The Role of HIV-1 Recombination and APOBEC3F/G-Mediated Hypermutation in HIV-1 Pathogenesis

by

Allison M. Land

A thesis submitted to the Faculty of Graduate Studies of
the University of Manitoba
in partial fulfillment of the degree of

Doctor of Philosophy

Department of Medical Microbiology
University of Manitoba
Winnipeg

The Role of HIV-1 Recombination and APOBEC3F/G-Mediated Hypermutation in HIV-1 Pathogenesis

By

Allison M. Land

A Thesis/Practicum submitted to the Faculty of Graduate Studies of The University of Manitoba in partial fulfillment of the requirement of the degree

Of

Doctor of Philosophy

Abstract

HIV-1, causative agent of the devastating pandemic of AIDS, shows incredible sequence variation, providing a substantial challenge for vaccine design. The aim of this thesis is to characterize HIV-1 diversity by examining levels of inter-subtype recombination and APOBEC-mediated proviral hypermutation, and determine its importance, by associating diversity with pathogenesis.

The results of this thesis were obtained by sequencing samples collected from subjects enrolled in two cohorts located in Nairobi, Kenya, composed of individuals involved in commercial sex work (CSW) and those who are not. We found, in agreement with previous studies, that CSW were more likely to be infected with recombinant virus than non-sex workers. This suggests that all individuals at high risk for HIV acquisition may be important drivers of viral diversity in the global pandemic and are thus an important target for prevention and intervention strategies.

We identified a subset of individuals with high levels of proviral HIV-1 APOBEC-mediated hypermutation, which correlated with $CD4^+$ count. This indicated that APOBEC3F/3G hypermutation may be important in controlling disease progression. Sequencing the APOBEC3G gene revealed three polymorphisms that were significantly associated with hypermutation; two of these were located in the region 5' of the APOBEC3G gene and may control expression. This data suggests that contrary to previous studies, APOBEC-mediated hypermutation may be controlled by increased activity of host APOBEC3G rather than by defects in the viral Vif. We thus suggest that increases in APOBEC3F/G activity may play a protective role in disease progression.

The exploitation of these findings may aid in the development of HIV prevention and therapeutic strategies.

Acknowledgments

There are numerous people who have helped me with this thesis; I attempt to acknowledge them here, but am conscious that this thank you inadequately matches the contribution.

First thanks to my supervisor, Dr. Frank Plummer, who has the amazing ability to fit things into the larger picture, and thus transform the feeling that one's work is trivial and unimportant into a feeling that it may actually be worthwhile.

Next thanks to my mentor, Dr. Blake Ball. He has helped me with every stage of this journey, providing honest input and allowing me to be honest in return. He has provided advice when asked for and motivation when needed; both were necessary.

Thank you to Dr. Ma Luo, my lab supervisor. She has been encouraging and supportive, from the very beginning.

Thank you to my advisory committee, especially Dr. Deb Court who has been influencing my scientific career since I was a second year undergrad, and Dr. Keith Fowke who challenges me.

Thank you to the grad students, post-docs, technicians and undergrads who have worked in the NML laboratory over the years. I loved coming to work every morning, and that has everything to do with you.

Thank you to all the grad students in the Plummer/Fowke labs. Thank you for beers and coffee. Thank you for sharing the good lab days, and especially for commiserating with me on the bad lab days. Thank you to the ones who graduated first for setting such a good example. Best of luck to the ones who will graduate after me.

Thank you to all the technicians and administrative staff at the NML and in Med Micro. Thank you for being able to help me get things done.

Finally, thank you to all the women who make this work possible. *Assante sana.*

Dedication

I would like to dedicate this thesis first to my parents, who instilled a sense of curiosity and a love of all things science in me, from an early age: you support and encourage me, put my happiness first and do all you can to help. Thank you.

Next to my husband, who has agreed to be taken along on this ride: Thank you for putting up with this life choice, which drives me crazy and makes me happy in often rapid succession. You support me in so many ways, from listening to me practice presentations countless times, to reassuring me that this is all possible. You are my rock.

Finally, to my daughter: your arrival has provided a keen motivation to wrap this up. Thank you for taking long naps when I needed to focus on writing, and for being wakeful when I needed something to pull me away. You remind me what is truly important and keep me grounded.

Table of Contents

List of Tables

List of Figures

Introduction

1.1 Origin of HIV

HIV-1 (human immunodeficiency virus 1) is a member of the *Retroviridae* family and the *Lentivirinae* genus, which is characterized by a long incubation period. It is part of the primate lentivirus group, which includes HIV-1, HIV-2 and SIV (simian immunodeficiency virus). HIV causes AIDS (acquired immunodeficiency syndrome), which is a disease of the human immune system characterized by a decrease in $CD4^+$ T cell count to less than 200 cells/μL (40). Over forty different species of African nonhuman primates are estimated to be infected with different lentiviral SIV infections (295). Unlike HIV, SIVs do not cause a significant depletion of $CD4^+$ T cells in the peripheral blood nor cause AIDS-like illness in their natural hosts (108,306). This is despite the natural SIV hosts maintaining high viral loads and a short *in vivo* lifespan of SIV-infected cells, suggesting high cellular pathogenicity (103,242,265).

Interestingly, disease does occur in non-natural hosts, such as infection with SIV of Asian macaques. Cross-species transmission of SIV has been shown to occur when primates in captivity are housed in shared quarters. SIV-harbouring sooty mangabeys infected macaques on numerous occasions prior to the identification of SIV and thus before testing and precautions would have been undertaken (63,170,180,208,274). Cross-species transmission has also doubtlessly occurred in nature, as at least eight SIVs are recombinant (i.e. derived from multiple, genetically distinct ancestors), indicating that a single host was infected with multiple viruses (7). For example, SIVcpz, which infects

chimpanzees, is a recombinant virus derived from ancestral SIVs that currently infect red-capped mangabeys and *Cercopithecus* monkeys in west-central Africa (16).

There are four subspecies of chimpanzee: *Pan troglodytes verus, P .t. vellerosus, P. t. troglodytes* and *P. t. schweinfurthii*, the latter two of which can be infected with SIVcpz (261). The closest relative of HIV-1 is SIVcpz*Ptt*, which infects the chimpanzee sub-species *Pan troglodytes troglodytes* (137). Transmission of the virus from chimpanzees to humans likely occurred parenterally when chimpanzees were butchered for bushmeat (260). Indeed, parenteral transmission from bites and other wounds may be the major route of SIV transmission in non-human primates (116). This cross-species transmission of an SIV to humans has happened at least three separate times, giving rise to three genetically distinct groups, HIV-1 groups M, N and O. HIV-1 group M and N viruses are derived from SIVcpz*Ptt*, transmitted by two geographically distinct groups of chimpanzees (137). The origin of group O HIV-1 is less well understood; the closest extant SIV viruses are found in gorillas (292). These viruses form a phylogenetic cluster intermingled with SIVcpz strains, however, suggesting that chimpanzees may have transmitted the virus to gorillas, which in turn passed the virus on to humans; alternatively, the common ancestor virus infected humans and gorillas separately (292).

HIV-1 group M is the most prevalent and widespread HIV that currently circulates in the human population. It is found worldwide and causes more than 95% of HIV infections (260). Group M is genetically diverse and can be further divided in clades or subtypes, based on significant phylogenetic clustering across the genome (Figure 1) (243). The

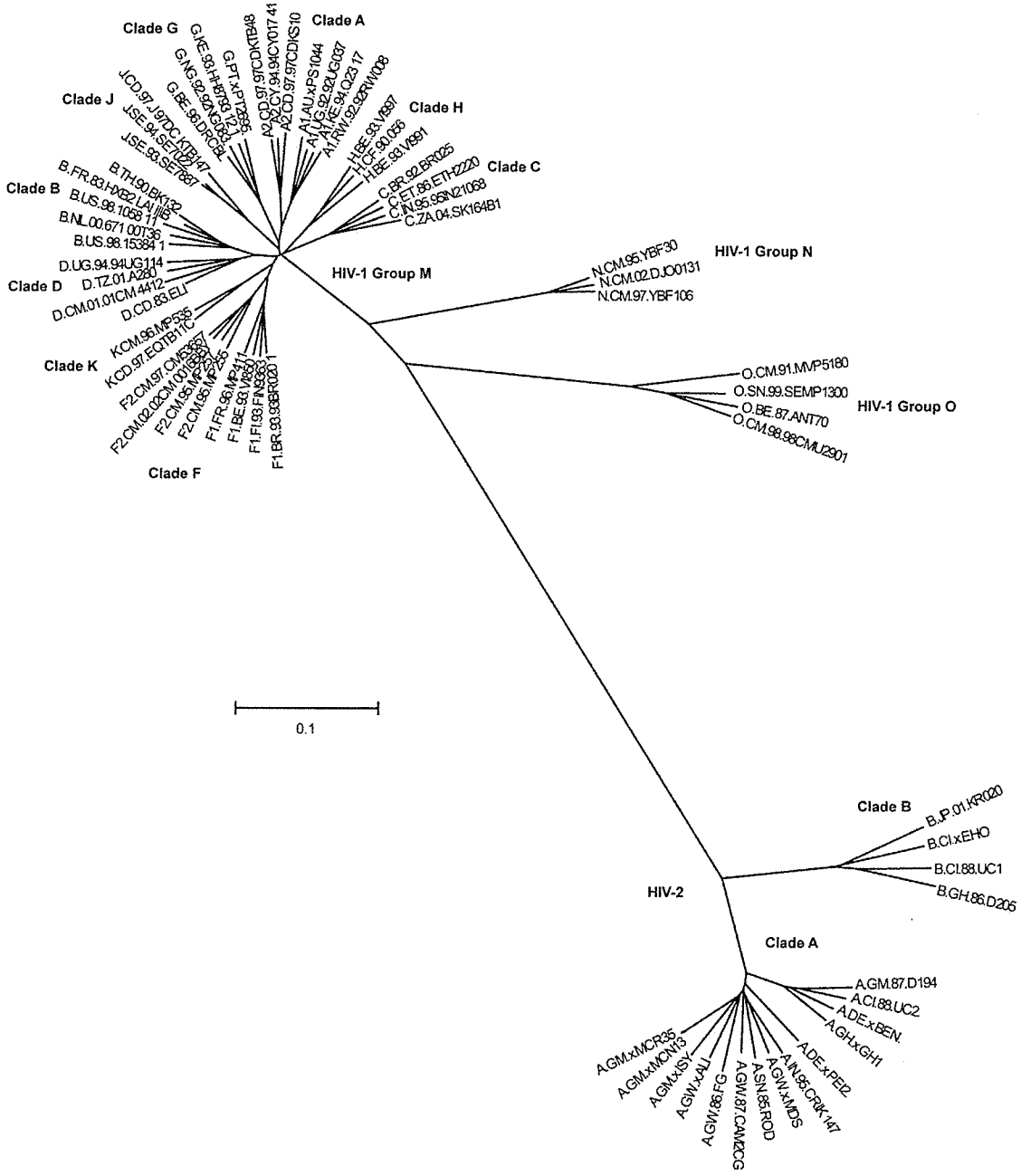


Figure 1. Neighbour-joining tree of full-length HIV-1 and HIV-2 reference sequences. Current reference sequences for HIV-1 group M, N and O and HIV-2 were obtained from the HIV Sequence Database (www.hiv.lanl.gov). The neighbour-joining tree was generated in MEGA 4 from a ClustalW alignment. Genetic distance is indicated.

clades currently identified as circulating in the global population are A, B, C, D, F, G, H, J and K. The different clades, many of which were in existence fifty years ago, likely arose due to population founder effects (260). These different viral subtypes, just like the different SIVs, can genetically recombine to form new recombinant viruses. Many different recombinant viruses have been described – over forty are actively circulating (each are found in at least three epidemiologically unlinked individuals) and are thus termed circulating recombinant forms (CRF), while many more have been described in only a single person (unique recombinant forms – URF) (189). HIV-1 group N has very limited spread, and thus far has only been isolated from individuals in Cameroon (189). HIV-1 group O is slightly more prevalent than HIV-1 group N, but is largely contained to West Central Africa (189).

The passage of SIV from a sooty mangabey (SIVsm) to humans also likely occurred by parenteral transmission (253). This cross-species transmission gave rise to HIV-2. Transmission of SIVsm to humans has occurred at least eight times, resulting in HIV-2 groups A-H, although only groups A and B show evidence of establishing epidemics (48,61,93,94,318). The distribution and prevalence of HIV-2 is limited compared to HIV-1, with the majority of infections occurring in West Africa (189).

The two oldest archived strains of HIV-1 were isolated from samples collected in 1959 and 1960 (ZR59 and DRC60, respectively). Yet, there is large genetic divergence between the two – ZR59 appears to be a subtype D virus, while DRC60 is a subtype A virus. This suggests that the HIV-1 genetic clades were already well established,

indicating that the introduction of HIV-1 into the human population is likely to have occurred at a much earlier time point. Worobey *et al.* used BEAST (Bayesian evolutionary analysis sampling trees) to determine the date of most recent common ancestor (TMRCA) for HIV-1, and found the range to be 1873-1933 (312).

For an AIDS epidemic to become established, HIV-1 likely needed a large population center to facilitate spread. Large cities were not present in central Africa before 1900, possibly explaining why the spread did not start much sooner. The earliest known strains of HIV-1 were isolated from Léopoldville (now Kinshasa, Democratic Republic of Congo) (312,337). Léopoldville was the largest city in central Africa at the time, and was on a main transportation route from Cameroon, where the SIV-infected chimpanzees were likely butchered for bushmeat and the zoonotic transmission likely occurred (260).

A new immunodeficiency was described in four American men in 1981; these were the first recognized cases of AIDS (104). A few years later, the causative agent, HIV, was isolated and identified (19,236). Subsequent studies have revealed that this viral infection was present in the population long before it was discovered. HIV was likely circulating in the United States twelve years before its discovery and in African countries clearly much longer (98).

1.2 HIV pathogenesis

HIV-1 can be spread between hosts by exposure to body fluids; significantly, via sexual intercourse (genital secretions), shared injection drug equipment (blood) and mother to

child transmission (MTCT) (blood, genital secretions and breast milk). The initial stage of infection, termed the acute stage, lasts about two months and is characterized by high levels of virus, leading to increased infectiousness (231,232,270,331). Also during this time, the $CD4^+$ cell levels drop as the virus replicates rapidly and disseminates to the various lymphoid tissue compartments (161,196). After the acute phase of HIV infection, most individuals enter an asymptomatic period where viral levels in the blood drop and reach a set point, usually below $4.3\log_{10}$ RNA copies/mL (160). This chronic phase of infection usually lasts 3 – 10 years. During this phase, HIV replication generally occurs at low levels in the lymph nodes and other tissues, and is seemingly controlled by antiviral immune responses. However, the $CD4^+$ levels continue to slowly drop and individuals not on therapy will begin to develop signs of infection and a loss of immune functions (34). In the absence of treatment at this stage, individuals can develop opportunistic infections and HIV-related cancers such as *Pneumocystis* pneumonia and Kaposi's sarcoma. $CD4^+$ levels below 200 cells/μL and these infections, known as AIDS-defining illnesses, signal the onset of AIDS (40). Destruction of $CD4^+$ T cells are a major reason for the immune dysfunction that causes AIDS. $CD4^+$ T cells are destroyed during HIV infection by both direct mechanisms, such as by the cytopathic effects of HIV, and indirect mechanisms, such as induction of apoptosis due to immune activation (4) and destruction of thymic lymphoid tissue leading to reduced production of new cells (188).

Not all individuals are equally susceptible to HIV disease. Some people, termed elite controllers, are able to suppress viral replication to non-detectable levels, in the absence

of anti-viral treatments (195). Other individuals progress slowly to AIDS and so are called slow progressors or long-term survivors (110). Still other people progress very quickly to AIDS after infection with HIV; these are referred to as rapid progressors (138). A very special subset of individuals do not become infected with HIV at all, despite high levels of exposure; these people are referred to as high-risk exposed HIV-seronegative individuals (234). There are a number of genetic factors that have been associated with differences in HIV susceptibility and disease progression. Certain HLA (human leukocyte antigen) alleles are associated with differences in disease progression as polymorphisms can affect the strength of the cellular immune anti-HIV response. Polymorphisms that affect the expression of the HIV co-receptor, CCR5, as well as certain cytokines, such as IL-10 and RANTES (regulated upon activation, normal T-cell expressed and secreted), are also associated with altered disease progression. Intracellular host antiviral factors, such as APOBEC3G, may also be important for controlling disease progression; this will discussed in detail in a subsequent section.

There is currently no HIV vaccine available, but drugs, termed antiretrovirals (ARVs) are used to inhibit HIV-1 replication and thus keep the infection under control. Different classes of drugs target various phases of the HIV replication cycle. Nucleoside reverse transcriptase inhibitors (NRTIs) and non-nucleoside reverse transcriptase inhibitors (NNRTIs) inhibit reverse transcription, protease inhibitors (PIs) inhibit viral assembly and entry inhibitors inhibit HIV-1 entry to the host cell. These drugs, however, are not perfect – toxicity to the host and development of viral resistance to the treatment are two

major challenges of ARV treatment, as is the fact that an estimated 6.7 million people are in need of these drugs, but are not receiving them (57).

1.3 HIV/AIDS in Sub-Saharan Africa

HIV/AIDS affects people across the globe; currently, 33 million people are infected worldwide (290). However, some areas are more affected than others – nowhere is harder hit by this epidemic than Sub-Saharan Africa, where 22 million people are currently infected with HIV (290). This region houses 67% of all people living with HIV and 75% of AIDS deaths, yet only contains just over 10% of the world's population (290). There are also other ways this region is affected: over 12 million children have been orphaned by the HIV/AIDS epidemic (290). As well, nearly 90% of children younger than fifteen living with HIV, live in sub-Saharan Africa (290). Life expectancy in sub-Saharan Africa has been dramatically affected by the HIV/AIDS epidemic; in countries with high HIV prevalence, life expectancy at birth has fallen to levels not seen since the 1950s. Indeed, the life expectancy is currently below 50 years for the region (290).

Half of all individuals living with HIV worldwide are women; however, in sub-Saharan Africa, nearly 60% of all people living with HIV are women (290). There are societal/cultural reasons for this disparity, such gender inequality and a lack of empowerment for women and girls (290). Additionally, women are actually physiologically more susceptible to the acquisition of HIV. The risk of HIV transmission is 2 – 3 times higher for a male to a female partner, than a female to a male partner

(66,216). There are a number of possible reasons for this difference. Anatomically, the female genital tract has a larger surface area and a more receptive contact surface than the male genital organs (235). Additionally, the pH of semen creates a favourable environment for HIV and prolongs viral survival (299). Also semen will remain in the female genital tract until it is absorbed, thus increasing the time of exposure. In contrast, the normally low vaginal pH makes an unfavourable environment for HIV and the duration of male exposure is determined by the duration of sexual intercourse (216). Hormonal changes, such as those caused by the use of oral contraceptive, also put women at increased risk of HIV acquisition, as does increased susceptibility to other STIs (sexually transmitted infections) (155,251).

1.4 HIV-1 Structure and Replication

The HIV-1 genome consists of two strands of positive-sense RNA. Each RNA molecule is 9.2 kb long, though the HIV-1 provirus is 9.7 kb, as during reverse transcription the ends of the RNA genome are duplicated to make LTRs (long terminal repeats) on each end. The genome contains three genes shared with all retroviruses: *gag, pol* and *env*, and an additional six genes: *nef, tat, rev, vif, vpr and vpu* (Figure 2). The RNA genome forms a complex with Nucleocapsid proteins (Gag p7), and is surrounded by a shell composed of Capsid protein (Gag p24), in a mature HIV-1 virion. Reverse transcriptase and Integrase are also associated with the ribonucleoprotein complex, as may be Protease. Matrix protein (Gag p17) forms a shell outside the core and a lipid membrane surrounds the Matrix shell. Envelope protein oligomers are associated with this membrane and dominate the surface of the HIV-1 virion (Figure 2) (56).

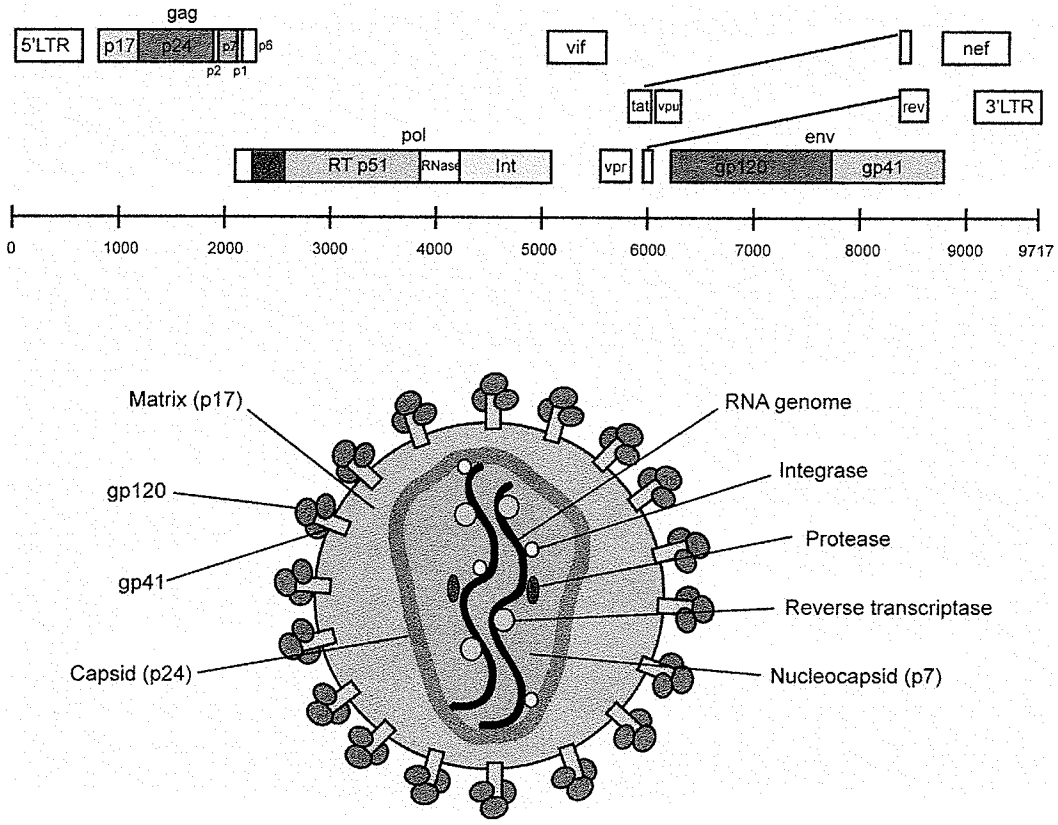


Figure 2. Representation of HIV-1 genome and structure. The top half of the figure shows the viral genome drawn to scale (nucleotide scale indicated). The genes are organized vertically by translation frame. The lower half of the figure depicts a cartoon of a mature virion. Gene products included in the cartoon are identically coloured in the genome map. The genome map is based on gene lengths publically available through the HIV Sequence Database (www.hiv.lanl.gov).

The HIV-1 replication cycle starts with the viral Envelope protein (Env) binding to a host CD4 receptor, which is expressed mainly on T lymphocytes and macrophages (30,31,114). Env is composed of a surface unit (gp120) noncovalently associated to a transmembrane unit (gp41). The gp120 unit interacts with CD4, resulting in a conformational change in the proteins that allow gp120 to interact with the primary HIV co- receptors, CCR5 or CXCR4 (31,149,314). The CD4-gp120-CCR5/CXCR4 complex stabilizes virus binding and allows for dissociation and conformational change of the gp41 unit. The gp41 N-terminus contains a fusion peptide that inserts into the cell membrane, setting off a chain of events that results in the formation of a fusion pore and cellular-viral membrane fusion (42,72,203,307).

Fusion of the viral and cellular membranes permits the viral core to be released into the host cell cytoplasm. The capsid is disassembled and the core proteins (Gag p24) are released in a process known as uncoating, which remains one of the least well understood steps of the viral life cycle (2). A reverse transcription complex is formed, which allows the reverse transcription process to initiate. The first step, RT-catalyzed synthesis of minus strand DNA, uses a host tRNA primer ($tRNA^{Lys,3}$) (14,176,183). During reverse transcription, the RNA of the RNA/DNA hybrid is degraded by the RNase H activity of RT. Two polypurine tract (PPT) sequences, one located in the middle of the genome, and thus called the central PPT, and the other located at the 3' end of the genome, and thus termed the 3' PPT, are resistant to RNase H cleavage and remain associated with the nascent DNA (70,91,102,201). Once the 5' end of the RNA genome is reached, the minus-sense single-stranded (ss) DNA transfers to the 3' end of the same, or the other co-

packaged, RNA genome molecule (281). Plus strand DNA synthesis is primed from the associated PPT sequences, and again, the plus-sense ssDNA must transfer to the 3' end of the minus strand DNA to complete viral replication. The result is double-stranded DNA with duplicated LTR ends.

The completed double-stranded (ds) DNA is transported to the host nucleus as part of the preintegration complex (PIC). The composition of the PIC is unclear, but it seems to include matrix protein (Gag p17), integrase, Vpr, and nucleocapsid (Gag p7) (89). The PIC is necessary, as HIV infects non-dividing cells, and thus the genome must be transported into the host nucleus in an active and energy-dependant manner. The precise viral factor responsible for guiding the PIC into the nucleus, remains to be identified (55).

Once the PIC enters the nucleus, viral integrase, encoded by *pol*, catalyzes the insertion of linear ds viral DNA into the host chromosome; this integrated viral DNA is referred to as provirus. Integrase binds the LTR at the each end of the viral genome and catalyzes the removal of a dinucleotide from the 3'end of each DNA strand; this is called 3' processing (67). Integrase next makes a five nucleotide staggered cleavage in the host DNA, and the 3' staggered ends of the viral DNA are joined to the host DNA by a strand transfer mechanism (67). Integration is completed by gap-filling between the viral and host DNA by cellular repair enzymes (89). It has been recently described that HIV prefers to integrate within symmetrical sequences in the host genome (106,120,313). Once the HIV-1 provirus is stably integrated, it becomes a permanent part of the host genome.

After the viral genome has been integrated as provirus, it can serve as a template for viral RNA production. Translation initiation occurs at the HIV-1 LTR, though basal translation levels are low. The viral protein Tat acts as a translational transactivator and enhances translation by more than two logs (65,84). Translation from the HIV-1 LTR generates more than thirty different types of viral RNA, which can be divided into three classes (238). The first class is composed of unspliced RNAs, which serve as mRNA for the Gag and GagPol proteins and also as genomic RNA for packaging into new virions. The second class is made up of partly spliced mRNAs of about 5 kb that encode the Env, Vif, Vpu and Vpr proteins. The final class is the small, multiply-spliced mRNAs of 1.7 – 2 kb, which encode the Rev, Tat and Nef proteins. Most cellular RNAs are fully spliced before transport out of the nucleus, whereas HIV requires some RNA to remain intact for Gag and GagPol translation and also to serve as viral genome. This is accomplished with the HIV-1 protein Rev, which binds the Rev responsive element (RRE) of *env*, present in all unspliced and partially spliced HIV RNAs, and interacts with the cellular nuclear export machinery, allowing unspliced and partially spliced viral RNA to enter the cytoplasm (89).

Viral proteins are translated from viral mRNA using the host cell's translation machinery. Once the viral proteins have been synthesized, viral assembly begins. The major factor in viral assembly is the Gag precursor polyprotein, $Pr55^{Gag}$, which is eventually processed into Matrix, Capsid, Nucleocapsid and p6 (88). Pr55 targets to and accumulates at the plasma membrane, promotes Gag-Gag interactions, encapsidates the viral RNA genome,

associates with Env glycoproteins and stimulates budding from the host cell (89). The Matrix domain of Gag is responsible for targeting and accumulates at the plasma cell membrane; binding occurs within minutes of synthesis. The Nucleocapsid domain of Gag interacts with the viral RNA packaging signal, consisting of four stem-loop structures, to initiate encapsidation of the viral genome. The Capsid domain, the p2 spacer peptide and the Nucleocapsid domain have all been shown to be important for the Gag-Gag interactions necessary for assembly (88). Both Env and Gag are targeted to lipid raft-like regions of the plasma membrane; Env to the outer leaflet and Gag to the inner leaflet. The viral core acquires Env proteins as it buds out of the plasma membrane (97,118,191). The mechanism that drives incorporation of Env glycoproteins into virus particles is not well understood, but it is hypothesized that an interaction between gp41 and the Matrix domain of Gag is important (88). Both Gag and GagPol are incorporated into nascent viral particles as polyproteins. The viral Protease cleaves these precursors, either during or shortly after virus release, to generate mature proteins. This triggers a cascade of structural rearrangements in the virion that leads to virus maturation (89).

Four HIV-1 proteins not yet mentioned also have important roles in HIV-1 replication; because they are not needed for replication in all cell lines, they are termed accessory or auxiliary proteins. The Vpu protein has two major functions, the first of which is to induce degradation of CD4 that has been bound by gp160 in the endoplasmic reticulum (ER), thus liberating gp160 for its role in viral assembly (311). Vpu also promotes viral budding and release, which recently has been determined to be due to Vpu antagonism of tetherin, a host restriction factor that acts by tethering newly formed viruses to the

infected cell's surface, thus preventing release (213,294). The Vpr protein has four known roles in HIV-1 replication, including weak stimulation of gene expression from the HIV-1 LTR and nuclear import of the PIC. Vpr arrests cells in the G_2 phase of the cell cycle (115,133,239,244), the significance of which is unclear, however G_2 arrest is thought to promote transcription from the viral LTR (99,121,338). Additionally, Vpr is believed to promote apoptosis by the intrinsic signalling pathway (210), but it is not clear if this activity is independent of Vpr's induction of cell cycle arrest (6). The Vif protein is important for counteracting the host restriction factors APOBEC3F and APOBEC3G, which will be discussed in greater detail in the following subsections. The Nef protein has four major functions. Like Env and Vpu, it is involved in downregulating CD4 on host cells, by causing the viral receptor to be internalized and subsequently degraded (95). Nef also downregulates cell surface expression of MHC-I (major histocompatibility) molecules, facilitating viral immune evasion (259). Nef mediates cell signalling and activation, which may enhance viral replication, especially in memory T cells (11,241,266). Finally, Nef enhances viral infectivity by an unknown CD4 independent mechanism (54,198).

In vivo, the average generation time for HIV-1 is 2.6 days, and each day an estimated 10.6×10^9 virions are produced in a chronically-infected patient (226). However, only about 1 in 60,000 virus particles are infectious (143,229). HIV-1 replication is thus a complex process, which is still not completely understood.

1.5 HIV Genetic Variation

There are a number of reasons for the high level of genetic diversity in HIV. One reason is that HIV is an RNA virus and RNA viruses have generally higher mutation rates than organisms with DNA genomes. The mutation rate is 0.76 nucleotide errors per genome per replication for riboviruses (RNA viruses excluding retroviruses), while the rate for DNA-based microbes is 0.0034 nucleotide errors per genome per replication (75,76). The mutation rate for HIV-1 is close to 1 nucleotide error per genome per replication (218). This difference in mutation rates is largely due to a lack of proof-reading ability in RNA viruses' RNA polymerase, the enzyme responsible for copying the viral RNA genome.

The RNA polymerase of retroviruses such as HIV is called reverse transcriptase (RT). RT is distinctive as it uses an RNA template (the viral genome) to transcribe a DNA copy; it is thus an RNA-dependent DNA polymerase. In HIV-1, RT is encoded by the *pol* gene and is composed of a p66/51 heterodimer (144). The high mutation rate of one nucleotide error per genome replication, in combination with the rapid viral replication rate, which is estimated at 1×10^{10} virions per day, means that in the roughly 10 kb HIV-1 genome, one error is incorporated at each position in an infected individual, per day (226). Of course, not every incorporated mutation yields an infectious virus, which in part contributes to estimates of infectious unit/particle ratios of as low as 1 in 60,000 (143,229). Nevertheless, the error-prone RT is an important source for HIV diversity.

RT does not just contribute to HIV's genetic diversity by causing point mutations, but also by causing genetic recombination. Each infectious viral particle is packaged with two strands of genomic RNA. RT is responsible for converting the RNA genome into DNA, which is then incorporated into the host genome. This process involves two obligatory strand transfer events when RT reaches the 5' ends of the plus and minus strand templates. The 5' end of the template can transfer to the 3' end of either the same, or the other co-packaged, RNA genome molecule and thus the transfer can cause either inter- or intramolecular switches (293,325). Additionally, RT can switch templates during replication of internal regions. Rates of recombination are estimated to range between three and thirty crossover events per generation of viral DNA (45,128,159,222,339). If the two packaged viral RNA strands originated from two different viruses, as may occur if the host cell was infected with multiple viruses, the resulting viral DNA will be recombinant. This recombination is an important source of viral diversity and will be discussed further in introduction subsection 1.6.

APOBEC3 proteins, especially APOBEC3G and APOBEC3F, are also important contributors to HIV diversity. These proteins are part of the host innate immune response and target HIV as well as some other RNA viruses and retroelements (subclass of genetic transposon that self-amplifies via an RNA intermediate). They bind single-stranded RNA viral genome, and act on the newly synthesized single-stranded DNA (24,113,124,163,332,336). APOBEC3G and APOBEC3F are cytidine deaminases, removing the amine group from cytidine in the nascent DNA strand, resulting in a uridine base. Uridine forms base pairs with adenine, causing an adenine to be incorporated

instead of a guanine when the second, positive-sense DNA strand of the proto-provirus is copied from the first, negative-sense DNA strand. Thus, the hallmark of APOBEC3F/G activity is guanine to adenine hypermutation. This hypermutation is believed to play an important role in HIV-1 diversity and may explain the general enrichment for adenosine in the HIV-1 genome (22,268).

The host adaptive immune response influences viral diversity by leading to the dynamic process of immune escape: randomly generated mutations, caused by the previously described mechanisms, can allow for decreased immune recognition and thus are positively selected (i.e. are retained in the population, as they aid the virus to evade the host immune system). Antibodies are generated by B cells and released into host plasma; thus they target extracellular antigens. Antibodies can help combat HIV in three main ways: they can neutralize the virus by binding to the viral components responsible for host cell entry (gp120 and gp41), thus hindering infection; antibodies can coat the virus and stimulate destruction by effector cells such as macrophages; and antibodies can bind the virus and trigger the complement system, which will also clear the virus. Escape mutations to antibodies occur in the exposed proteins of HIV-1, such the viral envelope protein gp120, and are a major driver of patient and global envelope diversity (305). HIV-1, like other viruses, is an intracellular pathogen and thus is also combated by cytotoxic T lymphocyte (CTL) adaptive immune responses. During HIV replication, short segments of viral protein, called epitopes, are presented on the infected cell's surface by human leukocyte antigen (HLA) class I molecules (190). Randomly generated viral mutations can be selected to decrease antigen presentation on the HLA-I molecule

and/or CTL recognition of the epitope/HLA-I complex, leading to CTL escape (105). CTL escape is seen across the viral genome, as epitopes are generated against the entire viral proteome. CTL responses are thought to better contribute to immune control than antibodies, making them an important genome-wide influence against viral evolution (37,157).

1.6 The Importance of HIV-1 Clades and Recombination

HIV-1 has become increasingly diverse since its introduction into the human population an estimated hundred years ago. Individual strains of group M virus, for example, may differ by 35 – 40% (96). This diversity has implications for host response and adaptation to the virus, as well as vaccine and therapeutic design. As previously mentioned, the different clades seem to have developed due to founder effects. For example, most infections in North America and Europe are caused by clade B HIV-1 (125,189). Subtype B HIV-1 likely migrated from Africa to Haiti in the 1960s, where it established a localized epidemic (98). The pandemic clade B that is found in North America and Europe emerged from a single migration of the virus out of Haiti a few years later (98). Most HIV infections in India are caused by clade C virus, while in some regions, a recombinant virus predominates: in West and West-central Africa, CRF02_AG is the most prevalent form of HIV-1 (189). Again, this predominance is likely due to the introduction of clade C and recombinant AG virus, respectively, to these areas.

Although the clades likely arose due to population founder effects, there are significant differences between them. Clades are phylogenetically classified based on *env*, which

may genetically differ by 20 – 50%. The *pol* region, however, is much less divergent, as it encodes two enzymes, RT and protease, which are critical for viral replication. However, small differences in this region are still significant, as many ARVs target these gene products. For example, HIV-1 group O and HIV-2 are resistant to all NNRTIs (68,69,280). Additionally, some studies have indicated that clade C virus develops resistance to NNRTIs more quickly than clade B viruses (167,168). The LTR regions, which contain transcriptional promoters of HIV replication, also differ significantly between the clades (127,202,329). For example, the copy number of NF-κB binding sites varies between the different clades, and thus the response to this transcriptional factor also differs (127). Co-receptor usage also differs between the clades. Clade B HIV-1 uses the CCR5 co-receptor during the early stages of disease, and switches to the CXCR4 co-receptor during the end stages (60,138,217). Clades A and C viruses, however, favour CCR5 use throughout disease, while subtype D virus uses both CCR5 and CXCR4 throughout the infection (1,47,224,285,333).

The effect of HIV clade on disease progression is less well understood than the genetic differences between the clades, though some studies have found an *in vivo* significance to HIV-1 subtype. This evidence, however, is not always consistent. For example, follow-up of individuals infected by HIV-1 envelope subtypes A and D revealed that subtype D was associated with lower $CD4^+$ count and increased disease progression (134). Another study found that subjects infected with clade C virus had higher plasma viral loads, lower $CD4^+$ counts and faster progression to AIDS than individuals infected with clade A or D virus (214). A large study compared progression rates between individuals infected with

clade A, C, D and recombinant virus, and found that subjects infected with subtype D virus had the fastest disease progression (297). Clade has also been associated with transmission; a study of HIV-positive women in an antenatal clinic in Kenya found increased MTCT (mother to child transmission) with subtype D compared to subtype A (321), yet a similar study in Uganda did not find a significant difference (78). It has also been observed that increased pathogenicity of a subtype does not necessarily indicate greater prevalence. Clade C accounts for over 50% of global HIV-1 infections, yet, pairwise viral fitness competitions revealed that clade C is in fact the least replicatively fit of the group M viruses by 100-fold (9). The transmission of clade C virus is just as effective as other group M viruses, however, and indeed has a higher viremia set point and higher virus levels in genital fluids (131,240,301). This suggests that although clade C virus is as efficiently transmitted, it is less virulent than other group M viruses. Indeed, decreased virulence could lead to longer asymptomatic and transmission periods, suggesting a mechanism for the increased prevalence of clade C (10).

Recombinant subtypes also play a significant role in the HIV-1 pandemic. In areas where multiple clades co-circulate, such as East Africa, nearly 30% of viral strains are URF (unique recombinant forms) (189). Additionally, some CRF have become widespread: CRF02_AG in Western Africa, as mentioned above, CFR01_AE in Southeast Asia, CRF07_BC and CRF08_BC among intravenous drug users (IDU) in China and CFR12_BF in Argentina (189,283). In some regions a CRF has replaced a non-recombinant clade as the dominant HIV form in a population. This is true amongst the IDU of Thailand; in the late 1980's, clade B was predominant, but by the mid-1990's,

CRF01_AE become the most prevalent form (135). This suggests that recombination may allow a virus to quickly attain advantage under specific circumstances, such as by increasing pathogenicity or transmission. In addition to intersubtype recombination, intrasubtype recombination, occurring between different quasi-species of the same viral clade, has also been shown to be important for generating and maintaining viral diversity (43). Certainly, recombination is a rapid and efficient mechanism for viral evolution.

In order for recombination to occur, a cell must be co-infected by two different HIV-1 strains. This may occur if a patient is infected with multiple strains, which is defined as coinfection if a patient is infected with a second strain of HIV-1 at the same time or shortly after the first, or defined as superinfection if a patient is infected with a second strain of HIV-1 after the first strain has already established infection (132). Superinfection has been observed in IDU and commercial sex workers (CSW), both of which groups are typically highly exposed to multiple strains of HIV (81,122,286). Coinfection and superinfection are generally difficult to detect and distinguish in cross-sectional studies. The identification of recombination, however, suggests that coinfection or superinfection has previously occurred. Recent studies have shown that people highly exposed to HIV are at a greater risk of dual infection and recombination due to their increased exposure (12,13,117).

Recombination in HIV-1 can be difficult to detect. Although putative hotspots for recombination have been identified, recombination does not confine itself to these regions (17,185,199). Thus, the only unambiguous way to determine recombination is to

perform full-length HIV-1 genomic sequencing, allowing the entire length of genome to be examined. For example, studies of clade distribution in Kenya based on partial genome sequencing have estimated the frequency of recombination to be 0 – 25% (164,209,214,237,269,322). In contrast, a survey that examined full-length HIV-1 genomic sequences found that 40% of the sequences were recombinant (74).

1.7 HIV Diversity in Kenya

Kenya is located in Sub-Saharan Africa, in a region of high HIV prevalence and where multiple clades co-circulate. A significant number of molecular epidemiology studies have been published describing the distribution of HIV clades and recombination in this region. These studies were consistent in finding that clade A is the most prevalent form of HIV-1 in Kenya. One of the earliest published studies examined a portion of *env* from seventeen patients and found that 71% of the sequences were clade A and 29% were clade D (237). A larger study performed in rural Western Kenya examined sequence from the C2-V3 region of *env* in thirty patients and also found that the majority belonged to clade A1 (67%), while only two sequences were clade D, one was clade C and one clade G (269). Partial *env* sequence performed on HIV-1 isolated from 320 women from Nairobi revealed that 70% of the sequences were clade A, 20% were clade D, 7% were clade C and a single clade G sequence was discovered (214). A study in Kisumu, in Western Kenya, looked at both *env* gp41 and a *gag-pol* segment for 460 individuals. The authors found that 75% of the sequences had the same clade in both genomic regions; 59% of all sequences were clade A1 in both regions, 10% clade D, 3% C and 2% G (322). An epidemiological study of the mother to child transmission cohort examined in

this study used RFLP (restriction fragment length polymorphism) analysis of *protease* and p24 to determine clade and found that the majority of the 130 mothers were infected with clade A (58%), while 20% were clade D and 2% clade C (209). The most recent study published from Kenya examined *integrase* from 140 HIV positive sexually transmitted infection patients in Nairobi, and found 64% of sequences were clade A1, 17% D, 9% C and 1% G (164). The only previous survey from Kenya that generated full-length sequence data examined forty-one HIV-1 sequences collected from patients across Southern Kenya (74). The authors reported that 56% of the sequences were non-recombinant clade A, but found only a single non-recombinant clade C and a single non-recombinant clade D sequence. Thus, to date, all Kenyan epidemiological surveys found that more than half of the HIV-1 sequences were clade A1, with a minority of clade D and clade C, and only occasionally identified clade G.

The studies summarized above differed in the amount of recombination that was reported. Those that examined a single HIV-1 genomic region either did not identify recombination (237), or found low levels (164,209,214), to a maximum of 17% recombinant viruses (269). The study that examined two genomic regions identified a higher proportion of recombinants: 25% (322). The study that examined forty-one full-length HIV-1 sequences found that nearly 40% of the sequences were recombinant, most of which were URF, generating the highest estimation of recombination prevalence in Kenya (74). Thus the identification of recombination is facilitated by examining multiple regions of HIV-1 genome and full-length sequencing is likely the most accurate method.

1.8 Discovery of APOBEC3G and the APOBEC Family

Prior to the discovery of the restriction function of APOBEC3G, scientists suspected the existence of a cellular restriction factor counteracted by the HIV-1 protein Vif. Vif mutation caused defects in viral productivity: Vif-deficient viruses replicate, but the resulting progeny viruses are not replicatively competent. However, it was noted that this observation was cell line specific; in some cell lines, such as HeLa, COS, 293T, SupT1, CEM-SS and Jurkat, Vif-deficient HIV-1 were able to replicate normally over multiple passages, thus these cell lines were "permissive" for Vif mutants (89). The major cellular targets of HIV-1 replication, primary lymphocytes and macrophages, as well as a few cell lines, such as HUT78 and CEM, cannot support the replication of Vif defective viruses; thus they are "non-permissive" (262). The construction of heterokaryons formed between non-permissive and permissive cells revealed that the non-permissive phenotype was dominant, suggesting the presence of an anti-viral factor expressed by the non-permissive cells that is suppressed by Vif (175,267).

To identify the candidate gene for this anti-viral factor, Sheehy *et al.* employed a complementary DNA subtraction strategy using the pair of genetically related cell lines, CEM-SS and CEM, which display the permissive and non-permissive phenotype, respectively (262). The authors identified cDNA belonging to a gene they named *CEM15*, which was present in all the non-permissive cell lines tested, but not in any of the permissive cell lines. They verified the anti-viral function of CEM15 by transducing the cDNA into CEM-SS cells and found that CEM15 was capable of converting permissive cells to the non-permissive phenotype (262). Sheehy *et al.* noted sequence

homology between CEM15 and the RNA-editing protein APOBEC1 (262). A later study by Kao *et al.* confirmed that CEM15 was the previously identified and characterized protein, APOBEC3G, which was known to have cytidine deaminase activity, but not previously known to have a role in host restriction of viral replication (113,136).

The first member of the APOBEC family to be identified was APOBEC1 (**apo**lipoprotein **B** mRNA-**e**diting enzyme, **c**atalytic polypeptide-**1**), which mediates cytidine to uridine editing of apolipoprotein B mRNA and is located on chromosome 12 (282). APOBEC2 is located on chromosome 6 and has an unknown physiological function (162,186). Jarmuz *et al.* identified an additional APOBEC loci, the APOBEC3 cluster found on chromosome 22, which contains genes and pseudogenes APOBEC3A to 3G (126). APOBEC3H, which is located downstream of APOBEC3G, was not identified as a member of the APOBEC3 family until the human genome project was complete (59,304). APOBEC4, the last member to be identified, is located on chromosome 12 (245,304). Activation-induced deaminase (AID) is also considered to be a member of this family cytidine deaminases; AID uses cytidine to uridine deamination to initiate immunoglobulin gene diversification (207,211).

This family of cytidine deaminases is characterized by a zinc-binding deaminase motif, with the consensus active site of His-X-Glu-X_{23-28}-Pro-Cys-X_{2-4}-Cys (where X is any amino acid) (126). AID, APOBEC1, APOBEC2 and APOBEC4 all have a single cytidine deaminase domain. Some of the APOBEC3 members, however, are double deaminase domain proteins. These include APOBEC3B, APOBEC3DE, APOBEC3F

and APOBEC3G. It is interesting to note that the proteins most efficient at restricting HIV-1 (i.e. APOBEC3B, APOBEC3F and APOBEC3G) have two deaminase domains (101).

APOBEC3G was the first APOBEC3 protein to be discovered as having antiretroviral activity, specifically against HIV-1. APOBEC3A does not have anti-HIV activity, though it blocks replication of adeno-associated virus (AAV) and retrotransposons (26,27,44,100,205). APOBEC3B is a strong inhibitor of HIV-1, causing GA to AA hypermutation, but unlike APOBEC3G, it is not expressed in $CD4^+$ T cells or macrophages (24,126). It is also not neutralized by the HIV-1 Vif protein (24,71,326). APOBEC3C does not inhibit HIV-1 replication, but it is capable of causing hypermutation in the virus, preferring a GA over a GG context (113,154). Furthermore, APOBEC3C is expressed in PBMCs (peripheral blood mononuclear cells), macrophages and thymocytes and like APOBEC3B, is resistant to Vif (126,154). APOBEC3D and APOBEC3E combine to form a single protein, APOBEC3DE (59). There is some controversy over the antiretroviral activity of APOBEC3DE, with Stenglein *et al.* not finding any, while Dang *et al.* determined that APOBEC3DE did weakly block HIV-1 replication, cause hypermutation in a GA, GG and novel GC context, and was blocked by Vif (62,271). The publication by Dang *et al.* noted that the HIV-1 inhibition was dose dependent; the two groups used different systems to express APOBEC3DE, and so the discrepancy may be due to differences in the amount of APOBEC3DE present in the two systems (62,271). APOBEC3F and APOBEC3G are the most potent inhibitors of HIV-1 replication and will be discussed in further detail in the next section. APOBEC3H

generally has only poor antiretroviral activity; however, certain single nucleotide polymorphisms (SNPs) improve the antiretroviral and hypermutating activity of this Vif-resistant protein (62,109,221). Interestingly, these SNPs are prevalent in certain African populations (109). Hypermutation by APOBEC3H occurs preferentially in a GA context (109).

In addition to some of the APOBEC3 family members having antiretroviral activity, many have other physiological functions as well. Indeed, data suggests that the APOBEC3 family evolved to protect the genome from endogenous retrotransposons (255). As described, APOBEC3A is a strong suppressor of AAV and retrotransposons, such as LINE-1 and Alu (26,27,44,100,205). APOBEC3B, 3C and 3F also inhibit retrotransposition by LINE-1 (205). Furthermore, the antiviral activity of the APOBEC3 proteins is not limited to retroviruses and retrotransposons. APOBEC3A, 3C and 3H show evidence of causing hypermutation in human papillomavirus (HPV) (296). Many of the APOBEC3 proteins also have activity against hepatitis B virus (HBV). APOBEC3B, 3C, 3F and 3G cause HBV hypermutation, while APOBEC3B, 3F and 3G also inhibit HBV reverse transcription (20,28,215,248,276,287). Furthermore, APOBEC3B inhibits HBV gene transcription (335).

The APOBEC family of proteins are found in all vertebrates, but double deaminase domain proteins are only found in mammals (8,113). The APOBEC3 proteins show a relatively recent expansion in mammals, as rodents have only a single APOBEC3 gene, while humans and chimpanzees encode for proteins APOBEC3A-H (126,304). Further

differences can be observed within the different primates. For example, APOBEC3H is a potent restriction factor in old world monkeys such as rhesus macaques, but in humans, the activity of this protein is largely lost (221). Additionally, APOBEC3A is not found in rhesus macaques (221). These observations suggest that the APOBEC3 family of proteins have undergone recent evolutionary pressures.

1.9 Antiretroviral functions of APOBEC3G and APOBEC3F

APOBEC3G is a 384 amino acid, ~46 kDa cytoplasmic protein; immunocytochemistry shows both diffuse cytoplasmic staining of APOBEC3G, as well as accumulation in P bodies (cellular structures involved in mRNA processing) (92,146,308,309). The APOBEC3G gene contains 8 exons and is ~10 kb long (126). APOBEC3F is a 373 amino acid, ~45 kDa protein, with similar cellular localization to APOBEC3G (309). The gene contains 7 exons (it is missing the last exon of APOBEC3G) and is ~12 kb (126). In the APOBEC3F and 3G genes, exons 5, 6 and 7 are duplications of 2, 3 and 4, respectively (126). The amino-terminal cytidine deaminase domain extends from residues 65 – 104 in APOBEC3G, while the carboxy terminal cytidine deaminase domain, which mediates cytidine deamination, spans residues 257 – 291. The crystal structure of a portion of APOBEC3G has recently been solved, showing a core β-sheet composed of five β -strands, surrounded by six α-helices (119). The structure shows the zinc atom of the cytidine deaminase active site coordinated by three residues: H257, C288 and C291 (119).

Residues outside of the active site have been shown to be important for APOBEC3G activity. A single amino acid difference between human APOBEC3G (hAPOBEC3G) and African green monkey (AGM) APOBEC3G confers species-specificity to interaction with the host immunodeficiency virus (i.e. HIV-1 vs. SIV_{AGM}, respectively) (25,179,257,317). The human protein has an aspartic acid at position 128 (D128), whereas the AGM protein has a lysine at this position. Additionally, a four amino acid region N-terminal to position 128, most importantly a tyrosine (Y124) and tryptophan (W127), is needed for virion packaging of the innate protein, and thus is also important for antiviral function (123). Site-directed mutagenesis of the hAPOBEC3G gene revealed that a three amino acid motif from positions 128 – 130 consisting of aspartic acid – proline – aspartic acid is necessary for interaction with HIV-1 Vif (123). Amino acid residues 54 to 124 of APOBEC3G are sufficient for coimmunoprecipitation with HIV-1 Vif, suggesting that there may in fact be multiple sites of protein interaction (58). Zhang *et al.* mapped the region of APOBEC3G required for its degradation by Vif to amino acids 105 – 245; the authors note that while residues 105 – 156 are required for Vif interaction, they are not sufficient for degradation, which needs the residues 157 – 245 as well (334). This region of APOBEC3G maps to the linker region between the two cytidine deaminase motifs.

In a cell, APOBEC3G resides in either a high molecular mass (HMM) complex, or a low molecular mass (LMM) complex. The HMM complex can be larger than 700 kDa and is found in activated $CD4^+$ T cells; however, the APOBEC3G in these complexes is inactive. In contrast, the LMM complex (~100 kDa, which corresponds to dimer

formation) is found in unstimulated $CD4^+$ T cells and monocytes, and contains active APOBEC3G that can block HIV-1 replication. The HMM complex can be converted to an LMM form with the treatment of RNase (51). The HMM complex is not well defined, containing 95 unique proteins, such as the ribonuclear proteins Staufen and Ro, and multiple RNAs, including retroelement Alu and hY RNAs (52). APOBEC3F also forms LMM complexes, corresponding to dimer formation, as well as HMM complexes; however, these HMM complexes are resistant to RNase treatment (303). Interestingly, APOBEC3F and APOBEC3G have been shown to form heterodimers (310).

APOBEC3G expression can be affected by mitogens and cytokines. Activation of T cells and dendritic cells by stimulation of CCR5 and CD40 with the ligands CCL3 and CD40L, respectively, increases APOBEC3G expression, as does treatment with HSP70 (230). Treatment of peripheral blood lymphocytes (PBLs) with phytohemagglutiin (PHA) and IL-2 also increases expression of the innate protein (272). Treatment of the H9 T cell line with phorbol myristate acetate (PMA) induces expression of APOBEC3G mRNA and protein (246). However, induction does not necessarily mean increased restriction by APOBEC3G, for mitogenic activation shifts APOBEC3G from its active LMM form into inactive HMM complexes (52). Certain cytokines, such as IL-2, IL-7 and IL-15 also activate APOBEC3G gene expression in PBLs and recruit LMM APOBEC3G into HMM complexes (273). In contrast, IFN-α has also been shown to increase expression of APOBEC3G in resting T cells and dendritic cells, but this expression is associated with enhancement of APOBEC3G in the LMM form and an inhibitory effect on HIV-1 replication (46,51,302). The APOBEC3G promoter itself is

not inducible by mitogenic stimulation, but is activated by the ubiquitous transcription factors Sp1 and Sp3 (206). The level of APOBEC3G protein within the cell is thus controlled at multiple levels.

The primary antiretroviral activity of APOBEC3F and 3G to be identified was guanine to adenine hypermutation. APOBEC3G causes GG to AG mutations, while APOBEC3F causes GA to AA mutations, in both instances where the third nucleotide is not a C (21,141,154,163,326). Hypermutation can cause detrimental viral mutations, such as the generation of premature stop codons, especially at tryptophan residues, where the transition of a G to an A causes the Trp codon (UGG) to change to a stop codon (UAG, UAA or UGA) (50,300). Additionally, hypermutation can cause mutations that lead to other amino acid substitutions – indeed, it is believed that APOBEC3-mediated hypermutation is an important driving force of HIV-1 viral diversity (233).

In the absence of the HIV-1 protein Vif, cytoplasmic host APOBEC3G can be packaged into newly formed viral particles (111,178,181,262). Xu *et al.* determined that approximately seven APOBEC3G molecules are packaged in each virion (316). The mechanism that mediates this packaging, however, remains unclear (275). Some authors propose that APOBEC3G is packaged through an interaction with Gag in an RNA-independent manner (3,41,73,171). Others propose interaction with host RNA, of which 7SL RNA (a non-coding RNA that is part of the signal recognition particle) is a favorite candidate (35,256,278,330). Also proposed is interaction with viral genomic RNA (15,139,140). Regardless of how APOBEC3G is packaged, the inclusion of this innate

protein in the viral particle allows it access to the RNA genome as it is reverse transcribed into DNA prior to proviral integration. APOBEC3G binds the ssRNA genome and acts on the ssDNA copy, deaminating cytosine to uracil (113,124,332). This results in guanine to adenine hypermutation in the corresponding positive DNA strand. This evidence of proviral hypermutation becomes archived in the infected cells, and can be examined for evidence of the antiviral activity, as has been described by Kijak *et al* (142). Hypermutation in the viral RNA genome, however, is rarely found (141,249). The Vif protein of wildtype HIV-1 counteracts APOBEC3G activity by hijacking the cellular Cullin5-ElonginB-ElonginC E3 ubiquitin ligase to target it for proteasomal degradation, and by inhibiting APOBEC3G incorporation into newly formed viral particles (58,165,166,182,192,263,272,327). APOBEC3F acts in a similar mechanism and is likewise inhibited by Vif (24,163,310,336).

Interestingly, the level of hypermutation is not uniform across the proviral genome. Rather, two gradients can be observed, with local maxima of hypermutation levels occurring 5' of the central PPT and the 3' PPT (142,277). As described in section 1.5, this corresponds to the priming sites for second strand synthesis of the viral DNA, thus higher levels of hypermutation are observed where the minus strand DNA remains single stranded for a longer period of time (277).

The *in vivo* anti-viral mechanism of APOBEC3F/G has not been completely defined. Certainly, guanine to adenine hypermutation can cause detrimental mutations in the viral genome, reducing or abolishing the production of viable progeny, but other antiviral

APOBEC3 functions have also been described. APOBEC3-mediated hypermutation may trigger degradation of the nascent viral DNA by host uracil glycosylases and apurininc/apyrimidinic endonucleases, thereby inhibiting the production of provirus (319). APOBEC3G has also been suggested to interfere with the removal of primer tRNA, as well as to inhibit DNA strand transfer and integration (172,187). Some anti-viral effects have been described for APOBEC3G in the absence of hypermutation. Guo *et al.* reported reduced reverse transcription priming and reduced levels of viral DNA in Vif-negative virus, in the absence of APOBEC3G deamination and hypermutation (107). APOBEC3G was also found to interfere with proviral integration in the absence of hypermutation (172). However, a recent publication stated that deaminase-defective APOBEC3G had antiviral activity only when expressed at high levels, questioning the physiological relevance of deaminase-independent activities of APOBEC3F/3G (200).

The role of APOBEC3F/G in HIV disease progression and pathogenesis is similarly unclear, although a relationship between the two has been suggested (23,53,129,130,177,223). The APOBEC3G mRNA expression level has been associated with decreased viremia and long term non-progression (129) and with HIV-exposed seronegative individuals (23). However, these studies have not been readily replicated (53), resulting in significant controversy over the role of APOBEC3 in HIV disease progression *in vivo*.

To date, 134 APOBEC3G SNPs have been described, of which nine are in coding regions (www.ncbi.nlm.nih.gov/SNP). A minority of SNPs have been associated with

HIV/AIDS disease. Valcke *et al* described a C40693T SNP, located in intron four, that was associated with an increased risk of HIV-1 infection (291). This SNP is located close to an exon/intron boundary, thus the authors suggest alternative splicing as a mechanism for altering APOBEC3G to result in increased susceptibility to infection. An *et al* described a SNP that causes a codon change in exon four: H186R (5). The variant 186R allele was associated with a decline in $CD4^+$ T cells and accelerated progression to AIDS (5). The polymorphism is located in a region required for degradation by Vif, leading to the hypothesis that this SNP may lead to decreased APOBEC3G degradation.

1.10 HIV-1 Vif

HIV-1 Vif is a 23 kDa cytoplasmic protein that counteracts the antiviral functions of host APOBEC3F and 3G proteins by hijacking the cellular Cullin5-ElonginB-ElonginC E3 ubiquitin ligase to target the proteins for proteasomal degradation, and by inhibiting their incorporation into newly formed viral particles (58,165,166,182,192,263,272,327). Vif proteins from different HIV-1 strains are highly conserved, indicating the importance of this viral protein (219). The N-terminus (residues 1 – 21) contains a tryptophan-rich stretch that is conserved; some of these tryptophan residues are required for Vif recognition of APOBEC3F and 3G (284). Residues 63 – 70 and 86 – 89 are also well conserved and are predicted to be important for β-strand formation (90). A charged central hydrophilic region ^{88}EWRKKR^{93} is thought to enhance steady-state expression of Vif; within this region, the glutamic acid at position 88 and tryptophan at position 89 are conserved among HIV-1 strains (90). Two pairs of conserved histidine/cystine residues (H108, C114, C133 and H139) flank an α-helix and coordinate a Zn^{2+} ion, making up the

HCCH motif which directly binds Cullin5 (193,315). The BC-box motif is responsible for binding ElonginC and is defined by the highly conserved ^{144}SLQ(Y/F)LA^{149} motif (328). Four major phosphorylation sites have been identified in Vif: T96, S144, T155 and T188 (323,324). The conserved proline-rich ^{161}PPLP^{164} domain is important for Vif multimerization, which in turn is important for viral infectivity and preventing APOBEC3G incorporation into viral particles (197,320).

The binding of Vif to APOBEC3F/3G is essential for HIV-1 to overcome this host defence protein. The main binding determinants are in the N-terminal region of Vif (58,182,194,250,252,258,268,284,308). However, amino acids 169-192 of the C-terminal region also mediate APOBEC3G interaction (252). Other residues and motifs important for APOBEC3G binding include residues 85 – 95, ^{40}YRHHY^{44} (250,258), the tryptophan residues W5, W21, W38 and W89 (284), I9 (308), K22, E45 and N48 (250,268). Important sites for APOBEC3F binding include the tryptophan residues W11 and W79 (284), and Q12 (250,268). This data suggests that the APOBEC3F and 3G binding sites are non-linear, relying on electrostatic charge and conformation (18).

1.11 ML and MCH Cohorts

The work described in this thesis is based on patient samples obtained from two cohorts located in Nairobi, Kenya. The ML cohort was established in 1985 and is composed of women actively involved in commercial sex work (86,87). This cohort is well characterized and studied, and has allowed for important findings, including the identification of individuals resistant to HIV infection, despite high levels of exposure

(234). The MCH (mother child health) cohort is located in the same area as the ML clinic, but these female attendees are not involved in commercial sex work and are at lower risk of HIV acquisition (32,77,174). This cohort has also been well characterized and studied, and has contributed to understanding the transmission of HIV from mother to infant (64).

Hypothesis and Specific Objectives

Rationale: Previous studies have shown that individuals at high risk of acquiring HIV, such as sex workers, are more likely to be infected with recombinant virus, if they perform their high risk activities in an area where multiple subtypes circulate (12,13,117). The high prevalence of HIV infection in Kenya and co-circulation of multiple subtypes makes this a likely place to find recombinant HIV. The advantage of full-length HIV sequence analysis for recombination detection is that one can determine the presence of breakpoints across the entire length of the genome, and thus detect all or most breakpoints. However, the advantage of partial genome sequencing is that a smaller sequence amplicon allows one to expand the study to include more subjects. Proviral HIV sequence provides a wealth of information about the infecting virus. As well as being used for determining viral subtype and discerning and characterizing recombination, proviral HIV also provides an archive of hypermutation activity (142). Hypermutation, as caused by the various APOBEC3 proteins, has been recently described as an important mechanism of viral restriction and its role in disease progression is currently being debated in the literature. Previous studies have indicated that increased levels of proviral hypermutation are caused by deleterious mutations in Vif, which make it unable to counteract APOBEC3F/G (223). However, the major factors controlling the APOBEC/Vif balance are still to be determined.

Based on the above rationale, we developed the following hypotheses:

- We hypothesize that HIV-1 proviral sequence analysis will reveal a higher proportion of recombination in the high-risk ML cohort than what has been previously described in the general Kenyan population. We furthermore hypothesize that within this highly-exposed population, increased duration of commercial sex work will be correlated with increased likelihood of infection with recombinant HIV, as individuals who practice sex work for a longer time will have more chances for exposure to different viral subtypes.
    - To address this hypothesis, we will:
        - Generate full-length HIV-1 proviral sequences from ten subjects from the ML cohort, determine the subtype of the HIV-1 sequences and characterize any recombination
        - Generate partial HIV-1 genomic sequences from 240 individuals from the ML and MCH cohorts, determine the viral subtype and characterize any recombination
        - Correlate HIV-1 recombination with participation in and duration of commercial sex work, and any other pertinent epidemiological factors

- We hypothesize that hypermutation will be present in the HIV-1 proviral sequences obtained from members of the ML and MCH cohorts, and that as APOBEC-mediated hypermutation is a viral restriction mechanism, increased hypermutation will be associated with decreased disease progression, as measured by $CD4^+$ count and viral load. We further hypothesize that the viral Vif protein from the hypermutated viruses will not be significantly different than the Vif from non-hypermutated viruses.
    - To address this hypothesis, we will:
        - Identify APOBEC-mediated hypermutation in proviral HIV-1 sequences
        - Determine if hypermutation associates with $CD4^+$ count and viral load
        - Sequence the HIV-1 *vif* from hypermutated and non-hypermutated proviruses and compare the gene and protein sequences

- We hypothesize that host factors are more important than viral factors in controlling hypermutation, and thus in individuals superinfected with HIV-1, the level of proviral hypermutation will be constant despite superinfection.
    - To address this hypothesis, we will:
        - Identify a prospective cohort of individuals who became infected and subsequently superinfected with HIV and for which there is longitudinal HIV proviral sequence data; determine levels of hypermutation in these sequences
        - Compare hypermutation levels in the original virus before and after the superinfection event and with the superinfecting virus

- We hypothesize that subjects infected with significantly hypermutated HIV-1 provirus have polymorphisms in the APOBEC3G gene that make this host defence more active against HIV-1, possibly by increasing the protein's deamination activity or by decreasing its susceptibility to Vif-mediated degradation. We expect these mutations to be present in all or most of the subjects with significantly hypermutated HIV-1 provirus, but in none or few of the subjects with non-significantly hypermutated HIV-1 provirus.
    - To address this hypothesis, we will:
        - Amplify and sequence the APOBEC3G gene and surrounding region for subjects with differing levels of HIV-1 proviral hypermutation
        - Identify polymorphisms in the APOBEC3G genomic region and compare their allelic frequency between the differentially hypermutated populations

These specific hypotheses will help us to address the overall hypothesis of this thesis; that HIV-1 sequence diversity is an important factor of HIV-1 pathogenesis.

Materials and Methods

1. Solutions

Cell Culture Media for PBMCs:

RPMI (Roswell Park Memorial Institute medium) – pH 7.2

10% fetal-calf serum (FCS)(heat inactivated at 56°C for 30 minutes)

2% penicillin/streptomycin

All reagents obtained from Invitrogen, Burlington, Ontario

Phosphate-Buffered Saline (PBS):

48.5 grams PBS per 1 litre dd H_2O (final concentration: 137.93 mM NaCl, 2.67 mM KCl, 8.1 mM Na_2HPO_4, 1.47 mM KH_2P0_4) (Invitrogen, Burlington, Ontario)

pH 7.4

2. Commercial Kits

DNA Isolation:

The QIAamp DNA Mini Kit (Qiagen, Mississauga, Ontario) was used for isolating genomic (including proviral) DNA from PBMCs.

PCR Reagents:

The full-length HIV-1 provirus was amplified with the Expand Long Template PCR System (Roche Diagnostics, Mannheim Germany). Subsequent reactions used the Expand High Fidelity PCR System (Roche Diagnostics, Mannheim Germany).

Amplification of genomic APOBEC3G DNA used the Expand Long Range, dNTPack System (Roche Diagnostics, Mannheim Germany).

Viral Load:

Viral load determinations were performed on 200µl heparin-treated plasma using the NucliSens® HIV-1 QT assay (bioMérieux, Marcy l'Etoile, France).

Sequencing:

HIV-1 was sequenced using the ABI Prism BigDye Terminator Cycle Sequencing Ready Reaction Kit v.3.1 (Applied Biosystems, Streetsville, Ontario). APOBEC3G was sequenced using the GS FLX Standard Series Kit for the Genome Sequencer FLX Instrument (454 Life Sciences, Branford, Conneticut).

3. Methods

3.1 Subject Selection

3.1.1 Pumwani sex worker (ML) cohort

HIV-positive women enrolled in a well described, highly exposed, commercial sex-worker (CSW) cohort in the Pumwani area of Nairobi, Kenya were randomly selected for inclusion in this study based on sample availability (86,87). Subjects were determined to be HIV infected by detection of HIV antibody. All subjects are believed to have been infected with HIV by heterosexual contact and all subjects were ARV naive at the time blood samples were collected. Each year, the sex worker cohort enrollees complete an administered, self-reported survey, describing sexual practices and other epidemiologic

factors. Both the University of Manitoba and University of Nairobi ethics review panels have approved these studies.

3.1.2 Mother child health (MCH) cohort

A mother/child health clinic operates in the same area of Pumwani as the CSW cohort. The non-sexworker attendees of this clinic are of a similar socio-economic status and ethnicity to the members of the ML cohort but with lower exposure to HIV due to a lower number of sexual partners (32,77,174). Demographic, social and clinical data are collected from these women at every visit. Subjects from this cohort were similarly included in this study.

3.1.3 HIV negative PBMC donors

HIV-1 negative blood donors were recruited from the University of Manitoba to provide fresh PBMCs for co-culturing HIV-1.

3.2 Sample Collection and PBMC Isolation

Peripheral blood samples were obtained by venipuncture into vacutainer tubes containing sodium heparin as an anticoagulant. A separate tube containing EDTA as an anticoagulant was collected for $CD4^+/CD8^+$ T cell enumeration. The blood was processed in an aseptic manner using universal precautions. Tubes containing blood samples were initially spun in a centrifuge for 7 minutes at 1500 RPM to separate out the plasma fraction. The plasma was removed and aliquoted into cryovials, one of which was sent for HIV-1 serology, while the rest were catalogued and stored at -80°C for

future studies. The remaining blood was diluted 1:2 with PBS-2%FCS and layered over ficol-hypaque (Invitrogen), using twice the volume of diluted blood as the ficol. Layered bloods were spun for 25 minutes at 1400 rpm to separate the PBMCs. After separation, the PBMC layer was removed into a new tube. The PBMCs were diluted with PBS-2% FCS and gently mixed prior to centrifugation at 1600 rpm for 10 minutes. After centrifugation, the PBS was removed and discarded, while the cells were resuspended in RPMI media. An aliquot was removed for counting and viability testing using a haemocytometer and trypan blue exclusion stain. The PBMCs were ultimately resuspended in RPMI containing 10% FCS and 2% penicillin/streptomycin at a final concentration of $5x10^6$ cells per mL.

Samples that were not destined for immediate use were stored in 90% FCS with 10% DMSO. These were then placed in an isopropyl freezing chamber at -80°C overnight to allowing cooling at approximately -1°C/min, and then were stored in liquid nitrogen.

When needed, samples were thawed in a 37°C waterbath, then immediately 10mL of RPMI + FCS + Pen/Strep media was added. The cells were then washed twice and resuspended in fresh RPMI + FCS + Pen/Strep media.

3.3 HIV Testing and Confirmation

All subjects in the ML and MCH clinics are routinely tested for HIV-1 serostatus at every visit. Plasma is tested using a Recombigen ELISA (Trinity Biotech, Carlsbad California). Samples with negative results in this assay are considered HIV-1 negative.

Samples with positive results are confirmed with a second immunoassay, Detect HIV1/2 (Adaltis, Montreal Quebec). Samples that have positive results in both assays are considered HIV-1 positive.

3.4 $CD4^+$ and $CD8^+$ T Cell Counts

Whole blood collected for $CD4^+$ and $CD8^+$ T cell counts was labeled using antibodies specific for CD4 and CD8 with the Tritest flow cytometry assay (BD Pharmingen, Mississauga Ontario). These samples were then processed with a FACScan flow cytometer (BD, Mississauga Ontario), allowing both $CD4^+$ and $CD8^+$ cell counts to be measured, as well as $CD4^+$percentage.

3.5 Viral Co-culture

HIV was expanded by primary co-culture with phytohemagglutinin (PHA) (Sigma-Aldrich, Oakville Ontario) stimulated PBMCs, as described by Lane (153). Fresh PBMCs were isolated from HIV negative, low-risk donors as described above. The PBMCs were grown in RPMI containing 10% FCS and 2% penicillin/streptomycin, with the addition of 5 μg/mL PHA. The cells were incubated at 37°C with CO_2 for 72 hours. A frozen aliquot of PBMCs (containing 2 – 4 x 10^6 cells) from an HIV-infected patient of interest was thawed and resuspended in 2 mL RPMI containing 10% FCS, 2% penicillin/streptomycin, 10 Units/mL IL-2 and 2 μg/mL polybrene. To this was added an equal amount of PHA-stimulated PBMCs from an HIV-negative, low-risk donor, in an equal volume. This coculture was maintained in media containing IL-2, with the periodic addition of fresh PHA-stimulated HIV-negative cells from the same donor. Supernatant

was removed at regular intervals and virus production was monitored by p24 ELISA. Cells were harvested at peak p24 production. The virus was collected and archived for virological studies.

3.6 DNA Isolation and Full-length HIV-1 Genome Amplification from Provirus

DNA was isolated from both PBMCs obtained directly from HIV-positive study subjects and from co-cultured PBMCs using the QIAamp DNA Mini Kit (Qiagen Inc, Mississauga Ontario). HIV-1 provirus was amplified by nested PCR reactions. The first reaction generated a nearly 9 kb amplicon, using published primers MSF12b (HXB2 location 623-649) and ofm19 (HXB2 location 9632-9604) and the Expand Long Template PCR System (Roche Diagnostics GmbH, Mannheim Germany) at the recommended conditions (Figure 3) (38). An initial denaturation cycle of two minutes at 94°C was followed by thirty amplification cycles, consisting of ten seconds of denaturing at 94°C, thirty seconds for primer annealing at 60°C and eight minutes of DNA elongation at 68°C. This was concluded with a final extension cycle of thirty minutes at 68°C, before cooling the reactions to 4°C.

Three separate secondary reactions, each using the Expand High Fidelity PCR System (Roche Diagnostics GmbH, Mannheim, Germany) at the recommended conditions, generated overlapping amplicons to span the first PCR product. The first of these amplicons ("gag") was 2 kb, produced with forward primer alnf1 (5'-GCCCGAACAGGGACYYGAAAGCGAAAG-3') and reverse primer GagRT (5'-CCATTGTTTAACCTTTGGGCCATCCA-3'). An initial denaturation cycle of two

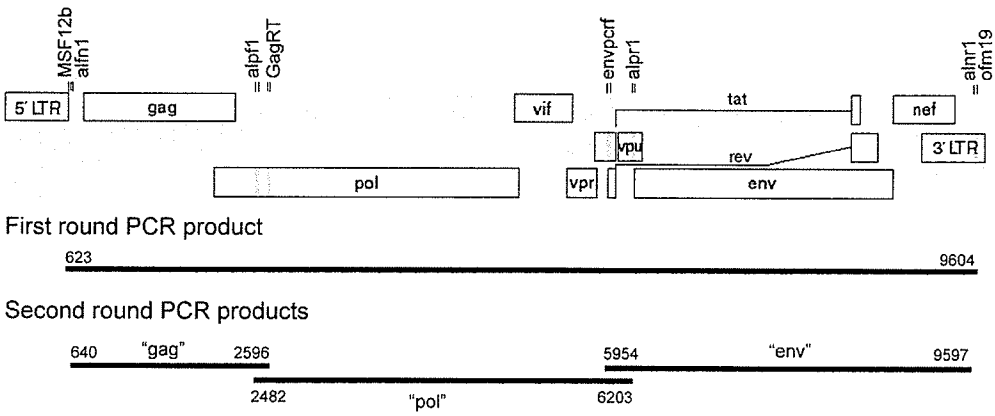


Figure 3. HIV-1 proviral PCR amplification scheme. The relative location of primers is indicated on top of the HIV-1 genome representation. The top black bar indicates the size and location of the primary PCR product, numbered according to HXB2 location, generated by primers MSF12b and ofm19 (38). The lower three black bars indicate the size and location of the secondary PCR products, numbered according to HXB2 location, generated by primer pairs alnf1 and GagRT, alpf1 and alpr1, and envpcrf and alnr1, respectively.

minutes at 94°C was followed by twenty-five amplification cycles, consisting of fifteen seconds of denaturing at 94°C, thirty seconds for primer annealing at 60°C and one and a half minutes of DNA elongation at 72°C. This protocol was concluded with a final extension cycle of five minutes at 72°C, before cooling the reactions to 4°C.

The second amplicon ("pol") was 3.7 kb, generated with forward primer alpf1 (5'-TAGGACCTACACCTGTCAACATAATTG-3') and reverse primer alpr1 (5'-TCATTGCCACTGTCTTCTGCTCTTTC-3'). Thermocycling commenced with an initial denaturation cycle of two minutes at 94°C. Twenty-five amplification cycles followed, consisting of fifteen seconds of denaturing at 94°C, thirty seconds for primer annealing at 57°C and three minutes of DNA elongation at 68°C. This protocol was concluded with a final extension cycle of seven minutes at 68°C, before cooling the reactions to 4°C.

The third amplicon ("env") was 3.6 kb, generated with forward primer envpcrf (5'-GGCTTAGGCATCTCCTATGGCAGGAAGAAG-3') and reverse primer alnr1 (5'-GGCAAGCTTTATTGAGGCTTAAGCAGTG-3'). An initial denaturation cycle of two minutes at 94°C was followed by twenty-five amplification cycles, comprised of fifteen seconds of denaturation at 94°C, thirty seconds for primer annealing at 60°C and three minutes of DNA elongation at 68°C. Thermocycling was concluded with a final extension cycle of seven minutes at 68°C, before cooling the reactions to 4°C. The first and second amplicon had a 114 base-pair overlap. The second and third amplicon had a 249 base-pair overlap.

3.7 Partial HIV-1 Genome Amplification from Provirus

3.7.1 Amplification of vpu/env *region*

DNA was isolated from PBMCs obtained from HIV-positive subjects using the QIAamp DNA Mini Kit (Qiagen Inc, Mississauga Ontario). Nested PCR reactions amplified proviral *vpu* and the first 349 nucleotides of *env*. The first reaction created an amplicon that was nearly 2 kb, using primers polseqf6 (5'-CAAGCAGGACATAACAAGGTAG-3') and envseqr4.5 (5'-TGTTATTTCTAGATCCCCTCCTG-3') and the Expand High Fidelity PCR System (Roche Diagnostics GmbH, Mannheim, Germany) at the recommended conditions. Thermocycling consisted of an initial denaturation cycle of two minutes at 94°C. Thirty amplification cycles followed, comprised of fifteen seconds of denaturation at 94°C, thirty seconds for primer annealing at 53°C and two minutes of DNA elongation at 72°C. This protocol was concluded with a final extension cycle of seven minutes at 72°C, before cooling the reactions to 4°C.

The secondary nested reaction specifically amplified the region of interest, using primers envpcrf (5'-GGCTTAGGCATCTCCTATGGCAGGAAGAAG-3') and envseqr6 (5'-CGAGTGGGGTTAACTTTACACATG-3'), also with the Expand High Fidelity PCR System at the recommended conditions. Thermocycling commenced with an initial denaturation cycle of two minutes at 94°C. Twenty-five amplification cycles followed, consisting of fifteen seconds of denaturing at 94°C, thirty seconds for primer annealing at 55°C and one and a half minutes of DNA elongation at 72°C. This protocol was

concluded with a final extension cycle of five minutes at 72°C, before cooling the reactions to 4°C.

3.7.2 Amplification of vif

Nested PCR reactions were also used to amplify the proviral *vif* region, using outer primers polseqf3.6 (5'-AGTTATCCCAGCAGAAACAGGAC-3') and polseqr1 (5'-TCGCTGTCTCCGCTTCTTCCTG-3') for the first reaction, which generated a 1.46 kb product. The thermocycling used the Expand High Fidelity PCR System and commenced with an initial denaturation cycle of two minutes at 94°C. Thirty amplification cycles followed, consisting of fifteen seconds of denaturing at 94°C, thirty seconds for primer annealing at 56°C and two minutes of DNA elongation at 72°C. This protocol was concluded with a final extension cycle of seven minutes at 72°C, before cooling the reactions to 4°C.

Primers polseqf4 (5'-CTGCAGTTAAGGCAGCCTGTTG-3') and polseqr1.5 (5'-CTTCAACTCCTGCCCAAGTATC-3') were used to generated the secondary, nested product of 1.09 kb. The Expand High Fidelity PCR System was used, with thermocycling conditions starting with an initial denaturing cycle of two minutes at 94°C. Twenty-five amplification cycles followed, consisting of fifteen seconds of denaturing at 94°C, thirty seconds for primer annealing at 55°C and one and a half minutes of DNA elongation at 72°C. This protocol was concluded with a final extension cycle of five minutes at 72°C, before cooling the reactions to 4°C.

3.8 HIV-1 Sequence and Phylogenetic Analysis

The nested PCR amplicons were purified using an AcroPrep™ 96 filter plate (Pall Life Sciences, Ville St. Laurent, Quebec), prior to sequencing. The purified PCR products were directly sequenced, using ABI Prism BigDye Terminator Cycle Sequencing Ready Reaction Kit v.3.1 (Applied Biosystems, Streetsville, Ontario). The sequencing reactions were resolved on an ABI Prism 3100 Genetic Analyser (Hitachi, Japan) and the resulting trace files were assembled into overlapping, double-stranded contigs, using Sequencher v.4.0.5 (Gene Codes Corporation, Ann Arbor, USA).

The full-length proviral sequence was generated by sequencing three overlapping amplicons. The first, termed the gag amplicon, was sequenced using the primers listed in Table 1. The second amplicon, pol, was sequenced with the primers listed in Table 2. The third and final amplicon, pol, was sequenced with the primers listed in Table 3.

The *vpu/env* segment was sequenced using the nested PCR primers envpcrf and envseqr6. The *vif* segment was sequenced using the nested PCR primers polseqf4 and polseqr1.5, as well as the primers polseqf5 and MM4b/polseqr2.5.

Regions from the full-length provirus that proved troublesome to resolve by direct sequencing of the PCR product were cloned into TOP10 Chemically Competent cells, using the pCR®4-TOPO® vector (TOPO TA Cloning® Kit for Sequencing, Invitrogen™ Life Technologies, Carlsbad, California). For each insert, ten to twenty clones were

Table 1: Gag amplicon sequencing primers

Name	Sequence	Orientation	Location[a]	Tm
alnf1	GCCCGAACAGGGACYYGAAAGCGAAAG	Forward	640-666	66.6
Gagseqf0	GAAAGCGAAAGTTCCAGAGAAG	Forward	656-680	58.2
Gagseqf1.1	GAGATGGGTGCGAGAGCGTC	Forward	787-806	64.0
Gagseqf1.5	YTRGTATGGGCAAGCAGGGAG	Forward	889-909	60.0
Gagseqf1.6	GATAGAGGTAAAAGACACCAAG	Forward	1062-1083	56.3
Gagseqf2[b]	CAGCATTATCAGAAGGAGCCAC	Forward	1307-1328	60.1
Gagseqf2.5	GARGYGAYATAGCAGGAACTAC	Forward	1487-1508	56.3
Gagseqf2.9	CTACATTAGAAGAAATGATGAC	Forward	1811-1832	52.6
Gagseqf3.5	GRGTKTTRGCYGAGGCAATGAG	Forward	1871-1892	58.2
Gagseqf3.6	CMAAYATAATGATGCAGAGAG	Forward	1910-1930	52.2
Gagseqf3.8	AGCCAACAGCCCCACCAGAGC	Forward	2150-2170	65.8
Gagseqf4[b]	TTAGATACAGGAGCAGATGATACAG	Forward	2322-2346	58.7
Gagseqr1.5	TCAGTGCAGTCTTTCATTTGGTG	Reverse	2072-2050	58.4
Gagseqr2[b]	CATTGCCTCAGCCAAAACTCTTGC	Reverse	1890-1867	62.0
Gagseqr2.5	ARATKTCTCCYACTGGGAYAG	Reverse	1573-1553	54.1
Gagseqr2.9	CAGCYTCCTCATTGATGGTATC	Reverse	1417-1396	58.2
Gagseqr3[b]	CATGGCTGCTTGATGTCCCCCCAC	Reverse	1383-1360	67.1
Gagseqr4[b]	AGCTCCCTGCTTGCCCATAC	Reverse	911-892	61.9
GagRT[b]	CCATTGTTTAACCTTTGGGCCATCCA	Reverse	2621-2596	62.0

[a]Location is given in HXB2 numbering

[b]These are previously designed lab primers

Table 2: Pol amplicon sequencing primers

Name	Sequence	Orientation	Location[a]	Tm
alpf1	TAGGACCTACACCTGTCAACATAATTG	Forward	2482-2508	60.5
alpr1	TCATTGCCACTGTCTTCTGCTCTTTC	Reverse	6228-6203	62.0
Polseqf0.5	GTTAAACAATGGCCAYTGACAG	Forward	2610-2631	56.3
Polseqf0	GYAYAAAYAATGAGACACCAG	Forward	2950-2970	52.2
Polseqf1	TATCAGTACAATGTGCTTCCAC	Forward	2979-3000	56.3
Polseqf1.5	GATGAYTTRTATGTAGGATCTG	Forward	3102-3123	52.6
Polseqf2.1	GAATTAGAATTGGCAGAGAACAG	Forward	3447-3469	56.6
Polseqf2.5	GAYRGACTAYTGGCAGGCTAC	Forward	3755-3775	58.0
Polseqf3	CAGACTCACAATATGCATTAG	Forward	4039-4059	54.1
MM3a[b]/ polseqf3.5	CATGGGTACCAGCACAGAAAGG	Forward	4150-4171	61.9
Polseqf3.6	AGTTATCCCAGCAGAAACAGGAC	Forward	4490-4512	60.2
Polseqf4	CTGCAGTTAAGGCAGCCTGTTG	Forward	4600-4621	61.9
Polseqf5	ATGGCAGGTGATGATTGTGTGGC	Forward	5052-5074	62.0
Polseqf6	CAAGCAGGACATAACAAGGTAG	Forward	5446-5467	58.2
HIVB10[c]/ Polseqf7	CTATGGCAGGAAGAAGCGGAGAC	Forward	5968-5990	63.7
Polseqr1	TCGCTGTCTCCGCTTCTTCCTG	Reverse	5995-5974	63.8
Polseqr1.5	CTTCAACTCCTGCCCAAGTATC	Reverse	5733-5712	60.1
polseqr2	ATCCTACCTTGTTATGTCCTG	Reverse	5470-5450	56.1
MM4b[b]/ Polseqr2.5	TGGATGTGTACTTCTGAACTTA	Reverse	5213-5192	54.5
NPol4481[b] /Polseqr3	CTGCTGTCCCTGTAATAAACCCG	Reverse	4921-4899	62.0
SEQ4485[b]/ Polseqr4	GTTTCTGCTGGGATAACTTCTGC	Reverse	4507-4485	60.2
MM3b[b]/ Polseqr4.5	CCTTTGTGTGCTGGTACCCATG	Reverse	4171-4150	61.9
SEQ3869[b]/ Polseqr5	TAGCTGCCCCATCTACATAG	Reverse	3888-3869	57.8
SEQ3630[b]/ Polseqr5.5	ATTGTTTTACATCATTAGTGTG	Reverse	3651-3630	50.8
Polseqr6	TCTGTATATCATTGACAGTCCAG	Reverse	3324-3302	56.6
Polseqr6.5	GTGTCTCATTGTTTGTACTAG	Reverse	2967-2947	54.1
Polseqr7	GCAAATACTGGAGTATTGTATG	Reverse	2734-2713	54.5

[a]Location is given in HXB2 numbering

[b]These primers were designed by Melanie Murray (unpublished data).

[c]This primer was published by Oelrichs *et al.* (220)

Table 3: Env amplicon sequencing primers

Name	Sequence	Orientation	Location[a]	Tm
Envpcrf[b]	GGCTTAGGCATCTCCTATGGCAGGAAGAAG	Forward	5954-5983	67.5
Envpcrf1[b]	GAAAGAGCAGAAGACAGTGGCAATG	Forward	6203-6227	62.0
Envseqf2[b]	TGGGCTACACATGCCTGTGTACC	Forward	6429-6451	63.7
Envseqf2.5	GAYSAAAGCCTAAAGCCATGTG	Forward	6561-6582	56.3
Envseqf3[b]	CCAATTCCCATACATTATTGTG	Forward	6858-6879	56.6
Envseqf4[b]	GGAGGGGACCTAGAAATTACAACACA	Forward	7320-7345	62.0
Envseqf5[b]	CAGCAGGAAGCACTATGGGCG	Forward	7798-7818	63.9
Envseqf5.5	TGCCTTGGAACTCTAGTTGGAG	Forward	8047-8068	60.1
Envseqf6[b]	GAGTTAGGCAGGGATACTCACC	Forward	8344-8365	61.9
Envseqf6.1	ATTCGMTTAGTGAGCGGATTC	Forward	8460-8480	56.1
Envseqf7[b]	ATACCATAGCAATAGCAGTAGCTG	Forward	8671-8694	58.5
Envseqf7.5	CTCAGGTACCTTTAAGACCAATG	Forward	9011-9033	58.4
Envseqf8[b]	GCCCGAGAGCTRCATCCGGAG	Forward	9379-9399	65.8
Envseqr0.1	GTGTAAYAGRCTGTTGTTCTCTC	Reverse	9294-9272	56.6
Envseqr1[b]	TTTGACCACTTGCCACCCAT	Reverse	8816-8797	57.8
Envseqr1.5	ATCGAATGGATCTGTCTCTG	Reverse	8466-8447	55.8
Envseqr2[b]	GGTGAGTATCCCTGCCTAACTC	Reverse	8365-8344	61.9
Envseqr2.5	CTTGTTCATTCTTTTCCTGCTG	Reverse	8199-8178	56.3
Envseqr3[b]	CAATAATTGTCTGGCCTGTACCGT	Reverse	7859-7836	60.3
Envseqr3.5	GTGGGTGCTACTCCTAGTGGTTC	Reverse	7720-7698	63.7
Envseqr4.5	TGTTATTTCTAGATCCCCTCCTG	Reverse	7340-7318	58.4
Envseqr5[b]	TTCCATGTGTACATTGTACTG	Reverse	6975-6955	54.1
Envseqr5.5	CATTAAGTGGTACTAWATCAAGT	Reverse	6783-6758	55.1
Envseqr6[b]	CGAGTGGGGTTAACTTTACACATG	Reverse	6600-6577	60.3
alnr1	GGCAAGCTTTATTGAGGCTTAAGCAGTG	Reverse	9624-9597	63.5

[a]Location is given in HXB2 numbering

[b]These are previously designed lab primers

sequenced. When clonal variability was observed, the clone that was most represented in the population filled the sequence gap.

The HIV-1 sequences from each subject (whether full-length or partial) were aligned against the 2005 (current) reference sequences from the Los Alamos HIV sequence database (hiv.lanl.gov) using ClustalW (49). MEGA 3.1 was employed for neighbour-joining phylogenetic tree analysis (nucleotide distance calculated by Kimura's two-parameter method), enabling clade status to be assigned to each HIV isolate or segment where appropriate (148). The strength of the branch nodes was tested with bootstrap analysis (500 replicates). Recombination breakpoint analysis was performed with the Recombination Identification Program: RIP 3.0 (www.hiv.lanl.gov). The sequences were realigned between the putatively identified breakpoints and the segments were re-analyzed through the construction of neighbour-joining phylogenetic trees with reference sequences, including previously characterized CRFs where appropriate.

3.9 Hypermutation detection

Hypermut 2.0, available from www.lanl.hiv.gov, was employed to detect APOBEC-type hypermutation (247). The sequences analyzed in each section were used to generate a population-specific consensus for each HIV-1 clade that was represented. The sequences were then compared on a clade by clade basis to this consensus using Hypermut 2.0, which identified guanine to adenine hypermutation, and also the nucleotide context of this hypermutation.

3.10 Determination of Viral Load

Viral nucleic acids were isolated from heparin plasma using the NucliSens® HIV-1 QT assay (bioMérieux, Marcy l'Etoile, France) according to the manufacturer's instructions. Briefly, 200 µL plasma was added to 3 mL NucliSens® lysis buffer (buffered guanidium thiocyanate). Silica was added and the silica-bound nucleic acids were washed sequentially with NucliSens® wash buffer, 70% ethanol and acetone. Finally, the nucleic acids were eluted into a low-salt buffer. The lower detection limit of the assay at this volume is 125 copies/mL (29).

3.11 Amplification of Genomic APOBEC3G Gene

DNA was isolated from subject PBMCs as described above. Unique PCR primers were designed to amplify the coding region of the gene and the surrounding region containing all known APOBEC3G SNPs. Forward primer A3Gfb (5'-CCCACACTTAAACAGTCAACTCTG-3') was located 2044 bases upstream of the APOBEC3G transcript, while the reverse primer A3Grb (5'-GTTGTGTGTAGTGCGAGTATTGTG-3') was located 829 bases downstream of the end of the transcript, resulting in an amplicon of 13486 nucleotides (Figure 4). The amplicon was generated using the Expand Long Range, dNTPack (Roche Diagnostics GmbH, Mannheim, Germany) at the recommended conditions. Primers were employed at a final concentration of 0.3 µM in a reaction volume of 50 µL, containing 3% DMSO. Thermocycling commenced with an initial denaturation cycle of two minutes at 92°C. This was followed by thirty amplification cycles, consisting of ten seconds of denaturing at 92°C, fifteen seconds for primer annealing at 56°C and fourteen minutes of DNA

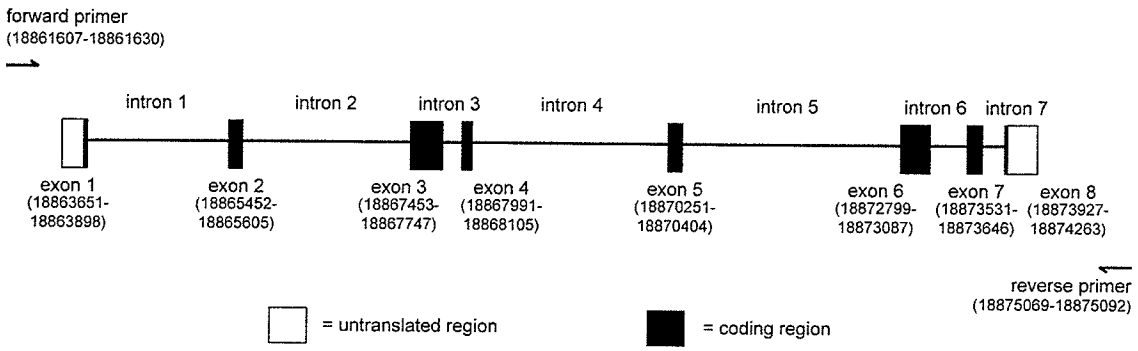


Figure 4. APOBEC3G PCR amplification scheme. The location of the primers is indicated. The exons (untranslated regions and coding regions) and introns are drawn to scale. The exon positions are shown. All positions are numbered in accordance with the NCBI SNP database for the human APOBEC3G gene.

elongation at 68°C. Thermocycling was concluded with a final extension cycle of seven minutes at 68°C, before cooling the reactions to 8°C. If non-specific bands were observed, the primer annealing temperature was increased incrementally to 60°C, on a sample-specific basis.

Some samples were further subjected to nested PCR amplification, to increase yield. Primers that overlapped the first round primers were designed. Nested forward primer A3GinF (5'-CAACTCTGTTCTGCCTGCCACAAG-3') was located 2028 bases upstream of the APOBEC3G transcript, while nested reverse primer A3GinR (5'-GTATTGTGTGTGCGTGTTGTGTGTG-3') was located 813 bases downstream of the end of the APOBEC3G transcript, resulting in a nested amplicon of 13453 nucleotides. The nested amplicon was also generated using the Expand Long Range, dNTPack (Roche Diagnostics GmbH, Mannheim, Germany) at the same conditions as the primary PCR reaction, with the exception that a higher primer annealing temperature of 68°C was employed.

3.12 APOBEC3G Pyrosequencing

Prior to sequencing, each APOBEC3G amplicon was purified and quantified. The amplified genomic DNA was subsequently sheared into 300 to 800 bp fragments, to facilitate even coverage of the sequencing. Up to 5000 ng of the samples to be sequenced individually was sheared. For the six samples pooled to make the intermediate hypermutation pool, 400 ng of each sample was pooled, and then 1650 ng from this pool was sheared. For the eighty-seven samples pooled to make the low hypermutation pool,

200 ng of each sample was pooled, and then 1570 ng from this pool was sheared. Adaptors were annealed to the sheared DNA, which was then denatured. The now single-stranded fragments were immobilized onto beads via the adaptor. The beads and DNA are mixed such that each bead carries only a single DNA molecule.

The beads were then emulsified with PCR amplification reagents in a water-in-oil mixture to create multiple microreactions, each of which contained a single bead, with a single DNA fragment. The emulsification PCR amplification resulted in each bead being covered with millions of identical DNA fragments. These beads were then loaded onto a PicoTiterPlate (454 Life Sciences, Branford, Connecticut) for sequencing – the diameter of the wells in the plate is such that only a single bead fits. Sequencing enzymes were added to each well and the Genome Sequencer FLX Instrument (454 Life Sciences) flowed individual nucleotides, in a fixed order, across the wells. Pyrosequencing is based on sequencing by synthesis – when a complementary nucleotide was added, there was a chemiluminescent signal, the intensity of which is proportional to the number of nucleotides incorporated. The signals were recorded and subsequently interpreted by the GS FLX System (454 Life Sciences) software as sequence information.

3.13 Pyrosequence Assembly and Analysis

The sequence reads captured by GS FLX System software were aligned against a reference sequence using GS Mapper Software (454 Life Sciences). The reference sequence used was nucleotides 37801000 to 37814600 of *Homo sapiens* chromosome 22, reference assembly (Accession number NC_000022.9), which includes APOBEC3G and

the surrounding region and spans the amplicon generated in this thesis. The GS Mapper Software identified sequence differences between the reference and sample sequences, and provided the sequence of individual reads that both supported, and did not support, the putative polymorphism. The sequences provided were a subset of the total reads that covered that particular region; only sequences that could support or refute the polymorphism with high confidence were retrieved (i.e. high quality read, the sequence neither started nor ended at the nucleotide in question).

Each polymorphism that the software identified was manually inspected. The number of sequences that supported and refuted each polymorphism was adjusted, as necessary, based on the manual inspection. Each polymorphism in the nine individual samples was reported, if determined to be valid after examination. However, for the pooled samples, only polymorphisms present in 10% or more of the individuals reads were reported, if they proved valid, unless the polymorphism was also found in other samples. This conservative approach to SNP discovery was applied to prevent the reporting of sequence anomalies as novel SNPs.

3.14 Statistics

3.14.1 Chi-square test

A chi-square test is used on categorical data. The chi-square test was used to test for significant differences in the distribution of HIV-1 clade and recombination with categorical measures of HIV exposure; specifically, whether the individuals were HIV negative at time of enrolment into the ML or MCH cohorts, whether the individuals

practiced commercial sex work, and the degree of condom usage. In this thesis, the statistical program SPSS for Windows 11.0.1 was employed for chi-square testing.

3.14.2 Power calculation

A power calculation is used to determine β error, where β is the likelihood that the study will not detect genuine differences (power = 1 – β). A power calculation for the Chi-square test was employed to determine the appropriate sample size needed to detect significant differences between markers of HIV-1 exposure and infection with a non-recombinant compared to a recombinant HIV-1 strain, using Java Applets for Power and Sample Size (156).

3.14.3 One-way ANOVA

A one-way ANOVA test is used to compare three or more groups of continuous data. A one-way ANOVA was used to test for significant differences in the distribution of HIV-1 clade and recombination with numerical measures of HIV exposure, specifically $CD4^+$ cell count, viral load, subject age, year of birth, years in sex work, years in the ML cohort, number of sex partners per day, and the year sex work was commenced. These tests were performed using SPSS for Windows 11.0.1.

Additionally, the Kruskal-Wallis test, which is a non-parametric one-way ANOVA, was used to test for significant differences in the APOBEC3F/G rate ratio values in proviral sequence obtained from the original infecting virus before and after superinfection, and the superinfecting virus. This test was performed with the statistical program GraphPad

Prism 4 for Windows. Dunn's post test was used to follow up on identified statistically significant differences from the Kruskal-Wallis test. A non-parametric test was chosen as biological data rarely follows a normal or Gaussian distribution, which would assume that the data extends infinitely in both directions. However, ANOVA is a relatively robust test for non-parametric data, providing the sample is reasonably large and the data follows an approximate Gaussian distribution. Unfortunately, non-parametric tests are less powerful than those that assume a Gaussian distribution, making it harder to detect real differences. This lack of power is mitigated with larger sample sizes (204).

3.14.4 Two-factor ANOVA

A two-factor ANOVA is used to determine how responses are affected by two factors, both independently and if they interact. A two-factor ANOVA was used to determine if age of subject was significantly associated with HIV-1 clade/recombination independently of involvement in sex work. This test was performed using SPSS for Windows 11.0.1.

3.14.5 Kolmogorov-Smirnov test

A Kolmogorov-Smirnov test examines data for deviations from a Gaussian distribution. The Kolmogorov-Smirnov test was employed to determine if the adenine proportions from 240 HIV-1 proviral sequences segments were normally distributed, using GraphPad Prism 4.

3.14.6 Mann-Whitney

A Mann-Whitney test is a non-parametric test used to compare two groups of continuous data. The Mann-Whitney test was used to determine if the adenine proportion in HIV-1 RNA sequence was significantly different from the adenine proportion in matched HIV-1 proviral sequence for individuals infected with hypermutated as well as non-hypermutated virus. Additionally, the Mann-Whitney test was used to test for significant differences in $CD4^+$ cell count, $CD4^+$ percentage, viral load and subject age between individuals infected with hypermutated HIV-1 and individuals infected with non-hypermutated HIV-1. The Mann-Whitney test was also used to determine if there was a significant difference between the APOBEC3F/G rate ratio values in proviral sequence obtained before superinfection compared to after superinfection. This test was performed using GraphPad Prism 4.

3.14.7 Test for correlation

A correlation test is used to determine if two variables have covariation, without discriminating cause and effect. The correlation coefficient, r, quantifies the direction and magnitude of the correlation (204). A Spearman (non-parametric) test for correlation was employed to determine if adenine proportion in HIV-1 proviral sequence was significantly correlated with $CD4^+$ count for 240 subjects. This test was performed using GraphPad Prism 4.

3.14.8 Binomial distribution

The binomial distribution is the discrete probability distribution of the number of successes (k) in a sequence of n independent experiments (which have a yes/no outcome), where success occurs with probability, p. The equation for this function is:

$$f(k;n,p) = \binom{n}{k} p^k (1-p)^{n-k}$$

where $\binom{n}{k} = (n!/(k!(n-k)!)$

A cumulative binomial distribution describes the probability of finding k or fewer successes in a sequence of n independent experiments (which have a yes/no outcome), where success occurs with probability, p. The equation for this function is:

$$F(k;n,p) = \sum_{0}^{k} f(k;n,p)$$

A binomial distribution was used to determine the maximum expected number of reads per APOBEC3G allele in the pool of 87 samples with low HIV-1 proviral hypermutation. A cumulative binomial distribution equation was employed for determining the minimum number of expected reads per APOBEC3G allele in the nine samples with high proviral hypermutation, as well as the pools of samples with intermediate and low proviral hypermutation. Microsoft Office Excel 2003 was used to calculate these distributions.

3.14.9 Chi-Square Test for Trend

A chi-square test for trend is an extension of the chi-square test, which determines if there is a linear trend in the data. The chi-square test for trend was used to test the APOBEC3G allelic SNP distribution for individuals infected with highly hypermutated HIV-1 provirus, intermediately hypermutated provirus and non-hypermutated provirus.

In this thesis, the software GraphPad Prism 4 was employed to determine chi-square test for trend.

Results

1. Full-length HIV-1 Proviral Sequencing of Ten Highly Exposed Women Reveals a High Proportion of Intersubtype Recombinants

1.1 Rationale

Full-length HIV sequencing is the best method for detecting viral intersubtype recombination. The high prevalence of HIV infection in Kenya and co-circulation of multiple subtypes makes this a likely place to find recombinant HIV. Despite this, the majority of sequence information from Kenya is based on partial genome sequencing, while the available full-length sequence information is limited, illustrating the need for increased full-length HIV sequences from this region. One of the reasons that partial genomic sequencing is performed is that HIV-1 provirus can be integrated at very low levels in host DNA, making it difficult to amplify full-length genomes. Co-culture of infected PBMCs with non-infected cells will amplify HIV-1, but may introduce sequence mutations. Previous studies have sampled blood bank donors in an effort to obtain a profile of infecting HIV in the general Kenyan population. It thus remains to be seen if a specific sub-population, such as a group at high risk for HIV infection, has a similar HIV sequence distribution to the general population.

1.2 Hypothesis

Previous studies have shown that individuals at high risk of acquiring HIV, such as CSW, are more likely to be infected with recombinant virus than those at a lower risk, if they undertake their high risk activities in an area where multiple subtypes circulate (12,13,117). We thus expect our sequence analysis of full-length HIV-1 provirus to

reveal a higher proportion of recombination than what has been previously described in the general Kenyan population. We furthermore hypothesize that within this highly-exposed population, increased duration of sex work will be correlated with increased likelihood of infection with recombinant HIV, as individuals who practice sex work for a longer time will have more chances for exposure to different viral subtypes.

1.3 Objectives

- Compare sequences generated from viral amplification using co-cultured subject PBMCs with sequences generated directly from non-cultured subject PBMCs to determine if significant sequence mutations are introduced
- Generate full-length HIV-1 proviral sequence from ten subjects from the ML cohort
- Determine the subtype of the HIV-1 sequences and characterize the extent of recombination
- Correlate HIV-1 recombination with duration of commercial sex work and other pertinent epidemiological factors

1.4 Study outline

PBMCs were collected from ten HIV-infected women from the ML cohort and co-cultured with HIV-negative cells to amplify the virus prior to genomic DNA isolation. Additionally, for two subjects, PBMCs were directly used as a substrate for isolating genomic DNA. Subjects were randomly selected based on sample availability and sample collection within a one-year period. Subjects included both new enrolees and

patients that had been followed for a number of years in the cohort. Proviral HIV-1 was amplified from the isolated genomic DNA. Multiple primers were used to generate amplicons that were sequenced and assembled into overlapping double-stranded contigs that spanned the full-length of the HIV genome. Each full-length genome was analyzed for recombination with the web-based RIP tool and by neighbour-joining tree generation. A self-reported survey of a multitude of epidemiological factors, including duration of sex work, number of partners, and condom usage, is administered to cohort participants bi-annually. Variables that may affect HIV exposure collected in these surveys were correlated with presence of recombination.

1.5 Evaluation of sequence after co-culture and multiple PCRs

To assess the role of short-term PBMC co-culture on the resulting progeny proviral sequences, we compared proviral sequence generated from DNA isolated directly from patient PBMCs, to proviral sequence generated from DNA isolated from co-cultured PBMCs for two subjects: ML1979 and ML1990. PBMCs for both viral co-culture and direct DNA isolation were isolated from the same blood sample. Total genomic DNA isolated directly from one aliquot of the PBMCs was used as a template for PCR amplification and proviral sequencing. A second aliquot of PBMCs was co-cultured with HIV-uninfected donor PBMCs to allow for amplification of the virus *in vitro*. After limited primary co-culture (cells were harvested at peak p24 production), genomic DNA from these PBMCs was also isolated for use as a template for PCR amplification and proviral sequencing. The sequences generated by direct DNA isolation and by primary co-culture prior to DNA isolation were compared by determining pairwise similarity

(Table 4). The sequences derived from patient ML1979 showed 97% identity, while the sequences derived from patient ML1990 showed 95% identity, showing that viral amplification by limited PBMC co-culture introduced a minimal number of nucleotide changes. Sequence analysis revealed that ML1979 was infected with a recombinant HIV virus; the recombination breakpoints in the sequence generated from direct DNA isolation, compared to the breakpoints in the sequence generated from primary co-culture prior to DNA isolation, were virtually identical (Table 5). The largest difference observed was in the second breakpoint, differing by 17 +/- 6 nucleotides, where the parental sequence changed from clade G to A1. Overall, minimal differences were observed between the sequences generated by direct DNA isolation compared to DNA isolation after amplifying co-culture.

To examine the effects of repeated long PCR amplification of the same proviral sample, HIV provirus was amplified from a single DNA sample from subject ML1901 in two separate PCR amplifications, using the Expand Long Template PCR System (Roche Diagnostics GmbH, Mannheim, Germany). These sequences were compared by determining their pairwise similarity. The proviral sequences, generated by separate PCR amplifications, were highly similar (99% identity) (Table 4). Thus, minimal differences were observed between the sequences generated by separate PCR amplifications.

To further assess the relationship between the sequences generated from direct DNA isolation compared to DNA isolation after amplifying co-culture, phylogenetic relationship was determined (Figure 5). The proviral sequences from subjects ML1979

Table 4. Pairwise similarity of full-length HIV-1 proviral sequences generated from a single subject by parallel treatments

Patient	Sample	Nucleotide Length	Identical nucleotides	% Identical
ML1979[a]	PCR	8963	8781	97
	Co-culture	8943		
ML1990[a]	PCR	8938	8547	95
	Co-culture	8947		
ML1901[b]	PCR 1	8940	8928	99
	PCR 2	8942		

[a]The same PBMCs were used for direct DNA isolation and viral co-culture prior to DNA isolation.

[b]Two separate long PCR amplifications were performed on the same DNA template.

Table 5. Differences in breakpoints between full-length HIV-1 proviral sequences from subject ML1979 generated by different methodologies

ML1979 PCR breakpoints[a]	ML1979 co-culture breakpoints[a]	Difference
2036	2035	1
3234 ± 5	3251 ± 1	17 ± 6
4102	4102	0
4553 ± 26	4553 ± 26	0 ± 52
5530 ± 5	5530 ± 5	0 ± 10
5999 ± 6	5991 ± 1	8 ± 7
6754 ± 3	6750	6 ± 3
8363 ± 40	8364 ± 44	1 ± 74

[a]All numbering is based on HXB2 coordinates.

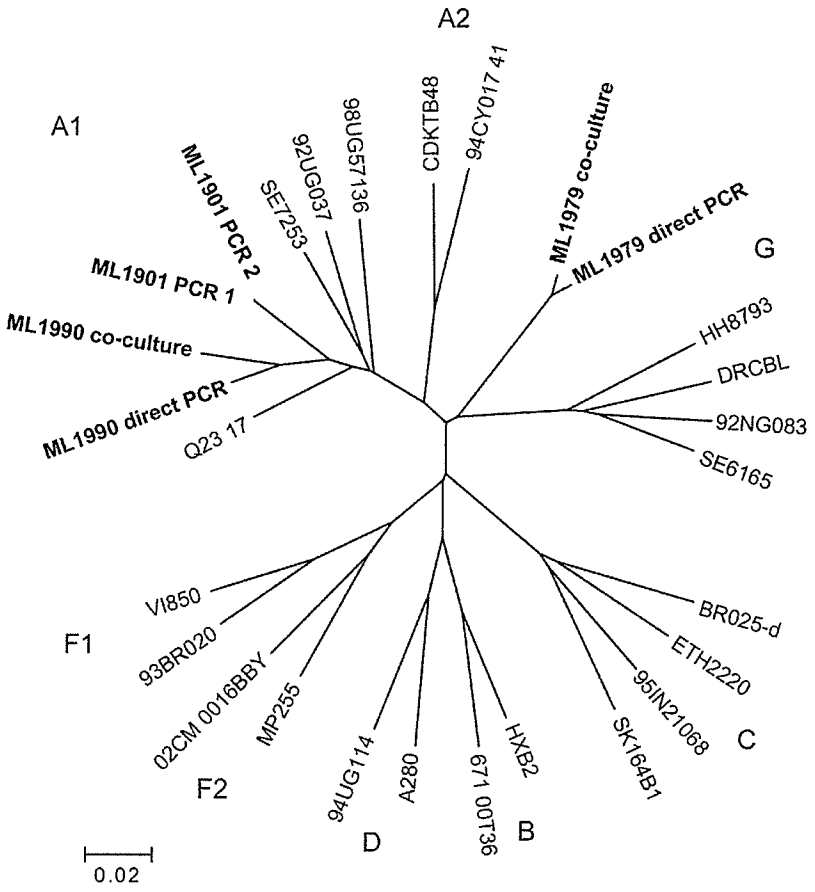


Figure 5. Neighbour-joining tree of HIV-1 sequences generated by parallel methods and reference sequences. Sequences were aligned with ClustalW and the phylogenetic tree was created using MEGA 3.1. 1901 PCR1 and 1901 PCR2 are sequences generated from the same patient by different long PCR amplifications. 1979 co-culture and 1979 direct PCR are sequences generated from the same patient, but the first sequence was generated from DNA isolated from PBMCs that were subjected to primary co-culture. Similarly, 1990 co-culture and 1990 direct PCR are generated from the same patient. Sequences generated in this thesis are indicated in bold. Reference sequences were obtained from the Los Alamos HIV Sequence Database. Scale of genetic distance is indicated.

and ML1990, generated directly from PBMC DNA and from PBMCs that were co-culture to amplify the virus, were aligned with reference sequences and this alignment generated a neighbour-joining tree. The sequence pairs from both subjects ML1979 and ML1990 clustered with a bootstrap value of 100%. The phylogenetic relationship between the sequences from ML1901 generated by separate PCR amplifications was assessed in the same manner. The pair of sequences from subject ML1901 also clustered with a bootstrap value of 100%. Thus, the phylogenetic analysis was in agreement with the pairwise similarity, that these sequence pairs have minimal differences.

1.6 Phylogenetic analysis of the full-length sequences

After determining that proviral sequences generated from viral ampflication by co-culture of PBMCs did not differ significantly from sequences generated directly from subject PBMCs, PBMCs from an additional eight study participants were co-cultured prior to DNA isolation and sequence generation. The HIV-1 clade infecting each subject was determined by generating a neighbour-joining tree using the ten full-length patient sequences, along with current reference sequences (www.hiv.lanl.gov) (Figure 6). Five of the viral sequences, derived from subjects ML752, ML1901, ML1945, ML1990 and ML2014, clustered within the A1 reference sequences with a bootstrap value of 100%, indicating that they are non-recombinant sequences belonging to the A1 subtype. We further attempted to identify recombination using Simplot v.3.5.1, and a deficiency of multiple parental sequences confirmed the absence of recombination (169). The other five sequences branched basal to the reference sequence clades, indicating that these were likely recombinant. The sequences derived from subjects ML1076 and ML1956

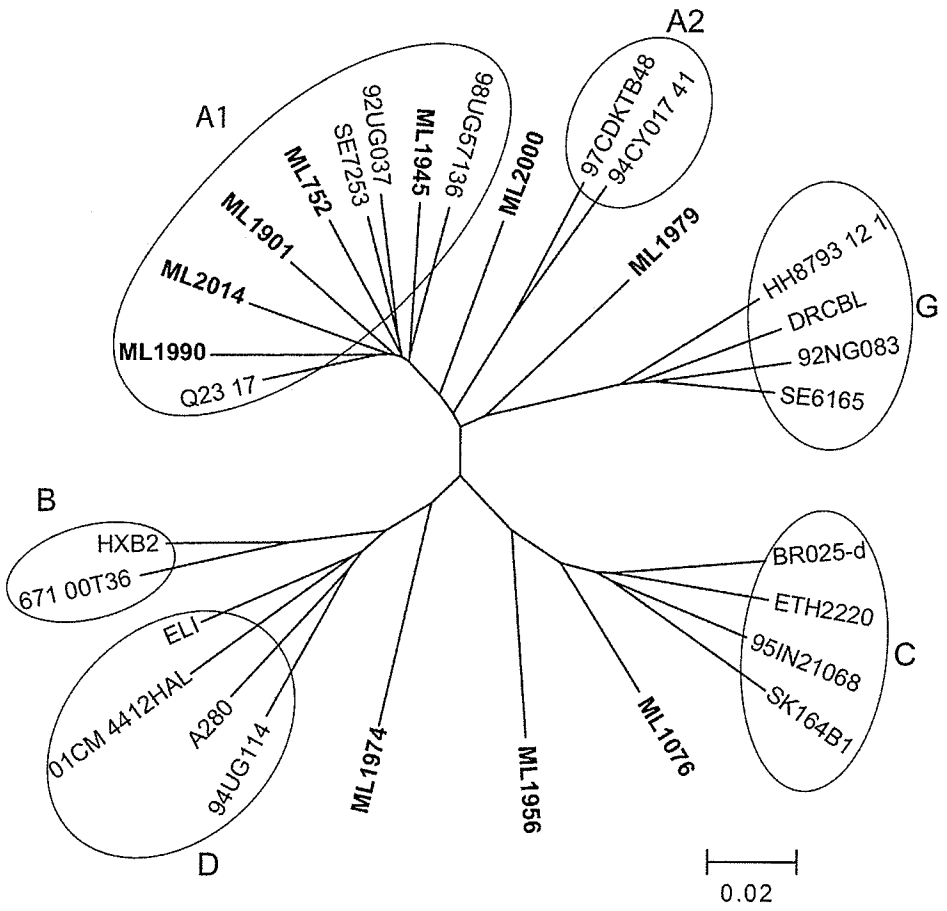


Figure 6. Neighbour-joining tree of ten full-length HIV-1 sequences and references. Full-length HIV-1 proviral sequences were aligned using ClustalW; the phylogenetic tree was created using MEGA 3.1. Sequences described in this publication, 756, 1076, 1901, 1945, 1956, 1977, 1979, 1990, 2000 and 2014 are indicated in boldface. Reference sequences were obtained from the Los Alamos HIV Sequence Database. Non-recombinant sequences are circled by clade. Scale of genetic distance is indicated.

branched basal to the subtype C group, the sequence from ML1974 branched basal to the B and D subtypes group, the sequence from ML1979 branched basal to the subtype G group, while the sequence from ML2000 branched basal to the subtype A1 group. Thus, preliminary analysis suggests that five proviral sequences were non-recombinant, while the other five were likely intersubtype recombinant.

To determine the degree of variability within the clade A1 sequences, they were aligned with all publically available clade A1 full-length HIV sequences from Kenya. This alignment was used to generate a neighbour-joining phylogenetic tree (Figure 7). The sequences that were multiply amplified and sequenced from the same individual clustered together with bootstrap values of 100. Additional sequences from this cohort, generated by previous studies, are included in the analysis. Two sequences generated from samples isolated on different dates from the same subject – ML752 – are present, with one sequence generated in this thesis and the other in a previous publication (80). These also cluster with a bootstrap value of 100. However, sequences generated from different subjects from the cohort, created in both this thesis and in a previous publication, are dispersed over the tree with the other sequences, suggesting that the virus infecting individuals of the ML cohort are not genetically distinct from virus infecting other Kenyan individuals.

To identify the presence of recombination and the location of breakpoints in the five likely recombinant HIV-1 sequences, we employed the RIP 3.0 recombination identification tool (www.lanl.gov). The sequences were split at the putatively identified

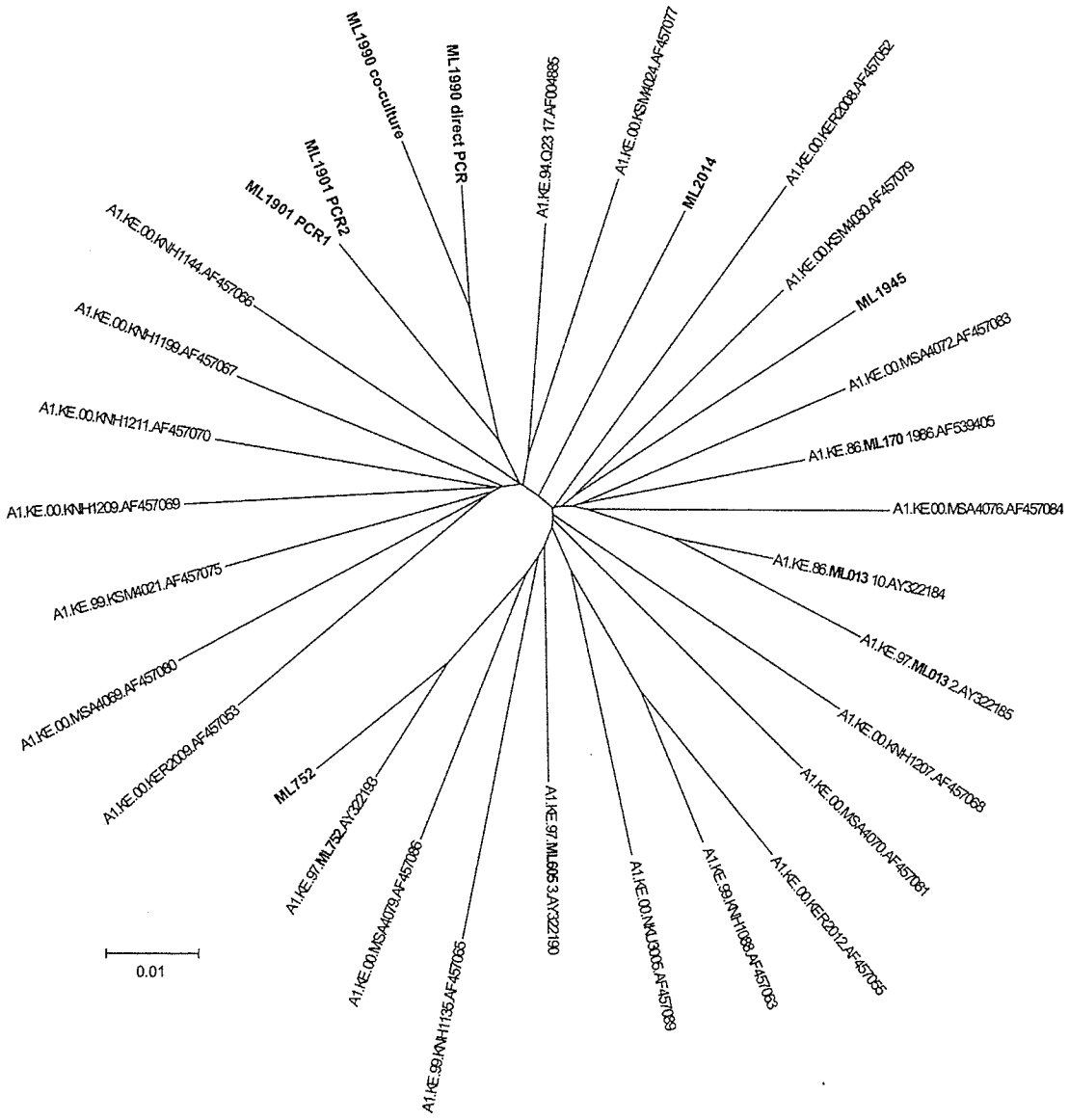


Figure 7. Neighbour-joining tree of full-length clade A1 sequences generated in this thesis and all other publically available full-length HIV-1 clade A1 sequences from Kenya (74,80,237). Sequences generated in this thesis are labelled in bold. Sequences that were downloaded from GenBank are listed as HIV-1_clade.country_name.sample_year.sample_name.accession_number. Other publicly available sequences from the ML cohort have the sample name (ML number) bolded. Sequences were aligned with ClustalW and the phylogenetic tree was created using Mega 4. Scale of genetic distance is indicated.

breakpoints, realigned with the reference sequences, including CRFs where appropriate, and examined by neighbour-joining bootstrap analysis, as described by Siepel and Korber (264). The sequence from ML1974 had three breakpoints, and alternated between clade D and clade A1 across the genome (Figure 8). The sequence from ML1076 also was comprised of two parental subtypes, clade C and clade D, but had four breakpoints. The sequence from ML1956 had four breakpoints across the genome, and three parental subtypes: clades A2, D and C. The sequence from ML2000 was also composed of three parental subtypes: clades A1, C, and D and had five breakpoints. The HIV-1 isolate from subject ML1979 displayed the highest level of recombination, with eight breakpoints observed along the genome and three parental subtypes: clades A1, G and C. Some common recombination breakpoints were identified between different viral sequences. Viral sequences derived from ML1076, ML1956 and ML1979 all had breakpoints at the end of *gag p7*, between nucleotide positions 2035-2066 (numbered relative to the HXB2 genome). The sequences from ML1974 and ML1979 both had breakpoints in *pol integrase*, between positions 4526-4553. In totality, however, each sequence displayed a unique and complex recombination pattern.

1.7 Association of recombination with epidemiological markers of sexual exposure

To determine if recombination is more prevalent in subjects that have higher exposure to HIV, epidemiologic measures of sexual exposure were correlated with the presence of recombination (Table 6). No exposure characteristics significantly differed between women infected with recombinant virus, compared to those who were infected with non-recombinant virus, suggesting that exposure and recombination may not be related. This

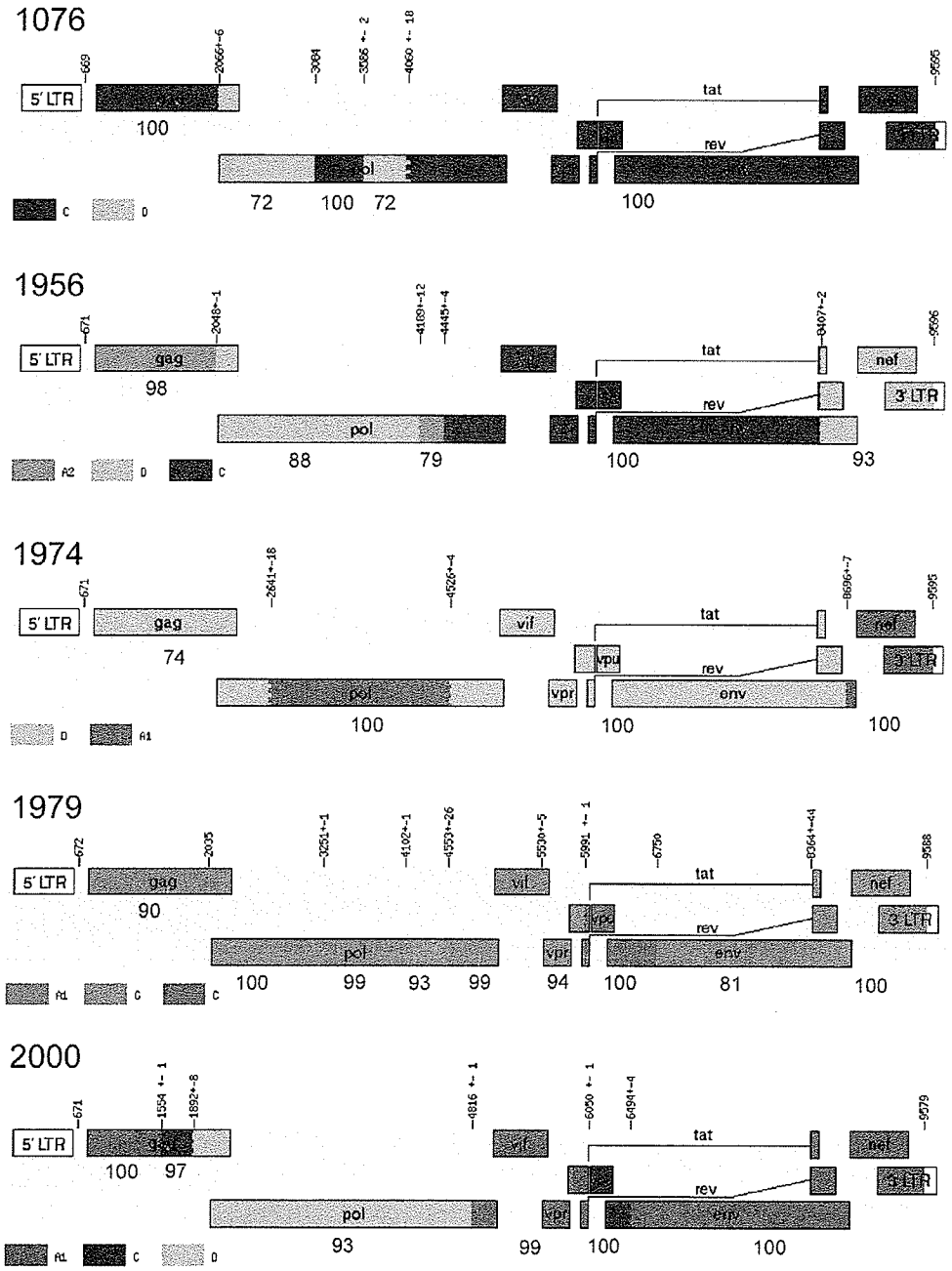


Figure 8. Representation of breakpoints in recombinant full-length HIV-1 proviral sequences. Genome segments are coloured according to parental clade (see legend below each sequence). Numbers below the coloured sections indicate the level of bootstrap support. Numbers above the junction between colours represent the sequence start and stop locations, and the recombination breakpoints. All numbers are given as HXB2 co-ordinates. The figure was generated using Recombinant HIV-1 Drawing Tool (www.hiv.lanl.gov).

Table 6. Patient epidemiological and HIV-1 clade data

Patient	Sample Date (mm/dd/yy)	Birth Year	Years CSW[a]	Sex Partners/ Day	Condom Usage[b]	HIV Clade
ML752	11/27/01	1957	22	4	often	A1
ML1901	06/13/01	1967	8	4	always	A1
ML1945	04/18/01	1959	5	5	never	A1
ML1990	07/10/02	1964	8	3	often	A1
ML2014	04/22/02	1974	1	1	always	A1
ML1076	08/28/01	1967	14	3	always	C/D
ML1956	08/15/01	1964	8	3	always	A2/C/D
ML1974	03/06/02	1962	5	3	often	A1/D
ML1979	03/06/02	1969	7	3	always	A1/C/G
ML2000	03/19/02	1969	7	2	always	A1/C/D

[a]Number of years prior to sample collection the subject was involved in active sex work.
[b]The participants were asked to rank their condom usage habits as either always, often, seldom or never using condoms.

study, however, is based upon only a limited number of sequences and likely lacks the power to determine significant differences in exposure between subjects infected with recombinant HIV-1 compared to subjects infected with non-recombinant virus. Indeed, the comparison of two populations of five subjects, assuming a large difference in the proportion of recombinants infecting the two groups of 0.8 and 0.2 yields a power of only 0.2 (i.e. unlikely to detect significant differences) (156). To have the power of the study be greater or equal to 0.8, assuming the same proportions of recombinant virus, we would need to compare two populations of thirteen or more subjects each.

1.8 Summary

Overall, there are disproportionately few published full-length HIV sequences from Kenya. This was the second population survey describing full-length HIV-1 sequences from Kenya, and the first in a high-risk Kenyan population, and showed that the level of HIV recombination in this population was slightly higher than, though similar to, the rate of recombination described in the general Kenyan population. Despite this higher rate of recombination, the HIV sequences from this high-risk population were not genetically distinct from other HIV sequences from Kenya. The work from this chapter has been published in the journal *AIDS Research and Human Retroviruses* (151).

2. High Prevalence of Genetically Similar HIV-1 Recombinant Viruses among Infected Sex Workers

2.1 Rational

The advantage of full-length HIV sequence analysis for recombination detection is that one can determine the presence of breakpoints across the entire length of the genome, and thus detect all or most breakpoints. However, the advantage of partial genome sequencing is that a smaller sequence amplicon allows one to expand the study to include more subjects. Additionally, full-length genome can be difficult to amplify from subjects where the provirus is integrated at a low level. As the study in the previous section lacked the power to detect differences in recombination frequency between high-risk subjects, a new study was undertaken with the approach of examining a smaller genome segment and a greater number of subjects.

2.2 Hypothesis

The previous section indicated that the ML cohort, a high-risk population, had a higher frequency of recombination than what has been previously reported for a Kenyan population. We wanted to determine if this is true when we compare the ML cohort, a high-risk population to the MCH cohort, a low-risk population, by examining partial HIV-1 sequence. We furthermore wished to increase our sample size and re-address our hypothesis that there will be an increased prevalence of recombinant HIV in individuals with higher exposure.

2.3 Objectives

- Generate partial HIV-1 proviral sequences from 240 individuals from the ML and MCH cohorts
- Analyse the sequences to determine clade and characterize any recombination
- Correlate HIV-1 recombination with participation in sex work, duration of sex work and other epidemiological factors associated with HIV exposure

2.4 Study outline

PBMCs were collected from 215 HIV-infected women from the ML cohort and 25 HIV positive women from the MCH cohort and used for isolating genomic DNA. Subjects were randomly selected based on sample availability. Individuals included both new enrolees to the cohort and patients that had been followed for a number of years. A 590 nucleotide region of the HIV-1 provirus that included the *vpu* gene and the 5' end of the *env* gene was amplified with a nested PCR approach, using genomic DNA as a substrate. The amplicon was sequenced and analyzed for recombination with the RIP web-based tool and by neighbour-joining tree generation. A self-reported survey is administered to cohort participants bi-annually; the information collected in these surveys was correlated with presence of recombination.

2.5 Phylogenetic analysis of the proviral sequences

To determine the HIV-1 clade of the virus infecting each subject, the proviral sequences were aligned with the current reference sequences (www.hiv.lanl.gov) and subsequently used to generate a series of neighbour-joining trees (Figure 9). Of the 240 sequences


A
clade A1
94
0.02
clade A2
clade D
clade B
clade C

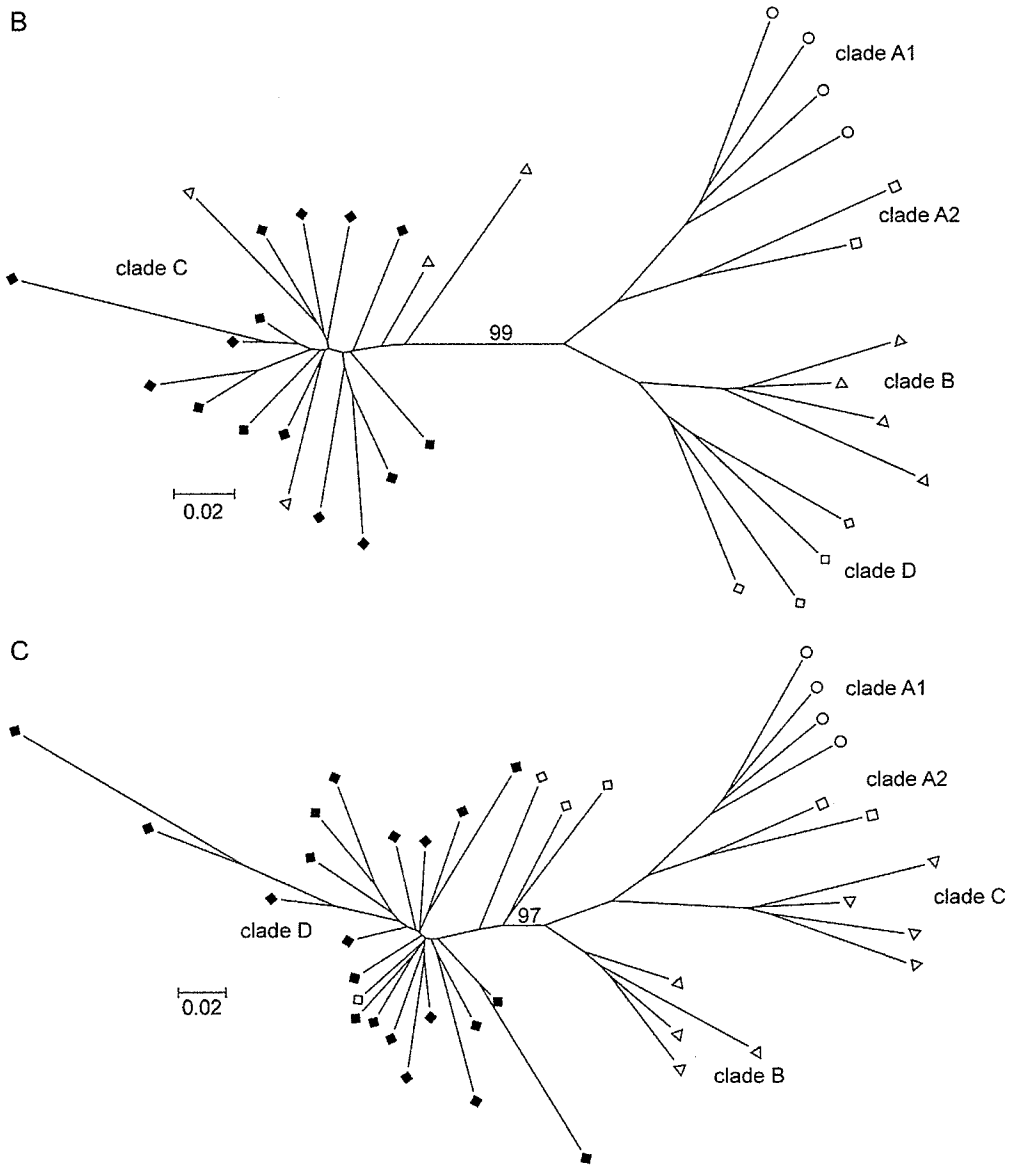


Figure 9. Neighbour-joining trees of proviral HIV-1 sequence segments that cluster tightly with a single clade. The 590 nucleotide long regions were aligned with reference sequences using ClustalW and the phylogenetic trees were created using MEGA 4.0. Sequences generated in this study are labelled with black diamonds. Reference sequences are indicated by open symbols: clade A1 by circles, clade A2 by squares, clade B by triangles, clade C by inverted triangles and clade D by diamonds. The bootstrap value that supports the clade of interest is marked. Scale of genetic distance is indicated. A (previous page). Sequences that clustered strongly with A1 reference sequences. B. Sequences that clustered strongly with clade C reference sequences. C. Sequences that clustered strongly with clade D reference sequences.

examined, 167 clustered with the clade A1 reference sequences with a bootstrap value of 94. 15 sequences clustered within the clade C references with a bootstrap value of 99, while 21 sequences clustered within the clade D references with a bootstrap value of 97. Thirty-seven sequences were recombinant in the region examined (Figure 10a).

The RIP 3.0 recombination identification tool (www.lanl.gov) was employed to characterize the recombination patterns. The sequences were excised at the putatively identified breakpoints, realigned with the reference sequences and examined by neighbour-joining bootstrap analysis, as described by Siepel and Korber (264). The recombinant sequences were comprised of the prominent clades circulating in Kenya: A1, C and D, as well as A2. The majority of recombinant isolates contained clades A1 and D as parental sequences (Figure 10b).

2.6 Characterization of recombination

Examination of the 37 recombinant HIV-1 sequences revealed that most isolates (n = 33) had shared recombination patterns and thus formed groups. Recombination groups were composed of sequences that shared a common breakpoint between common clades. Group 1 was composed of seven sequences that had a breakpoint between clades D and C at 6217 ± 13 (HXB2 coordinates) (Table 7 and Figure 11). Group 2 was composed of eleven sequences that had a breakpoint between clades D and A1 at 6195 ± 20. Group 3 was composed of three sequences that had a breakpoint between clades C and A1 at 6201 ± 4. Group 4 was composed of seven sequences that had a breakpoint between clades C and A1 at 6368 ± 53. Group 5 was composed of five sequences that had a breakpoint

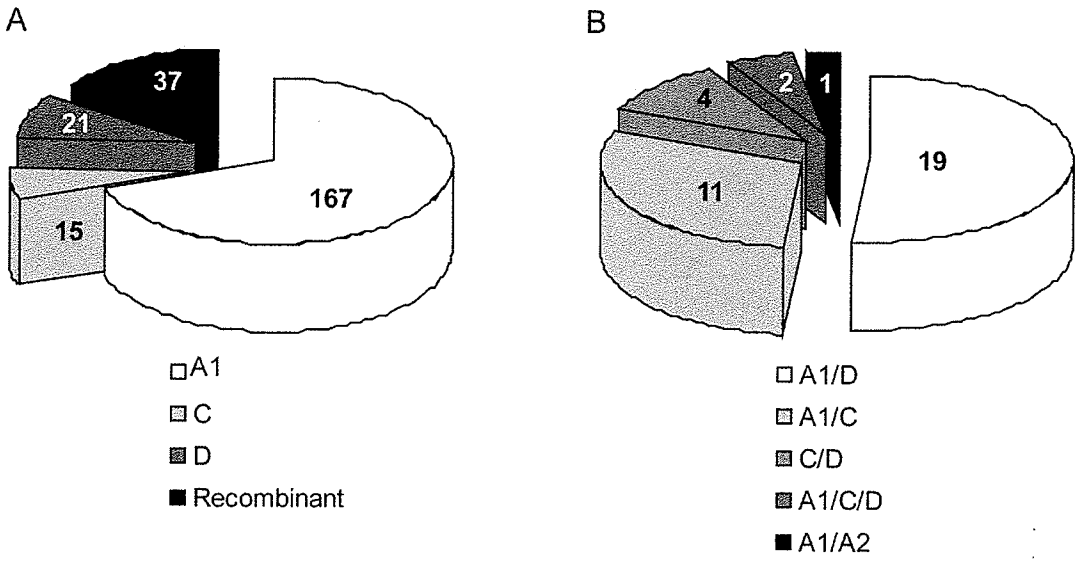


Figure 10. Distribution of HIV-1 subtype and recombination for 240 HIV-1 proviral segments. A. The proportion of the 240 examined sequences classified as each clade, or as recombinant, is displayed in a pie chart. B. The subtype composition of the 37 identified recombinant sequences is displayed in a pie chart.

Table 7. Clade breakpoints for 37 identified recombinant HIV-1 sequences

Group[a]	Patient	Start[b]	Stop	Clade	Bootstrap[c]
1a	ML7	6050	6211	D	98
		6212	6362	C	99
		6363	6475	D	89
	ML215	6001	6204	D	89
		6205	6467	C	99
	ML1451	6001	6209	D	99
		6210	6349	C	99
		6350	6545	D	94
	ML1999	5984	6209	D	100
		6210	6327	C	99
1b	ML888	6029	6213	D	92
		6214	6327	C	98
		6328	6554	A1	100
	ML2170	5984	6207	D	90
		6208	6360	C	99
		6361	6576	A1	99
1c	ML2019	5984	6115	C	98
		6116	6230	D	97
		6231	6364	C	99
		6365	6576	D	90
2a	ML210	6001	6200	D	85
		6201	6549	A1	100
	ML602	6010	6204	D	99
		6205	6556	A1	100
	ML825	6019	6175	D	84
		6176	6541	A1	99
	ML1155	6025	6204	D	94
		6205	6516	A1	100
	ML1411	6002	6214	D	99
		6215	6557	A1	100
	ML1419	5984	6209	D	98
		6210	6416	A1	99
	ML1649	6002	6211	D	98
		6212	6555	A1	100
	ML1941	5984	6194	D	95
		6195	6576	A1	100
	ML2020	5988	6213	D	94
		6214	6539	A1	99
2b	ML293	6017	6102	A1	93
		6103	6200	D	79
		6201	6549	A1	100
	ML1443	6036	6110	A1	98
		6111	6201	D	96
		6202	6268	A1	96
3	ML313	6017	6205	C	99
		6206	6314	A1	83
	ML1335	6012	6197	C	99
		6198	6545	A1	100
	MCH5384	5992	6206	C	100
		6207	6554	A1	100
4	ML264	5999	6332	C	100
		6333	6556	A1	99
	ML1008	6006	6353	C	100
		6354	6547	A1	99
	ML1203	5984	6421	C	100
		6422	6574	A1	99
	ML1322	6007	6413	C	100
		6414	6544	A1	99
	ML1979	5984	6361	C	100
		6368	6576	A1	78
	ML2000	5984	6367	C	100
		6368	6576	A1	99
	ML2227	5999	6314	C	100
		6315	6554	A1	100
5a	ML790	5999	6322	D	100
		6323	6555	A1	99
	ML1337	5984	6318	D	100
		6319	6576	A1	100
	ML1578	5984	6342	D	100
		6343	6576	A1	99
	ML1660	6011	6336	D	100
		6337	6489	A1	96
5b	ML616	6004	6115	A1	98
		6116	6347	D	99
		6348	6555	A1	100
orphan	ML216	6011	6113	D	98
		6114	6553	A1	100
	ML1391	6030	6362	A1	94
		6363	6528	C	62
	ML1818	5984	6327	A1	99
		6328	6574	D	99
	ML2185	5995	6211	A2	75
		6212	6487	A1	91

[a]Groups are composed of sequences that shared a common breakpoint between common clades. Subgroups are assigned based on additional recombination breakpoints.

[b]All numbering is based on HXB2 coordinates.

[c]Bootstrap values are a percentage of 500 replicates. Sequences were aligned with reference sequences from clades A1, A2 (for ML2185 only), C and D.

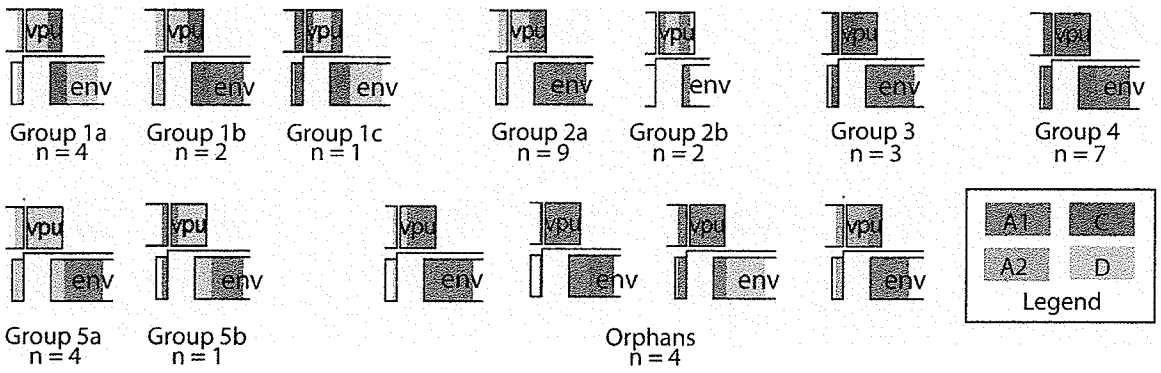


Figure 11. Representation of recombination identified in the examined 590 nucleotide *vpu/env* HIV-1 proviral fragment. Coloured segments represent the parental clades (see legend in lower right corner). A representative sequence for each group/subgroup is shown. Groups were composed of sequences that shared a common breakpoint between common clades. Subgroups were assigned based on additional recombination breakpoints. The figure was generated using Recombinant HIV-1 Drawing Tool (www.hiv.lanl.gov).

between clades C and A1 at 6333 ± 15. Subgroups (example 1a, 1b, 1c) were assigned based on additional recombination breakpoints (Table 7 and Figure 11). Only four orphan (unique) recombination patterns were identified.

In addition to the groups being related based on shared recombination breakpoints, some groups were additionally phylogenetically related. Group1 isolates clustered most closely with clade C viruses, but when examined in combination with reference sequences and the clade C viruses identified in this study, they formed a distinct cluster, with recombinant isolates from group1a and 1b forming a sub-cluster with a bootstrap value of 91 (Figure 12a). This suggests that group1a and 1b isolates are related, and although they do not have identical breakpoints, they may be derived from the same ancestor virus. Similarly, Group3 isolates clusters most closely with clade C viruses, and when examined with reference sequences and clade C viruses identified in this study, they also formed a distinct cluster, with two of the three isolates forming a sub-cluster with a bootstrap value of 99 (Figure 12b). These recombination groups are also distinctive when examined together, with reference sequences and clade C isolates from this study (Figure 12c). The sample dates of these phylogenetically related recombinant sequences were examined, to determine if they were also temporally related (Table 8). The sample dates within groups are dispersed by eight to sixteen years, while the first HIV positive dates within the groups are dispersed by five to eighteen years. Thus, neither the sample dates nor the first HIV positive serology dates for the subjects formed clusters by recombination group, suggesting that the viruses were not directly transmitted between the individuals sampled.

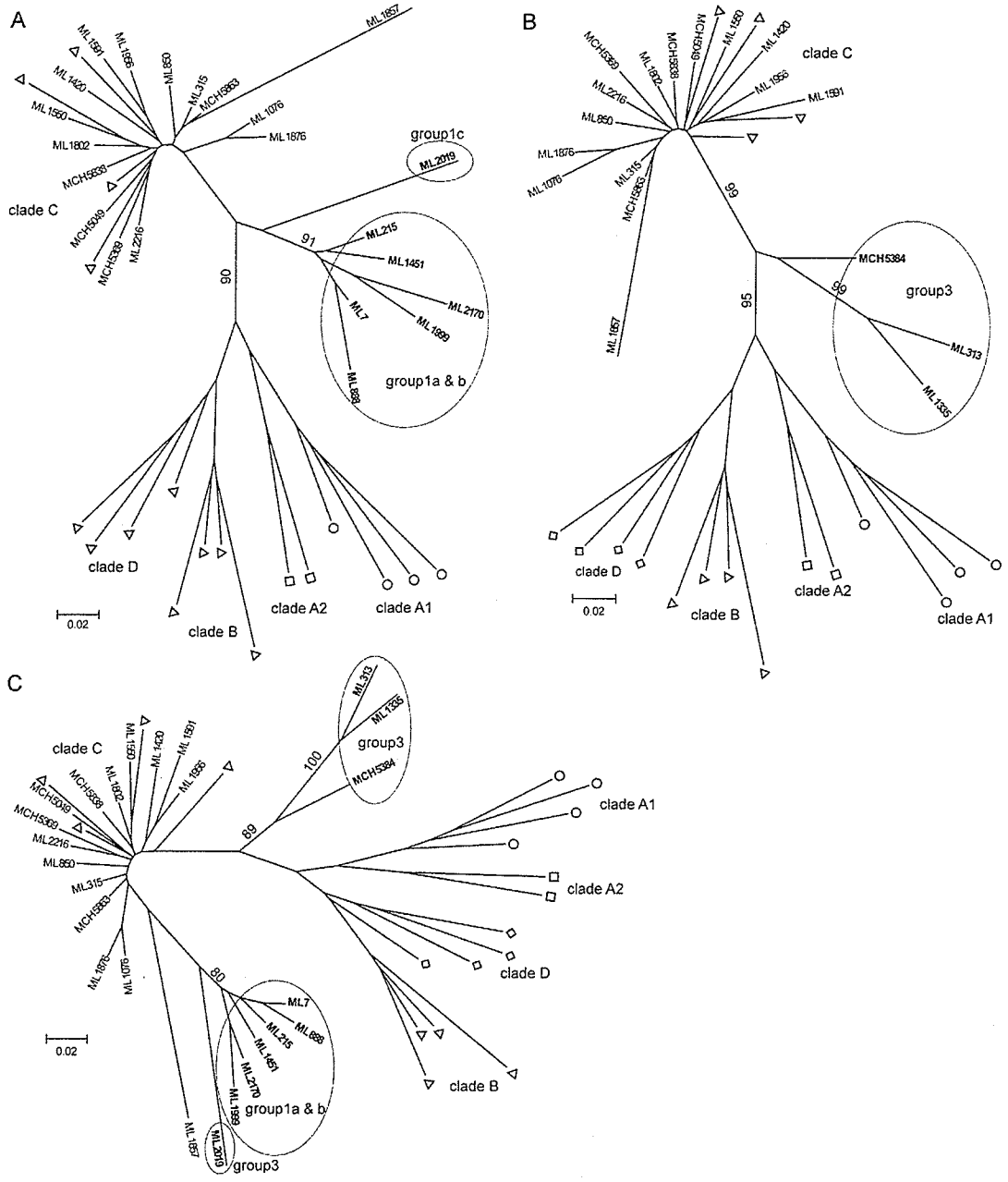


Figure 12. Neighbour-joining trees of proviral HIV-1 sequences from recombinant groups that showed phylogenetic relatedness. A. Group1 recombinant isolates were aligned and examined with current reference sequences (www.lanl.hiv.gov) and clade C isolates identified in this study. B. Group 3 recombinant isolates were aligned and examined with current reference sequences (www.lanl.hiv.gov) and clade C isolates identified in this study. C. Group 1 and group 3 recombinant isolates were aligned and examined with current reference sequences (www.lanl.hiv.gov) and clade C isolates identified in this study. Recombinant isolates are indicated in bold text. Reference sequences are labelled with circles (clade A1), squares (clade A2), triangles (clade B), inverted triangles (clade C) and diamonds (clade D). Bootstrap values of interest are indicated. Scale of genetic distance is shown.

Table 8. HIV-infection dates and sampling dates do not correlate for subjects infected with phylogenetically-related recombinant HIV-1 sequences

Group	Sample	Last HIV- Date[a]	First HIV+ Date	Sampled Date[b]
1a	ML7	-	Mar. 28, 1985	Sep. 7, 1987
	ML215	-	Mar. 7, 1985	May 15, 1987
	ML1451	-	Feb. 11, 1992	May 28, 1992
	ML1999	-	Mar. 18, 2002	Mar. 18, 2002
1b	ML888	-	Jul. 7, 1987	Jul. 17, 1998
	ML2170	Apr. 12, 2002	Jun. 3, 2003	Jun. 3, 2003
1c	ML2019	-	May 23, 2002	May 23, 2002
3	ML313	Mar. 14, 1985	Apr. 21, 1985	Jul. 16, 1987
	ML1335	-	Dec. 6, 1990	Aug. 11, 1995
	MCH5384	-	-	Feb. 20, 1990

[a]This date is only available for subjects who were enrolled into the ML or MCH cohorts when they were HIV seronegative.

[b]This is the date of the sample used for DNA isolation and subsequent sequencing.

2.7 *Association of clade/recombination with epidemiological characteristics*

In order to determine if epidemiological characteristics were associated with infecting clade or recombination status, the distribution of these characteristics was examined using univariate analysis. These variables included if the subject was HIV negative upon enrolment into either of the high-risk ML or low-risk MCH cohorts, the subjects' $CD4^+$ cell count and viral load at the time the sample was collected, age, year of birth and whether the subject participated in sex work. For those that participated in sex work, additional data was examined: years of sex work, years of follow up in the cohort, number of sex partners per day and condom usage (Table 9). The average $CD4^+$ count was 326 cells/mL and the average viral load was $4.75\log_{10}$. For the 215 subjects that actively participated in commercial sex work the average number of sexual partners was 4.4 per day and the average duration of sex work was 7.7 years. Most of these epidemiological characteristics were not significantly associated with HIV-1 clade/recombination. However, participation in sex work was significantly associated with HIV-1 clade distribution ($p = 0.038$, Chi-square), such that C clade viruses were more likely to be isolated from subjects that did not participate in sex work, while recombinant viruses were more likely to be isolated from subjects that did participate in sex work (36 of the 215 HIV isolates from sex workers were recombinant, while 1 of the 25 isolates from non-sex workers were recombinant). Additionally, age was significantly distributed between the clades ($p = 0.048$, ANOVA), with the highest average age (33.5 years) found in subjects infected with recombinant virus. However, a two-factor ANOVA examining age, clade/recombination and sex work revealed that while age and sex work are linked ($p = 7.36 \times 10^{-7}$), age and clade/recombination in fact are not ($p =$

Table 9. Association of epidemiological characteristics for 240 examined subjects with infecting HIV-1 clade

	Total	Clade A1	Clade C	Clade D	Recombinant	*p* value
n	240	167	15	21	37	
$CD4^+$ count, cells/μL[a]	326	330 (10-1488)	305 (160-757)	296 (53-537)	332 (10-948)	0.932
Viral load, copies/mL[a]	4.75 $\log_{10}$	4.64 $\log_{10}$ (2.1-6.3)	4.20 $\log_{10}$ (2.1-4.8)	5.10 $\log_{10}$ (2.1-5.8)	4.85 $\log_{10}$ (2.1-5.9)	0.489
HIV negative enrolment[b]	58	40	2	4	12	0.450[d]
Age[a]	32.3	32.7 (18 – 58)	28.4 (19 – 38)	30.3 (16 – 43)	33.5 (18 – 59)	0.048[NS]
Year of birth[a]	1963	1962 (1942 – 1978)	1966 (1955 – 1976)	1964 (1953 – 1976)	1964 (1944 – 1982)	0.066
Low risk group	**25**	**16**	**4**	**4**	**1**	**0.038[d,*]**
Sex worker group	**215**	**151**	**11**	**17**	**36**	
Years sex work[a]	7.7	7.8 (1 – 20)	6.7 (1 – 18)	7.7 (1 – 20)	7.8 (1 – 20)	0.942
Years in cohort[a]	4.1	4.2 (0 –15)	3.3 (0 – 6)	3 (0 – 8)	4.3 (0 – 15)	0.589
Sex partners/day[a]	4.4	4.5 (1 – 10)	4.7 (2 – 10)	4 (1 – 10)	4.3 (1 – 10)	0.771
Year started sex work[a]	1987	1987 (1973 – 2003)	1989 (1975 – 1997)	1989 (1977-1999)	1988 (1976 – 2002)	0.595
Condom usage[c]	often/always	often/always	always	often/always	often/always	0.058[d]

[a]The average and range (in brackets) for each group is shown. Association was tested with a one-way ANOVA.
[b]These subjects become HIV positive after enrolment into the cohort.
[c]The participants were asked to rank their condom usage habits as either always, often, seldom or never using condoms. The average for each group is shown.
[d]Association was tested with a Chi-Square test.
[NS]Follow-up testing with a two-way ANOVA revealed that age was not associated with clade independent of involvement in sex work.
*This *p* value is statistically significant (< 0.05).

0.56). Thus, commercial sex workers in this sample set are older then women who do not engage in high-risk activity and this is the likely reason that age was initially identified as being associated with clade/recombination. This leaves participation in sex work as the only factor examined in this study that was associated with clade/presence of recombination.

2.8 Summary

This work contributes to the HIV-1 sequence data available from Kenya and confirms previous studies that found a difference in recombination prevalence in high-risk compared to low-risk populations. Additionally, these findings suggest that many unique recombinant sequences may in fact represent as yet unidentified ciruculating recombinant forms of HIV. The work from this chapter has been published in the journal *AIDS Research and Human Retroviruses* (152).

3. HIV-1 Proviral Hypermutation correlates with $CD4^+$ Count

3.1 Rationale

Proviral HIV sequence provides a wealth of information about the infecting virus. As shown in the previous sections, it can be used for determining viral subtype and discerning and characterizing recombination. Proviral HIV also provides an archive of hypermutation activity (142). Hypermutation caused by the various APOBEC3 proteins has been recently described as an important mechanism of viral restriction and its role in disease progression is currently being debated in the literature.

3.2 Hypothesis

APOBEC-mediated hypermutation leaves a characteristic sequence pattern; we hypothesize that examination of the 240 proviral HIV-1 sequences generated in the previous section will reveal the presence of hypermutation. We furthermore hypothesize that as APOBEC works as a viral restriction mechanism, hypermutation will be associated with disease progression, as measured by $CD4^+$ count and viral load. We believe that increased levels of hypermutation will be associated with mutations in the autologous viral protein Vif, which normally counteracts APOBEC3.

3.3 Objectives

- Identify APOBEC-mediated hypermutation in proviral HIV-1 sequences
- Confirm that the genomic area examined (*vpu/env*) is sensitive for detecting hypermutation by comparing hypermutation levels in autologous *gag* for a subset of viral sequences

- Characterize the hypermutation as being the result of APOBEC3G and/or APOBEC3F activity
- Examine the clonal diversity of the hypermutation
- Determine if hypermutation correlates with $CD4^+$ count and viral load
- Sequence *vif* from a subset of proviruses and determine if there are mutations that suggest decreased anti-APOBEC activity in the Vif from hypermutated proviruses

3.4 Study outline

The 590 nucleotide proviral sequences obtained from 240 HIV positive women, described in the previous section, were examined for hypermutation by quantifying the proportion of adenine nucleotide residues. These proviral sequences were compared to corresponding RNA sequences from the same region for a subset of patients. Additionally, using the web tool Hypermut, the sequences were examined for hypermutation by comparing the patient sequences to a consensus and examining the context of G to A transitions. Upon identifying a subset of sequences as being significantly hypermutated, multiples clones of these samples were sequenced to access the level of interclonal variability. The level of hypermutation was correlated with $CD4^+$ count and plasma viral load to determine if a relationship exists between hypermutation and disease progression. Vif sequence was compared between a subset of viruses that had extensive proviral hypermutation and a subset of viruses with low levels of proviral hypermutation.

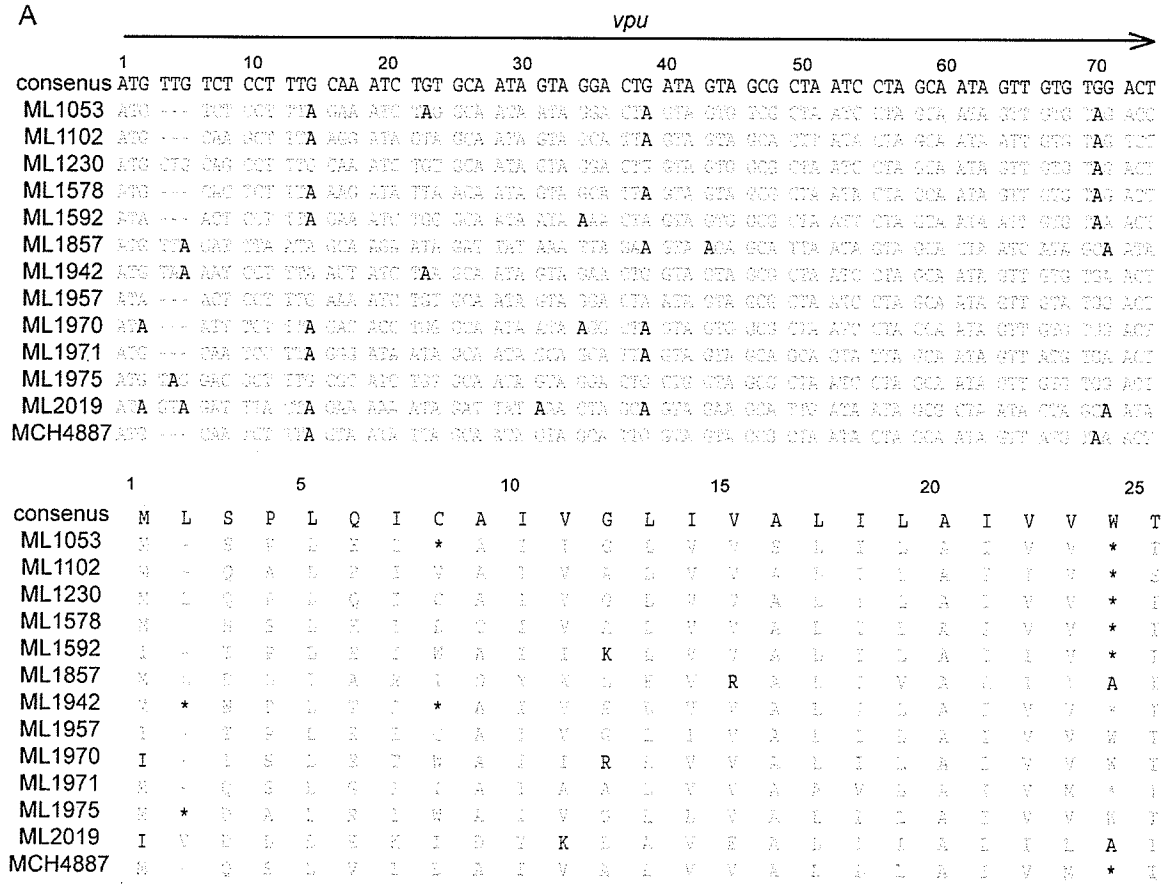
3.5 Proviral sequences contain premature stop codons due to hypermutation

HIV-1 sequence diversity was assessed by examining the predominant proviral *vpu/env* DNA sequence from 240 subjects from the ML and MCH cohorts, as described in the previous section. Fifteen of the 240 sequences had significant amino acid mutations in Vpu, such as missing start codons and premature stop codons, which would presumably prevent protein translation and lead to a lethal viral phenotype. Twenty sequences had significant amino acid mutations in Env and thirteen of these had significant amino acid mutations in both Vpu and Env (Figure 13 A and B, respectively). Previous studies have demonstrated that hypermutation can cause premature stop codons, especially at tryptophan codons, where the transition of a G to A causes the Trp codon (UGG) to change to a stop codon (UAG, UAA or UGA) (50,300). To assess the possibility that the identified lethal mutations in Vpu and Env could be the result of hypermutation, the proviral patient sequences were compared to a cohort-specific consensus sequence (Figure 13). G to A nucleotide transitions were noted and determined to be responsible for the majority of observed detrimental amino acid mutations.

3.6 A subset of proviral sequences has elevated adenine proportion

As APOBEC3G and APOBEC3F are known to cause G to A hypermutation in HIV-1 provirus, the possibility of hypermutation due to APOBEC cytidine deamination was investigated by examining the sequences for adenine enrichment (24,163,310,327,336). The proportion of adenine nucleotides in all 240 sequences was determined and revealed a generally symmetrical bell-shaped distribution with a mean adenine proportion of 0.36 and a standard deviation of 0.018 (Figure 14). Indeed, testing for deviations from

A


vpu
1 10 20 30 40 50 60 70
consensus ATG TTG TCT CCT TTG CAA ATC TGT GCA ATA GTA GGA CTG ATA GTA GCG CTA ATC CTA GCA ATA GTT GTG TGG ACT
ML1053
ML1102
ML1230
ML1578
ML1592
ML1857
ML1942
ML1957
ML1970
ML1971
ML1975
ML2019
MCH4887
1 5 10 15 20 25
consensus M L S P L Q I C A I V G L I V A L I L A I V V W T
ML1053
ML1102
ML1230
ML1578
ML1592
ML1857
ML1942
ML1957
ML1970
ML1971
ML1975
ML2019
MCH4887

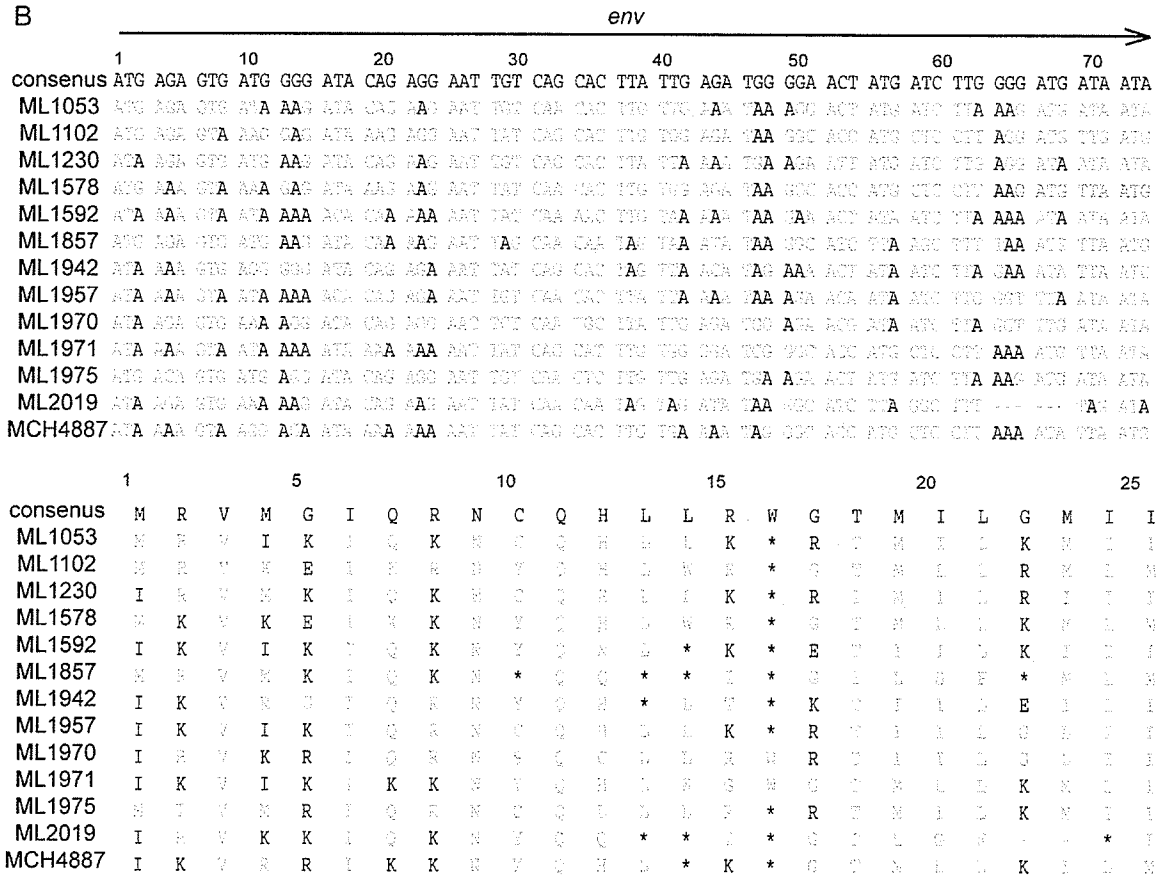


Figure 13. Sequence context of identified G to A HIV-1 proviral hypermutation. A (previous page). The first 75 nucleotides of the *vpu* ORF from thirteen dramatically mutated proviruses. B. The first 75 nucleotides of the *env* ORF from the same thirteen proviruses. In each panel, the top half displays the nucleotide sequence, while the bottom half shows the corresponding amino acid sequence. The consensus sequence for both the DNA and protein was generated based on the 240 proviral sequences examined in this thesis. The sequences are numbered from the HIV-1 *vpu* and *env* start, respectively. The proviral sequence was compared to the corresponding non-hypermutated plasma RNA sequence, where available, or else to a clade-specific consensus. Darkened nucleotide letters in the sample sequences indicate changes from G to A in either a GGD or GAD context, where D is equal to G, A or T (i.e. proposed APOBEC3G or APOBEC3F hypermutation). Darkened amino acid letters in the sample sequences indicate where proposed hypermutation at the nucleotide level caused an amino acid change at the protein level.

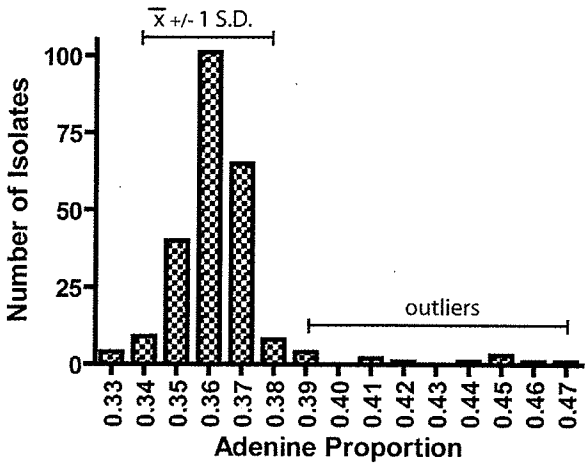


Figure 14. Distribution of proviral HIV-1 adenine proportion in a 590 nucleotide fragment spanning *vpu* and the 5' end of *env* for 240 HIV-1 isolates. The adenine proportion is based on proviral sequence isolated from HIV infected women from two Kenyan cohorts.

Gaussian distribution using the Kolmogorov-Smirnov test shows that the data set as a whole is not normally distributed, but upon removing the top outliers from the analysis, there is no longer any significant deviation. Fourteen sequences had values higher than one standard deviation above the mean. Thirteen of these were also identified as having lethal amino acid mutations in Vpu and Env. This finding provided further evidence that the lethal mutations were the result of hypermutation and merited further investigation.

3.7 Free virus RNA is not hypermutated

To determine if hypermutation was restricted to provirus or could also be observed in the RNA genomes of free virus, sequences were generated from date-matched plasma RNA for a subset of eighteen subjects, including four patients with hypermutated proviral sequence. Phylogenetic analysis of the sequences confirmed that both samples originated from the same individual for fifteen pairs, as they clustered with a bootstrap value greater than 80% of the 500 replicates (Figure 15). For the other three sequence pairs, the sequences did not cluster together tightly, suggesting that the patients may be dually infected with distinct viruses. All the viral RNA sequences had adenine proportions similar to the overall mean observed in the proviruses, and did not show evidence of hypermutation (Figure 16). Furthermore, the plasma RNA from non-hypermutated proviral sequences had similar adenine proportions to the corresponding proviral DNA. However, the four sequences with hypermutated proviral sequences had significantly higher adenine proportions than the matched plasma virus sequence ($p = 0.029$, Mann-Whitney) suggesting that the observed hypermutation was restricted to provirus. This

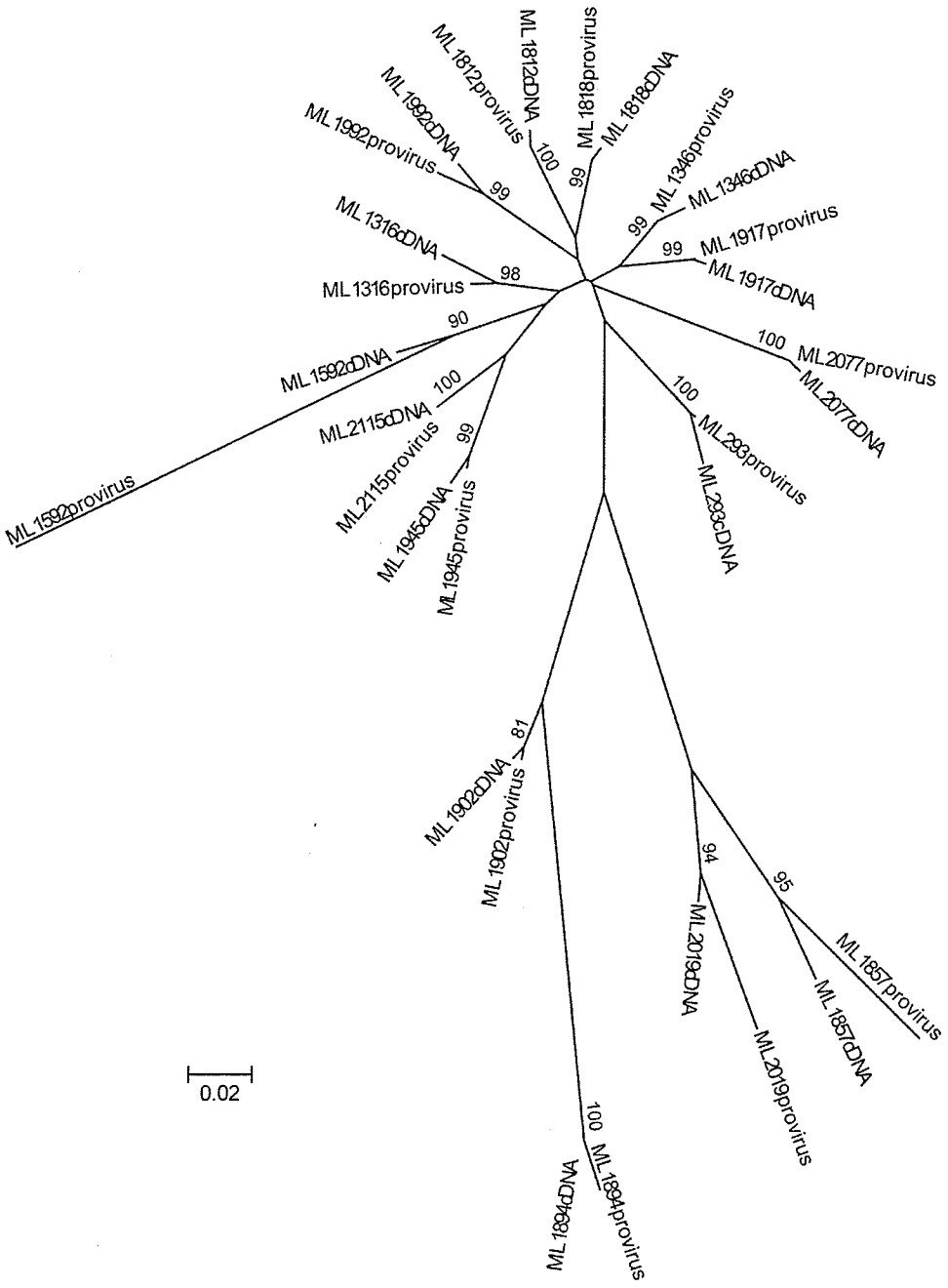


Figure 15: Neighbour-joining tree of matched proviral and cDNA HIV-1 sequences. Fifteen paired samples had the identical 590 nucleotide region sequenced from both provirus and from cDNA generated from plasma virus. The sequences cluster together in a neighbour-joining tree generated in MEGA 4 from a ClustalW alignment. Each pair had a bootstrap value supporting the cluster of greater than 80% (out of 500 replicates). Genetic distance is indicated.

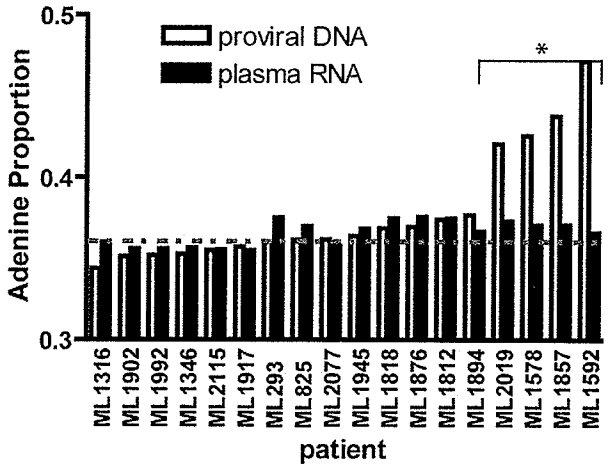
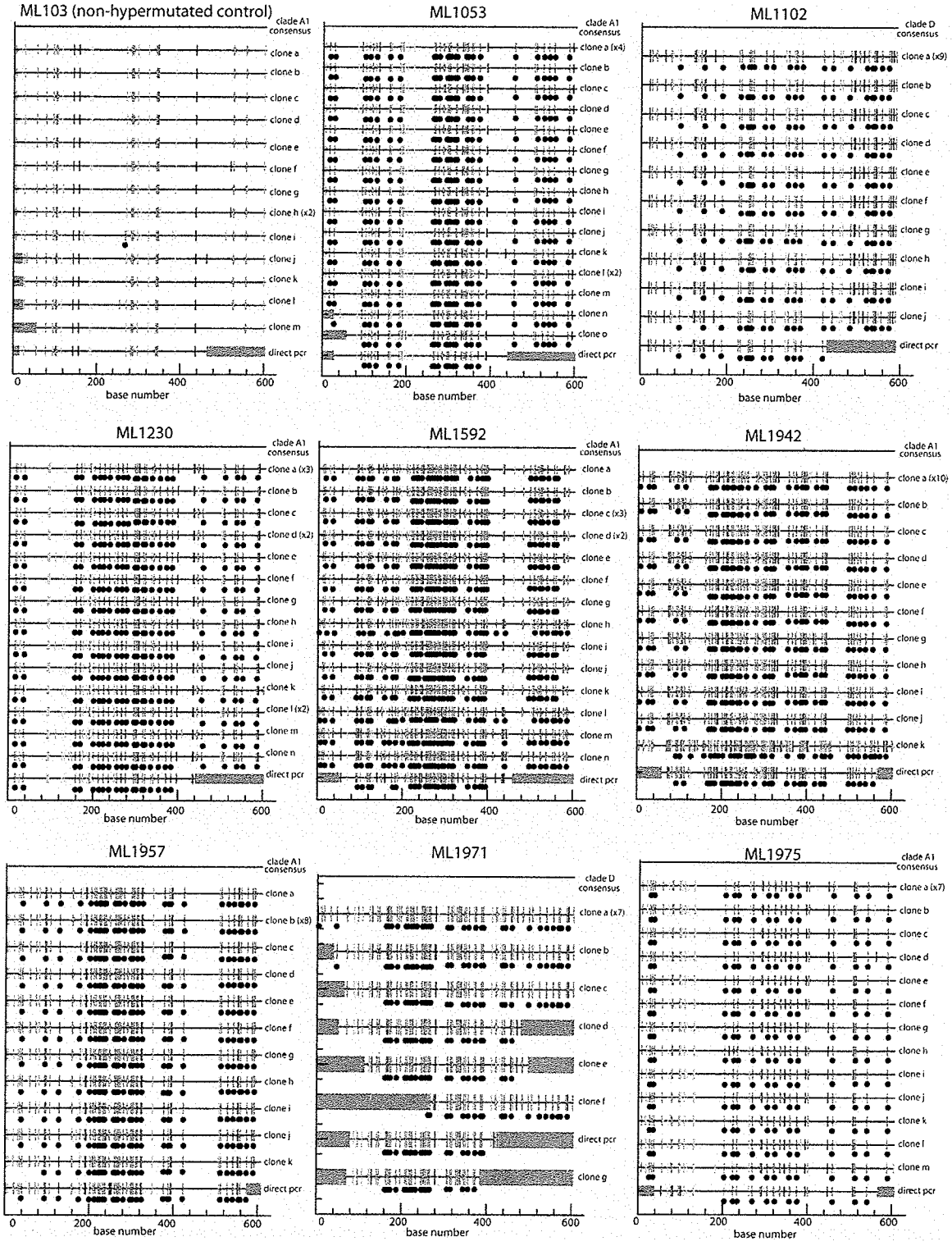


Figure 16. Adenine proportion in proviral HIV-1 sequences and plasma-derived viral RNA sequences in eighteen subject and date matched samples. The same 590 nucleotide HIV-1 genomic region was examined for both proviral and viral RNA sequences. The four sequences indicated with the asterisk had hypermutated proviral *vpu/env* sequences. The mean adenine proportion for the 240 proviral sequences is indicated on the graph with a dashed horizontal line.

finding further indicates that APOBEC3-type hypermutation may be a potential explanation for these findings.

3.8 Directly sequenced PCR products show extensive hypermutation and are representative of proviral sequence

Direct sequencing of a PCR amplicon will reveal the major proviral sequence, but may not be representative of intrapatient diversity. In order to examine proviral diversity within individual subjects, we sequenced thirteen to twenty clones from twenty-three patients, including all thirteen with dramatically hypermutated proviral sequence and a random subset of ten subjects with proviral sequences that were not dramatically hypermutated. The sequence data generated from the multiple clones for each subject confirmed that the detrimental mutations in Vpu and Env were not the result of sequencing or sampling errors, but were in fact representative of the majority of the subjects' proviral population. For the majority of hypermutated samples, the clonal sequences show levels of hypermutation similar to that observed in the sequence generated directly from the proviral PCR product (Figure 17). Fifteen clonal populations showed identical hypermutation patterns to the direct PCR sequence for all the examined clones (including 9 out of the 13 extremely hypermutated sequences). Additionally, two clonal populations had the majority of cloned sequences (85% or greater) match the direct PCR sequence (the hypermutated sequences ML1578 and ML1970). In three clonal populations, half of the clones matched the direct PCR sequence, including the hypermutated sequence ML1857. In another three clonal populations the hypermutation patterns in the clones were generally similar, though not identical, to that observed in the


ML103 (non-hypermutated control)
clade A1 consensus
clone a
clone b
clone c
clone d
clone e
clone f
clone g
clone h (x2)
clone i
clone j
clone k
clone l
clone m
direct pcr
0
200
400
600
base number
ML1053
clade A1 consensus
clone a (x4)
clone b
clone c
clone d
clone e
clone f
clone g
clone h
clone i
clone j
clone k
clone l (x2)
clone m
clone n
clone o
direct pcr
0
200
400
600
base number
ML1102
clade D consensus
clone a (x9)
clone b
clone c
clone d
clone e
clone f
clone g
clone h
clone i
clone j
direct pcr
0
200
400
600
base number
ML1230
clade A1 consensus
clone a (x3)
clone b
clone c
clone d (x2)
clone e
clone f
clone g
clone h
clone i
clone j
clone k
clone l (x2)
clone m
clone n
direct pcr
0
200
400
600
base number
ML1592
clade A1 consensus
clone a
clone b
clone c (x3)
clone d (x2)
clone e
clone f
clone g
clone h
clone i
clone j
clone k
clone l
clone m
clone n
direct pcr
0
200
400
600
base number
ML1942
clade A1 consensus
clone a (x10)
clone b
clone c
clone d
clone e
clone f
clone g
clone h
clone i
clone j
clone k
direct pcr
0
200
400
600
base number
ML1957
clade A1 consensus
clone a
clone b (x8)
clone c
clone d
clone e
clone f
clone g
clone h
clone i
clone j
clone k
direct pcr
0
200
400
600
base number
ML1971
clade D consensus
clone a (x7)
clone b
clone c
clone d
clone e
clone f
direct pcr
clone g
0
200
400
600
base number
ML1975
clade A1 consensus
clone a (x7)
clone b
clone c
clone d
clone e
clone f
clone g
clone h
clone i
clone j
clone k
clone l
clone m
direct pcr
0
200
400
600
base number

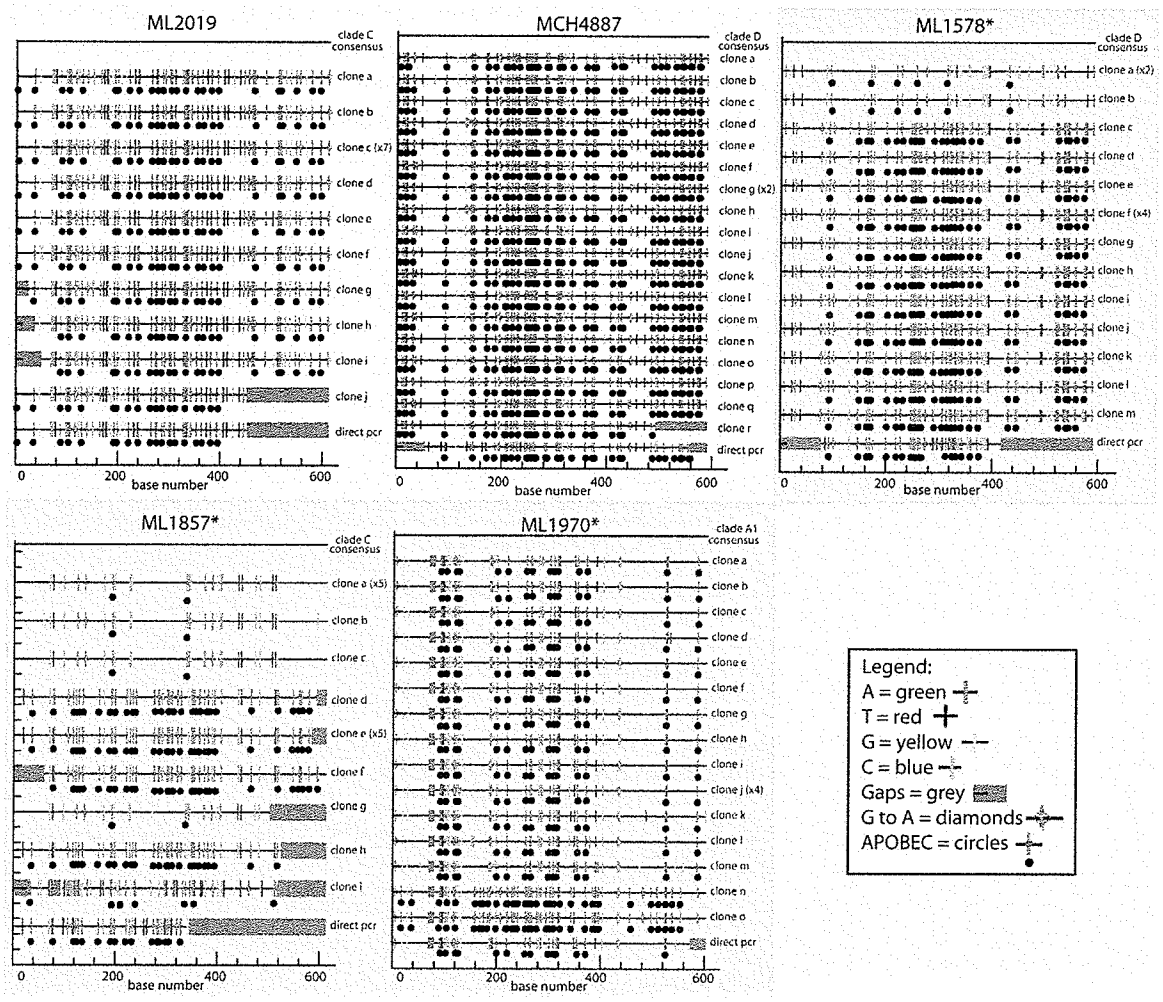


Figure 17. Directly sequenced PCR product and clones are compared to a population-specific, clade-specific HIV-1 consensus sequence. ML103 was selected as a representative non-hypermutated sequence. Changes in the patient sequence compared to the consensus sequence are represented with a coloured bar, as indicated in the legend. The representations were generated with the Highlighter tool available from hiv.lanl.gov. Black circles below hypermutation sites were manually added to ease visualization. The sequences are ordered from top to bottom as most to least similar to the consensus. Identical sequences are shown once, with the number of times the sequence was retrieved listed beside the clone name. Subject identifiers indicated with an asterisk (*) show interclonal diversity.

direct PCR sequence. Overall, the vast majority of samples' clonal and proviral sequences display almost identical patterns of hypermutation; for example, ML103 shows no hypermutation in the direct PCR sequence, and only one site of hypermutation in one of the fourteen examined clones (Figure 17). Similarly, ML1053 shows many hypermutation sites in the direct PCR sequence, and these are shared with and between all nineteen examined clones. However, three out of sixteen clonal sequences from ML1578 and nine out of seventeen clonal sequences from ML1857 showed lower levels of hypermutation than the proviral sequence. Conversely, two out of nineteen clonal sequences from ML1970 show higher levels of hypermutation than the proviral sequence. This data suggests that the direct PCR sequence is generally representative of the predominant proviral sequence; however as expected, in some patients subtle diversity can be observed.

3.9. HIV-1 proviral vpu/env *region is sensitive for detecting hypermutation*

Previous literature has shown that hypermutation does not occur uniformly across the HIV proviral genome, but instead has localized highs and lows. This study examines a single region of the HIV-1 genome as an estimate of overall hypermutation levels. To determine if this region is informative for detecting hypermutation, the levels of hypermutation, as estimated by adenine proportion, were compared in 53 patient and date matched sequences for the 590 nucleotide *vpu/env* region with *gag p24* (693 nucleotides) sequences that were previously published (Figure 18) (227). This comparison included three subjects with high levels of hypermutation in *vpu/env* region. The *vpu/env* region is seen to have a larger range of hypermutation levels than *gag*. Indeed, the adenine levels

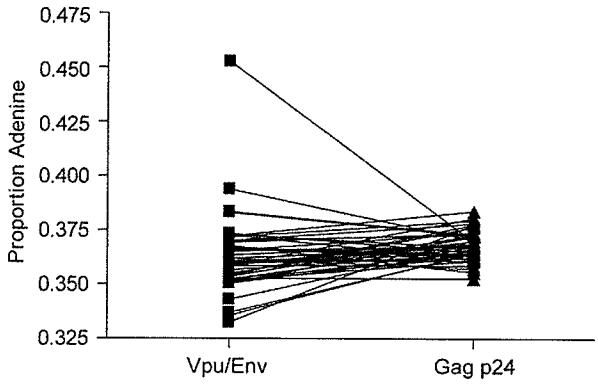


Figure 18. Comparison of hypermutation, as approximated by proportion adenine, in two HIV-1 proviral regions. For fifty-six samples, sequence data was available from both a 590 nucleotide region encompassing *vpu* and the first 349 nucleotides of *env* and the 1502 nucleotide *gag* gene. Each data point represents a single sequence for the given region; the paired sequences from each patient are joined by a line.

for *gag* are tightly clustered, and do not show any outliers that would correspond to hypermutation. This data implies that the *vpu/env* region is appropriate for identifying proviral hypermutation; indeed, it lies in a region of the HIV-1 proviral genome that shows higher levels of predicted hypermutation than *gag* (142,277).

3.10 Observed hypermutation is characteristic of APOBEC3F/3G

All 240 proviral *vpu/env* sequences were examined for hypermutation characteristic of APOBEC3F/G deamination activity using Hypermut 2.0 (hiv.lanl.gov), which examines the sequence connotation of nucleotide changes (247). Using the conservative default settings to examine APOBEC-type G to A hypermutation, we identified seventeen sequences out of our 240 that had significant APOBEC-type hypermutation ($p < 0.05$). Thirteen of the seventeen sequences were previously identified as both having an adenine proportion higher than one standard deviation above the mean and containing lethal mutations in Vpu and Env. Thus, these thirteen represent the patients with the most hypermutated dominant proviral sequences in this population (Table 10).

In addition to examining the 240 sequences for general APOBEC activity with Hypermut 2.0, we also looked for APOBEC3F and APOBEC3G activity individually, as these two enzymes have distinct sequence specificities; APOBEC3G causes GG to AG mutations, while APOBEC3F causes GA to AA mutations; for both enzymes, where the third nucleotide can not be a C (21,141,154,163,328). This approach has not been validated by Hypermut 2.0, so although suggestive, the results are likely less informative than those generated by examining the sequences for general APOBEC3F/G activity. Twenty-four

Table 10: Hypermutation criteria for proviral HIV-1 sequences that had significant hypermutation by at least one assessment

Patient Identifier	Fatal Mutations in Vpu and Env	Adenine Proportion	General Hypermut Significance	Specific Hypermut Significance		All Categories
				A3G	A3F	
MCH4887	Yes [a]	0.4560 [b]	1.012 E-13 [c]	0.00560 [d]	2.709 E-14 [d]	Yes
ML1592	Yes [a]	0.4688 [b]	2.660 E-10 [c]	0.0175 [d]	9.277 E-11 [d]	Yes
ML1942	Yes [a]	0.4510 [b]	6.420 E-10 [c]	0.0042 [d]	5.693 E-10 [d]	Yes
ML1957	Yes [a]	0.4353 [b]	5.911 E-9 [c]	0.0251 [d]	6.44 E-10 [d]	Yes
ML1971	Yes [a]	0.4533 [b]	3.085 E-8 [c]	0.4574	2.039 E-10 [d]	Yes
ML1230	Yes [a]	0.3937 [b]	2.772 E-7 [c]	1.734 E-9 [d]	0.0013 [d]	Yes
ML1578	Yes [a]	0.4212 [b]	1.708 E-5 [c]	1.214 E-5 [d]	0.0014 [d]	Yes
ML1975	Yes [a]	0.3865 [b]	8.894 E-5 [c]	5.915 E-5 [d]	0.0038 [d]	Yes
ML1857	Yes [a]	0.4520 [b]	1.425 E-4 [c]	1.456 E-4 [d]	0.0038 [d]	Yes
ML2019	Yes [a]	0.4148 [b]	6.901 E-4 [c]	1.361 E-7 [d]	0.1554	Yes
ML1053	Yes [a]	0.3932 [b]	0.0012 [c]	6.704 E-5 [d]	0.0644	Yes
ML1102	Yes [a]	0.3828 [b]	0.0029 [c]	1.466 E-4 [d]	0.1321	Yes
ML1970	Yes [a]	0.3861 [b]	0.0058 [c]	0.00350 [d]	0.0633	Yes
ML2209	No	0.3635	0.0030 [c]	0.0910	0.0017 [d]	No
ML1649	No	0.3605	0.0113 [c]	0.0494 [d]	0.0228 [d]	No
ML1903	Only Env	0.3784	0.0153 [c]	0.0021 [d]	0.2276	No
ML1419	No	0.3643	0.0287 [c]	0.2019	0.0243 [d]	No
ML1992	No	0.3519	0.0508	1	0.0149 [d]	No
ML602	No	0.3596	0.0619	0.0397 [d]	0.2397	No
ML1155	No	0.3727	0.0764	0.5218	0.0356 [d]	No
ML1894	Only Vpu	0.3728	0.0982	0.0277 [d]	0.4758	No
ML2115	No	0.3551	0.1039	1	0.0268 [d]	No
ML790	No	0.3513	0.1076	0.0229 [d]	0.5607	No
ML49	No	0.3539	0.1163	1	0.0453 [d]	No
MCH5736	Only Vpu	0.4061 [b]	0.3199	0.1815	0.5891	No
Total meeting criteria:	13	14	17	24		13

[a] These samples had mutations such as missing start codons and premature stop codons in the proviral ORFs examined.

[b] Adenine proportion was higher than one standard deviation above the mean (i.e. > 0.382751).

[c] Hypermut 2.0 analysis for general APOBEC-type hypermutation (default settings of G to A substitutions with a downstream context of RD, where R is either A or G and D is A, T or G) gave a significant p value (<0.05).

[d] Hypermut 2.0 analysis for specific APOBEC3G (settings of G to A substitutions with a downstream context of GD) or APOBEC3F (settings of G to A substitutions with a downstream context of AD) hypermutation gave a significant p value (<0.05).

sequences were identified as having significant hypermutation characteristic of APOBEC3F and/or APOBEC3G activity. All seventeen sequences previously identified as containing general APOBEC-type hypermutation by Hypermut 2.0 also had significant specific hypermutation for APOBEC3F and/or APOBEC3G, suggesting that the observed hypermutation was due to one or both of these enzymes (Table 10).

To further examine the role of these enzymes in the thirteen proviral sequences with the most pronounced hypermutation, we examined the dinucleotide sequence connotation of the G to A hypermutation. Figure 19 shows that most of the G to A nucleotide changes occurred where the original sequence was a GA or a GG, suggestive of APOBEC3F or 3G involvement, respectively. Comparatively fewer changes occurred at GC or GT dinucleotides. It is interesting to note that in most patients a high proportion of G to A nucleotide changes occurring at a GG dinucleotide context co-occurs with a comparatively lower proportion of changes occurring at a GA dinucleotide context. The reverse is also true. In patient ML1970, however, the proportions of G to A nucleotide changes occurring at GG and GA dinucleotides are similar. This suggests that in the majority of subjects, either APOBEC3G or APOBEC3F shows dominant hypermutation activity.

3.11 Proviral hypermutation is associated with increased $CD4^+$ counts

To ascertain if the observed hypermutation has clinical significance, the subjects' $CD4^+$ counts, measured at the time the blood samples were collected, were examined as a marker for disease progression. $CD4^+$ counts from the seventeen patients with significant

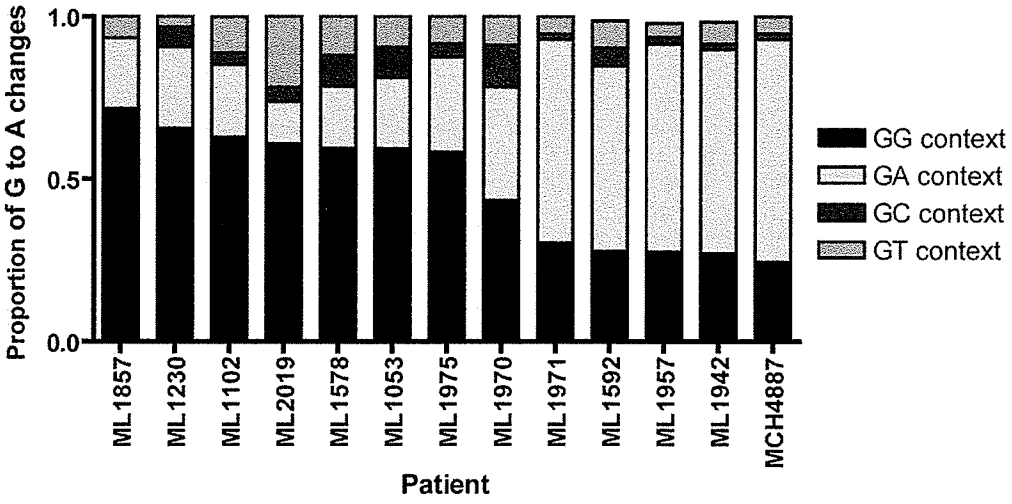


Figure 19. Dinucleotide context of G to A hypermutation in thirteen patients with significant levels of HIV-1 proviral hypermutation. The proportion of G to A hypermutation in each proviral sequence occurring at the dinucleotides GG, GA, GC and GT is indicated. Hypermutation context was determined through comparison of subject sequences to a clade-specific, population-specific consensus, using Hypermut 2.0.

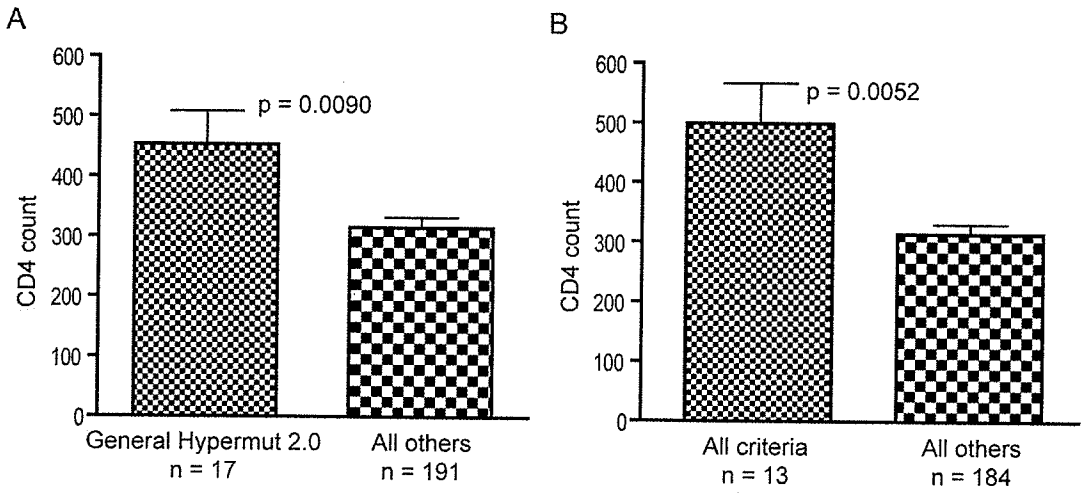


Figure 20. Association of higher $CD4^+$ count with hypermutated HIV-1 proviral sequences. $CD4^+$ counts are compared between patients with hypermutated provirus and patients without hypermutated provirus. A. Hypermutation was defined as significant general APOBEC cytidine deaminase activity, as determined by Hypermut 2.0. B. Hypermutation was defined as sequences with detrimental mutations in Vpu and Env, proviral adenine proportion greater than one standard deviation above the mean, and significant Hypermut 2.0 hypermutation. Groups are compared by a Mann-Whitney test. The bar height corresponds to the mean and the error bar represents the standard error of the mean.

hypermutation, as identified by Hypermut 2.0, were significantly higher than the other subjects ($p = 0.009$) (Figure 20A). Thirteen patients met all of the following criteria for hypermutation: general APOBEC3F/G-type hypermutation as determined by Hypermut 2.0 analysis, adenine proportion higher than one standard deviation above the mean for the 240 proviral sequences examined, and lethal mutations in the two examined ORFs. These thirteen patients also had significantly higher $CD4^+$ counts than the other participants ($p = 0.005$) (Figure 20B).

After determining that there was a relationship between hypermutation and $CD4^+$ count for the subset of patients with evidence of dramatic proviral hypermutation, the entire data set was examined to see if the relationship persisted. $CD4^+$ count and proviral adenine proportion were correlated in an unbiased, casewise comparison for all samples where $CD4^+$ counts were available ($n = 208$), and found to be significantly correlated with adenine proportion ($p = 0.042$) with a Spearman *r* value of 0.1411 (Figure 21).

Additionally, $CD4^+$ percentage (of total lymphocytes) was examined, as this lymphocyte measure is less variable than $CD4^+$ count (288). Comparison of the 17 subjects with significant hypermutation, as identified by Hypermut 2.0 to the remaining subjects revealed that they had significantly higher % $CD4^+$ levels ($p = 0.0395$), supporting the $CD4^+$ count data (Figure 22). Comparison of the 13 subjects that met all the hypermutation criteria to the remaining subjects, however, showed border-line non-significance ($p = 0.0540$).

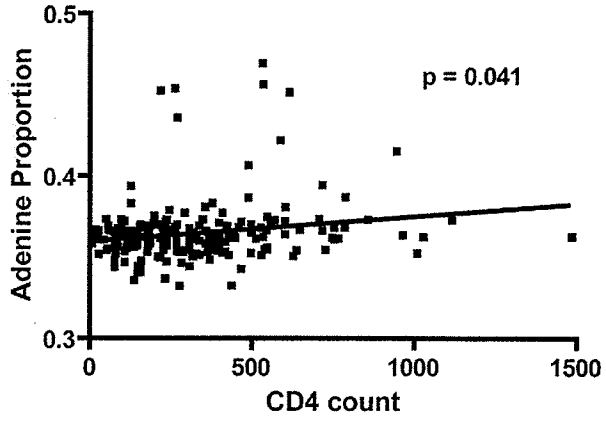


Figure 21. Correlation of HIV-1 proviral adenine proportion with $CD4^+$ count. $CD4^+$ counts were available for 208 subjects. Adenine proportion was measured from the 590 nucleotide *vpu/env* HIV-1 proviral region. Each data point represents a single subject. These two measures were significantly positively correlated ($p = 0.041$).

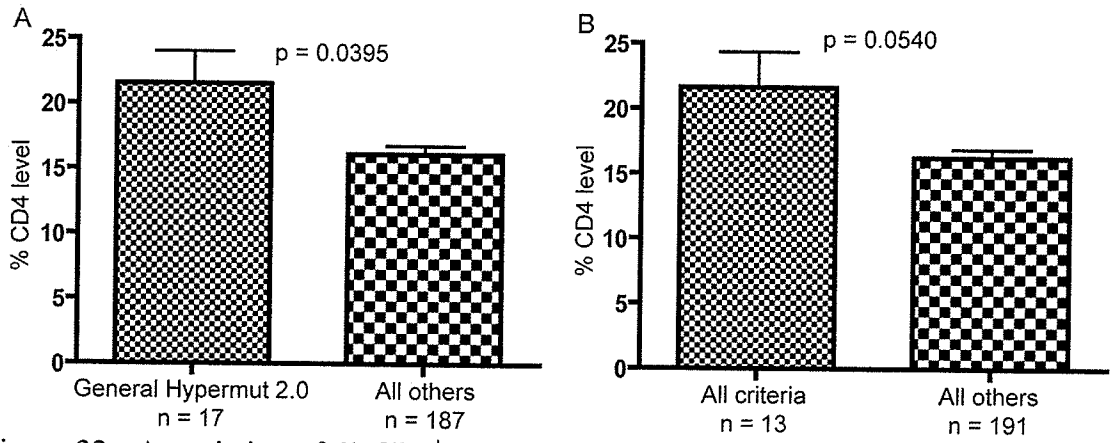


Figure 22. Association of % $CD4^+$ count with hypermutated HIV-1 proviral sequences. $CD4^+$ percentages of total lymphocytes are compared between patients with hypermutated provirus and patients without hypermutated provirus. A. Hypermutation was defined as significant general APOBEC cytidine deaminase activity, as determined by Hypermut 2.0. B. Hypermutation was defined as sequences with detrimental mutations in Vpu and Env, proviral adenine proportion greater than one standard deviation above the mean, and significant Hypermut 2.0 hypermutation. Groups are compared by a Mann-Whitney test. The bar height corresponds to the mean and the error bar represents the standard error of the mean.

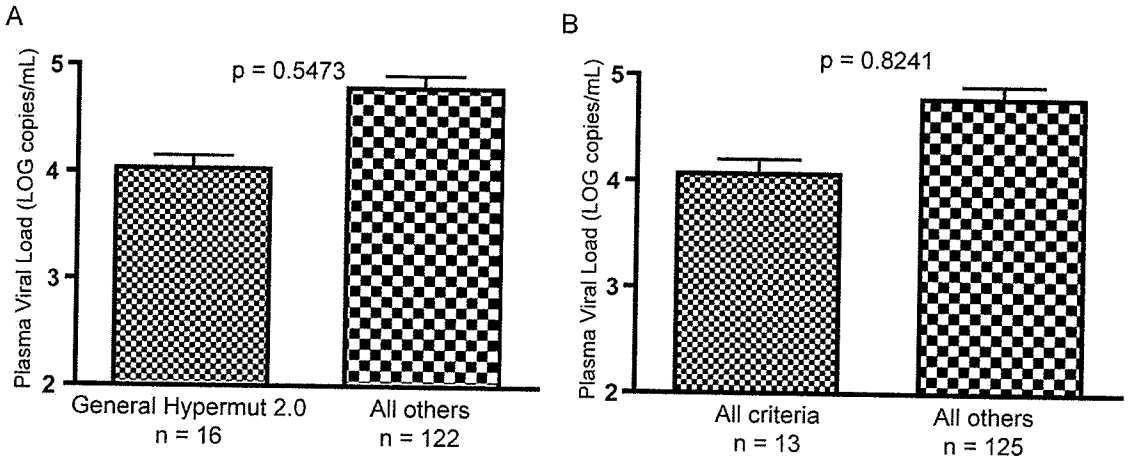


Figure 23. Lack of statistically significant association of plasma viral load with hypermutated HIV-1 proviral sequences. Plasma viral loads are compared between patients with hypermutated provirus and patients without hypermutated provirus. A. Hypermutation was defined as significant general APOBEC cytidine deaminase activity, as determined by Hypermut 2.0. B. Hypermutation was defined as sequences with detrimental mutations in Vpu and Env, proviral adenine proportion greater than one standard deviation above the mean, and significant Hypermut 2.0 hypermutation. Groups are compared by a Mann-Whitney test. The bar height corresponds to the mean and the error bar represents the standard error of the mean.

Viral load was measured from archived samples where date-matched plasma samples were available (n=138) (Figure 23). The same groups were compared as for the $CD4^+$ data. Although in both comparisons the patients with hypermutated proviral sequence had lower mean viral loads (for comparison of patients with General Hypermut 2.0 significance: mean = 4.03 $\log_{10}$ copies/mL compared to mean = 4.79 $\log_{10}$ copies/mL, for comparison of patients that met all the hypermutation criteria: mean = 4.07 $\log_{10}$ copies/mL compared to mean = 4.78 $\log_{10}$ copies/mL), the distribution was not statistically significant ($p = 0.55$ and $p = 0.82$, respectively). However, viral load determination was conducted on archived samples, not freshly isolated plasma. Sample storage is known to affect viral load measurement, whereas the $CD4^+$ counts were measured on site from fresh blood and are likely more accurate (39).

$CD4^+$ count is known to vary with age, with $CD4^+$ counts decreasing with increasing age (212). We therefore wished to ensure that the differences in $CD4^+$ count were not dependant on age. Age was available for 235 of the 240 subjects. The thirteen individuals with significant hypermutation in all tested categories had a mean age of 30 years, while the remaining subjects had a mean age of 32 years. This distribution was not determined to be significant with a Mann-Whitney comparison test ($p = 0.244$). Similarly, the seventeen individuals with significant hypermutation as determined by Hypermut 2.0 had a mean age of 31 years, while the remaining subjects has a mean age of 32 years. This was also not determined to be a significant difference by Mann-Whitney comparison test ($p = 0.260$).

3.12 Hypermutation is not associated with obvious Vif mutations

In order to determine the role Vif sequence polymorphisms may be playing in the observed hypermutation, a subset of Vif sequences were examined from both highly hypermutated and non-hypermutated proviruses. Intact starting methionine residues and no premature stop substitutions were found in the Vif of all examined samples (Figure 24). Additionally, these samples were examined for evidence of APOBEC activity with Hypermut 2.0, as described for the *vpu/env* sequences (Table 11). Non-significant levels of hypermutation were found in all *vif* sequences, regardless of whether hypermutation was identified in the *vpu/env* region. Amino acid polymorphisms in the Vif sequences were identified, but none in regions that had been previously identified as critical for Vif interaction with APOBEC (172,173,182,192,250,284,328). Additionally, only conservative substitutions were identified at single amino acid residue locations that had been suggested to have an impact on APOBEC interaction (268). Overall, Vif from highly hypermutated provirus was markedly similar to Vif from the other samples in this study, and to published consensus sequences, suggesting that Vif polymorphisms may not be responsible for the increased APOBEC3 type hypermutation observed within this population. Further studies are necessary, however, as subtle changes in Vif sequence may have an effect on function.

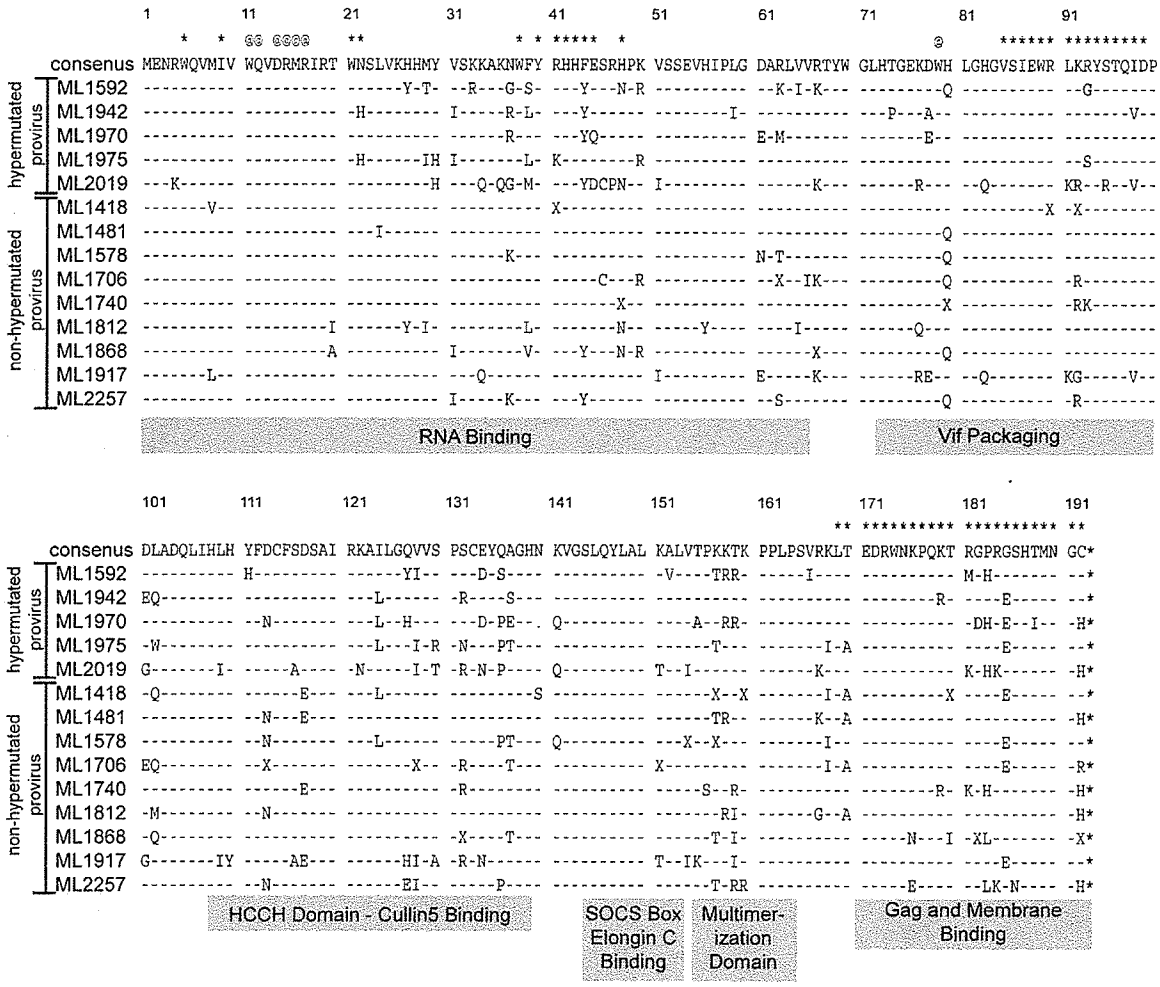


Figure 24. Comparison of Vif sequences from isolates with highly-hypermutated and non-hypermutated proviral *vpu/env* sequence. The consensus for this group of sequences is shown on the top line; matched residues are indicated in subsequent sequences with a dash (-), while non-identical residues are shown with the one-letter amino acid code. Residues labelled X are ambiguous due to multiple DNA nucleotides in the corresponding codon. The first five sequences (ML1592, ML1942, ML1970, ML1975 and ML2019) represent isolates with highly hypermutated proviral *vpu/env* sequences. The next nine sequences represent isolates with non-hypermutated proviral *vpu/env* sequences. The entire (193 amino acid) Vif ORF is shown, with numbering from the +1 of translation. Functional domains are highlighted under the approximate sequence regions. Residues that have been reported to be important for APOBEC3G interaction are marked with an asterisk (*) while residues that have been reported to be important for APOBEC3F interaction are marked with an at symbol (@) (172,173,182,192,250,284,328).

Table 11. Lack of significant hypermutation in Vif

Patient Identifier	Hypermutated Vpu/Env Region[a]	Hypermut Significance in Vif[b]
ML1592	Yes	0.974
ML1942	Yes	0.646
ML1970	Yes	0.971
ML1975	Yes	1
ML2019	Yes	0.199
ML1418	No	0.511
ML1481	No	0.798
ML1578	No	0.889
ML1706	No	0.973
ML1740	No	0.571
ML1812	No	1
ML1868	No	0.979
ML1917	No	0.618
ML2257	No	0.969

[a]Hypermutation of these samples was determined as outlined in Table 10.
[b]Hypermut 2.0 analysis for general APOBEC-type hypermutation (default settings of G to A substitutions with a downstream context of RD, where R is either A or G and D is A, T or G). A significant value would be considered as <0.05.

3.13 Summary

The role of APOBEC-mediated hypermutation in disease progression is not well defined. Previous publications have documented correlation of hypermutation with $CD4^+$ count and viral load in the presence of detrimental Vif mutations (223) and a lack of correlation between hypermutation and viral load (289). This, however, is the first account of hypermutation being associated with disease progression (as measured by $CD4^+$ count) in the absence of mutations in Vif. The work from this chapter has been published in the *Journal of Virology* (150).

4. Longitudinal Analysis of Subjects Superinfected with HIV-1 Reveals Changes in Hypermutation Levels in a Subset of Individuals

4.1 Rationale

The previous section showed that proviral HIV-1 sequence data provides a good indication of APOBEC3F/G mediated hypermutation, and that this hypermutation correlates with $CD4^+$ count, a measure of disease progression. This correlation was in the absence of obvious sequence defects in Vif, HIV's defence against host APOBEC, contrary to other published studies that found a correlation between hypermutation and disease progression. The cause of the increased hypermutation thus remains to be determined, as it is unknown what causes the shift from Vif to APOBEC control.

4.2 Hypothesis

With each viral infection, there are a multitude of variables that will determine the infection's progression; these variables pertain to both the virus and the host. With viral superinfection, however, the host is the same (excepting changes caused by the initial infection), while the incoming virus is the major variable. Therefore, if the infecting virus, via the activity of the viral Vif protein, is responsible for determining the level of APOBEC-mediated hypermutation, one could hypothesize that upon superinfection with a new virus, there would be altered levels of hypermutation, at least in the provirus of the superinfecting virus. However, if host factors are responsible for determining the level of APOBEC-mediated hypermutation, one would hypothesize that the level of proviral hypermutation would be constant despite superinfection, both before and after the superinfection event and between the initial and the superinfecting viruses. Based on the

results of the previous section, which indicated that Vif does not seem to be the major player in determining hypermutation levels, we hypothesized that host factors are more important in controlling hypermutation, and thus examination of the HIV provirus from subjects before and after superinfection would reveal no changes in hypermutation levels.

4.3 Objectives

- Identify a prospective cohort of individuals who became infected and subsequently superinfected with HIV and for which there is longitudinal HIV proviral sequence data
- Retrieve the available sequences from the superinfected individuals and determine levels of hypermutation in the sequences
- Compare hypermutation levels in the original virus before and after the superinfection event and with the superinfecting virus

4.4 Study outline

A recently published study by Piantadosi *et al.* describes the prospective follow up and sequence analysis from seven patients that were determined to be superinfected by HIV-1 (228). Proviral sequences from the *env* V1 – V5 region, sampled over the course of HIV infection for each subject, were publically available through the GenBank genetic sequence database (Accession Numbers EU163983-EU164399). These sequences were retrieved and analysed for hypermutation, before and after the superinfection event, to determine if changes in hypermutation levels could be observed.

4.5 The published study

The subjects described in the publication by Piantadosi *et al.* are from a prospective cohort of high-risk by heterosexual contact HIV-1 seronegative women from Mombasa, Kenya (184,228). Blood samples from the subjects were obtained approximately monthly and tested for HIV-1 infection. The samples described in the superinfection study were obtained from 1993 – 2004. Nested PCR was used to amplify a 1.2 kb region of *env*. Superinfection was determined by examining proviral HIV-1 sequences from *env* and *gag,* where available, from two different time points – one corresponding to early after initial infection, and the other from a later, chronic period (3.6 – 7.2 years post-infection). All sequences from the initial time point were found to be monophylogenetically clustered. Potential superinfection cases were identified when sequences from the chronic phase of infection did not form a single cluster with the sequences from the early phase of infection. Seven cases of superinfection were identified (Table 12).

4.6 Phylogenetic analysis of proviral HIV-1 env *sequences obtained before and after superinfection*

To determine which HIV sequences obtained from each subject represented the original infecting virus and which represented the new, superinfecting virus, the phylogenetic analysis described in the original publication was repeated (228). Briefly, for each of the seven patients, all proviral HIV-1 *env* sequences were aligned by ClustalW. The resulting alignment was used to generate a neighbour-joining tree with MEGA 4. The superinfecting virus sequences are expected to form a distinct branch from the original

Table 12: Superinfection cases identified by Piantadosi *et al.*[1] (228)

Subject ID	Estimated Time of Superinfection (days post-infection)	Initial *env* Subtype	Chronic *env* Subtype
QA413	714-1007	A	A
QB045	1680-2048	A2	A2 + A1
QB685	303-1453	A	A
QB726	749-1031	A	A
QB850	52-73	A	A + A/D
QC885	58-152	A	A
QD022	1832-1957	A	C

[1]The data in this table was obtained from Table 1 of the reference publication (228).

virus sequences. In contrast, sequences from the original virus before the superinfection event are expected to intermingle with the sequences from the same virus after the superinfection event. This pattern was observed for patient QA413 (Figure 25, panel A). The sequences from virus sampled prior to the superinfection event all cluster together with a bootstrap value of 100, and also include a subset of the sequences from virus sampled after the superinfection event. These samples thus likely represent the same virus. The other half of the tree contains only viral sequences sampled after the superinfection event – these likely represent the superinfecting virus. The samples from patients QB045, QB726 and QB850 (Figure 25, panel B, D and E) display a similar pattern, with bootstrap values of 99-100 that support the branches that separate the original virus from the superinfecting virus. In patient QB685, the sequences from viral samples obtained prior to superinfection are distinct from the sequences sampled after the superinfection event, separated with a bootstrap value of 100 (Figure 25, panel C). This suggests that the superinfecting virus has outcompeted the original infecting virus, such that the original virus is no longer easily detected after the superinfection event. Similar patterns were observed for patients QC885 and QD022, with bootstrap values of 90 and 100, respectively (Figure 25, panels F-G).

4.7 Examination of proviral HIV-1 env *sequences for changes in APOBEC-mediated hypermutation with superinfection*

As described previously, APOBEC3F and APOBEC3G are known to cause G to A hypermutation in HIV-1 provirus; this allows for the detection of hypermutation by examining the proviral HIV sequences (24,163,310,327,336). For each of the seven

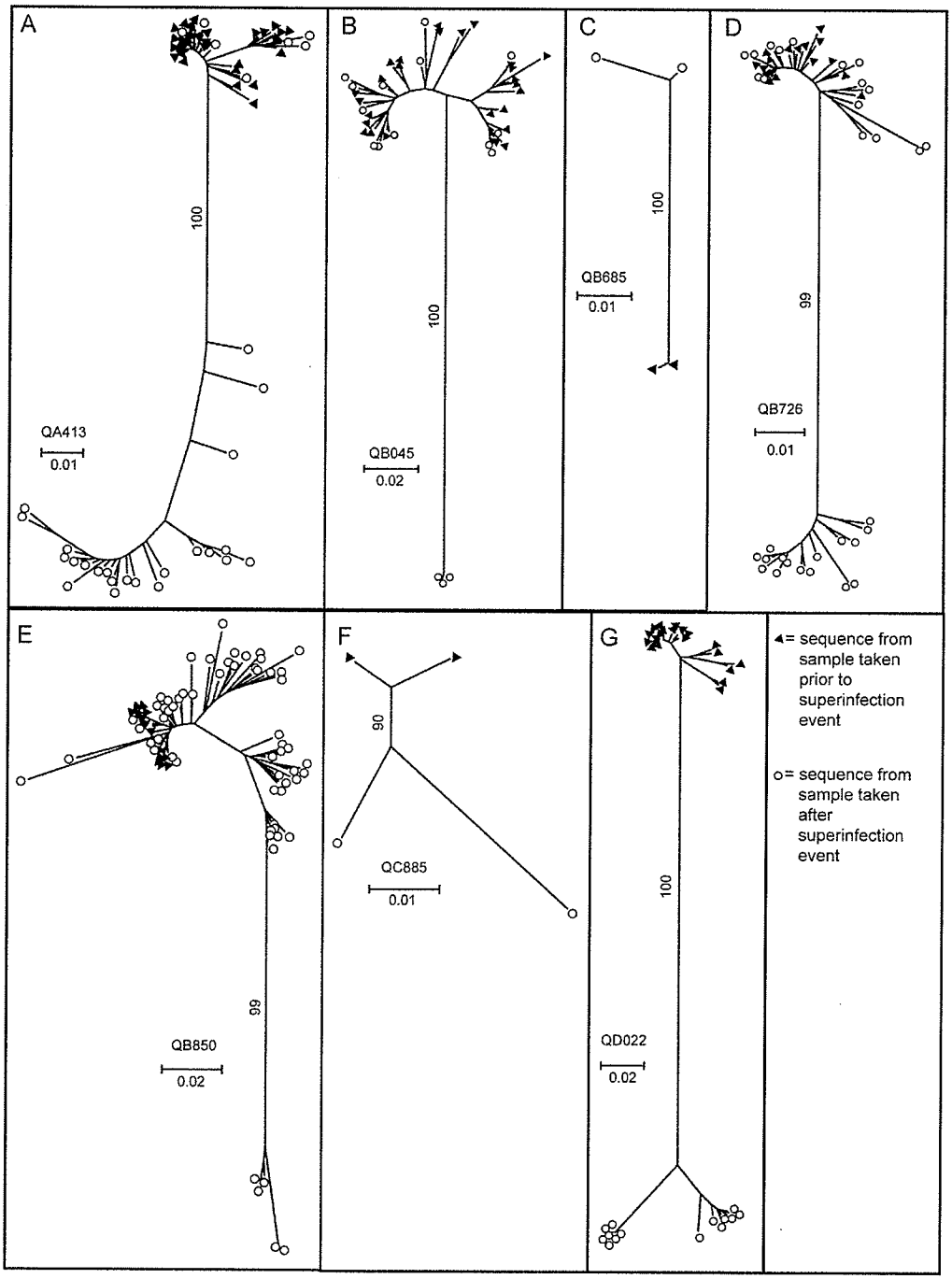


Figure 25. Neighbour-joining trees of proviral HIV-1 sequences from subjects that experienced a superinfection event. For each of the seven subjects (panels A-G), all the *env* sequences, both prior and subsequent to the superinfection event, were aligned and used to construct a neighbour-joining tree. Sequences from HIV-1 samples collected prior to the superinfection event are represented by closed triangles, while sequences from HIV-1 samples collected after the superinfection event are represented by open circles. The bootstrap value on the branch that likely separates the original viral sequences from the superinfecting viral sequences is indicated on each tree. The scale of genetic distance is indicated in each panel.

subjects with identified superinfection, the proviral adenine proportion was determined for each sequence, from each time point (Figure 26). Additionally, these viral sequences were examined for hypermutation specifically characteristic of APOBEC3F/G deamination activity using Hypermut 2.0 (hiv.lanl.gov), as described in Results section 3.10 (247). The viral sequences were compared to clade-specific consensus sequences generated from this population. The rate ratio, which measures the amount of APOBEC-mediated hypermutation in a given sequence, with higher values indicating higher levels of hypermutation, was plotted for the seven superinfected subjects (Figure 27). Rate ratios were compared for viral sequences obtained prior to superinfection, sequences from the original infecting virus sampled after superinfection and sequences from the superinfecting virus, where all three categories were applicable. Otherwise, the sequences were simply compared before and after the superinfection event.

4.7.1 Subject QA413

Patient QA413 (Figures 26 and 27, panel A) showed low levels of HIV-1 proviral adenine proportion prior to superinfection; however after superinfection, the sequences from the original virus maintained the low level of adenine proportion (mean = 37.8%), while the sequences from the superinfecting virus had higher adenine proportion levels (mean = 39.4%). The sequences from the superinfecting virus also had a significantly higher rate ratio than the sequences from the original virus, both pre and post superinfection ($p < 0.0001$, Kruskal-Wallis, $p < 0.001$ and $p < 0.001$, respectively, Dunn's multiple comparison test). The

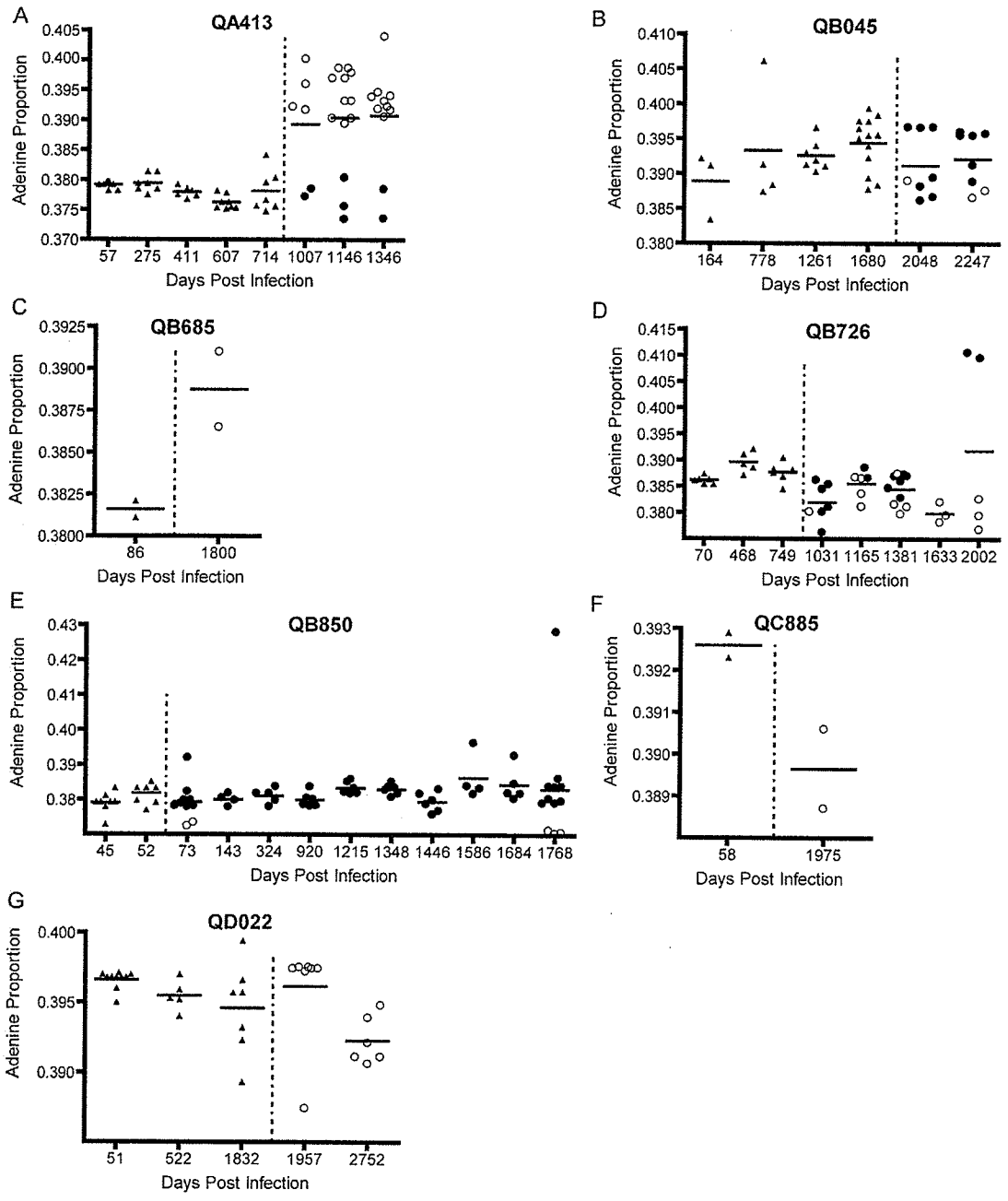


Figure 26. Levels of HIV-1 proviral adenine proportion before and after superinfection. Seven subjects with documented cases of superinfection were examined for hypermutation in an *env* region using adenine proportion (228). The sequences were examined at each of the available times points, measured in days post initial HIV-1 infection. Each data point represents an individual sequence. The horizontal bar represents the mean adenine proportion for each time point. The vertical dashed line indicates the time point of superinfection. Data points from the original viral sequences are represented by closed triangles, while data points from the superinfecting viral sequences are represented by open circles.

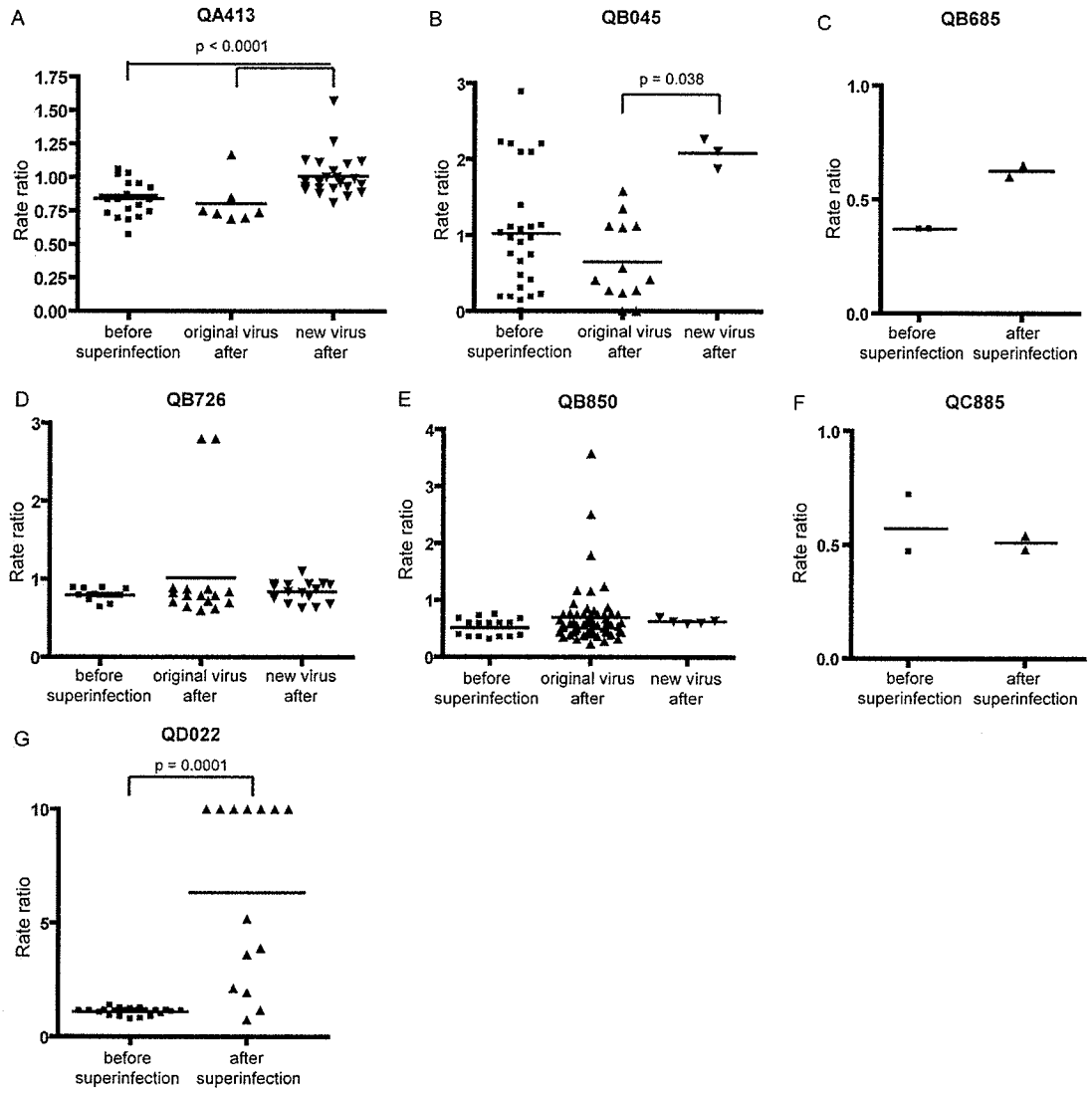


Figure 27. G to A hypermutation rate ratios for the seven patients with HIV-1 superinfection. The rate ratio is defined by the expression $^{(a/b)}/_{(c/d)}$, where (a) is the number of G to A mutations that occur with a downstream context of RD, where R is either A or G and D is A, T or G; (b) is the total number sites that could be potentially hypermutated (i.e. the number of sites that are GRD); (c) is the number of control mutations, defined as G to A mutations that occur with a downstream context of YN or RC, where N is any nucleotide; (d) is the total number of sites that could be potential control sites (i.e. the number of sites that are GYN or GRC). Each available *env* sequence for the seven patients with documented cases of superinfection (228) was examined for hypermutation using the Hypermut 2.0 tool available at www.hiv.lanl.gov by comparing the sequences with a population and clade specific consensus. The rate ratio values were then compared for sequences obtained before the superinfection event, sequences from the original infecting virus after the superinfection event and sequences from the superinfecting virus (A, B, D, E). In panels C, F, and G, sequences before superinfection are simply compared to sequences after superinfection, as there does not appear to be a mix of sequences after the superinfection event.

samples from the superinfecting virus were maintained over follow-up: at the first sampling point post superinfection, the superinfecting virus is 66.7% of the sampled sequences, increasing to 76.9% and 81.8% at the two subsequent follow-ups. Thus, the superinfecting viral sequences maintained higher levels of hypermutation than the original viral sequences.

4.7.2 Subject QB045

The sequences from the superinfecting virus in patient QB0445 (Figures 26 and 27, panel B) also had a significantly higher rate ratio than the sequences from the original virus after superinfection, but they were not significantly different from the sequences prior to the superinfection event ($p = 0.038$, Kruska-Wallis, $p < 0.05$ and $p > 0.05$, respectively, Dunn's multiple comparison test). Additionally, this hypermutation difference was not indicated by the adenine proportion data. However, it should be noted that greater variability in rate ratio was observed in the viral sequences sampled prior to superinfection – seemingly, superinfection triggered a cull of cells infected with comparatively hypermutated virus, or perhaps the replication pressure exerted by an additional virus selected for the most fit (i.e. less hypermutated) viruses. Increased viral sequencing at the time points after superinfection would be required to test this hypothesis.

4.7.3 Subjects QB685 andQC885

These two superinfected subjects (Figures 26 and 27, panels C and F) show neither changes in adenine proportion nor hypermutation rate ratio. However, they also have the

least sampling (two sequences were available at each time point) and follow-up (one visit pre superinfection and one visit post superinfection). This underscores the importance of adequate testing to detect hypermutation.

4.7.4 Subjects QB726 and QB850

In two superinfection cases, QB726 and QB850, an emergence of a subset of proviruses with higher adenine proportions at the last sampling time point was observed (Figures 26 and 27, panels D and E). In QB726, the hypermutated samples were derived from the original virus, appearing over 2000 days after the initial infection. Similarly, at the last sampling point for subject QB850, which corresponded to over 1700 days after the initial infection, one of the sequences derived from the original virus showed increased adenine proportion. This may represent an eventual emergence of hypermutated virus given sufficient time. However, there were other subjects in this group that had similar follow up periods, without the emergence of highly hypermutated proviral sequences, so this time period may not be sufficient to generate hypermutated proviruses in all individuals. Alternatively, the proportion of hypermutated sequences may be quite low, such that increased sampling at each time point would be required to gain an accurate assessment.

4.7.5 Subject QD022

The sequences from the superinfecting virus in patient QD022 (Figures 26 and 27, panel G) also had a significantly higher rate ratio than the original viral sequences ($p = 0.0001$, Mann-Whitney), although this difference in hypermutation was not detected by examining adenine proportion. Interestingly, in this patient the superinfecting virus

seemingly out-competed the original infecting virus, such that after the superinfection event, the original virus was no longer detected. However, there were only two sampling points after superinfection, so perhaps samples from additional time points would have revealed the presence of the original virus. Indeed, it seems unusual that a virus more susceptible to APOBEC-mediated hypermutation would be able to outcompete and replace a virus that was comparatively resistant to APOBEC-mediated hypermutation.

4.8 Summary

This examination of hypermutation looked at viral population dynamics within patients to determine whether host or viral factors are responsible for determining APOBEC-mediated hypermutation levels, a topic that has yet to be resolved. These results indicated that in at least two instances, viral factors were responsible for controlling hypermutation levels, as the superinfecting virus had significantly higher levels of hypermutation than the original infecting virus. In the other examined cases, however, host factors seemed more important in controlling hypermutation levels as they remained constant in both viruses after superinfection. The work from this chapter is the subject of a manuscript in preparation.

5. Discovery of SNPs Associated with Differing Levels of HIV-1 Proviral Hypermutation by Pyrosequencing the APOBEC3G Gene

5.1 Rationale

Results section 3 of this thesis showed that a subset of subjects were infected with HIV-1 that was significantly hypermutated. However, hypermutation occurred in the absence of obvious mutations in Vif, the HIV-1 protein responsible for combating the host innate proteins, APOBEC3G and APOBEC3F, which cause this hypermutation. The previous thesis section indicates that APOBEC-mediated hypermutation may be controlled by both host and viral factors. The reason for the increased proviral hypermutation in these patients is thus unclear.

5.2 Hypothesis

As the viral Vif protein did not seem to be responsible for the observed increases in HIV-1 proviral hypermutation, we thought it prudent to examine host factors that may be involved. APOBEC3G is the most well described of the APOBEC3 proteins and has the highest level of proviral hypermutation activity. We hypothesize that subjects infected with significantly hypermutated HIV-1 provirus will have mutations in APOBEC3G that make this host defence more active against HIV-1, possibly by increasing the protein's expression, deamination activity or by decreasing its susceptibility to Vif-mediated degradation. We expect these mutations to be over-represented in the subjects with significantly hypermutated HIV-1 provirus, but under-represented in the subjects with non-significantly hypermutated provirus.

5.3 Objectives

- Identify subjects that have dramatically hypermutated HIV-1 provirus, intermediately hypermutated provirus and non-hypermutated provirus (as determined in results section 3 of this thesis)
- Amplify the APOBEC3G gene and surrounding region for a subset of these subjects
- Sequence the APOBEC3G amplicon using pyrosequencing technology
- Identify SNPs and compare frequency between the differentially hypermutated populations

5.4 Study Outline

The APOBEC3G gene and surrounding region, encompassing all published SNPs (13.5 kb) was amplified for nine individuals with dramatically hypermutated HIV-1 provirus and sequenced using the GS FLX pyrosequencing platform (Roche Diagnostics). Additionally, the APOBEC3G amplicon from six individuals with intermediate levels of proviral hypermutation was pooled in equimolar amounts and sequenced. Similarly, the APOBEC3G amplicon from 87 individuals with non-hypermutated provirus was pooled in equimolar quantities and sequenced. GS FLX Data Analysis Software (Roche Diagnostics) was used to identify SNPs by comparing the resulting sequences to the APOBEC3G region of the *Homo sapiens* chromosome 22 reference assembly. Identified SNPs were manually validated prior to comparing SNP frequency between the various proviral hypermutation groups.

5.5 Overview of GS FLX Sequencing results

The APOBEC3G gene region was amplified from nine individuals with highly hypermutated HIV-1 provirus, as identified in results section 3 of this thesis (Table 13). Additionally, two pooled samples, one consisting of amplicons from six individuals with intermediate levels of proviral hypermutation (Table 14) and one consisting of amplicons from eighty-seven individuals with low levels of proviral hypermutation (Table 15), were sequenced. The samples with intermediate levels of hypermutation met some, but not all of the criteria for having significant levels of HIV-1 proviral hypermutation; the criteria they did meet are listed in Table 14. The criteria these samples did not meet included a high adenine proportion in the proviral segment examined and the presence of non-functional mutations in the examined ORFs. The samples with low levels of hypermutation did not meet any of the criteria established for hypermutation in results section 3 of this thesis.

The total sample size of this study was 102 individuals, which corresponds to 204 alleles. The pyrosequencing data generated for these samples was validated by examining the quality control data (Table 16). The sequence reads for each sample or pool formed a single contig that spanned the length of the amplicon. The depth of coverage was sufficient to sample both alleles in the nine individual samples. Similarly, the depth was sufficient to sample the twelve alleles in the pool of six samples. However, the pool of 87 samples did not have a great enough depth to detect a SNP present in only a single allele with high confidence.

Table 13: Samples with significantly hypermutated HIV-1 provirus for which the APOBEC3G genomic region was sequenced

Patient Identifier	Fatal Mutations in Vpu and Env[a]	Adenine Proportion[b]	General Hypermut Significance[c]	Specific Hypermut Significance[d]	
				A3G	A3F
ML1053	Yes	0.3932	0.0012	6.704 E-5	0.0644
ML1102	Yes	0.3828	0.0029	1.466 E-4	*0.1321*
ML1578	Yes	0.4212	1.708 E-5	1.214 E-5	0.0014
ML1592	Yes	0.4688	2.660 E-10	0.0175	9.277 E-11
ML1857	Yes	0.4520	1.425 E-4	1.456 E-4	0.0038
ML1957	Yes	0.4353	5.911 E-9	0.0251	6.44 E-10
ML1970	Yes	0.3861	0.0058	0.00350	*0.0633*
ML1975	Yes	0.3865	8.894 E-5	5.915 E-5	0.0038
ML2019	Yes	0.4148	6.901 E-4	1.361 E-7	*0.1554*

[a]These samples had mutations such as missing start codons and premature stop codons in the proviral ORFs examined.

[b]Adenine proportion was higher than one standard deviation above the mean (i.e. >0.382751).

[c]Hypermut 2.0 analysis for general APOBEC-type hypermutation (default settings of G to A substitutions with a downstream context of RD, where R is either A or G and D is A, T or G) gave a significant p value (<0.05).

[d]Hypermut 2.0 analysis for specific APOBEC3G (settings of G to A substitutions with a downstream context of GD) and/or APOBEC3F (settings of G to A substitutions with a downstream context of AD) hypermutation gave a significant p value (<0.05). Non-significant values are italicized.

Table 14: Composition of intermediate pool (pool of six samples with intermediately hypermutated HIV-1 provirus)

Sample	General Hypermut Significance[a]	Specific Hypermut Significance[b]	
		A3G	A3F
ML602	0.0619	**0.0397**	0.2397
ML790	0.1076	**0.0229**	0.5607
ML1419	**0.0287**	0.2019	**0.0243**
ML1649	**0.0113**	**0.0494**	**0.0228**
ML1992	0.0508	1	**0.0149**
ML2209	**0.0030**	0.0910	**0.0017**

[a]Hypermut 2.0 analysis for general APOBEC-type hypermutation (default settings of G to A substitutions with a downstream context of RD, where R is either A or G and D is A, T or G); significant values ($p < 0.05$) are bolded.

[b]Hypermut 2.0 analysis for specific APOBEC3G (settings of G to A substitutions with a downstream context of GD) or APOBEC3F (settings of G to A substitutions with a downstream context of AD) hypermutation; significant values ($p < 0.05$) are bolded.

Table 15: Composition of low pool (pool of 87 samples with non-significantly hypermutated HIV-1 provirus)

Sample	Sample	Sample	Sample	Sample	Sample
MCH5570	ML385	ML631	ML1005	ML1418	ML1870
MCH6099	ML387	ML639	ML1008	ML1443	ML1885
ML17	ML389	ML657	ML1075	ML1450	ML1917
ML48	ML416	ML718	ML1076	ML1481	ML1990
ML76	ML431	ML752	ML1199	ML1591	ML2000
ML122	ML435	ML781	ML1203	ML1594	ML2020
ML157	ML478	ML825	ML1227	ML1640	ML2166
ML215	ML509	ML864	ML1292	ML1667	ML2185
ML216	ML525	ML888	ML1296	ML1694	ML2203
ML264	ML546	ML897	ML1316	ML1706	ML2204
ML265	ML571	ML959	ML1335	ML1740	ML2227
ML288	ML589	ML960	ML1337	ML1783	ML2245
ML327	ML590	ML995	ML1346	ML1789	
ML331	ML603	ML1000	ML1390	ML1790	
ML353	ML605	ML1004	ML1404	ML1812	

Table 16: Pyrosequencing quality control data

Sample	Number of Contigs	Length of Contig (kb)	Average Sequencing Depth[a]	Average Sequencing Depth per Allele[b]	Minimum Expected Reads per Allele[c]	Average Read Length (bases)
ML1053	1	13.481	52	26	20	225
ML1102	1	13.508	50	25	19	217
ML1578	1	13.481	52	26	20	216
ML1592	1	13.517	51	26	20	224
ML1857	1	13.477	41	21	15	216
ML1957	1	13.496	50	25	19	223
ML1970	1	13.478	53	27	21	225
ML1975	1	13.480	57	29	22	212
ML2019	1	13.480	75	38	30	222
Intermediate Pool	1	13.515	169	14	8	215
Low Pool	1	13.523	253	1	<1 [d]	220

[a]This represents the average of reads that could reliably support or refute a SNP (i.e. the read is of good quality and neither begins nor ends at the SNP in question).

[b]This column is the average sequencing depth divided by the total number of alleles in the sample.

[c]A cumulative binomial distribution was calculated to determine this number (k), where fewer reads per allele had a probability less than 0.05. Binomial distribution parameters were: n is the average sequencing depth and p is the inverse of the total number of alleles in the sample.

[d]The maximum expected reads for any one allele in this pool is 3. A non-cumulative (discrete) binomial distribution was calculated to determine this number (k), where more reads per allele had a probability less than 0.05. Binomial distribution parameters were as above: n is the average sequencing depth (i.e. 253 reads) and p is the inverse of the total number of alleles in the sample (i.e. 174^{-1}).

In total, 89 high-confidence SNPs were mapped in these samples; 39 of them were novel (Figure 28) (Appendix A). The newly described SNPs were in general found in a low allelic frequency in this population, though two of these SNPs were found in greater than 25% of the alleles. The majority of SNPs, both novel and previously described, were in non-coding regions; only four were in exons (Table 17). Two of these, however, were synonymous changes (i.e. did not cause an amino acid change). One of these SNPs had not been previously described. The two SNPs that changed the encoded amino acid were located in exon 4 and exon 6 and caused histidine to arginine and glutamine to glutamic acid changes, respectively.

5.6 APOBEC3G SNPs were present in differential frequencies between the hypermutation groups

To determine whether specific APOBEC3G polymorphisms were associated with elevated HIV-1 proviral hypermutation, the distribution of the 89 identified SNPs in the different hypermutation groups was compared. Some of the observed SNPs seemed to be over-represented in the samples with highly hypermutated HIV-1 provirus; eight SNPs fit this criterion (Table 18). Analysis of the frequencies using a Chi-square test for trend, however, only confirmed that one of the SNPs was significantly associated with hypermutation. This SNP was a G to C nucleotide change observed 571 nucleotides upstream of the APOBEC3G gene. This SNP was present in all of the nine samples with highly hypermutated HIV-1 provirus, in both alleles of each subject. It was also observed in the pool of samples with intermediate levels of proviral hypermutation, and in the pool of samples without proviral hypermutation, though in significantly fewer alleles ($p =$

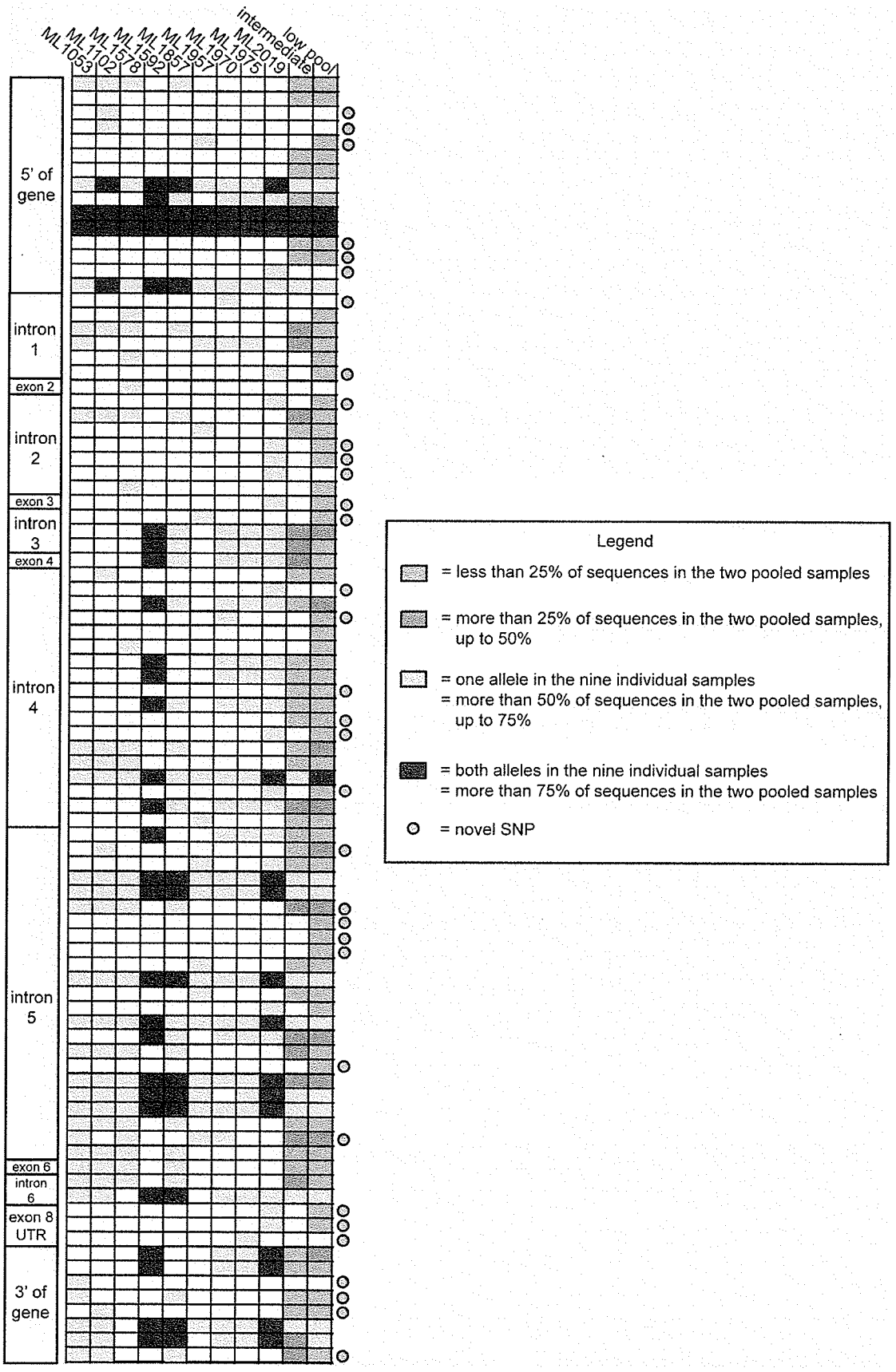

ML1053
ML1102
ML1578
ML1592
ML1857
ML1957
ML1970
ML1975
ML2019
intermediate
low pool
5' of gene
intron 1
exon 2
intron 2
exon 3
intron 3
exon 4
intron 4
intron 5
exon 6
intron 6
exon 8 UTR
3' of gene
Legend
= less than 25% of sequences in the two pooled samples
= more than 25% of sequences in the two pooled samples, up to 50%
= one allele in the nine individual samples
= more than 50% of sequences in the two pooled samples, up to 75%
= both alleles in the nine individual samples
= more than 75% of sequences in the two pooled samples
= novel SNP

Figure 28 (previous page). Heat map indicating the prevalence of observed SNPs in the APOBEC3G genomic region. Each column is a different sample; each row is a different SNP. The SNPs are plotted from 5' to 3' of the APOBEC3G amplicon; the region in which the SNPs are located is indicated to the left of the rows. The significance of each colour is indicated in the legend; an empty box indicates the SNP was not present in that sample. SNPs not previously described are also indicated, to the right of the row.

Table 17: Identified SNPs in APOBEC3G exons

Location	Base Change	mRNA Position	Amino Acid Change	Amino Acid Position	Samples with SNP	Est. Allelic f.[a]	Published Allelic f.[b]
Exon 2	C to T	372	Synonymous (Pro)	47	ML1578	0.005	0.042
Exon 3	C to T	411	Synonymous (Ser)	60	ML2019 Pool87	0.02	*Unpublished SNP*
Exon 4	A to G	788[c]	His to Arg	186	ML1592 ML1857 ML1970 ML1975 ML2019 Pool6 Pool87	0.24	0.467
Exon 6	C to G	1054	Gln to Glu	275	ML1053 ML1102 ML1578 ML1854 Pool6 Pool87	0.25	0.167

[a]The estimated allelic frequency of the SNP was determined by multiplying the observed frequency of sequences with the SNP by the total number of alleles for the two pools and rounding to the nearest whole number to arrive at the estimated number of alleles with the SNP, and then dividing the total number of alleles sampled (204) by the estimated number of alleles with the SNP.

[b]The allelic frequency of the non-reference SNP in a Sub-Saharan African population is given.

[c]This SNP has been associated with a decline in $CD4^+$ T cells and accelerated progression to AIDS (5).

Table 18: SNPs that are over-represented in the samples with significantly hypermutated HIV-1 provirus

Position[a]	Gene Location	Base Change	Allelic f. in 9 hypermut. samples[b] (N)	Allelic f. in inter-mediate pool[c] (est. N)[e]	Allelic f. in low pool[d] (est. N)[e]	Published Allelic f.[f]	p value[g]
18862771	5' of gene	C to T	0.33 (6/18)	0.04 (1/12)	0.17 (30/174)	N.D.	0.1912
18863080	5' of gene	G to C	1.0 (18/18)	0.82 (10/12)	0.80 (139/174)	0.917	**0.0413**
18863561	5' of gene	C to G	0.67 (12/18)	0.60 (7/12)	0.53 (92/174)	0.562	0.2503
18868856	Intron 4	T to A[h]	0.28 (5/18)	0.08 (1/12)	0.15 (26/174)	0.261	0.2631
18868857	Intron 4	C to A[h]	0.28 (5/18)	0.08 (1/12)	0.15 (26/174)	0.261	0.2631
18868942	Intron 4	G to A	0.33 (6/18)	0.23 (3/12)	0.24 (42/174)	0.310	0.4165
18872375	Intron 5	C to G	0.67 (12/18)	0.45 (5/12)	0.45 (78/174)	0.469	0.1181
18874359	3' of gene	C to T	0.33 (6/18)	0.06 (1/12)	0.18 (31/174)	0.283	0.2215

[a]Position is numbered in accordance with the NCBI SNP database for the human APOBEC3G gene.
[b]These nine samples were obtained from patients with significantly hypermutated HIV-1 provirus and were sequenced individually. N = the total number of alleles.
[c]Intermediate pool is composed of 6 samples from patients with intermediately hypermutated HIV-1 provirus.
[d]Low pool is composed of 87 samples from patients with non-significantly hypermutated HIV-1 provirus.
[e]The estimated number of alleles was determined by multiplying the observed frequency of sequences with the SNP by the total number of alleles in the pool (i.e. twice the sample number) and rounding to the nearest whole number.
[f]The allelic frequency of the non-reference allele in a Sub-Saharan African population is given.
[g]Significance in allelic distribution was determined with a Ch-square test for trend. A significant p value was determined to be less than 0.05.
[h]These SNPs are linked.

0.0413). This SNP has previously been found at an allelic frequency of 0.917 in a Sub-Saharan African population.

Six SNPs seemed to be under-represented in the highly hypermutated samples (Table 19). Analysis of the frequencies using a Chi-square test for trend confirmed that two of the SNPs were indeed significantly depleted from the samples with high levels of hypermutation. One of these SNPs was a nucleotide change from G to C observed in the region 5' of the APOBEC3G gene (1.5 kb upstream). This SNP was not present in any of the samples with highly hypermutated HIV-1 provirus, but was observed at a low level in the pool of samples with intermediate levels of proviral hypermutation, and at an approximately six times greater frequency in the samples without proviral hypermutation ($p = 0.0276$). The allelic frequency of this SNP in the pool of samples with intermediate proviral hypermutation was similar to the allelic frequency described in the NCBI SNP database for a Sub-Saharan African population. The other significantly distributed SNP was a nucleotide change from C to T located in intron 5 of the APOBEC3G gene, 326 bases downstream from the end of exon 5. This SNP was only present in one of the nine samples with highly hypermutated HIV-1 provirus, in one of the subject's APOBEC3G alleles. It was observed at a nearly three-fold greater allelic frequency in the pool with intermediate levels of proviral hypermutation, and at a greater than four-fold frequency in the pool with low levels of proviral hypermutation ($p = 0.0437$). Interestingly, this SNP has not been previously described.

Table 19: SNPs that are under-represented in the samples with significantly hypermutated HIV-1 provirus

Position[a]	SNP Location	Base Change	Allelic f. in 9 hypermut. samples[b] (N)	Allelic f. in inter-mediate pool[c] (est. N)[e]	Allelic f. in low pool[d] (est. N)[e]	Published Allelic f.[f]	p value[g]
18862077	5' of gene	G to C	0 (0/18)	0.03 (1/12)	0.19 (33/174)	0.023	**0.0276**
18862429	5' of gene	G to A	0 (0/18)	0.10 (1/12)	0.08 (14/174)	0.033	0.2557
18868173	Intron 4	C to T[h]	0.06 (1/18)	0.23 (3/12)	0.14 (24/174)	0.042	0.5883
18870730	Intron 5	C to T[i]	0.06 (1/18)	0.16 (2/12)	0.26 (45/174)	–	**0.0437**
18870776	Intron 5	C to T[i]	0.22 (4/18)	0.38 (5/12)	0.32 (56/174)	–	0.4090
18874456	3' of gene	- to CCTCACTC ACAGAGCC CCGCCCA[i]	0.06 (1/18)	0.20 (2/12)	0.10 (17/174)	–	0.7807

[a]Position is numbered in accordance with the NCBI SNP database for the human APOBEC3G gene.
[b]These nine samples were obtained from patients with significantly hypermutated HIV-1 provirus and were sequenced individually. N = the total number of alleles.
[c]Intermediate pool is composed of 6 samples from patients with intermediately hypermutated HIV-1 provirus.
[d]Low pool is composed of 87 samples from patients with non-significantly hypermutated HIV-1 provirus.
[e]The estimated number of alleles with the SNP was determined by multiplying the observed frequency of sequences with the SNP by the total number of alleles in the pool (i.e. twice the sample number) and rounding to the nearest whole number.
[f]The allelic frequency of the non-reference allele in a Sub-Saharan African population is given.
[g]Significance in allelic distribution was determined with a Chi-square test for trend. A significant p value was determined to be less than 0.05.
[h]This SNP has been previously documented to be associated with increased susceptibility to HIV-1 infection (ref).
[i]These SNPs have not been previously identified.

5.7 Summary

Polymorphisms in APOBEC3G have been correlated with HIV infection and disease progression, but these SNPs have not been previously correlated with levels of HIV-1 hypermutation (5,291). This study of an HIV-positive Sub-Saharan African population identified 3 SNPs that were significantly associated with differing levels of HIV-1 proviral hypermutation; one of these SNPs had not been previously identified. This suggests that differences in APOBEC3G expression and/or activity may indeed be responsible for the identified differences in HIV-1 proviral hypermutation. The results outlined in this chapter are the subject of a manuscript in preparation.

Discussion

Over twenty-five years have passed since the virus now known as HIV was first described, yet HIV and its sequela, AIDS, continue to cause health problems of international concern. Treatment options are available to those who can afford it, but most experts agree that a vaccine is needed to stop this pandemic. Many candidate vaccines have been tried, and all have failed – some spectacularly. This has prompted an appeal within the scientific community to go back to basics so that we can better engineer a response (82,83). The viral sequence is likely an important source of information; it is a record of epidemiology, evolution and immunological influences. We can use the viral sequence to determine viral clade and recombination, and to monitor the efforts of host immune molecules such as APOBEC3F and 3G. The work in this thesis was focused on gleaning information from viral genomic sequence in order to better understand viral pathogenesis. The overall hypotheses of the research are presented here:

- Individuals that are highly exposed to HIV are more likely to be infected with a recombinant virus than those that are not highly exposed. Furthermore, within a group of highly exposed individuals, those more highly exposed will be more likely to be infected with recombinant virus.

- APOBEC-mediated hypermutation will be present at variable levels in a population of HIV-infected individuals. This HIV-1 proviral hypermutation will be an important factor to describe and quantify, as it will be associated with disease progression.

- Host factors are largely responsible for controlling APOBEC-mediated hypermutation of HIV, therefore in subjects superinfected with HIV, hypermutation levels will not change with superinfection with a distinct virus.

- Polymorphisms in host APOBEC3G genes are responsible for the high levels of APOBEC3G hypermutation activity observed in some individuals.

This study expanded the repertoire of full-length HIV-1 sequences from Kenya that are publically available. Further sequencing of a smaller region of HIV-1 on a larger sample set contributed to the understanding of recombination in a highly exposed population. Continued sequence analysis provided a convincing link between APOBEC3F/G hypermutation and HIV-1 disease progression *in vivo*. To further explore APOBEC-mediated hypermutation, intrapatient sequence diversity was examined pre and post HIV superinfection, which was the first analysis of superinfection and hypermutation. This thesis also provides the first description of APOBEC3G polymorphisms linked to HIV-1 proviral hypermutation sequence data. Together, the work described in this thesis has advanced the field by adding to the understanding of how innate host immunity can affect viral diversity, both of which in turn influence viral pathogenesis. These factors are critical to understand for the development of effective anti-HIV therapeutics, such as a vaccine.

1. Full-length HIV-1 Proviral Sequencing of Ten Highly Exposed Women Reveals a High Proportion of Intersubtype Recombinants

Previous HIV-1 sequence surveys in Kenya have largely focussed on particular genes or genomic regions of the viral genome (164,209,214,237,269,322), although one previous publication did survey full-length genomes (74). This has made it difficult to arrive at a good estimate of overall HIV recombination rates in Kenya and furthermore makes it difficult to compare rates between different at risk populations. The results of sequence surveys suggest that full-length sequence data allows for more accurate estimation and characterization of recombination. Partial genome sequencing, in comparison, tends to underestimate the total recombination present.

In results section 1, I examined HIV-1 sequence diversity and the prevalence of intersubtype recombination in a commercial sex worker cohort, by generating full-length proviral sequence from ten HIV positive subjects. To maximize the success rate, provirus was amplified by short term PBMC co-culture. Comparison of HIV-1 proviral sequences generated from limited co-culture of subject PBMCs to sequences generated directly from patient PBMCs showed the sequences to be highly homologous, with shared identity ranging from 95-97%. This observation validates sequencing proviruses that have been derived from limited primary co-culture for viral genotyping studies. Co-culture facilitates full-length sequence success as during the process the viral genome integrates multiple times into the host genome, resulting in a higher copy number and increased template (85). Although the data suggests that short term co-culture introduces minimal sequence change, multiple passages may result in greater HIV-1 sequence diversification

and so caution should be exercised when employing this strategy to not over-amplify the virus and thus needlessly allow the introduction of replication errors.

Previous HIV studies performed in Kenya have determined that the predominant circulating form is HIV-1 clade A1 (74,164,209,214,237,269,322). These studies examined several different groups, such as STI patients (164), pregnant or postpartum women (209,269,322), blood bank donors (74), and breast-feeding mothers (214). The findings are based on sequence data from a single or multiple regions of the genome, and only in one instance, full genome sequencing (74). Not surprisingly, publications that examined full HIV-1 genome sequence or multiple genomic regions identified higher levels of recombination. Studies that examined only a single HIV gene (*env* or *pol*) generated results varying from no recombination or very low levels of recombination (2.2% and 4%), to moderate levels of recombination (10% and 17%) (164,209,214,237,269). In comparison, a study that examined portions of both *gag* and *env* found a recombination rate of 25%, while full-length genomic sequencing found a recombination rate of 39% (74,322). For this thesis, full-length HIV-1 proviral sequences were generated from ten HIV positive sex workers from Nairobi, Kenya. Although the study size was small, our findings agree with previous studies from Kenya. HIV-1 clade A1 was the most prevalent subtype. In fact, all non-recombinant proviruses in this study were clade A1 (n = 5). Additionally, the level of recombination observed from full length sequencing in this population (50%), though higher than has previously been described in Kenya, is in general agreement with previous data based on full-length sequencing (74). Half of the genomes examined were unique, newly identified

intersubtype recombinants. It is interesting to note that no full-length clade C or D sequences were detected in this study, despite their presence in recombinant sequences identified in both this and other Kenyan surveys. Indeed, only two Kenyan full-length clade C and three Kenyan full-length clade D sequences have been published, making the source of the clade C and D in these recombinants unknown (36,74,80,298). One can speculate the source may be other recombinant viruses that contain clade C and D genomic regions, or the source may be viruses from other geographic regions.

Phylogenetic analysis of the non-recombinant clade A1 sequences from this cohort revealed significant viral diversity. This was observed despite the cohort members all practicing sex work in the same small community of Pumwani, Kenya, which suggests that the HIV-1 transmission events may have occurred in close geographical proximity (147). Phylogenetic analysis of these sequences with all the other Kenyan full-length clade A1 sequences available revealed that they did not form a distinct cluster, but rather were interspersed with other Kenyan sequences, as were other previously published sequences from this cohort (Figure 7). A study by Harris *et al* showed that full-length clade C HIV-1 sequence forms phylogenetic clusters that correlate with geography, although the authors noted that the relationship was not maintained when sub-genomic regions were examined (111). In contrast, we found no evidence of CSW specific circulating viruses, rather, someone who is highly exposed to HIV in this area is likely exposed to a variety of different strains from multiple geographic regions. This demonstrates that CSW cohorts encounter several HIV isolates and as such are a good sentinel population for observing HIV-1 diversity.

The recombinant subtypes identified in this study displayed even greater diversity. The extensive recombination (three to eight breakpoints) observed in the viral isolates from these highly exposed subjects suggests either multiple recombination events or a single recombination event between unidentified recombinant viruses circulating in this population (Figure 8). Although A1 is the most prevalent subtype in Kenya, only three of the five recombinant genomes contained the A1 subtype. A significant minority of HIV-1 sequences in Kenya are clade D and clade C. Estimates for clade D range from 29% (237) to 2.4% (74), while estimates for clade C range from 9% (164) to undetectable (237). Despite the lower overall prevalence of clade C and clade D viruses, four of the five recombinant isolates identified in this theses contained each of clade C and clade D components. This may reflect a change in the relative prevalence of subtypes in Kenya, highlighting the importance of longitudinal and cross-sectional sequence analysis. This alternatively suggests that the non-recombinant subtype distribution may not necessarily be similar to the distribution and relative proportion of subtypes that make up the recombinant forms. It is possible that a characteristic of subtype A1 HIV-1, or the immune response the subtype generates, may make it less prone to intersubtype recombination.

In addition to describing and characterizing the clades and recombination of full-length HIV-1 proviral sequence, this section of the thesis aimed to correlate increased sexual exposure among highly exposed individuals with recombination. However, exposure did not correlate with recombination, likely due to lack of power in the study due to an

insufficient sample size, thus leaving the aim to be tackled in the next section of this thesis.

2. High Prevalence of Genetically Similar HIV-1 Recombinant Viruses among Infected Sex Workers

Multiple studies have shown that individuals at high risk of acquiring multiple HIV-1 infections, such as sex workers and IDUs, are more likely to be infected with recombinant virus, especially if they participate in these high risk activities in an area where multiple HIV-1 strains co-circulate (12,13,117). We hypothesized that this would also be true within a highly exposed group: the most at risk and therefore the most highly exposed individuals would be more likely to be infected with recombinant HIV-1. Our full-length sequence analysis lacked the power to detect this difference. In order to increase our sample size, the length of the segment under study was decreased in this thesis section, although this reduces our ability to detect recombinants; thus, estimations of recombination generated in this thesis section will be conservative.

In results section 2, I continued my examination of HIV-1 sequence diversity and recombination in the ML cohort and also included subjects from the MCH cohort, also from Nairobi, Kenya. A 590 nucleotide fragment of the HIV-1 proviral genome, which includes *vpu* and the first 349 nucleotides of *env*, was examined for 240 subjects from these two cohorts. Sequence analysis of these fragments revealed that the majority were clade A1 (n = 167), while the rest were clade C (n = 15), clade D (n = 21) and recombinant (n = 37). This clade proportion is in general agreement with previous

Kenyan studies. The reported amount of recombinant sequences has varied widely, largely dependant on the size of the sequence segment examined, or whether full-length HIV-1 genome was analyzed. The proportion of recombinant sequences found in this study was 15%, but it is likely an underestimate of the true percentage, as only a relatively small segment of the HIV-1 genome was examined, to allow for a greater number of samples to be analysed.

This genomic section was chosen for in-depth analysis as the full-length sequencing (section 1) suggested the presence of mutations in *vpu*. The full-length HIV-1 sequence from ML1990 did not encode a functional gene product. Additionally, the HIV-1 sequence from ML1974 showed nucleotide ambiguity in *vpu,* some of which ambiguities could have encoded a truncating mutation. Although this genomic section was chosen to investigate mutations (explored in section 3 of this thesis), the region does span previously identified recombination breakpoints: of the five recombinant viruses identified in section 1 of this thesis, one has a breakpoint in this region. Similarly, of the sixteen recombinant isolates identified in the previous full-length survey from Kenya, four contained breakpoints in the region examined in this thesis section (74). Thus, though the size of the fragment examined is only ~5% of the total HIV-1 genome, it is predicted, based on the results of previous surveys, to detect ~20% of recombinant isolates.

The majority of the recombinant isolates identified in this study were composed of clades A1 and D (n = 19/37), the two most prevalent clades in Kenya

(74,164,209,214,237,269,322). Indeed, in the region examined, all but four recombinant isolates contained clade A1, the most prevalent clade. Thus the clade composition of the recombinant isolates resembles the overall composition of circulating clades in this region. This finding is in contrast to what was observed based on full-length sequence analysis, where the clade composition of the recombinant viruses was relatively deficient in clade A1. It is also interesting to note that a significant number of the isolates had two or more breakpoints in the relatively small genomic region (n = 8/37), highlighting the complexity of the recombination in this population.

Surprisingly, many of the recombinant isolates identified in this study formed distinct groups based upon common breakpoints between common clades. Within some of the groups, subgroups were formed based on additional breakpoints. Groups 1 and 3 were phylogenetically related, suggesting that they may represent as yet unidentified CRFs. Furthermore, a range of first HIV positive dates and sample dates were recorded for the subjects within these groups, spanning five to eighteen years and eight to sixteen years, respectively. This suggests that the strains are not closely epidemiologically linked, but rather may represent circulating recombinant isolates. In total, of the 37 HIV-1 isolates that were identified as recombinant, only four were unique. This questions the prevalence of unique recombinants in the pandemic, especially in people highly exposed to HIV-1 and in areas where multiple clades circulate, though it must be cautioned that only a segment of the HIV-1 genome was examined and thus the strains may have unique breakpoints outside the examined area. However, Konings *et al.* also speculated on the emergence and eventual evolution of URFs to CRFs in areas where multiple viral clades

co-circulate (145). Furthermore, new CRFs are continuously being described – in 2006 there were sixteen, yet now over forty have been published (189)(www.hiv.lanl.gov).

The epidemiological characteristics of the 240 subjects under study were examined using univariate analysis to determine if subjects infected with different viral clades had differing epidemiological characteristics (Table 9). Interestingly, of the characteristics examined, only active participation in sex work was significantly associated with clade distribution. Subjects that participated in sex work were less likely to be infected with clade C HIV-1, but were more likely to be infected with HIV that was recombinant in the region examined ($p = 0.038$). The women who actively participated in commercial sex work had, on average, 4.4 partners per day. In comparison, it has been previously reported that women in the lower risk cohort had an average of less than three partners in the last five years (174). As heterosexual transmission is these women's primary route of infection, individuals from the commercial sex worker cohort are clearly at a higher risk for HIV and other sexually transmitted infections. Thus, in agreement with other studies, individuals who were highly exposed to HIV-1 were more likely to be infected with recombinant virus, which is expected, as recombination is dependant upon infection with multiple strains of virus (12,13,117). This suggests that high-risk HIV transmitter groups may be driving global HIV diversity, thus making them an important target for intervention.

The relative enrichment of clade C infection in the women that did not participate in sex work is more difficult to understand. Clade C infection has spread faster than other

subtypes, and currently accounts for over half of all HIV-1 infections (10,79). The finding may suggest that participation in sex work results in exposure to a different subset of viruses, and highlights the need for continuing epidemiological surveys. The results of the previous thesis section, however, demonstrated that full-length HIV-1 clade A1 sequences from ML cohort members were phylogenetically interspersed with other clade A1 sequences from Kenya, which suggests that in fact sex workers are not infected with distinct viruses. However, there may be a difference between the way recombinant and non-recombinant isolates are distributed among risk groups, in addition to their enrichment within risk groups.

3. HIV-1 Proviral Hypermutation correlates with $CD4^+$ Count

APOBEC3F/G are innate immune molecules that combat HIV through a number of postulated mechanisms, one of which is hypermutation of the HIV-1 provirus (113,124,200,332). This proviral hypermutation is an important driver of viral diversity and can be used to identify previous APOBEC activity, as the provirus preserves an archive of the proteins' activity (113,142). Studies have examined HIV-1 provirus for hypermutation and attempted to correlate it with measures of HIV/AIDS disease progression, such as $CD4^+$ count and/or plasma viral load; however not all studies found a correlation (223,289). Thus there is significant controversy over both the role of hypermutation in HIV defence and its correlation with disease progression.

Our study of viral diversity in 240 HIV-1 proviral sequences identified premature stop codons in both of the examined open reading frames, *vpu* and *env,* for thirteen isolates.

Further examination revealed that the detrimental amino acids changes were in fact the result of guanine to adenine hypermutation in the DNA coding sequence. Additionally, we observed increased overall hypermutation, as measured by percentage adenine, in fourteen isolates. Examination of the corresponding viral RNA, however, showed a lack of detrimental mutations and a lack of elevated adenine proportion. This suggested that the mechanism of hypermutation was not global, such as an imbalance in nucleotide pools, but rather a targeted phenomenon, such as APOBEC mediated hypermutation. This finding agrees with a survey of over two thousand plasma virus sequences from nine HIV positive patients that found no evidence of hypermutation in RNA sequences, although over 6% of the proviral DNA sequences from the same patients were hypermutated (141). We agree with the authors' hypothesis that the hypermutated proviruses were unable to produce virus particles, due to the fatal mutations caused by the hypermutation, explaining the absence of hypermutated plasma RNA sequences. These non-hypermutated RNA sequences that were isolated from subjects infected with hypermutated provirus may have originated in an unsampled tissue compartment or reservoir where hypermutation has not occurred.

The sequence context of the observed hypermutation further suggested that the host antiviral APOBEC3F and 3G proteins were involved. The Hypermut 2.0 tool identified significant general APOBEC3-type hypermutation in these sequences, as well as specific APOBEC3F and APOBEC3G hypermutation; APOBEC3F causes GA to AA nucleotide changes, while APOBEC3G causes GG to AG transitions (163,328). It is interesting to note that ten samples showed significant Hypermut 2.0 hypermutation characteristic of

both APOBEC3G and of APOBEC3F, suggesting that these proteins may be working in concert in these patients (Table 10). Examination of the dinucleotide context for the guanine to adenine hypermutation in thirteen patients with dramatic hypermutation supported these observations. An elevated GG context of the hypermutation would be expected to be due to APOBEC3G involvement, while an elevated GA context would be caused by APOBEC3F involvement. In contrast, GC and GT context should be representative of the background mutation rate. In most of the examined sequences, both the GG and GA context were elevated compared to GC and GT, again suggesting co-involvement of APOBEC3F and 3G. Indeed, it has been shown that these proteins can form heteromultimers (112,309,310).

We based our analysis thus far on sequence derived directly from PCR amplicons, which we presumed to be reflective of the dominant proviral species. To ascertain the truth of our assumption, we sequenced a population of clones from twenty-three individuals. The clonal sequencing largely showed that the direct PCR sequence was indeed representative of the dominant proviral species (Figure 17). The presence of identical clones suggests that the proviruses arose as the result of host cell replication, which is plausible, as if the provirus is in fact a dead product with missing ORFs, it is not likely to kill the infected cell as the cell is not exposed to the cytopathic effects of replicating virus. Similar clones may be the result of progeny viruses produced by the same cell incorporating APOBEC3F/G proteins at similar amounts. Thus, when these viruses infect new cells and are reverse transcribed, APOBEC3F/G proteins from the same cell are acting on the nascent reverse transcripts, causing similar levels of hypermutation. There were a few

instances, however, where the cloned sequences showed distinct sub-populations with differing levels of hypermutation. The infecting viruses maybe have been generated in a cell with a low level of cytoplasmic APOBEC3F/G, such that not all nascent particles packaged APOBEC3F/G. Alternatively, the viruses may have been produced in different cellular compartments, where the expression levels of APOBEC3F/G were also different.

A subset of subjects that were determined to have dramatic levels of hypermutation (as defined by the presence of premature stop codons, elevated adenine proportion and a significant APOBEC3F/G sequence connotation to the hypermutation) had significantly higher $CD4^+$ cell counts than the other subjects (Figure 20). These subjects represent an extreme, where severely hypermutated provirus appears as the predominant proviral species. A study of nine patients with long term viral suppression due to HAART demonstrated that hypermutation could be identified, albeit at a minority, in the proviral sequence of all the subjects (141). Entertaining the likelihood, therefore, that all HIV positive patients have a variable amount of proviral hypermutation, adenine proportion in the dominant proviral sequence was used to estimate the level of hypermutation in all subjects and was indeed found to be positively correlated with $CD4^+$ cell count in all 240 subjects. This indicates that as the dominant proviral HIV-1 sequence is increasingly hypermutated, the subjects' $CD4^+$ counts similarly increase, for both patients with dramatic hypermutation and patients with subtle hypermutation. Although correlation does not necessarily indicate causation, this supports the hypothesis that with increasing hypermutation, fewer viable viral progeny are produced, protecting infected cells from direct viral cytopathic effects (141). This work thus also supports the anti-viral role of

APOBEC-mediated hypermutation, which has been recently debated (107,172,187,200,319).

An alternative explanation of the association of hypermutation with $CD4^+$ count is that a strong host CTL response targets cells that are infected with replication competent virus for destruction. As a result, cells infected with non-replication competent HIV-1, such as virus affected by the observed hypermutation, would be comparatively enriched in patients with strong, protective immune responses and therefore higher $CD4^+$ cell levels. However, if this were true, a general enrichment of mutation would be expected, due to the error-prone nature of HIV-1 reverse transcriptase, not just APOBEC-mediated hypermutation. Yet, all the proviruses that had detrimental mutations in Vpu and Env were hypermutated and all the mutations were due to guanine to adenine hypermutation. Furthermore, G to A changes in the dinucleotide context of GG and GA, hallmarks of APOBEC3G and 3F activity, respectively, were elevated above the dinucleotide context of GC and GT, which would represent non-APOBEC mediated mutations (163,328). In the absence of APOBEC-specific hypermutation, the levels would presumably be similar. Furthermore, Vif, which is located at a local minimum of APOBEC hypermutation, was not significantly mutated (277). This also would not be expected if another mutating factor had a significant effect on these viruses. Nevertheless, APOBEC may work in cooperation with strong CTL responses to control HIV-1 replication, highlighting the need for continued research about the cooperation between host innate, intrinsic and adaptive immunity.

The HIV-1 protein Vif is known to counteract APOBEC3F and APOBEC3G, but examination of viral Vif sequences did not reveal any obvious detrimental mutations that would explain the increased APOBEC3F/G hypermutation activity. Furthermore, hypermutation was not observed in the proviral *vif* sequences, although sequences were obtained from subjects that had significant proviral hypermutation in the *vpu/env* portion of the HIV-1 genome. This lack of Vif mutation is in contrast to a study by Pace *et al* of 127 clade B HIV-1 infected patients from an Australian cohort, which found a correlation between hypermutated proviral sequences and viral load, and attributed the relationship to stop mutations due to hypermutation in the infecting viral Vif that disabled this viral defence against cellular APOBEC (223). The authors note, however, that it is difficult to determine if these mutations are the cause of the increased levels of observed hypermutation, or an effect (223). Priming for DNA synthesis occurs at fixed locations in the HIV-1 genome and therefore not all sites are single-stranded for the same period of time, suggesting that as APOBEC3G acts only on single-stranded DNA, hypermutation will not occur uniformly across the genome (124,277). In fact, Vif is located at a local minimum of predicted mutation, suggesting a mechanism that explains our finding a lack of *vif* hypermutation and consequent lack of detrimental amino acid mutations (277).

We hypothesize as a mechanism to explain our observations that subjects infected with more dramatically hypermutated HIV-1 provirus have elevated APOBEC3F/G activity. Potential causes for this elevated activity include upregulation of APOBEC3F/G, such as by IFN-α, which has been recently shown to be a potent inducer, allowing APOBEC3F/G to overcome Vif inhibition (28,225,254,279) or the enzyme may exist in a more

enzymatically active form (for example, a higher proportion of APOBEC3G may be present in its low molecular mass, active form) (51). Alternatively, the infecting viruses may encode a Vif that is inefficient or unable to target APOBEC3F and APOBEC3G for degradation or prevent their packaging into nascent viral particles, although our data does not support this hypothesis (58,165,172,182,192,263,284,327,328). In either scenario, we theorize that as a result of increased APOBEC3F/G and/or decreased Vif levels/activity, the newly budded viral particles contain higher levels of APOBEC3F and APOBEC3G, ultimately resulting in hypermutation during infection of new cells. It seems likely that the resulting provirus is sufficiently hypermutated that replication competent, hypermutated progeny cannot be generated, as illustrated by the lack of hypermutated RNA sequences. Indeed, the presence of non-hypermutated viral RNA sequences argues strongly in favour of host factors controlling the hypermutation; APOBEC3F and 3G may be differentially expressed in PBMCs compared to non-circulating HIV-target cells. Hypermutation in HIV-1 *tat* may render the viral activator of transcription non-functional, reducing the amount of hypermutated and possibly truncated transcripts, which in turn would reduce the number of viral epitopes presented on the infected cell's surface. Importantly, these factors would lead to the cells infected with hypermutated provirus not being subjected to HIV's direct cytopathic effects and would decrease their likelihood of being targeted by CTLs and for immune activation induced apoptosis, likely explaining the observed higher $CD4^+$ counts and presumable decrease in disease progression in patients with increased proviral hypermutation (141).

Conflicting publications highlight that the role of APOBEC3 in HIV disease progression has not yet been resolved. In their study using stimulated PBMCs, Jin *et al.* found that APOBEC3G mRNA levels were positively correlated with patient $CD4^+$ count and inversely correlated with viral load, while Cho *et al.*, testing unstimulated PBMCs, found that neither APOBEC3G nor APOBEC3F mRNA correlated with $CD4^+$ count and viral load (53,129). Hypermutation observed in the present study is likely a direct effect of APOBEC3F/G cytidine deamination, whereas mRNA expression levels, which were examined in these previous studies, may not directly correlate with enzymatic activity, due to translational and post-translational regulation. This direct assessment of sequence hypermutation, likely the result of APOBEC3F/G cytidine deaminase activity, found that as the level of hypermutation in the subjects' predominant proviral sequence increased, so did the subjects' $CD4^+$ counts, suggesting an *in vivo* role for APOBEC in disease progression.

APOBEC3F/G proviral hypermutation likely exists in a spectrum. The thirteen subjects with dramatically hypermutated provirus had detectable viral loads, intact viral RNA sequence and significantly depressed $CD4^+$ counts due to HIV infection compared to HIV uninfected patients, suggesting that there may be a tissue or cellular compartment where APOBEC3 is not packaged into viral progeny, leading to the production of non-hypermutated virus, which in turn leads to non-hypermutated infectious viral progeny, sustaining the infection. On the other end of the spectrum, patients without hypermutation in the dominant provirus may have minority sequences that are hypermutated (141). Indeed, the examination of clonal sequences indicated some degree

of variability within patients. The data presented in this section of the thesis illustrates a correlative relationship between hypermutation, likely the direct result of APOBEC3F/G activity and $CD4^+$ cell count *in vivo*, in the absence of obvious Vif polymorphism. These findings highlight the potential for enhancing host APOBEC3F and APOBEC3G for therapeutic purposes, and suggest that even small increases in APOBEC activity may attenuate HIV-1 replication and disease progression.

4. Longitudinal Analysis of Subjects Superinfected with HIV-1 Reveals Changes in Hypermutation Levels in a Subset of Individuals

Superinfection studies are a powerful tool for examining the effects of infection within a host. The role of APOBEC3F/G in AIDS disease progression is neither well understood nor described. Furthermore, the key determinants that allow this host defence system, in a minority of cases, to overcome the viral counter-defence are also unknown. Thus, examining the levels of APOBEC-mediated hypermutation in subjects that have been superinfected with HIV will generate relevant information that can be applied to these important questions. Based on the results of the previous thesis section, we hypothesize that host factors are the major components that determine if APOBEC3F/G can be overcome by viral Vif, and thus predict that HIV proviral hypermutation levels would be unaffected by HIV superinfection.

Our analyses of HIV superinfection is based on data made available from a study that identified seven individuals superinfected with HIV in a prospective cohort of highly exposed women from Mombasa, Kenya (228). Proviral *env* sequences were available

both prior to and post superinfection. In all seven cases, the superinfecting virus was phylogenetically distinct from the original infecting virus, separated by bootstrap values ranging from 90 to 100 percent. This distinctness allowed us to assume that the superinfecting virus was sufficiently different from the original infecting virus that the viruses may be capable of exerting distinct pressures on the host. An important caveat, however, is that we determined differences based on the *env* region of the virus, while we are most interested in potential *vif* differences.

In results section 3.6 we looked for population-level differences in hypermutation by examining proviral sequences from 240 individuals for proviral hypermutation. Using a similar approach, we looked for differences in hypermutation within an individual, pre and post superinfection in results section 4.7. For many of the seven examined subjects, hypermutation levels were constant over the course of infection. Subject QA413 and QD022, however, were superinfected with viruses that maintained higher hypermutation levels than the original infecting virus. For subject QA413, this was confirmed by both adenine proportion and Hypermut 2.0 rate ratio, while for subject QD022, the higher hypermutation levels were only indicated by rate ratio. Subject QB045 also showed changes in hypermutation, but it is less clear if it is caused by the superinfecting virus, or by the superinfection event, as the hypermutation levels in the original virus seemed changed upon superinfection. Perhaps superinfection in this subject triggered a cull of cells infected with comparatively hypermutated virus, or perhaps the replication pressure exerted by an additional virus selected for the most fit (less hypermutated) viruses.

Three post-superinfection follow-up data points were available for subject QA413, which spanned 339 days (i.e. nearly a year), while two post-superinfection follow-ups were available for subject QD022, which spanned 795 days (i.e. over two years). This is doubtlessly long enough for any effects residual from the viral donor to have been washed out, as the average generation time for HIV-1 *in vivo* is 2.6 days (226). The maintenance of a distinct level of hypermutation argues strongly that the viruses are having markedly different effects on combating the host APOBEC defence mechanism; the original virus seems to be adequately quenching the APOBEC defence, whereas the superinfecting virus is comparably vulnerable. These observations are best explained by hypothesizing that the superinfecting virus had determinants, possibly in Vif, distinct from the original virus that were unable to combat the host APOBEC3F/G defence mechanisms. Returning to our original hypothesis, it seems that in this case, contrary to what we predicted, viral determinants were responsible at least in part for determining the level of APOBEC3F/G hypermutation.

This difference in susceptibility to the host APOBEC defence mechanism leads one to speculate if over time the superinfecting virus would eventually succumb to APOBEC, allowing the original virus to out-compete it. Assuming non-biased sampling of the proviral sequences, this does not seem to have occurred in either subject. Subject QA413 shows an increasing proportion of hypermutated superinfecting viral sequences, while in subject QD022 only the hypermutated superinfecting viral sequences are present post superinfection. We hypothesize that the superinfecting virus is succumbing to the host APOBEC defence during the follow-up period; once the virus entered host cells, it

became hypermutated to the extent that the provirus was stuck. The cells infected with non-functional virus would not be exposed to the cytopathic effects of replicating virus, resulting in relative enrichment for cells harbouring this dead-end product. To test this hypothesis, the replicating virus fraction could be sampled (i.e. the plasma RNA) to determine the relative proportions of the original and superinfecting virus.

This examination of superinfection cases revealed that in at least two subjects, viral factors were responsible for controlling hypermutation levels, as the superinfecting virus had significantly higher levels of hypermutation than the original infecting virus. In the majority of the other examined cases, however, host factors seemed more important in controlling hypermutation levels. The control of HIV-1 proviral hypermutation is important to understand so that APOBEC-mediated hypermutation may be exploited for therapeutic purposes.

5. Discovery of SNPs Associated with Differing Levels of HIV-1 Proviral Hypermutation by Pyrosequencing the APOBEC3G Gene

APOBEC3G restricts HIV-1 replication by causing hypermutation of the viral genome as it is reverse transcribed from RNA to DNA, and may also have hypermutation-independent antiviral properties, though these are less well defined. Previous studies have examined APOBEC3G polymorphisms and linked them to increased HIV-1 infection and to increased disease progression (5,291). However, no study has linked APOBEC3G polymorphisms to levels of HIV-1 proviral hypermutation. We hypothesized that the increased levels of HIV-1 proviral hypermutation identified in

results section 3 of this thesis would be correlated with polymorphisms in the APOBEC3G gene of the affected subject.

We examined APOBEC3G genes from HIV-1 infected Kenyan subjects with significantly hypermutated HIV-1 provirus, intermediately hypermutated HIV-1 provirus and non-hypermutated HIV-1 provirus. In this population of 102 individuals, we found 89 different SNPs in the 13.5 kb APOBEC3G region. Of these, 39 had not been previously described. The majority of the 89 SNPs identified in this study were found in non-coding regions. This is also true of SNPs identified in other studies; only a minority of polymorphisms have been identified in APOBEC3G exons. Half of the SNPs identified within exons in this study were synonymous. This illustrates the important function of APOBEC3G; SNPs in the coding regions which cause amino acid changes do not seem to be well tolerated.

The SNPs identified in this study were examined to see if any were expressed in differential frequencies based on the level of HIV-1 proviral hypermutation in the corresponding subject. One SNP was found to be over-represented in the individuals infected with highly hypermutated HIV-1 provirus, while another two SNPs were found to be under-represented in these individuals.

The G18863080C SNP is located in the 5' region of the APOBEC3G gene (Table 19). This SNP was over-represented in the individuals infected with highly hypermutated HIV-1 provirus, and was present in both alleles for all nine of these examined subjects. It

was also present in the majority of alleles in the pools of intermediately and non-hypermutated samples, but at a significantly lower frequency ($p = 0.0413$). Interestingly, the G18863080C SNP destroys a potential sterol regulatory element binding protein (SREBP) site, as identified by SiteGA (158). SREBPs are transcription factors; destruction of the site would presumably lead to a decrease in upregulation of the APOBEC3G gene by this transcription factor. However, this SNP was associated with increased hypermutation, and therefore likely increased APOBEC3G expression. It is possible that destruction of this SREBP site may abolish steric hindrance to nearby sites, generated by SREBP binding the DNA upstream of the APOBEC3G gene and therefore blocking access to the neighbouring genomic regions. Thus, destruction of the SREBP site may allow access of more potent transcription factors to the DNA.

The G18862077C mutation is located in the 5' region of the APOBEC3G gene (Table 18). The SNP was under-represented in the individuals infected with highly hypermutated HIV-1 provirus, and in fact, was not present in any of these examined individuals ($p = 0.0276$). The presence of this SNP was thus associated with decreased APOBEC3G hypermutation activity; possibly due to disruption of a transcription factor binding site upstream of the APOBEC3G gene. However, a scan of the region with the SiteGA transcription factor binding sites recognition program did not indicate any changes in known binding sites with this SNP (158). Alternatively, this SNP may be linked to another polymorphism, outside the examined region, with an important function.

The C18870730T SNP is located in intron 5 of the APOBEC3G gene (Table 18). This SNP was also under-represented in the individuals infected with highly hypermutated HIV-1 provirus and associated with decreased APOBEC3G hypermutation activity ($p = 0.0437$). Interestingly, this SNP has not been previously described. Intron 5 is 2.4 kb long; the SNP is located over 300 bp from the end of exon 5. The polymorphism may cause alternative splicing, leading to a form of the protein with less hypermutation activity. However, analysis of the region using the NetGene 2 server to predict donor and acceptor human mRNA splice sites did not reveal any changes with this SNP (33). This SNP may alternatively be linked to another polymorphism that causes decreased APOBEC3G activity.

The SNP that was over-represented in the subjects infected with highly hypermutated virus, G18863080C, is located 1 kb from the other 5' region SNP, G18862077C, which is significantly distributed among the different hypermutation groups. The sum of these two SNPs' allelic frequencies is close to 1 in each of the different sub-populations. It is thus tempting to hypothesize one of these SNPs is always or often present in an allele, but that the two SNPs are not found in the same allele. However, due to the pooling of samples, this cannot be determined from the current data. It is also an important limitation of this data that the Chi Square analysis for the SNPs was performed on estimated numbers of alleles, which may not be the same as the actual number of alleles with each SNP.

6. Summary

This thesis has used sequence analysis techniques to answer three major questions that address the overall hypothesis: HIV-1 sequence diversity is an important factor of HIV-1 pathogenesis.

1. Is HIV exposure correlated with HIV recombination?

   To answer this first question, we determined that half of the examined full-length HIV-1 proviral sequences were unique intersubtype recombinants. However, recombination was not found to correlate with exposure in this group, as measured by years involved in commercial sex work. Thus, we expanded our study to 240 individuals, representing both highly exposed and non-highly exposed subjects and examined a smaller, 590 nucleotide region that included *vpu* and part of *env*. We demonstrated that recombinant isolates were more likely to be isolated from the study subjects that were actively involved in sex work, in agreement with previous studies (12,13,117). However, within the highly exposed population, individuals at increased risk of HIV-acquisition were not more likely to be infected with recombinant HIV. This suggests that all high-risk individuals are an important source for viral diversity, and thus an important group to target with prevention strategies to minimize increasing global viral diversity, which presents a challenge to developing an effective HIV vaccine (96).

2. Does HIV proviral hypermutation correlate with measures of disease progression?

   To answer the second question, we showed that thirteen individuals had highly hypermutated proviral HIV-1 sequences. These thirteen individuals also had significantly higher $CD4^+$ counts than the other subjects ($p = 0.0052$). A similar trend was observed for $CD4^+$ percentage and viral load. Furthermore, adenine proportion for all individuals correlated with $CD4^+$ count ($p = 0.041$). Thus, this thesis found that APOBEC-mediated hypermutation did indeed correlate with measures of disease progression. Furthermore, we suggest that increasing APOBEC3G hypermutation activity may protect against HIV disease progression, and thus is an attractive target for developing therapeutic and preventative strategies.

3. Are hypermutation levels controlled by the virus, or the host?

   To answer the third question, we determined that there were no dramatic differences in the Vif sequences from highly hypermutated HIV-1 provirus and non-hypermutated provirus. We furthermore demonstrated that in seven individuals superinfected with HIV, only two subjects clearly showed different hypermutation levels in the original and superinfecting virus, indicating that viral factors (i.e. Vif) were controlling hypermutation. Returning our attention to the individuals identified with highly hypermutated HIV-1 provirus, we found three SNPs that were significantly associated with altered HIV-1 proviral hypermutation. This indicates that host factors may play an important role in controlling APOBEC-mediated HIV-1 proviral hypermutation, which until now, is a topic unexplored in the literature.

7. Future Work

The work described in this thesis answers interesting questions involving HIV genetic variability and has contributed to the description of HIV viral diversity in Kenya and in high-risk groups. It has provided evidence that APOBEC-mediated hypermutation is associated with disease progression (cited by four publications to date, including a review by a leader in the field) and suggested that this hypermutation is controlled by host factors such as increased expression of APOBEC3G. Importantly, this work also leads to questions of interest and relevance which can be used to direct future research on this topic.

Certainly, more HIV-1 sequence data should be collected, especially full-length genomic sequences from regions, such as Kenya, that are highly infected but not highly sampled. An interesting and novel approach to examining clade and recombination diversity would be to overlay sequence data with sexual network data. Previous studies and this thesis have shown that recombinant HIV is more likely to be found in highly exposed populations. However, this thesis has shown that higher exposure within a highly exposed population is not associated with increased recombination. Furthermore, this thesis demonstrated that the non-recombinant full-length HIV-1 sequences sampled from a highly exposed population are phylogenetically interspersed with sequences from the general population. This begs the question – where is the disconnect? Are people highly exposed to HIV more likely to be infected with recombinant HIV because they are in a distinct sexual network, where multiple subtypes and recombinant isolates circulate (i.e. the recombination is occurring at a population level)? Or, are people highly exposed to

HIV more likely to be infected with recombinant HIV because of their individual exposures (i.e. the recombination is occurring at the individual level due to super/co-infection)? The combination of sequence and sexual network data may be able to help answer these questions, which would provide important information about the transmission of viral isolates.

Regarding hypermutation, there are number of topics to be explored. This thesis described a non-biased patient sampling approach to examining the correlation between APOBEC-mediated hypermutation and disease progression. It would also be valuable to compare APOBEC-mediated hypermutation and APOBEC expression levels between subjects of known disease progression groups, such as rapid progressors, long-term non-progressors, elite controllers and even resistant individuals. APOBEC3G targets other viruses for hypermutation, such as hepatitis B virus; the hypermutation levels in this virus could be examined in the absence of HIV infection. This examination would indicate the level of disease attenuation capable of being attained via altered APOBEC activity.

The genetic variability of APOBEC3G was explored in this thesis; three SNPs were found to be significantly associated with HIV-1 proviral hypermutation levels. This finding should be validated by expanding the population size and sequencing the SNP regions on an individual basis to determine linkage. This work could also be expanded to a different population, to determine if the association is true for a different ethnic background. To characterize the role of the significantly associated SNPs, exon arrays could be used to determine if alternative splicing is occurring. Additionally, binding

assays could be performed to determine if indeed the G18863080C SNP destroys a SREBP site, and if this destruction allows for increased binding by other transcription factors at nearby sites.

The genetic variability in APOBEC3F would also doubtlessly be interesting and informative, as would genetic variability of the other, less well studied APOBEC3 proteins. Furthermore, Cullin5, ElonginB and ElonginC are important for Vif/APOBEC interaction, and variability in these proteins could also contribute to APOBEC's anti-viral activity.

The anti-viral function of the APOBEC proteins has only recently been described. Thus, there is still much to be characterized in order to fully understand the role and mechanism of these innate proteins in attenuating HIV-1 replication. The findings will be important for exploiting this natural defence mechanism for anti-HIV therapeutic purposes.

In closing, this thesis showed that viral diversity, as measured by clade/recombination status and proviral hypermutation, is important for understanding HIV disease. Furthermore, this thesis suggests that host factors regulating viral diversity may play an important role in protection.

References

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

2. **Aiken, C.** 2009. Cell-Free Assays for HIV-1 Uncoating. Methods Mol. Biol. **485**:41-53.

3. **Alce, T. M. and W. Popik**. 2004. APOBEC3G is incorporated into virus-like particles by a direct interaction with HIV-1 Gag nucleocapsid protein. J. Biol. Chem. **279**:34083-34086.

4. **Ameisen, J. C.** 1994. Programmed cell death (apoptosis) and cell survival regulation: relevance to AIDS and cancer. AIDS **8**:1197-1213.

5. **An, P., G. Bleiber, P. Duggal, G. Nelson, M. May, B. Mangeat, I. Alobwede, D. Trono, D. Vlahov, S. Donfield, J. J. Goedert, J. Phair, S. Buchbinder, S. J. O'Brien, A. Telenti, and C. A. Winkler**. 2004. APOBEC3G genetic variants and their influence on the progression to AIDS. J. Virol. **78**:11070-11076.

6. **Andersen, J. L., J. L. DeHart, E. S. Zimmerman, O. Ardon, B. Kim, G. Jacquot, S. Benichou, and V. Planelles**. 2006. HIV-1 Vpr-induced apoptosis is cell cycle dependent and requires Bax but not ANT. PLoS. Pathog. **2**:e127.

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

8. **Arakawa, H., J. Hauschild, and J. M. Buerstedde**. 2002. Requirement of the activation-induced deaminase (AID) gene for immunoglobulin gene conversion. Science **295**:1301-1306.

9. **Arien, K. K., A. Abraha, M. E. Quinones-Mateu, L. Kestens, G. Vanham, and E. J. Arts**. 2005. The replicative fitness of primary human immunodeficiency virus type 1 (HIV-1) group M, HIV-1 group O, and HIV-2 isolates. J. Virol. **79**:8979-8990.

10. **Arien, K. K., G. Vanham, and E. J. Arts**. 2007. Is HIV-1 evolving to a less virulent form in humans? Nat. Rev. Microbiol. **5**:141-151.

11. **Arora, V. K., R. P. Molina, J. L. Foster, J. L. Blakemore, J. Chernoff, B. L. Fredericksen, and J. V. Garcia**. 2000. Lentivirus Nef specifically activates Pak2. J. Virol. **74**:11081-11087.

12. **Arroyo, M. A., M. Hoelscher, W. Sateren, E. Samky, L. Maboko, O. Hoffmann, G. Kijak, M. Robb, D. L. Birx, and F. E. McCutchan**. 2005. HIV-1 diversity and prevalence differ between urban and rural areas in the Mbeya region of Tanzania. AIDS **19**:1517-1524.

13. **Arroyo, M. A., W. B. Sateren, D. Serwadda, R. H. Gray, M. J. Wawer, N. K. Sewankambo, N. Kiwanuka, G. Kigozi, F. Wabwire-Mangen, M. Eller, L. A. Eller, D. L. Birx, M. L. Robb, and F. E. McCutchan**. 2006. Higher HIV-1 incidence and genetic complexity along main roads in Rakai District, Uganda. J. Acquir. Immune. Defic. Syndr. **43**:440-445.

14. **Arts, E. J. and S. F. Le Grice**. 1998. Interaction of retroviral reverse transcriptase with template-primer duplexes during replication. Prog. Nucleic Acid Res. Mol. Biol. **58**:339-393.

15. **Bach, D., S. Peddi, B. Mangeat, A. Lakkaraju, K. Strub, and D. Trono**. 2008. Characterization of APOBEC3G binding to 7SL RNA. Retrovirology. **5**:54.

16. **Bailes, E., F. Gao, F. Bibollet-Ruche, V. Courgnaud, M. Peeters, P. A. Marx, B. H. Hahn, and P. M. Sharp**. 2003. Hybrid origin of SIV in chimpanzees. Science **300**:1713.

17. **Baird, H. A., Y. Gao, R. Galetto, M. Lalonde, R. M. Anthony, V. Giacomoni, M. Abreha, J. J. Destefano, M. Negroni, and E. J. Arts**. 2006. Influence of sequence identity and unique breakpoints on the frequency of intersubtype HIV-1 recombination. Retrovirology. **3**:91.

18. **Barraud, P., J. C. Paillart, R. Marquet, and C. Tisne**. 2008. Advances in the structural understanding of Vif proteins. Curr. HIV. Res. **6**:91-99.

19. **Barre-Sinoussi, F., J. C. Chermann, F. Rey, M. T. Nugeyre, S. Chamaret, J. Gruest, C. Dauguet, C. Axler-Blin, F. Vezinet-Brun, C. Rouzioux, W. Rozenbaum, and L. Montagnier**. 1983. Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). Science **220**:868-871.

20. **Baumert, T. F., C. Rosler, M. H. Malim, and W. F. von**. 2007. Hepatitis B virus DNA is subject to extensive editing by the human deaminase APOBEC3C. Hepatology **46**:682-689.

21. **Beale, R. C., S. K. Petersen-Mahrt, I. N. Watt, R. S. Harris, C. Rada, and M. S. Neuberger**. 2004. Comparison of the differential context-dependence of DNA deamination by APOBEC enzymes: correlation with mutation spectra in vivo. J. Mol. Biol. **337**:585-596.

22. **Berkhout, B., A. Grigoriev, M. Bakker, and V. V. Lukashov**. 2002. Codon and amino acid usage in retroviral genomes is consistent with virus-specific nucleotide pressure. AIDS Res. Hum. Retroviruses **18**:133-141.

23. **Biasin, M., L. Piacentini, C. S. Lo, Y. Kanari, G. Magri, D. Trabattoni, V. Naddeo, L. Lopalco, A. Clivio, E. Cesana, F. Fasano, C. Bergamaschi, F. Mazzotta, M. Miyazawa, and M. Clerici**. 2007. Apolipoprotein B mRNA-editing enzyme, catalytic polypeptide-like 3G: a possible role in the resistance to HIV of HIV-exposed seronegative individuals. J. Infect. Dis. **195**:960-964.

24. **Bishop, K. N., R. K. Holmes, A. M. Sheehy, N. O. Davidson, S. J. Cho, and M. H. Malim**. 2004. Cytidine deamination of retroviral DNA by diverse APOBEC proteins. Curr. Biol. **14**:1392-1396.

25. **Bogerd, H. P., B. P. Doehle, H. L. Wiegand, and B. R. Cullen**. 2004. A single amino acid difference in the host APOBEC3G protein controls the primate species specificity of HIV type 1 virion infectivity factor. Proc. Natl. Acad. Sci. U. S. A **101**:3770-3774.

26. **Bogerd, H. P., H. L. Wiegand, B. P. Doehle, K. K. Lueders, and B. R. Cullen**. 2006. APOBEC3A and APOBEC3B are potent inhibitors of LTR-retrotransposon function in human cells. Nucleic Acids Res. **34**:89-95.

27. **Bogerd, H. P., H. L. Wiegand, A. E. Hulme, J. L. Garcia-Perez, K. S. O'Shea, J. V. Moran, and B. R. Cullen**. 2006. Cellular inhibitors of long interspersed element 1 and Alu retrotransposition. Proc. Natl. Acad. Sci. U. S. A **103**:8780-8785.

28. **Bonvin, M., F. Achermann, I. Greeve, D. Stroka, A. Keogh, D. Inderbitzin, D. Candinas, P. Sommer, S. Wain-Hobson, J. P. Vartanian, and J. Greeve**. 2006. Interferon-inducible expression of APOBEC3 editing enzymes in human hepatocytes and inhibition of hepatitis B virus replication. Hepatology **43**:1364-1374.

29. **Boom, R., C. J. Sol, M. M. Salimans, C. L. Jansen, P. M. Wertheim-van Dillen, and N. J. van der**. 1990. Rapid and simple method for purification of nucleic acids. J. Clin. Microbiol. **28**:495-503.

30. **Bour, S., R. Geleziunas, and M. A. Wainberg**. 1994. The role of CD4 and its downmodulation in establishment and maintenance of HIV-1 infection. Immunol. Rev. **140**:147-171.

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

32. **Braddick, M. R., J. K. Kreiss, J. B. Embree, P. Datta, J. O. Ndinya-Achola, H. Pamba, G. Maitha, P. L. Roberts, T. C. Quinn, and K. K. Holmes**. 1990. Impact of maternal HIV infection on obstetrical and early neonatal outcome. AIDS **4**:1001-1005.

33. **Brunak, S., J. Engelbrecht, and S. Knudsen**. 1991. Prediction of human mRNA donor and acceptor sites from the DNA sequence. J. Mol. Biol. **220**:49-65.

34. **Buchbinder, S. P., M. H. Katz, N. A. Hessol, P. M. O'Malley, and S. D. Holmberg**. 1994. Long-term HIV-1 infection without immunologic progression. AIDS **8**:1123-1128.

35. **Burnett, A. and P. Spearman**. 2007. APOBEC3G multimers are recruited to the plasma membrane for packaging into human immunodeficiency virus type 1 virus-like particles in an RNA-dependent process requiring the NC basic linker. J. Virol. **81**:5000-5013.

36. **Burns, C. C., L. M. Gleason, A. Mozaffarian, C. Giachetti, J. K. Carr, and J. Overbaugh**. 2002. Sequence variability of the integrase protein from a diverse collection of HIV type 1 isolates representing several subtypes. AIDS Res. Hum. Retroviruses **18**:1031-1041.

37. **Carlson, J. M. and Z. L. Brumme**. 2008. HIV evolution in response to HLA-restricted CTL selection pressures: a population-based perspective. Microbes. Infect. **10**:455-461.

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

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

40. **Castro, K. G., J. W. Ward, L. Slutsker, J. W. Buehler, H. W. Jaffe, R. L. Berkelman, and J. W. Curran**. 1992. 1993 Revised Classification System for HIV Infection and Expanded Surveillance Case Definintion for AIDS Among Adolescents and Adults. MMWR **41**:1-19.

41. **Cen, S., F. Guo, M. Niu, J. Saadatmand, J. Deflassieux, and L. Kleiman**. 2004. The interaction between HIV-1 Gag and APOBEC3G. J. Biol. Chem. **279**:33177-33184.

42. **Chan, D. C., D. Fass, J. M. Berger, and P. S. Kim**. 1997. Core structure of gp41 from the HIV envelope glycoprotein. Cell **89**:263-273.

43. **Charpentier, C., T. Nora, O. Tenaillon, F. Clavel, and A. J. Hance**. 2006. Extensive recombination among human immunodeficiency virus type 1 quasispecies makes an important contribution to viral diversity in individual patients. J. Virol. **80**:2472-2482.

44. **Chen, H., C. E. Lilley, Q. Yu, D. V. Lee, J. Chou, I. Narvaiza, N. R. Landau, and M. D. Weitzman**. 2006. APOBEC3A is a potent inhibitor of adeno-associated virus and retrotransposons. Curr. Biol. **16**:480-485.

45. **Chen, J., T. D. Rhodes, and W. S. Hu**. 2005. Comparison of the genetic recombination rates of human immunodeficiency virus type 1 in macrophages and T cells. J. Virol. **79**:9337-9340.

46. **Chen, K., J. Huang, C. Zhang, S. Huang, G. Nunnari, F. X. Wang, X. Tong, L. Gao, K. Nikisher, and H. Zhang**. 2006. Alpha interferon potently enhances the anti-human immunodeficiency virus type 1 activity of APOBEC3G in resting primary CD4 T cells. J. Virol. **80**:7645-7657.

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

48. **Chen, Z., A. Luckay, D. L. Sodora, P. Telfer, P. Reed, A. Gettie, J. M. Kanu, R. F. Sadek, J. Yee, D. D. Ho, L. Zhang, and P. A. Marx**. 1997. Human immunodeficiency virus type 2 (HIV-2) seroprevalence and characterization of a distinct HIV-2 genetic subtype from the natural range of simian immunodeficiency virus-infected sooty mangabeys. J. Virol. **71**:3953-3960.

49. **Chenna, R., H. Sugawara, T. Koike, R. Lopez, T. J. Gibson, D. G. Higgins, and J. D. Thompson**. 2003. Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res. **31**:3497-3500.

50. **Cheynier, R., S. Gratton, J. P. Vartanian, A. Meyerhans, and S. Wain-Hobson**. 1997. G --> A hypermutation does not result from polymerase chain reaction. AIDS Res. Hum. Retroviruses **13**:985-986.

51. **Chiu, Y. L., V. B. Soros, J. F. Kreisberg, K. Stopak, W. Yonemoto, and W. C. Greene**. 2005. Cellular APOBEC3G restricts HIV-1 infection in resting CD4+ T cells. Nature **435**:108-114.

52. **Chiu, Y. L., H. E. Witkowska, S. C. Hall, M. Santiago, V. B. Soros, C. Esnault, T. Heidmann, and W. C. Greene**. 2006. High-molecular-mass APOBEC3G complexes restrict Alu retrotransposition. Proc. Natl. Acad. Sci. U. S. A **103**:15588-15593.

53. **Cho, S. J., H. Drechsler, R. C. Burke, M. Q. Arens, W. Powderly, and N. O. Davidson**. 2006. APOBEC3F and APOBEC3G mRNA levels do not correlate with human immunodeficiency virus type 1 plasma viremia or CD4+ T-cell count. J. Virol. **80**:2069-2072.

54. **Chowers, M. Y., C. A. Spina, T. J. Kwoh, N. J. Fitch, D. D. Richman, and J. C. Guatelli**. 1994. Optimal infectivity in vitro of human immunodeficiency virus type 1 requires an intact nef gene. J. Virol. **68**:2906-2914.

55. **Christ, F., W. Thys, R. J. De, R. Gijsbers, A. Albanese, D. Arosio, S. Emiliani, J. C. Rain, R. Benarous, A. Cereseto, and Z. Debyser**. 2008. Transportin-SR2 imports HIV into the nucleus. Curr. Biol. **18**:1192-1202.

56. **Coffin, J. M., S. H. Hughes, and H. E. Varmus**. 1997. Retroviruses. Cold Spring Harbor Laboratory Press.

57. **Cohen, J.** 2008. AIDS research. Treat everyone now? A 'radical' model to stop HIV's spread. Science **322**:1453.

58. **Conticello, S. G., R. S. Harris, and M. S. Neuberger**. 2003. The Vif protein of HIV triggers degradation of the human antiretroviral DNA deaminase APOBEC3G. Curr. Biol. **13**:2009-2013.

59. **Conticello, S. G., C. J. Thomas, S. K. Petersen-Mahrt, and M. S. Neuberger**. 2005. Evolution of the AID/APOBEC family of polynucleotide (deoxy)cytidine deaminases. Mol. Biol. Evol. **22**:367-377.

60. **Cornelissen, M., G. Mulder-Kampinga, J. Veenstra, F. Zorgdrager, C. Kuiken, S. Hartman, J. Dekker, H. L. van der, C. Sol, R. Coutinho, and .** 1995. Syncytium-inducing (SI) phenotype suppression at seroconversion after intramuscular inoculation of a non-syncytium-inducing/SI phenotypically mixed human immunodeficiency virus population. J. Virol. **69**:1810-1818.

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

62. **Dang, Y., X. Wang, W. J. Esselman, and Y. H. Zheng**. 2006. Identification of APOBEC3DE as another antiretroviral factor from the human APOBEC family. J. Virol. **80**:10522-10533.

63. **Daniel, M. D., N. L. Letvin, N. W. King, M. Kannagi, P. K. Sehgal, R. D. Hunt, P. J. Kanki, M. Essex, and R. C. Desrosiers**. 1985. Isolation of T-cell tropic HTLV-III-like retrovirus from macaques. Science **228**:1201-1204.

64. **Datta, P., J. E. Embree, J. K. Kreiss, J. O. Ndinya-Achola, M. Braddick, M. Temmerman, N. J. Nagelkerke, G. Maitha, K. K. Holmes, P. Piot, and .** 1994. Mother-to-child transmission of human immunodeficiency virus type 1: report from the Nairobi Study. J. Infect. Dis. **170**:1134-1140.

65. **Dayton, A. I., J. G. Sodroski, C. A. Rosen, W. C. Goh, and W. A. Haseltine**. 1986. The trans-activator gene of the human T cell lymphotropic virus type III is required for replication. Cell **44**:941-947.

66. **de, V., I.** 1994. A longitudinal study of human immunodeficiency virus transmission by heterosexual partners. European Study Group on Heterosexual Transmission of HIV. N. Engl. J. Med. **331**:341-346.

67. **Delelis, O., K. Carayon, A. Saib, E. Deprez, and J. F. Mouscadet**. 2008. Integrase and integration: biochemical activities of HIV-1 integrase. Retrovirology. **5**:114.

68. **Descamps, D., G. Collin, F. Letourneur, C. Apetrei, F. Damond, I. Loussert-Ajaka, F. Simon, S. Saragosti, and F. Brun-Vezinet**. 1997. Susceptibility of human immunodeficiency virus type 1 group O isolates to antiretroviral agents: in vitro phenotypic and genotypic analyses. J. Virol. **71**:8893-8898.

69. **Descamps, D., G. Collin, I. Loussert-Ajaka, S. Saragosti, F. Simon, and F. Brun-Vezinet**. 1995. HIV-1 group O sensitivity to antiretroviral drugs. AIDS **9**:977-978.

70. **Destefano, J. J., R. G. Buiser, L. M. Mallaber, R. A. Bambara, and P. J. Fay**. 1991. Human immunodeficiency virus reverse transcriptase displays a partially processive 3' to 5' endonuclease activity. J. Biol. Chem. **266**:24295-24301.

71. **Doehle, B. P., A. Schafer, and B. R. Cullen**. 2005. Human APOBEC3B is a potent inhibitor of HIV-1 infectivity and is resistant to HIV-1 Vif. Virology **339**:281-288.

72. **Doms, R. W. and J. P. Moore**. 2000. HIV-1 membrane fusion: targets of opportunity. J. Cell Biol. **151**:F9-14.

73. **Douaisi, M., S. Dussart, M. Courcoul, G. Bessou, R. Vigne, and E. Decroly**. 2004. HIV-1 and MLV Gag proteins are sufficient to recruit APOBEC3G into virus-like particles. Biochem. Biophys. Res. Commun. **321**:566-573.

74. **Dowling, W. E., B. Kim, C. J. Mason, K. M. Wasunna, U. Alam, L. Elson, D. L. Birx, M. L. Robb, F. E. McCutchan, and J. K. Carr**. 2002. Forty-one near full-length HIV-1 sequences from Kenya reveal an epidemic of subtype A and A-containing recombinants. AIDS **16**:1809-1820.

75. **Drake, J. W., B. Charlesworth, D. Charlesworth, and J. F. Crow**. 1998. Rates of spontaneous mutation. Genetics **148**:1667-1686.

76. **Drake, J. W. and J. J. Holland**. 1999. Mutation rates among RNA viruses. Proc. Natl. Acad. Sci. U. S. A **96**:13910-13913.

77. **Embree, J., J. Bwayo, N. Nagelkerke, S. Njenga, P. Nyange, J. Ndinya-Achola, H. Pamba, and F. Plummer**. 2001. Lymphocyte subsets in human immunodeficiency virus type 1-infected and uninfected children in Nairobi. Pediatr. Infect. Dis. J. **20**:397-403.

78. **Eshleman, S. H., L. A. Guay, A. Mwatha, E. Brown, P. Musoke, F. Mmiro, and J. B. Jackson**. 2005. Comparison of mother-to-child transmission rates in Ugandan women with subtype A versus D HIV-1 who received single-dose nevirapine prophylaxis: HIV Network For Prevention Trials 012. J. Acquir. Immune. Defic. Syndr. **39**:593-597.

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

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

81. **Fang, G., B. Weiser, C. Kuiken, S. M. Philpott, S. Rowland-Jones, F. Plummer, J. Kimani, B. Shi, R. Kaul, J. Bwayo, O. Anzala, and H. Burger**. 2004. Recombination following superinfection by HIV-1. AIDS **18**:153-159.

82. **Fauci, A. S.** 2008. 25 years of HIV. Nature **453**:289-290.

83. **Fields, B. N.** 1994. AIDS: time to turn to basic science. Nature **369**:95-96.

84. **Fisher, A. G., M. B. Feinberg, S. F. Josephs, M. E. Harper, L. M. Marselle, G. Reyes, M. A. Gonda, A. Aldovini, C. Debouk, R. C. Gallo, and .** 1986. The trans-activator gene of HTLV-III is essential for virus replication. Nature **320**:367-371.

85. **Fontenot, G., K. Johnston, J. C. Cohen, W. R. Gallaher, J. Robinson, and R. B. Luftig**. 1992. PCR amplification of HIV-1 proteinase sequences directly from lab isolates allows determination of five conserved domains. Virology **190**:1-10.

86. **Fowke, K. R., R. Kaul, K. L. Rosenthal, J. Oyugi, J. Kimani, W. J. Rutherford, N. J. Nagelkerke, T. B. Ball, J. J. Bwayo, J. N. Simonsen, G. M. Shearer, and F. A. Plummer**. 2000. HIV-1-specific cellular immune responses among HIV-1-resistant sex workers. Immunol. Cell Biol. **78**:586-595.

87. **Fowke, K. R., N. J. Nagelkerke, J. Kimani, J. N. Simonsen, A. O. Anzala, J. J. Bwayo, K. S. MacDonald, E. N. Ngugi, and F. A. Plummer**. 1996. Resistance to HIV-1 infection among persistently seronegative prostitutes in Nairobi, Kenya. Lancet **348**:1347-1351.

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

89. **Freed, E. O.** 2001. HIV-1 replication. Somat. Cell Mol. Genet. **26**:13-33.

90. **Fujita, M., H. Akari, A. Sakurai, A. Yoshida, T. Chiba, K. Tanaka, K. Strebel, and A. Adachi**. 2004. Expression of HIV-1 accessory protein Vif is controlled uniquely to be low and optimal by proteasome degradation. Microbes. Infect. **6**:791-798.

91. **Furfine, E. S. and J. E. Reardon**. 1991. Reverse transcriptase.RNase H from the human immunodeficiency virus. Relationship of the DNA polymerase and RNA hydrolysis activities. J. Biol. Chem. **266**:406-412.

92. **Gallois-Montbrun, S., B. Kramer, C. M. Swanson, H. Byers, S. Lynham, M. Ward, and M. H. Malim**. 2007. Antiviral protein APOBEC3G localizes to ribonucleoprotein complexes found in P bodies and stress granules. J. Virol. **81**:2165-2178.

93. **Gao, F., L. Yue, D. L. Robertson, S. C. Hill, H. Hui, R. J. Biggar, A. E. Neequaye, T. M. Whelan, D. D. Ho, G. M. Shaw, and .** 1994. Genetic diversity of human immunodeficiency virus type 2: evidence for distinct sequence subtypes with differences in virus biology. J. Virol. **68**:7433-7447.

94. **Gao, F., L. Yue, A. T. White, P. G. Pappas, J. Barchue, A. P. Hanson, B. M. Greene, P. M. Sharp, G. M. Shaw, and B. H. Hahn**. 1992. Human infection by genetically diverse SIVSM-related HIV-2 in west Africa. Nature **358**:495-499.

95. **Garcia, J. V. and A. D. Miller**. 1991. Serine phosphorylation-independent downregulation of cell-surface CD4 by nef. Nature **350**:508-511.

96. **Gaschen, B., J. Taylor, K. Yusim, B. Foley, F. Gao, D. Lang, V. Novitsky, B. Haynes, B. H. Hahn, T. Bhattacharya, and B. Korber**. 2002. Diversity considerations in HIV-1 vaccine selection. Science **296**:2354-2360.

97. **Gelderblom, H. R., E. H. Hausmann, M. Ozel, G. Pauli, and M. A. Koch**. 1987. Fine structure of human immunodeficiency virus (HIV) and immunolocalization of structural proteins. Virology **156**:171-176.

98. **Gilbert, M. T., A. Rambaut, G. Wlasiuk, T. J. Spira, A. E. Pitchenik, and M. Worobey**. 2007. The emergence of HIV/AIDS in the Americas and beyond. Proc. Natl. Acad. Sci. U. S. A **104**:18566-18570.

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

100. **Goila-Gaur, R., M. A. Khan, E. Miyagi, S. Kao, and K. Strebel**. 2007. Targeting APOBEC3A to the viral nucleoprotein complex confers antiviral activity. Retrovirology. **4**:61.

101. **Gooch, B. D. and B. R. Cullen**. 2008. Functional domain organization of human APOBEC3G. Virology **379**:118-124.

102. **Gopalakrishnan, V., J. A. Peliska, and S. J. Benkovic**. 1992. Human immunodeficiency virus type 1 reverse transcriptase: spatial and temporal relationship between the polymerase and RNase H activities. Proc. Natl. Acad. Sci. U. S. A **89**:10763-10767.

103. **Gordon, S. N., R. M. Dunham, J. C. Engram, J. Estes, Z. Wang, N. R. Klatt, M. Paiardini, I. V. Pandrea, C. Apetrei, D. L. Sodora, H. Y. Lee, A. T. Haase, M. D. Miller, A. Kaur, S. I. Staprans, A. S. Perelson, M. B. Feinberg, and G. Silvestri**. 2008. Short-lived infected cells support virus replication in sooty mangabeys naturally infected with simian immunodeficiency virus: implications for AIDS pathogenesis. J. Virol. **82**:3725-3735.

104. **Gottlieb, M. S., R. Schroff, H. M. Schanker, J. D. Weisman, P. T. Fan, R. A. Wolf, and A. Saxon**. 1981. Pneumocystis carinii pneumonia and mucosal candidiasis in previously healthy homosexual men: evidence of a new acquired cellular immunodeficiency. N. Engl. J. Med. **305**:1425-1431.

105. **Goulder, P. J. and D. I. Watkins**. 2004. HIV and SIV CTL escape: implications for vaccine design. Nat. Rev. Immunol. **4**:630-640.

106. **Grandgenett, D. P.** 2005. Symmetrical recognition of cellular DNA target sequences during retroviral integration. Proc. Natl. Acad. Sci. U. S. A **102**:5903-5904.

107. **Guo, F., S. Cen, M. Niu, J. Saadatmand, and L. Kleiman**. 2006. Inhibition of formula-primed reverse transcription by human APOBEC3G during human immunodeficiency virus type 1 replication. J. Virol. **80**:11710-11722.

108. **Hahn, B. H., G. M. Shaw, K. M. De Cock, and P. M. Sharp**. 2000. AIDS as a zoonosis: scientific and public health implications. Science **287**:607-614.

109. **Harari, A., M. Ooms, L. C. Mulder, and V. Simon**. 2009. Polymorphisms and splice variants influence the antiretroviral activity of human APOBEC3H. J. Virol. **83**:295-303.

110. **Hardy, A. M.** 1991. Characterization of long-term survivors of acquired immunodeficiency syndrome. The Long-term Survivor Collaborative Study Group. J. Acquir. Immune. Defic. Syndr. **4**:386-391.

111. **Harris, R. S., K. N. Bishop, A. M. Sheehy, H. M. Craig, S. K. Petersen-Mahrt, I. N. Watt, M. S. Neuberger, and M. H. Malim**. 2003. DNA deamination mediates innate immunity to retroviral infection. Cell **113**:803-809.

112. **Harris, R. S. and M. T. Liddament**. 2004. Retroviral restriction by APOBEC proteins. Nat. Rev. Immunol. **4**:868-877.

113. **Harris, R. S., S. K. Petersen-Mahrt, and M. S. Neuberger**. 2002. RNA editing enzyme APOBEC1 and some of its homologs can act as DNA mutators. Mol. Cell **10**:1247-1253.

114. **Harrison, S. C.** 2005. Mechanism of membrane fusion by viral envelope proteins. Adv. Virus Res. **64**:231-261.

115. **He, J., S. Choe, R. Walker, M. P. Di, D. O. Morgan, and N. R. Landau**. 1995. Human immunodeficiency virus type 1 viral protein R (Vpr) arrests cells in the G2 phase of the cell cycle by inhibiting p34cdc2 activity. J. Virol. **69**:6705-6711.

116. **Heeney, J. L., E. Rutjens, E. J. Verschoor, H. Niphuis, P. ten Haaft, S. Rouse, H. McClure, S. Balla-Jhagjhoorsingh, W. Bogers, M. Salas, K. Cobb, L. Kestens, D. Davis, G. G. van der, V. Courgnaud, M. Peeters, and K. K. Murthy**. 2006. Transmission of simian immunodeficiency virus SIVcpz and the evolution of infection in the presence and absence of concurrent human immunodeficiency virus type 1 infection in chimpanzees. J. Virol. **80**:7208-7218.

117. **Herbinger, K. H., M. Gerhardt, S. Piyasirisilp, D. Mloka, M. A. Arroyo, O. Hoffmann, L. Maboko, D. L. Birx, D. Mmbando, F. E. McCutchan, and M. Hoelscher**. 2006. Frequency of HIV type 1 dual infection and HIV diversity: analysis of low- and high-risk populations in Mbeya Region, Tanzania. AIDS Res. Hum. Retroviruses **22**:599-606.

118. **Hockley, D. J., R. D. Wood, J. P. Jacobs, and A. J. Garrett**. 1988. Electron microscopy of human immunodeficiency virus. J. Gen. Virol. **69 (Pt 10)**:2455-2469.

119. **Holden, L. G., C. Prochnow, Y. P. Chang, R. Bransteitter, L. Chelico, U. Sen, R. C. Stevens, M. F. Goodman, and X. S. Chen**. 2008. Crystal structure of the anti-viral APOBEC3G catalytic domain and functional implications. Nature **456**:121-124.

120. **Holman, A. G. and J. M. Coffin**. 2005. Symmetrical base preferences surrounding HIV-1, avian sarcoma/leukosis virus, and murine leukemia virus integration sites. Proc. Natl. Acad. Sci. U. S. A **102**:6103-6107.

121. **Hrimech, M., X. J. Yao, F. Bachand, N. Rougeau, and E. A. Cohen**. 1999. Human immunodeficiency virus type 1 (HIV-1) Vpr functions as an immediate-early protein during HIV-1 infection. J. Virol. **73**:4101-4109.

122. **Hu, D. J., S. Subbarao, S. Vanichseni, P. A. Mock, A. Ramos, L. Nguyen, T. Chaowanachan, F. Griensven, K. Choopanya, T. D. Mastro, and J. W. Tappero**. 2005. Frequency of HIV-1 dual subtype infections, including intersubtype superinfections, among injection drug users in Bangkok, Thailand. AIDS **19**:303-308.

123. **Huthoff, H. and M. H. Malim**. 2007. Identification of amino acid residues in APOBEC3G required for regulation by human immunodeficiency virus type 1 Vif and Virion encapsidation. J. Virol. **81**:3807-3815.

124. **Iwatani, Y., H. Takeuchi, K. Strebel, and J. G. Levin**. 2006. Biochemical activities of highly purified, catalytically active human APOBEC3G: correlation with antiviral effect. J. Virol. **80**:5992-6002.

125. **Jameel, S., M. Zafrullah, M. Ahmad, G. S. Kapoor, and S. Sehgal**. 1995. A genetic analysis of HIV-1 from Punjab, India reveals the presence of multiple variants. AIDS **9**:685-690.

126. **Jarmuz, A., A. Chester, J. Bayliss, J. Gisbourne, I. Dunham, J. Scott, and N. Navaratnam**. 2002. An anthropoid-specific locus of orphan C to U RNA-editing enzymes on chromosome 22. Genomics **79**:285-296.

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

128. **Jetzt, A. E., H. Yu, G. J. Klarmann, Y. Ron, B. D. Preston, and J. P. Dougherty**. 2000. High rate of recombination throughout the human immunodeficiency virus type 1 genome. J. Virol. **74**:1234-1240.

129. **Jin, X., A. Brooks, H. Chen, R. Bennett, R. Reichman, and H. Smith**. 2005. APOBEC3G/CEM15 (hA3G) mRNA levels associate inversely with human immunodeficiency virus viremia. J. Virol. **79**:11513-11516.

130. **Jin, X., H. Wu, and H. Smith**. 2007. APOBEC3G levels predict rates of progression to AIDS. Retrovirology. **4**:20.

131. **John-Stewart, G. C., R. W. Nduati, C. M. Rousseau, D. A. Mbori-Ngacha, B. A. Richardson, S. Rainwater, D. D. Panteleeff, and J. Overbaugh**. 2005. Subtype C Is associated with increased vaginal shedding of HIV-1. J. Infect. Dis. **192**:492-496.

132. **Jost, S., M. C. Bernard, L. Kaiser, S. Yerly, B. Hirschel, A. Samri, B. Autran, L. E. Goh, and L. Perrin**. 2002. A patient with HIV-1 superinfection. N. Engl. J. Med. **347**:731-736.

133. **Jowett, J. B., V. Planelles, B. Poon, N. P. Shah, M. L. Chen, and I. S. Chen.** 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**:6304-6313.

134. **Kaleebu, P., N. French, C. Mahe, D. Yirrell, C. Watera, F. Lyagoba, J. Nakiyingi, A. Rutebemberwa, D. Morgan, J. Weber, C. Gilks, and J. Whitworth.** 2002. Effect of human immunodeficiency virus (HIV) type 1 envelope subtypes A and D on disease progression in a large cohort of HIV-1-positive persons in Uganda. J. Infect. Dis. **185**:1244-1250.

135. **Kalish, M. L., A. Baldwin, S. Raktham, C. Wasi, C. C. Luo, G. Schochetman, T. D. Mastro, N. Young, S. Vanichseni, H. Rubsamen-Waigmann, and .** 1995. The evolving molecular epidemiology of HIV-1 envelope subtypes in injecting drug users in Bangkok, Thailand: implications for HIV vaccine trials. AIDS **9**:851-857.

136. **Kao, S., M. A. Khan, E. Miyagi, R. Plishka, A. Buckler-White, and K. Strebel.** 2003. The human immunodeficiency virus type 1 Vif protein reduces intracellular expression and inhibits packaging of APOBEC3G (CEM15), a cellular inhibitor of virus infectivity. J. Virol. **77**:11398-11407.

137. **Keele, B. F., F. Van Heuverswyn, Y. Li, E. Bailes, J. Takehisa, M. L. Santiago, F. Bibollet-Ruche, Y. Chen, L. V. Wain, F. Liegeois, S. Loul, E. M. Ngole, Y. Bienvenue, E. Delaporte, J. F. Brookfield, P. M. Sharp, G. M. Shaw, M. Peeters, and B. H. Hahn.** 2006. Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. Science **313**:523-526.

138. **Keet, I. P., P. Krijnen, M. Koot, J. M. Lange, F. Miedema, J. Goudsmit, and R. A. Coutinho.** 1993. Predictors of rapid progression to AIDS in HIV-1 seroconverters. AIDS **7**:51-57.

139. **Khan, M. A., R. Goila-Gaur, S. Opi, E. Miyagi, H. Takeuchi, S. Kao, and K. Strebel.** 2007. Analysis of the contribution of cellular and viral RNA to the packaging of APOBEC3G into HIV-1 virions. Retrovirology. **4**:48.

140. **Khan, M. A., S. Kao, E. Miyagi, H. Takeuchi, R. Goila-Gaur, S. Opi, C. L. Gipson, T. G. Parslow, H. Ly, and K. Strebel.** 2005. Viral RNA is required for the association of APOBEC3G with human immunodeficiency virus type 1 nucleoprotein complexes. J. Virol. **79**:5870-5874.

141. **Kieffer, T. L., P. Kwon, R. E. Nettles, Y. Han, S. C. Ray, and R. F. Siliciano.** 2005. G-->A hypermutation in protease and reverse transcriptase regions of human immunodeficiency virus type 1 residing in resting CD4+ T cells in vivo. J. Virol. **79**:1975-1980.

142. **Kijak, G. H., L. M. Janini, S. Tovanabutra, E. Sanders-Buell, M. A. Arroyo, M. L. Robb, N. L. Michael, D. L. Birx, and F. E. McCutchan.** 2008. Variable

contexts and levels of hypermutation in HIV-1 proviral genomes recovered from primary peripheral blood mononuclear cells. Virology **376**:101-111.

143. **Kimpton, J. and M. Emerman**. 1992. Detection of replication-competent and pseudotyped human immunodeficiency virus with a sensitive cell line on the basis of activation of an integrated beta-galactosidase gene. J. Virol. **66**:2232-2239.

144. **Kohlstaedt, L. A., J. Wang, J. M. Friedman, P. A. Rice, and T. A. Steitz**. 1992. Crystal structure at 3.5 A resolution of HIV-1 reverse transcriptase complexed with an inhibitor. Science **256**:1783-1790.

145. **Konings, F. A., G. R. Haman, Y. Xue, M. M. Urbanski, K. Hertzmark, A. Nanfack, J. M. Achkar, S. T. Burda, and P. N. Nyambi**. 2006. Genetic analysis of HIV-1 strains in rural eastern Cameroon indicates the evolution of second-generation recombinants to circulating recombinant forms. J. Acquir. Immune. Defic. Syndr. **42**:331-341.

146. **Kozak, S. L., M. Marin, K. M. Rose, C. Bystrom, and D. Kabat**. 2006. The anti-HIV-1 editing enzyme APOBEC3G binds HIV-1 RNA and messenger RNAs that shuttle between polysomes and stress granules. J. Biol. Chem. **281**:29105-29119.

147. **Kreiss, J. K., D. Koech, F. A. Plummer, K. K. Holmes, M. Lightfoote, P. Piot, A. R. Ronald, J. O. Ndinya-Achola, L. J. D'Costa, P. Roberts, and .** 1986. AIDS virus infection in Nairobi prostitutes. Spread of the epidemic to East Africa. N. Engl. J. Med. **314**:414-418.

148. **Kumar, S., K. Tamura, and M. Nei**. 2004. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief. Bioinform. **5**:150-163.

149. **Kwong, P. D., R. Wyatt, J. Robinson, R. W. Sweet, J. Sodroski, and W. A. Hendrickson**. 1998. Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. Nature **393**:648-659.

150. **Land, A. M., T. B. Ball, M. Luo, R. Pilon, P. Sandstrom, J. E. Embree, C. Wachihi, J. Kimani, and F. A. Plummer**. 2008. Human immunodeficiency virus (HIV) type 1 proviral hypermutation correlates with CD4 count in HIV-infected women from Kenya. J. Virol. **82**:8172-8182.

151. **Land, A. M., T. B. Ball, M. Luo, J. Rutherford, C. Sarna, C. Wachihi, J. Kimani, and F. A. Plummer**. 2008. Full-length HIV type 1 proviral sequencing of 10 highly exposed women from Nairobi, Kenya reveals a high proportion of intersubtype recombinants. AIDS Res. Hum. Retroviruses **24**:865-872.

152. **Land, A. M., M. Luo, R. Pilon, P. Sandstrom, J. Embree, C. Wachihi, J. Kimani, F. A. Plummer, and T. B. Ball**. 2008. High prevalence of genetically

similar HIV-1 recombinants among infected sex workers in Nairobi, Kenya. AIDS Res. Hum. Retroviruses **24**:1455-1460.

153. **Lane, J. R.** 1999. Isolation and Expansion of HIV from Cells and Body Fluids by Coculture, p. 3-10. *In* N. L. Michael and J. H. Kim (ed.), Methods in Molecular Medicine, Vol. 17: HIV Protocols. Humana Press Inc., Totowa, New Jersey.

154. **Langlois, M. A., R. C. Beale, S. G. Conticello, and M. S. Neuberger**. 2005. Mutational comparison of the single-domained APOBEC3C and double-domained APOBEC3F/G anti-retroviral cytidine deaminases provides insight into their DNA target site specificities. Nucleic Acids Res. **33**:1913-1923.

155. **Lavreys, L., J. M. Baeten, H. L. Martin, Jr., J. Overbaugh, K. Mandaliya, J. Ndinya-Achola, and J. K. Kreiss**. 2004. Hormonal contraception and risk of HIV-1 acquisition: results of a 10-year prospective study. AIDS **18**:695-697.

156. **Lenth, R. V.** 2007. Statistical power calculations. J. Anim Sci. **85**:E24-E29.

157. **Letvin, N. L.** 2006. Progress and obstacles in the development of an AIDS vaccine. Nat. Rev. Immunol. **6**:930-939.

158. **Levitsky, V. G., E. V. Ignatieva, E. A. Ananko, I. I. Turnaev, T. I. Merkulova, N. A. Kolchanov, and T. C. Hodgman**. 2007. Effective transcription factor binding site prediction using a combination of optimization, a genetic algorithm and discriminant analysis to capture distant interactions. BMC. Bioinformatics. **8**:481.

159. **Levy, D. N., G. M. Aldrovandi, O. Kutsch, and G. M. Shaw**. 2004. Dynamics of HIV-1 recombination in its natural target cells. Proc. Natl. Acad. Sci. U. S. A **101**:4204-4209.

160. **Levy, J. A.** 2009. HIV pathogenesis: 25 years of progress and persistent challenges. AIDS **23**:147-160.

161. **Li, Q., L. Duan, J. D. Estes, Z. M. Ma, T. Rourke, Y. Wang, C. Reilly, J. Carlis, C. J. Miller, and A. T. Haase**. 2005. Peak SIV replication in resting memory CD4+ T cells depletes gut lamina propria CD4+ T cells. Nature **434**:1148-1152.

162. **Liao, W., S. H. Hong, B. H. Chan, F. B. Rudolph, S. C. Clark, and L. Chan**. 1999. APOBEC-2, a cardiac- and skeletal muscle-specific member of the cytidine deaminase supergene family. Biochem. Biophys. Res. Commun. **260**:398-404.

163. **Liddament, M. T., W. L. Brown, A. J. Schumacher, and R. S. Harris**. 2004. APOBEC3F properties and hypermutation preferences indicate activity against HIV-1 in vivo. Curr. Biol. **14**:1385-1391.

164. **Lihana, R. W., S. A. Khamadi, M. K. Kiptoo, J. G. Kinyua, N. Lagat, G. N. Magoma, M. M. Mwau, E. P. Makokha, V. Onyango, S. Osman, F. A. Okoth, and E. M. Songok**. 2006. HIV type 1 subtypes among STI patients in Nairobi: a genotypic study based on partial pol gene sequencing. AIDS Res. Hum. Retroviruses **22**:1172-1177.

165. **Liu, B., P. T. Sarkis, K. Luo, Y. Yu, and X. F. Yu**. 2005. Regulation of Apobec3F and human immunodeficiency virus type 1 Vif by Vif-Cul5-ElonB/C E3 ubiquitin ligase. J. Virol. **79**:9579-9587.

166. **Liu, B., X. Yu, K. Luo, Y. Yu, and X. F. Yu**. 2004. Influence of primate lentiviral Vif and proteasome inhibitors on human immunodeficiency virus type 1 virion packaging of APOBEC3G. J. Virol. **78**:2072-2081.

167. **Loemba, H., B. Brenner, M. A. Parniak, S. Ma'ayan, B. Spira, D. Moisi, M. Oliveira, M. Detorio, M. Essex, and M. A. Wainberg**. 2002. Polymorphisms of cytotoxic T-lymphocyte (CTL) and T-helper epitopes within reverse transcriptase (RT) of HIV-1 subtype C from Ethiopia and Botswana following selection of antiretroviral drug resistance. Antiviral Res. **56**:129-142.

168. **Loemba, H., B. Brenner, M. A. Parniak, S. Ma'ayan, B. Spira, D. Moisi, M. Oliveira, M. Detorio, and M. A. Wainberg**. 2002. Genetic divergence of human immunodeficiency virus type 1 Ethiopian clade C reverse transcriptase (RT) and rapid development of resistance against nonnucleoside inhibitors of RT. Antimicrob. Agents Chemother. **46**:2087-2094.

169. **Lole, K. S., R. C. Bollinger, R. S. Paranjape, D. Gadkari, S. S. Kulkarni, N. G. Novak, R. Ingersoll, H. W. Sheppard, and S. C. Ray**. 1999. Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. J. Virol. **73**:152-160.

170. **Lowenstine, L. J., N. W. Lerche, J. L. Yee, A. Uyeda, M. B. Jennings, R. J. Munn, H. M. McClure, D. C. Anderson, P. N. Fultz, and M. B. Gardner**. 1992. Evidence for a lentiviral etiology in an epizootic of immune deficiency and lymphoma in stump-tailed macaques (Macaca arctoides). J. Med. Primatol. **21**:1-14.

171. **Luo, K., B. Liu, Z. Xiao, Y. Yu, X. Yu, R. Gorelick, and X. F. Yu**. 2004. Amino-terminal region of the human immunodeficiency virus type 1 nucleocapsid is required for human APOBEC3G packaging. J. Virol. **78**:11841-11852.

172. **Luo, K., T. Wang, B. Liu, C. Tian, Z. Xiao, J. Kappes, and X. F. Yu**. 2007. Cytidine deaminases APOBEC3G and APOBEC3F interact with HIV-1 integrase and inhibit proviral DNA formation. J. Virol.

173. **Luo, K., Z. Xiao, E. Ehrlich, Y. Yu, B. Liu, S. Zheng, and X. F. Yu**. 2005. Primate lentiviral virion infectivity factors are substrate receptors that assemble

with cullin 5-E3 ligase through a HCCH motif to suppress APOBEC3G. Proc. Natl. Acad. Sci. U. S. A **102**:11444-11449.

174. **MacDonald, K. S., J. E. Embree, N. J. Nagelkerke, J. Castillo, S. Ramhadin, S. Njenga, J. Oyug, J. Ndinya-Achola, B. H. Barber, J. J. Bwayo, and F. A. Plummer**. 2001. The HLA A2/6802 supertype is associated with reduced risk of perinatal human immunodeficiency virus type 1 transmission. J. Infect. Dis. **183**:503-506.

175. **Madani, N. and D. Kabat**. 1998. An endogenous inhibitor of human immunodeficiency virus in human lymphocytes is overcome by the viral Vif protein. J. Virol. **72**:10251-10255.

176. **Mak, J. and L. Kleiman**. 1997. Primer tRNAs for reverse transcription. J. Virol. **71**:8087-8095.

177. **Malim, M. H.** 2006. Natural resistance to HIV infection: The Vif-APOBEC interaction. C. R. Biol. **329**:871-875.

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

179. **Mangeat, B., P. Turelli, S. Liao, and D. Trono**. 2004. A single amino acid determinant governs the species-specific sensitivity of APOBEC3G to Vif action. J. Biol. Chem. **279**:14481-14483.

180. **Mansfield, K. G., N. W. Lerch, M. B. Gardner, and A. A. Lackner**. 1995. Origins of simian immunodeficiency virus infection in macaques at the New England Regional Primate Research Center. J. Med. Primatol. **24**:116-122.

181. **Mariani, R., D. Chen, B. Schrofelbauer, F. Navarro, R. Konig, B. Bollman, C. Munk, H. Nymark-McMahon, and N. R. Landau**. 2003. Species-specific exclusion of APOBEC3G from HIV-1 virions by Vif. Cell **114**:21-31.

182. **Marin, M., K. M. Rose, S. L. Kozak, and D. Kabat**. 2003. HIV-1 Vif protein binds the editing enzyme APOBEC3G and induces its degradation. Nat. Med. **9**:1398-1403.

183. **Marquet, R., C. Isel, C. Ehresmann, and B. Ehresmann**. 1995. tRNAs as primer of reverse transcriptases. Biochimie **77**:113-124.

184. **Martin, H. L., Jr., P. M. Nyange, B. A. Richardson, L. Lavreys, K. Mandaliya, D. J. Jackson, J. O. Ndinya-Achola, and J. Kreiss**. 1998. Hormonal contraception, sexually transmitted diseases, and risk of heterosexual transmission of human immunodeficiency virus type 1. J. Infect. Dis. **178**:1053-1059.

185. **Martins, L. O., E. Leal, and H. Kishino**. 2008. Phylogenetic detection of recombination with a Bayesian prior on the distance between trees. PLoS. ONE. **3**:e2651.

186. **Matsumoto, T., H. Marusawa, Y. Endo, Y. Ueda, Y. Matsumoto, and T. Chiba**. 2006. Expression of APOBEC2 is transcriptionally regulated by NF-kappaB in human hepatocytes. FEBS Lett. **580**:731-735.

187. **Mbisa, J. L., R. Barr, J. A. Thomas, N. Vandegraaff, I. J. Dorweiler, E. S. Svarovskaia, W. L. Brown, L. M. Mansky, R. J. Gorelick, R. S. Harris, A. Engelman, and V. K. Pathak**. 2007. HIV-1 cDNAs Produced in the Presence of APOBEC3G Exhibit Defects in Plus-Strand DNA Transfer and Integration. J. Virol.

188. **McCune, J. M.** 2001. The dynamics of CD4+ T-cell depletion in HIV disease. Nature **410**:974-979.

189. **McCutchan, F. E.** 2006. Global epidemiology of HIV. J. Med. Virol. **78 Suppl 1**:S7-S12.

190. **McMichael, A. J. and S. L. Rowland-Jones**. 2001. Cellular immune responses to HIV. Nature **410**:980-987.

191. **Meerloo, T., H. K. Parmentier, A. D. Osterhaus, J. Goudsmit, and H. J. Schuurman**. 1992. Modulation of cell surface molecules during HIV-1 infection of H9 cells. An immunoelectron microscopic study. AIDS **6**:1105-1116.

192. **Mehle, A., J. Goncalves, M. Santa-Marta, M. McPike, and D. Gabuzda**. 2004. Phosphorylation of a novel SOCS-box regulates assembly of the HIV-1 Vif-Cul5 complex that promotes APOBEC3G degradation. Genes Dev. **18**:2861-2866.

193. **Mehle, A., E. R. Thomas, K. S. Rajendran, and D. Gabuzda**. 2006. A zinc-binding region in Vif binds Cul5 and determines cullin selection. J. Biol. Chem. **281**:17259-17265.

194. **Mehle, A., H. Wilson, C. Zhang, A. J. Brazier, M. McPike, E. Pery, and D. Gabuzda**. 2007. Identification of an APOBEC3G binding site in human immunodeficiency virus type 1 Vif and inhibitors of Vif-APOBEC3G binding. J. Virol. **81**:13235-13241.

195. **Migueles, S. A., M. S. Sabbaghian, W. L. Shupert, M. P. Bettinotti, F. M. Marincola, L. Martino, C. W. Hallahan, S. M. Selig, D. Schwartz, J. Sullivan, and M. Connors**. 2000. HLA B*5701 is highly associated with restriction of virus replication in a subgroup of HIV-infected long term nonprogressors. Proc. Natl. Acad. Sci. U. S. A **97**:2709-2714.

196. **Miller, C. J., Q. Li, K. Abel, E. Y. Kim, Z. M. Ma, S. Wietgrefe, L. La Franco-Scheuch, L. Compton, L. Duan, M. D. Shore, M. Zupancic, M.**

Busch, J. Carlis, S. Wolinsky, and A. T. Haase. 2005. Propagation and dissemination of infection after vaginal transmission of simian immunodeficiency virus. J. Virol. **79**:9217-9227.

197. **Miller, J. H., V. Presnyak, and H. C. Smith**. 2007. The dimerization domain of HIV-1 viral infectivity factor Vif is required to block virion incorporation of APOBEC3G. Retrovirology. **4**:81.

198. **Miller, M. D., M. T. Warmerdam, I. Gaston, W. C. Greene, and M. B. Feinberg**. 1994. The human immunodeficiency virus-1 nef gene product: a positive factor for viral infection and replication in primary lymphocytes and macrophages. J. Exp. Med. **179**:101-113.

199. **Minin, V. N., K. S. Dorman, F. Fang, and M. A. Suchard**. 2007. Phylogenetic mapping of recombination hotspots in human immunodeficiency virus via spatially smoothed change-point processes. Genetics **175**:1773-1785.

200. **Miyagi, E., S. Opi, H. Takeuchi, M. Khan, R. Goila-Gaur, S. Kao, and K. Strebel**. 2007. Enzymatically Active APOBEC3G Is Required for Efficient Inhibition of Human Immunodeficiency Virus Type 1. J. Virol. **81**:13346-13353.

201. **Mizrahi, V.** 1989. Analysis of the ribonuclease H activity of HIV-1 reverse transcriptase using RNA.DNA hybrid substrates derived from the gag region of HIV-1. Biochemistry **28**:9088-9094.

202. **Montano, M. A., V. A. Novitsky, J. T. Blackard, N. L. Cho, D. A. Katzenstein, and M. Essex**. 1997. Divergent transcriptional regulation among expanding human immunodeficiency virus type 1 subtypes. J. Virol. **71**:8657-8665.

203. **Moore, J. P. and R. W. Doms**. 2003. The entry of entry inhibitors: a fusion of science and medicine. Proc. Natl. Acad. Sci. U. S. A **100**:10598-10602.

204. **Motulsky, H. J.** 2003. Prism 4 Statistics Guide - Statistical analysis for laboratory and clinical researchers. GraphPad Software Inc., San Diego, CA.

205. **Muckenfuss, H., M. Hamdorf, U. Held, M. Perkovic, J. Lower, K. Cichutek, E. Flory, G. G. Schumann, and C. Munk**. 2006. APOBEC3 proteins inhibit human LINE-1 retrotransposition. J. Biol. Chem. **281**:22161-22172.

206. **Muckenfuss, H., J. K. Kaiser, E. Krebil, M. Battenberg, C. Schwer, K. Cichutek, C. Munk, and E. Flory**. 2007. Sp1 and Sp3 regulate basal transcription of the human APOBEC3G gene. Nucleic Acids Res. **35**:3784-3796.

207. **Muramatsu, M., K. Kinoshita, S. Fagarasan, S. Yamada, Y. Shinkai, and T. Honjo**. 2000. Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. Cell **102**:553-563.

208. **Murphey-Corb, M., L. N. Martin, S. R. Rangan, G. B. Baskin, B. J. Gormus, R. H. Wolf, W. A. Andes, M. West, and R. C. Montelaro**. 1986. Isolation of an HTLV-III-related retrovirus from macaques with simian AIDS and its possible origin in asymptomatic mangabeys. Nature **321**:435-437.

209. **Murray, M. C., J. E. Embree, S. G. Ramdahin, A. O. Anzala, S. Njenga, and F. A. Plummer**. 2000. Effect of human immunodeficiency virus (HIV) type 1 viral genotype on mother-to-child transmission of HIV-1. J. Infect. Dis. **181**:746-749.

210. **Muthumani, K., D. S. Hwang, B. M. Desai, D. Zhang, N. Dayes, D. R. Green, and D. B. Weiner**. 2002. HIV-1 Vpr induces apoptosis through caspase 9 in T cells and peripheral blood mononuclear cells. J. Biol. Chem. **277**:37820-37831.

211. **Muto, T., M. Muramatsu, M. Taniwaki, K. Kinoshita, and T. Honjo**. 2000. Isolation, tissue distribution, and chromosomal localization of the human activation-induced cytidine deaminase (AID) gene. Genomics **68**:85-88.

212. **Nagel, J. E., F. J. Chrest, R. S. Pyle, and W. H. Adler**. 1983. Monoclonal antibody analysis of T-lymphocyte subsets in young and aged adults. Immunol. Commun. **12**:223-237.

213. **Neil, S. J., T. Zang, and P. D. Bieniasz**. 2008. Tetherin inhibits retrovirus release and is antagonized by HIV-1 Vpu. Nature **451**:425-430.

214. **Neilson, J. R., G. C. John, J. K. Carr, P. Lewis, J. K. Kreiss, S. Jackson, R. W. Nduati, D. Mbori-Ngacha, D. D. Panteleeff, S. Bodrug, C. Giachetti, M. A. Bott, B. A. Richardson, J. Bwayo, J. Ndinya-Achola, and J. Overbaugh**. 1999. Subtypes of human immunodeficiency virus type 1 and disease stage among women in Nairobi, Kenya. J. Virol. **73**:4393-4403.

215. **Nguyen, D. H., S. Gummuluru, and J. Hu**. 2007. Deamination-independent inhibition of hepatitis B virus reverse transcription by APOBEC3G. J. Virol. **81**:4465-4472.

216. **Nicolosi, A., M. L. Correa Leite, M. Musicco, C. Arici, G. Gavazzeni, and A. Lazzarin**. 1994. The efficiency of male-to-female and female-to-male sexual transmission of the human immunodeficiency virus: a study of 730 stable couples. Italian Study Group on HIV Heterosexual Transmission. Epidemiology **5**:570-575.

217. **Nielsen, C., C. Pedersen, J. D. Lundgren, and J. Gerstoft**. 1993. Biological properties of HIV isolates in primary HIV infection: consequences for the subsequent course of infection. AIDS **7**:1035-1040.

218. **Nowak, M.** 1990. HIV mutation rate. Nature **347**:522.

219. **Oberste, M. S. and M. A. Gonda**. 1992. Conservation of amino-acid sequence motifs in lentivirus Vif proteins. Virus Genes **6**:95-102.

220. **Oelrichs, R. B., V. A. Lawson, K. M. Coates, C. Chatfield, N. J. Deacon, and D. A. McPhee**. 2000. Rapid full-length genomic sequencing of two cytopathically heterogeneous Australian primary HIV-1 isolates. J. Biomed. Sci. **7**:128-135.

221. **OhAinle, M., J. A. Kerns, H. S. Malik, and M. Emerman**. 2006. Adaptive evolution and antiviral activity of the conserved mammalian cytidine deaminase APOBEC3H. J. Virol. **80**:3853-3862.

222. **Onafuwa, A., W. An, N. D. Robson, and A. Telesnitsky**. 2003. Human immunodeficiency virus type 1 genetic recombination is more frequent than that of Moloney murine leukemia virus despite similar template switching rates. J. Virol. **77**:4577-4587.

223. **Pace, C., J. Keller, D. Nolan, I. James, S. Gaudieri, C. Moore, and S. Mallal**. 2006. Population level analysis of human immunodeficiency virus type 1 hypermutation and its relationship with APOBEC3G and vif genetic variation. J. Virol. **80**:9259-9269.

224. **Peeters, M., R. Vincent, J. L. Perret, M. Lasky, D. Patrel, F. Liegeois, V. Courgnaud, R. Seng, T. Matton, S. Molinier, and E. Delaporte**. 1999. Evidence for differences in MT2 cell tropism according to genetic subtypes of HIV-1: syncytium-inducing variants seem rare among subtype C HIV-1 viruses. J. Acquir. Immune. Defic. Syndr. Hum. Retrovirol. **20**:115-121.

225. **Peng, G., K. J. Lei, W. Jin, T. Greenwell-Wild, and S. M. Wahl**. 2006. Induction of APOBEC3 family proteins, a defensive maneuver underlying interferon-induced anti-HIV-1 activity. J. Exp. Med. **203**:41-46.

226. **Perelson, A. S., A. U. Neumann, M. Markowitz, J. M. Leonard, and D. D. Ho**. 1996. HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. Science **271**:1582-1586.

227. **Peters, H. O., M. G. Mendoza, R. E. Capina, M. Luo, X. Mao, M. Gubbins, N. J. Nagelkerke, I. Macarthur, B. B. Sheardown, J. Kimani, C. Wachihi, S. Thavaneswaran, and F. A. Plummer**. 2008. An integrative bioinformatic approach for studying escape mutations in human immunodeficiency virus type 1 gag in the Pumwani Sex Worker Cohort. J. Virol. **82**:1980-1992.

228. **Piantadosi, A., B. Chohan, V. Chohan, R. S. McClelland, and J. Overbaugh**. 2007. Chronic HIV-1 infection frequently fails to protect against superinfection. PLoS. Pathog. **3**:e177.

229. **Piatak, M., Jr., M. S. Saag, L. C. Yang, S. J. Clark, J. C. Kappes, K. C. Luk, B. H. Hahn, G. M. Shaw, and J. D. Lifson**. 1993. High levels of HIV-1 in

plasma during all stages of infection determined by competitive PCR. Science **259**:1749-1754.

230. **Pido-Lopez, J., T. Whittall, Y. Wang, L. A. Bergmeier, K. Babaahmady, M. Singh, and T. Lehner**. 2007. Stimulation of cell surface CCR5 and CD40 molecules by their ligands or by HSP70 up-regulates APOBEC3G expression in CD4(+) T cells and dendritic cells. J. Immunol. **178**:1671-1679.

231. **Pilcher, C. D., J. J. Eron, Jr., S. Galvin, C. Gay, and M. S. Cohen**. 2004. Acute HIV revisited: new opportunities for treatment and prevention. J. Clin. Invest **113**:937-945.

232. **Pilcher, C. D., H. C. Tien, J. J. Eron, Jr., P. L. Vernazza, S. Y. Leu, P. W. Stewart, L. E. Goh, and M. S. Cohen**. 2004. Brief but efficient: acute HIV infection and the sexual transmission of HIV. J. Infect. Dis. **189**:1785-1792.

233. **Pillai, S. K., J. K. Wong, and J. D. Barbour**. 2008. Turning up the volume on mutational pressure: is more of a good thing always better? (A case study of HIV-1 Vif and APOBEC3). Retrovirology. **5**:26.

234. **Plummer, F. A., T. B. Ball, J. Kimani, and K. R. Fowke**. 1999. Resistance to HIV-1 infection among highly exposed sex workers in Nairobi: what mediates protection and why does it develop? Immunol. Lett. **66**:27-34.

235. **Polk, B. F.** 1985. Female-to-male transmission of AIDS. JAMA **254**:3177-3178.

236. **Popovic, M., M. G. Sarngadharan, E. Read, and R. C. Gallo**. 1984. Detection, isolation, and continuous production of cytopathic retroviruses (HTLV-III) from patients with AIDS and pre-AIDS. Science **224**:497-500.

237. **Poss, M., J. Gosink, E. Thomas, J. K. Kreiss, J. Ndinya-Achola, K. Mandaliya, J. Bwayo, and J. Overbaugh**. 1997. Phylogenetic evaluation of Kenyan HIV type 1 isolates. AIDS Res. Hum. Retroviruses **13**:493-499.

238. **Purcell, D. F. and M. A. Martin**. 1993. Alternative splicing of human immunodeficiency virus type 1 mRNA modulates viral protein expression, replication, and infectivity. J. Virol. **67**:6365-6378.

239. **Re, F., D. Braaten, E. K. Franke, and J. Luban**. 1995. Human immunodeficiency virus type 1 Vpr arrests the cell cycle in G2 by inhibiting the activation of p34cdc2-cyclin B. J. Virol. **69**:6859-6864.

240. **Renjifo, B., P. Gilbert, B. Chaplin, G. Msamanga, D. Mwakagile, W. Fawzi, and M. Essex**. 2004. Preferential in-utero transmission of HIV-1 subtype C as compared to HIV-1 subtype A or D. AIDS **18**:1629-1636.

241. **Renkema, G. H., A. Manninen, D. A. Mann, M. Harris, and K. Saksela**. 1999. Identification of the Nef-associated kinase as p21-activated kinase 2. Curr. Biol. **9**:1407-1410.

242. **Rey-Cuille, M. A., J. L. Berthier, M. C. Bomsel-Demontoy, Y. Chaduc, L. Montagnier, A. G. Hovanessian, and L. A. Chakrabarti**. 1998. Simian immunodeficiency virus replicates to high levels in sooty mangabeys without inducing disease. J. Virol. **72**:3872-3886.

243. **Robertson, D. L., J. P. Anderson, J. A. Bradac, J. K. Carr, B. Foley, R. K. Funkhouser, F. Gao, B. H. Hahn, M. L. Kalish, C. Kuiken, G. H. Learn, T. Leitner, F. McCutchan, S. Osmanov, M. Peeters, D. Pieniazek, M. Salminen, P. M. Sharp, S. Wolinsky, and B. Korber**. 2000. HIV-1 nomenclature proposal. Science **288**:55-56.

244. **Rogel, M. E., L. I. Wu, and M. Emerman**. 1995. The human immunodeficiency virus type 1 vpr gene prevents cell proliferation during chronic infection. J. Virol. **69**:882-888.

245. **Rogozin, I. B., M. K. Basu, I. K. Jordan, Y. I. Pavlov, and E. V. Koonin**. 2005. APOBEC4, a new member of the AID/APOBEC family of polynucleotide (deoxy)cytidine deaminases predicted by computational analysis. Cell Cycle **4**:1281-1285.

246. **Rose, K. M., M. Marin, S. L. Kozak, and D. Kabat**. 2004. Transcriptional regulation of APOBEC3G, a cytidine deaminase that hypermutates human immunodeficiency virus. J. Biol. Chem. **279**:41744-41749.

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

248. **Rosler, C., J. Kock, M. H. Malim, H. E. Blum, and W. F. von**. 2004. Comment on "Inhibition of hepatitis B virus replication by APOBEC3G". Science **305**:1403.

249. **Russell, R. A., M. D. Moore, W. S. Hu, and V. K. Pathak**. 2009. APOBEC3G induces a hypermutation gradient: purifying selection at multiple steps during HIV-1 replication results in levels of G-to-A mutations that are high in DNA, intermediate in cellular viral RNA, and low in virion RNA. Retrovirology. **6**:16.

250. **Russell, R. A. and V. K. Pathak**. 2007. Identification of Two Distinct HIV-1 Vif Determinants Critical for Interactions with Human APOBEC3G and APOBEC3F. J. Virol.

251. **Sagar, M., L. Lavreys, J. M. Baeten, B. A. Richardson, K. Mandaliya, J. O. Ndinya-Achola, J. K. Kreiss, and J. Overbaugh**. 2004. Identification of modifiable factors that affect the genetic diversity of the transmitted HIV-1 population. AIDS **18**:615-619.

252. **Santa-Marta, M., F. A. da Silva, A. M. Fonseca, and J. Goncalves**. 2005. HIV-1 Vif can directly inhibit apolipoprotein B mRNA-editing enzyme catalytic polypeptide-like 3G-mediated cytidine deamination by using a single amino acid interaction and without protein degradation. J. Biol. Chem. **280**:8765-8775.

253. **Santiago, M. L., F. Range, B. F. Keele, Y. Li, E. Bailes, F. Bibollet-Ruche, C. Fruteau, R. Noe, M. Peeters, J. F. Brookfield, G. M. Shaw, P. M. Sharp, and B. H. Hahn**. 2005. Simian immunodeficiency virus infection in free-ranging sooty mangabeys (Cercocebus atys atys) from the Tai Forest, Cote d'Ivoire: implications for the origin of epidemic human immunodeficiency virus type 2. J. Virol. **79**:12515-12527.

254. **Sarkis, P. T., S. Ying, R. Xu, and X. F. Yu**. 2006. STAT1-independent cell type-specific regulation of antiviral APOBEC3G by IFN-alpha. J. Immunol. **177**:4530-4540.

255. **Sawyer, S. L., M. Emerman, and H. S. Malik**. 2004. Ancient adaptive evolution of the primate antiviral DNA-editing enzyme APOBEC3G. PLoS. Biol. **2**:E275.

256. **Schafer, A., H. P. Bogerd, and B. R. Cullen**. 2004. Specific packaging of APOBEC3G into HIV-1 virions is mediated by the nucleocapsid domain of the gag polyprotein precursor. Virology **328**:163-168.

257. **Schrofelbauer, B., D. Chen, and N. R. Landau**. 2004. A single amino acid of APOBEC3G controls its species-specific interaction with virion infectivity factor (Vif). Proc. Natl. Acad. Sci. U. S. A **101**:3927-3932.

258. **Schrofelbauer, B., T. Senger, G. Manning, and N. R. Landau**. 2006. Mutational alteration of human immunodeficiency virus type 1 Vif allows for functional interaction with nonhuman primate APOBEC3G. J. Virol. **80**:5984-5991.

259. **Schwartz, O., V. Marechal, G. S. Le, F. Lemonnier, and J. M. Heard**. 1996. Endocytosis of major histocompatibility complex class I molecules is induced by the HIV-1 Nef protein. Nat. Med. **2**:338-342.

260. **Sharp, P. M. and B. H. Hahn**. 2008. AIDS: prehistory of HIV-1. Nature **455**:605-606.

261. **Sharp, P. M., G. M. Shaw, and B. H. Hahn**. 2005. Simian immunodeficiency virus infection of chimpanzees. J. Virol. **79**:3891-3902.

262. **Sheehy, A. M., N. C. Gaddis, J. D. Choi, and M. H. Malim**. 2002. Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. Nature **418**:646-650.

263. **Sheehy, A. M., N. C. Gaddis, and M. H. Malim**. 2003. The antiretroviral enzyme APOBEC3G is degraded by the proteasome in response to HIV-1 Vif. Nat. Med. **9**:1404-1407.

264. **Siepel, A. C. and B. T. Korber**. 1995. Scanning the Database for Recombinant HIV-1 Genomes, *In* Human Retroviruses and AIDS 1995. Los Alamos National Laboratory, Theroretical Biology and Biophysics, Los Alamos, New Mexico.

265. **Silvestri, G., D. L. Sodora, R. A. Koup, M. Paiardini, S. P. O'Neil, H. M. McClure, S. I. Staprans, and M. B. Feinberg**. 2003. Nonpathogenic SIV infection of sooty mangabeys is characterized by limited bystander immunopathology despite chronic high-level viremia. Immunity. **18**:441-452.

266. **Simmons, A., V. Aluvihare, and A. McMichael**. 2001. Nef triggers a transcriptional program in T cells imitating single-signal T cell activation and inducing HIV virulence mediators. Immunity. **14**:763-777.

267. **Simon, J. H., N. C. Gaddis, R. A. Fouchier, and M. H. Malim**. 1998. Evidence for a newly discovered cellular anti-HIV-1 phenotype. Nat. Med. **4**:1397-1400.

268. **Simon, V., V. Zennou, D. Murray, Y. Huang, D. D. Ho, and P. D. Bieniasz**. 2005. Natural variation in Vif: differential impact on APOBEC3G/3F and a potential role in HIV-1 diversification. PLoS. Pathog. **1**:e6.

269. **Songok, E. M., R. W. Lihana, M. K. Kiptoo, I. O. Genga, R. Kibaya, F. Odhiambo, K. Kobayashi, Y. Ago, N. Ndembi, F. Okoth, Y. Fujiyama, J. Imanishi, and H. Ichimura**. 2003. Identification of env CRF-10 among HIV variants circulating in rural western Kenya. AIDS Res. Hum. Retroviruses **19**:161-165.

270. **Stekler, J., B. J. Sycks, S. Holte, J. Maenza, C. E. Stevens, J. Dragavon, A. C. Collier, and R. W. Coombs**. 2008. HIV dynamics in seminal plasma during primary HIV infection. AIDS Res. Hum. Retroviruses **24**:1269-1274.

271. **Stenglein, M. D. and R. S. Harris**. 2006. APOBEC3B and APOBEC3F inhibit L1 retrotransposition by a DNA deamination-independent mechanism. J. Biol. Chem. **281**:16837-16841.

272. **Stopak, K., C. de Noronha, W. Yonemoto, and W. C. Greene**. 2003. HIV-1 Vif blocks the antiviral activity of APOBEC3G by impairing both its translation and intracellular stability. Mol. Cell **12**:591-601.

273. **Stopak, K. S., Y. L. Chiu, J. Kropp, R. M. Grant, and W. C. Greene**. 2007. Distinct patterns of cytokine regulation of APOBEC3G expression and activity in primary lymphocytes, macrophages, and dendritic cells. J. Biol. Chem. **282**:3539-3546.

274. **Stowell, R. E., E. K. Smith, C. Espana, and V. G. Nelson**. 1971. Outbreak of malignant lymphoma in rhesus monkeys. Lab Invest **25**:476-479.

275. **Strebel, K. and M. A. Khan**. 2008. APOBEC3G encapsidation into HIV-1 virions: which RNA is it? Retrovirology. **5**:55.

276. **Suspene, R., D. Guetard, M. Henry, P. Sommer, S. Wain-Hobson, and J. P. Vartanian**. 2005. Extensive editing of both hepatitis B virus DNA strands by APOBEC3 cytidine deaminases in vitro and in vivo. Proc. Natl. Acad. Sci. U. S. A **102**:8321-8326.

277. **Suspene, R., C. Rusniok, J. P. Vartanian, and S. Wain-Hobson**. 2006. Twin gradients in APOBEC3 edited HIV-1 DNA reflect the dynamics of lentiviral replication. Nucleic Acids Res. **34**:4677-4684.

278. **Svarovskaia, E. S., H. Xu, J. L. Mbisa, R. Barr, R. J. Gorelick, A. Ono, E. O. Freed, W. S. Hu, and V. K. Pathak**. 2004. Human apolipoprotein B mRNA-editing enzyme-catalytic polypeptide-like 3G (APOBEC3G) is incorporated into HIV-1 virions through interactions with viral and nonviral RNAs. J. Biol. Chem. **279**:35822-35828.

279. **Tanaka, Y., H. Marusawa, H. Seno, Y. Matsumoto, Y. Ueda, Y. Kodama, Y. Endo, J. Yamauchi, T. Matsumoto, A. Takaori-Kondo, I. Ikai, and T. Chiba**. 2006. Anti-viral protein APOBEC3G is induced by interferon-alpha stimulation in human hepatocytes. Biochem. Biophys. Res. Commun. **341**:314-319.

280. **Tantillo, C., J. Ding, A. Jacobo-Molina, R. G. Nanni, P. L. Boyer, S. H. Hughes, R. Pauwels, K. Andries, P. A. Janssen, and E. Arnold**. 1994. Locations of anti-AIDS drug binding sites and resistance mutations in the three-dimensional structure of HIV-1 reverse transcriptase. Implications for mechanisms of drug inhibition and resistance. J. Mol. Biol. **243**:369-387.

281. **Telesnitsky, A. and S. P. Goff**. 1993. Two defective forms of reverse transcriptase can complement to restore retroviral infectivity. EMBO J. **12**:4433-4438.

282. **Teng, B., C. F. Burant, and N. O. Davidson**. 1993. Molecular cloning of an apolipoprotein B messenger RNA editing protein. Science **260**:1816-1819.

283. **Thomson, M. M. and R. Najera**. 2005. Molecular epidemiology of HIV-1 variants in the global AIDS pandemic: an update. AIDS Rev. **7**:210-224.

284. **Tian, C., X. Yu, W. Zhang, T. Wang, R. Xu, and X. F. Yu**. 2006. Differential requirement for conserved tryptophans in human immunodeficiency virus type 1 Vif for the selective suppression of APOBEC3G and APOBEC3F. J. Virol. **80**:3112-3115.

285. **Tscherning, C., A. Alaeus, R. Fredriksson, A. Bjorndal, H. Deng, D. R. Littman, E. M. Fenyo, and J. Albert**. 1998. Differences in chemokine coreceptor usage between genetic subtypes of HIV-1. Virology **241**:181-188.

286. **Tsui, R., B. L. Herring, J. D. Barbour, R. M. Grant, P. Bacchetti, A. Kral, B. R. Edlin, and E. L. Delwart**. 2004. Human immunodeficiency virus type 1 superinfection was not detected following 215 years of injection drug user exposure. J. Virol. **78**:94-103.

287. **Turelli, P., B. Mangeat, S. Jost, S. Vianin, and D. Trono**. 2004. Inhibition of hepatitis B virus replication by APOBEC3G. Science **303**:1829.

288. **Turner, B. J., F. M. Hecht, and R. B. Ismail**. 1994. CD4+ T-lymphocyte measures in the treatment of individuals infected with human immunodeficiency virus type 1. A review for clinical practitioners. Arch. Intern. Med. **154**:1561-1573.

289. **Ulenga, N. K., A. D. Sarr, D. Hamel, J. L. Sankale, S. Mboup, and P. J. Kanki**. 2008. The level of APOBEC3G (hA3G)-related G-to-A mutations does not correlate with viral load in HIV type 1-infected individuals. AIDS Res. Hum. Retroviruses **24**:1285-1290.

290. **UNAids**. 2008. 2008 Report on the global AIDS epidemic, *In* .

291. **Valcke, H. S., N. F. Bernard, J. Bruneau, M. Alary, C. M. Tsoukas, and M. Roger**. 2006. APOBEC3G genetic variants and their association with risk of HIV infection in highly exposed Caucasians. AIDS **20**:1984-1986.

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

293. **van Wamel, J. L. and B. Berkhout**. 1998. The first strand transfer during HIV-1 reverse transcription can occur either intramolecularly or intermolecularly. Virology **244**:245-251.

294. **Van, D. N., D. Goff, C. Katsura, R. L. Jorgenson, R. Mitchell, M. C. Johnson, E. B. Stephens, and J. Guatelli**. 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**:245-252.

295. **VandeWoude, S. and C. Apetrei**. 2006. Going wild: lessons from naturally occurring T-lymphotropic lentiviruses. Clin. Microbiol. Rev. **19**:728-762.

296. **Vartanian, J. P., D. Guetard, M. Henry, and S. Wain-Hobson**. 2008. Evidence for editing of human papillomavirus DNA by APOBEC3 in benign and precancerous lesions. Science **320**:230-233.

297. **Vasan, A., B. Renjifo, E. Hertzmark, B. Chaplin, G. Msamanga, M. Essex, W. Fawzi, and D. Hunter**. 2006. Different rates of disease progression of HIV type 1 infection in Tanzania based on infecting subtype. Clin. Infect. Dis. **42**:843-852.

298. **Visawapoka, U., S. Tovanabutra, J. R. Currier, J. H. Cox, C. J. Mason, M. Wasunna, M. Ponglikitmongkol, W. E. Dowling, M. L. Robb, D. L. Birx, and F. E. McCutchan**. 2006. Circulating and unique recombinant forms of HIV type 1 containing subsubtype A2. AIDS Res. Hum. Retroviruses **22**:695-702.

299. **Voeller, B. and D. J. Anderson**. 1992. Heterosexual transmission of HIV. JAMA **267**:1917-1918.

300. **Wain-Hobson, S.** 1996. G -> A Hypermutation, *In* G. Myers, B. T. Korber, B. T. Foley, K.-T. Jeang, J. W. Mellors, and S. Wain-Hobson (ed.), Human Retroviruses and AIDS 1996. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM.

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

302. **Wang, F. X., J. Huang, H. Zhang, X. Ma, and H. Zhang**. 2008. APOBEC3G upregulation by alpha interferon restricts human immunodeficiency virus type 1 infection in human peripheral plasmacytoid dendritic cells. J. Gen. Virol. **89**:722-730.

303. **Wang, X., P. T. Dolan, Y. Dang, and Y. H. Zheng**. 2007. Biochemical differentiation of APOBEC3F and APOBEC3G proteins associated with HIV-1 life cycle. J. Biol. Chem. **282**:1585-1594.

304. **Wedekind, J. E., G. S. Dance, M. P. Sowden, and H. C. Smith**. 2003. Messenger RNA editing in mammals: new members of the APOBEC family seeking roles in the family business. Trends Genet. **19**:207-216.

305. **Wei, X., J. M. Decker, S. Wang, H. Hui, J. C. Kappes, X. Wu, J. F. Salazar-Gonzalez, M. G. Salazar, J. M. Kilby, M. S. Saag, N. L. Komarova, M. A. Nowak, B. H. Hahn, P. D. Kwong, and G. M. Shaw**. 2003. Antibody neutralization and escape by HIV-1. Nature **422**:307-312.

306. **Weiss, R. A.** 2001. Gulliver's travels in HIVland. Nature **410**:963-967.

307. **Weissenhorn, W., A. Dessen, S. C. Harrison, J. J. Skehel, and D. C. Wiley**. 1997. Atomic structure of the ectodomain from HIV-1 gp41. Nature **387**:426-430.

308. **Wichroski, M. J., K. Ichiyama, and T. M. Rana**. 2005. Analysis of HIV-1 viral infectivity factor-mediated proteasome-dependent depletion of APOBEC3G: correlating function and subcellular localization. J. Biol. Chem. **280**:8387-8396.

309. **Wichroski, M. J., G. B. Robb, and T. M. Rana**. 2006. Human retroviral host restriction factors APOBEC3G and APOBEC3F localize to mRNA processing bodies. PLoS. Pathog. **2**:e41.

310. **Wiegand, H. L., B. P. Doehle, H. P. Bogerd, and B. R. Cullen**. 2004. A second human antiretroviral factor, APOBEC3F, is suppressed by the HIV-1 and HIV-2 Vif proteins. EMBO J. **23**:2451-2458.

311. **Willey, R. L., F. Maldarelli, M. A. Martin, and K. Strebel**. 1992. Human immunodeficiency virus type 1 Vpu protein induces rapid degradation of CD4. J. Virol. **66**:7193-7200.

312. **Worobey, M., M. Gemmel, D. E. Teuwen, T. Haselkorn, K. Kunstman, M. Bunce, J. J. Muyembe, J. M. Kabongo, R. M. Kalengayi, E. Van Marck, M. T. Gilbert, and S. M. Wolinsky**. 2008. Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. Nature **455**:661-664.

313. **Wu, X., Y. Li, B. Crise, S. M. Burgess, and D. J. Munroe**. 2005. Weak palindromic consensus sequences are a common feature found at the integration target sites of many retroviruses. J. Virol. **79**:5211-5214.

314. **Wyatt, R., P. D. Kwong, E. Desjardins, R. W. Sweet, J. Robinson, W. A. Hendrickson, and J. G. Sodroski**. 1998. The antigenic structure of the HIV gp120 envelope glycoprotein. Nature **393**:705-711.

315. **Xiao, Z., E. Ehrlich, Y. Yu, K. Luo, T. Wang, C. Tian, and X. F. Yu**. 2006. Assembly of HIV-1 Vif-Cul5 E3 ubiquitin ligase through a novel zinc-binding domain-stabilized hydrophobic interface in Vif. Virology **349**:290-299.

316. **Xu, H., E. Chertova, J. Chen, D. E. Ott, J. D. Roser, W. S. Hu, and V. K. Pathak**. 2007. Stoichiometry of the antiviral protein APOBEC3G in HIV-1 virions. Virology **360**:247-256.

317. **Xu, H., E. S. Svarovskaia, R. Barr, Y. Zhang, M. A. Khan, K. Strebel, and V. K. Pathak**. 2004. A single amino acid substitution in human APOBEC3G antiretroviral enzyme confers resistance to HIV-1 virion infectivity factor-induced depletion. Proc. Natl. Acad. Sci. U. S. A **101**:5652-5657.

318. **Yamaguchi, J., S. G. Devare, and C. A. Brennan**. 2000. Identification of a new HIV-2 subtype based on phylogenetic analysis of full-length genomic sequence. AIDS Res. Hum. Retroviruses **16**:925-930.

319. **Yang, B., K. Chen, C. Zhang, S. Huang, and H. Zhang**. 2007. Virion-associated uracil DNA glycosylase-2 and apurinic/apyrimidinic endonuclease are involved in the degradation of APOBEC3G-edited nascent HIV-1 DNA. J. Biol. Chem. **282**:11667-11675.

320. **Yang, B., L. Gao, L. Li, Z. Lu, X. Fan, C. A. Patel, R. J. Pomerantz, G. C. DuBois, and H. Zhang**. 2003. Potent suppression of viral infectivity by the peptides that inhibit multimerization of human immunodeficiency virus type 1 (HIV-1) Vif proteins. J. Biol. Chem. **278**:6596-6602.

321. **Yang, C., M. Li, R. D. Newman, Y. P. Shi, J. Ayisi, A. M. van Eijk, J. Otieno, A. O. Misore, R. W. Steketee, B. L. Nahlen, and R. B. Lal**. 2003. Genetic diversity of HIV-1 in western Kenya: subtype-specific differences in mother-to-child transmission. AIDS **17**:1667-1674.

322. **Yang, C., M. Li, Y. P. Shi, J. Winter, A. M. van Eijk, J. Ayisi, D. J. Hu, R. Steketee, B. L. Nahlen, and R. B. Lal**. 2004. Genetic diversity and high proportion of intersubtype recombinants among HIV type 1-infected pregnant women in Kisumu, western Kenya. AIDS Res. Hum. Retroviruses **20**:565-574.

323. **Yang, X. and D. Gabuzda**. 1998. Mitogen-activated protein kinase phosphorylates and regulates the HIV-1 Vif protein. J. Biol. Chem. **273**:29879-29887.

324. **Yang, X., J. Goncalves, and D. Gabuzda**. 1996. Phosphorylation of Vif and its role in HIV-1 replication. J. Biol. Chem. **271**:10121-10129.

325. **Yu, H., A. E. Jetzt, Y. Ron, B. D. Preston, and J. P. Dougherty**. 1998. The nature of human immunodeficiency virus type 1 strand transfers. J. Biol. Chem. **273**:28384-28391.

326. **Yu, Q., D. Chen, R. Konig, R. Mariani, D. Unutmaz, and N. R. Landau**. 2004. APOBEC3B and APOBEC3C are potent inhibitors of simian immunodeficiency virus replication. J. Biol. Chem. **279**:53379-53386.

327. **Yu, X., Y. Yu, B. Liu, K. Luo, W. Kong, P. Mao, and X. F. Yu**. 2003. Induction of APOBEC3G ubiquitination and degradation by an HIV-1 Vif-Cul5-SCF complex. Science **302**:1056-1060.

328. **Yu, Y., Z. Xiao, E. S. Ehrlich, X. Yu, and X. F. Yu**. 2004. Selective assembly of HIV-1 Vif-Cul5-ElonginB-ElonginC E3 ubiquitin ligase complex through a novel SOCS box and upstream cysteines. Genes Dev. **18**:2867-2872.

329. **Zacharova, V., M. L. Becker, V. Zachar, P. Ebbesen, and A. S. Goustin**. 1997. DNA sequence analysis of the long terminal repeat of the C subtype of human immunodeficiency virus type 1 from Southern Africa reveals a dichotomy between B subtype and African subtypes on the basis of upstream NF-IL6 motif. AIDS Res. Hum. Retroviruses **13**:719-724.

330. **Zennou, V., D. Perez-Caballero, H. Gottlinger, and P. D. Bieniasz**. 2004. APOBEC3G incorporation into human immunodeficiency virus type 1 particles. J. Virol. **78**:12058-12061.

331. **Zetola, N. M. and C. D. Pilcher**. 2007. Diagnosis and management of acute HIV infection. Infect. Dis. Clin. North Am. **21**:19-48, vii.

332. **Zhang, H., B. Yang, R. J. Pomerantz, C. Zhang, S. C. Arunachalam, and L. Gao**. 2003. The cytidine deaminase CEM15 induces hypermutation in newly synthesized HIV-1 DNA. Nature **424**:94-98.

333. **Zhang, L., Y. Huang, T. He, Y. Cao, and D. D. Ho**. 1996. HIV-1 subtype and second-receptor use. Nature **383**:768.

334. **Zhang, L., J. Saadatmand, X. Li, F. Guo, M. Niu, J. Jiang, L. Kleiman, and S. Cen**. 2008. Function analysis of sequences in human APOBEC3G involved in Vif-mediated degradation. Virology **370**:113-121.

335. **Zhang, W., X. Zhang, C. Tian, T. Wang, P. T. Sarkis, Y. Fang, S. Zheng, X. F. Yu, and R. Xu**. 2008. Cytidine deaminase APOBEC3B interacts with heterogeneous nuclear ribonucleoprotein K and suppresses hepatitis B virus expression. Cell Microbiol. **10**:112-121.

336. **Zheng, Y. H., D. Irwin, T. Kurosu, K. Tokunaga, T. Sata, and B. M. Peterlin**. 2004. Human APOBEC3F is another host factor that blocks human immunodeficiency virus type 1 replication. J. Virol. **78**:6073-6076.

337. **Zhu, T., B. T. Korber, A. J. Nahmias, E. Hooper, P. M. Sharp, and D. D. Ho**. 1998. An African HIV-1 sequence from 1959 and implications for the origin of the epidemic. Nature **391**:594-597.

338. **Zhu, Y., H. A. Gelbard, M. Roshal, S. Pursell, B. D. Jamieson, and V. Planelles**. 2001. Comparison of cell cycle arrest, transactivation, and apoptosis induced by the simian immunodeficiency virus SIVagm and human immunodeficiency virus type 1 vpr genes. J. Virol. **75**:3791-3801.

339. **Zhuang, J., A. E. Jetzt, G. Sun, H. Yu, G. Klarmann, Y. Ron, B. D. Preston, and J. P. Dougherty**. 2002. Human immunodeficiency virus type 1 recombination: rate, fidelity, and putative hot spots. J. Virol. **76**:11273-11282.

Appendices

Appendix A: List of APOBEC3G polymorphisms found in all samples

Table A1: SNPs in the 5' region of the APOBEC3G gene

Position [a]	Gene Location	Base Change	% of Reads Supporting Base Change (estimated number of alleles [c])										
			ML1053	ML1102	ML1578	ML1592	ML1857	ML1957	ML1970	ML1975	ML2019	Intermed. Pool [d]	Low Pool [e]
18862066	5' of gene	A to T	43 (1)	58 (1)	46 (1)		33 (1)					25 (3)	19 (33)
18862077	5' of gene	G to C										3 (1)	19 (33)
18862092	5' of gene	C to T[b]		40 (1)									
18862094	5' of gene	C to T[b]		40 (1)									
18862383	5' of gene	G to A[b]						56 (1)					2 (3)
18862429	5' of gene	G to A										10 (1)	8 (14)
18862689	5' of gene	C to G										10 (1)	7 (12)
18862692	5' of gene	C to A	54 (1)	100 (2)	29 (1)	100 (2)	100 (2)	53 (1)	65 (1)	55 (1)	98 (2)	61 (7)	74 (129)
18862771	5' of gene	C to T				100 (2)			65 (1)	43 (1)	50 (1)	4 (1)	17 (30)
18863058	5' of gene	A to G	100 (2)	100 (2)	100 (2)	100 (2)	100 (2)	100 (2)	100 (2)	100 (2)	100 (2)	99 (12)	100(174)
18863080	5' of gene	G to C	100 (2)	100 (2)	100 (2)	100 (2)	100 (2)	100 (2)	100 (2)	100 (2)	100 (2)	82 (10)	80 (139)
18863485	5' of gene	A to T[b]										10 (1)	7 (12)
18863488	5' of gene	A to C[b]										10 (1)	7 (12)
18863557	5' of gene	C to T[b]									51 (1)		
18863561	5' of gene	C to G	46 (1)	100 (2)	29 (1)	100 (2)	97 (2)	61 (1)	72 (1)	55 (1)	100 (2)	60 (7)	53 (92)

[a] Position is numbered in accordance with the NCBI SNP database for the human APOBEC3G gene.
[b] These SNPs have not been previously identified.
[c] The estimated number of alleles was determined by multiplying the observed frequency of sequences with the SNP by the total number of alleles in the pool (i.e. twice the sample number) and rounding to the nearest whole number.
[d] Intermediate Pool is composed of 6 samples from patients with intermediately hypermutated HIV-1 provirus.
[e] Low Pool is composed of 87 samples from patients with non-significantly hypermutated HIV-1 provirus.

Table A2: SNPs in intron 1, exon 2, exon 3, intron 3 and exon 4 of the APOBEC3G gene

Position [a]	Gene Location	Base Change	% of Reads Supporting Base Change (estimated number of alleles [c])										
			ML1053	ML1102	ML1578	ML1592	ML1857	ML1957	ML1970	ML1975	ML2019	Intermed. Pool [d]	Low Pool [e]
18863953	intron 1	G to A[b]							50 (1)				
18864201	intron 1	C to G			72 (1)								5 (9)
18864362	intron 1	A to G	46 (1)	45 (1)	38 (1)		34 (1)					29 (3)	17 (30)
18864507	intron 1	C to T	55 (1)					49 (1)	42 (1)	56 (1)		26 (3)	13 (23)
18864507	intron 1	C to G			55 (1)								4 (7)
18865364	intron 1	G to A[b]									49 (1)		2 (3)
18865575	exon 2	C to T			38 (1)								
18865988	intron 2	C to T[b]									53 (1)		1 (2)
18866274	intron 2	C to G	48 (1)	41 (1)	22 (1)		24 (1)					27 (3)	18 (31)
18866479	intron 2	G to A						43 (1)				6 (1)	12 (21)
18866626	intron 2	C to G[b]									53 (1)		3 (5)
18866680	intron 2	T to C[b]									56 (1)		1 (2)
18866848	intron 2	G to A[b]									51 (1)		
18866897	intron 2	A to G			63 (1)								7 (12)
18867461	exon 3	C to T[b]									46 (1)		2 (3)
18867767	intron 3	G to A[b]						51 (1)					2 (3)
18867892	intron 3	C to G				100 (2)	58 (1)		67 (1)	54 (1)	44 (1)	28 (3)	26 (45)
18867906	intron 3	T to C				100 (2)	58 (1)		66 (1)	54 (1)	43 (1)	27 (3)	25 (44)
18868081	exon 4	A to G				100 (2)	70 (1)		64 (1)	59 (1)	45 (1)	32 (4)	22 (38)

[a] Position is numbered in accordance with the NCBI SNP database for the human APOBEC3G gene.
[b] These SNPs have not been previously identified.
[c] The estimated number of alleles was determined by multiplying the observed frequency of sequences with the SNP by the total number of alleles in the pool (i.e. twice the sample number) and rounding to the nearest whole number.
[d] Intermediate Pool is composed of 6 samples from patients with intermediately hypermutated HIV-1 provirus.
[e] Low Pool is composed of 87 samples from patients with non-significantly hypermutated HIV-1 provirus.

Table A3: SNPs in intron 4 of the APOBEC3G gene

Position [a]	Gene Location	Base Change	% of Reads Supporting Base Change (estimated number of alleles [c])										
			ML1053	ML1102	ML1578	ML1592	ML1857	ML1957	ML1970	ML1975	ML2019	Intermed. Pool [d]	Low Pool [e]
18868173	intron 4	C to T		47 (1)								23 (3)	14 (24)
18868466	intron 4	C to A[b]									52 (1)		
18868487	intron 4	G to T				100 (2)	63 (1)		75 (1)	44 (1)	50 (1)	23 (3)	28 (49)
18868539	intron 4	A to T[b]							79 (1)				2 (3)
18868582	intron 4	T to C											16 (28)
18868816	intron 4	G to A			65 (1)								9 (16)
18868856	intron 4	T to A				100 (2)			57 (1)	53 (1)	44 (1)	8 (1)	15 (26)
18868857	intron 4	C to A				100 (2)			57 (1)	53 (1)	44 (1)	8 (1)	15 (26)
18868907	intron 4	T to G[b]										4 (1)	10 (17)
18868942	intron 4	G to A				98 (2)	51 (1)		60 (1)	43 (1)	61 (1)	23 (3)	24 (42)
18868944	intron 4	C to G[b]										4 (1)	12 (21)
18869511	intron 4	C to T[b]									59 (1)		2 (3)
18869620	intron 4	C to G	48 (1)	58 (1)	44 (1)							21 (3)	27 (47)
18869867	intron 4	G to A	32 (1)	53 (1)	41 (1)		32 (1)					25 (3)	25 (44)
18869868	intron 4	T to A	32 (1)	53 (1)	41 (1)	100 (2)	98 (2)	54 (1)	46 (1)	42 (1)	100 (2)	53 (6)	76 (132)
18869924	intron 4	G to A[b]									49 (1)		1 (2)
18870035	intron 4	G to T				100 (2)	68 (1)		70 (1)	38 (1)	49 (1)	26 (3)	37 (64)
18870089	intron 4	C to G						39 (1)				9 (1)	11 (19)

[a] Position is numbered in accordance with the NCBI SNP database for the human APOBEC3G gene.
[b] These SNPs have not been previously identified.
[c] The estimated number of alleles was determined by multiplying the observed frequency of sequences with the SNP by the total number of alleles in the pool (i.e. twice the sample number) and rounding to the nearest whole number.
[d] Intermediate Pool is composed of 6 samples from patients with intermediately hypermutated HIV-1 provirus.
[e] Low Pool is composed of 87 samples from patients with non-significantly hypermutated HIV-1 provirus.

Table A4: SNPs in intron 5 of the APOBEC3G gene

Position [a]	Gene Location	Base Change	% of Reads Supporting Base Change (estimated number of alleles [c])										
			ML1053	ML1102	ML1578	ML1592	ML1857	ML1957	ML1970	ML1975	ML2019	Intermed. Pool [d]	Low Pool [e]
18870678	intron 5	G to A				100 (2)			63 (1)	45 (1)	47 (1)	9 (1)	23 (40)
18870730	intron 5	C to T[b]		51 (1)								16 (2)	26 (45)
18870760	intron 5	C to T						48 (1)			59 (1)	12 (1)	11 (19)
18870774	intron 5	C to T	54 (1)	49 (1)	33 (1)	100 (2)	86 (2)	50 (1)	62 (1)	54 (1)	100 (2)	61 (7)	69 (120)
18870775	intron 5	A to G	54 (1)	49 (1)	33 (1)	100 (2)	91 (2)	50 (1)	62 (1)	54 (1)	100 (2)	61 (7)	69 (120)
18870776	intron 5	C to T[b]	54 (1)	49 (1)	33 (1)		34 (2)					38 (5)	32 (56)
18870844	intron 5	A to G[b]											10 (17)
18870882	intron 5	T to C[b]											12 (21)
18870945	intron 5	A to G[b]											13 (23)
18871178	intron 5	C to T						56 (1)				9 (1)	15 (26)
18871212	intron 5	A to G	49 (1)	49 (1)	40 (1)	100 (2)	98 (1)	58 (1)	55 (1)	57 (1)	100 (2)	56 (7)	73 (127)
18871702	intron 5	C to T						64 (1)				9 (1)	10 (17)
18871810	intron 5	A to G									59 (1)		2 (3)
18871908	intron 5	C to G	51 (1)	46 (1)	41 (1)	99 (2)	89 (2)	64 (1)	58 (1)	48 (1)	100 (2)	60 (7)	68 (118)
18871910	intron 5	G to A				99 (2)	62 (1)		58 (1)	48 (1)	36 (1)	28 (3)	43 (75)
18872177	intron 5	C to T	34 (1)	44 (1)	41 (1)		33 (1)					27 (3)	22 (38)
18872210	intron 5	A to C[b]											12 (21)
18872375	intron 5	C to G	45 (1)	32 (1)	42 (1)	98 (2)	95 (2)	49 (1)	54 (1)	48 (1)	98 (2)	45 (5)	45 (78)
18872410	intron 5	G to A	45 (1)	42 (1)	41 (1)	100 (2)	95 (2)	54 (1)	33 (1)	59 (1)	100 (2)	55 (7)	60 (104)
18872538	intron 5	C to A	54 (1)	50 (1)	46 (1)	100 (2)	97 (2)	67 (1)	74 (1)	49 (1)	96 (2)	60 (7)	62 (64)
18872543	intron 5	G to A	53 (1)	55 (1)	46 (1)		22 (1)					21 (3)	17 (30)
18872768	intron 5	A to C[b]	46 (1)		38 (1)			41 (1)	56 (1)			48 (6)	38 (66)
18872782	intron 5	C to T	34 (1)	39 (1)	27 (1)		22 (1)					25 (3)	24 (42)

[a] Position is numbered in accordance with the NCBI SNP database for the human APOBEC3G gene.
[b] These SNPs have not been previously identified.
[c] The estimated number of alleles was determined by multiplying the observed frequency of sequences with the SNP by the total number of alleles in the pool (i.e. twice the sample number) and rounding to the nearest whole number.
[d] Intermediate Pool is composed of 6 samples from patients with intermediately hypermutated HIV-1 provirus.
[e] Low Pool is composed of 87 samples from patients with non-significantly hypermutated HIV-1 provirus.

Table A5: SNPs in exon 6, intron 6, the untranslated region of exon 8 and 3' of the APOBEC3G gene

Position [a]	Gene Location	Base Change	% of Reads Supporting Base Change (estimated number of alleles [c])										
			ML1053	ML1102	ML1578	ML1592	ML1857	ML1957	ML1970	ML1975	ML2019	Intermed Pool [d]	Low Pool [e]
18872886	exon 6	C to G	41 (1)	41 (1)	38 (1)		37 (1)					22 (3)	25 (44)
18873142	intron 6	G to A	42 (1)	48 (1)	43 (1)		29 (1)					31 (4)	14 (24)
18873324	intron 6	G to A	35 (1)	46 (1)		100 (2)	97 (2)		67 (1)	46 (1)	48 (1)	53 (6)	51 (89)
18873947	exon 8 UTR	G to A									51 (1)		5 (8)
18873974	exon 8 UTR	A to T									51 (1)		5 (8)
18874234	exon 8 UTR	G to C								45 (1)			
18874347	3'	C to T				97 (2)			68 (1)	53 (1)	100 (2)	7 (1)	31 (54)
18874359	3'	C to T				100 (2)			68 (1)	55 (1)	100 (2)	6 (1)	18 (31)
18874438	3'	- to ACCCTC ACTCAC AGAGCC CCGCCC	22 (1)										
18874465	3'	T to C					31 (1)					13 (2)	15 (26)
18874465	3'	- to CCTCAC TCACAG AGCCCC GCCCAC		19 (1)			41 (1)					20 (2)	10 (17)
18874482	3'	T to C	39 (1)	45 (1)		100 (2)	94 (2)	62 (1)	68 (1)	46 (1)	100 (2)	58 (7)	72 (125)
18874693	3'	C to G	37 (1)	54 (1)		97 (2)	98 (2)		70 (1)	48 (1)	100 (2)	39 (5)	52 (90)
18874840	3'	G to T	43 (1)	53 (1)			29 (1)					26 (3)	17 (30)

[a] Position is numbered in accordance with the NCBI SNP database for the human APOBEC3G gene.
[b] These SNPs have not been previously identified.
[c] The estimated number of alleles was determined by multiplying the observed frequency of sequences with the SNP by the total number of alleles in the pool (i.e. twice the sample number) and rounding to the nearest whole number.
[d] Intermediate Pool is composed of 6 samples from patients with intermediately hypermutated HIV-1 provirus.
[e] Low Pool is composed of 87 samples from patients with non-significantly hypermutated HIV-1 provirus.

Appendix B: Sequences submitted to Genbank

The full-length HIV-1 proviral sequences from 10 different subjects have been deposited in GenBank with accession numbers EU110085 – EU110097.

The 590 nucleotide long HIV-1 proviral sequences from 240 different subjects have been deposited in GenBank with accession numbers EU836326 – EU836565.

The 590 nucleotide long matched HIV-1 plasma RNA sequences from 18 different subjects have been deposited in GenBank with accession numbers EU838566 – EU836593.

Unique clones of the 590 nucleotide long HIV-1 proviral sequence from one individual with non-hypermutated provirus and all thirteen individuals with highly hypermutated provirus have been deposited in GenBank with accession numbers EU875081 – EU875368.

The HIV-1 proviral *vif* sequences from six highly hypermutated viruses and eight non-hypermutated viruses have been deposited in GenBank with accession numbers EU839404 – EU839417.

Appendix C: Abbreviations

A	adenine
AAV	adeno-associated virus
AGM	African green monkey
AID	activation-induced deaminase
AIDS	acquired immunodeficiency syndrome
APOBEC	apolipoprotein B mRNA-editing enzyme, catalytic polypeptide
ARV	antiretroviral
BEAST	Bayesian evolutionary analysis sampling trees
C	cytosine
cpz	chimpanzee
CRF	circulating recombinant form
CSW	commercial sex worker
CTL	cytotoxic T lymphocyte
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid
ds	double strand
FCS	fetal calf serum
G	guanine
gp	glycoprotein
HBV	hepatitis B virus
HIV	human immunodeficiency virus
HLA	human leukocyte antigen
HMM	high molecular mass
HPV	human papillomavirus
IDU	intravenous drug user
IL	interleukin
INF	interferon
kb	kilobase
kDa	kilodalton
LINE	long interspersed nuclear element
LMM	low molecular mass
LTR	long terminal repeat
MHC	major histocombatibility
mRNA	messenger RNA
MTCT	mother to child transmission
NF-κB	nuclear factor kappa-light-chain-enhancer of activated B cells
NNRTI	non-nucleoside reverse transcription inhibitors
NRTI	nucleoside reverse transcription inhibitors
ORF	open reading frame
PBL	peripheral blood lymphocyte
PBS	phosphate buffered saline
PHA	phytohemagglutiin
PI	protease inhibitors
PIC	preintegration complex

PMA	phorbol myristate acetate
PMBC	peripheral blood mononuclear cell
PPT	polypurine tract
Ptt	*Pan troglodytes troglodytes*
RANTES	regulated upon activation, normal T-cell expressed and secreted
RFLP	restriction fragment length polymorphism
RNA	ribonucleic acid
RRE	Rev responsive element
RT	reverse transcriptase
SIV	simian immunodeficiency virus
sm	sooty mangabey
SNP	single nucleotide polymorphism
SREBP	sterol regulatory element binding protein
ss	single strand
STI	sexually transmitted infection
T	thymine
TMRCA	the most recent common ancestor
tRNA	transfer RNA
URF	unique recombinant form