-43.

Meyer, B. E., Meinkoth, J. L., and Malim, M. H. (1996). Nuclear transport of human immunodeficiency virus type 1, visna virus, and equine infectious anemia virus Rev proteins: identification of a family of transferable nuclear export signals. *J Virol* **70**(4)**,** 2350-9.

Michler, K., Connell, B. J., Venter, W. D., Stevens, W. S., Capovilla, A., and Papathanasopoulos, M. A. (2008). Genotypic characterization and comparison of full-length envelope glycoproteins from South African HIV type 1 subtype C primary isolates that utilize CCR5 and/or CXCR4. *AIDS Res Hum Retroviruses* **24**(5)**,** 743-51.

Migueles, S. A., Osborne, C. M., Royce, C., Compton, A. A., Joshi, R. P., Weeks, K. A., Rood, J. E., Berkley, A. M., Sacha, J. B., Cogliano-Shutta, N. A., Lloyd, M., Roby, G., Kwan, R., McLaughlin, M., Stallings, S., Rehm, C., O'Shea, M. A., Mican, J., Packard, B. Z., Komoriya, A., Palmer, S., Wiegand, A. P., Maldarelli, F., Coffin, J.

M., Mellors, J. W., Hallahan, C. W., Follman, D. A., and Connors, M. (2008). Lytic granule loading of CD8+ T cells is required for HIV-infected cell elimination associated with immune control. *Immunity* **29**(6)**,** 1009-21.

Milich, L., Margolin, B. H., and Swanstrom, R. (1997). Patterns of amino acid variability in NSI-like and SI-like V3 sequences and a linked change in the CD4-binding domain of the HIV-1 Env protein. *Virology* **239**(1)**,** 108-18.

Miller, M. D., Farnet, C. M., and Bushman, F. D. (1997). Human immunodeficiency virus type 1 preintegration complexes: studies of organization and composition. *J Virol* **71**(7)**,** 5382-90.

Miura, T., Brockman, M. A., Brumme, C. J., Brumme, Z. L., Carlson, J. M., Pereyra, F., Trocha, A., Addo, M. M., Block, B. L., Rothchild, A. C., Baker, B. M., Flynn, T., Schneidewind, A., Li, B., Wang, Y. E., Heckerman, D., Allen, T. M., and Walker, B. D. (2008). Genetic characterization of human immunodeficiency virus type 1 in elite controllers: lack of gross genetic defects or common amino acid changes. *J Virol* **82**(17)**,** 8422-30.

MMWR Weekly (1981). Kaposi's Sarcoma and Pneumocystis Pneumonia amoung Homosexual Men - New York City and California. *MMWR Weekly* **30**(4)**,** 305-308.

MMWR Weekly (1982). Epidemiologic notes and Reprots Pneumocystis carinii Pneumonia amoung persons with hemophilia A. *MMWR Weekly* **31**(27)**,** 365-367.

Mologni, D., Citterio, P., Menzaghi, B., Zanone Poma, B., Riva, C., Broggini, V., Sinicco, A., Milazzo, L., Adorni, F., Rusconi, S., Galli, M., and Riva, A. (2006). Vpr and HIV-1 disease progression: R77Q mutation is associated with long-term control of HIV-1 infection in different groups of patients. *Aids* **20**(4)**,** 567-74.

Montagnier, L., Clavel, F., Krust, B., Chamaret, S., Rey, F., Barre-Sinoussi, F., and Chermann, J. C. (1985). Identification and antigenicity of the major envelope glycoprotein of lymphadenopathy-associated virus. *Virology* **144**(1)**,** 283-9.

Montavon C, N. V., Vergne L, et al (1999). *6th Annual Internation Discussion Meeting on HIV Dynamics and Evolution, Atlanta*.

Morris, L., Cilliers, T., Bredell, H., Phoswa, M., and Martin, D. J. (2001). CCR5 is the major coreceptor used by HIV-1 subtype C isolates from patients with active tuberculosis. *AIDS Res Hum Retroviruses* **17**(8)**,** 697-701.

Motulsky, A. G., Vandepitte, J., and Fraser, G. R. (1966). Population genetic studies in the Congo. I. Glucose-6-phosphate dehydrogenase deficiency, hemoglobin S, and malaria. *Am J Hum Genet* **18**(6)**,** 514-37.

Moutouh, L., Corbeil, J., and Richman, D. D. (1996). Recombination leads to the rapid emergence of HIV-1 dually resistant mutants under selective drug pressure. *Proc Natl Acad Sci U S A* **93**(12)**,** 6106-11.

Nahmias, A. J., Weiss, J., Yao, X., Lee, F., Kodsi, R., Schanfield, M., Matthews, T., Bolognesi, D., Durack, D., Motulsky, A., and et al. (1986). Evidence for human infection with an HTLV III/LAV-like virus in Central Africa, 1959. *Lancet* **1**(8492)**,** 1279-80.

Navarro, F., and Landau, N. R. (2004). Recent insights into HIV-1 Vif. *Curr Opin Immunol* **16**(4)**,** 477-82.

Ndung'u, T., Sepako, E., McLane, M. F., Chand, F., Bedi, K., Gaseitsiwe, S., Doualla-Bell, F., Peter, T., Thior, I., Moyo, S. M., Gilbert, P. B., Novitsky, V. A., and Essex, M. (2006). HIV-1 subtype C in vitro growth and coreceptor utilization. *Virology* **347**(2)**,** 247-60.

Neel, C., Etienne, L., Li, Y., Takehisa, J., Rudicell, R. S., Bass, I. N., Moudindo, J., Mebenga, A., Esteban, A., Van Heuverswyn, F., Liegeois, F., Kranzusch, P. J., Walsh, P. D., Sanz, C. M., Morgan, D. B., Ndjango, J. B., Plantier, J. C., Locatelli, S., Gonder, M. K., Leendertz, F. H., Boesch, C., Todd, A., Delaporte, E., Mpoudi-Ngole, E., Hahn, B. H., and Peeters, M. (2010). Molecular epidemiology of simian immunodeficiency virus infection in wild-living gorillas. *J Virol* **84**(3)**,** 1464-76.

Negroni, M., and Galetto, R. (2009). Retroviruses. *In* "Viral Genome Replication" (C. E. Cameron, M. Gotte, and K. D. Raney, Eds.), Vol. Part 1, pp. 109-128. Springer Science and Business Media, New York.

Neil, S. J., Zang, T., and Bieniasz, P. D. (2008). Tetherin inhibits retrovirus release and is antagonized by HIV-1 Vpu. *Nature* **451**(7177)**,** 425-30.

Neilson, J. R., John, G. C., Carr, J. K., Lewis, P., Kreiss, J. K., Jackson, S., Nduati, R. W., Mbori-Ngacha, D., Panteleeff, D. D., Bodrug, S., Giachetti, C., Bott, M. A., Richardson, B. A., Bwayo, J., Ndinya-Achola, J., and Overbaugh, J. (1999). Subtypes of human immunodeficiency virus type 1 and disease stage among women in Nairobi, Kenya. *J Virol* **73**(5)**,** 4393-403.

Nekhai, S., and Jeang, K. T. (2006). Transcriptional and post-transcriptional regulation of HIV-1 gene expression: role of cellular factors for Tat and Rev. *Future Microbiol* **1**(4)**,** 417-26.

Nisole, S., and Saib, A. (2004). Early steps of retrovirus replicative cycle. *Retrovirology* **1,** 9.

Nora, T., Charpentier, C., Tenaillon, O., Hoede, C., Clavel, F., and Hance, A. J. (2007). Contribution of recombination to the evolution of human immunodeficiency viruses expressing resistance to antiretroviral treatment. *J Virol* **81**(14)**,** 7620-8.

Novitsky, V., Smith, U. R., Gilbert, P., McLane, M. F., Chigwedere, P., Williamson, C., Ndung'u, T., Klein, I., Chang, S. Y., Peter, T., Thior, I., Foley, B. T., Gaolekwe, S., Rybak, N., Gaseitsiwe, S., Vannberg, F., Marlink, R., Lee, T. H., and Essex, M. (2002). Human immunodeficiency virus type 1 subtype C molecular phylogeny: consensus sequence for an AIDS vaccine design? *J Virol* **76**(11)**,** 5435-51.

Novitsky, V. A., Montano, M. A., McLane, M. F., Renjifo, B., Vannberg, F., Foley, B. T., Ndung'u, T. P., Rahman, M., Makhema, M. J., Marlink, R., and Essex, M. (1999). Molecular cloning and phylogenetic analysis of human immunodeficiency virus type 1 subtype C: a set of 23 full-length clones from Botswana. *J Virol* **73**(5)**,** 4427-32.

Oelrichs, R., Tsykin, A., Rhodes, D., Solomon, A., Ellett, A., McPhee, D., and Deacon, N. (1998). Genomic sequence of HIV type 1 from four members of the Sydney Blood Bank Cohort of long-term nonprogressors. *AIDS Res Hum Retroviruses* **14**(9)**,** 811-4.

Onafuwa-Nuga, A., and Telesnitsky, A. (2009). The remarkable frequency of human immunodeficiency virus type 1 genetic recombination. *Microbiol Mol Biol Rev* **73**(3)**,** 451-80, Table of Contents.

Ou, C. Y., Takebe, Y., Luo, C. C., Kalish, M., Auwanit, W., Bandea, C., de la Torre, N., Moore, J. L., Schochetman, G., Yamazaki, S., and et al. (1992). Wide distribution of two subtypes of HIV-1 in Thailand. *AIDS Res Hum Retroviruses* **8**(8)**,** 1471-2.

Ou, C. Y., Takebe, Y., Weniger, B. G., Luo, C. C., Kalish, M. L., Auwanit, W., Yamazaki, S., Gayle, H. D., Young, N. L., and Schochetman, G. (1993). Independent introduction of two major HIV-1 genotypes into distinct high-risk populations in Thailand. *Lancet* **341**(8854)**,** 1171-4.

Oyama, F., Kikuchi, R., Crouch, R. J., and Uchida, T. (1989). Intrinsic properties of reverse transcriptase in reverse transcription. Associated RNase H is essentially regarded as an endonuclease. *J Biol Chem* **264**(31)**,** 18808-17.

Pancera, M., Majeed, S., Ban, Y. E., Chen, L., Huang, C. C., Kong, L., Kwon, Y. D., Stuckey, J., Zhou, T., Robinson, J. E., Schief, W. R., Sodroski, J., Wyatt, R., and Kwong, P. D. (2010). Structure of HIV-1 gp120 with gp41-interactive region reveals layered envelope architecture and basis of conformational mobility. *Proc Natl Acad Sci U S A* **107**(3)**,** 1166-71.

Pantaleo, G., and Fauci, A. S. (1996). Immunopathogenesis of HIV infection. *Annu Rev Microbiol* **50,** 825-54.

Papathanasopoulos, M. A., Patience, T., Meyers, T. M., McCutchan, F. E., and Morris, L. (2003). Full-length genome characterization of HIV type 1 subtype C isolates from two slow-progressing perinatally infected siblings in South Africa. *AIDS Res Hum Retroviruses* **19**(11)**,** 1033-7.

Parkin, N. T., Chamorro, M., and Varmus, H. E. (1992). Human immunodeficiency virus type 1 gag-pol frameshifting is dependent on downstream mRNA secondary structure: demonstration by expression in vivo. *J Virol* **66**(8)**,** 5147-51.

Pastore, C., Nedellec, R., Ramos, A., Pontow, S., Ratner, L., and Mosier, D. E. (2006). Human immunodeficiency virus type 1 coreceptor switching: V1/V2 gain-of-fitness mutations compensate for V3 loss-of-fitness mutations. *J Virol* **80**(2)**,** 750-8.

Paxton, W., Connor, R. I., and Landau, N. R. (1993). Incorporation of Vpr into human immunodeficiency virus type 1 virions: requirement for the p6 region of gag and mutational analysis. *J Virol* **67**(12)**,** 7229-37.

Peeters, M., Gueye, A., Mboup, S., Bibollet-Ruche, F., Ekaza, E., Mulanga, C., Ouedrago, R., Gandji, R., Mpele, P., Dibanga, G., Koumare, B., Saidou, M., Esu-Williams, E., Lombart, J. P., Badombena, W., Luo, N., Vanden Haesevelde, M., and Delaporte, E. (1997). Geographical distribution of HIV-1 group O viruses in Africa. *Aids* **11**(4)**,** 493-8.

Pessler, F., and Cron, R. Q. (2004). Reciprocal regulation of the nuclear factor of activated T cells and HIV-1. *Genes Immun* **5**(3)**,** 158-67.

Pettit, S. C., Moody, M. D., Wehbie, R. S., Kaplan, A. H., Nantermet, P. V., Klein, C. A., and Swanstrom, R. (1994). The p2 domain of human immunodeficiency virus type 1 Gag regulates sequential proteolytic processing and is required to produce fully infectious virions. *J Virol* **68**(12)**,** 8017-27.

Phillips, D. M., and Bourinbaiar, A. S. (1992). Mechanism of HIV spread from lymphocytes to epithelia. *Virology* **186**(1)**,** 261-73.

Pierson, T., McArthur, J., and Siliciano, R. F. (2000). Reservoirs for HIV-1: mechanisms for viral persistence in the presence of antiviral immune responses and antiretroviral therapy. *Annu Rev Immunol* **18,** 665-708.

Piguet, V., Schwartz, O., Le Gall, S., and Trono, D. (1999). The downregulation of CD4 and MHC-I by primate lentiviruses: a paradigm for the modulation of cell surface receptors. *Immunol Rev* **168,** 51-63.

Pillai, S., Good, B., Richman, D., and Corbeil, J. (2003). A new perspective on V3 phenotype prediction. *AIDS Res Hum Retroviruses* **19**(2)**,** 145-9.

Ping, L. H., Nelson, J. A., Hoffman, I. F., Schock, J., Lamers, S. L., Goodman, M., Vernazza, P., Kazembe, P., Maida, M., Zimba, D., Goodenow, M. M., Eron, J. J., Jr., Fiscus, S. A., Cohen, M. S., and Swanstrom, R. (1999). Characterization of V3 sequence heterogeneity in subtype C human immunodeficiency virus type 1 isolates from Malawi: underrepresentation of X4 variants. *J Virol* **73**(8)**,** 6271-81.

Piot, P., Plummer, F. A., Rey, M. A., Ngugi, E. N., Rouzioux, C., Ndinya-Achola, J. O., Veracauteren, G., D'Costa, L. J., Laga, M., Nsanze, H., and et al. (1987).

Retrospective seroepidemiology of AIDS virus infection in Nairobi populations. *J Infect Dis* **155**(6)**,** 1108-12.

Piyasirisilp, S., McCutchan, F. E., Carr, J. K., Sanders-Buell, E., Liu, W., Chen, J., Wagner, R., Wolf, H., Shao, Y., Lai, S., Beyrer, C., and Yu, X. F. (2000). A recent outbreak of human immunodeficiency virus type 1 infection in southern China was initiated by two highly homogeneous, geographically separated strains, circulating recombinant form AE and a novel BC recombinant. *J Virol* **74**(23)**,** 11286-95.

Pizzato, M., Helander, A., Popova, E., Calistri, A., Zamborlini, A., Palu, G., and Gottlinger, H. G. (2007). Dynamin 2 is required for the enhancement of HIV-1 infectivity by Nef. *Proc Natl Acad Sci U S A* **104**(16)**,** 6812-7.

Plantier, J. C., Leoz, M., Dickerson, J. E., De Oliveira, F., Cordonnier, F., Lemee, V., Damond, F., Robertson, D. L., and Simon, F. (2009). A new human immunodeficiency virus derived from gorillas. *Nat Med* **15**(8)**,** 871-2.

Pollard, S. R., Rosa, M. D., Rosa, J. J., and Wiley, D. C. (1992). Truncated variants of gp120 bind CD4 with high affinity and suggest a minimum CD4 binding region. *EMBO J* **11**(2)**,** 585-91.

Ponnighaus, J. M., Fine, P. E., Bliss, L., Gruer, P. J., Kapira-Mwamondwe, B., Msosa, E., Rees, R. J., Clayton, D., Pike, M. C., Sterne, J. A., and et al. (1993). The Karonga Prevention Trial: a leprosy and tuberculosis vaccine trial in northern Malawi. I. Methods of the vaccination phase. *Lepr Rev* **64**(4)**,** 338-56.

Ponninghaus, J. M., Fine, P. E., Bliss, L., Sliney, I. J., Bradley, D. J., and Rees, R. J. (1987). The Lepra Evaluation Project (LEP), an epidemiological study of leprosy in Northern Malawi. I. Methods. *Lepr Rev* **58**(4)**,** 359-75.

Popov, S., Rexach, M., Zybarth, G., Reiling, N., Lee, M. A., Ratner, L., Lane, C. M., Moore, M. S., Blobel, G., and Bukrinsky, M. (1998). Viral protein R regulates nuclear import of the HIV-1 pre-integration complex. *EMBO J* **17**(4)**,** 909-17.

Poropatich, K., and Sullivan, D. J., Jr. (2011). Human immunodeficiency virus type 1 long-term non-progressors: the viral, genetic and immunological basis for disease non-progression. *J Gen Virol* **92**(Pt 2)**,** 247-68.

Posada, D. (2008). jModelTest: phylogenetic model averaging. *Mol Biol Evol* **25**(7)**,** 1253-6.

Prince, A. M., Brotman, B., Lee, D. H., Andrus, L., Valinsky, J., and Marx, P. (2002). Lack of evidence for HIV type 1-related SIVcpz infection in captive and wild chimpanzees (Pan troglodytes verus) in West Africa. *AIDS Res Hum Retroviruses* **18**(9)**,** 657-60.

Purcell, D. F., and Martin, M. A. (1993). Alternative splicing of human immunodeficiency virus type 1 mRNA modulates viral protein expression, replication, and infectivity. *J Virol* **67**(11)**,** 6365-78.

Rambaut, A., Posada, D., Crandall, K. A., and Holmes, E. C. (2004). The causes and consequences of HIV evolution. *Nat Rev Genet* **5**(1)**,** 52-61.

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

Raymond, S., Delobel, P., Mavigner, M., Ferradini, L., Cazabat, M., Souyris, C., Sandres-Saune, K., Pasquier, C., Marchou, B., Massip, P., and Izopet, J. (2010). Prediction of HIV type 1 subtype C tropism by genotypic algorithms built from subtype B viruses. *J Acquir Immune Defic Syndr* **53**(2)**,** 167-75.

Reil, H., Bukovsky, A. A., Gelderblom, H. R., and Gottlinger, H. G. (1998). Efficient HIV-1 replication can occur in the absence of the viral matrix protein. *EMBO J* **17**(9)**,** 2699-708.

Reinhart, T. A., Rogan, M. J., Huddleston, D., Rausch, D. M., Eiden, L. E., and Haase, A. T. (1997). Simian immunodeficiency virus burden in tissues and cellular compartments during clinical latency and AIDS. *J Infect Dis* **176**(5)**,** 1198-208.

Renjifo, B., Fawzi, W., Mwakagile, D., Hunter, D., Msamanga, G., Spiegelman, D., Garland, M., Kagoma, C., Kim, A., Chaplin, B., Hertzmark, E., and Essex, M. (2001). Differences in perinatal transmission among human immunodeficiency virus type 1 genotypes. *J Hum Virol* **4**(1)**,** 16-25.

Rhodes, D. I., Ashton, L., Solomon, A., Carr, A., Cooper, D., Kaldor, J., and Deacon, N. (2000). Characterization of three nef-defective human immunodeficiency virus type 1 strains associated with long-term nonprogression. Australian Long-Term Nonprogressor Study Group. *J Virol* **74**(22)**,** 10581-8.

Richman, D. D., and Bozzette, S. A. (1994). The impact of the syncytium-inducing phenotype of human immunodeficiency virus on disease progression. *J Infect Dis* **169**(5)**,** 968-74.

Robertson, D. L., Anderson, J. P., Bradac, J. A., Carr, J. K., Foley, B., Funkhouser, R. K., Gao, F., Hahn, B. H., Kalish, M. L., Kuiken, C., Learn, G. H., Leitner, T., McCutchan, F., Osmanov, S., Peeters, M., Pieniazek, D., Salminen, M., Sharp, P. M., Wolinsky, S., and Korber, B. (2000). HIV-1 nomenclature proposal. *Science* **288**(5463)**,** 55-6.

Robertson, D. L., Sharp, P. M., McCutchan, F. E., and Hahn, B. H. (1995). Recombination in HIV-1. *Nature* **374**(6518)**,** 124-6.

Rodenburg, C. M., Li, Y., Trask, S. A., Chen, Y., Decker, J., Robertson, D. L., Kalish, M. L., Shaw, G. M., Allen, S., Hahn, B. H., and Gao, F. (2001). Near full-length clones and reference sequences for subtype C isolates of HIV type 1 from three different continents. *AIDS Res Hum Retroviruses* **17**(2)**,** 161-8.

Rodriguez, M. A., Chen, Y., Craigo, J. K., Chatterjee, R., Ratner, D., Tatsumi, M., Roy, P., Neogi, D., and Gupta, P. (2006). Construction and characterization of an infectious molecular clone of HIV-1 subtype A of Indian origin. *Virology* **345**(2)**,** 328-36.

Rogel, M. E., Wu, L. I., and Emerman, M. (1995). The human immunodeficiency virus type 1 vpr gene prevents cell proliferation during chronic infection. *J Virol* **69**(2)**,** 882-8.

Romani, B., and Engelbrecht, S. (2009). Human immunodeficiency virus type 1 Vpr: functions and molecular interactions. *J Gen Virol* **90**(Pt 8)**,** 1795-805.

Ross, E. K., Fuerst, T. R., Orenstein, J. M., O'Neill, T., Martin, M. A., and Venkatesan, S. (1991). Maturation of human immunodeficiency virus particles assembled from the gag precursor protein requires in situ processing by gag-pol protease. *AIDS Res Hum Retroviruses* **7**(5)**,** 475-83.

Rousseau, C. M., Birditt, B. A., McKay, A. R., Stoddard, J. N., Lee, T. C., McLaughlin, S., Moore, S. W., Shindo, N., Learn, G. H., Korber, B. T., Brander, C., Goulder, P. J., Kiepiela, P., Walker, B. D., and Mullins, J. I. (2006). Large-scale amplification, cloning and sequencing of near full-length HIV-1 subtype C genomes. *J Virol Methods* **136**(1-2)**,** 118-25.

Rowland-Jones, S. L., and Whittle, H. C. (2007). Out of Africa: what can we learn from HIV-2 about protective immunity to HIV-1? *Nat Immunol* **8**(4)**,** 329-31.

Saez-Cirion, A., Pancino, G., Sinet, M., Venet, A., and Lambotte, O. (2007). HIV controllers: how do they tame the virus? *Trends Immunol* **28**(12)**,** 532-40.

Salminen, M. O., Carr, J. K., Robertson, D. L., Hegerich, P., Gotte, D., Koch, C., Sanders-Buell, E., Gao, F., Sharp, P. M., Hahn, B. H., Burke, D. S., and McCutchan, F. E.

(1997). Evolution and probable transmission of intersubtype recombinant human immunodeficiency virus type 1 in a Zambian couple. *J Virol* **71**(4)**,** 2647-55.

Salvi, R., Garbuglia, A. R., Di Caro, A., Pulciani, S., Montella, F., and Benedetto, A. (1998). Grossly defective nef gene sequences in a human immunodeficiency virus type 1-seropositive long-term nonprogressor. *J Virol* **72**(5)**,** 3646-57.

Sankaran, S., Guadalupe, M., Reay, E., George, M. D., Flamm, J., Prindiville, T., and Dandekar, S. (2005). Gut mucosal T cell responses and gene expression correlate with protection against disease in long-term HIV-1-infected nonprogressors. *Proc Natl Acad Sci U S A* **102**(28)**,** 9860-5.

Santiago, M. L., Bibollet-Ruche, F., Gross-Camp, N., Majewski, A. C., Masozera, M., Munanura, I., Kaplin, B. A., Sharp, P. M., Shaw, G. M., and Hahn, B. H. (2003). Noninvasive detection of Simian immunodeficiency virus infection in a wild-living L'Hoest's monkey (Cercopithecus Ihoesti). *AIDS Res Hum Retroviruses* **19**(12)**,** 1163-6.

Sato, A., Igarashi, H., Adachi, A., and Hayami, M. (1990). Identification and localization of vpr gene product of human immunodeficiency virus type 1. *Virus Genes* **4**(4)**,** 303-12.

Scarlatti, G., Tresoldi, E., Bjorndal, A., Fredriksson, R., Colognesi, C., Deng, H. K., Malnati, M. S., Plebani, A., Siccardi, A. G., Littman, D. R., Fenyo, E. M., and Lusso, P. (1997). In vivo evolution of HIV-1 co-receptor usage and sensitivity to chemokine-mediated suppression. *Nat Med* **3**(11)**,** 1259-65.

Schuitemaker, H., Koot, M., Kootstra, N. A., Dercksen, M. W., de Goede, R. E., van Steenwijk, R. P., Lange, J. M., Schattenkerk, J. K., Miedema, F., and Tersmette, M. (1992). Biological phenotype of human immunodeficiency virus type 1 clones at different stages of infection: progression of disease is associated with a shift from monocytotropic to T-cell-tropic virus population. *J Virol* **66**(3)**,** 1354-60.

Schultz, A. K., Zhang, M., Bulla, I., Leitner, T., Korber, B., Morgenstern, B., and Stanke, M. (2009). jpHMM: improving the reliability of recombination prediction in HIV-1. *Nucleic Acids Res* **37**(Web Server issue)**,** W647-51.

Schultz, S. J., Zhang, M., and Champoux, J. J. (2004). Recognition of internal cleavage sites by retroviral RNases H. *J Mol Biol* **344**(3)**,** 635-52.

Schwartz, O., Dautry-Varsat, A., Goud, B., Marechal, V., Subtil, A., Heard, J. M., and Danos, O. (1995). Human immunodeficiency virus type 1 Nef induces accumulation of CD4 in early endosomes. *J Virol* **69**(1)**,** 528-33.

Schwartz, O., Marechal, V., Le Gall, S., Lemonnier, F., and Heard, J. M. (1996). Endocytosis of major histocompatibility complex class I molecules is induced by the HIV-1 Nef protein. *Nat Med* **2**(3)**,** 338-42.

Seager, I., Leeson, M. D., Crampin, A. C., Mulawa, D., French, N., Glynn, J. R., Travers, S. A., and McCormack, G. P. (2011). HIV type 1 mutational patterns in HIV type 1 subtype C-infected long-term survivors in Karonga District Malawi: further analysis and correction. *AIDS Res Hum Retroviruses* **28**(3)**,** 308-13.

Serwadda, D., Mugerwa, R. D., Sewankambo, N. K., Lwegaba, A., Carswell, J. W., Kirya, G. B., Bayley, A. C., Downing, R. G., Tedder, R. S., Clayden, S. A., and et al. (1985). Slim disease: a new disease in Uganda and its association with HTLV-III infection. *Lancet* **2**(8460)**,** 849-52.

Shankarappa, R., Chatterjee, R., Learn, G. H., Neogi, D., Ding, M., Roy, P., Ghosh, A., Kingsley, L., Harrison, L., Mullins, J. I., and Gupta, P. (2001). Human immunodeficiency virus type 1 env sequences from Calcutta in eastern India:

identification of features that distinguish subtype C sequences in India from other subtype C sequences. *J Virol* **75**(21)**,** 10479-87.

Shankarappa, R., Margolick, J. B., Gange, S. J., Rodrigo, A. G., Upchurch, D., Farzadegan, H., Gupta, P., Rinaldo, C. R., Learn, G. H., He, X., Huang, X. L., and Mullins, J. I. (1999). Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *J Virol* **73**(12)**,** 10489-502.

Sharp, P. M., Bailes, E., Chaudhuri, R. R., Rodenburg, C. M., Santiago, M. O., and Hahn, B. H. (2001). The origins of acquired immune deficiency syndrome viruses: where and when? *Philos Trans R Soc Lond B Biol Sci* **356**(1410)**,** 867-76.

Sharp, P. M., and Hahn, B. H. (2010). The evolution of HIV-1 and the origin of AIDS. *Philos Trans R Soc Lond B Biol Sci* **365**(1552)**,** 2487-94.

Sharp, P. M., Shaw, G. M., and Hahn, B. H. (2005). Simian immunodeficiency virus infection of chimpanzees. *J Virol* **79**(7)**,** 3891-902.

Sheehy, A. M., Gaddis, N. C., Choi, J. D., and Malim, M. H. (2002). Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. *Nature* **418**(6898)**,** 646-50.

Siepel, A. C., Halpern, A. L., Macken, C., and Korber, B. T. (1995). A computer program designed to screen rapidly for HIV type 1 intersubtype recombinant sequences. *AIDS Res Hum Retroviruses* **11**(11)**,** 1413-6.

Silvestri, G. (2005). Naturally SIV-infected sooty mangabeys: are we closer to understanding why they do not develop AIDS? *J Med Primatol* **34**(5-6)**,** 243-52.

Simon, F., Mauclere, P., Roques, P., Loussert-Ajaka, I., Muller-Trutwin, M. C., Saragosti, S., Georges-Courbot, M. C., Barre-Sinoussi, F., and Brun-Vezinet, F. (1998). Identification of a new human immunodeficiency virus type 1 distinct from group M and group O. *Nat Med* **4**(9)**,** 1032-7.

Simon-Loriere, E., Martin, D. P., Weeks, K. M., and Negroni, M. (2010). RNA structures facilitate recombination-mediated gene swapping in HIV-1. *J Virol* **84**(24)**,** 12675-82.

Sing, T., Low, A. J., Beerenwinkel, N., Sander, O., Cheung, P. K., Domingues, F. S., Buch, J., Daumer, M., Kaiser, R., Lengauer, T., and Harrigan, P. R. (2007). Predicting HIV coreceptor usage on the basis of genetic and clinical covariates. *Antivir Ther* **12**(7)**,** 1097-106.

Skasko, M., Tokarev, A., Chen, C. C., Fischer, W. B., Pillai, S. K., and Guatelli, J. (2011). BST-2 is rapidly down-regulated from the cell surface by the HIV-1 protein Vpu: evidence for a post-ER mechanism of Vpu-action. *Virology* **411**(1)**,** 65-77.

Smit-McBride, Z., Mattapallil, J. J., McChesney, M., Ferrick, D., and Dandekar, S. (1998). Gastrointestinal T lymphocytes retain high potential for cytokine responses but have severe CD4(+) T-cell depletion at all stages of simian immunodeficiency virus infection compared to peripheral lymphocytes. *J Virol* **72**(8)**,** 6646-56.

Smith, D. M., Richman, D. D., and Little, S. J. (2005). HIV superinfection. *J Infect Dis* **192**(3)**,** 438-44.

Soares, E. A., Martinez, A. M., Souza, T. M., Santos, A. F., Da Hora, V., Silveira, J., Bastos, F. I., Tanuri, A., and Soares, M. A. (2005). HIV-1 subtype C dissemination in southern Brazil. *Aids* **19 Suppl 4,** S81-6.

Soto, P. C., Stein, L. L., Hurtado-Ziola, N., Hedrick, S. M., and Varki, A. (2010). Relative over-reactivity of human versus chimpanzee lymphocytes: implications for the human diseases associated with immune activation. *J Immunol* **184**(8)**,** 4185-95.

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**(21)**,** 2688-90.

Steegen, K., Luchters, S., Demecheleer, E., Dauwe, K., Mandaliya, K., Jaoko, W., Plum, J., Temmerman, M., and Verhofstede, C. (2007). Feasibility of detecting human immunodeficiency virus type 1 drug resistance in DNA extracted from whole blood or dried blood spots. *J Clin Microbiol* **45**(10)**,** 3342-51.

Stoneburner, R. L., and Low-Beer, D. (2004). Population-level HIV declines and behavioral risk avoidance in Uganda. *Science* **304**(5671)**,** 714-8.

Strack, B., Calistri, A., Craig, S., Popova, E., and Gottlinger, H. G. (2003). AIP1/ALIX is a binding partner for HIV-1 p6 and EIAV p9 functioning in virus budding. *Cell* **114**(6)**,** 689-99.

Strebel, K. (2003). Virus-host interactions: role of HIV proteins Vif, Tat, and Rev. *Aids* **17 Suppl 4,** S25-34.

Strebel, K., Klimkait, T., Maldarelli, F., and Martin, M. A. (1989). Molecular and biochemical analyses of human immunodeficiency virus type 1 vpu protein. *J Virol* **63**(9)**,** 3784-91.

Strebel, K., Klimkait, T., and Martin, M. A. (1988). A novel gene of HIV-1, vpu, and its 16-kilodalton product. *Science* **241**(4870)**,** 1221-3.

Streeck, H., Li, B., Poon, A. F., Schneidewind, A., Gladden, A. D., Power, K. A., Daskalakis, D., Bazner, S., Zuniga, R., Brander, C., Rosenberg, E. S., Frost, S. D., Altfeld, M., and Allen, T. M. (2008). Immune-driven recombination and loss of control after HIV superinfection. *J Exp Med* **205**(8)**,** 1789-96.

Su, L., Graf, M., Zhang, Y., von Briesen, H., Xing, H., Kostler, J., Melzl, H., Wolf, H., Shao, Y., and Wagner, R. (2000). Characterization of a virtually full-length human immunodeficiency virus type 1 genome of a prevalent intersubtype (C/B') recombinant strain in China. *J Virol* **74**(23)**,** 11367-76.

Suphaphiphat, P., Essex, M., and Lee, T. H. (2007). Mutations in the V3 stem versus the V3 crown and C4 region have different effects on the binding and fusion steps of human immunodeficiency virus type 1 gp120 interaction with the CCR5 coreceptor. *Virology* **360**(1)**,** 182-90.

Switzer, W. M., Parekh, B., Shanmugam, V., Bhullar, V., Phillips, S., Ely, J. J., and Heneine, W. (2005). The epidemiology of simian immunodeficiency virus infection in a large number of wild- and captive-born chimpanzees: evidence for a recent introduction following chimpanzee divergence. *AIDS Res Hum Retroviruses* **21**(5)**,** 335-42.

Takehisa, J., Kraus, M. H., Ayouba, A., Bailes, E., Van Heuverswyn, F., Decker, J. M., Li, Y., Rudicell, R. S., Learn, G. H., Neel, C., Ngole, E. M., Shaw, G. M., Peeters, M., Sharp, P. M., and Hahn, B. H. (2009). Origin and biology of simian immunodeficiency virus in wild-living western gorillas. *J Virol* **83**(4)**,** 1635-48.

Tatt, I. D., Barlow, K. L., and Clewley, J. P. (2000). A gag gene heteroduplex mobility assay for subtyping HIV-1. *J Virol Methods* **87**(1-2)**,** 41-51.

Tatt, I. D., Barlow, K. L., Nicoll, A., and Clewley, J. P. (2001). The public health significance of HIV-1 subtypes. *Aids* **15 Suppl 5,** S59-71.

Taube, R., Fujinaga, K., Wimmer, J., Barboric, M., and Peterlin, B. M. (1999). Tat transactivation: a model for the regulation of eukaryotic transcriptional elongation. *Virology* **264**(2)**,** 245-53.

Tavassoli, A. (2010). Targeting the protein-protein interactions of the HIV lifecycle. *Chem Soc Rev* **40**(3)**,** 1337-46.

Telesnitsky, A., and Goff, S. P. (1993). Two defective forms of reverse transcriptase can complement to restore retroviral infectivity. *EMBO J* **12**(11)**,** 4433-8.

Temin, H. M. (1993). Retrovirus variation and reverse transcription: abnormal strand transfers result in retrovirus genetic variation. *Proc Natl Acad Sci U S A* **90**(15)**,** 6900-3.

The Global Fund (2010). The Global Fund 2010: Innovation and impact: Geneva. .

Travers, S. A., Clewley, J. P., Glynn, J. R., Fine, P. E., Crampin, A. C., Sibande, F., Mulawa, D., McInerney, J. O., and McCormack, G. P. (2004). Timing and reconstruction of the most recent common ancestor of the subtype C clade of human immunodeficiency virus type 1. *J Virol* **78**(19)**,** 10501-6.

Triques, K., Bourgeois, A., Vidal, N., Mpoudi-Ngole, E., Mulanga-Kabeya, C., Nzilambi, N., Torimiro, N., Saman, E., Delaporte, E., and Peeters, M. (2000). Near-full-length genome sequencing of divergent African HIV type 1 subtype F viruses leads to the identification of a new HIV type 1 subtype designated K. *AIDS Res Hum Retroviruses* **16**(2)**,** 139-51.

Truant, R., and Cullen, B. R. (1999). The arginine-rich domains present in human immunodeficiency virus type 1 Tat and Rev function as direct importin beta-dependent nuclear localization signals. *Mol Cell Biol* **19**(2)**,** 1210-7.

Tully, D. C., and Wood, C. (2010). Chronology and evolution of the HIV-1 subtype C epidemic in Ethiopia. *Aids* **24**(10)**,** 1577-82.

Tzitzivacos, D. B., Tiemessen, C. T., Stevens, W. S., and Papathanasopoulos, M. A. (2009). Viral genetic determinants of nonprogressive HIV type 1 subtype C infection in antiretroviral drug-naive children. *AIDS Res Hum Retroviruses* **25**(11)**,** 1141-8.

UNAIDS (2009a). Global Report - Sub Saharan Africa.

UNAIDS (2009b). World Wide AIDS epidemic update.

UNAIDS (2010). UNIAIDS Report on the Global AIDS epidemic.

UNAIDS, a. W. (2006). Annex 1: country profiles. 2006 report on Global AIDS epidemic Geneva, Switzerland.

UNAIDS, a. W. (2007). Sub-Saharan Africa AIDS epidemic update, regional summary.

UNAIDS/WHO (2007). "Towards Universal Access, Scaling up priority HIV/AIDS interventions in the health sector" Progress Report.

(2010). Country Progress Report: Malawi HIV and AIDS Monitoring and Evaluation Report: 2008-2009. UNGASS.

Vallari, A., Bodelle, P., Ngansop, C., Makamche, F., Ndembi, N., Mbanya, D., Kaptue, L., Gurtler, L. G., McArthur, C. P., Devare, S. G., and Brennan, C. A. (2010a). Four new HIV-1 group N isolates from Cameroon: Prevalence continues to be low. *AIDS Res Hum Retroviruses* **26**(1)**,** 109-15.

Vallari, A., Holzmayer, V., Harris, B., Yamaguchi, J., Ngansop, C., Makamche, F., Mbanya, D., Kaptue, L., Ndembi, N., Gurtler, L., Devare, S., and Brennan, C. A. (2010b). Confirmation of putative HIV-1 group P in Cameroon. *J Virol* **85**(3)**,** 1403-7.

Van Damme, N., Goff, D., Katsura, C., Jorgenson, R. L., Mitchell, R., Johnson, M. C., Stephens, E. B., and Guatelli, J. (2008). The interferon-induced protein BST-2 restricts HIV-1 release and is downregulated from the cell surface by the viral Vpu protein. *Cell Host Microbe* **3**(4)**,** 245-52.

van Harmelen, J., Williamson, C., Kim, B., Morris, L., Carr, J., Karim, S. S., and McCutchan, F. (2001). Characterization of full-length HIV type 1 subtype C sequences from South Africa. *AIDS Res Hum Retroviruses* **17**(16)**,** 1527-31.

Van Harmelen, J. H., Van der Ryst, E., Loubser, A. S., York, D., Madurai, S., Lyons, S., Wood, R., and Williamson, C. (1999). A predominantly HIV type 1 subtype C-

restricted epidemic in South African urban populations. *AIDS Res Hum Retroviruses* **15**(4)**,** 395-8.

Van Heuverswyn, F., Li, Y., Neel, C., Bailes, E., Keele, B. F., Liu, W., Loul, S., Butel, C., Liegeois, F., Bienvenue, Y., Ngolle, E. M., Sharp, P. M., Shaw, G. M., Delaporte, E., Hahn, B. H., and Peeters, M. (2006). Human immunodeficiency viruses: SIV infection in wild gorillas. *Nature* **444**(7116)**,** 164.

van't Wout, A. B., Kootstra, N. A., Mulder-Kampinga, G. A., Albrecht-van Lent, N., Scherpbier, H. J., Veenstra, J., Boer, K., Coutinho, R. A., Miedema, F., and Schuitemaker, H. (1994). Macrophage-tropic variants initiate human immunodeficiency virus type 1 infection after sexual, parenteral, and vertical transmission. *J Clin Invest* **94**(5)**,** 2060-7.

Vandegraaff, N., Devroe, E., Turlure, F., Silver, P. A., and Engelman, A. (2006). Biochemical and genetic analyses of integrase-interacting proteins lens epithelium-derived growth factor (LEDGF)/p75 and hepatoma-derived growth factor related protein 2 (HRP2) in preintegration complex function and HIV-1 replication. *Virology* **346**(2)**,** 415-26.

Velazquez-Campoy, A., Todd, M. J., Vega, S., and Freire, E. (2001). Catalytic efficiency and vitality of HIV-1 proteases from African viral subtypes. *Proc Natl Acad Sci U S A* **98**(11)**,** 6062-7.

Vidal, N., Peeters, M., Mulanga-Kabeya, C., Nzilambi, N., Robertson, D., Ilunga, W., Sema, H., Tshimanga, K., Bongo, B., and Delaporte, E. (2000). Unprecedented degree of human immunodeficiency virus type 1 (HIV-1) group M genetic diversity in the Democratic Republic of Congo suggests that the HIV-1 pandemic originated in Central Africa. *J Virol* **74**(22)**,** 10498-507.

Vogt, P. (1973). "The genome of avian RNA tumor viruses: a discussion of four models." In Proceddings of the Fourth Lepetit Colloquium, North Holland Publishing Co, Amsterdam.

Wain-Hobson, S., Sonigo, P., Danos, O., Cole, S., and Alizon, M. (1985). Nucleotide sequence of the AIDS virus, LAV. *Cell* **40**(1)**,** 9-17.

Walker, B. D. (2007). Elite control of HIV Infection: implications for vaccines and treatment. *Top HIV Med* **15**(4)**,** 134-6.

Walker, P. R., Pybus, O. G., Rambaut, A., and Holmes, E. C. (2005). Comparative population dynamics of HIV-1 subtypes B and C: subtype-specific differences in patterns of epidemic growth. *Infect Genet Evol* **5**(3)**,** 199-208.

Wang, B., Ge, Y. C., Jozwiak, R., Bolton, W., Palasanthiran, P., Ziegler, J., Chang, J., Xiang, S. H., Cunningham, A. L., and Saksena, N. K. (1997). Molecular analyses of human immunodeficiency virus type 1 V3 region quasispecies derived from plasma and peripheral blood mononuclear cells of the first long-term-nonprogressing mother and child pair. *J Infect Dis* **175**(6)**,** 1510-5.

Wang, B., Mikhail, M., Dyer, W. B., Zaunders, J. J., Kelleher, A. D., and Saksena, N. K. (2003). First demonstration of a lack of viral sequence evolution in a nonprogressor, defining replication-incompetent HIV-1 infection. *Virology* **312**(1)**,** 135-50.

Wang, B., Spira, T. J., Owen, S., Lal, R. B., and Saksena, N. K. (2000). HIV-1 strains from a cohort of American subjects reveal the presence of a V2 region extension unique to slow progressors and non-progressors. *Aids* **14**(3)**,** 213-23.

Weidner, J., Cassens, U., Gohde, W., Sibrowski, W., Odaibo, G., Olaleye, D., Reichelt, D., and Greve, B. (2011). An improved PCR method for detection of HIV-1 proviral

DNA of a wide range of subtypes and recombinant forms circulating globally. *J Virol Methods* **172**(1-2)**,** 22-6.

Weisman, Z., Kalinkovich, A., Borkow, G., Stein, M., Greenberg, Z., and Bentwich, Z. (1999). Infection by different HIV-1 subtypes (B and C) results in a similar immune activation profile despite distinct immune backgrounds. *J Acquir Immune Defic Syndr* **21**(2)**,** 157-63.

Weiss, E. R., and Gottlinger, H. (2011). The role of cellular factors in promoting HIV budding. *J Mol Biol* **410**(4)**,** 525-33.

White, R. G., Vynnycky, E., Glynn, J. R., Crampin, A. C., Jahn, A., Mwaungulu, F., Mwanyongo, O., Jabu, H., Phiri, H., McGrath, N., Zaba, B., and Fine, P. E. (2007). HIV epidemic trend and antiretroviral treatment need in Karonga District, Malawi. *Epidemiol Infect* **135**(6)**,** 922-32.

Willey, R. L., Maldarelli, F., Martin, M. A., and Strebel, K. (1992). Human immunodeficiency virus type 1 Vpu protein induces rapid degradation of CD4. *J Virol* **66**(12)**,** 7193-200.

Williams, S. G., Madan, R., Norris, M. G., Archer, J., Mizuguchi, K., Robertson, D. L., and Lovell, S. C. (2011). Using knowledge of protein structural constraints to predict the evolution of HIV-1. *J Mol Biol* **410**(5)**,** 1023-34.

Wintersberger, U. (1990). Ribonucleases H of retroviral and cellular origin. *Pharmacol Ther* **48**(2)**,** 259-80.

Wisniewski, M., Balakrishnan, M., Palaniappan, C., Fay, P. J., and Bambara, R. A. (2000). The sequential mechanism of HIV reverse transcriptase RNase H. *J Biol Chem* **275**(48)**,** 37664-71.

Wolinsky, S. M., Korber, B. T., Neumann, A. U., Daniels, M., Kunstman, K. J., Whetsell, A. J., Furtado, M. R., Cao, Y., Ho, D. D., and Safrit, J. T. (1996). Adaptive evolution of human immunodeficiency virus-type 1 during the natural course of infection. *Science* **272**(5261)**,** 537-42.

Woo, J., Robertson, D. L., and Lovell, S. C. (2010). Constraints on HIV-1 diversity from protein structure. *J Virol* **84**(24)**,** 12995-3003.

Worobey, M., Gemmel, M., Teuwen, D. E., Haselkorn, T., Kunstman, K., Bunce, M., Muyembe, J. J., Kabongo, J. M., Kalengayi, R. M., Van Marck, E., Gilbert, M. T., and Wolinsky, S. M. (2008). Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature* **455**(7213)**,** 661-4.

Xiao, L., Owen, S. M., Goldman, I., Lal, A. A., deJong, J. J., Goudsmit, J., and Lal, R. B. (1998a). CCR5 coreceptor usage of non-syncytium-inducing primary HIV-1 is independent of phylogenetically distinct global HIV-1 isolates: delineation of consensus motif in the V3 domain that predicts CCR-5 usage. *Virology* **240**(1)**,** 83-92.

Xiao, L., Rudolph, D. L., Owen, S. M., Spira, T. J., and Lal, R. B. (1998b). Adaptation to promiscuous usage of CC and CXC-chemokine coreceptors in vivo correlates with HIV-1 disease progression. *Aids* **12**(13)**,** F137-43.

Xiao, Y., Chen, G., Richard, J., Rougeau, N., Li, H., Seidah, N. G., and Cohen, E. A. (2008). Cell-surface processing of extracellular human immunodeficiency virus type 1 Vpr by proprotein convertases. *Virology* **372**(2)**,** 384-97.

Xu, S., Huang, X., Xu, H., and Zhang, C. (2007). Improved prediction of coreceptor usage and phenotype of HIV-1 based on combined features of V3 loop sequence using random forest. *J Microbiol* **45**(5)**,** 441-6.

Yamada, T., and Iwamoto, A. (2000). Comparison of proviral accessory genes between long-term nonprogressors and progressors of human immunodeficiency virus type 1 infection. *Arch Virol* **145**(5)**,** 1021-7.

Yang, W., Bielawski, J. P., and Yang, Z. (2003). Widespread adaptive evolution in the human immunodeficiency virus type 1 genome. *J Mol Evol* **57**(2)**,** 212-21.

Yoshimura, F. K., Diem, K., Learn, G. H., Jr., Riddell, S., and Corey, L. (1996). Intrapatient sequence variation of the gag gene of human immunodeficiency virus type 1 plasma virions. *J Virol* **70**(12)**,** 8879-87.

Zaitseva, L., Cherepanov, P., Leyens, L., Wilson, S. J., Rasaiyaah, J., and Fassati, A. (2009). HIV-1 exploits importin 7 to maximize nuclear import of its DNA genome. *Retrovirology* **6,** 11.

Zhang, L. Q., MacKenzie, P., Cleland, A., Holmes, E. C., Brown, A. J., and Simmonds, P. (1993). Selection for specific sequences in the external envelope protein of human immunodeficiency virus type 1 upon primary infection. *J Virol* **67**(6)**,** 3345-56.

Zhang, M., Schultz, A. K., Calef, C., Kuiken, C., Leitner, T., Korber, B., Morgenstern, B., and Stanke, M. (2006). jpHMM at GOBICS: a web server to detect genomic recombinations in HIV-1. *Nucleic Acids Res* **34**(Web Server issue)**,** W463-5.

Zhou, Q., Chen, D., Pierstorff, E., and Luo, K. (1998). Transcription elongation factor P-TEFb mediates Tat activation of HIV-1 transcription at multiple stages. *EMBO J* **17**(13)**,** 3681-91.

Zhou, Q., and Sharp, P. A. (1995). Novel mechanism and factor for regulation by HIV-1 Tat. *EMBO J* **14**(2)**,** 321-8.

Zhou, W., Parent, L. J., Wills, J. W., and Resh, M. D. (1994). Identification of a membrane-binding domain within the amino-terminal region of human immunodeficiency virus type 1 Gag protein which interacts with acidic phospholipids. *J Virol* **68**(4)**,** 2556-69.

Zhu, P., Liu, J., Bess, J., Jr., Chertova, E., Lifson, J. D., Grise, H., Ofek, G. A., Taylor, K. A., and Roux, K. H. (2006). Distribution and three-dimensional structure of AIDS virus envelope spikes. *Nature* **441**(7095)**,** 847-52.

Zhuang, J., Jetzt, A. E., Sun, G., Yu, H., Klarmann, G., Ron, Y., Preston, B. D., and Dougherty, J. P. (2002). Human immunodeficiency virus type 1 recombination: rate, fidelity, and putative hot spots. *J Virol* **76**(22)**,** 11273-82.

Appendices

Appendix 1

Table 1. List of the primer names used in the RT PCR from Chapter 2, section 2.2.4.1. The primer sequences are not available.

Primer name	RT PCR primer was used in
3' LTR B CAT	Reverse Transcriptase PCR (a)
Env End	Reverse Transcriptase PCR (b)

Table 2. A list of both the primary and secondary primers used to amplify the full HIV-1 genome in two overlapping fragments from Chapter 2, section 2.2.4.1. The primer sequences are not available.

Outside primary PCR primers		Inside secondary PCR primers	
5' half	Vif-Reverse 1	5' half	CMV-PBS
	HIV-R-Start		dmvifRT2
3' half	Sense 5193-5213.vif	3' half	dmvifRT1
	pREC-HIV-R-end-and-forRT-PCR		pREC-HIV-R-end-anti

Table 3. A list of the primers designed to amplify full *env* and *gag* from Chapter 2, section 2.2.4.1. The primer sequences are not available.

Gene	Outside primary PCR primers	Inside Secondary PCR primers
env	Env A B-nef.vif.nef	KpnI.Env Env End
	Env A Env End	KpnI.Env Env End
	Tat Rec Con Fwd 3 Tat Rec Con Bwd 7	Int Fwd gp120 3 Tat Rec Con Bwd 4
	Tat Rec Con Fwd 3 Tat Rec Con Bwd 7	F Env Ecto 1 B env MSD2
gag	Gag p17 short Fwd 7 B Gag p6.5	Gag p17 rec Fwd Gag p6 rec Bwd

Appendices 2 – 7

Appendix 2 (On disc provided)

Phylogenetic reconstruction of *env* identifying LTS30 as an unclassifiable subtype.

Maximum Likelihood tree of *env* sequences from Karonga along with subtype reference sequences from the LANL database (http://www.hiv.lanl.gov). Branches leading to the unclassifiable subtype are red.

Appendix 3 (On disc provided)

Phylogenetic reconstruction of *gag* identifying LTS30 as an unclassifiable subtype.

Maximum Likelihood tree of *gag* sequences from Karonga along with subtype reference sequences from the LANL database (http://www.hiv.lanl.gov). Branches leading to the unclassifiable subtype are red.

Appendix 4 (On disc provided)

Phylogenetic reconstruction of *env* clonal and consensus sequences from the Karonga LTS.

Maximum Likelihood tree of *env* consensus and clonal sequences from the Subtype C LTS and 40 control subtype C sequences from Karonga. The subtype C ancestral sequence from Travers et al. (2004) was used as an out-group. ● = clonal sequences generated from samples retrieved in 2004, □ = clonal sequences generated from samples retrieved in 2008, ■ = clonal sequences generated from samples retrieved in 2010. LTS sequences are labelled with the LTS number and the year of sample collection.

Appendix 5 (On disc provided)

Phylogenetic reconstruction of *gag* clonal and consensus sequences from the Karonga LTS.

Maximum Likelihood tree of *gag* consensus and clonal sequences from the Subtype C LTS and 40 control subtype C sequences from Karonga. The subtype C ancestral sequence from Travers et al. (2004) was used as an out-group. ● = clonal sequences generated from

samples retrieved in 2004, ■ = clonal sequences generated from samples retrieved in 2010. LTS sequences are labelled with the LTS number and the year of sample collection.

Appendix 6 (On disc provided)

Phylogenetic reconstruction of *env* sequences from Karonga from between 1983-2009.

Maximum Likelihood tree of all *env* subtype C sequences from Karonga including epidemiologically linked individuals (e.g. spouse pairs). The subtype C ancestral sequence from Travers et al. (2004) was used as an out-group. Red branches represent groupings that received over 70 % bootstrap support.

Appendix 7 (On disc provided)

Phylogenetic reconstruction of *gag* sequences from Karonga from between 1983-2008.

Maximum Likelihood tree of all *gag* subtype C sequences from Karonga including epidemiologically linked individuals (e.g. spouse pairs). The subtype C ancestral sequence from Travers et al. (2004) was used as an out-group. Red branches represent groupings that received over 70 % bootstrap support.

Publication

AIDS RESEARCH AND HUMAN RETROVIRUSES
Volume 28, Number 3, 2012

DOI: 10.1089/aid.2011.0168

HIV Type 1 Mutational Patterns in HIV Type 1 Subtype C-Infected Long-Term Survivors in Karonga District Malawi: Further Analysis and Correction

Ishla Seager,[1] Michael D. Leeson,[1] Amelia C. Crampin,[2,3] Dominic Mulawa,[2] Neil French,[2,3] Judith R. Glynn,[3] Simon A.A. Travers,[1,4] and Grace P. McCormack[1]

Abstract

Here we present new sequence data from HIV-1 subtype C-infected long-term survivors (LTS) from Karonga District, Malawi. *Gag* and *env* sequence data were produced from nine individuals each of whom has been HIV-1 positive for more than 20 years. We show that the three amino acid deletion in *gag* p17 previously described from these LTS is not real and was a result of an alignment error. We find that the use of dried blood spots for DNA-based studies is limited after storage for 20 years. We also show some unlikely amino acid changes in *env* C2-V3 in LTS over time and different patterns of genetic divergence among LTS. Although no clear association between mutations and survival could be shown, amino acid changes that are present in more than one LTS may, in the future, be shown to be important.

The natural history of HIV-1 infection and disease progression has been well established in adults, with 10 years being the median time from initial infection to the development of AIDS in the absence of therapeutic intervention.[1,2] However, this can vary widely within individuals, with rapid progressors developing AIDS symptoms in as little as 6 months, and other individuals, such as long-term survivors (LTS), who progress to AIDS over a much longer period of time.[3] LTS are those individuals characterized as having survived for >10 years, without antiretroviral therapy (ART), but who also show a steady decline in the number of circulating CD4 cells.[4]

Some studies have shown viral factors play an important role in the survival of LTS, for example, viruses that contain defects in particular HIV-1 genes, such as *env, gag, nef, vpu, vif, rev,* and *tat*.[5] Currently there is very little information for LTS found in sub-Saharan Africa.[6] McCormack *et al*.[7] described a three amino acid deletion at the end of *gag* p17 found in 15 LTS from Karonga District, Malawi. In each LTS the deletion was observed in sequences dating from the late 1990s but was not present in any sequences dating from the 1980s. It was also described in two-thirds of the other HIV-1-positive individuals from Karonga District included in that study from the late 1990s and it was suggested that the deletion could be associated with longer survival and onward transmission.[7] In this work we sought to further characterize the viral factors involved in long-term survival in Karonga District, Malawi. We find that the previous observation of a three amino acid deletion[7] was, in fact, erroneous, and we describe our follow-up study of the 38 long-term survivors.

Thirty-eight HIV-positive individuals were seen in Karonga District Malawi in both the late 1980s and late 1990s.[2] Seventeen of them were still alive in 2004.[7] Fourteen of these were sought again in 2010 (three were not sought as they had refused to participate the previous two times) when it was found that three had died, one had left the region, and one refused to provide a sample. Of the nine individuals seen, eight were infected with subtype C and one (LTS30) was infected with the unclassifiable strain that was described for this region.[8] No amplification was possible from dried blood spots (DBS) from this latter individual prior to 2004. Five individuals had begun ART (one in 2005, two in 2006, one in 2008, and one in 2009). Four individuals had not begun ART and are thus HIV positive without treatment for a minimum of 21 years (although one of these has also now been referred) (Table 1).

DNA was reextracted from DBS collected between 1986 and 1989 from LTS and the wider population with a view to

[1]Molecular Evolution and Systematics Laboratory, Zoology, Ryan Institute, School of Natural Sciences, National University of Ireland, Galway, Ireland.
[2]Karonga Prevention Study, Chilumba, Malawi.
[3]Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, United Kingdom.
[4]South African National Bioinformatics Institute, University of the Western Cape, Bellville, South Africa.

Table 1. Summary of Sequence Data, CD4 Counts, and ART Information Available for Those LTS Found to Be Still Present in Karonga District in 2004 and in 2010

Survivor	*Year*	*Sequence data available*	*CD4 count (cells/mm^3)*	*ART start date*
LTS1	1999	*gag* and *env*		
	2004		586	
	2010	*gag* and *env*	449	
LTS2	1999	*gag* and *env*		
	2010	*gag* and *env*	244	ART 2009
LTS5	1989	*gag*		
	1999	*gag* and *env*		
	2004	*gag*		
	2010	*gag*	789	ART 2008
LTS8	1989	*gag* and *env*		
	1999	*gag* and *env*		
	2004	*gag* and *env*	265	
	2010	*gag* and *env*	734	ART 2005
LTS9[a]	1988	*gag* and *env*		
	1999	*gag* and *env*		
	2004	*gag* and *env*	56	
LTS10	1989	*gag* and *env*		
	1999	*gag* and *env*		
	2004	*gag* and *env*	328	
	2010	*gag* and *env*	138	Referred for ART 2010
LTS12	1988	*env*		
	1999	*gag* and *env*		
	2004	*gag* and *env*	439	
	2010	*gag* and *env*	390	ART 2006
LTS17[b]	1989	*gag* and *env*		
	1998	*gag* and *env*		
	2004	*gag* and *env*	452	
LTS20	1988	*gag* and *env*		
	2000	*gag* and *env*		
	2004	*gag* and *env*		
	2010	*gag* and *env*	139	ART 2006
LTS21	1989	*gag* and *env*		
	1999	*gag* and *env*		
	2004		47	
	2010	*gag* and *env*	32	Refused ART
LTS22[a]	1989	*gag* and *env*		
	1998	*gag* and *env*		
	2004	*gag* and *env*	475	
LTS30[c]	2004	*gag*	656	
	2010	*gag* and *env*	362	

[a]Had died by 2010.
[b]Had left the Karonga District by 2010.
[c]Unclassifiable subtype.
ART, antiretroviral therapy; LTS, long-term survivors.

cloning polymerase chain reaction (PCR) products to explore evidence of the three amino acid deletion in that time period. DBS were also available from 10 of the LTS seen in 2004 and cell pellets from 9 in 2010. Plasma or cell pellets were utilized from 100 individuals randomly chosen from samples collected between 2008 and 2010, from existing studies in the District to explore the frequency of the three amino acid deletion in the 2008–2010 time period. Proviral DNA was extracted from 200 μl of cell pellet or plasma using the QIAamp DNA Blood Mini Kit (Qiagen) blood and from the DBS using the QIAamp DNA Micro Kit (Qiagen). Nested PCR and sequencing of a 750-bp region of *gag* p17p24 and a 500-bp region of *env* C2V3 were carried out as previously described.[9] Amplification from the DBS from the 1980s was largely unsuccessful. DBS had been frozen at −20°C for over 20 years and exposed to a number of freeze-thaw events and it is highly probable that this has led to fragmentation of the DNA present. Previous studies found −20°C suitable for long-term storage of DBS (6 years)[10–13] but our work suggests an upper limit to the length of time such samples can be stored successfully in this way for DNA-based studies. Cloning was carried out using the Topo TA cloning kit (Biosciences). Automatic sequencing in both directions was carried out by Eurofins Genetic Services Ltd or by LGC Genomics. Sequence chromatographs were examined and manually edited in Seqman (DNAStar Inc.).

The three amino acid deletion was not found in any *gag* sequences produced from LTS in 2004 and 2010 or in over 100 sequences produced from blood samples collected in 2008–2010. Furthermore, none of 50 sequences from 50 clones produced from the 1998 sample from LTS2 (which had previously showed the deletion) contained the deletion. We then reexamined all of the raw data from the sequences used in McCormack *et al*.[7] and the deletion was not found in any of the sequences. We suggest that an alignment error was made at an early stage of the multiple alignment assembly of the relevant sequences in the original study. This serves as a stark reminder of the dangers of such errors when handling large numbers of sequences. All affected sequences from McCormack *et al*.[7] that were submitted to GenBank have been reexamined and the correct sequences redeposited.

Multiple alignments of all sequential *env* (74) and *gag* (65) sequences from subtype C-infected LTS along with 40 control sequences were assembled and optimized in MacClade 4 (Sinauer Associates). Phylogenetic trees were reconstructed under the GTR + gamma model of DNA substitution implemented by RAxML 7.0.3[14] with all parameters optimized by RAxML. Confidence levels in the groupings in the phylogeny were assessed using 1000 bootstrap replicates as part of the RAxML phylogeny reconstruction. The subtype C ancestral sequence derived in previous work[15] was employed as the out-group for both *gag* and *env* trees. Both gene trees showed that for most individuals the sequences from the different time points grouped together (8/10 for *env* and 10/11 for *gag*) but only half grouped with significant bootstrap support (Fig. 1a and b). Sequences from LTS21 formed multiple clusters on both gene trees, which is consistent with the pattern seen in McCormack *et al*.[7]

To further explore this, additional *gag* and *env* consensus sequences were produced from DNA reextracted from the DBS collected in 1989 and 1999 for this individual (LTS21) with additional *env* sequences produced from the 2010 DNA sample. In *gag* the 1990s sequences were ancestral to the 2010 sequences and the two 1989 sequences grouped distantly from them (Fig. 1b). In *env* the 2010 sequences showed further variation with two sequences grouping away from the 1999 sequences (Fig. 1a). The average genetic distance between the eight *env* sequences (across all time points) was 12%, which was higher than the 8.8% genetic distance between all other LTS sequences from all individuals at all the different time points. The genetic distance between two of the sequences collected from 2010 was even higher (17.5%) than between the sequences collected in the 1980s and 1990s (7.8%) from the same individual.

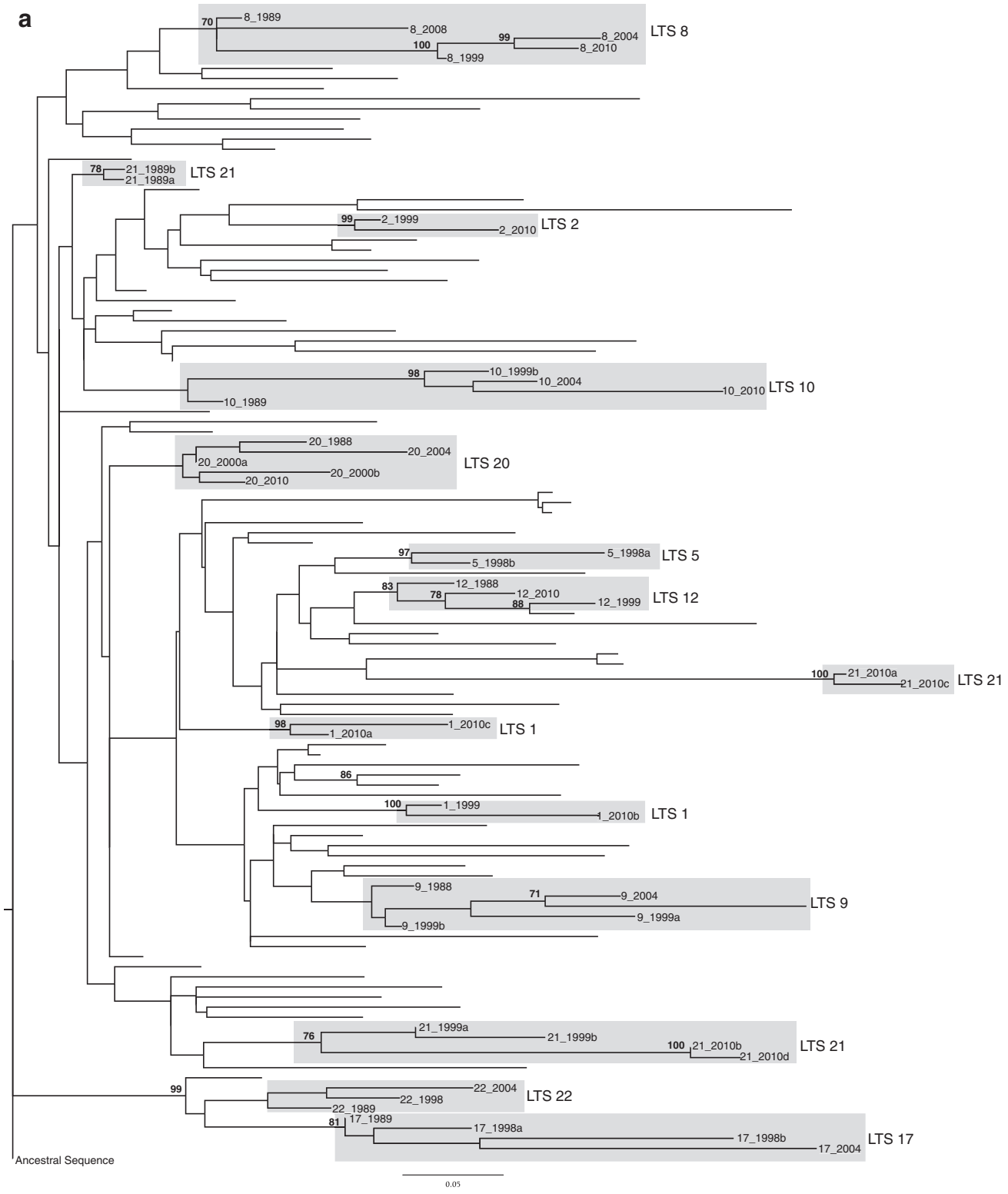


FIG. 1. Maximum Likelihood trees generated from (**a**) *env* and (**b**) *gag* gene sequences from long-term survivors (LTS) dating from 2004 and 2010 along with local control sequences. Only the LTS sequences are labeled and are named by the LTS number (see Table 1) and the year the sample was collected. Sequences labeled a, b, c, and d refer to multiple consensus sequences generated from the same time point. Bootstrap values of over 70 are marked on the relevant branches.

(*Figure Continued* →)

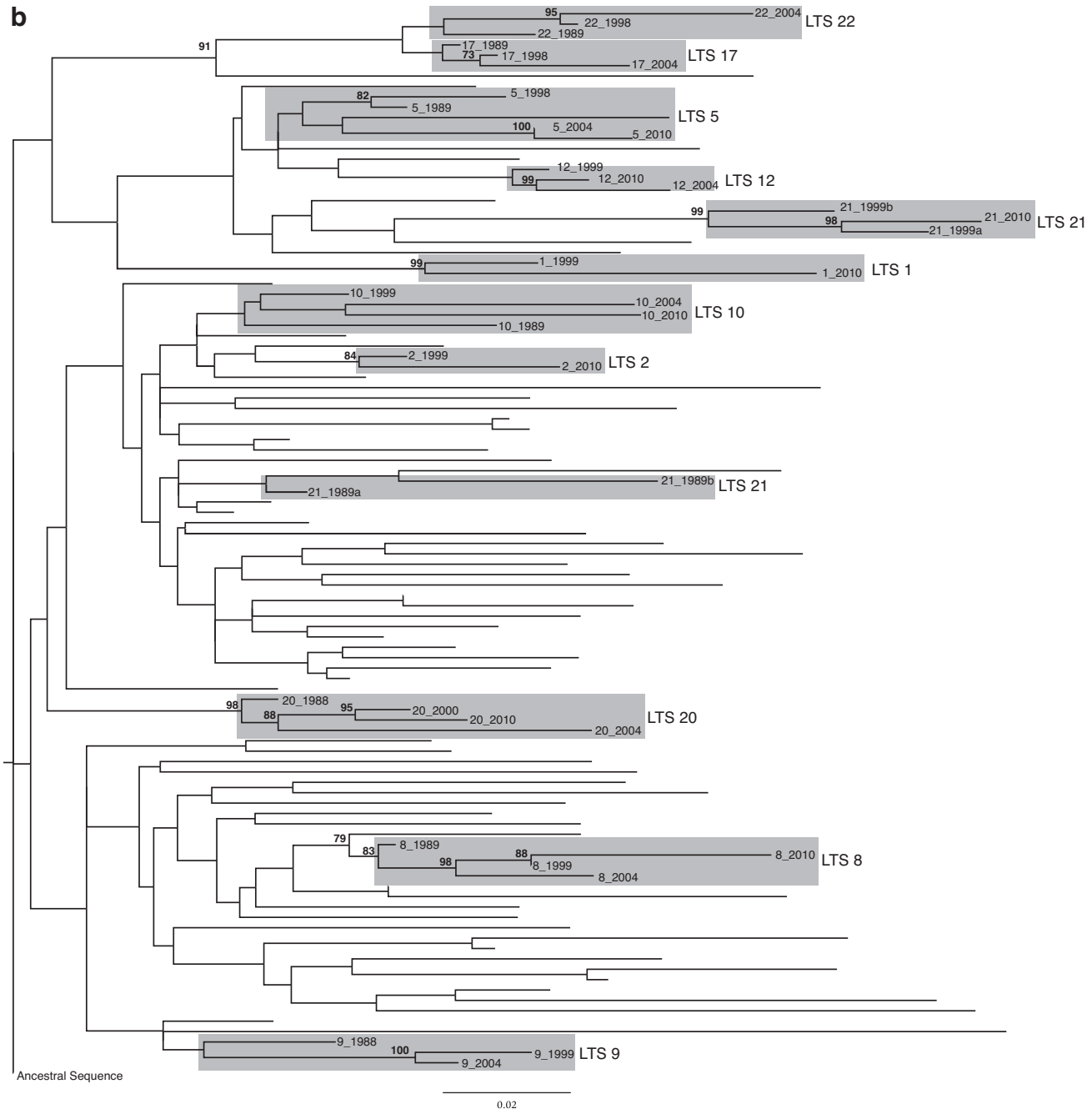


FIG. 1. (Continued).

Although sample mislabeling is very unlikely, as the individual's name was written on the filter paper as well as a unique identifier, we cannot exclude the possibility. Superinfection is also possible. The sequences came from a female who was 21 when she was first identified as being HIV-1 positive. She has maintained a very low CD4 count for the past 6 years, 47 cells/mm^3 in 2004 and 32 cells/mm^3 in 2010. At both visits she was described as being healthy and showing no signs of AIDS and she refused ART on both occasions. Divergent *env* sequences were also found among sequences from the 2010 sample of LTS1, a female who was 37 when she was first identified as being HIV-1 positive. Her CD4 count was 586 cells/mm^3 in 2004 and had fallen slightly to 449 cells/mm^3 in 2010 and at that time she had not been referred for ART.

The BLOSUM62 matrix was used to assess the likelihood of amino acid substitutions between sequential sequences of *env* for each LTS, with indels and shared deletions relative to the other sequences also noted. A graphic representation of these observed amino acid substitutions within each LTS sample from one time period to another shows the large amount of change that was apparent across most of *env* C2-V3 in all individuals (Fig. 2). There are a number of positions that

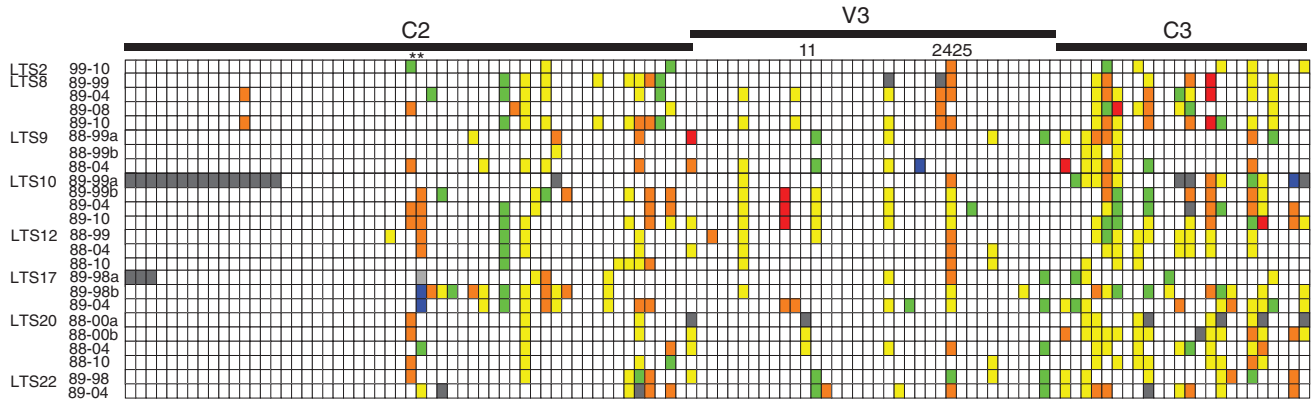


FIG. 2. A graphic representation of observed amino acid substitutions that have occurred in *env* within survivors over time. Substitutions are color coded by likelihood according to the BLOSUM62 matrix with green being the most likely and red the least (green > yellow > orange > red); blue indicates an insertion in one of the sequences relative to the other, pale gray indicates a shared deletion, and dark gray indicates an ambiguous site, e.g., a stop codon. The sequence collected at the earliest available time point was compared to all available *env* sequences from subsequent time points. The comparisons are labeled by the LTS number and the years being compared. Those labeled with an a or b refer to multiple sequences generated from the same time point. **Mark positions 268 and 269 in *env* using HXB2 numbering.

changed within nearly all of the LTS, e.g., HXB2 positions 268–269 showed substitutions in eight LTS (asterisk on Fig. 2). Some of these changes are less likely mutations, e.g., in six individuals (LTS8, 9, 10, 12, 20, and 22) there was a change from glycine to glutamic acid or vice versa. In one individual (LTS17) there a deletion at position 269 within the 1980s and one of the 1990s *env* sequences, which became a lysine in a second 1990s and 2004 sequence. Mutations at positions 11, 24, and 25 within the V3 loop have been associated with a change in coreceptor usage with a shift from negatively charged amino acids to positively charge amino acids being suggested to result in a switch from the use of CCR5 to CXCR4 usage. Only LTS17 showed evidence of a change to a positive charge in this region by 2004 but had left the district by 2010 and so we do not have any additional information on this person. A large amount of change was seen just after the V3 loop in the C3 region. Indeed the high degree of genetic divergence seen between multiple *env* sequences from the same time point in some LTS rendered comparisons of pairwise genetic distances between time points meaningless (even when we excluded the individuals with possible superinfection mentioned above). For example, two consensus sequences from LTS17 from 1999 showed a genetic distance of 4%. The population of viruses within individuals is currently being explored further by sequencing multiple clones.

Comparing amino acid substitutions in *gag* sequences showed higher numbers of substitutions within the *gag* p17

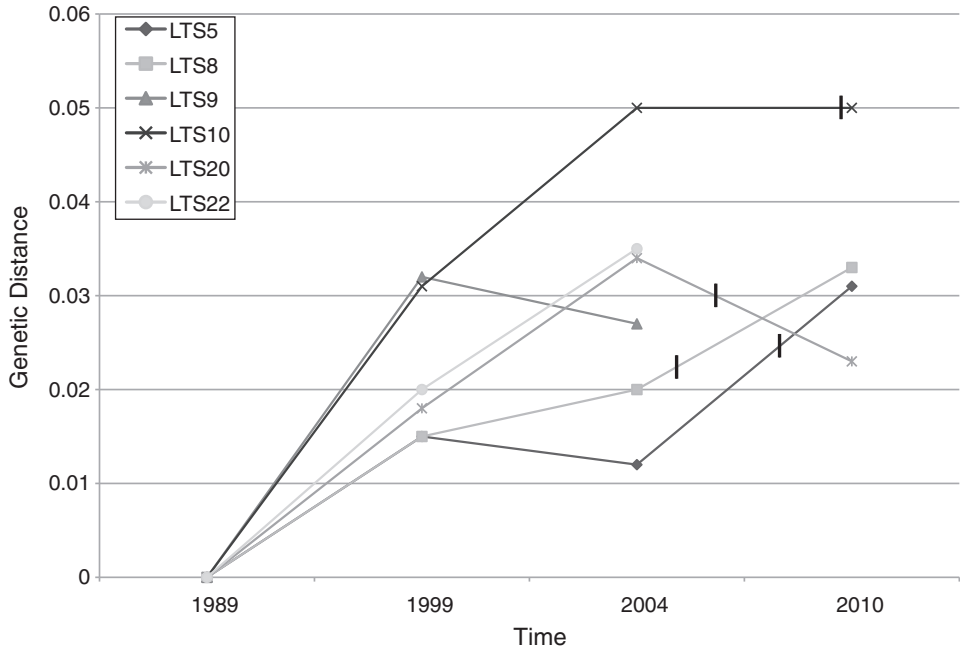


FIG. 3. The genetic divergence seen in *gag* over time in LTS5, 8, 9, 10, and 20. All sequences from one individual from different time points were compared to the sequence generated from the earliest time point available. LTS9 and 22 had died before 2010. LTS5 had began antiretroviral therapy (ART) in 2008, LTS8 in 2005, LTS20 in 2006, and LTS10 had been referred for ART in 2010 as indicated by the vertical lines.

domain when compared to p24 as might be expected and, using the BLOSUM62 matrix as a reference, most of these substitutions were changes that were more likely to occur (data not shown). Different patterns of genetic divergence in *gag*, calculated using pairwise distances, were apparent among the LTS over time. Two LTS (LTS5 and LTS8) showed an overall trend of an increase in genetic divergence in *gag* over time as can be seen in Fig. 3. Both LTS10 and LTS20 showed a general increase in genetic divergence from 1989 to 2004; however, in 2010 the amount of genetic divergence from 1989 decreased for LTS20 and plateaued in LTS10. Three of these individuals had begun ART between 2004 and 2010 and one has since been referred. Two individuals (LTS9 and LTS22) who were alive in 2004 but had died by 2010 showed very different patterns of divergence. LTS9 showed an increase in divergence from 1989 to 1999, which decreased in 2004, while in LTS22 there was a linear increase in divergence from 1989 to 2004 (Fig. 3).

In summary, we present new sequence data from HIV-1 subtype C-infected long-term survivors from Malawi, Africa. This work highlights some of the pitfalls associated with sequence analysis. We show that the three amino acid deletion in *gag* p17 previously described from these LTS[7] was the result of an alignment error. Extreme caution must be used while making inferences from sequence data as some errors may be very well hidden. However, data on long-term survivors infected with subtype C virus are important to accumulate. We show many amino acid changes in *env* C2-V3 in LTS over time. Although no clear association between mutations and survival could be shown, amino acid changes that are present in a number of LTS may in the future be shown to be important for survival, but future work in this regard will require data from virus and host.

Sequence Data

All sequences have been deposited into GenBank, accession numbers JN393505–393554, and raw data and alignments employed in this work are available from the authors on request. JN393505–393554

Acknowledgments

This material is based upon works supported by Science Foundation Ireland under Grant 07/RFP/EEEOBF424 and by an IRCSET EMBARK Scholarship to Ishla Seager (RS/2007/203). The Karonga Prevention Study is funded primarily by the Wellcome Trust, with contributions from LEPRA. Permission for the study was received from the National Health Sciences Research Committee, Malawi, and the Ethics Committee of the London School of Hygiene and Tropical Medicine, United Kingdom.

Author Disclosure Statement

No competing financial interests exist.

References

1. Langford SE, Ananworanich J, and Cooper DA: Predictors of disease progression in HIV infection: A review. AIDS Res Ther 2007;4:11.
2. Crampin AC, Floyd S, Glynn JR, *et al.:* Long-term follow-up of HIV-positive and HIV-negative individuals in rural Malawi. AIDS 2002;16(11):1545–1550.
3. Pantaleo G and Fauci AS: Immunopathogenesis of HIV infection. Annu Rev Microbiol 1996;50:825–854.
4. Learmont J, Tindall B, Evans L, *et al.:* Long-term symptomless HIV-1 infection in recipients of blood products from a single donor. Lancet 1992;340(8824):863–867.
5. Poropatich K and Sullivan DJ Jr: Human immunodeficiency virus type 1 long-term non-progressors: The viral, genetic and immunological basis for disease non-progression. J Gen Virol 2010;92(Pt 2):247–268.
6. Laeyendecker O, Redd AD, Lutalo T, *et al.:* Frequency of long-term nonprogressors in HIV-1 seroconverters from Rakai Uganda. J Acquir Immune Defic Syndr 2009;52(3):316–319.
7. McCormack GP, Glynn JR, Clewley JP, *et al.:* Emergence of a three codon deletion in gag p17 in HIV type 1 subtype C long-term survivors, and general population spread. AIDS Res Hum Retroviruses 2006;22(2):195–201.
8. McCormack GP, Glynn JR, Sibande F, *et al.:* Highly divergent HIV-1 group M sequences in Karonga District, Malawi in the early 1980s. AIDS Res Human Retroviruses 2003;19(5):441–445.
9. McCormack GP, Glynn JR, Crampin AC, *et al.:* Early evolution of the human immunodeficiency virus type 1 subtype C epidemic in rural Malawi. J Virol 2002;76(24):12890–12899.
10. Cassol S, Salas T, Arella M, Neumann P, Schechter MT, and O'Shaughnessy M: Use of dried blood spot specimens in the detection of human immunodeficiency virus type 1 by the polymerase chain reaction. J Clin Microbiol 1991;29(4):667–671.
11. Cassol S, Salas T, Gill MJ, *et al.:* Stability of dried blood spot specimens for detection of human immunodeficiency virus DNA by polymerase chain reaction. J Clin Microbiol 1992;30(12):3039–3042.
12. McNulty A, Jennings C, Bennett D, *et al.:* Evaluation of dried blood spots for human immunodeficiency virus type 1 drug resistance testing. J Clin Microbiol 2007;45(2):517–521.
13. Steegen K, Luchters S, Demecheleer E, *et al.:* Feasibility of detecting human immunodeficiency virus type 1 drug resistance in DNA extracted from whole blood or dried blood spots. J Clin Microbiol 2007;45(10):3342–3351.
14. Stamatakis A: RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 2006;22(21):2688–2690.
15. Travers SA, Clewley JP, Glynn JR, *et al.:* Timing and reconstruction of the most recent common ancestor of the subtype C clade of human immunodeficiency virus type 1. J Virol 2004;78(19):10501–10506.

Address correspondence to:
Grace P. McCormack
Molecular Evolution and Systematics Laboratory
Department of Zoology
Ryan Institute, School of Natural Sciences
National University of Ireland
Galway
Ireland

E-mail: grace.mccormack@nuigalway.ie