

**HIV Subtype C Diversity:  
Analysis of the Relationship of Sequence Diversity  
to Proposed Epitope Locations**



**A minithesis submitted in partial fulfillment of the requirements  
for the degree of Masters in Bioinformatics,  
University of the Western Cape.**

**Supervisor: Winston Hide and Cathal Seoighe**

**December 2002**

**HIV Subtype C Diversity:  
Analysis of the Relationship of Sequence Diversity to Proposed Epitope Locations**

**Elana Ann Ernstoff**

**Key Words**

HIV

Subtype C

Cytotoxic T Lymphocytes (CTL)

Human Leukocyte Antigen (HLA)

Positive Selection

Maximum Likelihood

Phylogenetic

Immunology

Evolution

Southern Africa



## Abstract

Southern Africa is facing one of the most serious HIV epidemics. This project contributes to the HIVNET, Network for Prevention Trials cohort for vaccine development. HIV's biology and rapid mutation rate have made vaccine design difficult. We examined HIV-1 subtype C diversity and how it relates to CTL epitope location along viral *gag* sequences. We found a negative correlation between codon sites under positive selection and epitope regions; suggesting epitope regions are evolutionarily conserved. It is possible that epitopes exist in non-conserved regions, yet fail to be detected due to the reference strain diverging from the circulating viral population. To test if CTL clustering is an artifact of the reference strain, we calculated differences between the *gag* codons and the reference strain. We found a weak negative correlation, suggesting epitopes in less conserved regions maybe evading detection. Locating conserved and optimal epitopes that can be recognized by CTLs is essential for the design of vaccine reagents.

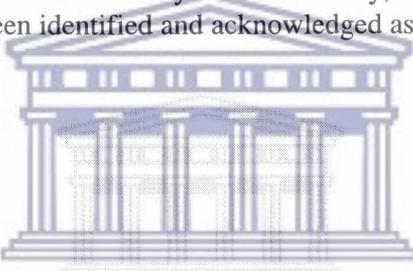
## Declaration

I declare that *HIV Subtype C Diversity: Analysis of the Relationship of Sequence Diversity to Proposed Epitope Locations* is my own work, that it has not been submitted before for any degree or examination in any other university, and that all the sources I have used or quoted have been identified and acknowledged as complete references.

Elana Ann Ernstoff

December 2002

signed:.....



UNIVERSITY of the  
WESTERN CAPE

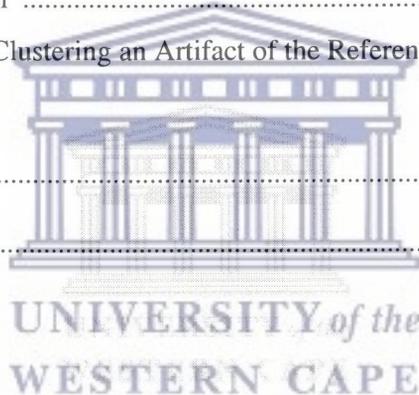
I would like to thank Cathal Seoighe and Winston Hide for their academic guidance, and Carolyn Williamson and Dr. Clive Grey for providing the sequence and epitope data.



## Table of Contents

Title Page .....	1
Key Words .....	2
Abstract .....	3
Declaration .....	4
Chapter 1    Introduction	
Origins and History of HIV .....	8
HIV Life Cycle and Biology .....	10
A Brief Review of the Human Immune System.....	14
Vaccine Design .....	16
Identifying CTL Epitopes and HLA Types .....	17
Human Immune System Response to HIV.....	18
Genomic bias for Adenines affects HIV mutation rates.....	19
Relationship of CTL Escape and Viral Load .....	20
Rare Allele and HLA Advantages .....	21
Obstacles to HIV Vaccine Development .....	22
Molecular Phylogenetic and Evolutionary Modeling .....	23
Codon Substitution Models .....	25
Chapter 2    Data and Methods	
The HIVNET Network for Prevention Trials.....	30
Data Set .....	30
Sequence Alignment and Phylogenetic Reconstruction.....	31
Positive Selection .....	31
Finding the Best Model .....	32
Measure of Variability .....	32

	Differences from the Reference Strain .....	32
	Calculation of correlation coefficient .....	33
Chapter 3	Results	
	CTL Clustering.....	34
	Positive Selection .....	34
	Measure of Variability .....	37
	Distance from the Reference Strain .....	38
Chapter 4	Discussion	
	Correlation between Peptide Density and Positive Selection .....	40
	Is CTL Clustering an Artifact of the Reference Strain? .....	40
Conclusion	.....	41
References	.....	42



## **Chapter 1:**

### **Introduction**

#### **Origin and History of the HIV Epidemic**

In 1983, scientists from the Pasteur Institute isolated a new virus from a patient. Three years later this novel virus was named the human immunodeficiency virus (HIV). This once unheard of virus is now the cause of a global epidemic affecting millions of lives. The most severely affected country is South Africa. In 1990 less than 5% of the population was seropositive, by the end of 2001 an estimated 5 million people were infected (UNAIDS, 2002). The first AIDS cases baffled doctors, and various theories were brought forth in an attempt to explain these abnormal occurrences of opportunistic infections. In 1987, AIDS became the first disease to be discussed by the United Nations general assembly ([www.avert.org](http://www.avert.org)). Finding a solution to HIV requires learning the virus' history, evolution, and molecular biology.

#### *HIV Origins*

Humans are not natural hosts of lentiviruses, our infection is a result of multiple cross-species (zoonosis) events (Hahn, 2000). Phylogenetic analysis of simian viruses can clarify the origins of AIDS and the factors contributing to the epidemic. There are two types of human AIDS viruses, HIV-1 and HIV-2, distinguished according to their genome organization and phylogenetic relationships. It is thought that HIV entered the Human population in at least seven different zoonosis events (Hahn, 2000). Strong evidence suggests that both virus types resulted from simian zoonosis events in Central Western Africa (Hahn, 2000).

Both HIV-1 and HIV-2 are further divided into subtypes. Subtypes are genetically distinct lineages that can be distinguished phylogenetically. HIV-1 has three distinct groups M, N, and O. HIV-2 has six distinct groups A thru F. Comparison of HIV subtypes to simian immunodeficiency virus (SIV) provides clues to the origins of HIV, as well as providing a valuable animal model. To date, five major lineages of primate lentiviruses have been fully sequenced helping to place the origins of HIV in Central Western Africa (Hahn, 2000).

The majority of individuals in the world are infected with HIV-1 group M. Group M includes 11 different subtypes, identified as A, A2, B, C, D, F1, F2, G, H, J, and K. These eleven subtypes are unevenly distributed around the world (Novitsky, 2001). It is well established that HIV rapidly evolves, but there is debate as to what stimulated the rise of multiple subtypes. Possibilities debated include multiple crossover events from the chimpanzee to human, founder-effect, or viral competition. A current theory is that group M is derived from a single chimpanzee to human transmission event, and host pressures are driving viral evolution (Hahn, 2000).

#### *Differences between Subtypes*

Due to the difficulty of developing a vaccine and because different subtypes can be circulating within a single population, it is important for a vaccine to be cross-reactive between subtypes. Therefore it is essential to understand the biological and geographical differences between subtypes. HIV-1 group M subtype C is the predominant subtype in Southern Africa and India. Europe and North America are mostly affected by subtype B. (McCutchan, 1999)

Different subtypes vary in their affect on their host. Variance in disease progression is

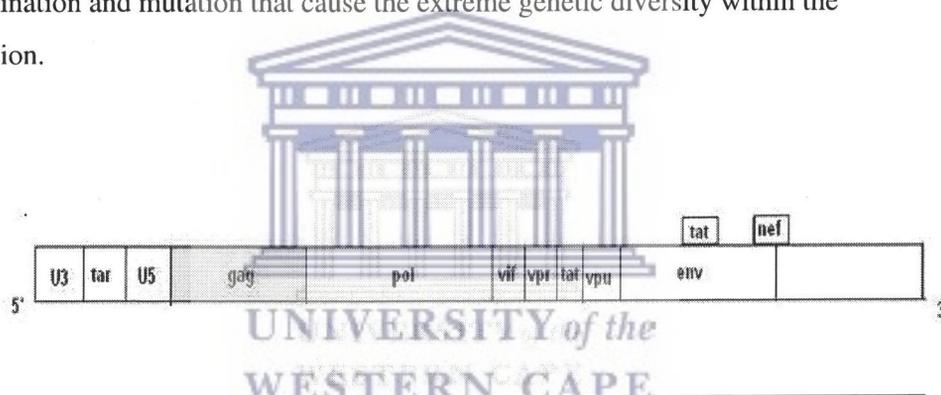
caused by the biological differences between strains and subtypes and the genetic difference between hosts. T-cell counts act as a measure of virus progression, and therefore are a common way to estimate the virulence of subtypes. (Kuiken, 1999). HIV utilizes a co-receptor in order to fuse into the host cell. The most common of these co-receptors are CCR5 or CXCR4. Subtype B can use either one of these co-receptors, whereas the other subtypes seemingly can only use one or the other. (Kuiken, 1999) There are also differences in the RNA's secondary structure, two examples are of the loop regions TAR and V3. TAR is a transcription domain found only in subtype A and subtype A/E mosaics. "The TAR element is a conserved, stable stem-loop structure required for Tat-mediated transactivation of HIV-1 gene expression, and the level of activity of this system may influence the rate of disease progression." (Kuiken, 1999). The V3 loop structurally varies between subtypes and is a region extensively studied. Functionally, it is an important domain of the viral envelope and influences the co-receptor usage (Kuby, 1997). The V3 loop region is one of the most variable regions in the Env protein. Is often used as a way to differentiate between subtypes (Kuiken, 1999) and is highly studied for vaccine purposes. However, there is still insufficient knowledge about the biological difference between subtypes and how these differences affect the epidemic.

## **HIV Life Cycle and Biology**

### *Classification and Genome Organization*

HIV is classified as a Cytopathic retrovirus, belonging to the genus of lentiviruses. Though it shares characteristics with other viruses in the group, such as equine infectious anemia virus and simian immunodeficiency virus (SIV), HIV possesses unique features that make it difficult to treat. Several characteristics differentiate retrovirus from other

viral groups. Their defining feature is their possession of an RNA genome and the enzyme reverse transcriptase, which facilitates the conversion of the RNA into DNA. Other distinguishing traits include a modified chemical structure to enable the different mode of RNA synthesis, a dimeric genome (two subunits which are almost identical in sequence, held together by base pairing), and a unique association with a specific molecule of tRNA, which serves as the primer for reverse transcription. HIV's genome consists of nine genes. Three of these, *gag*, *pol*, and *env* are common to all other retroviruses, the remaining six, *vif*, *vpr*, *tat*, *rev*, *nef*, and *vpu*, are unique to HIV. A list of the genes and the proteins they code for are listed in figure 2 (Kuby, 1997). The physical properties of the HIV viron structure and genome contribute to the high rates of recombination and mutation that cause the extreme genetic diversity within the population.



**Figure 1:** HIV genome organization

Gene	Protein Product	Function
<b>gag</b>	53-kDa precursor p17 p24 p9 p7	<i>Nucleocapsid proteins</i> Forms outer core-protein layer Forms inner core-protein layer Is component of nucleoid core Binds directly to genomic RNA
<b>env</b>	16-kDa precursor gp41 gp120	<i>Envelope glycoproteins</i> Is transmembrane protein associated with gp120 and required for fusion Protrudes from envelope and binds CD4
<b>pol</b>	Precursor p64 p51 p10 p32	<i>Enzymes</i> Has reverse transcriptase and RNase activity Have reverse transcriptase activity Is protease that cleaves gag precursor Is integrase
<b>vif</b>	p23	Promotes infectivity of viral particle

<b>vpr</b>	p15	weakly activates transcription of proviral DNA
<b>tat</b>	p14	Strongly activates transcription of proviral DNA
<b>rev</b>	p19	Allows export of unspliced and singly spliced mRNAs from nucleus
<b>nef</b>	p27	Increases viral replication; down-regulates host-cell CD4
<b>vpu</b>	p16	Is required for efficient viral assembly and budding

**Figure 2:** HIV protein functions.  
(Figure taken from Kuby, 1997)

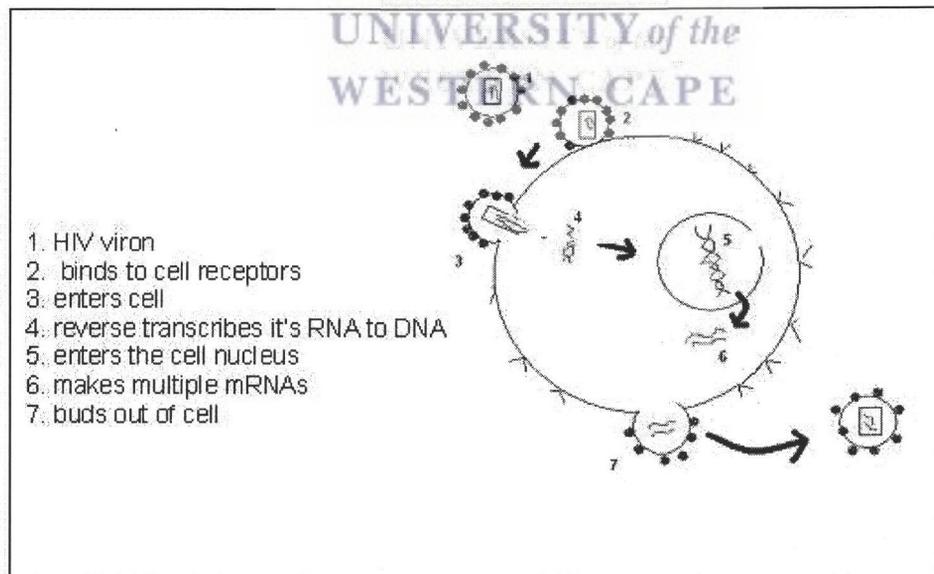
### *Life Cycle*

The life cycle of HIV can be described in six steps: binding to the target cell, fusion into the cell, reverse transcription, integration into the host's genome, replication, and budding of new virions (figure 3). HIV camouflages itself with host cell proteins, preventing the immune system from recognizing it as a foreign-body. The envelope that encases the viral core (nucleocapsid) is mostly covered with host-derived proteins, including major histocompatibility complex (MHC) proteins. Hidden amongst the host cells integrated on the envelope surface are two viral glycoproteins, gp120 and gp41. Glycoprotein 120 facilitates binding, and glycoprotein 41 enables fusion.

The primary host receptor for gp120 is CD4. CD4 receptors are most highly expressed on T-helper ( $T_H$ ) cells making them especially susceptible to HIV infection. However, macrophages, monocytes, dendritic cells, Langerhans cells, hematopoietic stem cells, certain rectal-lining cells, and microglial cells are also susceptible. Following the initial binding of gp120, a co-receptor and gp41 allow fusion of the virus into the target cell. Once inside the virus sheds its coat and reverse transcribes its RNA into DNA. This DNA then enters the host-cell nucleus, and with the help of an enzyme integrase becomes integrated into the host cell's genome, forming what is referred to as a provirus. Here the

virus can remain dormant for an undetermined amount of time. As long as the provirus remains in the latent state the viral genes are not expressed however, each time the target cell replicates the viral DNA is passed along to daughter cells. (Kuby, 1997)

Activation of the provirus marks the beginning of transcription of HIV's structural genes and the formation of new virions. Transcription results in mRNA, which is then translated into viral proteins and the synthesis of single stranded RNA (ssRNA). The virus assembles into a new viron, which will bud out of the cell in order to enter new cells. The host cell's membrane is modified by the insertion of gp41 and gp120. The viral ssRNA and core proteins assemble beneath the modified membrane and while budding acquire the modified host plasma membrane as its envelope. Budding often results in lyses of infected T-cells. T-cells also tend to bud a greater number of virions than other cells. Macrophages have lower levels of budding but usually do not lyse and therefore continue producing low levels of the virus.



**Figure 3:** HIV Life Cycle

### *HIV Mutation*

HIV's transcription is extremely error prone and mutates at an estimated rate of five to ten errors per replication cycle (Kurby, 1997). Furthermore the dimeric nature of HIV allows for high rates of recombination between the two RNA strands during viral DNA synthesis. The result is that no two HIV patients carry identical viruses, making vaccine design a complicated task. The medical field has to issue a new flu vaccine each year for it to be effective, and the mutation rate in HIV is 65 times greater than that of the influenza virus (Kuby, 1997). Less than a 2% amino acid change can cause a failure in the cross-reactivity of the flu vaccine (Gaschen, 2002). HIV's sequence diversity is constantly increasing and strains within the same subtype can have regions that differ up to 20% (Gashen, 2002). Now that more is understood about HIV molecular biology, and particularly with the drive to develop an effective vaccine, the challenge is to understand exactly how and why the virus is changing. It is crucial to understand the evolutionary pressures the immune system places on the virus, and to discover regions necessary for viral infection and replication in order to gain a better comprehension of viral escape from an immune response.

### **A Brief Review of the Human Immune System**

The immune system is a coordinated effort of numerous cell types, which together function to recognize and respond to foreign particles. The immune system has to be able to identify subtle physical and chemical differences between particles in order to distinguish between foreign cells (antigens) and self-cells. Once an antigen is recognized the appropriate effector cells must be activated in order to eliminate or neutralize the

foreign body. To avoid future infection, exposure to an antigen also induces a memory response (Kuby 1997).

### *Cell Types*

Two cell types are required to mount an effective immune response, lymphocytes and antigen presenting cells. Lymphocytes are one of many types of white blood cells; they process antigen-binding cell surface receptors. There are two main groups of lymphocytes, B lymphocytes (B cells) and T lymphocytes (T cells). The immune response is further divided into two types of responses, the humoral response and the cell-mediated response. The former includes the reaction of B cells with an antigen and their differentiation into plasma cells. The latter, the cell-mediated response includes the T cells responding to an antigen. There are two main types of T cells, T helper ( $T_H$ ) and T cytotoxic ( $T_C$ ) cells. It has been suggested that there may also be a third type, T suppressor cells ( $T_S$ ), but they are currently not well defined. Both  $T_H$  and CTL cells make up the effector cells of the cell-mediated immune response. (Kuby, 1997).

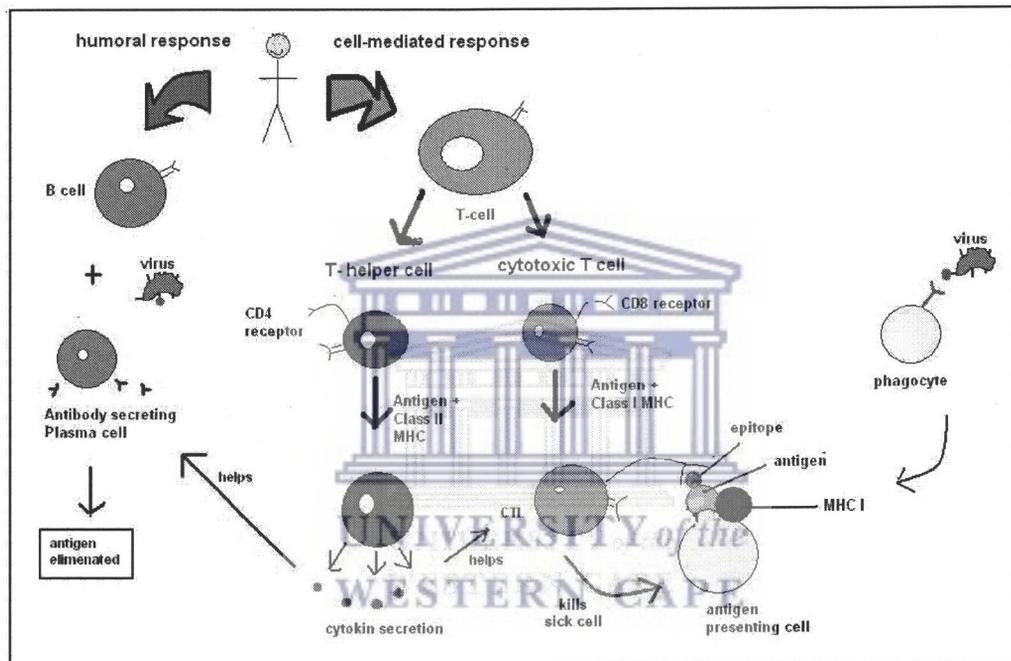
### *Major Histocompatibility Complex*

Both the humoral and cell-mediated responses require cytokines produced by the  $T_H$  cells. Cytokines are proteins important in assisting and regulating effector cells. Inappropriate use of cytokines can have severe autoimmune consequences; therefore activation is highly regulated. T-cells will only react with a complex consisting of an antigen and major histocompatibility complex (MHC) class II molecule presented on the surface of an antigen-presenting cell. MHC molecules in Humans are called Human Leukocyte Antigens (HLA), and are large glycoprotein complexes distributed over multiple loci. The body does not recognize an entire antigen only a discreet region. The site that the T or B cell recognizes is referred to as an epitope. The set of HLA types that a person possesses determines what epitopes a person can respond to. (Kuby, 1997)

#### *Antigen Presenting Pathways*

An antigen-presenting cell will ingest an antigen either by phagocytosis, endocytosis, or both, and present a peptide fragment (the epitope) on its surface. The antigen, coupled with a MHC molecule, can be recognized by a T cell. Antigens can be produced outside of a host cell (exogenous antigen) or within an infected cell (endogenous antigen). Viral proteins are considered endogenous. There are two separate pathways for dealing with intracellular (endogenous) and extracellular (exogenous) antigens. Endogenous antigens are processed in the cytosolic pathway and presented with class I MHC molecules and exogenous antigens are processed in the endocytic pathway and presented with class II MHC molecules. The class of MHC molecule determines the type of T cell it will be recognized by. T cells with a CD4 receptor, usually  $T_H$  cells, recognize an antigen presented with a class II MHC molecule. T cells with a CD8 receptor, which include  $T_C$  cells, recognize those with a class I MHC molecule. When a  $T_H$  cell is activated it becomes an effector cell that secretes various growth factors, cytokines, which play a role in activating B cells. When a  $T_C$  cell is activated it proliferates and differentiates into an

effector cell referred to as a cytotoxic T lymphocyte (CTL). CTLs do not secrete many cytokines, instead they exhibit cytotoxic activity. They are important for monitoring the cells of the body and eliminating any that display an antigen, such as a virus infected cell, tumor cell, or a cell of a foreign tissue graft. They are essential for recognizing altered self-cells, such as cells infected with a virus. (Kuby, 1997)



**Figure 4:**  
The human immune system.

### Vaccine Design

Vaccination is the most practical method of preventing and controlling viral diseases, including HIV. New discoveries concerning antigen processing pathways, T cells, and B cells are helping vaccine designers maximize vaccine effectiveness. A vaccine that

generates an immune response does not necessarily lead to immunization. A vaccine can induce a successful primary response but fail to produce memory cells leaving the individual susceptible to infection if exposed again. The role of memory cells is partially dependent on the period of incubation of the virus. For example, the influenza virus has a short incubation of less than three days. By the time the memory cells are activated disease symptoms are already in progress. The vaccine requires the maintenance of a high level of neutralizing antibodies and repeated immunization. In contrast, Polio requires at least three days for incubation providing sufficient time for memory cells to be activated. The vaccine approach is different and is designed to induce high levels of memory cells. HIV has an incubation period of approximately twelve hours (Kuby, 1997). HIV vaccine design efforts apply knowledge of the human immune system, HIV biology, and evolution.

### **Identifying CTL Epitopes and HLA Types**

The immune system does not recognize an entire antigen only the region that is the epitope. Identification of epitope regions is key to vaccine design. However, a problem arises in defining a three dimensional region according to a two dimensional sequence. The textbook definition of an epitope is the region to which an antibody binds but this is different to the definition often applied in laboratories. Is it right to consider an epitope to be strictly the residues which make contact with the ligand? (Greenspan, 1999) The debate lies in whether or not to include residues that contact the ligand but are energetically neutral, and the residues which have no physical contact with the ligand but whose substitution would affect the ligands ability of recognition. Epitope searching has been likened to looking for a needle in a haystack, but already a large number of CTL epitopes are documented. The time for a mutation to arise affects the viron strain's interaction with the host immune system. CTL epitope are not evenly distributed along the genome, but rather cluster in conserved regions (Yusim, 2002). Much more work is

required and a systematic mapping of relevant CTL epitopes is still in its early stages. For historical and economic reasons, the majority of past work is based on HIV-1B and HLA types found in Caucasians (Novitsky, 2001). To date, there have been few studies based on HIV-1C and other non-B subtypes. This is, however, changing with the increasing global research effort of non-B subtypes.

### **Human Immune System Response to HIV**

The immune system is in a continuous battle with HIV, with both sides suffering tremendous losses. The initial infection elicits a rapid increase in CTL cells. It is estimated that approximately half of all T lymphocytes are involved and rapid division occurs over the first two to three week period. CTL levels peak leading to a drop in viral load. A negative correlation between viral load and CTL response is then able to persist until a point. Something changes and CTL cells start losing their ability to contain the virus. Viral load dramatically increases and the immune system loses the battle.

CTL response is essential for controlling the virus (McMichael, 2002). A strong CTL response correlates with low plasma viremia, and a prolonged asymptomatic stage. HLA- type affects the effectiveness of the CTL response, as CTLs can only recognize an epitope bound to and HLA class I molecule. Although further study on the immune system's initial response to HIV is needed, McMichael et al. (McMichael, 2001) suggest that CD4<sup>+</sup>T cells are damaged early in primary infection and that this results in a suboptimal response by the CD8<sup>+</sup> T cells (the CTLs). It is possible that a vaccine, which can stimulate both a CD4<sup>+</sup> and CD8<sup>+</sup> T-cell response to HIV, might be an effective way of controlling the virus at an early stage so that damage to the immune system is minimal (McMichael, 2001).

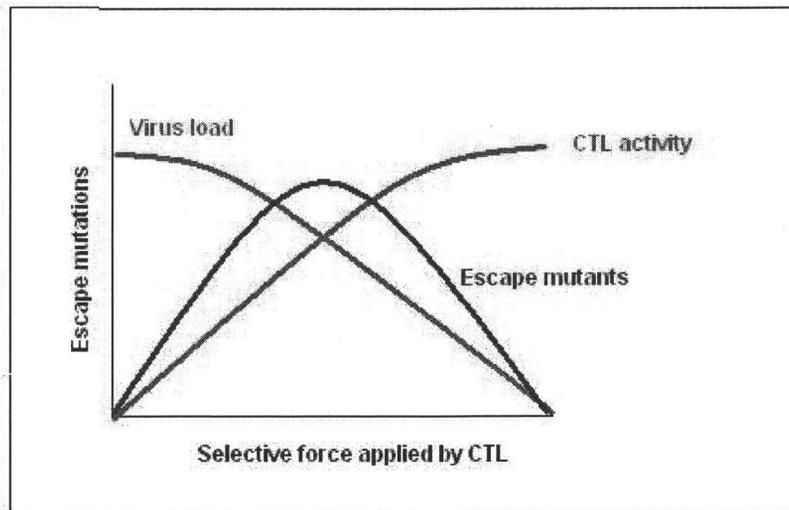
CTLs have multiple antiviral mechanisms. It is currently unclear which of these functions is most important for viral control. CTL cells have the ability to kill infected cells and

produce cytokines. Cytokines affect viral replication through their influence on T<sub>H</sub> cells activation and proliferation. The production of CC chemokines (a type of cytokine secreted by TH cells) (Kuby, 1997) such as MIP-1 $\alpha$ , MIP-1 and RANTES, suppress HIV replication by competing for or down-regulating CCR-5, a primary co-receptor for CD4 and necessary for fusion of the viron into the cell. CTLs secrete these antiviral factors at sites of viral replication. Perforin is another weapon of CTL cells. It is a protein made by CD8<sup>+</sup> T-cells and, together with granzymes, it is an important trigger of cell death (Viscidi, 1999). However, at some point during infection CTLs become incapable of controlling the virus. A possible contributing factor to this failure is other T-cells inability to help. The immune system is able to diminish the numbers of the virus initially, however what the immune system is effectively doing is positively selecting for CTL resistant HIV mutants. These mutants are not recognized by HLA molecule. Since only the virions capable of escape survive to reproduce eventually the entire viral population are CTL escape mutants. At this point the immune system is completely defenseless against the virus, and HIV is free to reproduce killing the immune system as it does so, thereby leaving the infected individual susceptible to opportunistic infections.



### **Genomic bias for Adenines affects HIV mutation rates**

HIV variation does not occur uniformly along the genome. The majority of HIV genetic variation is the result of point mutations, and to a lesser extent insertions and deletions (Viscidi, 1999). HIV has different nucleotide ratios than other retroviruses. HIV consists of approximately 40% adenines (A); this favoring of adenine is not seen in other retrovirus genomes, and has evolutionary consequences (Viscidi, 1999). During reverse transcription errors from G to A are the most frequent. The favoring of transitions to adenines affects which mutations are most likely. This can be seen clearly for example, in the case of the two drug resistant mutants, 184-I and 184-V. The wild type codon is



**Figure 5:**  
Representation of virus replication rate and CTL level  
(McMichael, 2001)

#### Rare Allele and HLA Advantages

With 5 to 10 mutations per replication and 180 generations per year, HIV generates enormous amounts of genetic variation (Shankarappa, 1999). CTL response is one of the key factors controlling the virus, and is also a driving factor behind its evolution. Mutations that do not functionally hinder the virus and enable them instead to escape a CTL response proliferate rapidly in the host. These escape mutations eventually will comprise the entire viral population in the host. How long the virus takes to develop escape mutants is dependent on the host HLA background and if the virus has encountered similar HLA types in recent previous infections. Since HLA type is inherited, mothers often transmit escape mutants to their unborn child in the case of mother to child HIV transmissions. The asymptomatic window is often not observed in these children. HLA B27 type is especially effective at controlling the virus (Goulder, 2001). Individuals with this HLA experience a slower period of disease progression.

However, children who inherited the HLA B27 allele from their mothers do not experience the advantage of having the HLA B27. This suggests that the mother is transmitting escape mutants. In a study by Goulder et al. one child inherited the HLA B27 paternally. In this case the child displayed what is often observed in adults, and had a slow progression towards AIDS, rarely seen in children (Goulder, 2001).

### *Natural Immunity*

Specific HLA types may slow progression to AIDS, however certain individuals appear to be resistant to the virus. The first genetic factor to show resistance to HIV is a deletion rendering the co-receptor CCR-5 inaccessible to HIV (Rowland-Jones, 1998). This mechanism has only been observed in Caucasians who are homozygous for a 32 base pair deletion in the coding sequence for CCR-5. This deletion prevents expression of the CCR-5 receptor, a primary co-receptor for HIV fusion. However, this deletion is not present in the prostitutes who are frequently exposed to HIV over years and have shown no signs of HIV (Rowland-Jones, 1998), suggesting that there may be multiple causes of natural immunity. Prostitutes are being carefully studied for clues of how to fight the virus. A study reported by Rowland-Jones et al. consisted of a group of Nairobi prostitutes. Each woman remained seronegative for twelve years despite being exposed to what is considered to be one of the highest HIV risk areas (Rowland-Jones, 1998). Cross-clade CTL activity may possibly be responsible for these women's protection. The group of women displayed a range of immune responses. It is unclear whether these responses prevent infection or if the detected antibodies (which were in conserved regions) are marks of numerous successful immune responses (Rowland-Jones, 1998).

### **Obstacles to HIV Vaccine Development**

In April of 1984 Margaret Heckler, Secretary of the United States Health and Human Services, announced "We hope to have a vaccine [against AIDS] ready for testing in about two years." (www.avert.org ). All types of vaccine approaches are being considered; yet certain obstacles stand in the way of success. For a vaccine to be effective it must be able to produce an immune response that protects the host from the pathogen. The type of immune response that would protect against HIV is still unknown; some antibodies can actually increase infection. HIV's mutation rate and the use of surface glycoproteins for immune escape pose another obstacle. Finding a single solution to these problems is not sufficient. A vaccine or a combination of vaccines needs to be developed which will be effective against different strains and preferably subtypes of HIV. Animal models are useful but have their limits and imperfections. For example, Chimpanzees do not develop an immune deficiency; therefore testing must focus on inhibition of viral replication (Kuby, 1997). Research is also limited by the size of the chimpanzee population. The development of SCID human mice might be a more promising animal model. SCID-hu mice are "reconstituted with human fetal liver, mesenteric lymph node, and thymus...[they] become populated with human T and B lymphocytes and other white blood cells" (Kuby, 1997). These mice have been useful for *in vivo* evaluation of antiviral agents and AIDS vaccines. Azidothymidine, AZT, the first therapeutic drug approved for HIV, inhibits HIV infection in SCID human mice (Gobbi, 2000). New developments in vaccine design are offering hope. However, debate persists about which biological mechanisms to target and which sequence regions will be the most useful for vaccine design.

### **Molecular Phylogenetics and Evolutionary Modeling**

HIV's fast evolutionary rate enables the application of phylogenetic methods to recently diverged viruses. The application of phylogenetics has shed light on the origins, diversity, and selective pressures acting on HIV. There are three main methods for

inferring phylogenetic relationships from molecular sequences, parsimony, maximum likelihood, and distance (neighbor joining). All three have their advantages and disadvantages; the one to choose depends on the data and the facilities available.

### *Parsimony*

Parsimony predicts an evolutionary tree that requires the smallest number of evolutionary changes. The algorithm is computationally simple, however it is not the most efficient, and is limited in the number of sequences it can handle. If there are large numbers of substitutions between the sequences, parsimony may underestimate the number of substitutions in the tree (Hillis, 1999). The parsimony approach is implemented in PAUP as well as DNAPARS, DNAPENNY that are included in Phylip.

### *Maximum Likelihood*

Maximum Likelihood methods search the space of possible phylogenetic trees and use probability calculations to determine which one is the most likely. As with the parsimony method, maximum likelihood performs an analysis on each column of a multiple sequence alignment. However, it has additional parameters relating variations in mutation rates and specific evolutionary models. It has the advantage that it can effectively analyze more divergent sequences than the parsimony method. Maximum likelihood's main drawback is its computational intensity; therefore there is an upper limit on how many sequences can be analyzed. This limit is however diminishing as computing power increases. DNAm1 and DNAm1k are two examples of programs that use the maximum likelihood method, both of which are included in the Phylip package.

The procedure to create a maximum likelihood tree is as follows. A sequence alignment

is created using estimated rates of substitution, from the alignment than DNAmI can be used to produce a maximum likelihood tree. Each tree has a calculated likelihood. "The probability of each tree is simply the product of the mutation rates in each branch of the tree, which itself is the product of the rate of substitution in each branch times the branch length." (Mount, 2001) Each tree is the product of its combined probability that it would produce the original inputted alignment (Mount, 2001). The tree with that has the highest likelihood is considered the maximum likelihood tree.

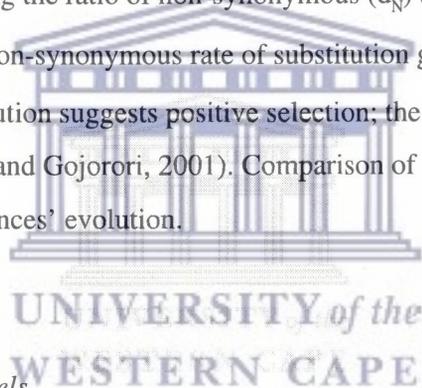
#### *Distance (Neighbor-Joining)*

The Distance methods for finding the best phylogenetic tree are based on distances between sequence pairs. It determines the number of changes between each pair in a group of sequences, starting with the most alike and working outward. The sequence pairs are considered 'neighbors' if they share a node and have the smallest number of changes between them. Phylip includes distance programs as well, called DNADIST and PROTDIST, for working with nucleic acids in the former and proteins in the later (Mount, 2001).

Distance methods are not computationally intense, however their accuracy is dependent on idealized evolutionary conditions. Because the method is dependent on branch lengths, which are determined by additivity, if two or more sequences diverge at approximately the same point in the evolutionary history, it is difficult to know which is the correct tree. Therefore, the accuracy of the topology is dependent on the degree to which the sequence set can be predicted by additivity (Mount, 2001).

#### **Codon Substitution Models**

“Positive selection is an evolutionary event in which a wild-type allele at a locus is replaced by a mutant allele with a higher fitness” (Suzuki and Nei, 2001). The amino acid code is degenerative. A nucleotide change that does not result in a change of amino acid is referred to as a synonymous substitution. Nonsynonymous substitutions are nucleotide changes that do result in a change of amino acid. Synonymous substitutions are said to be invisible to selective pressures (Yang, 2000). Whereas substitutions that lead to a change of amino acid (nonsynonymous substitutions) can alter a protein’s fitness, therefore are subject to natural selection. The observation that adaptive evolution (positive selection) occurs less frequently than purifying selection (random genetic drift) has led to the development of statistical models of codon substitution. Positive selection can be detected by analyzing the ratio of non-synonymous ( $d_N$ ) and synonymous evolutionary rates ( $d_S$ ). A non-synonymous rate of substitution greater than the synonymous rate of substitution suggests positive selection; the opposite suggests negative selection (Suzuki and Gojorori, 2001). Comparison of the substitution rates gives insight into the sequences’ evolution.



#### *The Yang and Nielsen Models*

Models of codon evolution can be a powerful tool for detecting sequence regions under positive selection. The comparison of the two fixation rates of non-synonymous to synonymous mutations is a means of quantifying the effect natural selection has on a particular codon. Nielson and Yang devised thirteen models (statistical distributions) that allow for heterogeneous ratio of synonymous to nonsynonymous substitutions among sequence sites (Yang, 2000). Each model has varying parameters and restrictions. They are effective, but due to the nature of the models can overestimate sites under positive selection.

The ratio of synonymous to nonsynonymous rates is represented by  $d_s/d_n$  (Yang, 2000). When  $d_s/d_n$  is found to be greater than one for a codon, that codon is a candidate for positive selection. Detecting positive selection can draw researchers attention to the importance of a region whose function maybe unknown. The disadvantage of applying this approach to the estimation of sites under selection is that it is heavily model dependent and can give false positives.

### *Nested models*

Two models of evolution are referred to as 'nested' if one is a special case of the other. For example, the neutral model of codon evolution used by Yang (Yang, 2000) is nested within the selection model (table 1). The neutral model has parameters which allow for  $d_s/d_n = 0$  and  $d_n/d_n = 1$ . The selection model includes a third parameter in which  $d_s/d_n$  is allowed to vary. By having nested models a likelihood ratio test can be done to determine if the extra parameter gives a significant improvement in the likelihood of the data. According to the likelihood ratio test, twice the log-likelihood difference ( $2\Delta l$ ) is approximately a  $\chi^2$  distribution. The degree of freedom,  $df$ , is equal to the difference in the number of parameters between the two models (Yang, 2000).

Model code	Number of parameters	Parameters	Notes
M0 (one-ratio)	1		one ratio for all sites
M1 (neutral)	1	$\theta$	$\theta_1 = 1 - \theta, \theta_0 = 0, \theta_2 = 1$
M2 (selection)	3	$\theta, \theta_1, \theta_2$	$\theta_2 = 1 - \theta - \theta_1, \theta_0 = 0, \theta_1 = 1$
M3 (discrete)	2K-1 (K = 3)	$\theta, \theta_1, \dots, \theta_{k-2}, \theta_k$ $\theta_1, \dots, \theta_{k-1}$	$\theta_{k-1} = 1 - \theta - \theta_1 - \dots - \theta_{k-2}$
M4 (freqs)	K - 1 (K = 5)	$\theta, \theta_1, \dots, \theta_{k-1}$	The $\theta_k$ are fixed at 0, $1/3, 2/3, 1$ , and 3
M5 (gamma)	2	$\alpha,$	From $(\alpha, )$
M6 (2gamma)	4	$\theta, \alpha_0, \theta, \alpha_1$	$\theta_0$ from $(\alpha, )$ and $\theta_1 = 1 - \theta_0$ from $(\alpha_1, \alpha_2)$
M7 ( )	2	$, q$	From $( , q)$
M8 ( & )	4	$\theta, p, q,$	$\theta_0$ from $( , q)$ and $\theta_1 - 1$ with
M9 ( & gamma)	5	$\theta, , q\alpha,$	$\theta_0$ from $( , q)$ and $\theta_1 - 1$ from $(\alpha, )$
M10 ( & gamma+1)	5	$\theta, , q\alpha,$	$\theta_0$ from $( , q)$ and $\theta_1 - 1$ from $1 + (\alpha, )$
M11 ( & normal>1)	5	$\theta, , q, \mu,$	$\theta_0$ from $( , q)$ and $\theta_1 - 1$ from $N(\mu, ^2)$ , truncated to $>1$
M12 (0&2normal>1)	5	$\theta, \mu_1, \mu_2, \theta, \theta_2$	$\theta_0$ with $\theta_0 = 0$ and $\theta_1 - \theta_0$ from the mixture: $\theta_1$ from $N(1, ^2_1)$ , and $1 - \theta_1$ from $N(\mu_2, ^2_2)$ , both normals truncated to $>1$
M13 (3normal>0)	6	$\theta, \mu_1, \mu_2, \theta, \theta_1, \theta_2$	$\theta_0$ from $N(0, ^2_1)$ and $\theta_2 = 1 - \theta_0 - \theta_1$ from $N(\mu_2, ^2_2)$ , all normals truncated to $>1$

**Table 1:**

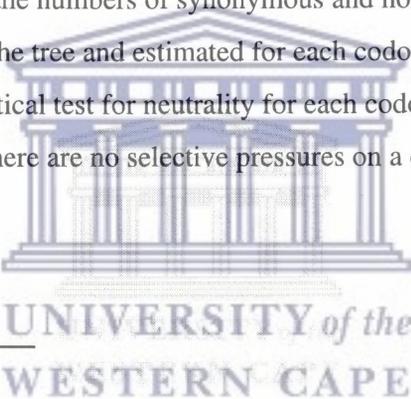
Summary of the Yang Codon Substitution models for Selective Pressure.  
(Table taken from Yang, 2000)



### *The Suzuki and Gojobori Models*

Suzuki and Gojobori also set out to mine DNA sequences in search of positively selected sites, however as opposed to the maximum likelihood approach of Yang and Nielsen, they chose to apply a parsimony based method.

The Suzuki and Gojobori method takes a multiple sequence alignment and takes it through five steps (Suzuki, 1999). Firstly, a phylogenetic tree is reconstructed. Second, the ancestral codon is inferred at each node of the tree for every codon site. Third, for each codon site throughout the tree, the average numbers of synonymous and nonsynonymous sites are estimated. Fourth, the numbers of synonymous and nonsynonymous substitutions are added for the tree and estimated for each codon site. The last step, the fifth one, is to apply a statistical test for neutrality for each codon site. The following equations will hold true if there are no selective pressures on a codon (Suzuki and Gojobori, 1999).


$$\frac{s_c}{S_c + n_c} = \frac{s_t}{S_t + n_t}$$

$$\frac{n_c}{s_c + n_c} = \frac{n_t}{s_t + n_t}$$

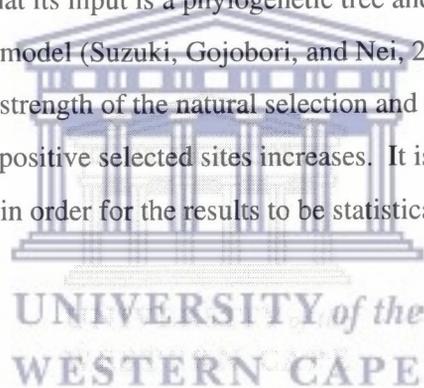
Where the variables,  $s_c$  stands for the total number of synonymous changes,  $n_c$  is the number of nonsynonymous changes, and  $s_t$  and  $n_t$  are the number synonymous and nonsynonymous sites (Suzuki and Gojobori, 1999).

The Suzuki and Gojobori method is not without its limitations. If positive selection occurs sporadically distributed along the phylogenetic tree, this method may fail to detect

it. Sequences that are closely related to each other may also create problems for generating the correct tree topology and would therefore have repercussions throughout the other steps resulting in an incorrect assessment of selective pressures (Suzuki and Gojobori, 1999). The method is most effective and reliable in cases where positive selection was very strong or functioned over a sustained period of time.

#### *ADAPSITE*

ADAPSITE is a program for detecting positive selection at single amino acid sites in a multiple sequence alignment using the parsimony method described above. It is simpler than the Yang models, in that its input is a phylogenetic tree and a sequence alignment, it does not utilize any sort of model (Suzuki, Gojobori, and Nei, 2001). The false positive rate is usually low. As the strength of the natural selection and the branch length increase the accuracy of picking up positive selected sites increases. It is a conservative model, but requires large data sets in order for the results to be statistically significant.

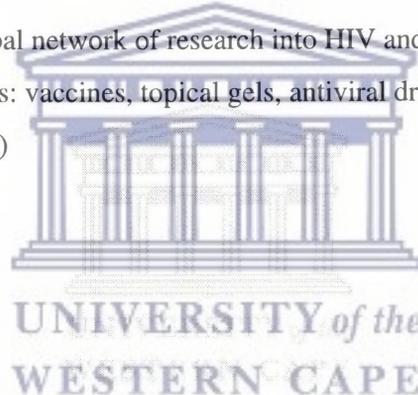


## **Chapter 2:**

### **Data and Methods**

#### **The HIVNET Network for Prevention Trials**

This project was carried out to support HIVNET sponsored research based at the University of Cape Town. The HIVNET, Network for Prevention Trials, is an international collaboration for HIV research. Established in 1993, by the UN Division of AIDS (DAIDS), it is a global network of research into HIV and AIDS. HIVNET sponsored research includes: vaccines, topical gels, antiviral drugs, and preventative behavior. (UNAIDS, 2002)



#### **Data set**

The data set consists of HIV-1 subtype C *gag* sequences and proposed epitope locations. There were 57 sub-Saharan African *gag* sequences sampled from seropositive individuals living in South Africa (Durban and Johannesburg), Malawi, Zambia, and Zimbabwe. Sequencing took place under the direction of Professor Carolyn Williamson at the University of Cape Town. Dr. Clive Grey at the South African National Institute for Communicable Diseases (NICD) provided the proposed epitope locations. The locations of the epitopes were not exactly determined. The epitopes were localized to regions of protein sequences 20 amino acid in length

## Sequence Alignment and Phylogenetic Reconstruction

An in-frame alignment of the *gag* sequences was generated using ClustalW and BioEdit (Hall, 1999). The programs fastDNAm1(Olsen) and DNArates (Felsenstein, 1981) were used to reconstruct a maximum likelihood tree of the *gag* sequences. The two programs are iterated until the log likelihood score associated with the tree stabilizes. FastDNAm1 is based on the Phylip program DNAm1, but is a faster algorithm. DNArates provides rate categories for fastDNAm1. It determines site-specific nucleotide substitution rates according to a maximum likelihood method.

### Positive Selection

Codeml (Yang, 2000) and Adapsite (Suzuki, 2001) were both used to infer sites of positive selection in the *gag* coding sequences. The Yang method can be a powerful tool to detecting positively selected sites but also has the disadvantage of being model rich and overestimating sites under selection. The method uses three types of input: a maximum likelihood tree, an in-frame alignment, and a model of sequence evolution. It estimates (using maximum likelihood) the prior probability that a given codon is evolving under positive selection and applies the prior to determine a posterior probability from Bayes' rule. Bayes' rule allows us to work out the probability of an event given observed data and a prior belief.

#### *Bayes' Rule*

$$\text{Posterior Probability} = \frac{\text{conditional likelihood} \times \text{prior}}{\text{likelihood}}$$

$$P(M|x,y) = \frac{P(x,y|M)P(M)}{P(x,y|M)P(M) + P(x,y|R)P(R)}$$

(Durbin, 1998)

### Finding the Best Model

Codeml allows a range of evolutionary models. Each model represents different distributions of the ratio of non-synonymous to synonymous substitution rates. We looked at four sets of nested models of codon evolution. The Yang models are abbreviated as M1 through M13. We ran M0 and M1, M5 and M6, M7 and M8, and lastly, M7 and M10 using the *gag* sequences (see table 1). The likelihood ratio test can then be used on the two models to determine whether there is statistical support for the use of the more general model (which often included positively selected sites).

likelihood ratio test

$$\chi^2 = 2 \log (ln_1 - ln_2)$$

### Measure of Sequence Variability

To measure amino acid variability at a given codon position we calculated entropy for each column in the protein alignment. Entropy takes into account the number of amino acids that occurred at a codon and their frequency. BioEdit (Hall, 1999) includes a program that calculates entropy. The equation for entropy is:

$$-\sum P_{aa} \log P_{aa}$$

Where the variable  $P_{aa}$  is the proportion of each amino acid in a particular position (Yusim, 2002).

### **Differences from the Reference Strain**

A program was written as part of this project (in the Perl scripting language) that looked at each codon site of all the sequences and compared it to the corresponding site of the reference strain. It counted the number of sequences that differ at each codon as compared to the reference strain.

### **Calculation of correlation coefficient**

The correlation coefficient,  $r$ , is used to determine if a correlation exists between a paired set of observations. The correlation coefficient can be positive or negative and shows the direction of the correlation. The coefficient of determination,  $R$ , is  $r$  squared. It can only have values from one to zero. One represents a perfect correlation (either a positive or negative) and zero representing a lack of correlation.  $R$  reveals the strength of the correlation. The two coefficients are used together to express both direction and strength of the relationship. Statistical significance in reference to correlation is a term used to describe if two values have relationship or if they are independent of each other. Significance of correlation is tested by the  $t$ - test.

### **Pearson Product Moment Correlation**

$$t = \frac{r}{\sqrt{[(1-r^2) / (N-2)]}}$$

Rho,  $r$ , is the correlation coefficient,  $N$  is the number of samples, and the degree of freedom is  $N-2$ . The test consists of looking at the distribution of the probabilities. The  $z$ -score must be determined. The  $z$ -score reveals which values of  $t$  lie outside of those likely to occur by chance (<http://faculty.vassar.edu/lowry/webtext.html>).

### Chapter 3:

#### Results

##### CTL Clustering

The regions of the proposed epitopes clustered along the HIV-1 *gag* sequence, forming distinct islands and valleys. (figure 6).

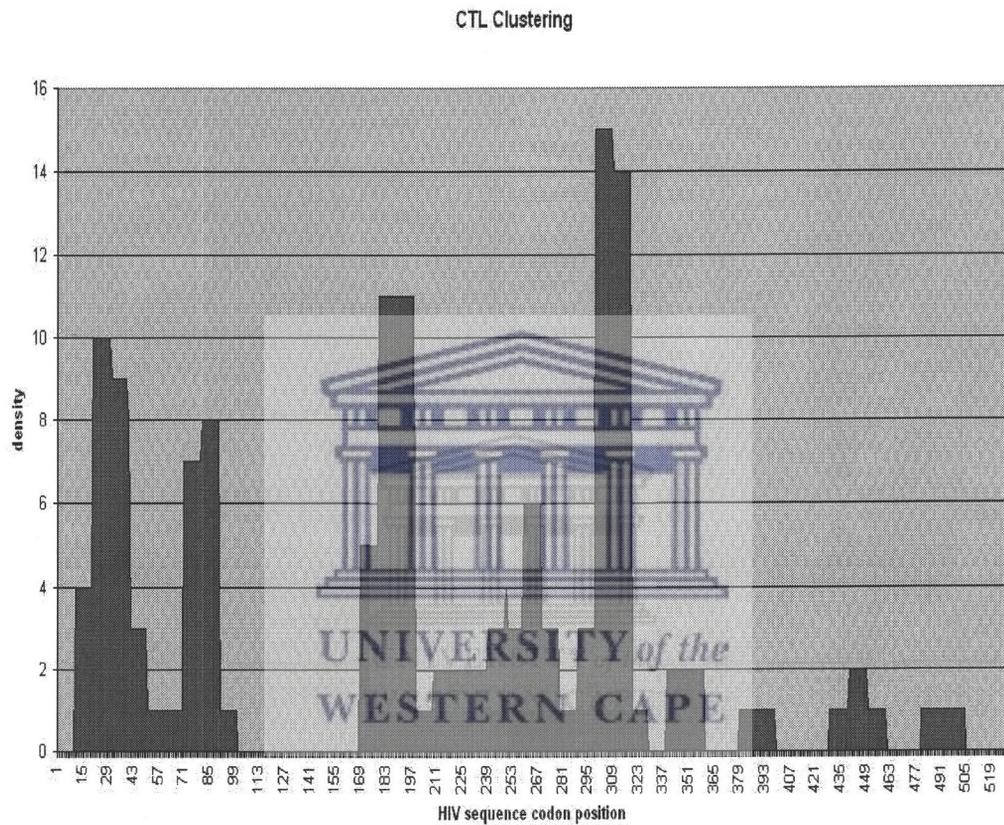


##### Positive Selection

Model 10 of codeml included sites with  $\omega = 3.29$ . A value of  $\omega$  ( $d/s$ ) exceeding one means that the rate nonsynonymous substitutions are significantly greater than the neutral rate. This suggests positive selection. The likelihood of the data was  $-16258.7$  under M10 and  $-16327.28$  under M7. Calculation of the likelihood ratio test revealed significant statistical support for a class of sites to be evolving under positive selective pressure.

Regions of positive selection found by ADAPSITE were negatively correlated with the

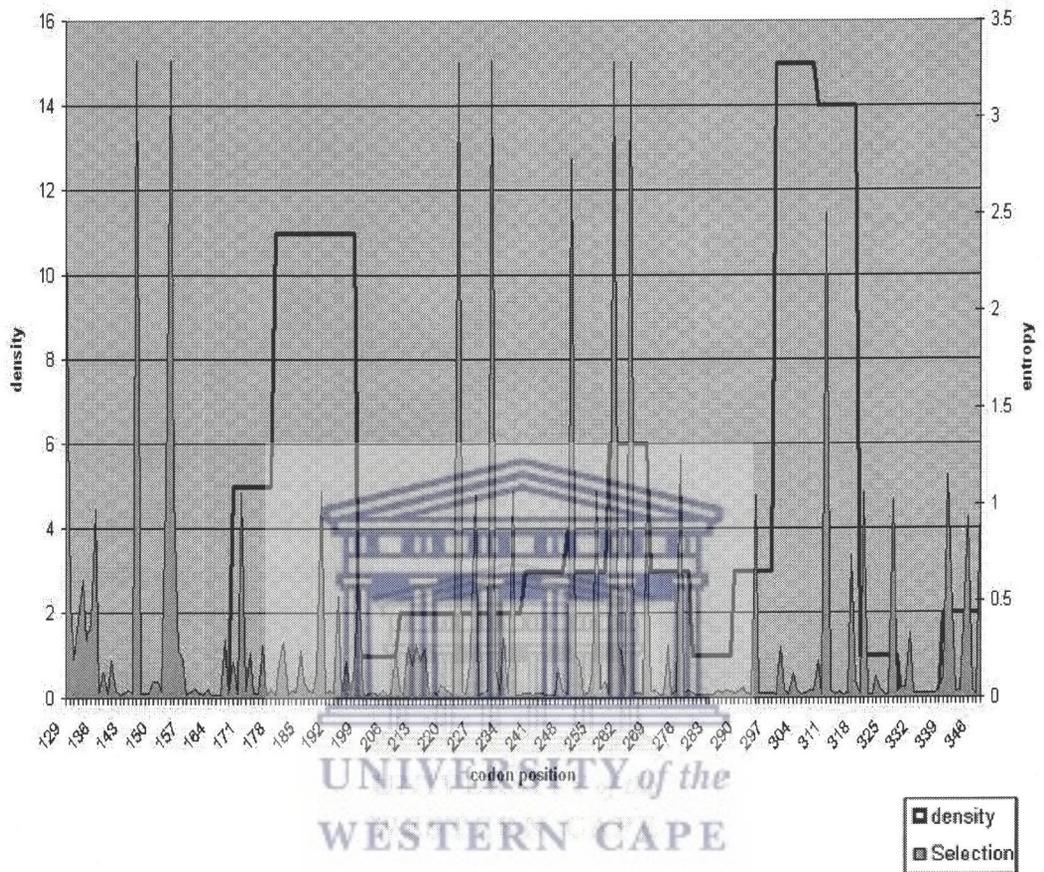
CTL epitope clusters (figure 6 and 7).



**Figure 6:**

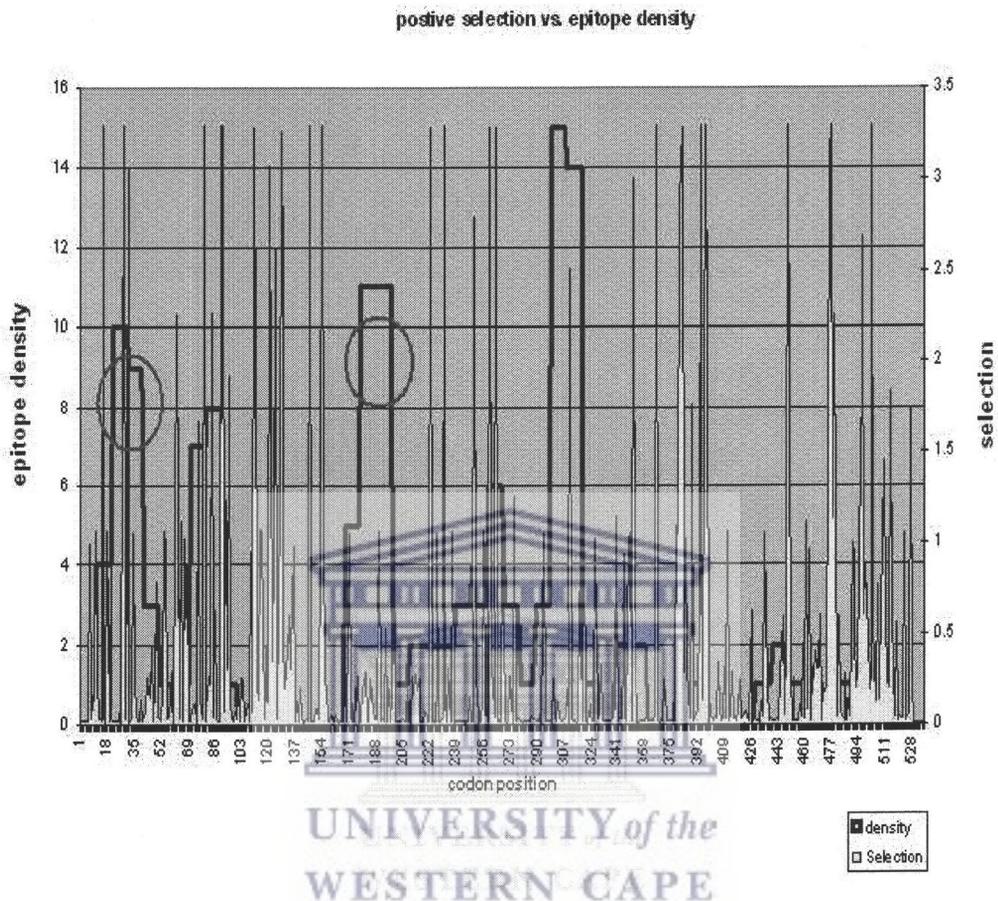
The CTL epitopes which individuals responded to cluster along the HIV sequence.

p24 positively selected sites vs. epitope density



**Figure 7:**

Region p24 of *gag*, positively selected regions vs. regions of CTL epitope response.



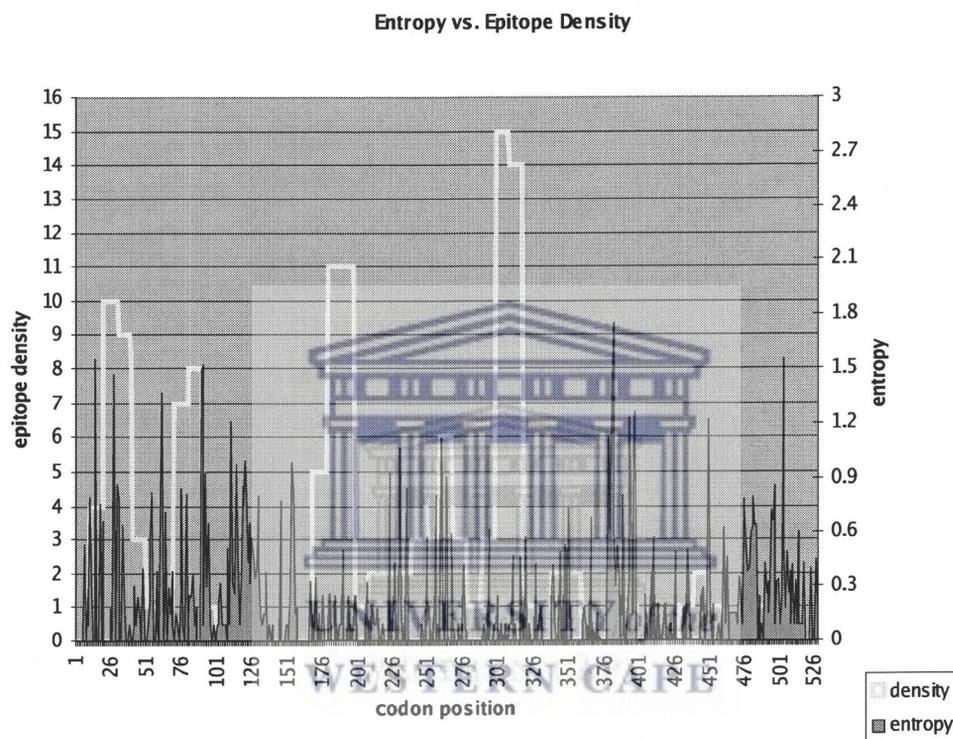
**Figure 8:**

The red circles represent regions of high density and a low level of positive selection; this may be a good vaccine candidate. The green circle emphasizes a region which although has a high level of epitopes also appears to be under strong selection.

### Measure of Variability

A larger value of entropy means a greater degree of variation within the sequences. It is a useful measurement of variability because is not linked to a specific tree topology. For

the 57 *gag* sequences we found higher levels of entropy in regions of lower epitope density (figure 9). This suggests that epitopes reside in evolutionary conserved regions.

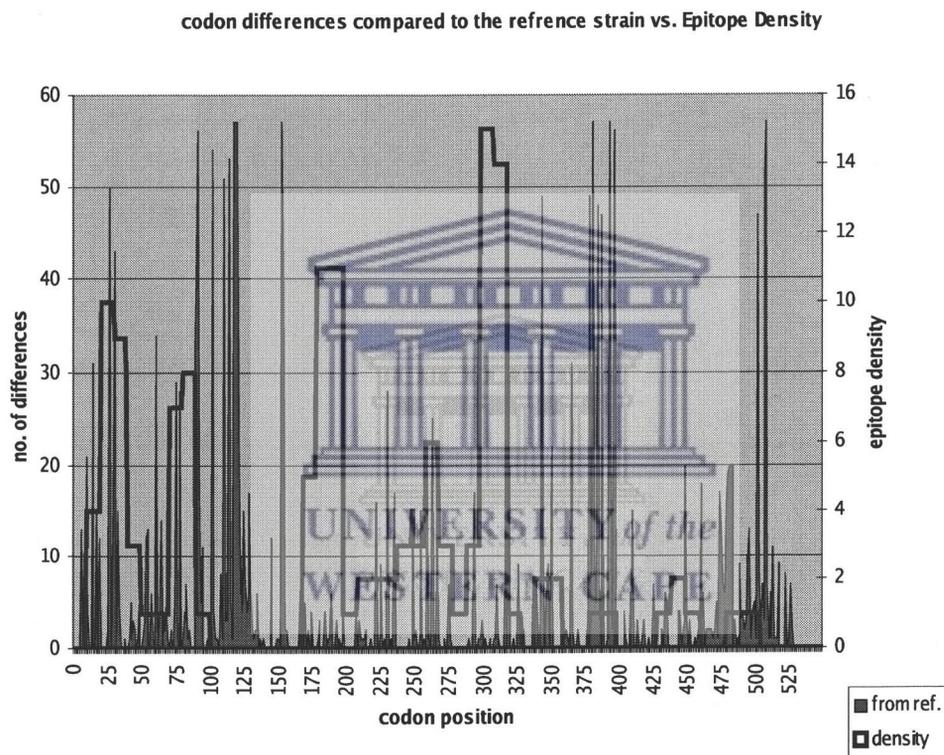


**Figure 9:**  
Entropy compared to regions where individuals responded to CTL epitopes.

### Distance from Reference Strain

We counted the number of sequences at each codon site that differed from the reference strain. There was a slight negative correlation between number of sequences differing from the reference strain and epitope density (figure 10). The correlation coefficient was

-0.09 and the coefficient of determination was 0.009. Divergence between the reference strain and the viral sequences circulating in the patient is a possible explanation for why epitopes have not been found in the regions with high divergence between reference strain and circulating virus.



**Figure 10:** The number of differences between the reference strain and locations of CTL epitope response.

## **Chapter 4:**

### **Discussion**

#### **Correlation between Peptide Density and Positive Selection**

The proposed epitopes form distinct clusters along the reference strain. The locations under positive selection found by ADAPTSITE show a negative correlation with regions of high epitope (the peptides containing the epitopes) density. This suggests that epitope regions are conserved, which is good news for vaccine design. Good candidates for vaccine design are the regions where a large number of individuals showed a CTL response but a low level of selection (such as the region circled in red in figure 8). As opposed to regions where epitope density may be high but selection is also high (the region circled in green in figure 8), which shows the virus is quickly evolving in that region and therefore may not be as effective in a vaccine.

#### **Is CTL clustering an artifact of the Reference Strain?**

Locating epitope regions is key to vaccine design, therefore it is important to know if the clustering of epitopes in conserved regions is because epitopes really do cluster, or is due to a failure to detect CTL responses to the reference strain in less conserved regions.

When looking for epitopes it is not feasible to search every sequence, therefore a reference sequence is chosen as a representative. However, if this reference has diverged significantly from the circulating viral sequences than it fails to be a good archetype. If this is the case, than epitopes may go undetected. Though the correlation between the *gag* sequences and their reference strain is weak, it is significant and suggests that some

epitopes have not been detected due to divergence between the reference strain and the viruses circulating within the patients.

## Conclusion

Genetic and evolutionary distances define HIV subtypes. Though only vaccine trials can truly determine the protective value of an epitope, the study of evolutionary trends and locations of CTL epitopes impacts the development of trial reagents. HIV mutations *in vitro* display different diversification trends than witnessed in infected individuals. *In vivo* the host's immune system exerts selective pressures on the virus, affecting its evolution.

We analyzed HIV-1 subtype C diversity and its relationship to epitope locations. We found a negative correlation between HIV-1 *gag* regions where individuals respond to an epitope and sites under positive selection. This suggests epitopes are evolutionarily conserved. It is possible that epitopes do exist in the regions between the islands of CTL clusters but fail to be detected. Possible causes for failure of detection are genetic distance of the reference strain to currently circulating viruses and sampling error. Locating conserved and advantageous epitopes that are successfully targeted by CTLs is crucial for the development of a HIV vaccine.

This project is a contribution to the HIVNET vaccine prevention trial.

## References

Avert. (2002) [online]. Available [http://www.avert.org/his81\\_86.htm](http://www.avert.org/his81_86.htm) The History of AIDS 1981-1986

*Concepts and Applications of Inferential Statistics* [online]. Available <http://faculty.vassar.edu/lowry/webtext.html>

Durbin, R., Eddy, S. Krogh, A. Mitchison, G. (1998). *Biological Sequence Analysis : Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press.

J. Felsenstein. (1981). 17: 368; Cladistics (1989) 5:164; Olsen, G., Matsuda, H., Hastrom, R., Overbeek, R., (1994). *Comput. Appl. Biosci.* 10:41

FastDNAm1 and DNArates were written by Gary Olsen and colleagues at the Ribosomal Database Project (RDP) at the University of Illinois at Urbana-Champaign. Available <http://geta.life.uiuc.edu/~gary/programs/fastDNAm1.html>

Gashen, B., Taylor, J., Yumim, K., Foley, B., Gao, F., Lang, D., Novitsky, V., Haynes, B., Hahn, B., Bhattacharya, T., Korber, B. (2002). *Diversity Consideration in HIV-1 Vaccine Selection*. *Science*. 296:2354-2360.

Gobbi, A. Stoddart, C., Locatelli, G., Santoro, F., Bare, C., Linqvist-Stepps, V., Moreno, M., Abbey, N., Herndier, B., Malnati, M., McCune, J., Lusso, P. (2000). *Coinfection of SCID-hu Thy/Liv Mice with Human Herpesvirus 6 and Human Immunodeficiency Virus Type 1*. *J. of Virology*. 74:8726-8731.

Goulder, P., Brander, C., Tang, Y., Tremblay, C., Colbert, R., Addo, M., Rosenberg, E., Nguyen, T., Allen, R., Trocha, A., Altfeld, M., He, S., Bunce, M., Funkhouser, R., Pelton, S., Burchett, S., McIntosh, K., Korber, B., Walker, B. (2001) *Evolution and transmission of stable CTL escape mutations in HIV infection*. *Nature*. 412:334-337.

Greenspan, N. and Cera, E. (1999). *Defining epitopes: It's not as easy as it seems*. *Nature Biotechnology*. 17: 936-937.

Hahn, B., Shaw, G., Cock, K., Sharp, P. (2000). *AIDS as a Zoonosis: Scientific and Public Health Implications*. 287: 607-614.

Hall, T.A. (1999). *BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT*. *Nucl. Acids. Symp. Ser.* 41:95-98.

- Hillis, D. (1999). The Evolution of HIV. *Phylogenetics and the Study of HIV*. The Johns Hopkins University Press: 105-121.
- Kuby, J. (1997). *Immunology*. New York: W.H. Freeman and Company.
- Kuiken, C., Foley, B., Guzman, E., Korber, B. (1999). The Evolution of HIV. *Determinants of HIV-1 Protein Evolution*. The Johns Hopkins University Press: 432-465.
- McCutchan, F. (1999). The Evolution of HIV. *Global Diversity in HIV*. The University Press: 41-72.
- McMichael, A., Rowland-Jones, S. (2001) *Cellular immune response to HIV*. Nature.419:980- 987.
- McMichael, A., Klenerman, P., (2002) *HLA Leaves Its Footprints on HIV*. Science.296:1410-1411.
- Moore, C., John, M., James, L., Christiansen, F., Witt, C., Mallal, S. (2002). *Evidence of HIV-1 Adaption to HLA-Restricted Immune Response at a Population Level*. 296: 1439-1442.
- Mount, D. (2001) *Bioinformatics: Sequence and Genome Analysis*. New York: Cold Spring Harbor Laboratory Press.
- Novitsky, V., Rybak, N., McLane, M.F., Gilbert, P., Chigwedere, P., Klein, I., Gaolekwe, S., Chang, Y., Peter, T., Thior, I., Ndung'u, T., Vannberg, F., Foley, B.T., Marlink, R., Lee, T.H., Essex, M. (2001). *Identification of Human Immunodeficiency Virus Type 1 Subtype C Gag-, Tat-, Rev, and Nef-Specific Elispot-Based Cytotoxic T-Lymphocyte Response for AIDS Vaccine Design*. J. of Virology. 75:9210-9228.
- Rowland-Jones, S., Dong, T., Fowke, K., Kimani, J. Krausa, P., Newell, H., Blanchard, T., Ariyoshi, K., Oyugi, J., Ngugi, E., Bwayo, J., MacDonald, K., McMichael, A., Plummer, F. (1998). *Cytotoxic T Cell Response to Multiple Conserved HIV Epitopes in HIV-Resistant Prostitutes in Nairobi*. 108:1758-1765.
- Shankarappa, R.(1999). The Evolution of HIV. *Evolution of HIV-1 Resistance to Antiviral Agents*. The Johns Hopkins University Press: 469-490.
- Suzuki, Y., Gojobori, T. (1999). *A Method for Detecting Positive Selection at Single Amino Acid Sites*. Mol. Biol. Evol. 16:1315-1328

Suzuki, Y., Gojobori, T., Nei, M. (2001). *ADAPTSITE: detecting natural selection at single amino acid sites*. *Bioinformatics*.17:660-661.

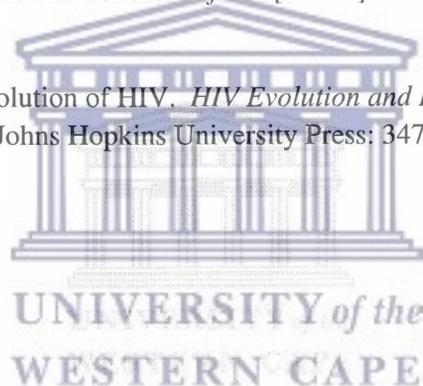
Suzuki, Y. and Nei, M. (2001). *Reliabilities of Parsimony-based and Likelihood-based Methods for Detecting Positive Selection at Single Amino Acid Sites*. *Mol. Biol. Evol.* 18(12):2179-2185.

Yang, Z., Nielsen, R., Goldman, N., Pedersen, A. (2000). *Codon-Substitution Models for Heterogeneous Selection Pressure at Amino Acid Sites*. *Genetics*. 155:431-449.

Yusim, K., Kesmir, C., Gaschen, B., Addo, M., Altfeld, M., Brunak, S. Chigaev, A. Detours, V., Korber, B. (2002). *Clustering Patterns of Cytotoxic T-Lymphocyte Epitopes in Human Immunodeficiency Virus Type 1 (HIV-1) Proteins Reveal Imprints of Immune Evasion on HIV-1 Global Variation*. *J. of Virology*. 76: 8757-8768.

UNAIDS (2002) *Fact Sheet Sub-Saharan Africa* [online]. Available <http://www.unaids.org/>

Viscidi, R. (1999). *The Evolution of HIV. HIV Evolution and Disease Progression via Longitudinal Studies*. The Johns Hopkins University Press: 347-389.





UNIVERSITY *of the*  
WESTERN CAPE