Examination of HIV-1 Diversity and Evolution by A Bioinformatics Approach

By

Binhua Liang

A Thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba

in partial fulfilment of the requirements of the degree of

Doctor of Philosophy

Department of Medical Microbiology

University of Manitoba

Winnipeg

Abstract

HIV-1 genetic diversity is a major obstacle for developing an effective vaccine. My hypothesis is that HIV-1 genetic diversity can be characterized and that cross-clade immunogens can be predicted at the population level.

To address this, I systematically investigated positive selection (PS) pressures on HIV-1 Env and Gag proteins based on the analysis of the sequences collected from the Los Alamos Sequence Database. I identified PS sites, investigated PS patterns, correlated PS with the known functional sites of the two proteins, calculated frequencies of HLA alleles targeting CTL epitopes, and compared PS patterns among major subtypes. The results showed that PS pressure was widely dispersed across the entire regions of both HIV-1 Env and Gag proteins, suggesting the conserved regions are under host immune response pressure. The neutralizing antibody, non-neutralizing antibody, and CTL responses were found to be the major forces driving genetic diversity of HIV-1 *env* and *gag* genes at population level. However, PS pressures on both Env and Gag proteins remain stable over time, suggesting genetic diversity of HIV-1 driven by host immune responses changed very little over the last 29 years. Furthermore, the results also demonstrated that up to 70% PS sites were shared among the major HIV-1 clades**,** implying the existence of cross-clade immunogenicity. A number of potential cross-clades immunogens were predicted to elicit CTL or neutralizing antibody responses from Env and Gag proteins. I also detected a significant correlation between HLA allele frequencies and host CTL responses elicited by Accessory/Regulator's proteins at population level. Moreover, I

detected an association between the frequency of HLA-B7 supertype and the number of identified optimal CTL epitopes. The results suggest HLA class I allele frequencies in a population influence the evolution of HIV-1. I also systematically evaluated the utility of ultra-deep pyrosequencing to characterize genetic diversity of HIV-1 *gag* genes within quasispecies. The results showed that ultra-deep pyrosequencing of amplified HIV genes is a better method than the traditional Sanger-clone-based method in the comprehensive characterization of genetic diversity of HIV-1 quasispecies, especially in detecting low frequency variations.

In conclusion, my thesis provides important information for rational design of an effective HIV-1 vaccine.

Dedication

I would like to give my dedication to my wife, Mrs. Manna Zhang, who have always been there for me and encouraged me. Without her support, I could not even pursue my PhD degree at my age and could not go through the hard time in the PhD program. I believe that she would be very proud of my achievements.

Acknowledgements

I would like to thank so many people who gave me so much help and supports. At first, to my supervisor, Dr. Francis Plummer, I thank him to give me the most valuable opportunity for taking my PhD program and also thank you for all your supports in the past. To my co-supervisor, Dr. Steven Jones, Steven has provided invaluable guidance and advice in my PhD project. Without his support, I would not accomplish my PhD project.

Thanks to the members in HIV research group in Winnipeg. You made our working environment wonderful and fun. Special thanks to Ma Luo and Blake. Ma Luo has always been there and supportive. I have learned about science, critical thoughts, and required skills from you. Your encouragements have helped me go through this journey. To Blake, thank you for all your help in the past, especially your persistent support to solve all of my problems during my PhD study. Your suggestions and insights are very helpful to my PhD program. Most importantly, you have made my graduate study very comfortable, much easier, and wonderful.

To all staff of Medical Microbiology Department, I was very happy to work with you during my graduate study. I would like to give special thanks to Dr. Xi Yang and Dr. Xiaojian Yao, who gave me helpful guidance on my PhD thesis project and advice on PhD program. Also, I cannot forget the moments during lunch and walking time with you in which I really enjoyed and shared with you. I also would like to thank my committee

members, Dr. Gary Van Domselaar and Dr. Brian Fristensky for their review of my thesis and advice which allow me accomplish my goal.

To my friend, Songok, you have provided lots of invaluable advice on my PhD program and your encouragements made me more confident in conducting my PhD project. I also miss the coffee time with you.

I thank the Canadian Institutes for Health Research (CIHR) for funding me to take my PhD program. I finally thank the women of the Pumwani cohorts who were involved in my study.

Table of Contents

Abstract …….. i

Dedication……...…… iii

Acknowledgements……. iv

Table of contents……. vi

List of Tables…….. x

List of Figures……. xii

Publications……..……..... xiii

1.0 Introduction

Overview……. 1
1.1 HIV-1 Discovery……. 3
1.2 HIV-1 Virology……... 4
1.3 HIV-1 Replication Circle…… 6
1.4 HIV-1 Transmission…… 10
1.5 The Pathogenesis of AIDS…….. 12
1.6 AIDS Epidemic……... 15
1.7 Host Immune Responses……. 16
1.7.1 Innate Immune Responses……... 17
1.7.2 Adaptive Immune Responses…….. 19
1.7.2.1 Cell-Medicated Immune Responses…… 19
1.7.2.2 Humoral Immune Responses…….. 23
1.8 HIV-1 Evolution and Diversity……... 26
1.9 Challenge of HIV-1 Vaccine Development…… 32
1.10 Project Goal and General Hypothesis……. 35
1.11 Thesis Outline …….……... 36
Part I Evolution and Characterization of HIV-1 *Env* Gene……..… 36
Part II Evolution and Characterization of HIV-1 *Gag* Gene……..… 37
Part III Impact of HLA Class I Allele Frequencies on CTL Responses…. 37
Part IV Impact of Second-Generation Technologies on Sequence Analysis of HIV-1 Genes…… 38

2.0 Materials and Methods

I. General Materials and Methods

2.1 HIV-1 *Env* Sequences……..... 39
2.2 HIV-1 *Gag* Sequences…… 41
2.3 Optimal CTL Epitopes & Corresponding Sequences……..... 43
2.4 Pumwani Commercial Sex Worker Cohort…… 45
2.5 Phylogenetic Analysis and Tree Building…….. 45
2.6 Statistical Analysis & Perl Scripts…….. 46

II Specific Methods

Part I. Evolution and Characterization of HIV *Env* Gene

2.I.1 Identification of Positive Selection Sites and Calculation of Their Frequencies……. 47
2.I.2 Analysis of Positive Selection…… 47

Part II. Evolution and Characterization of HIV-1 *Gag* gene

2.II.1 Identification of Positive Selection Sites and Calculation of Their Frequencies……. 49
2.II.2 Analysis of Positive Selection…… 49
2.II.3 Statistical Analysis……..... 50

Part III. Impact of HLA Class I Allele Frequencies on CTL Responses

2.III.1 Calculation of Average of HLA Class I Allele Frequencies…….. 51
2.III.2 Calculation of dN/dS Ratio of CTL Epitope Responses……. 52
2.III.3 Statistical Analysis……..... 52

Part IV. Impact of Second-Generation Sequencing Technologies on the Sequence Analysis of HIV-1 Genes

2.IV.1 Subjects……... 53
2.IV.2 PCR Amplification, Cloning, and Sequencing with Sanger Method…… 53
2.IV.3 Ultra-Deep Pyrosequencing with GS20……..... 54
2.IV.4 Sequence Alignment and Measure of Variants……... 54
2.IV.5 Measure of Amino Acid Variability and Determination of Correlation between Amino Acid Variations and Gag Functions……..... 55
2.IV.6 Map Consensus Differences to Gag Functional Sites and Positive Selection Analysis…….. 56
2.IV.7 Statistical Analysis and Perl Scripts……... 57

3.0 Results

Part I. Evolution and Characterization of HIV-1 *Env* Gene

3.I.1 Distribution of Positive Selection Sites in the Population…….. 58
3.I.2 Comparison of Positive Selection Sites across Major Subtypes…… 62
3.I.3 Changes of Positive Selection Pressure over Time……..... 65
3.I.4 Association of Positive Selection Sites with Host Immune Responses and Identification of Epitopes…… 67
3.I.5 Identification of Immunogenic Sequences for HIV-1 Vaccine Design…. 70

Part II. Evolution and Characterization of HIV-1 *Gag* Gene

3.II.1 Identification and Distribution of Positive Selection Sites……. 72
3.II.2 Comparison of PS Sites Identified between Los Alamos and Cohort Gag Sequences…… 76
3.II.3 Comparison of Positive Selection Sites across Major Subtypes……. 78
3.II.4 Association of Positive Selection Sites with Host Immune Responses and Identification of Epitopes……..... 80
3.II.5 Identification of Immunogenic Sequences for HIV-1 Vaccine Design….. 82

Part III. Impact of HLA Class I Allele Frequencies on CTL Responses

3.III.1 HLA Class I Allele Frequencies in the Population……..... 87
3.III.2 dN/dS Ratio of the CTL Epitopes across HIV-1 Proteins……..... 91
3.III.3 The Correlation of HLA Allele Frequency and dN/dS Ratio…..... 95
3.III.4 The impact of HLA Class I Supertypes on the Study of Disease Association……. 98

Part IV. Impact of Second-Generation Sequencing Technologies on the Sequence Analysis of HIV-1 Genes

3.IV.1 Characterization of 454 Pyrosequencing Data…… 101
3.IV.2 Distribution of Genetic Variants……..... 104
3.IV.3 Impact of Variations on the Functional Characterization of HIV-1 Gag Proteins……..... 106
3.IV.4 Impact of Variations on Studying Evolution of HIV-1 Gag……... 113

4.0 Discussion

4.1.0 Evolution and Characterization of HIV-1 *Env* Gene

4.1.1 Discussion……..…. 115
4.1.2 Summary……. 123

4.2.0 Evolution and Characterization of HIV-1 *Gag* Gene

4.2.1 Discussion ……..…..….. 124

4.2.2 Summary……. 129

4.3.0 Impact of HLA Class I Allele Frequencies on CTL responses

4.3.1 Discussion……..……. 130
4.3.2 Summary……. 133

4.4.0 Impact of Second-Generation Sequencing Technologies on the Sequence Analysis of HIV-1 Genes

4.4.1 Discussion……..... 134
4.4.2 Summary……. 138

4.5.0 Final Conclusion……... 139

5.0 References..... 143

6.0 Appendices..... 167

6.1 Appendix A Abbreviations ……..... 167
6.2 Appendix B Perl Scripts…….. 169

List of Tables

Table 1	HIV-1 Env sequence populations……..	40
Table 2	HIV-1 Gag sequence populations…….…….	42
Table 3	Optimal CTL epitopes used in this study…….	44
Table 4	Comparison of positive selection sites between each paired clades….....	63
Table 5	Comparison of the distribution patterns of PS sites among different clades……	64
Table 6	Association between PS and host immune responses……..	69
Table 7	Identified NAb and CTL epitopes on the regions of HIV-1 Env free from PS……...……...……..	71
Table 8	The shared PS sites among different *gag* subtypes..…….....	77
Table 9	Association between PS and host immune responses……..…	79
Table 10	Identified conserved CTL epitopes on HIV-1 Gag proteins…….....	81
Table 11	The match of the PS sites identified from Cohort and Los Alamos *gag* sequences……..	84
Table 12	Distribution of HLA class I alleles across HIV-1 proteins……..	90
Table 13	Association between HLA allele frequency and dN/dS ratio…….....	96
Table 14	HLA class I supertypes used in this study…….……..	99
Table 15	Association between HLA allele frequency of supertypes and dN/dS ratio……..	100
Table 16	The comparison of the detected variants by 454 and Sanger clone-based sequencing methods…….	103
Table 17	Entropy differences on the functional sites of HIV-1 Gag proteins….....	107

Table 18 Consensus differences of HIV-1 Gag proteins between 454 and Sanger cloning sequences overlapped with the functional sites in individuals…… 111

Table 19 The comparison of PS sites identified from 454 and Sanger cloning sequences populations……..... 114

List of Figures

Figure 1	Diagram of HIV-1 particle……..	5
Figure 2	HIV genome map…….	7
Figure 3	PS sites on HIV-1 Env glycoproteins……..	59
Figure 4	Comparison of PS site densities among different regions of gp120….....	61
Figure 5	Frequencies of PS sites on HIV-1 envelope glycoprotein over time…..	66
Figure 6	PS sites on the HIV-1 Gag Polyproteins……...……...	73
Figure 7	Distribution of PS sites on HIV-1 Gag proteins……..	75
Figure 8	Comparison of the identified PS sites from Los Alamos and Cohort HIV-1 Gag sequences……...…...	83
Figure 9	The match of the PS sites between Los Alamos and Cohort *gag* sequences……...…...	86
Figure 10	Comparison of HLA class I allele frequencies on the different restricted CTL epitopes on HIV-1 proteins…….	88
Figure 11	Comparison of dN/dS ratios on the CTL epitopes from different HIV-1 proteins……...……...	92
Figure 12	Comparison of dN/dS ratios on the CTL epitopes restricted by HLA-A and HLA-B alleles …….....	94
Figure 13	The correlation of HLA allele frequency and dN/dS ratio of CTL epitopes on HIV-1 accessory proteins….……	97
Figure 14	Distribution of the variants detected only by pyrosequencing method……..……	105
Figure 15	Correlation between entropy differences and epitope density in HIV-1 Gag……...	109

Publications arising from this thesis:

Liang B, Luo M, Ball TB, Jones SJM, and Plummer FA. Is evolution of HIV-1 toward its ancestor? *In preparation*

Liang B, Luo M, Scott-Herridge J, Semeniuk C, Mendoza M, Capina R, MacArthur I Sheardown B, Ji H, Kimani J, Ball TB, Van Domselaar G, Jones SJM, and Plummer FA. Characterization of Genetic Diversity and Insight into the Functions and Evolution of the HIV-1 *Gag* gene: A Comparison of Parallel Pyrosequencing and Sanger Dideoxy Clone-based Sequencing. *Submitted*

Liang B, Luo M, Ball TB, Jones SJM, and Plummer FA. QUASI Analysis of Host Immune Responses to Gag Polyproteins of Human Immunodeficiency Virus Type I by A Systematic Bioinformatics Approach. *Accepted*

Liang B, Luo M, Ball B, Jones S, Yao X, Domslelaar GV, Cheang M, and Plummer FA, Systematic Analysis of Host Immunological Pressure on the Envelope Gene of Human Immunodeficiency Virus Type 1 by an Immunological Bioinformatics Approach. *Current HIV Research, 2008, 6:284-93*

Liang B, Luo M, Ball B, and Plummer F
QUASI Analysis of the HIV-1 Envelope Sequences in Los Alamos National Laboratory HIV Sequence Database: Pattern and Distribution of PS Sites and their Frequencies over Years. *Biochemistry and Cell Biology, 2007, 85*(2)*:259-264*

1.0 Introduction

Overview

It has been 29 years since the first case of acquired immunodeficiency syndrome (AIDS) was recognized in the United States in 1981 (Gottlieb et al. 1981b) and human immunodeficiency virus type-1 (HIV-1) was discovered by French and American scientists as the causative agent for AIDS (Barre-Sinoussi et al. 1983; Popovic et al. 1984). During this period, approximately 25 million people have lost their lives to this disease, 33.2 million people are currently living with HIV-1, with the new infections occurring at a rate of between 1.8 and 4.1 million/yr (UNAIDS 2007). In addition to its impact on human health, HIV results in a dramatically reduced life expectancy, significant labor loss, and confers a large financial burden to society. Currently, the HIV epidemic continues to be uncontrolled with the global prevalence of HIV infection continuing to remain at the same level, especially in sub-Saharan African countries (UNAIDS 2007). There is neither cure nor effective vaccine for HIV/AIDS. Therefore, the HIV/AIDS pandemic is still causing an impact on society, the economy, and the politics of both the HIV epidemic regions and the rest of the world.

After 29 years of intense scientific research, significant gains have been made in our knowledge of HIV-1 in the areas of epidemiology, treatment and prevention strategies (Cohen et al. 2008; Wainberg and Jeang 2008; Walker and Burton 2008). One of the most outstanding achievements has been the development of antiviral therapies (ART) which have dramatically increased the survival rates for HIV-1 infected people and prevented

mother-to-child transmissions (Connor et al. 1994; Palella et al. 1998). However, more than 90 percent of the world's HIV infected people live in developing countries in which less than 10% of these individuals remain accessible to ART despite the fact of the great effort that was exerted in delivering and scaling up ART by WHO and others (http://www.who.int/hiv/topics/treatment/en/index.html, WHO). The development of a safe and effective HIV vaccine represents an obvious and ideal solution to end the HIV worldwide pandemic—an endeavor which draws much attention from HIV/AIDS researchers (Fauci 2008). To date, all the efforts in developing an HIV vaccine have been unsuccessful-partly due to the rapidly evolving nature of this virus and also to our incomplete understanding of its pathogenesis and the host's immune system's inappropriate response in the presence of HIV infection.

The genetic diversity of HIV-1 is one of the greatest challenges for developing a HIV vaccine (Brander et al. 2006; Gaschen et al. 2002; McBurney and Ross 2008). HIV-1 rapidly mutates and evolves within its hosts, resulting in a heterogeneous population of viruses (called quasispecies). Three groups (M-main, O-outlier, and N-non-M/non-O) of HIV-1 circulate in different geographical areas of the world. Between groups, viral nucleotide sequence differences can reach up to 50%. Within the M group, there are at least 13 clades and the variation between the amino acid sequences which construct structure proteins can be as high as 35% (Gaschen et al. 2002). Thus, vaccines designed to protect against a single virus may not provide protection against others (Brander et al. 2006). The strategies for reducing amino acid sequence differences such as the centralized sequences have been proposed in many gene products such as envelope

(Env), group antigen (Gag), polymerase enzymes (Pol), regulating late gene expression (Rev), negative replication factor (Nef), transactivator (Tat) and even whole proteome. These artificial proteins have shown to elicit strong cellular or humoral immune responses that recognize multiple HIV proteins from different clades (McBurney and Ross 2008). Identifying broadly reactive immunogens (intra clade /or inter clades) and determining the optimal sequences is one of the major goals of HIV research.

My project will characterize HIV-1 diversity, measure HIV-1 evolution, and identify potential cross-clade immunogenic sequences for effective vaccine design.

1.1 HIV-1 discovery

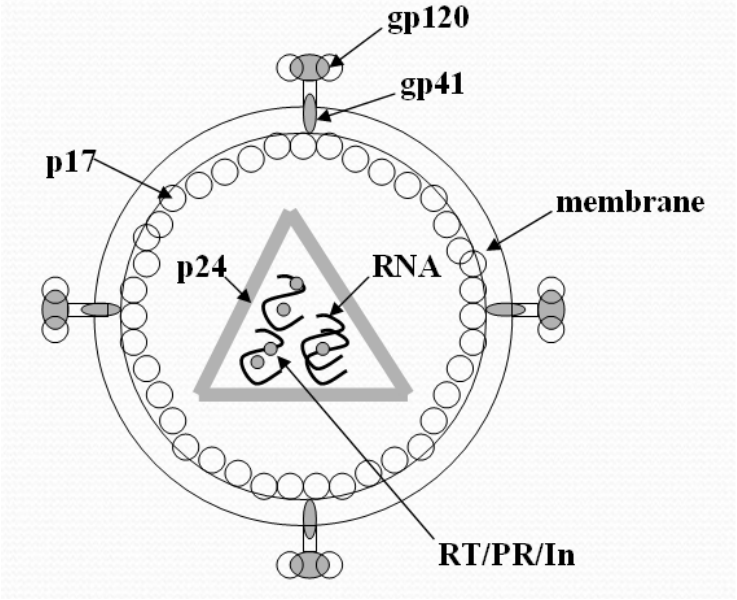
In June, 1981, the Centre for Disease Control (USA), first reported five cases of Pneumocystis carinii pneumonia (PCP) without a definite cause in a group of young people (Gottlieb et al. 1981a). All the patients displayed the signs of immune deficiency such as prolonged fever, lymphopenia, and anergy. But there were no clear causative agents identified. As more cases were discovered in 1982, CDC defined this unknown disease as Acquired Immune Deficiency Syndrome (AIDS) (CDC 1982a). With the accumulation of similar cases in other population groups, including homosexual men, hyperdermic drug users, Haitians, and haemophiliacs, an infectious agent was suspected as the cause of AIDS. This hypothesis was later supported by evidence that a child received blood transfusions and died from AIDS (CDC 1982b). In 1983, a French scientist, Dr. Luc Montagnier first isolated a T-lymphotropic retrovirus from a patient suspected to have AIDS and termed it as lymphadenopathy-associated virus (LAV)

(Barre-Sinoussi et al. 1983). In the following year, Dr. Rober Gallo of the National Cancer Institute also isolated the virus which causes AIDS and named it the human T-cell Leukemia virus III (HTLV-III) (Popovic et al. 1984). One year later, further details of LAV/HTLV-III were released which established that they were the same virus. In 1986, the newer name, Human Immunodeficiency Virus---HIV, was finally proposed by the International Committee on the Taxonomy of Viruses.

1.2 HIV-1 virology

HIV-1 is part of the lentivirus family. The mature HIV-1 virion is enveloped with a lipid bilayer membrane which surrounds a core, consisting of the genomic RNA molecules and proteins required for replication. This spherical morphology is 100-120 nm in diameter (Figure 1). The HIV-1 genome contains two identical strands of positive sense RNA, approximately 9 kb in size. The HIV-1 genome is classified into three major regions: *gag*, *pol*, and *env* which encode for a number of structural proteins and enzymes necessary for HIV-1 replication (Figure 2). Env proteins are produced from a gp160 precursor and form a non-covalent complex of gp120 and gp41 proteins on the cell surface. The gp120-gp41 proteins contain CD4 and chemokine receptor binding sites and mediate the viral attachment to the target cells (Richard Wyatt 2007) Gag proteins are derived from the processing of a 55 kDa precursor by the viral protease. There are four major Gag proteins: P17 (matrix-MA), P24 (capsid-CA), P7 (nucleocapsid-NC), and P6.

Figure 1 Diagram of HIV-1 particle



HIV-1 particle consists of two main parts: bilayer membrane and a core. The genomic RNA molecules and proteins required for replication form the core. Gp120/gp41: envelope protein ; p17: matrix protein; p24: capsid protein; RT/PR/In: reverse transcriptase/protease/integrase proteins.

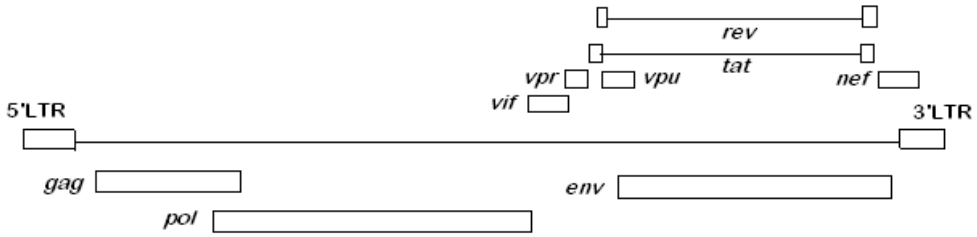
Gag proteins associate with the plasma membrane and play a key role in virus assembly (Gottlinger 2007). *Pol* is expressed as a 160 kDa Gag-Pol fusion protein which is further cleaved to produce three viral enzymes essential for HIV-1 replication, including a protease (PR), a reverse transcriptase (RT), and an integrase (IN). The PR is involved in processing the Gag and Gag-Pol precursor proteins. The RT reverse transcriptase transcribes the single stranded RNA into a double stranded DNA. The IN mediates the process of double stranded DNA integration into a host chromosome.

There are also proteins important in HIV-1 replication, which are termed accessory or auxiliary proteins. In general, accessory proteins are not absolutely necessary for viral propagation but optimize viral replication efficiency. So far, six accessory proteins have been identified. These are the Tat, Rev, the virus infectivity factor (Vif), the enhancing regulatory factor, virus protein R (Vpr), and the virus protein V (Vpu) (Figure 2).

1.3 HIV-1 replication circle

The HIV-1 replication life circle can be summarized by the following stages: virus entry, post-entry, gene expression, virus particle formation, and maturation (Freed 2001) (see "The HIV-1 Life Cycle" at http://www.mayo.edu, Mayo Foundation for Medical Education and Research) . Virus entry initiates with the attachment of its surface proteins, the gp120 and gp41 complex, to the CD4 molecules on host cells (Kwong et al. 1998). This binding leads to conformational changes within the gp41 protein which then induces the binding to host co-receptor molecules (CCR5 or CXCR4). HIV-1 can infect host T-cells/macrophages differentially by binding to CCR5 or CXCR4 receptors.

Figure 2 HIV Genome Map



HIV genome contains nine genes shown as rectangles. LTR: long terminal repeats; *gag*: group antigen; *pol*: polymerase enzyme; *env*: envelope; *vif*: viral infectivity protein; *vpr*: viral protein R; *vpu*: viral protein U; *tat*: transactivator protein; *rev*: regulator of expression of viral protein; *nef*: negative regulator protein.

Two types of HIV-1, R5 and X4, are recognized based on their co-receptor use and cellular tropism. The R5 viruses preferentially attach CCR5 co-receptors which are mainly expressed on macrophages (Alkhatib et al. 1996). The R5 viruses replicate efficiently in macrophage culture. In contrast, the X4 viruses grow well in T-cell culture and utilize CXCR4 co-receptors which are primarily expressed on T-cells.

Membrane fusion takes place between the membranes of the viruses and the host cell after binding to the CD4/co-receptors. Gp41 plays a key role in this fusion event. There is one hydrophobic fusion peptide with two heptad repeats on its ectodomain region. Upon virus binding to the CD4/co-receptor on the host cell, the gp41 fusion peptide is inserted into the host cell membrane and the two heptad repeats pack and fold into a six-helix bundle (Chan et al. 1997; Weissenhorn et al. 1997). Following virus entry into the host cell, the HIV-1 capsid is uncoated and the viral core is released into the cytoplasm. The RNA genome is first converted into double-stranded DNA (called proviral DNA) by the reverse transcriptase (RT). Reverse transcription proceeds in several steps, including DNA/RNA hybrid formation, degradation of the RNA, the first strand transfer, minus-strand DNA synthesis, and plus-strand DNA synthesis (Freed 2001). At this stage, the viruses are susceptible to host defense mechanisms such as the tripartite motif protein 5α (TRIM5α), the apolipoprotein B mRNA-editing enzyme: catalytic polypeptide-like 3G (APOBEC3G), and the methylase enzyme: protein arginine methyltransferase 6 (PRMT6) (Sheehy et al. 2002; Stremlau et al. 2004; Xie et al. 2007). The resulting proviral DNA is flanked by two long-terminal repeats (LTR) and remains associated with other proteins/DNA such as Gag MA, Gag NC, accessory protein Vpr, and a central DNA flap

to form the preintegration complex (PIC) (Bukrinsky et al. 1993; Freed 1998; Gallay et al. 1997; Zennou et al. 2000). With the aid of cellular factors, PIC is then directed to move into the nucleus of the host cell.

Upon PIC entry into the nucleus, IN catalyzes the insertion of the proviral DNA into the host cell by joining the 3'-recessed end (removing several nucleotides from the 3'-termini of both strands of viral DNAs) and the cleaved 3'-end cellular DNA. The integrated provirus is then used as the template for the synthesis of all the RNA related to virus replication. Provirus transcription depends on the host cell RNA polymerase II machinery and is initiated by cellular transcription factors such as Sp1 and NF-κB binding to LTR (Berkhout and Jeang 1992). The accessory protein Tat, an RNA-binding protein, recognizes the trans-activation responsive RNA leader and works with other cellular factors such as elongation factor b complex (cyclin-dependant kinase and cyclin T1) to trans-activate gene expression (Berkhout et al. 1989; Garber et al. 1998). This results in the generation of over 30 viral RNAs which are classified into three categories: unspliced mRNAs (encoding Gag and Gag-Pol precursors), partially spliced mRNAs (encoding Env, Vif, Vpu, and Vpr), and multiply spliced mRNAs (encoding Rev, Tat, and Nef). The transcribed unspliced and partially spliced RNAs are then exported by *Rev* from the nucleus into the cytoplasm for translation. This translocation requires the *Rev* to first bind to the Rev responsive element (RRE) present on both unspliced and partially spliced mRNAs and then interact with the cellular nuclear export machinery in order to deliver viral mRNAs into cytoplasm (Lever and Jeang 2006).

Within the cytoplasm, the transcribed mRNAs are translated into proteins with the aid of the host cellular machinery. The N-terminal of the Gag MA is modified by myristic acid which targets the Gag protein to bind the host cell membrane. At the same time, the Gag NC domain, particularly the two zinc-finger motifs, interact with the packaging signal of the genomic RNAs, resulting in the encapsulation of the RNAs (De Guzman et al. 1998; Zeffman et al. 2000). The assembly of the resulting Gag-lipid and Gag-RNA is then initiated and mediated by the interaction between the C-terminal of CA, p2, and the N-terminal of NC (Freed 1998). In parallel, the Env glycoprotein is first translated into gp160 within the rough endoplasmic reticulum (ER) and is then inserted into the lumen of the ER. In the Golgi, gp160 is cleaved by the host proteases to generate gp120 and gp41 which anchors into the membrane and also remains associated with gp120 to form a non-covalent complex. When the Env complex arrives in the cell surface, it is recruited to form virions by interacting with the Gag MA (Freed 1998). At this stage, the virion is able to pinch off from the host cell membrane. During this process, the "L" domain of p6 (Pro-Thr/Ser-Ala-Pro motif) plays a key role in stimulating virus release (Huang et al. 1995). After the virus particle is released out of the host cell, the Gag/Gag-Pol precursors are cleaved by PR to generate mature Gag and Pol proteins, leading to virion maturation.

1.4 HIV-1 transmission

There are three major transmission routes for HIV-1 infection: sexual transmission (ST), mother-to-child transmission (MTCT), and parenteral transmission (PT). ST happens when the secretions from an HIV-1 infected person come into contact with another's genital, oral, or rectal mucous membranes. ST is the cause for the majority of HIV-1

infections worldwide and its risk reaches 0.008 /per coital act during first five months after exposure (Wawer et al. 2005). HIV-1 transmission mainly depends on the concentration of HIV-1 within body (blood and genital secretions), virus-specific properties, and host susceptibility (Cohen et al. 2008). The risk of HIV-1 transmission is significantly correlated with the viral loads in blood and genital fluid (Chakraborty et al. 2001; Quinn et al. 2000). And any factors increasing the viral loads, especially other sexually transmitted diseases (STDs), also have an impact on HIV-1 transmission. STDs generally cause ulcers and inflammation, leading to the increased shedding and concentration of HIV-1 in the genital tract (Coombs et al. 2003; Ghys et al. 1997). In the cases in which the viral loads are equivalent, HIV-1 variants demonstrate a great variation in infectiousness primarily because of the different targeted chemokine co-receptors. R5 viruses can enhance the rate of ST when hosts are exposed to clade C with the commonly transmitted CCR5 variants (Ping et al. 1999). Human susceptibility to HIV-1 infection is also different due to host genetics factors, innate, and acquired immune responses. Deletion in CCR5 coding sequences (CCR5Δ32) substantially protect an individual from HIV-1 infection (Liu et al. 1996). Likewise, the alteration of mucosa by trauma or inflammation can increase susceptibility to infection with HIV-1 (Halperin 1999; Taha et al. 1998). Furthermore, resistance to HIV-1 infection has been reported in Africa sex workers and shown to be significantly associated with both cell-mediated and mucosal antibody responses (Kaul et al. 1999; Stranford et al. 1999). The application of condoms and circumcision of men can dramatically reduce the risk of ST (Bailey et al. 2007).

MTCT occurs when an HIV-1 infected woman is pregnant, labors, deliveries or breastfeeds her baby. MTCT accounts for over 90% of childhood HIV-1 infections and the chance of the virus being transmitted to the baby is estimated to be around 25% in an HIV-1 positive mother without treatment during pregnancy, labor and delivery. Breastfeeding will add an additional 12% chance of MTCT (Miotti et al. 1999). There are a variety of other factors increasing MTCT risks, including high viral load, low vitamin A, premature rupture of membranes, premature delivery, and exposure to maternal blood (Bulterys and Fowler 2000; Castetbon et al. 2000; Fowler et al. 2000). Antiretroviral therapy provided to pregnant women with HIV-1 can reduce rates of MTCT by two-thirds (Connor et al. 1994). MTCT can be further reduced by the safe, acceptable, feasible, and affordable formula feeding instead of breastfeeding.

PT occurs through infected blood/blood products on open cut and mainly accounts for HIV-1 infection in intravenous drug users (IDUs) and recipients of blood product transfusions. Less frequently, PT can be caused by receiving medical care in the use of the injection equipment in the developing countries. In addition, people receiving tattoos and piercings may also be at risk of PT (Berkley 1991).

1.5 The pathogenesis of AIDS

HIV-1 infection is characterized by the destruction of CD4+ T lymphocytes and the loss of host immune responses with the resultant increased opportunistic infections and malignancies. It can be classified into three stages: (1). primary HIV-1 infection, which is in the first 3-6 weeks after HIV-1 entry into host cells but before seroconversion. During

this stage, viral loads massively increase followed by a decrease to a set-point (usually below 20000 RNA copies/ml). Viral set-point indicates equilibrium between HIV-1 replication and host immune responses and is predictive of long-term prognosis of HIV-1 infection; (2). The chronic asymptomatic HIV-1 infection (also called latent HIV-1 infection), which may last 6 to 10 years with virus level at set-point; (3). The symptomatic HIV-1 infection (or AIDS stage), which is the failure of host immune responses and diseases development in hosts (Centlivre et al. 2007; Levy 2009).

During primary HIV-1 infection, viruses first enter into the mucosa through sexual contacts and cross the mucosal epithelium by at least four different mechanisms, including through breaches in the epithelium, transcytosis by M cells, transcytosis by epithelial cells, and through transportation by DCs (Centlivre et al. 2007). After HIV-1 crosses the epithelial barriers, it starts to primarily infect CD4+ resting memory T-cells there and replicate within them. CCR5 is the major co-receptor for HIV-1 entry into CD4+ T-cells and is very important for establishing infection during this early HIV-1 infection (Veazey et al. 1998). The infected CD4+ T-cells transmit the infection to the neighboring uninfected cells, including the activated CD4+ T-cells, through virologic synapse formation to maintain a continued viral transmission (Jolly et al. 2004). Compared to the resting CD4+ T-cells, the activated CD4+ T-cells can generate more viruses and thus disseminate the infection more rapidly (Zhang et al. 2004). At the same time, HIV-1 binds to DC-specific surface molecule, DC-SIGN. The DC-SIGN bound virus is capable of infecting CD4+ T-cells and also allows virus to survive outside of host cell until it migrates to draining lymph nodes (Geijtenbeek et al. 2000). From there, HIV-

1 is eventually and systematically disseminated to other parts of human body. Through a direct cytopathic effect on infected CD4+ T-cells and an indirect induction of apoptosis of uninfected CD4+ T-cells by HIV-1 gp120 Fas ligand, HIV infection causes a progressive loss of CD4+ T-cells (Centlivre et al. 2007). In the meantime, HIV-1 replication increases as more immune cells get infected and the numbers of infectious virions grow exponentially. The viral load can reach up to $1x10^8$ copies RNA/ml of blood (Lyles et al. 2000). On the other hand, HIV-1 infection induces host immune responses. Two kinds of cytotoxic effector cells: NK cells (the innate immune system) and CD8+ T-cells (the adaptive immune system) are mainly involved in the control of HIV-1 replication (Alter et al. 2007). The numbers of NK cells expand during the first 5 days of infection before CD8+ T-cells emerge and then decline as CD8+ T-cells start to dominate. NK cells can directly lyse the HIV-1 infected cells or produce antiviral cytokines such as INF-γ during the initial phase of infection. Anti-HIV-1 CD8+ T-cells arise between 20 to 25 days after infection, right after the peak of viraemia (Reynolds et al. 2005). Robust virus-specific CD8+ T-cells responses then develop to contain the massive HIV-1 replication and thus bring down the viral load (Koup et al. 1994; Kuroda et al. 1999). As a consequence, the viraemic peak falls until a "steady state" of viral load, called "viral set-point", is reached.

Following primary HIV-1 infection, the strong immune responses to HIV are induced, especially specific cell-mediated immunity, leading to control of viral replication at set-point for many years until progression to AIDS. However, the number of CD4+ T-cells decline slowly over time in a linear manner. The loss of CD4+ T-cells is believed to

come from either the increased death of CD4+ T-cell or reduced production of CD4+ T-cells or both (Weber 2001). Aside from the loss of CD4+ T-cells, their functions are also impaired, including defective proliferation and diminished secretion of IL-2 (Harari and Pantaleo 2008). During this period, the patients remain "healthy" but still infectious. When CD4+ T-cell numbers fall below 200 cells/μl, a variety of opportunistic infection and tumors will then develop due to the loss of the host cell-mediated immunity. At this stage, patients are diagnosed as AIDS and usually die within five years without treatment (Redfield et al. 1986).

1.6 AIDS epidemic

It is estimated that 33.2 million people are living with HIV infection, and 25 million people have already died from HIV/AIDS (UNAIDS 2007). Each day, more than six thousand of people are infected by HIV-1 and around the same number of people die from AIDS. Heterosexual transmission accounts for over 80% of all HIV-1 infection and thus is the main mode of HIV transmission in the world. The distribution of HIV infection is worldwide but uneven. Sub-Saharan Africa continues to be the region where most HIV-1 transmission occur (68%) and contains the majority of HIV-1 infected people (76%). Currently, the global prevalence of HIV infection remains at the level of 0.8% what it was in 2001 though there is a reduction from 5.8% to 5.0% in Sub-Saharan Africa. In North America, especially USA, most of HIV infected are men, mainly Africa American men who have sex with men-- MSM (Schneider E 2006). In Western Europe, the HIV-1 epidemic is primarily concentrated among the populations who have sex with men, inject drugs, and have multiple sexual partners. The targeted countries by HIV-1

infection in this region include Spain, France, and the United Kingdom. In Eastern Europe, only the Russian Federation and Ukraine are heavily affected and injection drug use is the dominant mode of transmission. In Latin America, over 30% of HIV infected people are from Brazil where MSM accounts for the most of transmission. In South and Southeast, HIV/AIDS are most distributed in India and the People's Republic of China. In this region, HIV-1 infection is concentrated in injecting drug users, female commercial sex workers/or their clients, and MSM (UNAIDS 2007).

Women currently make up 50% of the total HIV infected population. Around 61% of the infected women live in Sub-Saharan Africa. However, there is a trend that the proportions of women living with HIV are growing slowly in America, Asia, and Eastern Europe. At the same time, the total number of children (less than 15 years old) living with HIV is increasing, reaching 2.5 million in 2007. Most of HIV infected children live in Sub-Saharan Africa (UNAIDS 2007).

1.7 Host immune responses

HIV-1 infection induces host immune responses against HIV-1. Generally speaking, the host immune response consists of two arms: innate and adaptive. The innate immune system is the first line of defense in response to HIV-1 infection. The major differences between these two arms are the absence of reorganization of specific antigens and the lack of memory for the innate immune system.

1.7.1 The innate immune response

The innate immune response can limit the transmission and replication in the early stage of HIV-1 infection and activate the adaptive immune responses to mount a longer and effective protective response. There are three categories of the innate immune system: cellular, extracellular and intracellular (Lehner et al. 2008). The cellular components mainly include dendritic cells (DCs), Natural Killer (NK) cells, γð-T cells, and macrophages. All of them can detect HIV-1 by either recognizing single-strand RNA from HIV-1 through the use of their "pattern recognition" receptors 7/8 (Toll-like receptors-TLR 7/8) on DCs/macrophages or by interacting with HLA I on the HIV-1 infected cells with the killer immunoglobulin-like receptor (KIR) on NK/ γð-T cells (Colonna et al. 1997; Heil et al. 2004; Levy 2001; Martin et al. 2000). Both NK cells and γð-T cells are able to kill HIV-1 infected cells through cytolytic mechanisms (Biron and Brossay 2001; Selin et al. 2001).

Macrophages and DCs do not directly have a killing action but produce a large amount of extracellular factors, especially type I IFN (α and β), which can interfere with viral growth as well as activating T cells (Marrack et al. 1999; Rogge et al. 1998). Upon activation, the CC chemokines, RANTES, and the chemokine macrophage inflammatory protein 1 (MIP 1α and MIP 1β) are also induced. These extracellular factors are believed to be the important bridges that link the innate and adaptive immune systems (Lehner et al. 2008). In addition, these CC chemokines were shown to prevent HIV-1 infection by down-modulating the cell-surface expression of CCR5 (Wang et al. 1999).

Recently, intrinsic intracellular anti-viral factors were discovered as potential anti-HIV immune defense molecules. Such factors primarily comprise apolipoprotein B mRNA edition enzyme--catalytic peptide-like 3G (APOBEC3G), 3F (APOBEC3F), and tripartite motif 5 alpha (TRIM-5α). APOBEC3G belongs to the family of cytidine deaminases which is able to deaminate cytosine bases (converting cytosine to uridine) (Harris et al. 2002). In the absence of HIV-1 Vif, APOBEC3G in the cytoplasm of HIV-1 infected cells is packaged into newly generated HIV-1 virions and inhibit their replication in host cells. The inhibitory actions include: (1), APOBEC3G directly binds to HIV-1 RNA and physically block RT moving along the template, leading to the inhibition of reverse transcription; (2). APOBEC3G deaminates cytosine residues on the newly synthesized minus ssDNA and thus impairs DNA synthesis; (3). APOBEC3G results in the formation of the aberrant viral DNA, which affects the integration of the dsDNA into host cell genomes. However, these actions can be counteracted by HIV-1 Vif (Chiu and Greene 2009). TRIM-5α was first identified to inhibit HIV-1 replication in rhesus monkey cells (Stremlau et al. 2004). It belongs to the tripartite motif family which consists of 70 genes in humans. TRIM-5α contains a PRY/SPRY domain which recognizes HIV-1 capsids and is capable of recruiting TRIM-5α into the newly generated virion. The associated TRIM-5α-virus complex can either restrict the reverse transcription in host cells (Wu et al. 2006) or be directed to the proteasome for degradation (Campbell et al. 2008). Although, the rhesus macaque TRIM-5α presents a strong inhibition of HIV-1 infectivity, the human one does not work well with only weak inhibition.

Despite these innate immune responses playing an important role in control of HIV-1 replication and limitation of HIV-1 dissemination at the early stage, protection against virus infections is predominantly medicated by adaptive immunity, which is more specific, stronger, and lasts longer (Harari and Pantaleo 2008).

1.7.2 Adaptive immune responses

The adaptive immune system can also be divided into two arms: cell-mediated and humoral responses. These two arms control the infected cells or virions that escape the innate immune responses. Cell-mediated immune responses are activated in the first week post-infection and play a key role in control of chronic HIV-1 infection and viral set-point. In contrast, the humoral immune responses are induced several weeks later and are critical in preventing HIV-1 infection with poor effect on controlling HIV-1 infection (Harari and Pantaleo 2008).

1.7.2.1 Cell-mediated immune responses

Cellular immunity against HIV-1 is primarily mediated by CD8+ cytotoxic T-lymphocytes (CTL) with the aid of CD4+ T-helper cells. CTLs are generally primed by antigen-presenting cells (APCs) and target HIV infected cells by recognizing virus/ antigens complexed with class I MHC molecules, through T-cell receptors (TCRs) with co-stimulatory "confirmation" signals. After the interaction, CTLs then lyse the HIV infected cells by using granule-dependant or ligand-induced pathways (Lichterfeld et al. 2004). CTLs also secret a range of soluble non-cytolytic factors such as IFN-γ, TNF-α, CC chemokines, RANTES, MIP1α/MIPIβ, IL-2, and other soluble factors, which

migrate to the infection sites and inhibit HIV replication without killing infected cells. These suppressors are capable of suppressing diverse HIV-1 strains (DeVico and Gallo 2004).

HIV-1 antigen-specific CTL responses are significantly associated with suppression of viral load during primary infection (Alter et al. 2007). This is supported by the evidence that depletion of CD8+ T-cells in SIV-infected monkeys results in the loss of control of SIV replication, leading to high viral load (Schmitz et al. 2005). Furthermore, vigorous HIV-1 antigen-specific CTL responses were found in both Long-Term Non-Progressors (LTNPs) and people who remain uninfected after repeated exposure to HIV-1(Kaul et al. 2004; Schmitz et al. 2005), suggesting the critical role of CTLs in providing sterilizing immunity to HIV-1. However, it was also found that the functions of CD8+ T-cells were impaired during HIV-1 chronic infection. These impaired functions include both poor/or lack of cytolytic activity and suppressor activity as well as lack of IL-2 production and proliferation activity (Appay et al. 2000; Wherry et al. 2003; Zhang et al. 2003). In contrast, more functional CD8+ T-cells are found in LTNPs as well as in other virus infection in which viruses are eliminated or controlled (Harari et al. 2006; Schmitz et al. 2005). HIV-1 specific antigens are suspected to be the driving force for the loss of CD8+ T-cell functions (Wherry et al. 2003).

The presentation of HIV-1 epitopes is in the context of HLA class I alleles. Thus, CTL epitopes are restricted by HLA I alleles. These epitopes are under strong HLA-restricted immune selection pressure and evolves dynamically. As a consequence, escape mutations

in CTL epitopes inevitably develop during chronic infection. Mutations may abrogate the binding of CTL epitopes and HLA I molecules, resulting in the reduced TCR recognition and viral escape from CTL responses (McMichael and Rowland-Jones 2001). In fact, the mutations disrupting processing of CTL epitopes such as proteasome cleavage sites also affect the presentation of antigens by HLA I molecules and thus impair CTL activity (Del Val et al. 1991). The development of CTL escape mutations is one of the major factors responsible for the failure of control of HIV-1 replication. There is accelerated progression to AIDS when CTL escape mutations accumulate (Brown et al. 2005). This observation is supported by an observed reverse association that has been found between CTL escape mutations and viral load at the population level (Brumme et al. 2008). Recently, studies showed that only the CTL responses against Gag protein (but no Env or accessory proteins) are significantly correlated with lower level of viraemia (Kiepiela et al. 2007). This observation suggests that effective CTL responses only target specific regions of HIV-1 although many regions of HIV-1 elicit CTL immune responses.

CD8+ T-cell activities are strongly associated with HLA alleles and different HLA alleles can change the rate of AIDS progression. HLA-B27* and HLA-B*5701 were reported to contribute to CTL responses more than others and show protection against AIDS progression (Altfeld et al. 2006). HLA-B*27 and HLA-B*5701 mediate immunodominant CTL responses against two highly conserved peptides, termed KK10, and TW10, to confer protection (Altfeld et al. 2003; Goulder et al. 1997). In contrast, the HLA-B*35 is correlated with susceptibility to developing AIDS. Moreover, dominant

influence of HLA-B on HIV disease outcome was observed (Photinl kiepiela 2004). It seems that HLA-B alleles evolve faster than HLA-A alleles.

CD4+ T-cells play a key role in control of both arms of adaptive immune responses as well as in regulating innate immunity. CD4+ T-cells are important for the production of neutralizing antibodies in clearing viruses and critical for the maintenance of CTL functions in controlling viral replication (Janssen et al. 2003; Planz et al. 1997; Shedlock and Shen 2003; Sun and Bevan 2003). HIV-1 specific CD4+ T-cell responses can be detected in the majority of HIV-1 infected patients (Pitcher et al. 1999). CD4+ T-cell proliferative responses are significantly correlated with lower viral load and effective control of HIV-1 replication during both primary and chronic infections (Harari et al. 2002; McNeil et al. 2001; Younes et al. 2003). HIV-1 specific CD4+ T-cell responses are primarily elicited by epitopes found in the Gag and Nef proteins in chronic infection (Kaufmann et al. 2004).

During chronic infection, the number of CD4+ T-cells decline and their proliferative responses become weaker or absent due to lack of IL-2 (proliferation capacity). Comparison of CD4+ T-cell proliferative responses between LTNPs and subjects with progressive disease demonstrated that more frequent proliferative responses were detected in LTNPs (Harari et al. 2005; Tilton et al. 2007). Moreover, enhanced CD4+ T-cell proliferative responses were also found in patients receiving prolonged highly active antiretroviral therapy (HAART) (McNeil et al. 2001). These observations suggest functional defects of HIV-1 specific CD4+ T-cells in HIV-1 infection. On the other hand,

a switch of CD4+ T helper 1 (Th1) to CD4+ T helper 2 (Th2) has also been proposed: Th1 cell activities dominate at early stage whereas Th2 cell activities become prominent at late stage of HIV-1 infection (Clerici et al. 1993; Clerici and Shearer 1993). Th2 cells can produce cytokines such as IL-4, IL-5, and IL-10 which allow HIV-1 replicate more efficiently, whereas Th1 cells do not. Thus, the change of Th1 to Th2 is believed to be a critical step in the etiology of AIDS during chronic HIV infection.

Since HIV-1 preferentially infects HIV-1 specific CD4+ T-cells, the circulating HIV-1 specific CD4+ T-cells (producing cytokines) are at low frequency (around 5%) (Douek et al. 2002). The majorities of circulating HIV-1 specific CD4+ T-cells undergo apoptosis through Fas-mediated cascase-9 activation or by priming by HIV-1 gp120 envelope proteins (Ostrowski et al. 2006). HIV-1 specific memory CD4+ T-cells (CD4+ Tcm) are more susceptible to Fas-mediated apoptosis than either effector CD4+ T-cells (CD4+ Tem) or naïve T-cells during all stages of HIV infection (Douek et al. 2002). As a consequence, there may not be adequate HIV-1 CD4+ T-cells to "help" mount effective CTL and antibody responses against HIV-1 during chronic infection.

1.7.2.2 Humoral immune responses

The antibody responses against HIV envelope proteins gp120 and gp41 can be detected a few weeks after infection (Aasa-Chapman et al. 2004; Pilgrim et al. 1997; Richman et al. 2003). Thus, the presence of HIV-1 antibodies is an important diagnostic index for HIV-1 infection. Antibody responses to HIV can be further divided into two categories: neutralizing (NAb) and non-neutralizing (non-NAb).

In HIV infection, NAbs only account for a small portion of the total antibodies identified so far and are considered to be more efficient than non-NAbs in terms of fighting HIV infection. NAbs inhibit viral infection by either interfering with viral binding to CD4/co-receptors, or interrupting the fusion process through postreceptor engagement (Trkola et al. 1996; Wu et al. 1996; Wyatt and Sodroski 1998). This enables NAbs to preferentially target free virus particles but may not cell-cell spread (Bachmann and Zinkernagel 1997; Pantaleo et al. 1995; Zinkernagel et al. 2001). A number of NAbs with potent broad cross-neutralizing activities have been identified to target HIV envelope proteins, including IgG1b12, 2G12, 2F5, and 4E10 (Burton et al. 1994; Purtscher et al. 1994; Scanlan et al. 2002; Stiegler et al. 2001). Neutralizing antibody IgG1b12 and 2G12 recognize and interact with CD4-binding sites and high mannose-epitopes on gp120, respectively, whereas 2F5 and 4E10 identify and react with the membrane proximal external region (MPER) on gp41. These antibodies can provide protection from lentiviral challenge in non-human primates or delayed rebound following discontinued antiviral therapy after passive NAbs transfer (Trkola et al. 2005; Zwick et al. 2005). Higher titers of NAbs were also detected in LTNPs but not found in patients with progressive infection (Carotenuto et al. 1998).

However, as described before, many studies showed that a decrease in viral load was observed in the presence of cellular immunity but not NAbs in primary infection, suggesting that neutralizing antibody responses provide no protection at early stage of HIV-1 infection. In addition, HIV has developed several means to escape neutralizing antibody responses. Reduced accessibility to antibody-binding sites, the glycan shield on

the viral surface and the increased HIV diversity are comprehensively documented, and are considered to be responsible for the poor and narrowed neutralizing responses against HIV-1 primary isolates. To date, all attempted antibody-based vaccination strategies failed in eliciting protective neutralizing antibody responses against HIV-1 infection in humans.

Non-NAbs are abundant in HIV-1 infection and may outweigh NAbs in activating effector functions (Huber and Trkola 2007). Currently, the role of Non-Nabs is not clear in HIV-1 infection. It is proposed that Non-NAbs may contribute to virus elimination though the lysis of infected cells by activating killer cells (antibody-dependant cellular cytotoxicity, ADCC), lysis of viruses by complement, and through phagocytosis. ADCC can be triggered by the interaction between the Fc region of Non_NAbs and corresponding receptors on the effector cells (especially NK cells) to destroy the infected cells. The titers of ADCC are higher than NAbs responses and ADCC are relatively broad against different HIV-1 strains (Banks et al. 2002; Forthal et al. 2001). A correlation between ADCC and disease progression has been inferred (Baum et al. 1996; Sawyer et al. 1990).

The non-NAb induced complement lysis responses can be found in both primary and chronic infection with more responses in the latter (Huber et al. 2006; Spear et al. 2001). These responses have been shown to inhibit viral replication. It suggests that the complement lysis responses induced by non-NAbs are lower than ADCC responses in HIV-1 infection. Complement mediated lysis can either be down-regulated by

complement factor H (CFH) or interfered with by complement regulatory proteins CD46, CD55, and CD59 in HIV-1 infection (Pinter et al. 1995; Saifuddin et al. 1997; Schmitz et al. 1995).

1.8 HIV-1 evolution and diversity

The origin of HIV-1 was tracked to a simian virus: Simian Immuno-deficiency Virus (SIV) from both chimpanzees and sooty mangabeys (Hahn et al. 2000). These viruses are passed from chimpanzees to human through blood-borne transmission. It is estimated that HIV-1 group M was first introduced into human population around 1931 and rapidly evolved (Korber et al. 2000). The ongoing evolution of HIV-1 leads to substantial genetic diversity among virus isolates.

In vivo, it was estimated that the rate of point mutations of HIV-1 is 3.4 x 10^{-5}/per generation (Mansky 1996). In other words, on average 3-4 out of 10 daughter genomes contain a new mutation compared to the parental genome. HIV-1 employs two mechanisms to generate sequence diversity. One is the introduction of mutations into the genome during viral cDNA synthesis by error-prone reverse transcriptase (RT) since RT lacks DNA proof-reading activity (Bebenek et al. 1989). The other is the recombination, which occurs between two or more different viruses infecting the same cell in the same host (Groenink et al. 1992). In addition, the rate of mutations and generations in host also affect the rate of HIV-1 evolution. In practice, the rate of HIV-1 evolution can be estimated by phylogenic analysis of nucleotide sequences. The estimated rates of HIV-1 evolution are 0.24% (Env) and 0.19% substitutions per base pair per year (Gag) (Korber

et al. 2000). As a consequence, numbers of non-identical but closely related viral genomes, called "quasispecies", accumulate within the host.

Currently, the circulating viral strains are classified into three distinct groups (M, N, and O) based on their genetic differences. HIV-1 group M is the predominant circulating HIV-1 group and is responsible for the current pandemic. Phylogenetic analysis of group M viruses has further identified nine subtypes of HIV-1: A-D, F-H, J, K and over 20 circulating recombinant forms (CRFs). The genetic variability of these subtypes/or recombinant has a great effect on transmission and disease progression. For example, subtype C was reported to be transmitted from mother-to-child more frequently than subtype B (Renjifo et al. 2004). A higher mortality rate was found among people infected by subtype D viruses compared to those infected by other subtypes such as subtype C and B (Baeten et al. 2007; Kaleebu et al. 2002). The subtypes and recombinant forms of HIV-1 are globally dispersed and its distribution shows the complexity of the molecular epidemiology of HIV-1. The most prevalent HIV-1 subtypes are subtype C, A, B, D, G, CRF02_AG, and CRF01_AE (Geretti 2006; Taylor et al. 2008). Subtype C continues to be predominantly distributed in Eastern Africa, Southern Africa and India, which accounts for 50% of HIV-1 infection worldwide. Subtype A is primarily distributed in Eastern Africa, Central Africa, Eastern Europe, and Central Asia and is responsible for 12.3% of HIV-1 infections in the world. Subtype B is responsible for 10.2% of HIV-1 infections worldwide and is predominant in the Western Europe, America, Australia, and East Asia. The others such as subtype G (6.3%), D (2.5%), AG (4.8%), and AE (4.7%) are mainly concentrated in a specific geographic area: West Africa, Eastern Africa,

Southeast Asia, and West Africa, respectively (Taylor et al. 2008). The distribution of HIV-1 subtypes and recombinant forms results in a great genetic diversity of HIV-1 in the world.

HIV-1 genetic diversity seems to display a common pattern within a host during HIV-1 infection. At the earlier stage of HIV-1 infection, the diversity is very low (Grobler et al. 2004). After seroconversion, it starts to rise at a linear rate in untreated patients. When the point that CD4+ T-cells decline dramatically is reached, the diversity remains stable at a high level (Shankarappa et al. 1999). This pattern implies that the viruses are governed by their fitness in hosts. In fact, humans exert a great selective pressure on HIV-1 during infection. During primary infection, R5 viruses (using CCR5 chemokine receptors) predominate in the viral population. While X4 viruses usually appear in chronic infection. The transition of R5 to X4 has been shown to be associated with the progression of diseases (Schuitemaker et al. 1992). It was found that the mutations at V3 loop positions 11, 25, 24, and 27 sufficiently converted R5 to X4 viruses (de Jong et al. 1992; Fouchier et al. 1992; Milich et al. 1993).

HIV-1 diversity is also shaped by transmission between hosts. In comparing sequence variants between the recently infected recipients and transmitters, minor variants were found in recipients. The effect of transmission was found in detecting the variants in the gp120 hypervariable V3 loop region of HIV-1 Env protein, where fewer variants were detected in V3 than V4-5 during early infection, whereas above-average diversity was

shown compared to the other regions in chronic infection (Zhu et al. 1993). This transmission effect on HIV-1 diversity is called the "transmission bottleneck".

Adaptive immune responses are another factor, which are considered to exert the strongest pressure on the evolution of HIV-1 across almost all stages of infection. Neutralizing antibody and CTL responses are believed to be the main forces driving viral evolution by selecting variants. Cell-mediated immunity forces viruses to mutate in order not to be recognized by CTLs (CTL escape mutations). CTL escape mutants emerge within a month after primary infection and become dominant rapidly (Borrow et al. 1997). The appearance of CTL escape mutants has been significantly associated with the rapid progression to AIDS (Goulder et al. 1997). CTL escape mutations tend to be maintained after transmission if the new host shares the same HLA allele. Otherwise, escape mutations may revert to wild-type sequences based on their fitness cost (Leslie et al. 2004). In contrast, neutralization escape mutations appear slowly within chronic infection. It was shown that the escape mutations emerged about 15 months after the detection of neutralizing antibody responses in chimpanzees (Nyambi et al. 1997).

Escape mutations were reported to be across Env V1 to V5 loops (Burns and Desrosiers 1994). Evidence for neutralization escape mutations were also observed in HIV-1 infected patients (Albert et al. 1990; Arendrup et al. 1992; Lathey et al. 1997; Mahalanabis et al. 2009; Nakowitsch et al. 2005; Shibata et al. 2007). The emergence of neutralization escape mutations suggests that humoral immune selective pressure is

relative weak compared to cell-mediated immune selective pressure during HIV-1 infection.

Highly Active Antiretroviral Therapy (HAART) has been widely used for the treatment of HIV infected patients since 1996 and added extra pressure on viruses. The known six classes of antiretroviral (ARV) drugs approved by the US Food and Drug Administration (FDA) were all shown to select for drug resistance mutations. More than 200 mutations have been identified to be associated with antiretroviral resistance to drugs (Johnson et al. 2008). The majority of them are distributed in RT and PR proteins. Drug resistance mutations, especially multi-drug resistance mutations, were also shown to be transmitted in a population (Yerly et al. 2004). Furthermore, the appearance of resistance mutations to antiretroviral drugs may result in second mutations to compensate the lost fitness (Boyer et al. 1998; Nijhuis et al. 1999; Peters et al. 2001).

Determining the selection pressure exerted by host immunity on viral mutations is a major focus of evolutionary studies of HIV-1 (Frost et al. 2005; Moore et al. 2002; Ross and Rodrigo 2002; Wei et al. 2003); (Bonhoeffer et al. 1995; Dulioust et al. 1999; Leslie et al. 2005; Mohabatkar and Kar 2004). Many coding regions of HIV-1 are reported to be under positive selection (adaptive selection) pressures (Chen et al. 2004; Frost et al. 2001; Price et al. 1997; Travers et al. 2005; Yamaguchi-Kabata and Gojobori 2000). Positive selection could be exerted by neutralizing antibodies (Richman et al. 2003), T-helper lymphocytes (Ross and Rodrigo 2002) and cytotoxic T lymphocyte (CTL) responses (Choisy et al. 2004; da Silva and Hughes 1999), all of which involve Major

Histocompatibility Complex (MHC) class I and class II molecules (Moore et al. 2002) and are influenced by antigen processing through proteasomal cleavage (Zimbwa et al. 2006).

Detecting the presence of adaptive selection within protein-coding genes has traditionally been carried out by comparing synonymous and non-synonymous nucleotide substitution (dN/dS ratio: ω) with statistical analysis. Many models and parameter estimation methods have been used to detect PS sites (Chen et al. 2004; Choisy et al. 2004; de Oliveira et al. 2004; Huelsenbeck et al. 2001; Pond et al. 2006; Stewart et al. 2001; Suzuki and Gojobori 1999; Travers et al. 2005; Yamaguchi-Kabata et al. 2004), including Bayesian (Huelsenbeck et al. 2001), maximum likelihood (ML) (Yang 1997), and maximum parsimony (Suzuki and Gojobori 1999). Currently, maximum likelihood is the most commonly used method. It predicts PS sites according to the model of codon substitutions. Codons are subdivided into categories with varied ω which can be estimated by ML (Yang 1997). This method can predict PS at amino acid sites, and thus avoids consistent substitution rates among amino acid positions. ML method has been implemented in HYPHY package for use online (Pond and Frost 2005). The disadvantage of this method includes the practical limit on the number of sequences that can be analyzed, complexity of application, and time-consuming analysis. Alternatively, a codon-based method, QUASI, can be used to overcome these disadvantages (Liang et al. 2008; Stewart et al. 2001). QUASI also compares dN/dS ratio within a specific codon with a statistical analysis, in which, "empirical dN/dS ratio is compared to neutral dN/dS ratio by means of a two-side binomial distribution" (Stewart et al. 2001). The null

hypothesis that all mutations on that codon are equal will be rejected at a significant level if non-synonymous substitutions are overabundant. Thus, QUASI analysis is independent of phylogeny, which is properly applied to viral sequences whose ancestor is not clear or hard to approximate. QUASI can rapidly predict PS sites with thousands of sequences analyzed within seconds. The drawback of QUASI is the high false positive rate caused by ignoring phylogeny. However, longer sequences can dramatically reduce false positive rate (up to 2%). In our previous studies, QUASI has been applied to identify PS sites and shown to generate robust and reliable results (Liang et al. 2007; Liang et al. 2008; Peters et al. 2008).

1.9 Challenge of HIV-1 vaccine development

The development of an effective HIV-1 vaccine remains the priority for AIDS researchers. The optimal HIV-1 vaccine should elicit both strong neutralizing antibody and effective CTL immune responses. All the vaccines attempted so far failed to prevent HIV-1 infection. Most importantly, the current failure of the most promising vaccine candidate developed by Merck Research Laboratories cools our hope of success and questions our ability to develop an effective vaccine. The Merck candidate vaccine was designed using *gag*, *pol*, and *nef* genes from Clade B. It was unable to generate protective immunity against heterologous viruses in human (Cohen 2007).

One of the major challenges for the development of an HIV-1 vaccine is sequence diversity. For group M, nine subtypes and forty-three CRFs circulate in the different geographic regions of the world. Within the same subtype of viruses, genetic variation

can be up to 20%, whereas between subtypes, variation reaches 35% (Hemelaar et al. 2006). Thus, a vaccine based on a single isolate seems impossible to protect against a wide population of viral isolates. It poses a grand challenge for HIV-1 vaccine development. Secondly, HIV-1 infects critical immune cells, especially CD4+ T-cells, which are required for maintenance of vaccine specific immune response in the host. The loss of critical immune cells results in impaired effective immune responses against HIV-1. Thirdly, HIV-1 develops strategies to escape host immune responses. One of strategies is the masking of neutralization epitopes in gp120 by the dense glycosylation, which prevent antibodies from recognizing epitopes on the Env protein (Srivastava et al. 2005). In addition, mutations to escape CTL and neutralizing antibody epitopes are emerging constantly, resulting in escape of HIV-1 from host humoral and cell-mediated immune responses. Last, HIV-1 quickly establishes a latent reservoir by incorporating its DNA into the hosts' chromosomes, leading to an persistent infection for life (Chun et al. 1997).

Almost all vaccine efforts have been made to induce neutralizing antibodies (targeting the Env protein of HIV-1) or/and T-cell-mediated immune responses (target CTL epitopes). However, broadly neutralizing antibodies are proven to be very difficult to induce through immunization as the previously developed vaccines do not mimic the three-dimension trimetric structure of Env and cannot overcome the genetic diversity of Env (Srivastava et al. 2005). Novel approaches have been proposed to design Env immunogens, including native trimers as immunogens, targeting conserved conformational epitopes, redirecting responses away from variable loops (Johnston and Fauci 2007). Currently, these approaches are under investigation but have not achieved

much progress on gp120 alone. The diversity of Env in HIV-1 has also been addressed by two vaccine strategies (McBurney and Ross 2008). The first strategy is to centralize sequences in order to minimize the genetic distance between or within subtypes of the infecting viruses. The proposed three methods are ancestral, center of the tree, and consensus. Another approach is polyvalent vaccine which incorporates multiple components of antigens to have better coverage of the viral population. Although both strategies were shown to be immunogenic and generate a wider breadth of neutralizing antibody responses relative to monovalent vaccines (Chakrabarti et al. 2005; Weaver et al. 2006), the use of centralized and polyvalent vaccines either have not been completely tested or only protect against homologous viral challenges, respectively (Cho et al. 2001; Pal et al. 2006; Weaver et al. 2006).

The development of T-cell based vaccines also faces challenges. These vaccines, designed to help control infection, would likely not produce sterilizing immunity or clear HIV reservoirs. Rhesus macaques (vaccinated by *env*/*gag* plasmid DNA) were first challenged by SHIV-89.6P. The vaccinated monkeys developed strong virus-specific CTL responses and the viral replication was controlled without disease progression (Reimann et al. 1996). More studies generated similar results in nonhuman primates (Amara et al. 2001; Letvin et al. 2006). However, in the first human trial of T-cell based vaccine, STEP, there is no evidence showing the reduced infection rate or viraemia but a higher incidence of HIV-1 infection (Corey et al. 2009), In their studies, it was believed that both the magnitude and/or breadth of CTL responses were insufficient to control HIV-1 replication. The mechanisms are unknown and are currently under investigation. It

is expected that the vaccine does induce strong cell-mediated immune responses but the infecting viruses are no longer homologous to the vaccine. Thus, the concept of a T-cell based vaccine does not fail but instead it highlights the need for developing effective vaccines. To overcome the limitations, polyvalent protein cocktails were introduced to provide maximum coverage of potential T-cell epitopes (Fischer et al. 2007). This strategy has been tested in mice. It showed a significant increase in breadth of T-cell responses (Kong et al. 2009). Currently, several mosaic-based vaccines are in the development pipeline.

In summary, there is a consensus that HIV vaccines to major subtypes are a research priority and necessary to move into studies on nonhuman primates. Centralized and polyvalent vaccines as novel concepts may potentially overcome genetic diversity and move further into human clinical trials.

1.10 Project Goal and General Hypothesis

After the failure of two current promising vaccine candidates aimed at inducing neutralizing antibody or cellular immune responses in human, HIV research has shifted towards more basic scientific discovery to overcome the major underlying obstacles. As one of the greatest challenges for developing an effective HIV-1 vaccine is the genetic diversity of HIV-1, more efforts are needed to address viral sequence diversity in vaccine design. One of the major questions asked by HIV researchers is:"Are there any common sequences or immunogens among the diverse global circulating viruses that could be identified and targeted for vaccine design?" In another words, can the challenge of HIV -

1 sequence diversity be overcome by generating cross-clade protective immunity against HIV-1 infection? Supporting data to answer this question is sparse and confined to small-scale case studies or regional tests. There has been no systematic study addressing the issue of genetic diversity of HIV-1 at a global scale. We hypothesized that **HIV-1 sequence diversity can be characterized to identify regions relevant to immune selections and that potential cross-clade immunogenic sequences could be identified**. In this study, I analyzed the global sequences of two major structural genes of HIV-1, which are major targets of vaccine development to represent common circulating subtypes and cover pandemic geographic regions in the world. Using a bioinformatics approach and statistical analysis, I systematically investigated the molecular evolution and diversity of HIV-1 and correlated it with host immune responses. I have characterized host-virus interaction and identified the possible candidates for potential cross-clade immunogenic sequences for effective vaccine design.

1.11 Thesis Outline

The thesis includes the following four parts. Each part leads with a specific hypothesis, followed by data addressing the corresponding hypothesis.

Part I: Evolution and Characterization of HIV-1 *Env* Gene

Hypothesis: The molecular evolution of HIV-1 *env* gene can be measured at a global scale and this knowledge provides targets relevant to the development of an effective vaccine.

In this part, I analyzed all available full HIV-1 *env* sequences collected in the Los Alamos HIV database and determined the positive selection pressure on HIV-1 *env* at a population level. I focused on an overview of host immune responses to HIV-1 *Env* in the context of positive selection. This work was published in the journal of "Current HIV Research" in 2008 (Liang et al. 2008).

Part II: Evolution and Characterization of HIV-1 *Gag* Gene

Hypothesis: The molecular evolution of HIV-1 *gag* gene can also be measured at a global level. The knowledge derived from study would also allow us to understand the host immune responses shaping sequence diversity of HIV-1 Gag and help identify potential cross-clade immunogenic sequences for candidate vaccine design.

In this part, I analyzed all available full HIV-1 *gag* sequences collected in the Los Alamos HIV database and determined the positive selection pressure on HIV-1 Gag on a global scale. As I did with *env*, I also focused on the overview of host immune responses on HIV-1 Gag in the context of positive selection and compared it to the study of viral populations in the Pumwani Sex Workers Cohort.

Part III: Impact of HLA Class I Allele Frequencies on CTL responses

Hypothesis: There is a relationship between HLA class I allele frequencies, CTL responses and HIV-1 diversity at the population level. HLA class I allele frequencies shape HIV-1 diversity at the population level. By studying host-virus interaction and HLA class I allele frequencies at the population level, we should be able to predict viral

diversity and evolution. The results can provide important information for HIV-1 vaccine design.

In this study, I determined the relationship between average HLA class I allele frequencies and the optimal CTL epitopes identified and correlated them with the estimated immune pressure exerted by the host. This study investigated how HLA class I alleles restrict CTL responses at the population level, which is very important for universal HIV-1 T-cell based vaccine design. This is the first study of its kind.

Part IV: Impact of Second-Generation Sequencing Technologies on the Sequence Analysis of HIV-1 Genes

Hypothesis: Second-generation sequencing technologies will influence and improve the study of HIV function and evolution.

Here, I compared sequences generated by the 454 pyrosequencing technology and those generated by the clone-based Sanger sequencing method and assessed the possible impact of the sequence differences from two methods on the analysis and functional characterization of HIV-1 genes.

2.0 Materials and Methods

I. General Materials and Methods

2.1 HIV-1 *Env* Sequences

1100 nucleotide sequences of full-length HIV-1 *env* were obtained from the HIV Sequence Database (Los Alamos National Laboratory). Each sequence is from a different individual. We assumed that each submitted sequence is correct and represented the dominant HIV-1 variant in the given subject. Sequences containing stretches of unspecified nucleotides were removed from the alignments. A total of 1100 sequences were used for the study. The nucleotide sequences were converted to amino acids and aligned with ClustalW (Higgins et al. 1996; Thompson et al. 1994), and manually edited using "Molecular Evolutionary Genetics Analysis" software, version 3.1 (MEGA 3.1) (Kumar et al. 2004). Data subsets were retrieved as an alignment of sequences from the total aligned sequences (Table 1). Clade A, B, C, and D of HIV-1 were classified based on the classification of HIV Sequence Database (Los Alamos National Laboratory) and were further confirmed by using "Inter-subtype Recombination Analysis-RIP" tool (http://www.hiv.lanl.gov, Los Alomas National Lab) and by phylogenetic analysis. The aligned amino acids were reversely translated back to nucleotide sequences and then were converted into codon-format for QUASI analysis (Stewart et al. 2001) and in-frame stop codons were excluded from the alignments using Perl scripts. The deduced amino acids, including the conserved regions C1-5, variable regions V1-5 and gp41 were numbered according to the reference sequence, B.FR.83.HXB2_K03455 (HXB2).

Table 1 HIV-1 *Env* sequence populations[a]

Population	Lineage	#[b] of Sequences
Main pool	HIV-1 All[c]	1100
Clade A	HIV-1 M-A	96
Clade B	HIV-1 M-B	617
Clade C	HIV-1 M-C	206
Clade D	HIV-1 M-D	60
Others	HIV-1 others[d]	121
1990-1994	HIV-1 All	142
1995-1999	HIV-1 All	322
2000-2004	HIV-1 All	165
1985-1989	HIV-1 M-B	90
1990-1994	HIV-1 M-B	81
1995-1999	HIV-1 M-B	98

[a] All sequences used as a population in this study are derived from the main pool as an alignment of *env* nucleotide sequences. The sequences pool was provided by Los Alamos Laboratory (LANL). Main pool and Clade B were further classified into sub-populations based on five-year period of time for investigating the changes of PS over time.
[b] the number;
[c] HIV-1 All represents the combinations of M (main), O (outlier), and N (non-M/ non-O) groups of HIV-1;
[d] All of M, O and N except M: A, B, C, D

2.2 HIV-1 *Gag* Sequences

A total of 702 full-length HIV-1 *gag* nucleotide sequences were obtained from the HIV Sequence Database (Los Alamos National Laboratory, USA). Only one sequence was used for each individual in this study. The sequences were further divided into subtypes such as clade A (47), B (198), C (404), and D (24) based on the classification of HIV Sequence Database for the corresponding studies (Table 2). The aligned nucleotide sequences were converted into codon-format for QUASI analysis and known in-frame stop codons were excluded from the alignments by Perl scripts (see Appendix B). The sequences were numbered according to HXB2 and analyzed by the same methods described before.

Table 2 HIV-1 *Gag* sequence populations

Clade	Sequence[a]	Clade	Sequence[b]
A	47	A1	311
B	198	C	37
C	404	D	92
D	24	Others	28
Others	28		
Total:	702	Total:	486

[a] HIV-1 *Gag* sequences were obtained from HIV Sequence Database, Los Alamos National Laboratory, USA;
[b] HIV-1 G*ag* sequences were obtained from Pumwani Commercial Sex Worker Cohort, Kenya.

2.3 Optimal CTL Epitopes & Corresponding Sequences

182 optimal CTL epitopes and the corresponding HLA class I types were obtained from HIV Molecular Immunology 2005 (Nicole Frahm 2005). These CTL epitopes were selected based on the following criteria: (1). these epitopes are restricted by HLA class I A, B, and C; (2). CTLs have been shown to recognize the naturally processed epitopes; (3). Titration assays with truncated peptides show the shortest epitopes at the lowest peptide concentrations; (4). HLA class I molecules have to be identified and well defined for these CTL epitopes. These CTL epitopes are located across the entire HIV-1 genes, including *gag, protease*, *RT*, *integrase, vif, vpr, tat, rev, vpu, gp160*, and *nef*. A total of 9089 sequence fragments which contain these epitopes were also obtained from the HIV Sequence Database (Los Alamos National Laboratory, USA) (Table 3).

Table 3 Optimal CTL epitopes[a] used in this study

Location of CTLs[b]	# of CTL epitopes	HLA types[c]	# of sequences[d]
Gag	58	A, B, C	614
Gp160	28	A, B, C	796
Integrase	9	A, B, C	715
Nef	28	A, B, C	1290
Protease	4	A, B	715
Rev	2	A, B	739
RT	33	A, B	715
Tat	4	A, B, C	642
Vif	8	A, B	1015
Vpr	7	A, B	883
Vpu	1	A	965
Total	182	A, B, C	9089

[a] Optimal CTL epitopes were obtained from HIV Immunology Database 2005 (Los Alamos National Laboratory, USA);
[b] Locations of CTL eiptopes are referred to B.FR.83.HXB2_K03455(HXB2);
[c] HLA types were referred to the annotations of optimal CTL epitopes in HIV Immunology Database 2005;
[d] The corresponding sequences containing optimal CTL epitopes were obtained from HIV Sequence Database (Los Alamos National Laboratory, USA);

2.4 Pumwani Commercial Sex Worker Cohorts

The Cohort was established in the Pumwani district of Nairobi, Kenya in 1985 by a group of researchers from the University of Manitoba lead by Dr. Frank Plummer. More than 2000 commercial sex workers (CSW) have been enrolled in the Cohort for the last 29 years and over 800 remain visiting clinic regularly. Each enrolled women is assigned an ML number and received free health care including treatment of sexually transmitted disease (STD). Women are asked to visit clinic twice a year and biological samples including blood and cervical samples are taken at visit time. All personal and clinical information of participants is collected by the working staff and input into an online HIV database for research. Since 2004, all HIV positive women have received highly active antiviral therapy through the US President's Emergency Plan for AIDS Relief (PEPFAR) program.

2.5 Phylogenetic Analysis and Tree Building

Phylogenetic analyses were performed with MEGA 3.1 program based on the "distance method". The substitution mode "Kimura" was used for estimating distances and transition/transversion ratio was integrated. Neighbor-joining trees were constructed with a bootstrap test of inferred phylogeny at 1000 replications. A bootstrap over 70 was considered reliable.

2.6 Statistical Analysis & Perl Scripts

Simple 2x2 Chi-Squares tables were applied to test the association between the numbers of PS sites and various clade combinations of HIV-1 Env and Gag. P-values were adjusted to account for multiple testing. The relationship between various classes of immune responses (CTL, NAb, Ab, T-cell helper--Th, and proteasomal cleavage site--PCS) and numbers of PS sites was explored using contingency table analyses (PROC CATMOD, SAS). Various models were used. A model using only the main effects was used to focus on more detailed investigations of relevant interactions. All other tests were performed with either GraphPad Prism 4, or SPSS 16. Perl scripts developed in-house used in this study are available in Appendix B.

II. Specific Methods

Part I: Evolution and Characterization of HIV-1 *Env* Gene

2. I. 1 Identification of Positive Selection Sites and Calculation of Their Frequencies

Positive selection sites were identified in each data set using QUASI (Stewart et al. 2001) and confirmed based on the method as previously described (Chen et al. 2004). The identified positive selection sites derived from QUASI and dN/dS analysis were subsequently mapped along Env by Perl scripts and MEGA 3.1. Multiple selected amino acids on the same location of HXB2 were counted as a single positive selection at that amino acid. Positive selection sites on regions or sub-regions of data sets were numbered.

The frequency of positive selection sites in each region or sub-region of the data sets was calculated as the percentage of the number of positive selection sites of the studied region over the total number of amino acids in that region of the reference sequence (HXB2).

2. I. 2 Analysis of Positive Selection

To investigate the changes of PS pressure on the entire Env sequences at a population level over time, the aligned sequence pool and data subsets of clade B were divided into 3 groups by five-year intervals. Sequences without annotations of year were excluded. QUASI analysis was conducted on each group and frequencies of identified PS sites in each group were calculated and plotted over time. In order to determine the distribution pattern of PS sites among different subtypes, we performed QUASI analysis on clade A, B, C, D representing the major subtypes circulating worldwide. The locations of PS sites identified from each data set were mapped. The number of PS site at the same location

was determined between clades A, B, C, D or shared between at least three clades. The distribution pattern of PS sites was determined by comparing the average percentage of the same PS sites identified as above.

To illustrate the relationship between the PS and the host immune responses, the locations of known neutralizing antibody, cytotoxic-T lymphocyte (CTL), and T-helper (Th) epitopes (http://hiv-web.lanl.gov, Los Alomas National Lab) were collected. Neutralizing antibody (NAb) epitopes are defined as epitopes that are able to neutralize lab strains or at least some primary isolates of HIV-1. The CTL epitopes used in the study are the optimal CTL epitopes identified in HIV-1 clade B and non-clade B infections. All chosen epitopes in this study have been identified in human subjects (Nicole Frahm 2005). Proteasomal cleavage sites (PCS) were predicted by NetChop 3.0 (http://www.cbs.dtu.dk/, Center for Biological Sequence Analysis). The number of PS sites on the epitope regions or regions without epitopes were determined and analyzed to identify associations between the PS and the host immune responses. The HLA class I alleles were matched with the identified epitopes according to previous reports (Stephens 2005).

Part II: Evolution and Characterization of HIV-1 *Gag* Gene

2. II. 1 Identification of Positive Selection Sites and Calculation of Their Frequencies

As describe in **2.I.1**.

2. II. 2 Analysis of Positive Selection

We first compared the distribution of PS sites determined from the *gag* sequences of ML Cohort and those of Los Alamos HIV Sequence Database. Referring to HXB2, the *gag* gene was divided into p17, p24, p7, and p6 gene regions in which the percentages of PS sites were calculated and compared to ones calculated from the positive selection study of ML cohort. We also matched the PS sites determined from the *gag* sequence of Los Alamos Sequence Database to ones which were identified from the *gag* sequence of ML cohort and were shown to be correlated to HLA class I allele and disease progression. In order to determine the distribution pattern of PS sites among different subtypes at global scale, we performed QUASI analysis on clade A1 and D, which are the major circulating subtypes of HIV-1 in ML cohort, of the *gag* sequences of Los Alamos Sequence Database and ML cohort. The locations of PS sites identified from each data set were mapped. The number of PS sites at the same location was determined between clade A and D. The distribution pattern of PS sites was determined by comparing the average percentage of the same PS sites identified as described in **2.I.1**.

To illustrate the relationship between PS and host immune response, we utilized the same approaches as described in **2.I.1**.

2. II. 3 Statistical Analysis

50 random data sets of 30 PS sites were generated from the identified PS sites from QUASI analysis on the *gag* sequences of Los Alamos Sequence Database. The 30 PS sites of each random data set were matched to the PS sites identified from the previous study in ML cohort. The possibility of the match of PS sites in a range of 1 to 30 is calculated by the total numbers of the match of the corresponding PS sites (1 to 30) over 50.

Part III: Impact of HLA Class I Allele Frequencies on CTL responses

2. III. 1 Calculation of Average HLA Class I Allele Frequencies

HLA class I allele frequencies at population level in different regions of the world were obtained from Anthropology/Allele Frequencies, NCBI (http://www.ncbi.nlm.nih.gov/gv/mhc/main.fcgi? Cmd=init). To estimate an overall specific HLA class I allele frequency (OHF) restricting the corresponding optimal CTL epitope at a global scale, the HLA class I allele frequencies from different populations were summed based on the number of the corresponding CTL epitope sequences used in this study as equation (1):

$$OHF = \sum (n_1f_1 + n_2f_2 + \ldots\ldots + n_nf_n) / (n_1 + n_2 + \ldots\ldots n_n) \qquad \textbf{equation 1}$$

Where: *n* is the number of the optimal CTL epitope sequences from a specific population; *f* is the HLA class I allele frequency in a specific population

Sequence populations were classified according to the definition of worldwide populations (Middleton et al. 2003) and the number of the optimal CTL epitopes sequences from a specific population were then determined. Assuming the same impacts on restricting CTL epitopes, an average HLA class I allele frequency (AHF) was calculated on a corresponding epitope as equation (2):

$$AHF = (\ \sum(n_1f_1 + n_2f_2 + \ldots\ldots + n_nf_n) - \sum(R_1 + R_2 + \ldots\ldots + R_n)) / N \qquad \textbf{equation 2}$$

Where: *n* is the number of amino acids restricted by a specific HLA class I allele; *f* is the HLA class I allele frequency in a specific population; R: the product of the products of

any two HLA class I allele frequencies and the sum of their numbers of amino acids restricted by these two HLA class I alleles; N: the amino acid length of a specific optimal CTL epitope.

2. III. 2 Calculation of dN/dS ratio of CTL epitope sequences

Sequence fragments were obtained from 9098 virus sequences in HIV Sequence Database (Los Alamos National Laboratory, USA). The sequences overlap with 182 optimal CTL epitopes and were divided into data sets with each of which containing a determined CTL epitope restricted by a specific HLA class I allele. The classification of sequence fragments was referred to the annotation of sequences in HIV Sequence Database and the definition of optimal CTL epitopes in HIV Immunology Database 2007 (Los Alamos National Laboratory, USA). Each aligned data set of sequence fragments were then calculated for synonymous (dS) and non-synonymous distances (dN) using the Kumar method in MEGA 3.0. DN/dS ratio was thus determined for each CTL epitope restricted by a specific HLA class I allele at population level.

2. III. 3 Statistical Analyses

The statistical analysis was stratified by different factors, including HLA class I type (A, B, and C), HIV-1 protein type (Gag, Pol, Gp160, and Accessory proteins), and the number of HLA allele restrictions for the CTL epitope (single or multiple HLA alleles). The correlation between AHF and dN/dS ratio was assessed by Pearson Correlation Coefficients implemented in SAS v.9.1. The comparison of dN/dS ratios restricted by

HLA-A and HLA-B alleles was conducted by Student's *t* test implemented in GraphPad Prism 4.0.

Part IV: Impact of 2nd-Generation Sequencing Technologies on the Sequence Analysis of HIV-1 Genes

2. IV. 1 Subject

The study population for Part IV includes antiretroviral treatment-naïve HIV-1 positive women enrolled in the Pumwani Sex Worker cohort in Nairobi, Kenya. HLA class I genes had been previously typed in all subjects (Peters et al. 2008). Both the ethics committees of the University of the Manitoba and Kenya National Hospital approved this study. All patients provided informed consent for participation in this study.

2. IV. 2 PCR amplification, cloning and sequencing with Sanger method

Proviral DNA was isolated from 96 HIV-1-positive women and the *gag* gene was amplified using nested PCR. PCR amplification was confirmed using 1% agarose gel electrophoresis. The PCR products were purified using the MultiscreenHTS PCR plate (Millipore Corporation). The PCR products were then cloned using TOPO TA cloning kit (Invitrogen). BigDye Terminator v3.1 (Applied Biosystems) was used to sequence *gag* gene with specific primers T7 5'-TAATACGACTCACTATAGGG-3', T3 5'-ATTAACCCTCACTAAAGGGA-3', (GSF1.6) GAGSEQF1.6 5'-GATAGAGGTAAAA GACAC CAAG-3' (277-298), (GSF2) GAGSEQF2 5'- CAGCATTATCAGAAGGAGC CAC-3' (541-562), (GOR) GAGPCRRN 5'- CTCCAATTCCCCCTATCATTTTTGGT TTCC-3' (outside *gag*). The sequencing products were purified using ethanol-sodium

acetate precipitation. Purified sequencing products were analyzed with an ABI 3100 Genetic Analyzer (Applied Biosystems). Nucleotide sequences were assembled and edited with Sequencher 4.8 (Genecodes Corp.). An average of 26 clones per patient was sequenced and sequences from a total of 6401 full *gag* clones were generated.

2. IV. 3 Ultra-deep pyrosequencing with GS20

Ultra-deep pyrosequencing was carried out with GS20 sequencer (454 Life Sciences) by the DNA Core, National Microbiology Laboratory, Public Health Agency of Canada. Briefly, the purified 96 PCR products were mechanically sheared, ligated to the adaptors, and amplified on capture beads in high-density water-in-oil emulsion picolitre reactors (one million beads per picolitre plate) followed by pyrosequencing. The pyrosequencing yielded 2,360,500 sequence reads, 37.6% of which passed the quality control, with an average read length of 102 base pairs. The fasta format files of read sequence data from all patients were extracted. Human DNA contamination was filtered out by BLASTing (ftp://ftp.ncbi.nih.gov/blast) the qualified read sequence alignment to HIV-1 reference sequences (HIV Sequence Database, Los Alamos National Laboratory, USA), resulting in 771,011 read sequences remaining for this study.

2. IV. 4 Sequence Alignment and measure of variants

The sequences of clones from each patient were aligned to the HIV-1 HXB2 *gag* gene by ClustalW (Thompson et al. 1994). The variants and their frequencies in each sequence pool were identified by comparing the sequence of each clone to the consensus of the pool. The GS20 sequencer reads from each patient were mapped onto the corresponding

multiple aligned Sanger clone sequences by WUBLAST 2.0 (http://blast.wustl.edu, Washington University). The coordinate of each read to reference was recorded for building sequence consensus and calculating the frequency of nucleotide/or amino acid at each position by in-house developed Perl scripts (see Appendix B).

Variants were determined if their frequencies were less than 50% compared to the *gag* reference. Variants were considered valid using previously determined statistical methods and only variants with $P < 0.001$ were considered to be real (Wang et al. 2007). The number of nucleotide variants only detected by pyrosequencing was determined and plotted at each of the 1503 nucleotide positions of the *gag* gene. The average numbers of nucleotide variants per position on *p17, p24* and *p1p7p2p6* genes were compared using the student *t* test.

2. IV.5 Measurement of amino acid variability and determination of correlation between amino acid variations and defined Gag epitopes

An entropy score was calculated for each position for both the 454 and Sanger cloning amino acid consensuses (Korber et al. 1994). To study the effect of variations between the two methods, we compared the entropy scores on each paired functional sites of Gag p17 and p24 with a paired *t* test. The functional and immunologically relevant sites of Gag p17 and p24 were defined as the ones overlapping previously described optimal CTL epitopes, neutralizing antibody epitopes, viral replication sites, and virus particle formation sites (HIV Sequence Database and HIV Immunology Database, Los Alamos National Laboratory, USA). To study the correlation between amino acid variations and

CTL epitopes regions, we first determined the differences in entropy scores on each position from Gag p17 and Gag p24 between the viral populations generated by these two methods. The entropy score differences were then correlated with the CTL epitope-rich regions by statistical analysis (Yusim et al. 2002). The entropy score differences were smoothed by averaging them in a window of nine amino acids to match the typical size of a CTL epitope.

2. IV. 6 Map consensus differences to Gag functional and immunologically relevant sites and positive selection analysis

The amino acid consensus differences were first determined by comparing the consensus sequences generated from each patient by each of the two methods and then mapped to the HIV-1 HXB2 Gag reference (HIV Sequence Database, Los Alamos National Laboratory, USA). The consensus differences overlapping HIV-1 Gag functional and immunologically relevant sites were determined. Identification of positively selected sites was assessed by using four different approaches: Quasi analysis (Stewart et al. 2001), single-likelihood ancestor counting (SLAC), fixed-effects likelihood (FEL), and random-effects likelihood (REL) methods implemented in HyPhy package (Kosakovsky Pond and Frost 2005; Pond and Frost 2005). For the last three methods, the optimal time reversible substitution model was first determined for the applied sequence data and the maximum likelihood-based analysis was then carried out on the DATAMONKEY server (http://www.datamonkey.org, University of California).

2. IV.7 Statistical Analysis and Perl scripts

All statistical tests were performed with GraphPad Prism 4. Perl scripts developed by Binhua Liang in-house are attached in Appendix B.

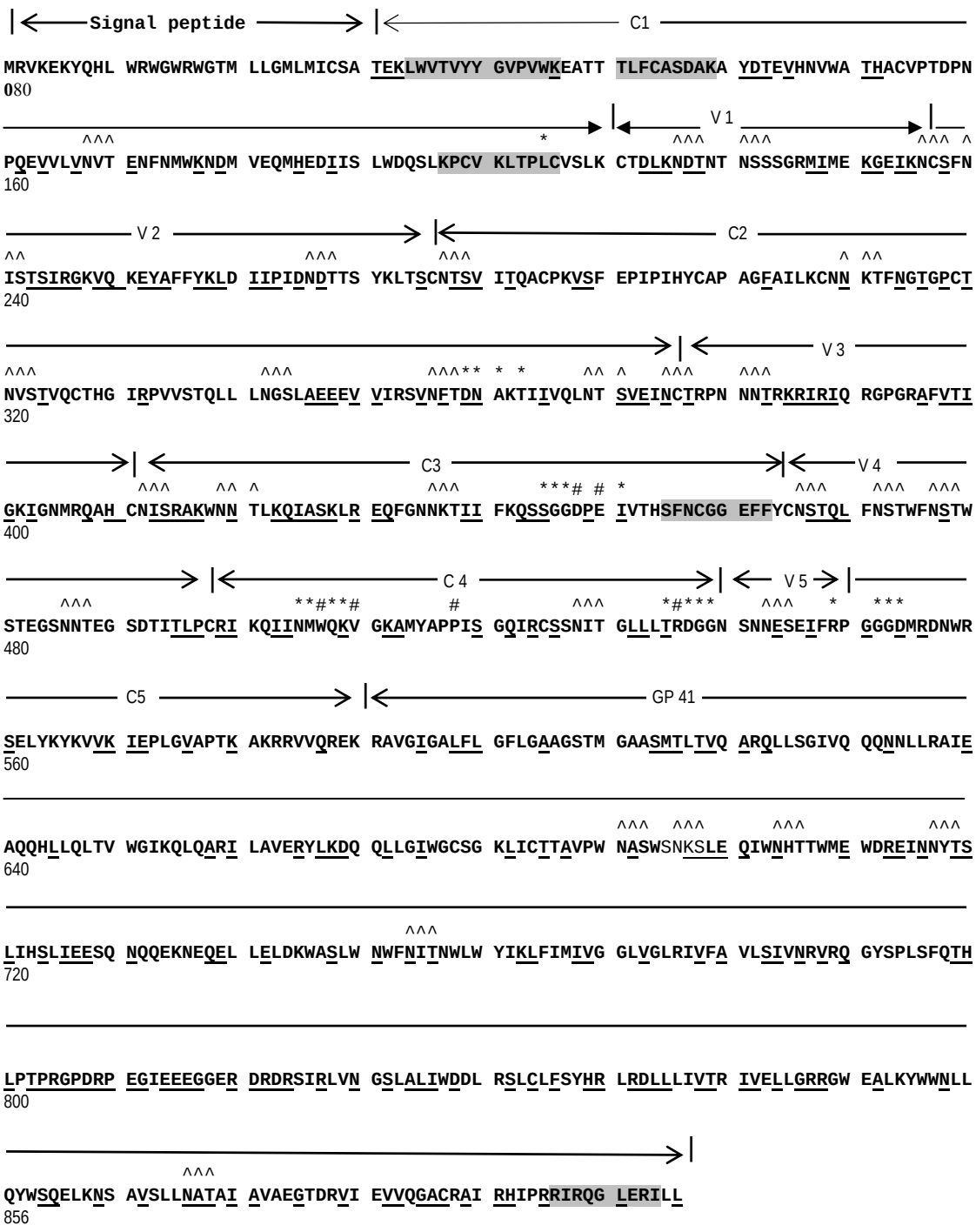
3.0 Results

Part I: Evolution and characterization of HIV-1 *Env* Gene

3. I. 1 Distribution of positive selection sites in the population

Knowing that identification of PS helps interpret host-virus interaction, examination of the distribution of PS patterns would allow us understand how host immune responses drive HIV-1 evolution at a population level. We conducted a QUASI analysis on the full-length Env sequences of HIV-1. A total of 288 positively selected amino acid sites were identified, accounting for 33.6% of amino acids within the Env protein. This suggests that the PS pressure on Env is very high in the population. Some QUASI identified positive selection sites were consistent with previous reports (de Oliveira et al. 2004; Travers et al. 2005; Yamaguchi-Kabata and Gojobori 2000), but more sites were identified and dispersed across the entire Env protein (Figure 3).

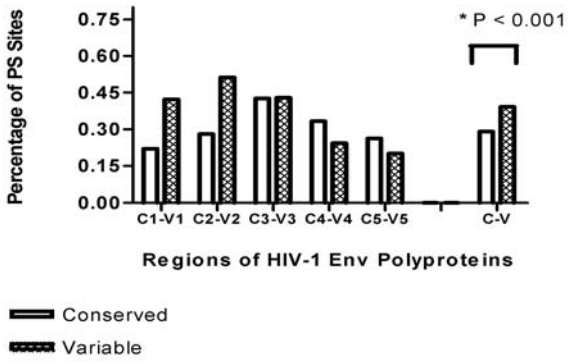
Figure 3 PS sites on the HIV-1 Env glycoproteins



PS sites were identified on the HIV-1 Env glycoproteins by QUASI approach. PS sites were underlined and numbered according to reference HXB2. The gp120 is further defined as the five conserved (C1-5) and five variable (V1-5) regions as indicated. The known neutralizing and CTL epitopes rarely positively selected are highlighted. ^: N-linked glycosylation site regions; #: critical CD4-binding sites; *: CD4-binding sites.

The majority of the identified positively selected sites (80.43%) on gp120 are on the solvent-accessible surface of the ternary complex of gp120 glycoprotein (Kwong et al. 1998; Yang et al. 2003). More PS sites were identified in the variable regions (39.16%) than in the conserved regions (28.99%) of Env ($P < 0.001$; Figure 4, V compared to C section), suggesting the variable regions of Env are under more selective pressure than the conserved regions of Env. Twenty PS sites were identified on N-linked glycosylation sites, which accounts for 68.97% of the total identified N-linked glycosylation sites, but only ten of these 20 PS sites disrupted required glycosylation patterns, implying that N-linked glycosylation sites are under some PS pressure. In addition, 4 out of 10 positively selected N-linked glycosylation sites on gp120 were within or neighboring the regions of the V3 (V3 loop), suggesting considerable PS on the V3 loop. Many regions involved in CD4 binding are under PS pressures (Figure 3). However, the critical contact residues such as Asp 368, Glu 370, and Trp 427 are conserved (Kwong et al. 1998). Some stretches of sequence were found to be free from PS, mainly located in the conserved regions of C1 and C2 (Figure 3).

Figure 4 Comparison of PS site densities among different regions of gp120



The percentages of PS sites were calculated as the number of PS sites over the total number of amino acids in each region of gp120. The percentages of PS sites paired in the defined conserved and variable regions of Env were plotted. C: conserved regions of HIV-1 Env; V: variable regions of HIV-1 Env; *: Significance was found.

3. I. 2 Comparison of positive selection sites across major subtypes

Since cross-clade immune responses exist, host immune responses possibly target the same regions of HIV-1 Env regardless of clade membership. To test this hypothesis, we identified the number of PS sites in each of the four major subtypes and compared them (Table 4). It was clear that there are significant variations in the percentage of PS sites among the four major HIV-1 subtypes, suggesting that different immune selection pressures may be exhibited on different subtypes. We then compared the locations of the PS sites among the major circulating subtypes. We found that many of the PS sites are located at the same locations (Table 5). Approximately 25% of PS sites are shared by clades A, B, C, and D, while 61% of PS sites were found to be shared among at least three of the four major clades. This suggests that at least for some regions, PS pressures are similar among different clades.

Table 4 Comparison of positive selection sites between paired clades

Clade	#[a] of the Unique PS Sites Between Each Pair (D)	# of the Shared PS Sites Across Each Pair (S)	D/S Ratio	*P*[b]	*P*-adj[c]
A-B	216	131	1.65	0.001	0.002
A-C	178	109	1.63	0.001	0.002
A-D	172	83	2.07	0.001	0.002
B-C	194	156	1.24	0.025	0.025
B-D	232	108	2.15	0.001	0.002
C-D	168	99	1.70	0.001	0.002

[a] the number;
[b] *Chi Square Test*;
[c] Adjusted for multiple testing by Bonferroni and Sidak;
PS: positive selection; D: non-synonymous; S: synonymous.

Table 5 Comparison of the distribution patterns of PS sites among different clades

Clade	Total number of PS sites	The number of PS sites shared across all clades	The number of PS sites shared among 3 clades
A	184	50(27.17)	117(63.59)
B	294	50(17.00)	143(48.64)
C	212	50(23.58)	131(61.79)
D	154	50(32.47)	108(70.12)

The locations of the identified PS sites were referred to HXB2. The numbers of PS sites at some locations among clades were calculated. Numbers in parenthesis indicates the percentages (%) of PS sites compared to the total number of PS sites in each clade. PS: positive selection.

3. I. 3 Changes of positive selection pressure over time

HIV-1 evolution is shaped by HLA allele frequencies within the human population, but HLA allele frequencies do not change much over time in the world. Thus, overall host immune pressure on HIV-1 Env stays stable over the past 2-3 dacades at a population level. We investigated PS pressure over time, and examined 2 groups of viruses where sufficient sequences were available: clade B and the entire Env sequence collection used in this study. We empirically divided Env sequences into 3 sub-groups based on five-year intervals and carried out QUASI analysis on each sub-group. While frequency of PS sites in clade B increased slightly from 1989 (19.7%) to 1999 (25%) (Figure 5), the frequency of PS sites remains stable from 1994 to 2004 if we include all subtypes in the analysis. The stability of PS pressures over time suggests that HIV viruses are under constant immune pressure, despite high mutation rates and rapid viral replication

Figure 5 Frequencies of PS sites on HIV-1 envelope glycoprotein over time

The frequencies of PS sites from the whole population, and clade B population were plotted based on five-year intervals. PS: positively selected.

3. I. 4 Association of positive selection sites with host immune responses and identification of epitopes

HIV-1 evolution is believed to be driven by host immune responses. Measurement of host immune pressure by determining PS on Env possibly predicts which host immune responses play key roles in HIV-1 Env evolution at a population level. We generated a system-level view of the host immune responses on HIV-1 Env in terms of PS by examining host immune responses likely responsible for PS. We correlated PS sites with experimentally defined antibody (Ab: 449), neutralizing Ab (NAb: 149), optimal CTL (35), and T-helper (Th: 380) epitopes based on the HIV Immunology Database 2005 from the Los Alamos National Laboratory. We also correlated PS sites with the predicted proteasomal cleavage sites on Env (Nielsen et al. 2005). When all the Ab, NAb, CTL, Th epitopes and cleavage sites were analyzed as independent factors, significant positive associations were identified between Th epitopes and PS (P = 0.0002; Table 6) and between neutralizing antibody response and PS (P = 0.0019; Table 6). When a model including the effects of only NAb and Th allowing interaction between them was run, this mode again showed that NAb (P = 0.013) and Th (P =0.0003) were associated with PS, but, against expectation, showed no significant interaction (P=0.6942) (Table 6). When NAb was analyzed in association with all other epitopes and cleavage sites except Th, and including interactions between NAb and these factors, we found a significant three-way interaction between NAb, CTL, and PCS (P=0.0310) (Table 6). Lastly, Th was also analyzed in association with all epitopes but not NAb, and including interactions between Th and these factors, we identified no significant interactions. These results suggest that the adaptive selection observed in HIV-1 Env is most likely determined by T helper

responses and neutralizing antibody responses, as well as the combined action of neutralizing antibody responses and $CD8^+$ cellular immune responses.

Table 6 Association between PS and host immune responses [§]

The involved Factors	The involved epitopes[a]	*P* values[b]
All factors only No interaction[c]	CTL NAb Ab Th PCS[d]	0.8962 **0.0019** 0.0955 **0.0002** 0.8501
NAb & Th With interaction[e]	NAb Th NAb*Th	**0.0130** **0.0003** 0.6942
NAb, PCS, & CTL With interaction[f]	NAb PCS CTL NAb*PCS NAb*CTL PCS*CTL NAb*CTL*PCS	0.2142 0.3479 0.6086 0.1947 0.2883 0.0995 **0.0310**
Th, CTL, & PCS With interaction[g]	Th CTL PCS Th*CTL Th*PCS Th*CTL*PCS	0.9802 0.9757 0.7716 0.9758 0.5681 0.7140

[a] The epitopes were derived from the HIV immunology Database 2005, Los Alamos National Laboratory.
[b] Significant values ($P < 0.05$) are indicated in boldface type.
[c] Only the main factors were tested independently in the statistical model.
[d] Predicted cleavage sites of proteasome.
[e] NAb and Th were tested independently in the statistical model. The interaction of NAb and Th was also tested, ignoring all other factors;
[f, g] Three-way interactions were also tested;
[§] Statistical analysis was conducted by contingency table analyses (PROC CATMOD, SAS).

3. I. 5 Identification of immunogenic sequences for HIV-1 vaccine design

Effective vaccines should be able to elicit protective immune responses for all HIV-1 subtypes (Gaschen et al. 2002). Thus, regions conserved across all clades and with defined neutralizing antibody epitopes and protective CTL epitopes would be good vaccine targets. We examined the conserved regions of Env where PS sites are rarely observed and identified PS sites with the experimentally defined neutralizing antibody epitopes and CTL epitopes. Using this approach focusing on the conserved region of Env we located epitopes of one neutralizing antibody and six CTL epitopes (Table 7), including LWVTVYYGVPVWK (34-46), TVYYGVPVWK (37-46) TLFCASDAK (51-59), KPCVKLTPLC (117-126), SFNCGGEFF (375-383), SFNCGGEFFY (375-384), and RIRQGLERA (846-854). The CTL epitopes are restricted by HLA A3, B*07, A29, and A*0205. Among them, A3 and A*0205 have been associated with low transmission of AIDS and with reduced disease progression in Long-Term Non-Progressors (LTNPs) in previous studies (Liu et al. 2003; Propato et al. 2001) and these epitopes might be good candidates for HIV-1 vaccine design. These HLA types are frequent in European, North American, as well as sub-Saharan populations with frequencies over 10% in each geographic region. However, B*07 allele and its affiliated supertype B7 have been reported in many studies to associate with more rapid disease progression (Gao et al. 2005; Scherer et al. 2004; Trachtenberg et al. 2003). If immune responses to these epitopes are ineffective or detrimental, the above identified epitope restricted by HLA-B*07 may not be an optimal epitope to include in an effective HIV vaccine. Clearly, the utility of epitopes identified through bioinformatical approaches must be carefully evaluated by further immunologic and epidemiological studies.

Table 7 Identified NAb and CTL epitopes on the regions of HIV-1 Env free from PS

The conserved epitope[a]	Epitope type	HLA or Ab types	HLA Allele[b]	Association with AIDS[c]
LWVTVYYGVPVWK(34-46)	NAb	4E10	NA	PR
TVYYGVPVWK (37-46)	CTL	A*0301	Europe (14%)	Unknown
TLFCASDAK (51-9)	CTL	A3	North America, Europe (10~20%)	LTNPs
KPCVKLTPLC (117-26)	CTL	B*07	Europe (14.1~17.3%)	FP
SFNCGGEFF (375-83)	CTL	B15, C*0401,Cw4, Cw*0401, C*0407	Philippines, Twain (~52%)	Unknown
SFNCGGEFFY(375-84)	CTL	A29	Zimbabwe (10%)	Unknown
RIRQGLERA(846-54)	CTL	A*0205	Kenyan (8.7%)	LT

[a] Locations of epitopes are referred to HXB2

[b] Population and frequencies of HLA alleles are referred to the Database of the human Major Histocompatibility Complex (dbMHC), the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/).

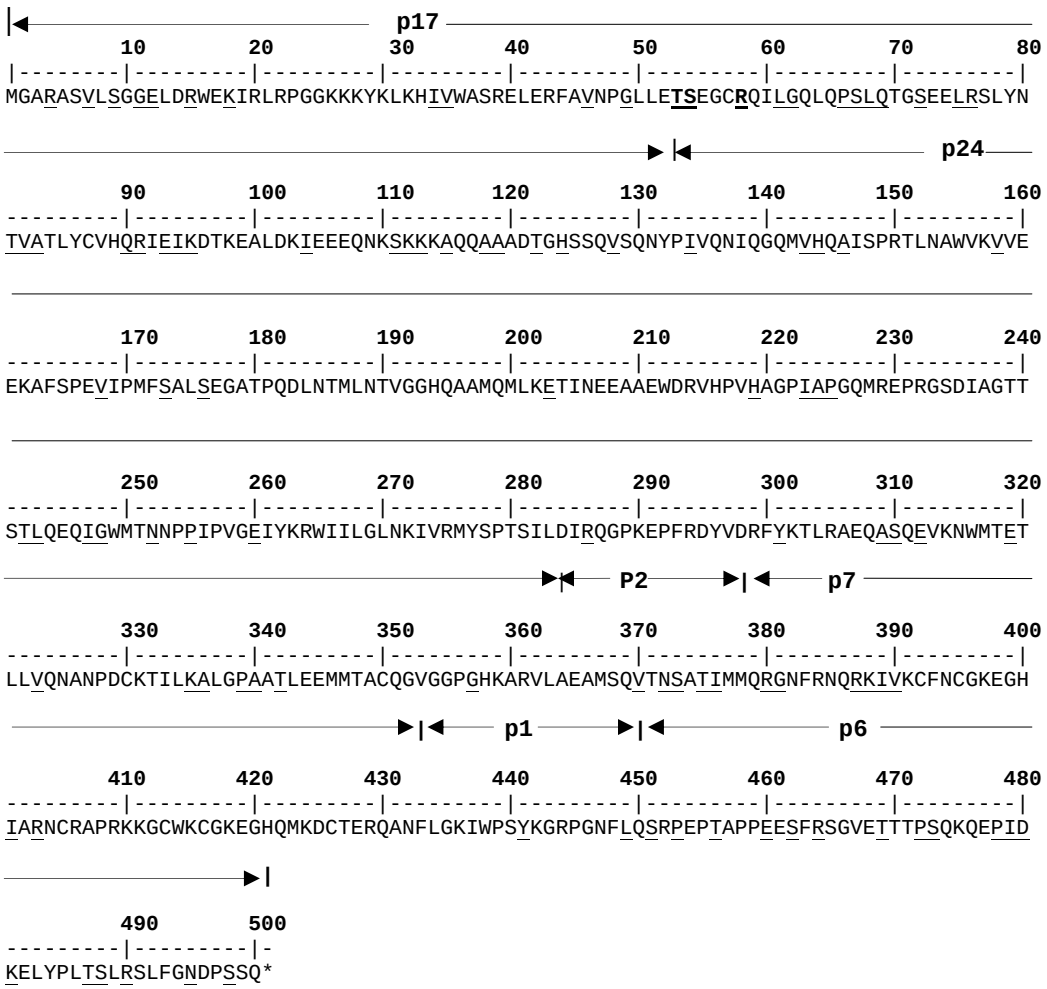
[c] FP: fast progression; PR: protective against HIV-1; LT : low transmission; LTNPs: Long-Term Nonprogressors.

Part II: Evolution and Characterization of HIV-1 *Gag* Gene

3. II.1 Identification and Distribution of positive selection sites

As with *env*, the full-length *gag* sequences of HIV-1 (Los Alamos Sequence Database) were subjected to QUASI analysis. A total of 107 positively selected amino acid sites were identified, accounting for 21.4% of amino acids within the Gag protein. This suggests that the PS pressure on Gag is common in the population but lower than on Env. It was found that PS sites dispersed across the entire Gag protein (Figure 6).

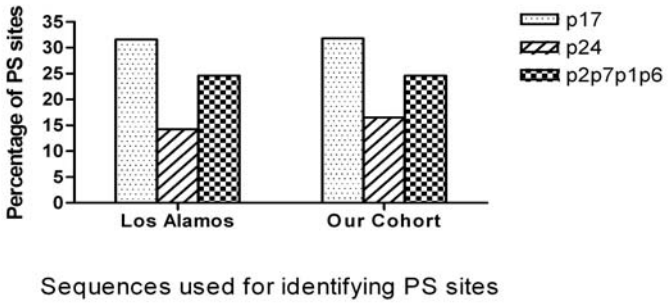
Figure 6 PS sites on the HIV-1 Gag Polyproteins



The HIV-1 Gag polyproteins were numbered according to reference HXB2. The HIV-1 Gag polyproteins are defined as p17, p24, p2, p7, p1, and p6 proteins. Positively selected amino acids are underlined.

The highest percentage of PS sites was found on p17 protein regions (31.06%), suggesting p17 protein sequences are under more positive selective pressure than other Gag regions at a global population level. In contrast, p24 was shown to be the most conserved with 14.29% of PS sites (Figure 7). The similar distribution pattern of PS sites was shown on Gag proteins from ML cohort (Figure 7). A total 46 of the identified PS sites overlap with the previously reported functional sites of Gag proteins such as CTL epitopes (42), neutralizing antibody epitopes (14), proteasomal cleavage sites (39), and particle packaging or formation of virion (9). Some stretches of sequence were found to be free from PS, mainly located in the regions of p24 and p7 (Figure 6).

Figure 7 Distribution of PS sites on HIV-1 Gag proteins



The percentages of PS sites were calculated as the number of PS sites over the total number of amino acids in each region of HIV-1 Gag proteins. The definitions of Gag proteins were referred to B.FR.83.HXB2_k03455 (HXB2). Los Alamos: *gag* sequences collected from HIV Sequence Database, Los Alamos National Laboratory, USA; Cohort: *gag* sequences derived from Pumwani Sex Worker Cohort, Nairobi, Kenya.

3. II.2 Comparison of positive selection sites across major subtypes

As with *Env*, we compared the number of PS sites in each of the three major subtypes (clade A, B, and C) (Table 8). There were significant variations in the percentage of PS sites among the three major HIV-1 subtypes, suggesting that different immune selection pressures may be exhibited on different subtypes of Gag. We then compared the locations of the PS sites among the major circulating subtypes. We found similar results to the analysis for Env that many of the PS sites were also located at the same locations on Gag. A range of 23.4% to 47.5% of PS sites is shared by clades A, B, and C in the different regions of Gag (Table 8). This result suggests that PS pressures are also similar among different clades of Gag, which is consistent with the result derived from the study on Env.

Table 8 Shared PS sites among different Gag subtypes

Clade	Total number of PS sites shared among all clades				Percentage of PS sites shared with other clades (%)			
	p17	p24	p7	p6	p17	p24	p6	p7
A	17(9)	25(10)	02(1)	11(6)	52.94	40.00	50.00	54.55
B	30(9)	22(10)	11(1)	12(6)	30.00	45.45	09.09	50.00
C	37(9)	31(10)	09(1)	16(6)	24.32	32.26	11.11	37.50
			Average:		37.75	39.23	23.4	47.35

The locations of PS sites were referred to HXB2. The numbers in parenthesis indicate the percentages (%) of PS sites compared to all the PS sites identified in each clade. PS: positive selection.

3. II.3 Association of positive selection sites with host immune responses and identification of epitopes

As Env, we also examined host immune responses likely responsible for PS at a system level. We correlated PS sites with both the predicted proteasomal cleavage sites (140) on Gag and experimentally defined NAb (4), Ab epitopes (125) and optimal CTL (39) epitopes on Gag based on the HIV Immunology Database 2008 from the Los Alamos National Laboratory, USA. When all the Ab, CTL, and proteasomal cleavage sites were analyzed as independent factors, no significant positive associations between them and PS sites were identified. When a model including the effects of all factors, allowing interaction between them, was run, this mode only showed a significant association between CTL & Ab and PS (P =0.0373) (Table 9). These results suggest that the adaptive selection observed in HIV-1 Gag is most likely determined by the combined action of $CD8^+$ cellular immune responses and Ab responses.

Table 9 Association between PS and host immune responses [§]

The involved factor	Epitopes involved[a]	P-value[b]
All factors only No interactions [c]	CTL	0.5771
	PCS	0.2876
	NAbs	0.3853
	Abs	0.5102
Ab, NAb, PCS, & CTL Interactions [d]	CTL*PCS	0.2121
	CTL*NAbs	0.4190
	CTL*Abs	**0.0373**
	PCS*Abs	0.2132
	PCS*NAb	0.8029
	NAbs*Abs	0.2719
	CTL*Abs*PCS	0.2220
	CTL*NAbs*PCS	0.2143
	CTL*NAbs*Abs	0.2112
	PCS*NAbs*Abs	0.9309

[a] The epitopes were derived from the HIV immunology Database 2006, Los Alamos National Laboratory, USA. CTL: cytotoxic T lymphocyte response epitopes; NAb: neutralizing antibody epitopes; Ab: non-neutralizing antibody epitopes; PCS: Predicted cleavage sites of proteasome;
[b] Significant values ($P<0.05$) are indicated in boldface type;
[c] Only the main factors were tested independently in the statistical model;
[d] Two-way interactions and three-way interactions were also tested;
[§] Statistical analysis was conducted by contingency table analyses (PROC CATMOD, SAS).

3. II.4 Identification of immunogenic sequences for HIV-1 vaccine design

From the previous analysis on Env, we identified several conserved epitope regions which may be potential candidates for effective vaccine design. In this study, we performed a similar analysis on Gag. The conserved regions of Gag where PS sites were not observed and identified PS sites with the experimentally defined CTL epitopes were examined. Using this approach and focusing on the conserved region of Gag, we identified epitopes of eighteen CTL epitopes across Gag proteins (Table 10). These CTL epitopes were primarily located on p24 (10), p17 (5), and others (3). The clades of these epitopes are mainly B (10) and C (5). These CTL epitopes are restricted by a variety of HLA class I A (2), B, (14) and C alleles. Among them, 10 HLA alleles were associated with favorable outcome of HIV infection such as slow progression to AIDS, low viral loads, and long-term non-progression. Nine of them are HLA type B and are frequent in European (11.8%), South Africa (7.5%), and North American (5.7%) populations. One CTL epitope, TPQDLNTML (p24 48-56), is restricted by multiple HLA alleles such as HLA B*0702, B*3901, B*4201, B*5301, B*8101, and Cw*0802 in which HLA B*3901, B*4201, and B*8101 were shown to be associated with slower progressions of AIDS. The utility of epitopes identified through bioinformatical approaches should be carefully evaluated by further immunologic and epidemiological studies.

Table 10 Identified conserved CTL epitopes on HIV-1 Gag proteins

Conserved epitope[a]	Clade	Gag protein[b]	HLA type	Distribution of HLA allele[c]	Association with AIDS[d]
AEWDRVHPV(78-86)	B	P24	B*4002	North America (15.9-63.9%)	Unknown
CRAPRKKGC(42-50)	B	P2p7p1p6	B14	North Africa (4.8 -21.8%)	Unknown
GGKKKYKLK(24-32)	B	P17	B*0801	Europe (11.8-37.3%)	SL(Turnbull et al. 2006)
GHQAAMQML(61-9)	C	P24	B*1510	South-Africa (< 4.3%)	LV(Kiepiela et al. 2007)
GLNKIVRMY(137-45)	B	P24	B*1501	North America (5.7-25.6%)	LV (Jones et al. 2004)
GQMREPRGSDI(94-104)	B, C	P24	B*13	South Africa (< 1.9%)	LV(Honeyborne et al. 2007)
IRLRPGGKK(19-27)	B	P17	B*2705	North America (9.1-19.3%)	Unknown
ISPRTLNAW(15-23)	B	P24	B*5701	Europe; North America (3.9-13.1%)	LTNP (Klein et al. 1998; Migueles and Connors 2001; Turnbull et al. 2006)
KRWIILGLNK(131-40)	B	P24	B*27	Europe (3.4-32.9%)	SL(Frater et al. 2007)
RLRPGGKKK(20-8)	B	P17	A*0301	Europe; North America (18.1-45.1%)	No (Altfeld et al. 2006)
RLRPGGKKKY(20-9)	B	P17	A*0301	Europe; North America (18.1- 45.1%)	No (Altfeld et al. 2006)
RMYSPTSI(143-50)	B	P24	B*5201	North America (2.0-19.3%)	Unknown
RQANFLGKI(66-74)	C	P2p7p1p6	B*13	South Africa (< 1.9%)	LV(Kiepiela et al. 2007)
SPRTLNAWV(16-24)	B	P24	B*0702	Europe; North America (18.1-41.5%)	No (Altfeld et al. 2006)
TERQANFL(64-71)	B	P2p7p1p6	B*4002	North America (15.9-63.9%)	Unknown
TPQDLNTML(48-56)	C	P24	B*0702	East Africa (13.9-41.5%)	No (Geldmacher et al. 2007)
	C		B*3901	South Africa (< 0.5%)	LV (Kiepiela et al. 2007)
	C		B*4201	South Africa (< 7.5%)	LV (Day et al. 2007)
	B		B*5301	North America (0.9%-16.8%)	Unknown
	C		B*8101	South Africa (< 3.0%)	LV (Kiepiela et al. 2007)
	B		Cw*0802	North America (0.9-(23.6%)	Unknow
VRMYSPVSI(142-50)	C	P24	Cw18	South Africa (< 3.6%)	SL (Kiepiela et al. 2007)
WASRELERF(36-44)	B	P17	B*3501	North America (13.5-(49.2%)	Unknown

[a,b] The definition of Gag proteins and the locations of epitopes on them are referred to HXB2;
[c] The population and frequencies of HLA alleles are referred to the Database of the human Major Histocompatibility Complex (dbMHC), the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/);
[d] SL: slow progression to AIDS; LV: low viral loads; LTNP: long-term non-progressor.

3. II.5 Comparison of PS sites identified between Los Alamos and Cohort Gag sequences

In order to provide evidence for PS sites identified from Los Alamos *gag* sequences, QUASI analysis was performed on the clade A *gag* sequences from Los Alamos Sequence Database. The identified PS sites were then compared to ones from the previous study on the clade A sequences from Pumwani Sex Worker Cohort. 54 out of the identified 61 PS sites from Los Alamos *gag* sequences were found to match those from the previous study (Figure 8). Furthermore, 3 out of 4 identified PS sites correlated with lower mean CD4+ counts were also identified by QUASI analysis on Los Alamos *gag* sequences. 82.35% of the identified PS sites correlated with HLA class I alleles matched ones from QUASI analysis of Los Alamos *gag* sequences (Table 11).

Figure 8 Comparison of the identified PS sites from Los Alamos[a] and cohort[b] HIV-1

Gag sequences

```
        1        10        20        30        40        50        60        70        80
        ---------|---------|---------|---------|---------|---------|---------|---------|
        MGARASVLSGGKLDRWEKIRLRPGGKKKYRLKHIVWASRELERFALNPSLLETAEGCQQMEGQLQPALKTGTEELKSLYN
Los
Alamos  ------I------------Q----------M----------K--I-G-----T-----------ST----S---R--F-
Cohort  ------I----Q-E-----Q-----P-Q-K-------G-MDK------G---ST-----------STIR-S---IR--F-

                 90        100       110       120       130       140       150       160
        ---------|---------|---------|---------|---------|---------|---------|---------|
        TVATLYCVHQRIDVKDTKEALDKIEEIQNKSKQKTQQAAADTGNSSKVSQNYPIVQNAQGQMIHQSLSPRTLNAWVKVIE
Los
Alamos  ------------EI---------V--M-------------A--------------------V--A------------V-
Cohort  AIV-I-I----MEI---------V-M-K-----RA------AESN-Q------V---V----V-AV-----------AV-

                 170       180       190       200       210       220       230       240
        ---------|---------|---------|---------|---------|---------|---------|---------|
        EKAFSPEVIPMFSALSEGATPQDLNMMLNIVGGHQAAMQMLDTTINEEAAEWDRLHPVHAGPIPPGQMREPRGSDIAGTT
Los
Alamos  -RG---------T----------------T-----------------------V-A-------A---------------
Cohort  -RG---------T--A---I---------V-----------------------V--AQV---VA---------------

                 250       260       270       280       290       300       310       320
        ---------|---------|---------|---------|---------|---------|---------|---------|
        STPQEQIGWMTSNPPIPVGDIYKRWIILGLNKIVRMYSRVSILDIKQGPKEPFRDYVDRFFKTLRTEQATQEVKGWMTET
Los
Alamos  --L----A---G--------------------------------VR----------------A------S--------D-
Cohort  --T---LA---G---V------RK--------------------V----------------A------S-D------D-

                 330       340       350       360       370       380       390       400
        ---------|---------|---------|---------|---------|---------|---------|---------|
        LLVQNANPDCKSILRALGAGATLEEMMTACQGVGGPGHKARVLAEAMSQVQQT-NIMMQRGNFRGQKRI-KCFNCGKEGH
Los
Alamos  --I--------T-KG---P--S--------------S------------A-HP------K---K----------------
Cohort  --I--------T-KG---P--S--------------S----------------------K---K--R-------------

                 410       420       430       440       450       460       470       480
        ---------|---------|---------|---------|---------|---------|---------|---------|
        LARNCRAPRKKGCWKCGKEGHQMKDCTERQANFLGKIWPSSKGRPGNFPQSRPEPTAPPAEIFGMGEEITSPPQKQEQKD
Los
Alamos  ----------------------------------R----N---------N-L-S------L---R---AP-L-------
Cohort  --K-K-----------------------------RL-SPN---------S-L--------LC-R---TTPSL------E

                 490       500
        ---------|---------|-
        REQ-PPLVSLKSLFGNDPLSQ*
Los
Alamos  K-----SI---------L---
Cohort  K-P-S-SI-R----SS-LSL-
```

QUASI analysis was performed on the clade A *gag* sequences from Los Alamos Sequence Database and the identified PS sites were aligned together along the consensus of *Gag* sequences from a previous study (Liang et al. 2008) and its PS sites. The positions of *gag* sequence were referred to HXB2. The upper and lower lines below each consensus represent PS sites identified from Los Alamos *gag* sequences and cohort sequences, respectively.

[a] HIV Sequence Database, Los Alamos National Laboratory, USA;

[b] Pumwani Sex Worker Cohort, Kenya.

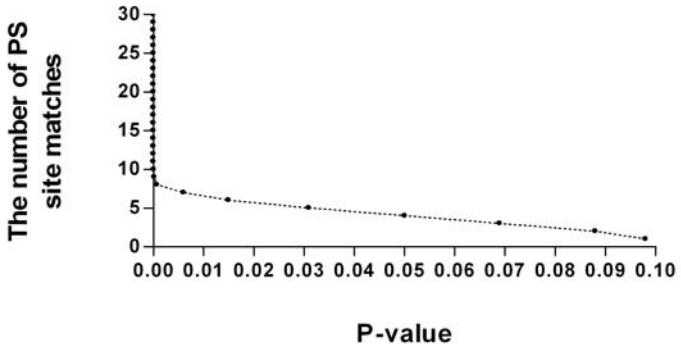
Table 11 the match of the PS sites identified from Cohort[a] and Los Alamos[b] *gag* sequences

protein	Mutation[c]	HLA Correlation[d]	CD4+ Count Correlation[e]	Match
P17 (16)	K12Q	POS	N/A	YES
	D14E	POS	N/A	NO
	R20Q	POS	N/A	YES
	K26R	POS	N/A	NO
	K28R	POS	N/A	NO
	E42D	POS	YES	NO
	R43K	POS	N/A	YES
	S49G	POS	N/A	YES
	L75I	POS	N/A	YES
	F79Y	NEG	YES	YES
	T81I	NEG	N/A	YES
	V82I	POS	N/A	YES
	V88I	POS	N/A	YES
	I92M	POS	N/A	NO
	D93E	NEG	N/A	YES
	T115A	POS	N/A	YES
			Percentage of matches:	69% (11/16)
P24 (10)	A163G	POS	YES	YES
	I190V	POS	YES	YES
	P243T	POS	N/A	YES
	I247L	POS	N/A	YES
	T303A/I	POS	N/A	YES
	T310S	POS	N/A	YES
	E312D	POS	N/A	YES
	E319D	POS	N/A	YES
	T242S	POS	N/A	YES
	G357S	POS	N/A	YES
			Percentage of matches:	100% (11/11)
P7 (3)	R384K	NO	YES	YES
	K387R	POS	N/A	YES
	R402K	POS	N/A	YES
			Percentage of matches:	100% (3/3)
P1 (2)	K435R	POS	N/A	YES
	S440R	POS	N/A	YES
			Percentage of matches:	100% (2/2)
P6 (2)	R472L/Q	POS	N/A	YES
	K489R	POS	N/A	YES
			Percentage of matches:	100% (2/2)

[a] Pumwani Sex Worker Cohort, Kenya;
[b] Los Alamos Sequence Database, National Laboratory, USA;
[c] The positions of mutations were referred to HXB2;
[d] The correlation of HLA class I alleles to positively selected amino acids in clade A1 and D Gag proteins from Cohort; POS: positive; NEG: negative;
[e] The correlation of lower mean CD4+ counts to positively selected amino acids in clade A1 and D Gag proteins from Cohort; N/A: not apply.

To further confirm these results, we conducted a statistical study testing the probability of PS site matches from the sequences from these populations. The result showed that the probability of PS site matches by random chance was very low, with the probability of five PS site matches between two PS site pools at the same time being less than 0.048. For over 10 PS site matches, the possibility was less than $10e^{-10}$ (Figure 9). However, the numbers of PS site matches between two PS site pools was 29 in this study, indicating that the PS sites identified from Los Alamos *gag* sequences are the same.

Figure 9 the match of the PS sites between Los Alamos[a] and Cohort[b] *gag* sequences



30 PS sites were generated at random from all the identified PS sites on Los Alamos *gag* sequences. These PS sites were then matched to ones identified from a previous study in Cohort (Peters et al. 2008) and the possibility of PS site matches from two sequence populations were calculated.

[a] Los Alamos Sequence Database, Los Alamos National Laboratory, USA;

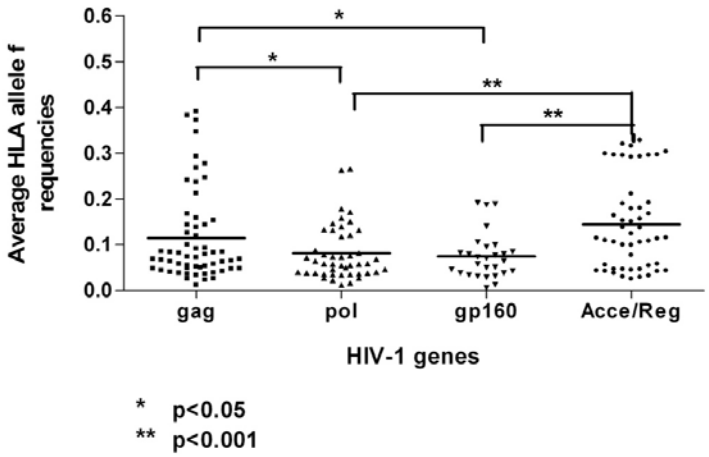
[b] Pumwani Sex Worker Cohort, Kenya.

Part III: Impact of HLA Class I Allele Frequencies on CTL responses

3. III. 1 HLA class I allele frequencies in the population

HLA class I alleles recognize HIV-1 specific CTL epitopes. Thus, determination of HLA allele (restricting the known common CTL epitopes in the world) frequencies could give an overview of which and to what extent HLA class I alleles regulate host CTL responses mediated by HIV-1 proteins at a population level. In my study, a total of 77 HLA class I alleles, including HLA-A (23), B (39), and C (15), were found to restrict 182 optimal CTL epitopes across HIV-1 proteins. The calculation of average HLA class I allele frequencies was carried out for all the restricted optimal CTL epitopes at population level. The results showed that there are more HLA alleles restricting the Gag CTL epitopes and accessory/regulatory proteins' CTL epitopes than the ones restricting epitopes of other HIV proteins, respectively (p <0.05 and p <0.001) (Figure 10). This result demonstrates that the Gag and accessory/regulatory protein's CTL epitopes are likely under more intensive restriction by HLA class I molecules.

Figure 10 Comparison of HLA class I allele frequencies on the different restricted CTL epitopes located on HIV-1 proteins



Optimal CTL epitopes from 2006 Immunology Database, Los Alamos National Laboratory, USA, were classified into four groups based on their locations on HIV-1 genes. The corresponding restricted average HLA allele frequencies were calculated for each group and were compared. Acce/Reg: accessory and regulatory genes. Y-axis: average HLA class I allele frequencies; X-axis: HIV-1 genes; *: $p<0.05$; **: $p<0.001$.

The distribution of epitopes of different HLA class I genes was analyzed among HIV-1 proteins. The epitopes of HLA-A alleles mostly concentrated on Pol protein, whereas the epitopes of HLA-B alleles were dominant on both Gag and accessory/regulator proteins. The epitopes of HLA-C alleles were not commonly observed on any HIV proteins (Table 12). In addition, 33 out of 182 CTL epitopes were found to be restricted by multiple HLA class I alleles range from 2 to 6. These include the epitopes in Gag (12), Pol (3), Gp160 (4), and accessory/regulator's proteins (14). Furthermore, eight of these epitopes were restricted by combination of alleles from A, B, and C. On the other hand, 25 HLA class I molecules restrict multiple different CTL epitopes, including HLA-A (10) and HLA-B (15) molecules. There is no difference in frequencies of alleles restricting the epitopes between HLA-A and HLA-B (data not shown).

Table 12 Distribution of HLA class I alleles across HIV-1 proteins.

HIV-1 proteins[a]		HLA-A	HLA-B	HLA-C
Gag	P17	11	10	2
	P24	6	39	6
	Others	0	4	4
	Total:	17 (20.73%)	53 (64.63%)	12 (14.63%)
Pol	RT	19	15	0
	IN	3	6	1
	Others	3	2	0
	Total:	25 (51.02%)	23 (46.94%)	1 (2.04%)
Gp160	Total:	16 (48.48%)	13 (39.39%)	4 (12.12%)
Accessory	Vif	3	6	0
	Vpr	3	5	0
	Tat	1	3	1
	Rev	1	3	0
	Vpu	1	0	0
	Nef	10	26	4
	Total:	29 (37.66%)	43 (55.84%)	5 (6.49%)

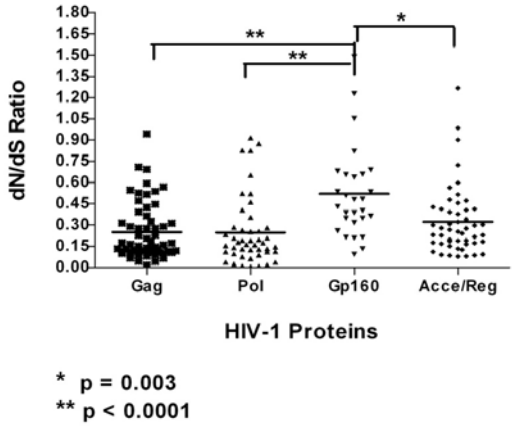
The percentage of dominant HLA class I type is highlighted on each HIV-1 protein.
[a] RT: reverse transcriptase; IN: integrase; Vif: virus infectivity factor; Vpr: virus protein R; Tat: transactivator protein; Rev: regulator of expression of virus protein; Vpu: virus protein V; Nef: enhancing regulatory factor.

3. III.2 dN/dS ratio of the CTL epitopes across HIV-1 proteins

By comparing synonymous and non-synonymous nucleotide substitution (dN/dS) with statistical analysis on CTL epitope sequences, we should be able to estimate host immune pressure exerted on CTL epitopes and determine what extent HIV-1 proteins are under host immune responses at a population level. We determined dN/dS ratio on 241 optimal CTL epitopes based on Los Alamos HIV-1 sequences overlapping the epitopes. The analysis showed that only four epitopes have dN/dS ratio large than 1, including FLKEKGGL (Nef, 90-97), EVAQRAYR (Gp160, 320-327), IVTRIVELL (Gp160, 777-785), AND RVKEKYQHL (Gp160, 2-10). The result suggests that these CTL epitopes are possibly under host positive selection pressures, accounting for only 2% of the studied CTL epitopes. Whereas, dN/dS ratios of other epitopes were less than 1, indicating that the majority of CTL epitopes were under natural selection.

We compared dN/dS ratios of the CTL epitopes derived from different HIV-1 proteins to have an overview of host selective forces on these protein epitopes at a population level. It was clear that the dN/dS ratio of the CTL epitopes located on Gp160 proteins was significantly higher than any one located on other proteins ($p <= 0.003$) (Figure 11), indicating that the CTL epitopes on Gp160 were likely under more host selection pressure.

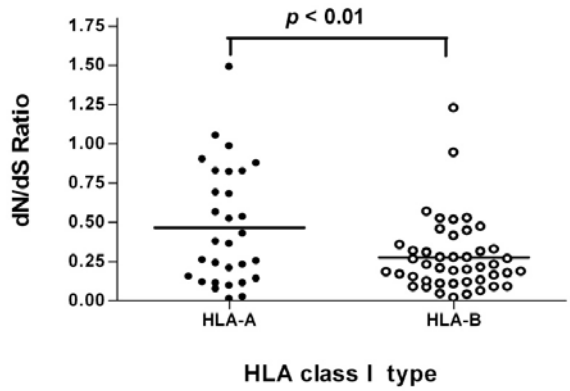
Figure 11 Comparison of dN/dS ratios on the CTL epitopes from the different HIV-1 proteins



The host immune response pressures, estimated as dN/dS ratio, on the restricted epitopes were compared among HIV-1 proteins. The optimal CTL epitopes, from 2006 Immunology Database, Los Alamos National Laboratory, USA, were classified into four groups based on their locations on HIV-1 proteins. The corresponding dN/dS ratios were calculated for each group and were compared. Acce/Reg: accessory and regulatory proteins. Y-axis: dN/dS ratio; X-axis: the studies HIV-1 proteins; *: p=0.003; **:p<0.0001.

There are no significant differences of dN/dS ratios among the CTL epitopes on Gag, Pol, and Accessory proteins. Further investigation of dN/dS ratios among the different epitopes showed significantly higher dN/dS ratio of the CTL epitopes restricted by HLA-A than ones restricted by HLA-B ($p < 0.01$) (Figure 12). In contrast, no significant differences in dN/dS ratios were found between the epitopes restricted by single HLA class I allele and those restricted by multiple HLA class I alleles (data not shown).

Figure 12 Comparison of dN/dS ratios on the CTL epitopes restricted by HLA-A and HLA-B alleles.



The dominant role of HLA class I types on restricting CTL epitopes was compared between HLA-A and HLA-B. The dN/dS ratios were calculated for the CTL epitopes which are restricted by HLA-A and HLA-B alleles. A Student *t* test was conducted between the dN/dS ratios on the CTL epitopes restricted by HLA-A and HLA-B alleles. Y-axis: dN/dS ratio; X-axis: HLA class I type.

3. III.3 The correlation of HLA allele frequency and dN/dS ratio

HLA allele frequency distributions are different among different populations in the world. Since CTL epitopes are restricted by HLA alleles, population HLA allele frequencies should influence host CTL responses against HIV-1. In another word, a correlation should exist between HLA allele frequencies (within population) and host CTL responses against HIV-1. To test this hypothesis, we examined the association of population HLA allele frequency and the corresponding selective force in the host populations. We calculated the average HLA allele frequencies that restricted certain CTL epitopes at a population level. Then, we measured the host selective pressure by estimating dN/dS ratio based on the entire sequences available that overlap the specific epitope and correlated it with average HLA allele frequency. At a population level, no correlation between HLA allele frequency and the corresponding dN/dS ratios on the restricted CTL epitopes was observed.

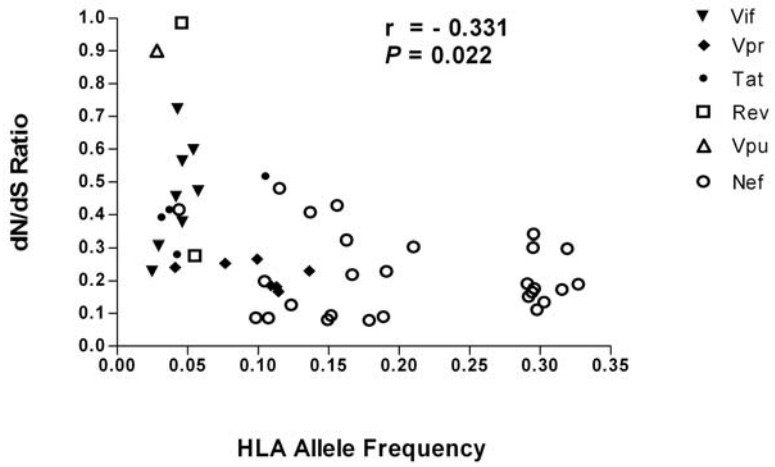
There are many "confounding factors" in studying HLA associations with disease progression, such as HLA and locations of CTL epitopes targeted. In order to remove the confounding effects, CTL epitopes were further classified into the data sets based on the HLA genes targeted and were examined as described above. Using stratified analysis, the significant correlation between HLA allele frequency and dN/dS ratio was found for the CTL epitopes located on the accessory genes (Table 13; Figure 13).

Table 13 Association between HLA allele frequency and dN/dS ratio

Study Factor		R[a]	P-value[b]
Genotype			
	Gag	0.1246	0.3514
	Accessory proteins	-0.4122	**0.0029**
	Gp160	0.0369	0.8523
	Pol	0.0219	0.8850
HLA type			
	Recombination	-0.0628	0.5127
	Non-recombination	0.0397	0.7536
HLA allele[c]			
	Single	-0.0858	0.5127
	Multiple	-0.0186	0.9129

[a] R: coefficient factor;
[b] The correlation between HLA allele frequency and dN/dS ratio was assessed using Pearson Correlation Coefficients in SAS v.9.1. The *p* value with the significance is highlighted.
[c] The epitopes restricted by single or multiple HLA alleles.

Figure 13 The correlation of HLA allele frequency and dN/dS ratio of CTL epitopes on HIV-1 accessory proteins



The dN/dS ratio on the corresponding CTL epitope on HI-1 accessory proteins was determined to estimate host CTL immune response on HIV-1 accessory proteins. dN/dS ratio was plotted along average HLA allele frequency. Each dot represents one epitope. The correlation of HLA allele frequency with dN/dS ratio was measured by Pearson's correlation coefficient method implemented in SPSS 16 and was tested at a two-tailed significance.

3. III.4 The impact of HLA class I supertypes on the study of disease association

HLA class I molecules are very polymorphic with over thousands allelic variants. Many of these variants are located in CTL epitope binding regions. HLA class I molecules are classified into supertypes based on sharing overlapping epitope binding specificities and thus immunological relevance. HLA class I supertypes are thought to increase the coverage of the population and associated with disease progression (Sette and Sidney 1999; Trachtenberg et al. 2003). Thus, classification of HLA class I molecules into supertypes could increase sensitivity for us to detect association between HLA allele frequencies and host immune responses. In this study, we classified CTL epitopes into 6 groups based on their restricted HLA alleles classified into HLA class I supertypes (Sidney et al. 2008) (Table 14). A statistical analysis was conducted to test the association between frequency of each HLA class I supertype and its corresponding dN/dS ratio. The result showed a significant correlation between HLA-B*7 supertype frequency and dN/dS ratio (Table 15).

Table 14 HLA class I supertypes used in this study

HLA Supertype[a]	HLA Alleles	[b]#HLA type	#HLA Alleles
A1	A*0101, 2501,2601, 2902,3002,3201	6	14
A2	A*0201,0202,0205,0207,6802	5	15
A3	A*0301,1101,3303,6801	4	29
B7	B*0702,3501,4201,5101,5501	5	29
B8	B*0801	1	9
B27	B*1402,1503,2705	3	15
B44	B*1801,4001,4002,4402,4501	5	15
B58	B*5701,5703,5801	3	18
B62	B1501,1503,5201	3	6

[a] The classification of HLA supertypes were based on the previous report (reference);
[b] The number.

Table 15 Association between HLA allele frequency of supertypes and dN/dS ratio

HLA class I supertype[a]	R[b]	P-value[c]
A1	0.028	0.928
A2	-0.166	0.555
A3	-0.258	0.177
B7	-0.4354	**0.019**
B8	-0.288	0.299
B27	0.188	0.502
B58	-0.089	0.725

[a] The classification of HLA supertypes were based on the previous report (reference);
[b] R: coefficient factor;
[c] The correlation between HLA allele frequency of each data set and its corresponding dN/dS ratio was assessed by using Pearson Correlation Coefficients implemented in SAS v.9.1. The *p* value with the significance was highlighted.

Part IV: Impact of 2nd-Generation Sequencing Technologies on the Sequence Analysis of HIV-1 Genes

3. IV.1 Characterization of 454 pyrosequencing data

Assuming that pyrosequencing technologies provides the ultra-deep coverage of target sequences and accurately and quantitatively assess HIV quasispecies, we should be able to detect more variants within HIV-1 *gag* quasispecies relative to Sanger. An average of 8031 sequence reads per sample was generated by the GS20 platform. The consensus *gag* sequences from all 96 samples generated from the GS20 platform matched the consensus sequences generated by Sanger sequencing. Twenty-six out of 96 samples with a redundancy of sequence reads lower than 100 fold or not consistent between regions of alignments, were excluded. For the remaining 70 samples, a total 85.8% of GS20 read sequences mapped to HIV-1 *gag* with an average depth of 384 sequence reads per position. The redundancy within the first 100 base pairs of amplified HIV-1 *gag* gene was lower than other part of *gag*.

We compared the two methods in their ability to characterize genetic diversity within HIV-1 *gag* quasispecies. Sanger clone-based method detected a total of 3632 variants over 1503 nucleotides of the HIV-1 *gag* gene with an average of 53 variants per sample. By comparison, 454 pyrosequencing detected 14034 variants at an average of 204 variants per sample (Table 16). The majority (2984/3632, 82.2%) of the variants detected by Sanger clone-based method were also detected by 454 pyrosequencing except for 642 variants. However, 11050 variants were only detected by 454 pyrosequencing. In addition, analysis of variant composition showed that 33.2% of variants (1205/3632) detected by

Sanger clone-based sequencing were present at a frequency of ≤ 20%. By contrast, 454 pyrosequencing detected 80.2% of variants (11262/14034) at frequency of ≤ 20% (Table 16).

For the variants only detected by 454 pyrosequencing, the majority of them (9504/11050, 86%) were present at frequency of <10%, and 38.1% of them (4215/11413) present at frequency of <2%. Furthermore, within 9504 variants, 4444 of them are non-synonymous mutations. All of these minor variants were confirmed by statistical analysis in which only those variants with their frequencies of occurrence yielding a *P* value of <0.001 were considered to be real based on the Poisson model (Wang et al. 2007).

Table 16 the comparison of the detected variants by 454 and Sanger clone-based sequencing methods

Method	Total number of minor variants[a]				The minor variants only detected by 454 or Sanger Clone-based methods		
	Total	Average/per	>20%	<20%	>10%	2-10%	<2%
454	14034	204	2772	11262	1546	5289	4215
Sanger	3632	53	2427	1205	642	6	N/A

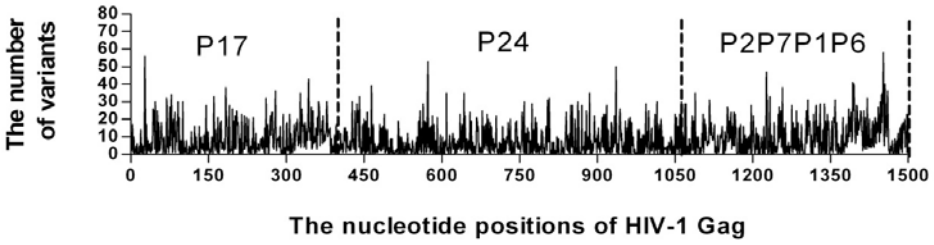
[a] A minor variant is defined as a nucleotide with a frequency less than 20% referred to the consensus population-based nucleotide sequence;
N/A : not applicable

3. IV.2 Distribution of genetic variants

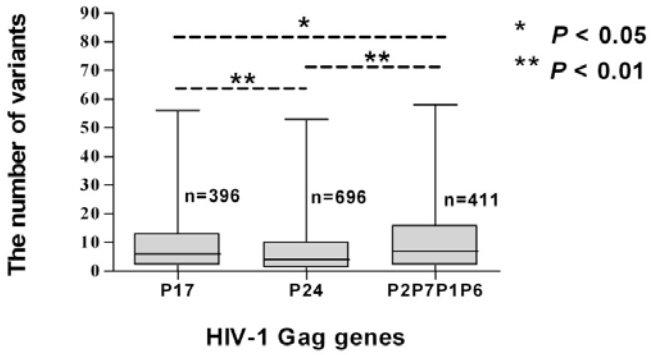
To explore the distribution of nucleotide variants, we determined the number of variants detected only by pyrosequencing at each nucleotide position of HIV-1 *gag* gene and plotted them. The nucleotide variants are distributed over entire *gag* gene with a range from 1 to 58 per position (Figure 14). Statistical analysis showed that the number of variants per position in the p24 gene region were significantly lower than the ones in p17 ($p = 0.007$) or p1p7p2p6 ($p < 0.0001$) (Figure 14).

Figure 14 Distribution of the variants detected only by pyrosequencing method

(A).



(B).



The variants were plotted along HIV-1 *gag* gene referred to HXB2. (A). a plot of the variants along the nucleotide positions of HIV-1 *gag*. (B). the number of variants among p17, p24, and p2p7p1p6 were compared with statistical analysis. n: number.

The observation suggests the relative conservation of p24 due to its functional constraints limiting its accrual of diversity. There was no significant difference between the numbers of variants per position in p17 compared to those in p1p7p2p6.

3. IV.3 Impact of variations on the functional characterization of HIV-1 Gag proteins

Viral sequence variants are generated by error prone reverse transcriptase and selected and maintained through interactions with host immune responses (Bonhoeffer et al. 1995; Liang et al. 2008; Yusim et al. 2002). The sequence differences between pyrosequencing and Sanger clone-based method could impact our understandings of viral-host interactions. To test this hypothesis, we compared the non-synonymous variations from 454 pyrosequencing to those determined by Sanger clone-based method in terms of their entropy scores. Since the entropy scores can quantify sequence variability, we compared them at the functional sites of HIV-1 Gag proteins, including the p17, p24, and p1p7p2p6 regions. In this case, we are able to predict if the sequences generated by two methods are different on functional and immunologically relevant sites of HIV-1 Gag proteins, such as NAb epitopes, CTL eipitoped, T-help epitopes, particle assembly sites, protease cleavage sites, cyclophilin A binding sites, etc. It was found that the entropy scores at the functional sites of the p17 and p24 generated by the two methods are significantly different (p=0.0273 and p=0.0302, respectively) (Table 17).

Table 17 Entropy differences[a] on the functional sites[b] of HIV-1 Gag proteins

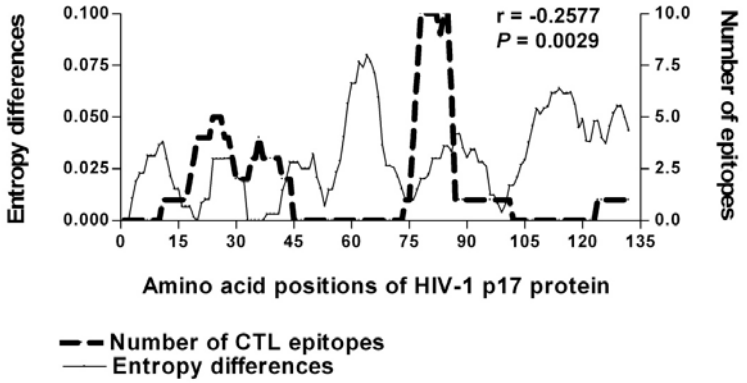
Gene	df[c]	Mean	95% CI[d]	*P*-value[e]
P17	75	1.05×10^{-2}	1.2×10^{-3} to 1.98×10^{-2}	**0.0273**
P24	192	7.10×10^{-3}	7.0×10^{-4} to 1.36×10^{-2}	**0.0302**
P1p7p2p6	78	7.70×10^{-3}	4.0×10^{-3} to 1.94×10^{-2}	0.1932

[a] Shannon entropy is calculated as a measure of variations in protein sequence alignments based on the method online (http://www.hiv.lanl.gov/, Los Alomas National Lab). Shannon differences were determined based on the entropy scores on the functional sites of HIV-1 Gag proteins derived from 454 and Sanger clone-based sequence populations;
[b] Gag protein functional sites are referred to HIV Sequence Databases and HIV Immunology Database (Los Alamos National Laboratory, USA);
[c] df: degrees of freedom;
[d] CI: confidence interval;
[e] Paired *t* test was conducted. *P* value with significance is highlighted.

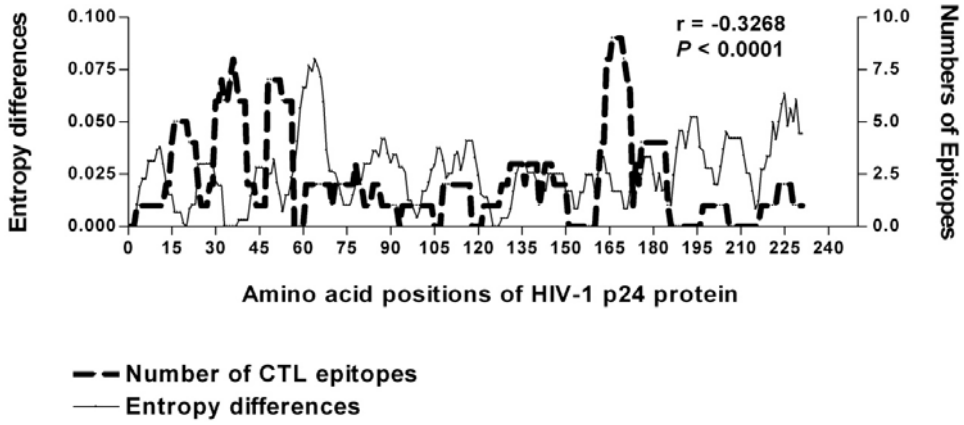
To further confirm this result, we correlated sequence variability with CTL responses. We found that sequence variability was negatively correlated with the CTL epitope density for both the p17 (r = -0.2577, p = 0.0029) and p24 regions (r = -0.3268, p < 0.0001). However, we observed no significant difference in entropy scores at the p1p7p2p6 region (Figure 15).

Figure 15 Correlation between entropy differences and epitope density in HIV-1 Gag.

(A).



(B).



Gag protein sequence positions are referred to HXB2. (A). Correlation between entropy differences and epitope density for p17. (B). Correlation between entropy differences and epitope density for p24. r: Spearman's correlation coefficient; *p*: Spearman's *p* value.

At the individual patient level, the non-synonymous variations between the two methods were also identified in 26 of 70 patients (25.7%) at 52 functional sites of HIV-1 Gag proteins (Table. 18). Of them, 50% (26/52) of them are CTL epitopes and 34.6% (18/52) are neutralizing antibody epitopes. The rest (8/52, 15.4%) include 4 cyclophilin A binding sites, 2 virus particle formatting, and 1 viral encapsulating site (Table.18).

Table 18 Consensus differences of HIV-1 Gag proteins between 454 and Sanger cloning sequences overlapped with the functional sites in individuals

ID[a]	Clade	HLA type	Epitope Sequence[b]	Function[c]
ML1111	A1	N/A	YSVHQRIDVKDTKEALEKIEEEQN(N/K) KSKKKA(T/P) (86-115)	NEUTRALIZING VIRUSES
		N/A	DTGNSS(S/N)QVSQNY (121-32)	NEUTRALIZING VIRUSES
		B*5703	K(K/R)AFS(N/S)PEVI (162-9)	CTL RESPONSES
		B*5703	K(K/R)AFS(N/S)PEVIPMF (162-72)	CTL RESPONSES
		B*4201	T(T/I)PQDLNTNL(180-8)	CTL RESPONSES
		N/A	PVH(Q/H)AGPIA (217-24)	CYCLOPHILIN A BINDING
ML1857	C	A*0201	SLY(Y/H)NT(T/A)VATL (77-85)	CTL RESPONSES
		B*3501	PPI(V/I)PVGD(E/D)IY (254-62)	CTL RESPONSES
		A*0201	FLGK(K/R)IWPS(P/S)Y(Y/H)K (433-42)	CTL RESPONSES
ML1992	A1	A*0801	EL(L/I)RSLYNTV (74-82)	CTL RESPONSES
		A*3002	RSLY(F/Y)NTVATLY (76-86)	CTL RESPONSES
		A*0201	SLY(F/Y)NTVATL (77-85)	CTL RESPONSES
		N/A	YSVHQRIDVKDTKEALEKIEEE Q(Q/K)NKSKKKA (86-115)	NEUTRALIZING VIRUSES
		N/A	KKAQQA(E/A)A(T/A)A(T/A)D(D/A)T (113-22)	NEUTRALIZING VIRUSES
		N/A	D(D/A)TGH(S/N)SSQ(N/Q)VSQNY(121-32)	NEUTRALIZING VIRUSES
		A*0801	DCK(K/R)TILKAL (329-37)	CTL RESPONSES
		A*0201	FLGK(K/R)IWPSYK (433-42)	CTL RESPONSES
ML1876	A1	A*2602	KYKL(L/M)KHI(I/L)VW (28-36)	CTL RESPONSES
		A*0202	SLY(Y/F)NTVATL (77-85)	CTL RESPONSES
		N/A	KKAQQA(A/E)A(A/T)A(A/T)D(A/D)T (113-22)	NEUTRALIZING VIRUSES
		N/A	D(A/D)TGH(S/N)SSQ(Q/N)VSQNY (121-32)	NEUTRALIZING VIRUSES
ML0795	A1	A*0202	SLYNTV(V/I)AT(T/V)L (77-85)	CTL RESPONSES
		N/A	CFNC(C/Y)GKEGHLARNC (392-407)	VIRAL ENCAPSIDATION
ML1003	A1	B*5301	QASQE(D/E)VKN(N/C)W (308-16)	CTL RESPONSES
ML1102	D	N/A	L(I/L)->X (64)	PARTICLE FORMATION
		N/A	KKAQQAA(T/A)ADT (113-21)	NEUTRALIZING VIRUSES
		N/A	DTG(R/G)H(N/H)SSQVSQNY (121-32)	NEUTRALIZING VIRUSES
		A*0201	FLGKIWPSY(H/Y)K (433-42)	CTL RESPONSES
ML1317	D	A*0201	FLGKIWPSY(H/Y)K (433-42)	CTL RESPONSES
ML1208	A1	N/A	L(I/L)->X (64)	PARTICLE FORMATION
		A*0201	SLYNTVATL(L/I) (77-85)	CTL RESPONSES
		N/A	DTGHSSQ(Q/K)VSQNY (121-32)	NEUTRALIZINGVIRUSES
ML1591	C	A*0201	SLYNTVAT(T/V)L (77-85)	CTL RESPONSES
		A*0201	FLGKIWPSY(N/H)K (433-42)	CTL RESPONSES
ML1660	D	A*2402	KYK(K/R)LKHIVW (28-36)	CTL RESPONSES
		N/A	YSVHQ(Q/E)R(R/K)ID(E/K)V(I/V)K(K/A) DTKEALEKIEEEQN(N/T)KSKKKA (86-115)	NEUTRALIZING VIRUSES
		N/A	KKAQQAA(A/T)ADT (113-22)	NEUTRALIZINGVIRUSES
		N/A	DTG(G/R)H(H/N)SSQVSQNY (121-32)	NEUTRALIZINGVIRUSES
ML1739	A1	N/A	KKAQQAAAD(D/G)T (113-22)	NEUTRALIZINGVIRUSES
ML0157	A1	A*0802	TPQDLNT(P/M)ML (180-8)	CTL RESPONSES
		N/A	PVHAQPIA(A/P)P (217-26)	CYCLOPHILIN A BINDING
ML0415	A1	A*0301	RLRPGGKKK(K/Q) (20-8)	CTL RESPONSES
		A*0301	RLRPGGKKK(K/Q)Y (20-9)	CTL RESPONSES
ML0548	A1	B*57	A(A/G)->G (248)	B*57 ESCAPING(Leslie et al. 2004)
ML1594	A1	N/A	EKIR(E/R)LR (17-22)	NEUTRALIZING VIRUSES(Papsidero et al. 1989)
		B*0801	GGK(R/K)KK(K/T)YK(K/R)L(M/L)K (24-32)	CTL RESPONSES
		B*0801	ELRSLYNT(T/A)V (74-82)	CTL RESPONSES
		N/A	DT(T/A)GH(H/S)SS(S/K)Q(K/Q)VSQNY (121-32)	NEUTRALIZING VIRUSES
		N/A	PVH(P/H)AQPI(V/I)AP (217-25)	CYCLOPHILIN A BINDING
ML1654	A1	N/A	DT(A/T)GH(S/N)SS(K/S)Q(Q/K)VSQNY (121-32)	NEUTRALIZING VIRUSES
ML1768	A1	N/A	D(A/D)TGHSSQ(Q/K)V(V/I)SQNY(121-32)	NEUTRALIZING VIRUSES
		N/A	PVHAGPI(I/A)AP (217-25)	CYCLOPHILIN A BINDING

[a] Patient IDs were from a cohort of the Pumwani Sex Worker in Nairobi, Kenya;
[b] Epitopes are derived from the best-defined CTL epitope summary (HIV Molecular Immunology 2006, Los Alamos National Laboratory, USA); Epitope positions were referred to HXB2 Gag proteins; Consensus differences were boldfaced;
[c] CTL: cytotoxic T lymphocyte response.

3. IV.4 Impact of variations on studying evolution of HIV-1 Gag

HIV-1 sequences have widely been used in molecular evolutionary studies of virus, potentially measuring selection pressures of host immune responses. Since molecular evolutionary studies on HIV-1 are completely based on its sequences, the differences of pyrosequencing-based sequences and Sanger clone-based sequences could affect our evolutionary study of HIV-1. We used four approaches to analyze the selection pressures on HIV-1 Gag to determine whether the sequence variant differences between pyrosequencing and Sanger clone-based methods influence PS selection analysis. First, we compared positively selected amino acids in Gag. Of 33 positively selected amino acids identified by the four different approaches, only two of them were identified by all four approaches and nine of them were identified by similar approaches in two sequence populations. However, the others were identified by different approaches and codons 75, 390, 474, 487 displayed a big difference in terms of the possibility of positive selection identification by any approach in two sequence populations (Table 19). Codons 75 and 487 are located in CTL epitope regions and codon 75 mutation in a CTL epitope was reported to be associated with HLA class I allele A*02012.

Table 19 the comparison of PS sites identified from 454[a] and Sanger cloning sequence populations

	454 sequence population				*Sanger Clone-based sequence population*				
Codon	Quasi [b]	SLAC[c]	FEL[d]	IFEL[e]	Quasi	SLAC	FEL	IFEL	Difference[f]
049	Positive	**0.200**	**0.130**	0.030	Positive	**0.250**	**0.240**	**0.600**	1/4
054	Positive	0.001	0.001	0.010	Positive	0.001	0.001	0.008	0
062	Positive	0.050	0.040	**0.190**	Positive	0.030	0.020	**0.080**	0
066	Positive	0.005	0.003	**0.290**	Positive	0.007	0.001	0.020	1/4
069	Positive	**0.130**	0.020	**0.140**	Positive	**0.190**	**0.050**	**0.260**	1/4
072	Positive	**0.080**	0.040	**0.170**	Positive	**0.080**	0.040	**0.130**	0
075	Positive	**0.150**	**0.280**	**0.460**	Positive	0.020	0.030	**0.280**	2/4
076	Positive	**0.080**	**0.110**	**0.050**	Positive	**0.060**	0.030	**0.070**	1/4
122	Positive	0.030	0.020	**1.000**	Positive	**0.080**	0.030	**0.150**	1/4
124	Positive	**0.310**	**0.230**	**0.270**	Positive	**0.260**	**0.170**	0.040	1/4
127	Positive	**0.270**	**0.120**	**0.200**	Positive	**0.140**	0.040	**0.340**	1/4
143	Positive	**0.090**	0.040	0.040	Positive	0.050	0.020	**0.130**	0
146	Positive	**0.100**	0.030	**0.110**	Positive	**0.060**	0.010	**0.150**	0
147	**Negative**	**0.200**	**0.080**	0.030	**Negative**	**0.150**	0.040	**0.100**	0
223	Positive	0.020	0.020	**0.170**	Positive	0.010	0.020	**0.780**	0
243	Positive	**0.070**	**0.250**	**0.560**	Positive	0.030	**0.100**	**0.680**	1/4
248	Positive	**0.070**	**0.140**	0.030	Positive	0.010	0.020	**0.100**	1/4
260	Positive	**0.160**	**0.060**	0.050	Positive	**0.140**	**0.060**	**0.120**	1/4
301	Positive	**0.450**	**0.240**	0.030	Positive	**0.340**	**0.170**	**0.070**	1/4
303	**Negative**	0.050	**0.110**	**0.580**	**Negative**	0.020	**0.060**	**0.750**	0
339	Positive	0.001	0.001	0.001	Positive	0.001	0.001	0.001	0
373	Positive	0.010	0.070	**0.760**	Positive	0.040	**0.050**	**0.750**	1/4
390	Positive	**0.080**	0.010	0.040	**Negative**	**0.470**	**0.260**	**0.190**	3/4
401	Positive	**0.150**	**0.280**	0.030	Positive	**0.340**	**0.510**	**0.070**	1/4
441	Positive	0.004	0.001	0.009	**Negative**	0.020	0.008	**0.080**	2/4
451	Positive	**0.050**	0.010	0.030	**Negative**	**0.080**	0.020	0.010	1/4
466	Positive	0.040	0.006	**0.120**	**Negative**	0.040	0.006	0.030	0
474	Positive	**0.110**	**0.060**	**0.050**	Positive	0.030	0.009	0.020	3/4
481	Positive	0.030	0.007	0.020	**Negative**	0.010	0.004	**0.140**	2/4
483	Positive	0.020	0.003	**0.110**	Positive	0.040	0.002	**0.140**	0
486	Positive	0.003	0.003	**0.500**	**Negative**	0.001	0.001	**0.230**	1/4
487	Positive	0.030	0.020	0.006	**Negative**	**0.130**	**0.200**	0.040	3/4
498	Positive	0.030	0.005	**0.060**	Positive	0.010	0.002	0.050	1/4

Positive selection sites were identified from two sequence populations: 454 and Sanger clone-based sequences; Bold values are deemed significant ($p < 0.05$). The possibly differentially positively selected sites between two sequence populations are highlighted in rows.

[a] 454 Life Science technology;

[b] Quasi analysis results;

[c:] Single Likelihood Ancestor Counting;

[d] Fixed Effects Likelihood;

[e] Internal Fixed Effects Likelihood;

[f] Difference of possibility of identification of positive selection sites by different approaches between pyrosequencing and Sanger clone-based sequence populations

4.0 Discussion

4.1.0 Evolution and Characterization of HIV-1 *Env* Gene

4.1.1 Discussion

Cross-clade immune responses to HIV-1 can be achieved by exposure to a single virus (Geels et al. 2005; Keating et al. 2002; McKinnon et al. 2005; Ross and Rodrigo 2002; Wainberg 2004; Wei et al. 2003) because it is the details of epitope and epitope context, not overall sequence similarities that define the specificity of host immune responses. Designing epitope-based vaccines against HIV has become one of the major strategies to combat the growing pandemic (Slobod et al. 2005). However, the results obtained from vaccines using any approach have not been encouraging: one reason for this is that HIV mutates rapidly and we might not know enough, especially at the population level, about how the virus evolves in the face of host immune responses and pathogenesis. Without this knowledge, it is unlikely we will be able to determine the best set of immunogens to induce strong cross-reactive protective immune responses (Gaschen et al. 2002; Herbeck et al. 2006; Mullins et al. 2004). Based on the assumption that host immune responses are the major selective pressures on HIV-1, we conducted a comprehensive analysis of all available Env sequences in an effort to gain an overview of host immunological pressure on the evolution of HIV-1 at a population level in a global context.

We found that PS sites were dispersed across the entire Env and the density of PS sites is much higher than previous reported (Fig 3) (de Oliveira et al. 2004; Travers et al. 2005; Yamaguchi-Kabata and Gojobori 2000). These results likely reflect the diversity of sequences from different clades, and geographic regions, ethnicity of individuals infected

over time in the population and demonstrates the power of our approach. Our finding that there are more PS sites in variable regions than in conserved regions (Fig 3) might be explained as follows: the variable regions of envelope are thought to form "surface-exposed loops" and be accessible to antigens, especially neutralizing antibodies, as the conserved regions are responsible for interacting with viral receptors and gp41 ectodomain (Burton et al. 2004; Kwong et al. 1998). Thus antibody responses to HIV-1 Env may play a more important role in the host immune responses against HIV-1. On the other hand, we also found that PS sites are in the conserved regions of Env with CD4 binding sites under PS pressure. Although the critical contact residues of CD4 (Asp 368, Glu 370, and Trp 427) to gp120 are free from PS, many gp120 residues that are thought to contact CD4 are under PS. This result can be explained by the fact that half of gp120 residues contacting CD4 lie within the interfacial cavity and there is a large water filled interfacial cavity between CD4 and gp120. In consequence, the interfacial cavity would accommodate some sequence variability, yet retain the CD4 binding specificity (Kwong et al. 1998). Thus, the conserved regions are thought to form the functional core of gp120, which determine the capability of the virus to infect cells. These regions are also targeted by many neutralizing and non-neutralizing antibodies (Binley et al. 1998; Kwong et al. 1998; Maksiutov et al. 2002; Moore and Sodroski 1996; Pantophlet et al. 2003; Richman et al. 2003; Wei et al. 2003; Yang et al. 2000a), so it is not surprising that the conserved regions of gp120 are under intensive PS, causing replacing mutations at the population level to accumulate in these regions over time. This phenomenon has also been observed in other studies (Mohabatkar and Kar 2004; Yamaguchi-Kabata and Gojobori 2000; Yamaguchi-Kabata et al. 2004).

N-linked glycans comprise about 50% of the mass of gp120 (Leonard et al. 1990) and they are essential for the correct folding and correct processing of HIV-1 Env protein (Li et al. 1993). They have been proposed to form a "glycan shield", which limits Env recognition by neutralizing antibodies while maintaining the ability of Env protein to interact with CD4 and co-receptors (Huang et al. 1997; Wei et al. 2003). Our results showed that the N-linked glycan sites are relatively conserved at a population level, which is consistent with a previous report (Louwagie et al. 1995). Despite intensive PS pressures from the host, the pattern of N-linked glycan sites does not change except in some of the variable regions, especially the V3 region (Fig 3). In V3 regions, N295 and N332 form the core sites for recognizing the global neutralizing antibody, 2G12 (Sanders et al. 2002; Scanlan et al. 2002) and the removal of N-linked glycan sites in V3 may confer a critical change to the conformation of the Env and thus reduce the accessibility and susceptibility of virus to neutralizing antibodies elicited by N-linked glycans. Thus, the lack of PS on the majority of N-linked glycan sites is likely the result of the protective effect of N-linked glycans against neutralizing antibodies. An alternative or complementary explanation is that the conservation of these regions is so critical for Env structure and function of the virus that any modifications in these regions would be lethal.

An effective HIV vaccine would need to generate cross-reactive immune responses regardless of infecting clade and geographic regions (Gaschen et al. 2002; Slobod et al. 2005). To test the possibility that the potential immunologically targeted sites could be identified in all major subtypes, we compared the PS patterns among four major subtypes

A, B, C, and D. We found that clade A, B, C, and D shared at least 25% of PS sites (Table.5), indicating host immune responses may be targeting the same regions of Env regardless of subtypes. Our results are consistent with previous studies (Choisy et al. 2004), and suggest that it is possible to identify potential immunologically targeted sites, and cross-clade immunity. We also noticed the differences in PS sites distribution among clade A, B, C, and D (Table.4). These differences may reflect host factors affecting PS such as the differences in population HLA allele frequencies in the region where a given subtype circulates and degree of diversity in a given clade (Choisy et al. 2004).

We investigated the trend of PS pressure over time at a population level as the ability to predict HIV-1 evolution is essential for the design of effective HIV vaccines. Although PS pressure may not be continuous in the course of the evolution of HIV-1 Env, the build-up of pressures in terms of accumulated replacement mutations over time could still provide information on the evolutionary direction of HIV-1 Env. The PS pressure was very stable over time when the entire collection of Env sequences was analyzed. For clade B, however, there is a trend that the PS increased from 1989 (19.7%) to 1999 (25%) (Figure 5). Since host immune responses are the driving force of PS of Env, the minor increase of PS sites over time for clade B may reflect increased diversity of the infected host population over time. As the overall diversity of host population in the world is stable and based on our analysis so is the immune pressure on the HIV-1 Env, an HIV vaccine is an achievable goal if we can understand how host population-based responses influence HIV-1 diversity and evolution.

To determine which host immune responses are strongly associated with PS, we correlated previously described host immune responses and factors affecting these responses (Ab, CTL, proteasomal processing and T helper cells) with PS pressure at a population level. A number of factors could influence this analysis, such as the dynamics of immune response in the course of evolution (da Silva 2003), recombination (Rambaut et al. 2004), individual differences in HLA haplotypes and the fact that only small subsets of epitopes have been defined (Leslie et al. 2005; Moore et al. 2002), the unknown HLA and Env sequence correlation at an individual level, and Ab recognition of non-linear epitopes. In recognizing these limitations, we explored the correlations between PS and host immune responses. We assumed that HIV-1 evolves continuously, and that the accumulated genetic divergence is maintained through transmission events regardless of the time each viral lineage remains within an individual. We also excluded recombined Env sequences from the study to avoid confounding the evolutionary history of analyzed sequences. We found that neutralizing antibody responses are significantly associated with PS pressure (P=0.0019; Table.6) and that T helper responses appears sufficient to account for the PS pressure (P=0.0002; Table.6). However, in a multivariate model, no interaction was identified between Th responses and others, including NAb. Th responses, whose main function is thought to modulate other cells involved in host immunity, including CTL response and Ab responses (Imami et al. 2001; Jansen et al. 2006), are thought to be incapable on their own of exerting such strong anti-viral effects. The question that arises here may be due to poorly defined Th epitopes in the Los Alamos HIV Sequence Database, in which Th epitopes cover the majority of HIV-1 Env sequence and overlap with many other epitopes, thus interfere with our statistical analysis.

Moreover, the neutralizing antibodies' response does not work alone; they interact with CTL responses and PCS activity to influence the evolution of HIV-1 Env. These results however strongly suggest that neutralizing antibody responses is one of the major forces driving the evolution of HIV-1 Env. The interactions observed in NAb*Th, Th*CTL, PCS*CTL, and Th*CTL*PCS of Env may change if the analysis is performed on additional viral proteins. Evolution of a predominant HIV-1 sequence in an individual may involve the interactions of immunity generated to the whole virus and immunity generated to Env could be the results of interactive immunity generated to other viral proteins. PS-based analysis of other viral proteins is necessary to fully understand host immunity to HIV-1 infection. With accumulation of the well defined epitopes and further classifications of epitopes such as Th1/2, and MHC class I/II restricted PSC, PS-based analysis could become a more robust, powerful tool, and generates more precise and reliable results to better understand host immunity to HIV-1 infection.

Previous molecular evolution studies have only analyzed a small number of Env sequences (Choisy et al. 2004), our analysis included all available Env sequences of different time periods, geographic region and subtypes. The PS pressure in our analysis covers the accumulation of mutations over time in different populations, which make a notable impact on detecting PS pressure (Anisimova et al. 2003; Herbeck et al. 2006; Shriner et al. 2003). In addition, the analysis of large amount of sequences collected over time increases statistical power of identifying PS sites under long-term recurrent selective pressure, which would not be detected with a small number of sequences.

A recently developed novel approach for designing vaccine is to construct a "centralized" sequence (minimizing genetic differences of virus) by computational methods (Gao et al. 2004; Gaschen et al. 2002; Nickle et al. 2003). The "centralized" sequences can be ancestral, consensus, and centre-of-tree sequences, which best represent the pool of variants, thus reducing the genetic variance of sequences from circulating viruses (Gao et al. 2004). The identified "immunogens" are then tested and confirmed by experimental methods. These methods, however, do not consider how host immune responses interact with "centralized" sequences, lessening their value. We have adduced information that could remedy this deficiency by identifying relatively "conserved" regions of Env that are free from PS (Figure 3). These regions could be potential immunogens that take into account both "centralized sequences" and the effects of immunological pressures from the host. So far, there are few previous publications that directly support this idea. A number of CTL and neutralizing antibody epitopes were reported to elicit strong immune responses which are correlated to either rapid progression to AIDS or reduced transmission of HIV-1 (Table. 7). The LWVTVYYGVPVWK epitope was reported to be recognized by the neutralizing antibody, 4E10 (Hager-Braun et al. 2006). While, LFCASDAK and RIRQGGLERA epitopes were restricted by HLA A*0301 and A*0205, respectively and have been associated with reduced HIV transmission and Long Term Non-Progression (Liu et al. 2003; Propato et al. 2001). Since these epitopes are conserved across clades and in conserved geographic regions in terms of PS, they may be good target for vaccines. The regions restricted by B*07 allele and its affiliated supertype B7 may elicit detrimental immune responses, as these alleles have been reported by many studies to associate with disease progression (Gao et al. 2005; Scherer et al. 2004;

Trachtenberg et al. 2003), and should therefore not be used for HIV-1 vaccines. The association of A29 and others with HIV-1 disease has not been well defined and their usefulness in the HIV-1 vaccine requires further study. Clearly, more work is needed to fully characterize the relationships between epitopes identified by this and other methods, their roles in HIV specific immune responses and clinical outcomes.

Ideally, an effective HIV-1 vaccine should elicit both protective humoral and cellular immune responses. Among HIV-1 viral proteins, only Env meets this requirement. It is the major protein on the surface of HIV-1 and mediates viral host entry. Env elicits not only neutralizing antibody responses (Richman et al. 2003) to reduce transmission/or limit primary infection but also cytotoxic T-cell lymphocyte (CTL) responses (Stanhope et al. 1993) to control the established infection. While other viral proteins, such as RT, Gag, and Nef, only elicit CTLs responses. Since our analysis is based on the HIV-1 sequences from HIV-1 infected individuals, the results might not reflect vaccine-like immunity generated in HIV-1 naïve subjects or ones who are highly exposed but resistant to HIV-1 infection. However, the PS-based analysis could be valuable once HIV sequence databases are more defined and the sequences of HIV-1 infected Rapid-Progressors and Long-Term Non-Progressors become available. Thus, PS-based analysis of HIV-1 sequences is one of the approaches to understand what types of anti-HIV-1 immunity (NAb, CD4+, CTL, TH, NK cells, etc) may be required for an effective HIV-1 vaccine.

4.1.2 Summary

This is the first study that extensively and systematically investigates PS pressure on the full HIV-1 Env using all available sequences. We found that PS pressure was widely dispersed across HIV-1 Env. The PS pressure on Env remains relatively stable at the population level over time. Moreover, neutralizing antibody responses are the major host immune responses that influence PS of HIV-1 Env. We also identified stretches of sequence rarely targeted by PS, and known epitopes eliciting strong cross-reactive protective humoral and cellular immune responses against HIV-1 Env in these regions. These findings will help in the design of more effective HIV-1 vaccines.

4.2.0 Evolution and Characterization of HIV-1 *Gag* Gene

4.2.1 Discussion

A better understanding of host immune responses at a population level is very important for an effective HIV vaccine. Currently, there is a focus on developing a T cell based HIV vaccine as more evidence accumulate on the key role of the CD8+ CTL responses in containing HIV infection (Korber et al. 2009). HIV Gag as one of major structure proteins of HIV is commonly targeted for a HIV vaccine (Johnston and Fauci 2007). In order to provide more information on both the extent to which host immune responses contribute to the diversity of HIV Gag and immunogenic potential of HIV Gag at a population level, we systematically conducted a QUASI analysis of all available *gag* sequences on the assumption that the major selection pressures are host cellular immune responses.

We found that PS sites were distributed across the whole Gag proteins with the lowest density of PS sites on p24 (14.29%). These results are the similar to previous reports (Addo et al. 2003; Peters et al. 2008). Since p24 Gag is highly immunogenic (elicits strong host immune responses), the low density of PS sites supports the hypothesis that p24 Gag escape confers significant fitness cost to the viruses and is not well tolerated by sequence variation (Payne et al. 2009). As described in the previous study on *env* (Liang et al. 2008), our analysis included all available *gag* sequences in Los Alamos HIV Sequence Database over 29 years time period, geographic regions, and subtypes. Therefore, our study has increased statistical power of measuring selection pressure presumably by the host immune response pressure. Our results would extend our

knowledge on how host immune responses operate on Gag and how host and viruses interact at a population level.

We performed QUASI analysis on the clade A *gag* sequences from Los Alamos Sequence Database and compared the identified PS sites to ones from a previous study on the clade A sequences from Pumwani Sex Worker Cohort (Peters et al. 2008).It was found that the majority of PS sites identified from Los Alamos *gag* sequences matched those detected from the previous study on *gag* sequences from Pumwani Sex Worker Cohort (Figure 8). Over 80% of the identified PS sites across Gag correlated to HLA class I alleles from the study in Pumwani Sex Worker Cohort were also matched to ones from QUASI analysis of Los Alamos *gag* sequences (Table 11). For p24, p1, p6, and p7 proteins, the percentage of this match is 100%. These results demonstrate that a high percentage of the identified PS site match between two populations. When the PS sites associated with lower CD4+ counts identified from Pumwani Sex Worker Cohort were searched from PS sites list from Los Alamos sequence database, high percentage of match (75%) was also found. Since this study took many factors, such as HLA allele restriction, CD4+ counts, and Gag proteins, into account in a comparison of PS sites between two populations, the high percentage of PS site match is seemingly not at random but more likely reflects a link between host and viruses.

To further confirm this hypothesis, we conducted a statistical analysis to exam the possibility that the PS site matched between the two populations. Thirty PS sites were generated at random from the PS site pool derived from Los Alamos *gag* sequences and

then were compared to those derived from the *gag* sequences of Pumwani Sex Worker Cohort. The possibility of five PS site matches between two populations is less than 0.05 (Figure 9). This result suggests that the identified PS sites on Gag proteins in our study are able to accurately indicate where host immune response pressures are exerted on Gag proteins at a population level. On the other hand, the estimation of PS on Gag proteins allows us to interpret the relationship between host factors such as HLA class I molecules and CTL responses and PS at a global scale.

An ideal T cell-based HIV vaccine should generate cross-reactive CTL responses regardless of infected clades and geographic regions. However, the considerable genetic and antigenic variation of HIV has become a major obstacle. Fortunately, many studies showed that cross-reactive immune responses exist (Coplan et al. 2005; De Groot et al. 2003; Ferrari et al. 1997; Geels et al. 2005; Gudmundsdotter et al. 2008). These cross-reactive epitopes appear to be able to tolerate some substitutions within epitopes other than at anchor positions (Malhotra et al. 2007). If this hypothesis held, our identification of cross-reactive epitopes according to amino acid identity would mislead our understanding of cross-reactive immunity at a population level. Also, a small numbers of sequences in previous studies limited our study power on the evolutionary process of HIV. To overcome this problem, we compared the PS sites derived from the entire sequence populations across major circulating clades A, B, and C in the world. In this case, we are able to identify and compare the patterns of antigenic epitopes cross subtypes. We found that more than 23 % of PS sites were matched at the same locations on clade A, B, and C Gag proteins, including p17, p24, p6, and p7 proteins.

For p24, around 40% PS sites were shared by clade A, B, and C. These observations demonstrate that PS pressures operate on the similar regions of clade A, B, and C Gag proteins (including p24), implying that, to some extent, the major circulating clades of HIV-1 possibly share similar antigenic immunogens such as epitopes. This result raises the possibility that cross-clade immunogens could be identified for a global HIV vaccine design. More experimental studies are necessary to screen and test cross-clade immunogenic epitopes at a population level.

As mentioned above, the assumption of our study is that host immune pressure is the major selection pressure. In fact, the factors involved in the processing of CTL epitopes were also shown to play crucial roles in cellular immune responses. These factors include proteasomal cleavage and HLA class I & II molecules (Payne et al. 2009; Peters et al. 2008). In order to determine which host immune responses are associated with PS, we correlated previously described host immune responses such as antibody (Ab), neutralizing antibody (NAb), CTL responses, and proteasomal processing with PS pressure at a population level. Many factors could affect our current study, including dynamics of immune responses through evolution of HIV-1, recombination, HLA allele frequencies, and non-linear Ab epitopes (Liang et al. 2008). On the assumption that HIV-1 evolves continuously, and that genetic variations are maintained through transmission in new hosts, we are able to infer a correlation between immune responses and PS as we did on Env (Liang et al. 2008), However, our study did not identify a significant association between CTL responses alone and PS ($p = 0.5771$). In contrast, many studies showed that CTL response driven CTL escape mutations on Gag sufficiently resulted in

loss of immune control and thus CTL responses were thought to predominantly drive evolution of HIV-1 in individuals (Payne et al. 2009). Our result implies that other factors may have an impact. We conducted an association study in a multivariate model and found a significant association between CTL & Ab responses and PS (p = 0.0373). This result supports the idea that CTL responses do not work alone but interact with others to influence the evolution of HIV-1 Gag at population level. Since Ab responses can not neutralize viruses, its role needs further clarification by experimental methods. Alternatively, the limited numbers of defined CTL, Ab, and NAb epitopes from Los Alamos Immunological Database could affect the results of our study. In addition, it was reported that HLA class I & II molecules imposed significant immune selective pressure on CTL epitopes, suggesting that the interaction of HLA and CTL responses drives evolution of HIV (Payne et al. 2009). In our study, HLA allele information was missing for many *gag* sequences and thus was not applied at a population level, which could hamper the ability of our approach to detect associations between CTL& HLA and PS at a population level.

We also investigated the immunogenic regions potential of eliciting cross-clade T cell responses and conserved irrespective of disease stage and antiretroviral treatment at a population level. Cross-clade immune responses have been widely reported at both individual and population level (De Groot et al. 2003; Gudmundsdotter et al. 2008; Liang et al. 2008; Malhotra et al. 2007; McKinnon et al. 2005). We searched the PS-free regions for previously identified immunogenic epitopes, including NAb and CTL epitopes. Eighteen CTL epitopes were identified, with 50% of them associated with the reduced

disease progression in previous reports (Table 10). Although the majority of these epitopes were tested in clade B, they are relatively conserved in evolutionary process of whole HIV-1 viruses for over 29 years and possibly could be recognized in patients with other clades of HIV-1. One example is GQMREPRGSDI which can be recognized by either clade B or C. Further experiments are necessary to test our hypothesis. In contrast, the centralized sequence (Gaschen et al. 2002) and polyvalent mosaic protein strategies (Fischer et al. 2007) focus on minimizing HIV diversity to increase coverage of potential T-cell epitopes. The disadvantages of these strategies are artificial sequence or increased numbers of sequences (increased cost) in vaccine design. Furthermore, of the variants within T-cell epitopes (other anchor positions) may not confer fitness cost to viruses. Thus, increased variants may not raise coverage of potential T-cell epitopes. In our study, we took host immune response pressure into account and were able to monitor evolutionary process of HIV, especially epitopes at a population level. For those variants which do not confer fitness cost (not under PS), there is no need to be included to increase coverage of potential epitopes.

4.2.2 Summary

We systematically measured host immune response pressure on HIV Gag protein. We found that the highly immunogenic p24 was the most conserved at a population level. The similar antigenic patterns were found across clade of HIV. Moreover, CTL responses alone do not drive the evolution of HIV at a population level. In the end, we identified some potential CTL epitopes as candidates for HIV vaccine design.

4.3.0 Impact of HLA Class I Allele Frequencies on CTL responses

4.3.1 Discussion

MHC class I loci (HLA-A, HLA-B, and HLA-C) in humans are essential to host immune responses since they encode molecules to present peptides to CD8+ T-cells. The importance of this is that some HLA class I molecules in host immune responses against infectious pathogens are shown by their association with differential disease outcomes. Among HLA class I alleles, HLA-B*1302, 2705, 5101, 5701, 5702, 5703, 5801, and 8101 have been reported to be associated with slower HIV disease progression and lower viral set-point while HLA-B*1801,3502, 3503, and 5802 were associated with rapid disease progression and the high viral set-point (Kiepiela et al. 2004; Lazaryan et al. 2006; Serwanga et al. 2009; Tang et al. 2002).

In order to explore the impact of HLA class I molecules on immune control of HIV at a population level, we first investigated the distribution of HLA class I alleles restricting 183 optimal epitopes across different HIV-1 proteins. We found that significantly more HLA class I alleles restricted epitopes on accessory/regulatory and *gag* gene products in comparison with Pol, and Gp160 (Figure 10). This observation supports the results of a previous report in which the greater fitness cost for escape within CTL epitopes was associated with HLA class I alleles with epitopes in accessory/regulatory and *gag* gene products (Rousseau et al. 2008). It suggests that HLA class I allele frequency might influence the evolution of CTL epitopes on Gag and accessory/regulatory proteins which are the likely major component candidates of an HIV-1 vaccine. If the variants escape from the host CTL responses on those proteins at population level, it would provide

important implications for HIV-1 vaccine design. In fact, in subtype C of a South Africa population, it was found that specific HIV-1 residues such as Gag (position 120, 184, 242, and 339) and Vif (position 725) restricted by HLA class I allele B*4201, B8*1, B*57, B*5801, B*08 and B*1503 were targets of the CTL responses and associated with low viral loads (Rousseau et al. 2008).

We also explored the distribution of HLA class I alleles (A, B, and C) restrictions among HIV-1 proteins. We found that many more HLA-B allele (54.78%) restricted the optimal epitopes than alleles of HLA-A (36%) and HLA-C (9.1%). Furthermore, we found that many more HLA-B alleles restricted optimal epitopes on Gag and accessory/regulatory proteins (Table 12). These observations are in line with a previous study showing that HLA B-restricted CTL epitopes were targeted more frequently and that HLA-B possibly played a dominant role in mediating evolution of HIV-1(Kiepiela et al. 2004). Our results provide a population-based approach to show the dominant influence of HLA-B on evolution of HIV-1. In addition, we found 25 HLA class I molecules restricting multiple CTL epitopes and 8 CTL epitopes restricted by multiple HLA class I molecules. This phenomenon was also shown in a previous study (Brumme et al. 2007). The role of these HLA class I molecules on mediating CTL responses has not been addressed by previous studies and requires further investigation.

In order to explore how HLA class I-restricted immune responses drive evolution of HIV-1 at population level, we correlated HLA class I allele frequency with host CTL response. Across the entire HIV genome, no correlation was identified between HLA class I allele

frequency and host CTL response. This result is possibly due to confounding factors such as HLA class I type, HIV genes where CTL epitopes are located, and how many HLA alleles restrict CTL epitopes. We removed the confounding factors from the analysis and identified the association of HLA class I allele frequency with host CTL response on Accessory/Regulatory proteins of HIV-1.

There are several limitations to this study. First, the numbers of optimal CTL epitopes listed in the HIV database are limited to the currently identified epitopes and the corresponding HLA alleles. Therefore, the results are limited to the epitopes and the HLA alleles analyzed and do not covers all the effect of host immune responses on the evolution of all HIV-1 genes or all sequences of a given HIV-1 gene. Second, host immune pressure exerted on each amino acid within CTL epitope is not equal. The determination of dN/dS ratio could not precisely measure host immune responses on CTL eiptopes. Finally, average HLA allele frequency may not accurately reflect the distribution of HLA allele frequency for a specific population. Those potential limitations are likely to influence the results of the analysis and possibly explain why no associations were identified on Gag, Env, Nef, Rev, and Pol proteins. The same result was also reported by other studies (Rousseau et al. 2008; Rousseau et al. 2009).

The classification of HLA class I supertypes increased general population coverage of CTL epitopes and the overall frequency of each supertype was relatively constant across different populations (Sette and Sidney 1999). In order to increase statistical power for the analysis, we grouped HLA alleles into supertypes based on their common motifs

(Sidney et al. 2008). We identified a significant association between the frequency of HLA–B7 supertype and selection of the corresponding CTL epitopes. This observation supports the hypothesis that HLA allele frequency has an impact on evolution of HIV-1. However, we did not observe the correlation between frequencies of other HLA supertypes with the CTL responses.

4.3.2 Summary

HLA class I allelic frequency in a population would be expected to have an impact on evolution of HIV-1 and different HLA alleles might have different influences on evolution of HIV-1. There are significantly higher numbers of HLA class I alleles targeting HIV-1 accessory/regulatory and Gag proteins. HLA-B restricted CTL epitopes are targeted more frequently. HLA class I alleles and their population frequencies should be taken into account for effective HIV-1 vaccine design.

4.4.0 Impact of 2nd-Generation Sequencing Technologies on the Sequence Analysis of HIV-1 Genes

4.4.1 Discussion

The genetic analysis of HIV-1, especially in estimating the diversity of quasispecieses and detecting mutations has traditionally been conducted by Cloning and subsequent Sanger dideoxy sequencing. However, polymerase induced sequence errors can confound results when sequencing cloned DNA. More importantly, the number of clones which can be affordably sequenced is unlikely to adequately represent the genetic variation of the viral population in a patient sample. More recent DNA sequencing approaches provide the potential to greatly reduce the cost, complexity, and time required to sequence DNA without the need for cloning and increase sensitivity of detecting variants (Margulies et al. 2005; Shendure and Ji 2008).

The Roche Genome sequencer GS20 (454) emerged as the first commercially available second-generation sequencer based on pyrosequencing technology (Margulies et al. 2005). A major advantage of the pyrosequencing technology is the ultra-deep coverage of target sequences (up to thousands), which enables researchers to estimate the genetic diversity of HIV-1 sequences, especially to accurately and quantitatively assess HIV quasispecies by detecting rare genetic variants in viral population such as rare drug resistant variants (Bushman et al. 2008; Rozera et al. 2009; Wang et al. 2007).

My study is based on the analysis of Sanger-based sequences. On assumption that 2nd-Generation Sequencing Technologies is superior to Sanger clone-based method

in characterizing genetic diversity of HIV-1 quasispecies, it is reasonable for us to hypothesize that 2nd-Generation Sequencing Technologies affects Sequence Analysis of HIV-1 Genes by detecting more variants.

To test this hypothesis, we compared the pyrosequencing with Sanger clone-based method in characterizing genetic diversity of the HIV-1 *gag* quasispecies from 96 patient samples and assessed the impact of pyrosequencing technology on studies of HIV-1 biology and evolution.

Our results showed that pyrosequencing technology triumphed over Sanger clone-based method in detecting low abundance viral variants, especially those with frequencies less than 2% in the population. In our study, the number of variants identified by pyrosequencing (14034) was more than 3 times of that identified by Sanger clone-based method (3632) (Table 16). Since the rate of GS20 sequence errors arising from upstream amplifications and other intrinsic platform properties were taken into account in determining minor variants in this study, the identified minor variants are unlikely to be false. Thus, these findings confirmed the advantage of applying this technology in detecting low abundant variants, especially the rare drug-resistant variants as previously reported (Hoffmann et al. 2007; Simen et al. 2009; Wang et al. 2007).

Several alternative methods have been developed before for detecting low abundant variants such as allele-specific sequencing (Cai et al. 2007; Halvas et al. 2006), heteroduplex tracking assay (HTA) (Delwart and Gordon 1997; Delwart et al. 1993;

Schnell et al. 2008), and single-genome amplification (SGA) (Palmer et al. 2005; Salazar-Gonzalez et al. 2008). All of them have the ability of detecting variants with frequencies in a range of 0.1 to 2 %. In comparison with pyrosequencing method, however, they are more complex and labor intensive. Moreover, allele-specific sequencing can only detect a small number of rare variants or investigate a fraction of interested sequences (Cai et al. 2007; Long et al. 2000; Ritola et al. 2004). SGA method is claimed to be able to accurately represent HIV-1 quasispecies and preclude *Taq*-inducted artifacts, template switching, resampling, and cloning bias which are produced in Sanger cloning sequencing (Salazar-Gonzalez et al. 2008). *Taq*-induced artifacts, template switching, and resampling have been reported during pyrosequencing [Dr. Wei Shao; personal communication]. These may explain the 642 sequence variants that were only determined by Sanger clone-based methods but not Pyrosequencing in our study. However, bioinformatics approaches can dramatically reduce the errors caused by these problems (Wang et al. 2007). For example, *Taq*-induced sequence errors can be corrected by statistical analysis. Therefore, pyrosequencing definitely is a better alternative for charactering genetic diversity of HIV-1, especially detecting minor variants in viral population. Our study shows that Sanger clone-based method underestimated the degree of HIV diversity/mutation. Consistent with a previous report, our analysis of sequences generated by pyrosequencing also showed that p24 is the least variant enriched region (Figure 14) (Yusim et al. 2002).

The difference of the detected sequence variants between pyrosequencing and Sanger clone-based method affected subsequent analysis of the Gag proteins. This can be shown by both the significant differences of the entropy s cores (Korber et al. 2000) on the functional sites of p17 (p = 0.0273) and p24 (p = 0.0302) and the significant correlation of amino acid variability (Shannon's entropy scores) between sequences generated by these two methods with the number of CTL epitope in p17 (r = -0.2577, p = 0.0029) and p24 (r = -0.3268, p < 0.0001) region (Figure 15). The negative correlation may due to "experimental selection effect" described previously (Yusim et al. 2002). Therefore, the impact of amino acid variations/ differences between pyrosequencing and Sanger clone-based method on functional analysis of HIV-1 Gag proteins at the population level should not be ignored.

In addition, our study showed that the sequence difference generated by these two methods also affected the HIV-1 Gag consensus at individual level. Especially, in 26.1% of the patients, the difference of Gag consensus generated by these two methods overlaps with the functional sites involving viral replication, assembling, packaging, CTL epitopes, neutralizing antibody epitopes, and cyclophilin A binding site (Table.18) and 50% of them are within CTL epitopes (Casement et al. 1995; McKinnon et al. 2005). Since these analyses did not include amino acid variants with frequencies less than 50%, the amino acid difference of HIV-1 Gag protein sequences generated by the two methods could well be underestimated. It was shown that CTL epitopes with high functional avidity rapidly select for escaping mutations, resulting in a low abundant variants which can be recognized by CTLs and elicit strong host immune responses (O'Connor et al. 2002;

Slifka and Whitton 2001). Thus, further experiments are necessary to examine CTL responses in these individuals.

The selection pressures by host immune responses shaped genetic variation of HIV-1 (Moore et al. 2002; Ross and Rodrigo 2002; Williamson 2003; Yang et al. 2003; Yang 2001; Yang et al. 2000b). Measuring and understanding the selection pressures is an important part of evolutionary biology (Bush 2001; Liang et al. 2008; Pond and Frost 2005). Current methods measuring selection pressures are based on the protein-coding sequences. The differences in sequences generated by pyrosequencing and Sanger clone-based methods could affect positive selection analysis of viral populations. Analysis using Quasi analysis, SLAC, FEL, and IFEL methods at given sites of HIV-1 Gag proteins between the two sequence populations showed that the positive selections on those unique sites of HIV-1 Gag proteins appear different, especially at codon 75, 390, 474, and 487 (Table 19) although there were no significant differences of positive selection measured by differential test (not shown) (Kosakovsky Pond and Frost 2005). Since codon 75 and 487 are within CTL epitopes, the difference of positive selection on these sites could mislead our interpretation of viral evolutionary processes and its interaction with host.

4.4.2 Summary

Our study showed that pyrosequencing technology is superior to Sanger clone-based method in characterizing genetic diversity of HIV-1 quasispecies. The difference in sequence variants could influence biological and evolutionary studies of HIV-1.

4.5.0 Final Conclusion

My study aimed to characterize genetic diversity of HIV-1 at a population level and to predict potential cross-clade immunogenic sequences for the design of an effective vaccine. I studied two structural proteins of HIV-1, Env and Gag, because they are the major targets of the current vaccine efforts.

Positive selection pressures on both proteins was systematically investigated and showed that genetic diversity of HIV-1 could be characterized in terms of PS pressures and PS patterns. First, PS pressure was found to be widely dispersed across both HIV-1 Env and Gag proteins, and the sequence variability driven by PS is not only in variable but also in conserved regions of HIV-1genes. The results suggest that the conserved regions of HIV-1 are also targeted by the host immune responses and should be considered as targets for vaccines. Second, humoral and cell-mediated immune responses, specially neutralizing antibody, non-neutralizing antibody, and CTL responses, were the major driving forces of genetic diversity within the HIV-1 *env* and *gag* genes at a population level. Furthermore, these responses act together to shape the evolution of HIV-1 worldwide. Third, PS pressure on both HIV-1 Env and Gag proteins remains relatively stable at the population level over time, showing that genetic diversity of HIV-1 driven by host immune responses has changed very little over the last 29 years. These results suggest that the functional constraints of HIV-1 restrict its diversity, as well as the adaptation of HIV-1 to its host. Studying adaptation of HIV-1 to host immune responses would contribute to designing effective HIV-1 vaccine candidates.

.

The PS patterns among four major subtypes (A, B, C, and D) on Env and Gag proteins were compared. The result showed that up to 70% PS sites were shared among the four HIV-1 clades, indicating the existence of the same target regions cross-clades by host immune responses. The result showed that cross-clade immunogenicity exists and there is some basis for developing a cross-clade vaccine. Furthermore, I predicted a number of potential immunogens from Env and Gag proteins as candidate components for a HIV-1 vaccine. These immunogens were previously reported CTL or neutralizing antibody epitopes and were conserved cross-clades worldwide. Most importantly, the host immune responses elicited by many of them were associated with the slower disease progression. Biological experiments should be conducted to confirm and understand their roles in HIV-1 specific immune responses and clinical outcomes.

I also showed that HLA class I allele frequencies in the population drive genetic diversity of HIV-1 and influence its evolution. There is significantly higher number of HLA alleles targeting CTL epitopes of HIV-1 Accessory/Regulator's proteins and a significant correlation was identified between HLA allele frequencies and host CTL responses elicited by Accessory/Regulator's proteins at population level. Moreover, the frequency of HLA class I supertypes such as HLA-B7 were significantly associated with the number of identified optimal CTL epitopes. Thus, HLA class I alleles and their population frequencies should be considered as one of major factors in designing an effective HIV-1 vaccine.

The impact of second-generation sequencing technologies on studying the HIV-1 biology and evolution was also examined. Ultra-deep pyrosequencing (454), one of second-generation sequencing technologies, was found to be superior to Sanger clone-based method in characterizing genetic diversity of HIV-1, especially low frequency (<10%) variants by detecting more sequence variants than Sanger clone-based method. The difference of sequence variants detected between these two methods could affect subsequent functional and evolutionary analysis of HIV-1. At a population level, there is a significant variation on functional sites of HIV-1 Gag proteins between sequences generated by 454 technology and those generated by Sanger clone-based method. This variation is significantly correlated with the density of CTL epitopes on HIV-1 Gag proteins. At an individual level, differences in consensus sequences were identified in more than 25% of patients between sequences generated by 454 technology and Sanger clone-base method, and these differences overlap with functional sites of HIV-1 Gag proteins, especially CTL epitopes.

Analysis of positive selection on HIV Gag sequences generated by the two methods identified several differences in PS sites. Two of them were located on immunogenic CTL epitope regions, and one PS site was associated with HLA class I allele. Thus, it is apparent that pyrosequencing technology is a better alternative for characterizing genetic diversity of HIV-1. The difference of sequence variants generated by two methods could impact our understanding of functions of HIV-1 and evolutionary process of molecular HIV-1 sequences.

The study of my thesis provides important information for developing an effective HIV-1 vaccine and demonstrates the potential of second-generation sequencing technologies to study function and evolution of HIV-1 viral quasispecies.

5.0 References

Aasa-Chapman, M.M., A. Hayman, P. Newton, D. Cornforth, I. Williams, P. Borrow, P. Balfe, and A. McKnight. 2004. Development of the antibody response in acute HIV-1 infection. *Aids* **18:** 371-381.

Addo, M.M., X.G. Yu, A. Rathod, D. Cohen, R.L. Eldridge, D. Strick, M.N. Johnston, C. Corcoran, A.G. Wurcel, C.A. Fitzpatrick, M.E. Feeney, W.R. Rodriguez, N. Basgoz, R. Draenert, D.R. Stone, C. Brander, P.J. Goulder, E.S. Rosenberg, M. Altfeld, and B.D. Walker. 2003. Comprehensive epitope analysis of human immunodeficiency virus type 1 (HIV-1)-specific T-cell responses directed against the entire expressed HIV-1 genome demonstrate broadly directed responses, but no correlation to viral load. *J Virol* **77:** 2081-2092.

Albert, J., B. Abrahamsson, K. Nagy, E. Aurelius, H. Gaines, G. Nystrom, and E.M. Fenyo. 1990. Rapid development of isolate-specific neutralizing antibodies after primary HIV-1 infection and consequent emergence of virus variants which resist neutralization by autologous sera. *Aids* **4:** 107-112.

Alkhatib, G., C. Combadiere, C.C. Broder, Y. Feng, P.E. Kennedy, P.M. Murphy, and E.A. Berger. 1996. CC CKR5: a RANTES, MIP-1alpha, MIP-1beta receptor as a fusion cofactor for macrophage-tropic HIV-1. *Science* **272:** 1955-1958.

Alter, G., N. Teigen, R. Ahern, H. Streeck, A. Meier, E.S. Rosenberg, and M. Altfeld. 2007. Evolution of innate and adaptive effector cell functions during acute HIV-1 infection. *J Infect Dis* **195:** 1452-1460.

Altfeld, M., M.M. Addo, E.S. Rosenberg, F.M. Hecht, P.K. Lee, M. Vogel, X.G. Yu, R. Draenert, M.N. Johnston, D. Strick, T.M. Allen, M.E. Feeney, J.O. Kahn, R.P. Sekaly, J.A. Levy, J.K. Rockstroh, P.J. Goulder, and B.D. Walker. 2003. Influence of HLA-B57 on clinical presentation and viral control during acute HIV-1 infection. *Aids* **17:** 2581-2591.

Altfeld, M., E.T. Kalife, Y. Qi, H. Streeck, M. Lichterfeld, M.N. Johnston, N. Burgett, M.E. Swartz, A. Yang, G. Alter, X.G. Yu, A. Meier, J.K. Rockstroh, T.M. Allen, H. Jessen, E.S. Rosenberg, M. Carrington, and B.D. Walker. 2006. HLA Alleles Associated with Delayed Progression to AIDS Contribute Strongly to the Initial CD8(+) T Cell Response against HIV-1. *PLoS Med* **3:** e403.

Amara, R.R., F. Villinger, J.D. Altman, S.L. Lydy, S.P. O'Neil, S.I. Staprans, D.C. Montefiori, Y. Xu, J.G. Herndon, L.S. Wyatt, M.A. Candido, N.L. Kozyr, P.L. Earl, J.M. Smith, H.L. Ma, B.D. Grimm, M.L. Hulsey, J. Miller, H.M. McClure, J.M. McNicholl, B. Moss, and H.L. Robinson. 2001. Control of a mucosal challenge and prevention of AIDS by a multiprotein DNA/MVA vaccine. *Science* **292:** 69-74.

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

Appay, V., D.F. Nixon, S.M. Donahoe, G.M. Gillespie, T. Dong, A. King, G.S. Ogg, H.M. Spiegel, C. Conlon, C.A. Spina, D.V. Havlir, D.D. Richman, A. Waters, P. Easterbrook, A.J. McMichael, and S.L. Rowland-Jones. 2000. HIV-specific

CD8(+) T cells produce antiviral cytokines but are impaired in cytolytic function. *J Exp Med* **192:** 63-75.

Arendrup, M., C. Nielsen, J.E. Hansen, C. Pedersen, L. Mathiesen, and J.O. Nielsen. 1992. Autologous HIV-1 neutralizing antibodies: emergence of neutralization-resistant escape virus and subsequent development of escape virus neutralizing antibodies. *J Acquir Immune Defic Syndr* **5:** 303-307.

Bachmann, M.F. and R.M. Zinkernagel. 1997. Neutralizing antiviral B cell responses. *Annu Rev Immunol* **15:** 235-270.

Baeten, J.M., B. Chohan, L. Lavreys, V. Chohan, R.S. McClelland, L. Certain, K. Mandaliya, W. Jaoko, and J. Overbaugh. 2007. HIV-1 subtype D infection is associated with faster disease progression than subtype A in spite of similar plasma HIV-1 loads. *J Infect Dis* **195:** 1177-1180.

Bailey, R.C., S. Moses, C.B. Parker, K. Agot, I. Maclean, J.N. Krieger, C.F. Williams, R.T. Campbell, and J.O. Ndinya-Achola. 2007. Male circumcision for HIV prevention in young men in Kisumu, Kenya: a randomised controlled trial. *Lancet* **369:** 643-656.

Banks, N.D., N. Kinsey, J. Clements, and J.E. Hildreth. 2002. Sustained antibody-dependent cell-mediated cytotoxicity (ADCC) in SIV-infected macaques correlates with delayed progression to AIDS. *AIDS Res Hum Retroviruses* **18:** 1197-1205.

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.

Baum, L.L., K.J. Cassutt, K. Knigge, R. Khattri, J. Margolick, C. Rinaldo, C.A. Kleeberger, P. Nishanian, D.R. Henrard, and J. Phair. 1996. HIV-1 gp120-specific antibody-dependent cell-mediated cytotoxicity correlates with rate of disease progression. *J Immunol* **157:** 2168-2173.

Bebenek, K., J. Abbotts, J.D. Roberts, S.H. Wilson, and T.A. Kunkel. 1989. Specificity and mechanism of error-prone replication by human immunodeficiency virus-1 reverse transcriptase. *J Biol Chem* **264:** 16948-16956.

Berkhout, B. and K.T. Jeang. 1992. Functional roles for the TATA promoter and enhancers in basal and Tat-induced expression of the human immunodeficiency virus type 1 long terminal repeat. *J Virol* **66:** 139-149.

Berkhout, B., R.H. Silverman, and K.T. Jeang. 1989. Tat trans-activates the human immunodeficiency virus through a nascent RNA target. *Cell* **59:** 273-282.

Berkley, S. 1991. Parenteral transmission of HIV in Africa. *Aids* **5 Suppl 1:** S87-92.

Binley, J.M., R. Wyatt, E. Desjardins, P.D. Kwong, W. Hendrickson, J.P. Moore, and J. Sodroski. 1998. Analysis of the interaction of antibodies with a conserved enzymatically deglycosylated core of the HIV type 1 envelope glycoprotein 120. *AIDS Res Hum Retroviruses* **14:** 191-198.

Biron, C.A. and L. Brossay. 2001. NK cells and NKT cells in innate defense against viral infections. *Curr Opin Immunol* **13:** 458-464.

Bonhoeffer, S., E.C. Holmes, and M.A. Nowak. 1995. Causes of HIV diversity. *Nature* **376:** 125.

Borrow, P., H. Lewicki, X. Wei, M.S. Horwitz, N. Peffer, H. Meyers, J.A. Nelson, J.E. Gairin, B.H. Hahn, M.B. Oldstone, and G.M. Shaw. 1997. Antiviral pressure exerted by HIV-1-specific cytotoxic T lymphocytes (CTLs) during primary infection demonstrated by rapid selection of CTL escape virus. *Nat Med* **3:** 205-211.

Boyer, P.L., H.Q. Gao, and S.H. Hughes. 1998. A mutation at position 190 of human immunodeficiency virus type 1 reverse transcriptase interacts with mutations at positions 74 and 75 via the template primer. *Antimicrob Agents Chemother* **42:** 447-452.

Brander, C., N. Frahm, and B.D. Walker. 2006. The challenges of host and viral diversity in HIV vaccine design. *Curr Opin Immunol* **18:** 430-437.

Brown, S.A., J.L. Hurwitz, X. Zhan, P.C. Doherty, and K.S. Slobod. 2005. CD8+ T-cells: are they sufficient to prevent, contain or eradicate HIV-1 infection? *Curr Drug Targets Infect Disord* **5:** 113-119.

Brumme, Z.L., C.J. Brumme, D. Heckerman, B.T. Korber, M. Daniels, J. Carlson, C. Kadie, T. Bhattacharya, C. Chui, J. Szinger, T. Mo, R.S. Hogg, J.S. Montaner, N. Frahm, C. Brander, B.D. Walker, and P.R. Harrigan. 2007. Evidence of differential HLA class I-mediated viral evolution in functional and accessory/regulatory genes of HIV-1. *PLoS Pathog* **3:** e94.

Brumme, Z.L., I. Tao, S. Szeto, C.J. Brumme, J.M. Carlson, D. Chan, C. Kadie, N. Frahm, C. Brander, B. Walker, D. Heckerman, and P.R. Harrigan. 2008. Human leukocyte antigen-specific polymorphisms in HIV-1 Gag and their association with viral load in chronic untreated infection. *Aids* **22:** 1277-1286.

Bukrinsky, M.I., S. Haggerty, M.P. Dempsey, N. Sharova, A. Adzhubel, L. Spitz, P. Lewis, D. Goldfarb, M. Emerman, and M. Stevenson. 1993. A nuclear localization signal within HIV-1 matrix protein that governs infection of non-dividing cells. *Nature* **365:** 666-669.

Bulterys, M. and M.G. Fowler. 2000. Prevention of HIV infection in children. *Pediatr Clin North Am* **47:** 241-260.

Burns, D.P. and R.C. Desrosiers. 1994. Envelope sequence variation, neutralizing antibodies, and primate lentivirus persistence. *Curr Top Microbiol Immunol* **188:** 185-219.

Burton, D.R., R.C. Desrosiers, R.W. Doms, W.C. Koff, P.D. Kwong, J.P. Moore, G.J. Nabel, J. Sodroski, I.A. Wilson, and R.T. Wyatt. 2004. HIV vaccine design and the neutralizing antibody problem. *Nat Immunol* **5:** 233-236.

Burton, D.R., J. Pyati, R. Koduri, S.J. Sharp, G.B. Thornton, P.W. Parren, L.S. Sawyer, R.M. Hendry, N. Dunlop, P.L. Nara, and et al. 1994. Efficient neutralization of primary isolates of HIV-1 by a recombinant human monoclonal antibody. *Science* **266:** 1024-1027.

Bush, R.M. 2001. Predicting adaptive evolution. *Nat Rev Genet* **2:** 387-392.

Bushman, F.D., C. Hoffmann, K. Ronen, N. Malani, N. Minkah, H.M. Rose, P. Tebas, and G.P. Wang. 2008. Massively parallel pyrosequencing in HIV research. *Aids* **22:** 1411-1415.

Cai, F., H. Chen, C.B. Hicks, J.A. Bartlett, J. Zhu, and F. Gao. 2007. Detection of minor drug-resistant populations by parallel allele-specific sequencing. *Nat Methods* **4:** 123-125.

Campbell, E.M., O. Perez, J.L. Anderson, and T.J. Hope. 2008. Visualization of a proteasome-independent intermediate during restriction of HIV-1 by rhesus TRIM5alpha. *J Cell Biol* **180:** 549-561.

Carotenuto, P., D. Looij, L. Keldermans, F. de Wolf, and J. Goudsmit. 1998. Neutralizing antibodies are positively associated with CD4+ T-cell counts and T-cell function in long-term AIDS-free infection. *Aids* **12:** 1591-1600.

Casement, K.S., P.N. Nehete, R.B. Arlinghaus, and K.J. Sastry. 1995. Cross-reactive cytotoxic T lymphocytes induced by V3 loop synthetic peptides from different strains of human immunodeficiency virus type 1. *Virology* **211:** 261-267.

Castetbon, K., V. Leroy, R. Spira, and F. Dabis. 2000. [Preventing the transmission of HIV-1 from mother to child in Africa in the year 2000]. *Sante* **10:** 103-113.

CDC. 1982a. Current Trends Update on Acquired Immune Deficiency Syndrome (AIDS)-United States. *MMWR Weekly* **31:** 507-508.

CDC. 1982b. Epidemiologic Notes and Reports Possible Transfusion-Associated Acquired Immune Deficiency Syndrome, AIDS-California. *MMWR Weekly* **31:** 652-654.

Centlivre, M., M. Sala, S. Wain-Hobson, and B. Berkhout. 2007. In HIV-1 pathogenesis the die is cast during primary infection. *Aids* **21:** 1-11.

Chakrabarti, B.K., X. Ling, Z.Y. Yang, D.C. Montefiori, A. Panet, W.P. Kong, B. Welcher, M.K. Louder, J.R. Mascola, and G.J. Nabel. 2005. Expanded breadth of virus neutralization after immunization with a multiclade envelope HIV vaccine candidate. *Vaccine* **23:** 3434-3445.

Chakraborty, H., P.K. Sen, R.W. Helms, P.L. Vernazza, S.A. Fiscus, J.J. Eron, B.K. Patterson, R.W. Coombs, J.N. Krieger, and M.S. Cohen. 2001. Viral burden in genital secretions determines male-to-female sexual transmission of HIV-1: a probabilistic empiric model. *Aids* **15:** 621-627.

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.

Chen, L., A. Perlina, and C.J. Lee. 2004. Positive selection detection in 40,000 human immunodeficiency virus (HIV) type 1 sequences automatically identifies drug resistance and positive fitness mutations in HIV protease and reverse transcriptase. *J Virol* **78:** 3722-3732.

Chiu, Y.L. and W.C. Greene. 2009. APOBEC3G: an intracellular centurion. *Philos Trans R Soc Lond B Biol Sci* **364:** 689-703.

Cho, M.W., Y.B. Kim, M.K. Lee, K.C. Gupta, W. Ross, R. Plishka, A. Buckler-White, T. Igarashi, T. Theodore, R. Byrum, C. Kemp, D.C. Montefiori, and M.A. Martin. 2001. Polyvalent envelope glycoprotein vaccine elicits a broader neutralizing antibody response but is unable to provide sterilizing protection against heterologous Simian/human immunodeficiency virus infection in pigtailed macaques. *J Virol* **75:** 2224-2234.

Choisy, M., C.H. Woelk, J.F. Guegan, and D.L. Robertson. 2004. Comparative study of adaptive molecular evolution in different human immunodeficiency virus groups and subtypes. *J Virol* **78:** 1962-1970.

Chun, T.W., L. Carruth, D. Finzi, X. Shen, J.A. DiGiuseppe, H. Taylor, M. Hermankova, K. Chadwick, J. Margolick, T.C. Quinn, Y.H. Kuo, R. Brookmeyer, M.A. Zeiger,

P. Barditch-Crovo, and R.F. Siliciano. 1997. Quantification of latent tissue reservoirs and total body viral load in HIV-1 infection. *Nature* **387:** 183-188.

Clerici, M., F.T. Hakim, D.J. Venzon, S. Blatt, C.W. Hendrix, T.A. Wynn, and G.M. Shearer. 1993. Changes in interleukin-2 and interleukin-4 production in asymptomatic, human immunodeficiency virus-seropositive individuals. *J Clin Invest* **91:** 759-765.

Clerici, M. and G.M. Shearer. 1993. A TH1-->TH2 switch is a critical step in the etiology of HIV infection. *Immunol Today* **14:** 107-111.

Cohen, J. 2007. AIDS research. Did Merck's failed HIV vaccine cause harm? *Science* **318:** 1048-1049.

Cohen, M.S., N. Hellmann, J.A. Levy, K. DeCock, and J. Lange. 2008. The spread, treatment, and prevention of HIV-1: evolution of a global pandemic. *J Clin Invest* **118:** 1244-1254.

Colonna, M., F. Navarro, T. Bellon, M. Llano, P. Garcia, J. Samaridis, L. Angman, M. Cella, and M. Lopez-Botet. 1997. A common inhibitory receptor for major histocompatibility complex class I molecules on human lymphoid and myelomonocytic cells. *J Exp Med* **186:** 1809-1818.

Connor, E.M., R.S. Sperling, R. Gelber, P. Kiselev, G. Scott, M.J. O'Sullivan, R. VanDyke, M. Bey, W. Shearer, R.L. Jacobson, and et al. 1994. Reduction of maternal-infant transmission of human immunodeficiency virus type 1 with zidovudine treatment. Pediatric AIDS Clinical Trials Group Protocol 076 Study Group. *N Engl J Med* **331:** 1173-1180.

Coombs, R.W., P.S. Reichelderfer, and A.L. Landay. 2003. Recent observations on HIV type-1 infection in the genital tract of men and women. *Aids* **17:** 455-480.

Coplan, P.M., S.B. Gupta, S.A. Dubey, P. Pitisuttithum, A. Nikas, B. Mbewe, E. Vardas, M. Schechter, E.G. Kallas, D.C. Freed, T.M. Fu, C.T. Mast, P. Puthavathana, J. Kublin, K. Brown Collins, J. Chisi, R. Pendame, S.J. Thaler, G. Gray, J. McIntyre, W.L. Straus, J.H. Condra, D.V. Mehrotra, H.A. Guess, E.A. Emini, and J.W. Shiver. 2005. Cross-reactivity of anti-HIV-1 T cell immune responses among the major HIV-1 clades in HIV-1-positive individuals from 4 continents. *J Infect Dis* **191:** 1427-1434.

Corey, L., M.J. McElrath, and J.G. Kublin. 2009. Post-step modifications for research on HIV vaccines. *Aids* **23:** 3-8.

da Silva, J. 2003. The evolutionary adaptation of HIV-1 to specific immunity. *Curr HIV Res* **1:** 363-371.

da Silva, J. and A.L. Hughes. 1999. Molecular phylogenetic evidence of cytotoxic T lymphocyte (CTL) selection on human immunodeficiency virus type 1 (HIV-1). *Mol Biol Evol* **16:** 1420-1422.

Day, C.L., P. Kiepiela, A.J. Leslie, M. van der Stok, K. Nair, N. Ismail, I. Honeyborne, H. Crawford, H.M. Coovadia, P.J. Goulder, B.D. Walker, and P. Klenerman. 2007. Proliferative capacity of epitope-specific CD8 T-cell responses is inversely related to viral load in chronic human immunodeficiency virus type 1 infection. *J Virol* **81:** 434-438.

De Groot, A.S., B. Jesdale, W. Martin, C. Saint Aubin, H. Sbai, A. Bosma, J. Lieberman, G. Skowron, F. Mansourati, and K.H. Mayer. 2003. Mapping cross-clade HIV-1 vaccine epitopes using a bioinformatics approach. *Vaccine* **21:** 4486-4504.

De Guzman, R.N., Z.R. Wu, C.C. Stalling, L. Pappalardo, P.N. Borer, and M.F. Summers. 1998. Structure of the HIV-1 nucleocapsid protein bound to the SL3 psi-RNA recognition element. *Science* **279:** 384-388.

de Jong, J.J., J. Goudsmit, W. Keulen, B. Klaver, W. Krone, M. Tersmette, and A. de Ronde. 1992. Human immunodeficiency virus type 1 clones chimeric for the envelope V3 domain differ in syncytium formation and replication capacity. *J Virol* **66:** 757-765.

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

Del Val, M., H.J. Schlicht, T. Ruppert, M.J. Reddehase, and U.H. Koszinowski. 1991. Efficient processing of an antigenic sequence for presentation by MHC class I molecules depends on its neighboring residues in the protein. *Cell* **66:** 1145-1153.

Delwart, E.L. and C.J. Gordon. 1997. Tracking changes in HIV-1 envelope quasispecies using DNA heteroduplex analysis. *Methods* **12:** 348-354.

Delwart, E.L., E.G. Shpaer, J. Louwagie, F.E. McCutchan, M. Grez, H. Rubsamen-Waigmann, and J.I. Mullins. 1993. Genetic relationships determined by a DNA heteroduplex mobility assay: analysis of HIV-1 env genes. *Science* **262:** 1257-1261.

DeVico, A.L. and R.C. Gallo. 2004. Control of HIV-1 infection by soluble factors of the immune response. *Nat Rev Microbiol* **2:** 401-413.

Douek, D.C., J.M. Brenchley, M.R. Betts, D.R. Ambrozak, B.J. Hill, Y. Okamoto, J.P. Casazza, J. Kuruppu, K. Kunstman, S. Wolinsky, Z. Grossman, M. Dybul, A. Oxenius, D.A. Price, M. Connors, and R.A. Koup. 2002. HIV preferentially infects HIV-specific CD4+ T cells. *Nature* **417:** 95-98.

Dulioust, A., S. Paulous, L. Guillemot, A.M. Delavalle, F. Boue, and F. Clavel. 1999. Constrained evolution of human immunodeficiency virus type 1 protease during sequential therapy with two distinct protease inhibitors. *J Virol* **73:** 850-854.

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

Ferrari, G., W. Humphrey, M.J. McElrath, J.L. Excler, A.M. Duliege, M.L. Clements, L.C. Corey, D.P. Bolognesi, and K.J. Weinhold. 1997. Clade B-based HIV-1 vaccines elicit cross-clade cytotoxic T lymphocyte reactivities in uninfected volunteers. *Proc Natl Acad Sci U S A* **94:** 1396-1401.

Fischer, W., S. Perkins, J. Theiler, T. Bhattacharya, K. Yusim, R. Funkhouser, C. Kuiken, B. Haynes, N.L. Letvin, B.D. Walker, B.H. Hahn, and B.T. Korber. 2007. Polyvalent vaccines for optimal coverage of potential T-cell epitopes in global HIV-1 variants. *Nat Med* **13:** 100-106.

Forthal, D.N., G. Landucci, and E.S. Daar. 2001. Antibody from patients with acute human immunodeficiency virus (HIV) infection inhibits primary strains of HIV type 1 in the presence of natural-killer effector cells. *J Virol* **75:** 6953-6961.

Fouchier, R.A., M. Groenink, N.A. Kootstra, M. Tersmette, H.G. Huisman, F. Miedema, and H. Schuitemaker. 1992. Phenotype-associated sequence variation in the third variable domain of the human immunodeficiency virus type 1 gp120 molecule. *J Virol* **66:** 3183-3187.

Fowler, M.G., R.J. Simonds, and A. Roongpisuthipong. 2000. Update on perinatal HIV transmission. *Pediatr Clin North Am* **47:** 21-38.

Frater, A.J., H. Brown, A. Oxenius, H.F. Gunthard, B. Hirschel, N. Robinson, A.J. Leslie, R. Payne, H. Crawford, A. Prendergast, C. Brander, P. Kiepiela, B.D. Walker, P.J. Goulder, A. McLean, and R.E. Phillips. 2007. Effective T-cell responses select human immunodeficiency virus mutants and slow disease progression. *J Virol* **81:** 6742-6751.

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

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

Frost, S.D., H.F. Gunthard, J.K. Wong, D. Havlir, D.D. Richman, and A.J. Leigh Brown. 2001. Evidence for positive selection driving the evolution of HIV-1 env under potent antiviral therapy. *Virology* **284:** 250-258.

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

Gallay, P., T. Hope, D. Chin, and D. Trono. 1997. HIV-1 infection of nondividing cells through the recognition of integrase by the importin/karyopherin pathway. *Proc Natl Acad Sci U S A* **94:** 9825-9830.

Gao, F., B.T. Korber, E. Weaver, H.X. Liao, B.H. Hahn, and B.F. Haynes. 2004. Centralized immunogens as a vaccine strategy to overcome HIV-1 diversity. *Expert Rev Vaccines* **3:** S161-168.

Gao, X., A. Bashirova, A.K. Iversen, J. Phair, J.J. Goedert, S. Buchbinder, K. Hoots, D. Vlahov, M. Altfeld, S.J. O'Brien, and M. Carrington. 2005. AIDS restriction HLA allotypes target distinct intervals of HIV-1 pathogenesis. *Nat Med* **11:** 1290-1292.

Garber, M.E., P. Wei, V.N. KewalRamani, T.P. Mayall, C.H. Herrmann, A.P. Rice, D.R. Littman, and K.A. Jones. 1998. The interaction between HIV-1 Tat and human cyclin T1 requires zinc and a critical cysteine residue that is not conserved in the murine CycT1 protein. *Genes Dev* **12:** 3512-3527.

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.

Geels, M.J., S.A. Dubey, K. Anderson, E. Baan, M. Bakker, G. Pollakis, W.A. Paxton, J.W. Shiver, and J. Goudsmit. 2005. Broad cross-clade T-cell responses to gag in individuals infected with human immunodeficiency virus type 1 non-B clades (A to G): importance of HLA anchor residue conservation. *J Virol* **79:** 11247-11258.

Geijtenbeek, T.B., D.S. Kwon, R. Torensma, S.J. van Vliet, G.C. van Duijnhoven, J. Middel, I.L. Cornelissen, H.S. Nottet, V.N. KewalRamani, D.R. Littman, C.G. Figdor, and Y. van Kooyk. 2000. DC-SIGN, a dendritic cell-specific HIV-1-binding protein that enhances trans-infection of T cells. *Cell* **100:** 587-597.

Geldmacher, C., J.R. Currier, E. Herrmann, A. Haule, E. Kuta, F. McCutchan, L. Njovu, S. Geis, O. Hoffmann, L. Maboko, C. Williamson, D. Birx, A. Meyerhans, J. Cox, and M. Hoelscher. 2007. CD8 T-cell recognition of multiple epitopes within specific Gag regions is associated with maintenance of a low steady-state viremia

in human immunodeficiency virus type 1-seropositive patients. *J Virol* **81:** 2440-2448.

Geretti, A.M. 2006. HIV-1 subtypes: epidemiology and significance for HIV management. *Curr Opin Infect Dis* **19:** 1-7.

Ghys, P.D., K. Fransen, M.O. Diallo, V. Ettiegne-Traore, I.M. Coulibaly, K.M. Yeboue, M.L. Kalish, C. Maurice, J.P. Whitaker, A.E. Greenberg, and M. Laga. 1997. The associations between cervicovaginal HIV shedding, sexually transmitted diseases and immunosuppression in female sex workers in Abidjan, Cote d'Ivoire. *Aids* **11:** F85-93.

Gottlieb, M., H. Schanker, P. Fan, A. Saxon, and J. Wisman. 1981a. Pneumocystis Pneumonia-Los Angeles. *MMWR Weekly* **30:** 1-3.

Gottlieb, M.S., R. Schroff, H.M. Schanker, J.D. Weisman, P.T. Fan, R.A. Wolf, and A. Saxon. 1981b. 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.

Gottlinger, H.G. 2007. HIV-1 Gag: a Molecular Machine Driving Viral Particle Assembly and Release. Los Alomas National Laboratory.

Goulder, P.J., R.E. Phillips, R.A. Colbert, S. McAdam, G. Ogg, M.A. Nowak, P. Giangrande, G. Luzzi, B. Morgan, A. Edwards, A.J. McMichael, and S. Rowland-Jones. 1997. Late escape from an immunodominant cytotoxic T-lymphocyte response associated with progression to AIDS. *Nat Med* **3:** 212-217.

Grobler, J., C.M. Gray, C. Rademeyer, C. Seoighe, G. Ramjee, S.A. Karim, L. Morris, and C. Williamson. 2004. Incidence of HIV-1 dual infection and its association with increased viral load set point in a cohort of HIV-1 subtype C-infected female sex workers. *J Infect Dis* **190:** 1355-1359.

Groenink, M., A.C. Andeweg, R.A. Fouchier, S. Broersen, R.C. van der Jagt, H. Schuitemaker, R.E. de Goede, M.L. Bosch, H.G. Huisman, and M. Tersmette. 1992. Phenotype-associated env gene variation among eight related human immunodeficiency virus type 1 clones: evidence for in vivo recombination and determinants of cytotropism outside the V3 domain. *J Virol* **66:** 6175-6180.

Gudmundsdotter, L., D. Bernasconi, B. Hejdeman, E. Sandstrom, A. Alaeus, K. Lidman, B. Ensoli, B. Wahren, and S. Butto. 2008. Cross-clade immune responses to Gag p24 in patients infected with different HIV-1 subtypes and correlation with HLA class I and II alleles. *Vaccine* **26:** 5182-5187.

Hager-Braun, C., H. Katinger, and K.B. Tomer. 2006. The HIV-neutralizing monoclonal antibody 4E10 recognizes N-terminal sequences on the native antigen. *J Immunol* **176:** 7471-7481.

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.

Halperin, D.T. 1999. Dry sex practices and HIV infection in the Dominican Republic and Haiti. *Sex Transm Infect* **75:** 445-446.

Halvas, E.K., G.M. Aldrovandi, P. Balfe, I.A. Beck, V.F. Boltz, J.M. Coffin, L.M. Frenkel, J.D. Hazelwood, V.A. Johnson, M. Kearney, A. Kovacs, D.R. Kuritzkes, K.J. Metzner, D.V. Nissley, M. Nowicki, S. Palmer, R. Ziermann, R.Y. Zhao, C.L. Jennings, J. Bremer, D. Brambilla, and J.W. Mellors. 2006. Blinded, multicenter

comparison of methods to detect a drug-resistant mutant of human immunodeficiency virus type 1 at low frequency. *J Clin Microbiol* **44:** 2612-2614.

Harari, A., V. Dutoit, C. Cellerai, P.A. Bart, R.A. Du Pasquier, and G. Pantaleo. 2006. Functional signatures of protective antiviral T-cell immunity in human virus infections. *Immunol Rev* **211:** 236-254.

Harari, A. and G. Pantaleo. 2008. HIV-1-specific immune response. *Adv Pharmacol* **56:** 75-92.

Harari, A., G.P. Rizzardi, K. Ellefsen, D. Ciuffreda, P. Champagne, P.A. Bart, D. Kaufmann, A. Telenti, R. Sahli, G. Tambussi, L. Kaiser, A. Lazzarin, L. Perrin, and G. Pantaleo. 2002. Analysis of HIV-1- and CMV-specific memory CD4 T-cell responses during primary and chronic infection. *Blood* **100:** 1381-1387.

Harari, A., F. Vallelian, P.R. Meylan, and G. Pantaleo. 2005. Functional heterogeneity of memory CD4 T cell responses in different conditions of antigen exposure and persistence. *J Immunol* **174:** 1037-1045.

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.

Heil, F., H. Hemmi, H. Hochrein, F. Ampenberger, C. Kirschning, S. Akira, G. Lipford, H. Wagner, and S. Bauer. 2004. Species-specific recognition of single-stranded RNA via toll-like receptor 7 and 8. *Science* **303:** 1526-1529.

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

Herbeck, J.T., D.C. Nickle, G.H. Learn, G.S. Gottlieb, M.E. Curlin, L. Heath, and J.I. Mullins. 2006. Human immunodeficiency virus type 1 env evolves toward ancestral states upon transmission to a new host. *J Virol* **80:** 1637-1644.

Higgins, D.G., J.D. Thompson, and T.J. Gibson. 1996. Using CLUSTAL for multiple sequence alignments. *Methods Enzymol* **266:** 383-402.

Hoffmann, C., N. Minkah, J. Leipzig, G. Wang, M.Q. Arens, P. Tebas, and F.D. Bushman. 2007. DNA bar coding and pyrosequencing to identify rare HIV drug resistance mutations. *Nucleic Acids Res* **35:** e91.

Honeyborne, I., A. Prendergast, F. Pereyra, A. Leslie, H. Crawford, R. Payne, S. Reddy, K. Bishop, E. Moodley, K. Nair, M. van der Stok, N. McCarthy, C.M. Rousseau, M. Addo, J.I. Mullins, C. Brander, P. Kiepiela, B.D. Walker, and P.J. Goulder. 2007. Control of human immunodeficiency virus type 1 is associated with HLA-B*13 and targeting of multiple gag-specific CD8+ T-cell epitopes. *J Virol* **81:** 3667-3672.

Huang, M., J.M. Orenstein, M.A. Martin, and E.O. Freed. 1995. p6Gag is required for particle production from full-length human immunodeficiency virus type 1 molecular clones expressing protease. *J Virol* **69:** 6810-6818.

Huang, X., J.J. Barchi, Jr., F.D. Lung, P.P. Roller, P.L. Nara, J. Muschik, and R.R. Garrity. 1997. Glycosylation affects both the three-dimensional structure and antibody binding properties of the HIV-1IIIB GP120 peptide RP135. *Biochemistry* **36:** 10846-10856.

Huber, M., M. Fischer, B. Misselwitz, A. Manrique, H. Kuster, B. Niederost, R. Weber, V. von Wyl, H.F. Gunthard, and A. Trkola. 2006. Complement lysis activity in

autologous plasma is associated with lower viral loads during the acute phase of HIV-1 infection. *PLoS Med* **3:** e441.

Huber, M. and A. Trkola. 2007. Humoral immunity to HIV-1: neutralization and beyond. *J Intern Med* **262:** 5-25.

Huelsenbeck, J.P., F. Ronquist, R. Nielsen, and J.P. Bollback. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* **294:** 2310-2314.

Imami, N., G. Hardy, A. Pires, C. Burton, A. Sullivan, and F. Gotch. 2001. Detection and quantification of HIV-1 specific CD4 helper and CD8 cytotoxic cells: their role in HIV-1-infected individuals and vaccine recipients. *HIV Med* **2:** 146-153.

Jansen, C.A., D. van Baarle, and F. Miedema. 2006. HIV-specific CD4+ T cells and viremia: who's in control? *Trends Immunol* **27:** 119-124.

Janssen, E.M., E.E. Lemmens, T. Wolfe, U. Christen, M.G. von Herrath, and S.P. Schoenberger. 2003. CD4+ T cells are required for secondary expansion and memory in CD8+ T lymphocytes. *Nature* **421:** 852-856.

Johnson, V.A., F. Brun-Vezinet, B. Clotet, H.F. Gunthard, D.R. Kuritzkes, D. Pillay, J.M. Schapiro, and D.D. Richman. 2008. Update of the Drug Resistance Mutations in HIV-1. *Top HIV Med* **16:** 138-145.

Johnston, M.I. and A.S. Fauci. 2007. An HIV vaccine--evolving concepts. *N Engl J Med* **356:** 2073-2081.

Jolly, C., K. Kashefi, M. Hollinshead, and Q.J. Sattentau. 2004. HIV-1 cell to cell transfer across an Env-induced, actin-dependent synapse. *J Exp Med* **199:** 283-293.

Jones, N.A., X. Wei, D.R. Flower, M. Wong, F. Michor, M.S. Saag, B.H. Hahn, M.A. Nowak, G.M. Shaw, and P. Borrow. 2004. Determinants of human immunodeficiency virus type 1 escape from the primary CD8+ cytotoxic T lymphocyte response. *J Exp Med* **200:** 1243-1256.

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.

Kaufmann, D.E., P.M. Bailey, J. Sidney, B. Wagner, P.J. Norris, M.N. Johnston, L.A. Cosimi, M.M. Addo, M. Lichterfeld, M. Altfeld, N. Frahm, C. Brander, A. Sette, B.D. Walker, and E.S. Rosenberg. 2004. Comprehensive analysis of human immunodeficiency virus type 1-specific CD4 responses reveals marked immunodominance of gag and nef and the presence of broadly recognized peptides. *J Virol* **78:** 4463-4477.

Kaul, R., J. Rutherford, S.L. Rowland-Jones, J. Kimani, J.I. Onyango, K. Fowke, K. MacDonald, J.J. Bwayo, A.J. McMichael, and F.A. Plummer. 2004. HIV-1 Env-specific cytotoxic T-lymphocyte responses in exposed, uninfected Kenyan sex workers: a prospective analysis. *Aids* **18:** 2087-2089.

Kaul, R., D. Trabattoni, J.J. Bwayo, D. Arienti, A. Zagliani, F.M. Mwangi, C. Kariuki, E.N. Ngugi, K.S. MacDonald, T.B. Ball, M. Clerici, and F.A. Plummer. 1999. HIV-1-specific mucosal IgA in a cohort of HIV-1-resistant Kenyan sex workers. *Aids* **13:** 23-29.

Keating, S.M., R.C. Bollinger, T.C. Quinn, J.B. Jackson, and L.M. Carruth. 2002. Cross-clade T lymphocyte-mediated immunity to HIV type 1: implications for vaccine design and immunodetection assays. *AIDS Res Hum Retroviruses* **18:** 1067-1079.

Kiepiela, P., A.J. Leslie, I. Honeyborne, D. Ramduth, C. Thobakgale, S. Chetty, P. Rathnavalu, C. Moore, K.J. Pfafferott, L. Hilton, P. Zimbwa, S. Moore, T. Allen, C. Brander, M.M. Addo, M. Altfeld, I. James, S. Mallal, M. Bunce, L.D. Barber, J. Szinger, C. Day, P. Klenerman, J. Mullins, B. Korber, H.M. Coovadia, B.D. Walker, and P.J. Goulder. 2004. Dominant influence of HLA-B in mediating the potential co-evolution of HIV and HLA. *Nature* **432:** 769-775.

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

Klein, M.R., S.H. van der Burg, E. Hovenkamp, A.M. Holwerda, J.W. Drijfhout, C.J. Melief, and F. Miedema. 1998. Characterization of HLA-B57-restricted human immunodeficiency virus type 1 Gag- and RT-specific cytotoxic T lymphocyte responses. *J Gen Virol* **79 (Pt 9):** 2191-2201.

Kong, W.P., L. Wu, T.C. Wallstrom, W. Fischer, Z.Y. Yang, S.Y. Ko, N.L. Letvin, B.F. Haynes, B.H. Hahn, B. Korber, and G.J. Nabel. 2009. Expanded breadth of the T-cell response to mosaic human immunodeficiency virus type 1 envelope DNA vaccination. *J Virol* **83:** 2201-2215.

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

Korber, B.T., N.L. Letvin, and B.F. Haynes. 2009. T cell Vaccine Strategies for HIV, the Virus With a Thousand Faces. *J Virol*.

Korber, B.T., K. MacInnes, R.F. Smith, and G. Myers. 1994. Mutational trends in V3 loop protein sequences observed in different genetic lineages of human immunodeficiency virus type 1. *J Virol* **68:** 6730-6744.

Kosakovsky Pond, S.L. and S.D. Frost. 2005. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol Biol Evol* **22:** 1208-1222.

Koup, R.A., J.T. Safrit, Y. Cao, C.A. Andrews, G. McLeod, W. Borkowsky, C. Farthing, and D.D. Ho. 1994. Temporal association of cellular immune responses with the initial control of viremia in primary human immunodeficiency virus type 1 syndrome. *J Virol* **68:** 4650-4655.

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

Kuroda, M.J., J.E. Schmitz, W.A. Charini, C.E. Nickerson, M.A. Lifton, C.I. Lord, M.A. Forman, and N.L. Letvin. 1999. Emergence of CTL coincides with clearance of virus during primary simian immunodeficiency virus infection in rhesus monkeys. *J Immunol* **162:** 5127-5133.

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.

Lathey, J.L., R.D. Pratt, and S.A. Spector. 1997. Appearance of autologous neutralizing antibody correlates with reduction in virus load and phenotype switch during primary infection with human immunodeficiency virus type 1. *J Infect Dis* **175:** 231-232.

Lazaryan, A., E. Lobashevsky, J. Mulenga, E. Karita, S. Allen, J. Tang, and R.A. Kaslow. 2006. Human leukocyte antigen B58 supertype and human immunodeficiency virus type 1 infection in native Africans. *J Virol* **80:** 6056-6060.

Lehner, T., Y. Wang, J. Pido-Lopez, T. Whittall, L.A. Bergmeier, and K. Babaahmady. 2008. The emerging role of innate immunity in protection against HIV-1 infection. *Vaccine* **26:** 2997-3001.

Leonard, C.K., M.W. Spellman, L. Riddle, R.J. Harris, J.N. Thomas, and T.J. Gregory. 1990. Assignment of intrachain disulfide bonds and characterization of potential glycosylation sites of the type 1 recombinant human immunodeficiency virus envelope glycoprotein (gp120) expressed in Chinese hamster ovary cells. *J Biol Chem* **265:** 10373-10382.

Leslie, A., D. Kavanagh, I. Honeyborne, K. Pfafferott, C. Edwards, T. Pillay, L. Hilton, C. Thobakgale, D. Ramduth, R. Draenert, S. Le Gall, G. Luzzi, A. Edwards, C. Brander, A.K. Sewell, S. Moore, J. Mullins, C. Moore, S. Mallal, N. Bhardwaj, K. Yusim, R. Phillips, P. Klenerman, B. Korber, P. Kiepiela, B. Walker, and P. Goulder. 2005. Transmission and accumulation of CTL escape variants drive negative associations between HIV polymorphisms and HLA. *J Exp Med* **201:** 891-902.

Leslie, A.J., K.J. Pfafferott, P. Chetty, R. Draenert, M.M. Addo, M. Feeney, Y. Tang, E.C. Holmes, T. Allen, J.G. Prado, M. Altfeld, C. Brander, C. Dixon, D. Ramduth, P. Jeena, S.A. Thomas, A. St John, T.A. Roach, B. Kupfer, G. Luzzi, A. Edwards, G. Taylor, H. Lyall, G. Tudor-Williams, V. Novelli, J. Martinez-Picado, P. Kiepiela, B.D. Walker, and P.J. Goulder. 2004. HIV evolution: CTL escape mutation and reversion after transmission. *Nat Med* **10:** 282-289.

Letvin, N.L., J.R. Mascola, Y. Sun, D.A. Gorgone, A.P. Buzby, L. Xu, Z.Y. Yang, B. Chakrabarti, S.S. Rao, J.E. Schmitz, D.C. Montefiori, B.R. Barker, F.L. Bookstein, and G.J. Nabel. 2006. Preserved CD4+ central memory T cells and survival in vaccinated SIV-challenged monkeys. *Science* **312:** 1530-1533.

Lever, A.M. and K.T. Jeang. 2006. Replication of human immunodeficiency virus type 1 from entry to exit. *Int J Hematol* **84:** 23-30.

Levy, J.A. 2001. The importance of the innate immune system in controlling HIV infection and disease. *Trends Immunol* **22:** 312-316.

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

Li, Y., L. Luo, N. Rasool, and C.Y. Kang. 1993. Glycosylation is necessary for the correct folding of human immunodeficiency virus gp120 in CD4 binding. *J Virol* **67:** 584-588.

Liang, B., M. Luo, T.B. Ball, and F.A. Plummer. 2007. QUASI analysis of the HIV-1 envelope sequences in the Los Alamos National Laboratory HIV sequence

database: pattern and distribution of positive selection sites and their frequencies over years. *Biochem Cell Biol* **85:** 259-264.

Liang, B., M. Luo, T.B. Ball, X. Yao, G. Van Domselaar, W.R. Cuff, M. Cheang, S.J. Jones, and F.A. Plummer. 2008. Systematic analysis of host immunological pressure on the envelope gene of human immunodeficiency virus type 1 by an immunobioinformatics approach. *Curr HIV Res* **6:** 370-379.

Lichterfeld, M., X.G. Yu, M.T. Waring, S.K. Mui, M.N. Johnston, D. Cohen, M.M. Addo, J. Zaunders, G. Alter, E. Pae, D. Strick, T.M. Allen, E.S. Rosenberg, B.D. Walker, and M. Altfeld. 2004. HIV-1-specific cytotoxicity is preferentially mediated by a subset of CD8(+) T cells producing both interferon-gamma and tumor necrosis factor-alpha. *Blood* **104:** 487-494.

Liu, C., M. Carrington, R.A. Kaslow, X. Gao, C.R. Rinaldo, L.P. Jacobson, J.B. Margolick, J. Phair, S.J. O'Brien, and R. Detels. 2003. Association of polymorphisms in human leukocyte antigen class I and transporter associated with antigen processing genes with resistance to human immunodeficiency virus type 1 infection. *J Infect Dis* **187:** 1404-1410.

Liu, R., W.A. Paxton, S. Choe, D. Ceradini, S.R. Martin, R. Horuk, M.E. MacDonald, H. Stuhlmann, R.A. Koup, and N.R. Landau. 1996. Homozygous defect in HIV-1 coreceptor accounts for resistance of some multiply-exposed individuals to HIV-1 infection. *Cell* **86:** 367-377.

Long, E.M., H.L. Martin, Jr., J.K. Kreiss, S.M. Rainwater, L. Lavreys, D.J. Jackson, J. Rakwar, K. Mandaliya, and J. Overbaugh. 2000. Gender differences in HIV-1 diversity at time of infection. *Nat Med* **6:** 71-75.

Louwagie, J., W. Janssens, J. Mascola, L. Heyndrickx, P. Hegerich, G. van der Groen, F.E. McCutchan, and D.S. Burke. 1995. Genetic diversity of the envelope glycoprotein from human immunodeficiency virus type 1 isolates of African origin. *J Virol* **69:** 263-271.

Lyles, R.H., A. Munoz, T.E. Yamashita, H. Bazmi, R. Detels, C.R. Rinaldo, J.B. Margolick, J.P. Phair, and J.W. Mellors. 2000. Natural history of human immunodeficiency virus type 1 viremia after seroconversion and proximal to AIDS in a large cohort of homosexual men. Multicenter AIDS Cohort Study. *J Infect Dis* **181:** 872-880.

Mahalanabis, M., P. Jayaraman, T. Miura, F. Pereyra, E.M. Chester, B. Richardson, B. Walker, and N.L. Haigwood. 2009. Continuous viral escape and selection by autologous neutralizing antibodies in drug-naive human immunodeficiency virus controllers. *J Virol* **83:** 662-672.

Maksiutov, A.Z., A.G. Bachinskii, and S.I. Bazhan. 2002. [Searching for local similarities between HIV-1 and human proteins. Application to vaccines]. *Mol Biol (Mosk)* **36:** 447-459.

Malhotra, U., J. Nolin, J.I. Mullins, and M.J. McElrath. 2007. Comprehensive epitope analysis of cross-clade Gag-specific T-cell responses in individuals with early HIV-1 infection in the US epidemic. *Vaccine* **25:** 381-390.

Mansky, L.M. 1996. Forward mutation rate of human immunodeficiency virus type 1 in a T lymphoid cell line. *AIDS Res Hum Retroviruses* **12:** 307-314.

Margulies, M., M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, L.A. Bemben, J. Berka, M.S. Braverman, Y.J. Chen, Z. Chen, S.B. Dewell, L. Du, J.M. Fierro, X.V.

Gomes, B.C. Godwin, W. He, S. Helgesen, C.H. Ho, G.P. Irzyk, S.C. Jando, M.L. Alenquer, T.P. Jarvie, K.B. Jirage, J.B. Kim, J.R. Knight, J.R. Lanza, J.H. Leamon, S.M. Lefkowitz, M. Lei, J. Li, K.L. Lohman, H. Lu, V.B. Makhijani, K.E. McDade, M.P. McKenna, E.W. Myers, E. Nickerson, J.R. Nobile, R. Plant, B.P. Puc, M.T. Ronan, G.T. Roth, G.J. Sarkis, J.F. Simons, J.W. Simpson, M. Srinivasan, K.R. Tartaro, A. Tomasz, K.A. Vogt, G.A. Volkmer, S.H. Wang, Y. Wang, M.P. Weiner, P. Yu, R.F. Begley, and J.M. Rothberg. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437:** 376-380.

Marrack, P., J. Kappler, and T. Mitchell. 1999. Type I interferons keep activated T cells alive. *J Exp Med* **189:** 521-530.

Martin, A.M., E.M. Freitas, C.S. Witt, and F.T. Christiansen. 2000. The genomic organization and evolution of the natural killer immunoglobulin-like receptor (KIR) gene cluster. *Immunogenetics* **51:** 268-280.

McBurney, S.P. and T.M. Ross. 2008. Viral sequence diversity: challenges for AIDS vaccine designs. *Expert Rev Vaccines* **7:** 1405-1417.

McKinnon, L.R., T.B. Ball, J. Kimani, C. Wachihi, L. Matu, M. Luo, J. Embree, K.R. Fowke, and F.A. Plummer. 2005. Cross-clade CD8(+) T-cell responses with a preference for the predominant circulating clade. *J Acquir Immune Defic Syndr* **40:** 245-249.

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

McNeil, A.C., W.L. Shupert, C.A. Iyasere, C.W. Hallahan, J.A. Mican, R.T. Davey, Jr., and M. Connors. 2001. High-level HIV-1 viremia suppresses viral antigen-specific CD4(+) T cell proliferation. *Proc Natl Acad Sci U S A* **98:** 13878-13883.

Middleton, D., L. Menchaca, H. Rood, and R. Komerofsky. 2003. New allele frequency database: http://www.allelefrequencies.net. *Tissue Antigens* **61:** 403-407.

Migueles, S.A. and M. Connors. 2001. Frequency and function of HIV-specific CD8(+) T cells. *Immunol Lett* **79:** 141-150.

Milich, L., B. Margolin, and R. Swanstrom. 1993. V3 loop of the human immunodeficiency virus type 1 Env protein: interpreting sequence variability. *J Virol* **67:** 5623-5634.

Miotti, P.G., T.E. Taha, N.I. Kumwenda, R. Broadhead, L.A. Mtimavalye, L. Van der Hoeven, J.D. Chiphangwi, G. Liomba, and R.J. Biggar. 1999. HIV transmission through breastfeeding: a study in Malawi. *Jama* **282:** 744-749.

Mohabatkar, H. and S.K. Kar. 2004. Prediction of exposed domains of envelope glycoprotein in Indian HIV-1 isolates and experimental confirmation of their immunogenicity in humans. *Braz J Med Biol Res* **37:** 675-681.

Moore, C.B., M. John, I.R. James, F.T. Christiansen, C.S. Witt, and S.A. Mallal. 2002. Evidence of HIV-1 adaptation to HLA-restricted immune responses at a population level. *Science* **296:** 1439-1443.

Moore, J.P. and J. Sodroski. 1996. Antibody cross-competition analysis of the human immunodeficiency virus type 1 gp120 exterior envelope glycoprotein. *J Virol* **70:** 1863-1872.

Mullins, J.I., D.C. Nickle, L. Heath, A.G. Rodrigo, and G.H. Learn. 2004. Immunogen sequence: the fourth tier of AIDS vaccine design. *Expert Rev Vaccines* **3:** S151-159.

Nakowitsch, S., H. Quendler, H. Fekete, R. Kunert, H. Katinger, and G. Stiegler. 2005. HIV-1 mutants escaping neutralization by the human antibodies 2F5, 2G12, and 4E10: in vitro experiments versus clinical studies. *Aids* **19:** 1957-1966.

Nickle, D.C., M.A. Jensen, G.S. Gottlieb, D. Shriner, G.H. Learn, A.G. Rodrigo, and J.I. Mullins. 2003. Consensus and ancestral state HIV vaccines. *Science* **299:** 1515-1518; author reply 1515-1518.

Nicole Frahm, C.B. 2005. Optimal CTL Epitope Identification in HIV Clade B and Non-Clade B Infection. *HIV Molecular Immunology 2005*.

Nielsen, M., C. Lundegaard, O. Lund, and C. Kesmir. 2005. The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage. *Immunogenetics* **57:** 33-41.

Nijhuis, M., R. Schuurman, D. de Jong, J. Erickson, E. Gustchina, J. Albert, P. Schipper, S. Gulnik, and C.A. Boucher. 1999. Increased fitness of drug resistant HIV-1 protease as a result of acquisition of compensatory mutations during suboptimal therapy. *Aids* **13:** 2349-2359.

Nyambi, P.N., P. Lewi, M. Peeters, W. Janssens, L. Heyndrickx, K. Fransen, K. Andries, M. Vanden Haesevelde, J. Heeney, P. Piot, and G. van der Groen. 1997. Study of the dynamics of neutralization escape mutants in a chimpanzee naturally infected with the simian immunodeficiency virus SIVcpz-ant. *J Virol* **71:** 2320-2330.

O'Connor, D.H., T.M. Allen, T.U. Vogel, P. Jing, I.P. DeSouza, E. Dodds, E.J. Dunphy, C. Melsaether, B. Mothe, H. Yamamoto, H. Horton, N. Wilson, A.L. Hughes, and D.I. Watkins. 2002. Acute phase cytotoxic T lymphocyte escape is a hallmark of simian immunodeficiency virus infection. *Nat Med* **8:** 493-499.

Ostrowski, M.A., Q. Yu, F.Y. Yue, J. Liu, B. Jones, X.X. Gu, M. Loutfy, C.M. Kovacs, and R. Halpenny. 2006. Why can't the immune system control HIV-1? Defining HIV-1-specific CD4+ T cell immunity in order to develop strategies to enhance viral immunity. *Immunol Res* **35:** 89-102.

Pal, R., V.S. Kalyanaraman, B.C. Nair, S. Whitney, T. Keen, L. Hocker, L. Hudacik, N. Rose, I. Mboudjeka, S. Shen, T.H. Wu-Chou, D. Montefiori, J. Mascola, P. Markham, and S. Lu. 2006. Immunization of rhesus macaques with a polyvalent DNA prime/protein boost human immunodeficiency virus type 1 vaccine elicits protective antibody response against simian human immunodeficiency virus of R5 phenotype. *Virology* **348:** 341-353.

Palella, F.J., Jr., K.M. Delaney, A.C. Moorman, M.O. Loveless, J. Fuhrer, G.A. Satten, D.J. Aschman, and S.D. Holmberg. 1998. Declining morbidity and mortality among patients with advanced human immunodeficiency virus infection. HIV Outpatient Study Investigators. *N Engl J Med* **338:** 853-860.

Palmer, S., M. Kearney, F. Maldarelli, E.K. Halvas, C.J. Bixby, H. Bazmi, D. Rock, J. Falloon, R.T. Davey, Jr., R.L. Dewar, J.A. Metcalf, S. Hammer, J.W. Mellors, and J.M. Coffin. 2005. Multiple, linked human immunodeficiency virus type 1 drug resistance mutations in treatment-experienced patients are missed by standard genotype analysis. *J Clin Microbiol* **43:** 406-413.

Pantaleo, G., J.F. Demarest, M. Vaccarezza, C. Graziosi, G.P. Bansal, S. Koenig, and A.S. Fauci. 1995. Effect of anti-V3 antibodies on cell-free and cell-to-cell human immunodeficiency virus transmission. *Eur J Immunol* **25:** 226-231.

Pantophlet, R., I.A. Wilson, and D.R. Burton. 2003. Hyperglycosylated mutants of human immunodeficiency virus (HIV) type 1 monomeric gp120 as novel antigens for HIV vaccine design. *J Virol* **77:** 5889-5901.

Papsidero, L.D., M. Sheu, and F.W. Ruscetti. 1989. Human immunodeficiency virus type 1-neutralizing monoclonal antibodies which react with p17 core protein: characterization and epitope mapping. *J Virol* **63:** 267-272.

Payne, R.P., P.C. Matthews, J.G. Prado, and P.J. Goulder. 2009. HLA-mediated control of HIV and HIV adaptation to HLA. *Adv Parasitol* **68:** 1-20.

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.

Peters, S., M. Munoz, S. Yerly, V. Sanchez-Merino, C. Lopez-Galindez, L. Perrin, B. Larder, D. Cmarko, S. Fakan, P. Meylan, and A. Telenti. 2001. Resistance to nucleoside analog reverse transcriptase inhibitors mediated by human immunodeficiency virus type 1 p6 protein. *J Virol* **75:** 9644-9653.

Photinl kiepiela, A.J.L., Isobell Honeyborne, Philip J.R.Goulder. 2004. Dominant influence of HLA-B in mediating the potential co-evolution of HIV and HLA. *Nature* **432:** 769-774.

Pilgrim, A.K., G. Pantaleo, O.J. Cohen, L.M. Fink, J.Y. Zhou, J.T. Zhou, D.P. Bolognesi, A.S. Fauci, and D.C. Montefiori. 1997. Neutralizing antibody responses to human immunodeficiency virus type 1 in primary infection and long-term-nonprogressive infection. *J Infect Dis* **176:** 924-932.

Ping, L.H., J.A. Nelson, I.F. Hoffman, J. Schock, S.L. Lamers, M. Goodman, P. Vernazza, P. Kazembe, M. Maida, D. Zimba, M.M. Goodenow, J.J. Eron, Jr., S.A. Fiscus, M.S. Cohen, and R. Swanstrom. 1999. Characterization of V3 sequence heterogeneity in subtype C human immunodeficiency virus type 1 isolates from Malawi: underrepresentation of X4 variants. *J Virol* **73:** 6271-6281.

Pinter, C., A.G. Siccardi, R. Longhi, and A. Clivio. 1995. Direct interaction of complement factor H with the C1 domain of HIV type 1 glycoprotein 120. *AIDS Res Hum Retroviruses* **11:** 577-588.

Pitcher, C.J., C. Quittner, D.M. Peterson, M. Connors, R.A. Koup, V.C. Maino, and L.J. Picker. 1999. HIV-1-specific CD4+ T cells are detectable in most individuals with active HIV-1 infection, but decline with prolonged viral suppression. *Nat Med* **5:** 518-525.

Planz, O., S. Ehl, E. Furrer, E. Horvath, M.A. Brundler, H. Hengartner, and R.M. Zinkernagel. 1997. A critical role for neutralizing-antibody-producing B cells, CD4(+) T cells, and interferons in persistent and acute infections of mice with lymphocytic choriomeningitis virus: implications for adoptive immunotherapy of virus carriers. *Proc Natl Acad Sci U S A* **94:** 6874-6879.

Pond, S.L. and S.D. Frost. 2005. Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics* **21:** 2531-2533.

Pond, S.L., S.D. Frost, Z. Grossman, M.B. Gravenor, D.D. Richman, and A.J. Brown. 2006. Adaptation to different human populations by HIV-1 revealed by codon-based analyses. *PLoS Comput Biol* **2:** e62.

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.

Price, D.A., P.J. Goulder, P. Klenerman, A.K. Sewell, P.J. Easterbrook, M. Troop, C.R. Bangham, and R.E. Phillips. 1997. Positive selection of HIV-1 cytotoxic T lymphocyte escape variants during primary infection. *Proc Natl Acad Sci U S A* **94:** 1890-1895.

Propato, A., E. Schiaffella, E. Vicenzi, V. Francavilla, L. Baloni, M. Paroli, L. Finocchi, N. Tanigaki, S. Ghezzi, R. Ferrara, R. Chesnut, B. Livingston, A. Sette, R. Paganelli, F. Aiuti, G. Poli, and V. Barnaba. 2001. Spreading of HIV-specific CD8+ T-cell repertoire in long-term nonprogressors and its role in the control of viral load and disease activity. *Hum Immunol* **62:** 561-576.

Purtscher, M., A. Trkola, G. Gruber, A. Buchacher, R. Predl, F. Steindl, C. Tauer, R. Berger, N. Barrett, A. Jungbauer, and et al. 1994. A broadly neutralizing human monoclonal antibody against gp41 of human immunodeficiency virus type 1. *AIDS Res Hum Retroviruses* **10:** 1651-1658.

Quinn, T.C., M.J. Wawer, N. Sewankambo, D. Serwadda, C. Li, F. Wabwire-Mangen, M.O. Meehan, T. Lutalo, and R.H. Gray. 2000. Viral load and heterosexual transmission of human immunodeficiency virus type 1. Rakai Project Study Group. *N Engl J Med* **342:** 921-929.

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

Redfield, R.R., D.C. Wright, and E.C. Tramont. 1986. The Walter Reed staging classification for HTLV-III/LAV infection. *N Engl J Med* **314:** 131-132.

Reimann, K.A., J.T. Li, R. Veazey, M. Halloran, I.W. Park, G.B. Karlsson, J. Sodroski, and N.L. Letvin. 1996. A chimeric simian/human immunodeficiency virus expressing a primary patient human immunodeficiency virus type 1 isolate env causes an AIDS-like disease after in vivo passage in rhesus monkeys. *J Virol* **70:** 6922-6928.

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.

Reynolds, M.R., E. Rakasz, P.J. Skinner, C. White, K. Abel, Z.M. Ma, L. Compton, G. Napoe, N. Wilson, C.J. Miller, A. Haase, and D.I. Watkins. 2005. CD8+ T-lymphocyte response to major immunodominant epitopes after vaginal exposure to simian immunodeficiency virus: too late and too little. *J Virol* **79:** 9228-9235.

Richard Wyatt, P.D.K., Wayne A. Hendrickson, and Joseph G. Sodroski. 2007. Structure of the Core of the HIV-1 gp120 Exterior Envelope Glycoprotein. Los Alomas National Laboratory.

Richman, D.D., T. Wrin, S.J. Little, and C.J. Petropoulos. 2003. Rapid evolution of the neutralizing antibody response to HIV type 1 infection. *Proc Natl Acad Sci U S A* **100:** 4144-4149.

Ritola, K., C.D. Pilcher, S.A. Fiscus, N.G. Hoffman, J.A. Nelson, K.M. Kitrinos, C.B. Hicks, J.J. Eron, Jr., and R. Swanstrom. 2004. Multiple V1/V2 env variants are frequently present during primary infection with human immunodeficiency virus type 1. *J Virol* **78:** 11208-11218.

Rogge, L., D. D'Ambrosio, M. Biffi, G. Penna, L.J. Minetti, D.H. Presky, L. Adorini, and F. Sinigaglia. 1998. The role of Stat4 in species-specific regulation of Th cell development by type I IFNs. *J Immunol* **161:** 6567-6574.

Ross, H.A. and A.G. Rodrigo. 2002. Immune-mediated positive selection drives human immunodeficiency virus type 1 molecular variation and predicts disease duration. *J Virol* **76:** 11715-11720.

Rousseau, C.M., M.G. Daniels, J.M. Carlson, C. Kadie, H. Crawford, A. Prendergast, P. Matthews, R. Payne, M. Rolland, D.N. Raugi, B.S. Maust, G.H. Learn, D.C. Nickle, H. Coovadia, T. Ndung'u, N. Frahm, C. Brander, B.D. Walker, P.J. Goulder, T. Bhattacharya, D.E. Heckerman, B.T. Korber, and J.I. Mullins. 2008. HLA class I-driven evolution of human immunodeficiency virus type 1 subtype c proteome: immune escape and viral load. *J Virol* **82:** 6434-6446.

Rousseau, C.M., D.W. Lockhart, J. Listgarten, S.N. Maley, C. Kadie, G.H. Learn, D.C. Nickle, D.E. Heckerman, W. Deng, C. Brander, T. Ndung'u, H. Coovadia, P.J. Goulder, B.T. Korber, B.D. Walker, and J.I. Mullins. 2009. Rare HLA drive additional HIV evolution compared to more frequent alleles. *AIDS Res Hum Retroviruses* **25:** 297-303.

Rozera, G., I. Abbate, A. Bruselles, C. Vlassi, G. D'Offizi, P. Narciso, G. Chillemi, M. Prosperi, G. Ippolito, and M.R. Capobianchi. 2009. Massively parallel pyrosequencing highlights minority variants in the HIV-1 env quasispecies deriving from lymphomonocyte sub-populations. *Retrovirology* **6:** 15.

Saifuddin, M., T. Hedayati, J.P. Atkinson, M.H. Holguin, C.J. Parker, and G.T. Spear. 1997. Human immunodeficiency virus type 1 incorporates both glycosyl phosphatidylinositol-anchored CD55 and CD59 and integral membrane CD46 at levels that protect from complement-mediated destruction. *J Gen Virol* **78 (Pt 8):** 1907-1911.

Salazar-Gonzalez, J.F., E. Bailes, K.T. Pham, M.G. Salazar, M.B. Guffey, B.F. Keele, C.A. Derdeyn, P. Farmer, E. Hunter, S. Allen, O. Manigart, J. Mulenga, J.A. Anderson, R. Swanstrom, B.F. Haynes, G.S. Athreya, B.T. Korber, P.M. Sharp, G.M. Shaw, and B.H. Hahn. 2008. Deciphering human immunodeficiency virus type 1 transmission and early envelope diversification by single-genome amplification and sequencing. *J Virol* **82:** 3952-3970.

Sanders, R.W., M. Venturi, L. Schiffner, R. Kalyanaraman, H. Katinger, K.O. Lloyd, P.D. Kwong, and J.P. Moore. 2002. The mannose-dependent epitope for neutralizing antibody 2G12 on human immunodeficiency virus type 1 glycoprotein gp120. *J Virol* **76:** 7293-7305.

Sawyer, L.A., D.A. Katzenstein, R.M. Hendry, E.J. Boone, L.K. Vujcic, C.C. Williams, S.L. Zeger, A.J. Saah, C.R. Rinaldo, Jr., J.P. Phair, and et al. 1990. Possible beneficial effects of neutralizing antibodies and antibody-dependent, cell-mediated cytotoxicity in human immunodeficiency virus infection. *AIDS Res Hum Retroviruses* **6:** 341-356.

Scanlan, C.N., R. Pantophlet, M.R. Wormald, E. Ollmann Saphire, R. Stanfield, I.A. Wilson, H. Katinger, R.A. Dwek, P.M. Rudd, and D.R. Burton. 2002. The broadly neutralizing anti-human immunodeficiency virus type 1 antibody 2G12 recognizes a cluster of alpha1-->2 mannose residues on the outer face of gp120. *J Virol* **76:** 7306-7321.

Scherer, A., J. Frater, A. Oxenius, J. Agudelo, D.A. Price, H.F. Gunthard, M. Barnardo, L. Perrin, B. Hirschel, R.E. Phillips, and A.R. McLean. 2004. Quantifiable cytotoxic T lymphocyte responses and HLA-related risk of progression to AIDS. *Proc Natl Acad Sci U S A* **101:** 12266-12270.

Schmitz, J., J.P. Zimmer, B. Kluxen, S. Aries, M. Bogel, I. Gigli, and H. Schmitz. 1995. Antibody-dependent complement-mediated cytotoxicity in sera from patients with HIV-1 infection is controlled by CD55 and CD59. *J Clin Invest* **96:** 1520-1526.

Schmitz, J.E., R.P. Johnson, H.M. McClure, K.H. Manson, M.S. Wyand, M.J. Kuroda, M.A. Lifton, R.S. Khunkhun, K.J. McEvers, J. Gillis, M. Piatak, J.D. Lifson, G. Grosschupff, P. Racz, K. Tenner-Racz, E.P. Rieber, K. Kuus-Reichel, R.S. Gelman, N.L. Letvin, D.C. Montefiori, R.M. Ruprecht, R.C. Desrosiers, and K.A. Reimann. 2005. Effect of CD8+ lymphocyte depletion on virus containment after simian immunodeficiency virus SIVmac251 challenge of live attenuated SIVmac239delta3-vaccinated rhesus macaques. *J Virol* **79:** 8131-8141.

Schneider E, G.M., Kajese T, McKenna MT. 2006. Epidemiology of HIV/AIDS--United States, 1981-2005. *MMWR Morb Mortal Wkly Rep* **55:** 589-592.

Schnell, G., W.L. Ince, and R. Swanstrom. 2008. Identification and recovery of minor HIV-1 variants using the heteroduplex tracking assay and biotinylated probes. *Nucleic Acids Res* **36:** e146.

Schuitemaker, H., M. Koot, N.A. Kootstra, M.W. Dercksen, R.E. de Goede, R.P. van Steenwijk, J.M. Lange, J.K. Schattenkerk, F. Miedema, and M. Tersmette. 1992. Biological phenotype of human immunodeficiency virus type 1 clones at different stages of infection: progression of disease is associated with a shift from monocytotropic to T-cell-tropic virus population. *J Virol* **66:** 1354-1360.

Selin, L.K., P.A. Santolucito, A.K. Pinto, E. Szomolanyi-Tsuda, and R.M. Welsh. 2001. Innate immunity to viruses: control of vaccinia virus infection by gamma delta T cells. *J Immunol* **166:** 6784-6794.

Serwanga, J., L.A. Shafer, E. Pimego, B. Auma, C. Watera, S. Rowland, D. Yirrell, P. Pala, H. Grosskurth, J. Whitworth, F. Gotch, and P. Kaleebu. 2009. Host HLA B*allele-associated multi-clade Gag T-cell recognition correlates with slow HIV-1 disease progression in antiretroviral therapy-naive Ugandans. *PLoS One* **4:** e4188.

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

Shankarappa, R., J.B. Margolick, S.J. Gange, A.G. Rodrigo, D. Upchurch, H. Farzadegan, P. Gupta, C.R. Rinaldo, G.H. Learn, X. He, X.L. Huang, and J.I. Mullins. 1999. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *J Virol* **73:** 10489-10502.

Shedlock, D.J. and H. Shen. 2003. Requirement for CD4 T cell help in generating functional CD8 T cell memory. *Science* **300:** 337-339.

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.

Shendure, J. and H. Ji. 2008. Next-generation DNA sequencing. *Nat Biotechnol* **26:** 1135-1145.

Shibata, J., K. Yoshimura, A. Honda, A. Koito, T. Murakami, and S. Matsushita. 2007. Impact of V2 mutations on escape from a potent neutralizing anti-V3 monoclonal antibody during in vitro selection of a primary human immunodeficiency virus type 1 isolate. *J Virol* **81:** 3757-3768.

Shriner, D., D.C. Nickle, M.A. Jensen, and J.I. Mullins. 2003. Potential impact of recombination on sitewise approaches for detecting positive natural selection. *Genet Res* **81:** 115-121.

Sidney, J., B. Peters, N. Frahm, C. Brander, and A. Sette. 2008. HLA class I supertypes: a revised and updated classification. *BMC Immunol* **9:** 1.

Simen, B.B., J.F. Simons, K.H. Hullsiek, R.M. Novak, R.D. Macarthur, J.D. Baxter, C. Huang, C. Lubeski, G.S. Turenchalk, M.S. Braverman, B. Desany, J.M. Rothberg, M. Egholm, and M.J. Kozal. 2009. Low-Abundance Drug-Resistant Viral Variants in Chronically HIV-Infected, Antiretroviral Treatment-Naive Patients Significantly Impact Treatment Outcomes. *J Infect Dis* **199:** 693-701.

Slifka, M.K. and J.L. Whitton. 2001. Functional avidity maturation of CD8(+) T cells without selection of higher affinity TCR. *Nat Immunol* **2:** 711-717.

Slobod, K.S., C. Coleclough, S.A. Brown, J. Stambas, X. Zhan, S. Surman, B.G. Jones, A. Zirkel, P.J. Freiden, B. Brown, R. Sealy, M. Bonsignori, and J.L. Hurwitz. 2005. Clade, Country and Region-specific HIV-1 Vaccines: Are they necessary? *AIDS Res Ther* **2:** 3.

Spear, G.T., G.G. Olinger, M. Saifuddin, and H.M. Gebel. 2001. Human antibodies to major histocompatibility complex alloantigens mediate lysis and neutralization of HIV-1 primary isolate virions in the presence of complement. *J Acquir Immune Defic Syndr* **26:** 103-110.

Srivastava, I.K., J.B. Ulmer, and S.W. Barnett. 2005. Role of neutralizing antibodies in protective immunity against HIV. *Hum Vaccin* **1:** 45-60.

Stanhope, P.E., A.Y. Liu, W. Pavlat, P.M. Pitha, M.L. Clements, and R.F. Siliciano. 1993. An HIV-1 envelope protein vaccine elicits a functionally complex human CD4+ T cell response that includes cytolytic T lymphocytes. *J Immunol* **150:** 4672-4686.

Stephens, H.A. 2005. HIV-1 diversity versus HLA class I polymorphism. *Trends Immunol* **26:** 41-47.

Stewart, J.J., P. Watts, and S. Litwin. 2001. An algorithm for mapping positively selected members of quasispecies-type viruses. *BMC Bioinformatics* **2:** 1.

Stiegler, G., R. Kunert, M. Purtscher, S. Wolbank, R. Voglauer, F. Steindl, and H. Katinger. 2001. A potent cross-clade neutralizing human monoclonal antibody against a novel epitope on gp41 of human immunodeficiency virus type 1. *AIDS Res Hum Retroviruses* **17:** 1757-1765.

Stranford, S.A., J. Skurnick, D. Louria, D. Osmond, S.Y. Chang, J. Sninsky, G. Ferrari, K. Weinhold, C. Lindquist, and J.A. Levy. 1999. Lack of infection in HIV-exposed individuals is associated with a strong CD8(+) cell noncytotoxic anti-HIV response. *Proc Natl Acad Sci U S A* **96:** 1030-1035.

Stremlau, M., C.M. Owens, M.J. Perron, M. Kiessling, P. Autissier, and J. Sodroski. 2004. The cytoplasmic body component TRIM5alpha restricts HIV-1 infection in Old World monkeys. *Nature* **427:** 848-853.

Sun, J.C. and M.J. Bevan. 2003. Defective CD8 T cell memory following acute infection without CD4 T cell help. *Science* **300:** 339-342.

Suzuki, Y. and T. Gojobori. 1999. A method for detecting positive selection at single amino acid sites. *Mol Biol Evol* **16:** 1315-1328.

Taha, T.E., D.R. Hoover, G.A. Dallabetta, N.I. Kumwenda, L.A. Mtimavalye, L.P. Yang, G.N. Liomba, R.L. Broadhead, J.D. Chiphangwi, and P.G. Miotti. 1998. Bacterial vaginosis and disturbances of vaginal flora: association with increased acquisition of HIV. *Aids* **12:** 1699-1706.

Tang, J., S. Tang, E. Lobashevsky, A.D. Myracle, U. Fideli, G. Aldrovandi, S. Allen, R. Musonda, and R.A. Kaslow. 2002. Favorable and unfavorable HLA class I alleles and haplotypes in Zambians predominantly infected with clade C human immunodeficiency virus type 1. *J Virol* **76:** 8276-8284.

Taylor, B.S., M.E. Sobieszczyk, F.E. McCutchan, and S.M. Hammer. 2008. The challenge of HIV-1 subtype diversity. *N Engl J Med* **358:** 1590-1602.

Thompson, J.D., D.G. Higgins, and T.J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22:** 4673-4680.

Tilton, J.C., M.R. Luskin, A.J. Johnson, M. Manion, C.W. Hallahan, J.A. Metcalf, M. McLaughlin, R.T. Davey, Jr., and M. Connors. 2007. Changes in paracrine interleukin-2 requirement, CCR7 expression, frequency, and cytokine secretion of human immunodeficiency virus-specific CD4+ T cells are a consequence of antigen load. *J Virol* **81:** 2713-2725.

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

Travers, S.A., M.J. O'Connell, G.P. McCormack, and J.O. McInerney. 2005. Evidence for heterogeneous selective pressures in the evolution of the env gene in different human immunodeficiency virus type 1 subtypes. *J Virol* **79:** 1836-1841.

Trkola, A., T. Dragic, J. Arthos, J.M. Binley, W.C. Olson, G.P. Allaway, C. Cheng-Mayer, J. Robinson, P.J. Maddon, and J.P. Moore. 1996. CD4-dependent, antibody-sensitive interactions between HIV-1 and its co-receptor CCR-5. *Nature* **384:** 184-187.

Trkola, A., H. Kuster, P. Rusert, B. Joos, M. Fischer, C. Leemann, A. Manrique, M. Huber, M. Rehr, A. Oxenius, R. Weber, G. Stiegler, B. Vcelar, H. Katinger, L. Aceto, and H.F. Gunthard. 2005. Delay of HIV-1 rebound after cessation of antiretroviral therapy through passive transfer of human neutralizing antibodies. *Nat Med* **11:** 615-622.

Turnbull, E.L., A.R. Lopes, N.A. Jones, D. Cornforth, P. Newton, D. Aldam, P. Pellegrino, J. Turner, I. Williams, C.M. Wilson, P.A. Goepfert, M.K. Maini, and P. Borrow. 2006. HIV-1 epitope-specific CD8+ T cell responses strongly associated

with delayed disease progression cross-recognize epitope variants efficiently. *J Immunol* **176:** 6130-6146.

UNAIDS. 2007. AIDS epidemic update.

Veazey, R.S., M. DeMaria, L.V. Chalifoux, D.E. Shvetz, D.R. Pauley, H.L. Knight, M. Rosenzweig, R.P. Johnson, R.C. Desrosiers, and A.A. Lackner. 1998. Gastrointestinal tract as a major site of CD4+ T cell depletion and viral replication in SIV infection. *Science* **280:** 427-431.

Wainberg, M.A. 2004. HIV-1 subtype distribution and the problem of drug resistance. *Aids* **18 Suppl 3:** S63-68.

Wainberg, M.A. and K.T. Jeang. 2008. 25 years of HIV-1 research - progress and perspectives. *BMC Med* **6:** 31.

Walker, B.D. and D.R. Burton. 2008. Toward an AIDS vaccine. *Science* **320:** 760-764.

Wang, C., Y. Mitsuya, B. Gharizadeh, M. Ronaghi, and R.W. Shafer. 2007. Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. *Genome Res* **17:** 1195-1201.

Wang, Y., L. Tao, E. Mitchell, C. Bravery, P. Berlingieri, P. Armstrong, R. Vaughan, J. Underwood, and T. Lehner. 1999. Allo-immunization elicits CD8+ T cell-derived chemokines, HIV suppressor factors and resistance to HIV infection in women. *Nat Med* **5:** 1004-1009.

Wawer, M.J., R.H. Gray, N.K. Sewankambo, D. Serwadda, X. Li, O. Laeyendecker, N. Kiwanuka, G. Kigozi, M. Kiddugavu, T. Lutalo, F. Nalugoda, F. Wabwire-Mangen, M.P. Meehan, and T.C. Quinn. 2005. Rates of HIV-1 transmission per coital act, by stage of HIV-1 infection, in Rakai, Uganda. *J Infect Dis* **191:** 1403-1409.

Weaver, E.A., Z. Lu, Z.T. Camacho, F. Moukdar, H.X. Liao, B.J. Ma, M. Muldoon, J. Theiler, G.J. Nabel, N.L. Letvin, B.T. Korber, B.H. Hahn, B.F. Haynes, and F. Gao. 2006. Cross-subtype T-cell immune responses induced by a human immunodeficiency virus type 1 group m consensus env immunogen. *J Virol* **80:** 6745-6756.

Weber, J. 2001. The pathogenesis of HIV-1 infection. *Br Med Bull* **58:** 61-72.

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.

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.

Wherry, E.J., J.N. Blattman, K. Murali-Krishna, R. van der Most, and R. Ahmed. 2003. Viral persistence alters CD8 T-cell immunodominance and tissue distribution and results in distinct stages of functional impairment. *J Virol* **77:** 4911-4927.

Williamson, S. 2003. Adaptation in the env gene of HIV-1 and evolutionary theories of disease progression. *Mol Biol Evol* **20:** 1318-1325.

Wu, L., N.P. Gerard, R. Wyatt, H. Choe, C. Parolin, N. Ruffing, A. Borsetti, A.A. Cardoso, E. Desjardin, W. Newman, C. Gerard, and J. Sodroski. 1996. CD4-induced interaction of primary HIV-1 gp120 glycoproteins with the chemokine receptor CCR-5. *Nature* **384:** 179-183.

Wu, X., J.L. Anderson, E.M. Campbell, A.M. Joseph, and T.J. Hope. 2006. Proteasome inhibitors uncouple rhesus TRIM5alpha restriction of HIV-1 reverse transcription and infection. *Proc Natl Acad Sci U S A* **103:** 7465-7470.

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

Xie, B., C.F. Invernizzi, S. Richard, and M.A. Wainberg. 2007. Arginine methylation of the human immunodeficiency virus type 1 Tat protein by PRMT6 negatively affects Tat Interactions with both cyclin T1 and the Tat transactivation region. *J Virol* **81:** 4226-4234.

Yamaguchi-Kabata, Y. and T. Gojobori. 2000. Reevaluation of amino acid variability of the human immunodeficiency virus type 1 gp120 envelope glycoprotein and prediction of new discontinuous epitopes. *J Virol* **74:** 4335-4350.

Yamaguchi-Kabata, Y., M. Yamashita, S. Ohkura, M. Hayami, and T. Miura. 2004. Linkage of amino acid variation and evolution of human immunodeficiency virus type 1 gp120 envelope glycoprotein (subtype B) with usage of the second receptor. *J Mol Evol* **58:** 333-340.

Yang, W., J.P. Bielawski, and Z. Yang. 2003. Widespread adaptive evolution in the human immunodeficiency virus type 1 genome. *J Mol Evol* **57:** 212-221.

Yang, X., M. Farzan, R. Wyatt, and J. Sodroski. 2000a. Characterization of stable, soluble trimers containing complete ectodomains of human immunodeficiency virus type 1 envelope glycoproteins. *J Virol* **74:** 5716-5725.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* **13:** 555-556.

Yang, Z. 2001. Maximum likelihood analysis of adaptive evolution in HIV-1 gp120 env gene. *Pac Symp Biocomput***:** 226-237.

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

Yerly, S., S. Jost, A. Telenti, M. Flepp, L. Kaiser, J.P. Chave, P. Vernazza, M. Battegay, H. Furrer, B. Chanzy, P. Burgisser, M. Rickenbach, M. Gebhardt, M.C. Bernard, T. Perneger, B. Hirschel, and L. Perrin. 2004. Infrequent transmission of HIV-1 drug-resistant variants. *Antivir Ther* **9:** 375-384.

Younes, S.A., B. Yassine-Diab, A.R. Dumont, M.R. Boulassel, Z. Grossman, J.P. Routy, and R.P. Sekaly. 2003. HIV-1 viremia prevents the establishment of interleukin 2-producing HIV-specific memory CD4+ T cells endowed with proliferative capacity. *J Exp Med* **198:** 1909-1922.

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

Zeffman, A., S. Hassard, G. Varani, and A. Lever. 2000. The major HIV-1 packaging signal is an extended bulged stem loop whose structure is altered on interaction with the Gag polyprotein. *J Mol Biol* **297:** 877-893.

Zennou, V., C. Petit, D. Guetard, U. Nerhbass, L. Montagnier, and P. Charneau. 2000. HIV-1 genome nuclear import is mediated by a central DNA flap. *Cell* **101:** 173-185.

Zhang, D., P. Shankar, Z. Xu, B. Harnisch, G. Chen, C. Lange, S.J. Lee, H. Valdez, M.M. Lederman, and J. Lieberman. 2003. Most antiviral CD8 T cells during chronic viral infection do not express high levels of perforin and are not directly cytotoxic. *Blood* **101:** 226-235.

Zhang, Z.Q., S.W. Wietgrefe, Q. Li, M.D. Shore, L. Duan, C. Reilly, J.D. Lifson, and A.T. Haase. 2004. Roles of substrate availability and infection of resting and activated CD4+ T cells in transmission and acute simian immunodeficiency virus infection. *Proc Natl Acad Sci U S A* **101:** 5640-5645.

Zhu, T., H. Mo, N. Wang, D.S. Nam, Y. Cao, R.A. Koup, and D.D. Ho. 1993. Genotypic and phenotypic characterization of HIV-1 patients with primary infection. *Science* **261:** 1179-1181.

Zimbwa, P., A. Milicic, J. Frater, T.J. Scriba, A. Willis, P.J. Goulder, T. Pillay, H. Gunthard, J.N. Weber, H.T. Zhang, and R.E. Phillips. 2006. Precise Identification of an Hiv-1 Antigen Processing Mutant. *J Virol* **81:** 2031-2038.

Zinkernagel, R.M., A. LaMarre, A. Ciurea, L. Hunziker, A.F. Ochsenbein, K.D. McCoy, T. Fehr, M.F. Bachmann, U. Kalinke, and H. Hengartner. 2001. Neutralizing antiviral antibody responses. *Adv Immunol* **79:** 1-53.

Zwick, M.B., R. Jensen, S. Church, M. Wang, G. Stiegler, R. Kunert, H. Katinger, and D.R. Burton. 2005. Anti-human immunodeficiency virus type 1 (HIV-1) antibodies 2F5 and 4E10 require surprisingly few crucial residues in the membrane-proximal external region of glycoprotein gp41 to neutralize HIV-1. *J Virol* **79:** 1252-1261.

6.0 Appendices

6.1 Appendix A

Abbreviations

Ab	antibody
ADCC	antibody-dependant cellular cytotoxicity
AHF	average HLA class I allele frequency
AIDS	acquired immunodeficiency syndrome
APCs	antigen-presenting cells
APOBEC3G	catalytic polypeptide-like 3G
ART	antiviral therapy
CFH	complement factor H
CRFs	circulating recombinant forms
CSW	commercial sex workers
CTL	CD8+ cytotoxic T-lymphocytes
DCs	dendric cells
dN	non-synonymous distances
DNA	deoxyribo nucleic acid
dS	synonymous distances
FDA	the US Food and Drug Administration
FEL	fixed-effects likelihood
HAART	highly active antiretroviral therapy
HIV-1	human immunodeficiency virus type 1
HLA	human leukocyte antigen
HTLV-III	human T-cell Leukemia virus III
HXB2	HIV-1 reference sequence
IDUs	intravenous drug users
IFN	interferin
IN	integrase
KIR	killer immunoglobulin-like receptor
LAV	lymphadenopathy-associated virus
LTNPs	long-term non progressors
LTR	long-terminal repeats
MEGA	Molecular Evolutionary Genetics Analysis
MHC	major histocompatibility complex
MIP	macrophage inflammatory protein
MPER	membrane proximal external region
MSM	men who have sex with men
MTCT	mother-to-child transmission
Nabs	neutralizing antibodies
NK	natural killer
OHF	overall specific HLA class I allele frequency
PCP	pneumocystis carinii pneumonia

PCS	proteasomal cleavage site
PIC	preintegration complex
PR	protease
PRE	*Rev* responsive element
PRMT6	protein arginine methyltransferase
PS	positive selection
PT	parenteral transmission
REL	random-effects likelihood
Rev	regulator of expression of virus protein
RNA	ribo nucleic acid
RT	reverse transcriptase
SIV	simian immunodeficiency virus
SLAC	single-likelihood ancestor counting
ST	sexual transmission
STDs	sexual transmitted diseases
Tat	transactivator protein
TCRs	T-cell receptors
Th	T helper
TRIM5α	tripartite motif protein 5α
Vif	infectivity factor
Vpu	virus protein
454	454 Life Science

6.2 Appendix B

Perl Script 1

```
#=============================================================================
#  This Perl Script will read sequences in a file and randomly choose the required number of sequences
#  and  print them out in anther file; developed by Binhua Liang;2007
#=============================================================================

# file pathes
$infile_name = "H:/In_Seq.fas";
$outfile_name="H:/Out_Seq.fas";

#declare variables
$writeoutSeq = "";          #  sequence holder, which will be printed out
$writeoutSeq_Name = "";  #  sequence name holder, which will be printed out
@seqArray = ();             #  array for holding numbers, which indicate the sequence positions in input file
$seqCount = 0;              #  count the numbers of sequences in input file
$position =0;               #  for the position of sequence in input file
$arrayLength = 0;

# open files
open INFILE, "<$infile_name" or die "$infile_name did not open";
open OUTFILE,">$outfile_name" or die "$outfile_name did not open";

# count the numbers of sequneces in the in_put file
while (<INFILE>){
        $line = $_;chop $line;$firstLetter = substr($line,0,1);
        if ($firstLetter eq ">"){
                $seqCount ++;
        }# end of if
}# end of while
close(INFILE);

# create an array holding non_redundent random numbers, which represent the positions of each sequence
$random = int( rand($seqCount))+1;push(@seqArray,$random);

# do the work
while ($arrayLength < 60){
        $ifDuplicate = 0;$random = int( rand($seqCount))+1;

        for ($i=0;$i < scalar(@seqArray); ++$i){# check just created random number against array,which
                                                # holds all created numbers and find out if it is a
                $item = $seqArray[$i];          # duplicate or not
                if ($random == $item){
                        $ifDuplicate++;
                }
        }
        if ($ifDuplicate == 0){                 # if not duplicate, add this random number into array
                push(@seqArray,$random);
        }
        $arrayLength = scalar(@seqArray);
}# end of while
```

```
# read in sequence in a order and print sequence out
for ($i=0;$i < scalar(@seqArray); ++$i){
	$position = $seqArray[$i];$isPrint = 0;
	open INFILE, "<$infile_name" or die "$infile_name did not open";
	$countSeq = 0;
	while (<INFILE>){
		$line = $_;chop $line;$firstLetter = substr($line,0,1);
		if ($isPrint == 0 && $firstLetter eq ">"){# read in the write out sequence name and print
								# it out
			$countSeq ++;
			if ($countSeq == $position){
				$writeoutSeq_Name = $line;$isPrint = 1;
				print OUTFILE ("$writeoutSeq_Name\n");
			}
		}# end of if
		elsif ($isPrint == 1) {			# read in the write out sequence and print it out
			$writeoutSeq = $line;$isPrint = 0;print OUTFILE ("$writeoutSeq\n");
		}# end of elsif
	}# end while loop
	close(INFILE);
}# end for loop
close(INFILE);close(OUTFILE);

# == end of script1 ==========================================================
```

Perl Script 2

```
#============================================================================
#  This Perl Script is used to calculate HLA allele frequencies in the whole population; developed by
#  Binhua Liang; 2008
#============================================================================

#  declare variables
$output = ""; $HLA_freq = 0; $numb_Aus = 15; $numb_Eur = 71;$numb_Nam = 64;$numb_Nea = 21;
$numb_Oce = 3;$numb_Sam = 42;$numb_Sea = 86;$numb_Swa = 47;$numb_Ssa = 265;
@seqArray = ();

#  files
$infile_name = "H:/frequency_table.txt";
$outfile_name="H:/frequency_out.txt";

open INFILE, "<$infile_name" or die "$infile_name did not open";
open OUTFILE,">$outfile_name" or die "$outfile_name did not open";

# Read in sequences and compares them and removes duplicates
while (<INFILE>){
	$line = $_;chop $line;@seqArray = split(/\t/, $line);
	for ($i = 0; $i < scalar(@seqArray); ++$i){
		if ($i == 0){
			$output = $seqArray[$i];print OUTFILE ("$output\n");
		}
		elsif ($i == 1){
```

```
                $HLA_freq = ($seqArray[$i])*$numb_Aus;print OUTFILE ("$HLA_freq\n");
        }
        elsif ($i == 2){
                $HLA_freq = $HLA_freq + $seqArray[$i]*$numb_Eur;
                print OUTFILE ("$HLA_freq\n");
        }
        elsif ($i == 3){
                $HLA_freq = $HLA_freq + ($seqArray[$i])*$numb_Nam;
                print OUTFILE ("$HLA_freq\n");
        }
        elsif ($i == 4){
                $HLA_freq = $HLA_freq + ($seqArray[$i])*$numb_Nea;
                print OUTFILE ("$HLA_freq\n");
        }
        elsif ($i == 5){
                $HLA_freq = $HLA_freq + ($seqArray[$i])*$numb_Oce;
                print OUTFILE ("$HLA_freq\n");
        }
        elsif ($i == 6){
                $HLA_freq = $HLA_freq + ($seqArray[$i])*$numb_Sam;
                print OUTFILE ("$HLA_freq\n");
        }
        elsif ($i == 7){
                $HLA_freq = $HLA_freq + ($seqArray[$i])*$numb_Sea;
                print OUTFILE ("$HLA_freq\n");
        }
        elsif ($i == 8){
                $HLA_freq = $HLA_freq + ($seqArray[$i])*$numb_Swa;
                print OUTFILE ("$HLA_freq\n");
        }
        else {
                $HLA_freq = $HLA_freq + ($seqArray[$i])*$numb_Ssa;
                print OUTFILE ("$HLA_freq\n");
        }
    }# end of for loop
    $HLA_freq = $HLA_freq/614;
    $output = $output."\t".$HLA_freq;print OUTFILE ("$output\n");$output = "";$HLA_freq = 0;
}# end of while loop
close(INFILE);close(OUTFILE);

# == end of script2 ============================================================
```

Perl Script 3

```
#==============================================================================
#   This Perl Script calculate average HLA frequency on each CTL epitope; developed by Binhua Liang;
#   2008
#==============================================================================

#  file pathes
$infile_name1 = "F:/HLA_freq_table.txt";    #  Frequency file
$infile_name2 = "F:/In.txt";                #  HLA allels map file
$outfile_name="F:/Out_HLA_Freq_result.txt";
```

```
#  declare variables
$output = "";$HLA_name = "";$HLA_name_temp = "";$HLA_freq = 0;$HLA_freq_temp = 0;
$seqLength = 0;$token = "";$token_HLA_name = "";$token_HLA_freq_numb = 0;
@frequency_name = ();@frequency = ();$seqArray = ();@tmp = ();

#open files
open INFILE1, "<$infile_name1" or die "$infile_name 1 did not open";
open INFILE2, "<$infile_name2" or die "$infile_name 2 did not open";
open OUTFILE,">$outfile_name" or die "$outfile_name did not open";

#  Read frequency from file 1 and put them into @frequency (array frequency)
while (<INFILE1>){
        $line = $_;chop $line;@tmp = split (/\t/,$line);$HLA_name = $tmp[0];$HLA_freq = $tmp[1];
        push(@frequency_name,$HLA_name);push(@frequency,$HLA_freq);
}# end of while1
close(INFILE1);

# Read in data and calculates average HLA frequency
while (<INFILE2>){
        $line = $_;chop $line;@seqArray = split(/\t/, $line);$HLA_name = $seqArray[0];
        $seqLength = $seqArray[1];

        for ($i =0; $i <scalar(@frequency_name); ++$i){
                $token_HLA_name = $frequency_name[$i];
                if ($HLA_name eq $token_HLA_name){
                        $HLA_freq = $frequency[$i];
                }# end of if
        }# end of for

        $HLA_freq = $HLA_freq*$seqLength;

        for ($i = 2; $i < scalar(@seqArray); ++$i){
                $token = $seqArray[$i];

                if ($token =~ /-/){
                        @tmp = split (/-/, $token);$token_HLA_name = $tmp[0];
                        $token_HLA_freq_numb = $tmp[1];
                        for ($j =0; $j <scalar(@frequency_name); ++$j){
                                $HLA_name_temp = $frequency_name[$j];
                                if ($token_HLA_name eq $HLA_name_temp){
                                        $HLA_freq_temp = $frequency[$j];
                                        $HLA_freq = $HLA_freq + ($HLA_freq_temp)*($token_HL
                                        A_freq_numb);
                                }# end of if
                        }# end of for
                }# end of if

                elsif ($token ne "x"){
                        for ($k = 0; $k < scalar(@frequency_name); ++$k){
                                if ($token eq $frequency_name[$k]){
                                        $HLA_freq_temp = $frequency[$k];
                                        $HLA_freq = $HLA_freq + $HLA_freq_temp*$seqLength;
                                }# end of if
                        }# end of for
```

```
                }# end of elsif
        }# end of for

        $HLA_freq = $HLA_freq/$seqLength;$output = $HLA_name."\t".$HLA_freq;
        print OUTFILE ("$output\n");$HLA_freq = 0;$output = "";
}# end of while loop
close(INFILE2);close(OUTFILE);

# == end of script3  ================================================================
```

Perl Script 4

```
#=============================================================================
#  This Perl Script is to retrieve unique header of 454 read; developed by Binhua Liang; 2009
#=============================================================================

#  file pathes
$infile_name = "F:/454Reads.hiv_hits.fna";   #  file containing the check list of sequences
$outfile_name="F:/454_clip.txt";             #  file for outputing sequences

#  open files
open INFILE, "<$infile_name" or die "$infile_name did not open";
open OUTFILE,">$outfile_name" or die "$outfile_name did not open";

#  only leave unique header and remove other items
while (<INFILE>){
        $line = $_;$first = substr($line,0,1);
        if ($first eq ">"){
                @array = split (/ /,$line);$firstSentence = $array[0];print OUTFILE ("$firstSentence\n");
        }
        else{
                print OUTFILE ("$line");
        }
}# end of while
close(INFILE);close(OUTFILE);

# == end of script 4 ================================================================
```

Perl Script 5

```
#=============================================================================
#  This script is used to select the best match of each 454 read against multiple clone alignments (from the
#  first run coord file); developed by Binhua Liang; 2009
#=============================================================================

#  file pathes
$infile_name = "G:/patientID.coord";
$outfile_name="G:/patientID_out.coord";
```

```
# declare variables
@array = ();                # working array

$identity = "";             # overlapping of 454 read and reference in coord
$query = "";                # query sequence in coord
$match = 0;                 # the percentage of 454 read matching reference

$seqID = "";                # 454 read ID in coord
$overlapping = "";
$max_match = 0;             # best match of 454 read against multiple refer
                            # seqs from each patient
$token = "";                # to recode the line with the best match in coord

# open files
open (INFILE, $infile_name) || die ("$infile_name did not open");
open OUTFILE,">$outfile_name" or die "$outfile_name did not open";

# Read in each line in coord
while (<INFILE>){
        $line = $_;@array = split (/:/,$line);$identity = $array[0];$query = $array[2];$match =$array[9];

        if ($seqID eq ""){                                  # fill in variables we set at beginning
                $overlapping = $identity;$seqID = $query;$max_match = $match;
        }
        elsif ($seqID eq $query){                           # find best match
                if ($match > $max_match){
                        $max_match = $match;$overlapping = $identity;
                        $token = $line;                     # find the best match line
                }
        }
        else {# before print out best match, make sure the overlapping and percentage of the best match
              # meet our requirements reset max_match, seqID, overlapping, and token

                if (($overlapping >= 75)&& ($max_match >=65)){
                        print OUTFILE ("$token");
                }
                $max_match = $match;$seqID = $query;$overlapping = $identity;$token = $line;
        }
}# end of while
Close (INFILE);Close (OUTFILE);

# == end of script 5 ==========================================================
```

Perl Script 6

```
#==============================================================================
# This Perl Script is used to map 454 reads onto reference; developed by Binhua Liang; 2009
#==============================================================================

# file pathes
$infile_name1  = "H:/patient_out.coord";     # Coord file
$infile_name2 = "H:/patient.fas";            # Aligned clones file
$outfile_name = "H:/patient_aligned.fas";    # Output file
```

```
# declare variables
@array = ();@ref_array = ();@aligned_ref_array = ();@query_array =();$ref_seq = "";$tmp_seq = "";
$aligned_ref_seq = "";$query_seq = "";$start = 0;$token = "";$token2 = "";$end = 0;$count = 0;$seqID =
"";$match =0;$item = "";$ref_count =0;$output="";$begin = 0;$length = 0;$aligned_ref_char="";
$ref_char="";

#  open files
open (INFILE1, $infile_name1) || die ("$infile_name1 did not open");
open OUTFILE,">$outfile_name" or die "$outfile_name did not open";

#  Start working
while (<INFILE1>){
        $count++;$line = $_;@array = split (/:/,$line);$seqID = $array[6];chop $seqID;$start = $array[7];
        $end = $array[8];$ref_seq = $array[12];$query_seq = $array[11];@ref_array = split (//,$ref_seq);
        @query_array = split (//,$query_seq);

        # Find the right seq which matches seqID in coord file
        open (INFILE2, $infile_name2) || die ("$infile_name2 did not open");
        while (<INFILE2>){
                $token = $_;chop $token;$firstletter=substr($token,0,1);
                if (($firstletter eq ">")&&($match == 0)){
                        $item=substr($token,1);
                        if ($item eq $seqID){
                                $match = 1;
                        }
                }
                elsif (($firstletter ne ">")&&($match == 1)){
                        $tmp_seq=$token;$match = 0;last;
                }
        }# end of while loop
        close(INFILE2);

        # Locate the reference and query seqs from coord file; extract them
        $begin = $start-1;$length = $end - $start +1; $aligned_ref_seq = substr($tmp_seq,$begin,$length);
        @aligned_ref_array = split(//,$aligned_ref_seq); print OUTFILE (">$count\n");

        # print "-" before sequence
        for ($i=0; $i<$start-1; $i++){
                print OUTFILE ("-");
        }

        # Label "-" (gaps) in the reference seq from coord file as "X"
        for ($j=0; $j < scalar(@ref_array); $j++){
                $aligned_ref_char = $aligned_ref_array[$ref_count];$ref_char = $ref_array[$j];
                if (($ref_char eq "-")&&($aligned_ref_char eq "-")){
                        $ref_count++;$ref_array[$j]="X";
                }
                elsif (($ref_char ne "-")&&($aligned_ref_char ne "-")){
                        $ref_count++;
                }
                elsif (($ref_char eq "-")&&($aligned_ref_char ne "-")){
                }
        }
        $ref_count=0;
```

```
	#  Match query seq (reads) back to the original aligned reference (gaps expressed as "X")
	#  by removing "-" from coord reference seq (ie: remove inserts in query seq)
	for ($k=0;$k<scalar(@ref_array);$k++){
		$token = $ref_array[$k];$token2 = $query_array[$k];
		if (($token ne "-")&&($token ne "X")&&($token2 eq "-")){
			$output=$output."X";
		}
		elsif ($token ne "-"){
			$token=$query_array[$k];$output=$output.$token;
		}
	}
	print OUTFILE ("$output");$output="";print OUTFILE ("\n");
}# end of while
close(INFILE1);close(OUTFILE);

# == end of script 6 ==========================================================
```

Perl Script 7

```
#==============================================================================
#  This script is used to find mismatch between 454 consensus and reference and locate position and print
#  print them out; developed by Binhua Liang; 2009
#==============================================================================

#  file pathes
$infile_name  = "H:/patient_nt_freq.txt";      #  run-time 454 output
$outfile_name = "H:/patient_mismatch.txt";

# declare variables
@array = ();                                   #  working array
$A = 0;                                        #  hold the number of As at each position of ref
$C = 0;                                        #  ---------- Cs ----------
$G = 0;                                        #  ---------- Gs ----------
$T = 0;                                        # ---------- Ts ----------
$N = 0;
$total = 0;                                    #  hold the number of total 454 nt counts at each position
$ref = "";                                     #  ref nt at each position
$count = 0;                                    #  indicate the position of ref

#  open files
open (INFILE, $infile_name) || die ("$infile_name did not open");
open OUTFILE,">$outfile_name" or die "$outfile_name did not open";

#  Read 454 output and find mismatch between 454 consensus and reference
while (<INFILE>){
	$count++;                              #  indicate the position of ref
	$line = $_;@array = split (/\t/,$line);
	$total = $array[3] + $array[5] + $array[7] + $array[9]+ $array[11];
                                               #  read in the total number of As,Cs,Gs,Ts at that position of
                                               # ref
	$ref = $array[0];                      #  ref nt at that position
	$A = $array[3];                        #  count 454 As at that position of ref
	$C = $array[5];                        #  count 454 Cs at that position of ref
```

```
        $G = $array[7];                 #  count 454 Gs at that position of ref
        $T = $array[9];                 #  count 454 Ts at that position of ref
        $X = $array[11];$N = $array[13];

        if ($total !=0){    #  find the dominant 454_nt at that position of ref and compare 454_nt to ref_nt;
                            #   if they are not same, print out "position,ref_nt,and 454_nt"

                if (($A > 0.5*$total)&($ref ne "A")){
                        print OUTFILE ("$count\t$ref\tA\n");
                }
                elsif (($C > 0.5*$total)&($ref ne "C")){
                        print OUTFILE ("$count\t$ref\tC\n");
                }
                elsif (($G > 0.5*$total)&($ref ne "G")){
                        print OUTFILE ("$count\t$ref\tG\n");
                }
                elsif (($T > 0.5*$total)&($ref ne "T")){
                        print OUTFILE ("$count\t$ref\tT\n");
                }
                elsif ($X > 0.5*$total){
                        print OUTFILE ("$count\t$ref\tX\n");
                }
                elsif ($N > 0.5*$total){
                        print OUTFILE ("$count\t$ref\tN\n");
                }
        }
        $total = 0;
}# end of while
close(INFILE);close(OUTFILE);

# == end of script 7 ==========================================================
```

Perl Script 8

```
#=============================================================================
#  This script is used to find calculate the coverage of 454 reads; developed by Binhua Liang; 2009
#=============================================================================

# file pathes
$infile_name  = "H:/Patient_nt_freq.txt";     #  run-time 454 output
$outfile_name = "H:/Patient_coverage.txt";

# declare variables
@array = ();                                  #  working array
$total = 0;                                   #  indicate the position of ref
$count = 0;

# open files
open (INFILE, $infile_name) || die ("$infile_name did not open");
open OUTFILE,">$outfile_name" or die "$outfile_name did not open";

# do the work
while (<INFILE>){
```

```
        $count++;$line = $_;@array = split (/\t/,$line);
        $total = $array[2] + $array[4] + $array[6] + $array[8]+ $array[12];
                                        # read in the total number of As,Cs,Gs,Ts at that position of
                                        # ref
        print OUTFILE ("$count\t$total\n");

}# end of while
close(INFILE);close(OUTFILE);

# == end of script 8 ==========================================================
```

Perl Script 9

```
#==============================================================================
# This Perl Script is used to determine the clone minor variants (synonimous/or non-synonimous) based on
#  the  nt_freq.txt file and print out variants and their frequencies; developed by Binhua Liang;2009
# ==============================================================================

# Variables
$n_A = ""; $n_B = ""; $n_C = "";        # the nucleotide 1,2,3
$n_numb = 0; $total_numb = 0;           # the number of variant occurrences and total number of
                                        # reads
$seqNum = 1;                            # the nucleotide indicator (1 or 2 or 3)
$prevN = "F";                           # the previous nucleotide (determine if the next nucleotide is
                                        # homopolymer or not)
$consensusAA = ""; $tmpAA = "";$varN = ""; # 454 consensus AA, variant AA, and variant nucleotide
$varFreq = 0;                           # the variant frequency

 $patient_list =
"p04:p06:p08:p09:p11:p13:p14:p19:p20:p21:p22:p23:p24:p25:p26:p27:p28:p29:p30:p31:p32:p33:p34:p35:
p36:p37:p38:p39:p40:p41:p43:p45:p46:p47:p49:p50:p51:p52:p53:p54:p55:p56:p57:p60:p61:p63:p64:p65:p
66:p67:p68:p69:p70:p71:p72:p73:p74:p75:p76:p77:p78:p79:p80:p81:p82:p83:p84:p85:p86:p87:p89:p90:p9
1:p92:p94:p95:p96";

@patientID = (); $patient_id = "";
@array1 = (); @array2 = (); @array3 = ();    # hold each line of nucleotide 1,2,and 3


#  get an amino acid by giving 3 nucleotides and return it
sub getAA {
        $A = $_[0];$B = $_[1];$C = $_[2];

        if ((($A eq "A") && ($B eq "T") && ($C eq "T")) or (($A eq "A") && ($B eq "T") && ($C eq
"C"))or (($A eq "A") && ($B eq "T") && ($C eq "A"))){
                return "I";
        }
        elsif ((($A eq "C") && ($B eq "T") && ($C eq "T")) or (($A eq "C") && ($B eq "T") && ($C eq
"C"))or (($A eq "C") && ($B eq "T") && ($C eq "A")) or (($A eq "C") && ($B eq "T") && ($C eq "G"))
or(($A eq "T") && ($B eq "T") && ($C eq "A")) or (($A eq "T") && ($B eq "T") && ($C eq "G"))){
                return "L";
        }
```

```
		elsif (((($A eq "G") && ($B eq "T") && ($C eq "T")) or (($A eq "G") && ($B eq "T") && ($C eq "C"))or (($A eq "G") && ($B eq "T") && ($C eq "A")) or (($A eq "G") && ($B eq "T") && ($C eq "G"))){
			return "V";
		}
		elsif (((($A eq "T") && ($B eq "T") && ($C eq "T")) or (($A eq "T") && ($B eq "T") && ($C eq "C"))){
			return "F";
		 }
		elsif (($A eq "A") && ($B eq "T") && ($C eq "G")){
			return "M";
		}
		 elsif (((($A eq "T") && ($B eq "G") && ($C eq "T")) or (($A eq "T") && ($B eq "G") && ($C eq "C"))){
			return "C";
		}
		elsif (((($A eq "G") && ($B eq "C") && ($C eq "T")) or (($A eq "G") && ($B eq "C") && ($C eq "C"))or (($A eq "G") && ($B eq "C") && ($C eq "A")) or (($A eq "G") && ($B eq "C") && ($C eq "G"))){
			return "A";
		}
		elsif (((($A eq "G") && ($B eq "G") && ($C eq "T")) or (($A eq "G") && ($B eq "G") && ($C eq "C"))or (($A eq "G") && ($B eq "G") && ($C eq "A")) or (($A eq "G") && ($B eq "G") && ($C eq "G"))){
			return "G";
		}
		elsif (((($A eq "C") && ($B eq "C") && ($C eq "T")) or (($A eq "C") && ($B eq "C") && ($C eq "C"))or (($A eq "C") && ($B eq "C") && ($C eq "A")) or (($A eq "C") && ($B eq "C") && ($C eq "G"))){
			return "P";
		}
		elsif (((($A eq "A") && ($B eq "C") && ($C eq "T")) or (($A eq "A") && ($B eq "C") && ($C eq "C"))or (($A eq "A") && ($B eq "C") && ($C eq "A")) or (($A eq "A") && ($B eq "C") && ($C eq "G"))){
			return "T";
		}
		elsif (((($A eq "T") && ($B eq "C") && ($C eq "T")) or (($A eq "T") && ($B eq "C") && ($C eq "C"))or (($A eq "T") && ($B eq "C") && ($C eq "A")) or (($A eq "T") && ($B eq "C") && ($C eq "G")) or (($A eq "A") && ($B eq "G") && ($C eq "T")) or (($A eq "A") && ($B eq "G") && ($C eq "C"))){
			return "S";
		}
		elsif (((($A eq "T") && ($B eq "A") && ($C eq "T")) or (($A eq "T") && ($B eq "A") && ($C eq "C"))){
			return "Y";
		}
		elsif (($A eq "T") && ($B eq "G") && ($C eq "G")){
			return "W";
		 }
		 elsif (((($A eq "C") && ($B eq "A") && ($C eq "A")) or (($A eq "C") && ($B eq "A") && ($C eq "G"))){
			return "Q";
		}
		elsif (((($A eq "A") && ($B eq "A") && ($C eq "T")) or (($A eq "A") && ($B eq "A") && ($C eq "C"))){
			 return "N";
```

```
		}
		elsif (((($A eq "C") && ($B eq "A") && ($C eq "T")) or ((($A eq "C") && ($B eq "A") && ($C eq "C"))){
			return "H";
		}
		elsif (((($A eq "G") && ($B eq "A") && ($C eq "A")) or ((($A eq "G") && ($B eq "A") && ($C eq "G"))){
			return "E";
		}
		elsif (((($A eq "G") && ($B eq "A") && ($C eq "T")) or ((($A eq "G") && ($B eq "A") && ($C eq "C"))){
			return "D";
		}
		elsif (((($A eq "A") && ($B eq "A") && ($C eq "A")) or ((($A eq "A") && ($B eq "A") && ($C eq "G"))){
			return "K";
		}
		elsif (((($A eq "C") && ($B eq "G") && ($C eq "T")) or ((($A eq "C") && ($B eq "G") && ($C eq "C"))or ((($A eq "C") && ($B eq "G") && ($C eq "A")) or ((($A eq "C") && ($B eq "G") && ($C eq "G")) or((($A eq "A") && ($B eq "G") && ($C eq "A")) or ((($A eq "A") && ($B eq "G") && ($C eq "G"))){
			return "R";
		}
		elsif (((($A eq "T") && ($B eq "A") && ($C eq "A")) or ((($A eq "T") && ($B eq "A") && ($C eq "G")) or ((($A eq "T") && ($B eq "G") && ($C eq "A"))){
			return "stop";
		}
		else {
			return "X";
		}
	}# end of subroutine getAA

# determine minor variants and print them out
@patientID = split(/:/,$patient_list);                    #  put patientID into @patientID array
for ($j = 0;$j<scalar(@patientID);$j++){
		$patient_id = $patientID[$j];                     #  get each patientID

		#Files
		$infile_name = "G:/".$patient_id."_Frequency.txt";  #  input file format: 454_consensus(0);
                                                          #  454_coverage(1); A(2); #A(3); C(4);
                                                          #  C(5); G(6); #G(7); T(8); #T(9)
		$outfile_name="G:/".$patient_id."_MinorVar.txt";

		#Open files
		open (INFILE, $infile_name) || die ("$infile_name did not open");
		open OUTFILE, ">$outfile_name" or die "$outfile_name didn't open";
		while (<INFILE>){
			$line = $_;chop $line;
			#== deal with nucleotide 1 ==============================
			if ($seqNum == 1){
				@array1 = split (/\t/,$line);$n_A = $array1[0];$seqNum++;
			}
			#== deal with nucleotide 2 ==============================
			elsif ($seqNum == 2){
				@array2 = split (/\t/,$line);$n_B = $array2[0];$seqNum++;
			}
```

```
#== deal with nucleotide 1,2,3 together(generate an AA)=========
else {
        @array3 = split (/\t/,$line);$n_C = $array3[0];$seqNum = 1;$prevN = $n_C;
        $consensusAA = &getAA($n_A,$n_B,$n_C);

        #== nucleotide 1 start ============================
        $total_numb = $array1[1];$varN = $array1[2];$n_numb = $array1[3];
         print OUTFILE ("$n_A\t$total_numb\t");
         if (($n_numb > 3)&&($varN ne $n_A)){
                $tmpAA = &getAA($varN,$n_B,$n_C);
                if ($tmpAA eq $consensusAA){
                        $varFreq = $n_numb/$total_numb;
                        if ($varFreq < 0.5){
                                print OUTFILE ("$varN\t$varFreq\t");
                         }
                }
                else {
                        $varFreq = $n_numb/$total_numb;
                        if ($varFreq < 0.5){
                                print OUTFILE ("$varN\t$varFreq\t");
                        }
                }
         }
        $varN = $array1[4];$n_numb = $array1[5];
        if (($n_numb > 3)&&($varN ne $n_A)){
                 $tmpAA = &getAA($varN,$n_B,$n_C);
                if ($tmpAA eq $consensusAA){
                        $varFreq = $n_numb/$total_numb;
                         if ($varFreq < 0.5){
                                print OUTFILE ("$varN\t$varFreq\t");
                         }
                 }
                else {
                        $varFreq = $n_numb/$total_numb;
                        if ($varFreq < 0.5){
                                print OUTFILE ("$varN\t$varFreq\t");
                        }
                }
         }
        $varN = $array1[6];$n_numb = $array1[7];
         if (($n_numb > 3)&&($varN ne $n_A)){
                $tmpAA = &getAA($varN,$n_B,$n_C);
                if ($tmpAA eq $consensusAA){
                        $varFreq = $n_numb/$total_numb;
                         if ($varFreq < 0.5){
                                print OUTFILE ("$varN\t$varFreq\t");
                        }
                 }
                else {
                        $varFreq = $n_numb/$total_numb;
                        if ($varFreq < 0.5){
                                print OUTFILE ("$varN\t$varFreq\t");
                        }
                }
        }
```

```
$varN = $array1[8];$n_numb = $array1[9];
        if (($n_numb > 3)&&($varN ne $n_A)){
                $tmpAA = &getAA($varN,$n_B,$n_C);
                 if ($tmpAA eq $consensusAA){
                        $varFreq = $n_numb/$total_numb;
                         if ($varFreq < 0.5){
                                print OUTFILE ("$varN\t$varFreq\t");
                        }
                }
        else {
                $varFreq = $n_numb/$total_numb;
                 if ($varFreq < 0.5){
                        print OUTFILE ("$varN\t$varFreq\t");
                }
        }
}
print OUTFILE ("\n");@array1 = ();

 #== nucleotide 2 start. ==================================
$total_numb = $array2[1];$varN = $array2[2];$n_numb = $array2[3];
print OUTFILE ("$n_B\t$total_numb\t");

if (($n_numb > 3)&&($varN ne $n_B)){
        $tmpAA = &getAA($n_A,$varN,$n_C);
         if ($tmpAA eq $consensusAA){
                $varFreq = $n_numb/$total_numb;
                if ($varFreq < 0.5){
                        print OUTFILE ("$varN\t$varFreq\t");
                }
        }
        else {
                $varFreq = $n_numb/$total_numb;
                if ($varFreq < 0.5){
                        print OUTFILE ("$varN\t$varFreq\t");
                }
        }
}
$varN = $array2[4];$n_numb = $array2[5];
if (($n_numb > 3)&&($varN ne $n_B)){
        $tmpAA = &getAA($n_A,$varN,$n_C);
         if ($tmpAA eq $consensusAA){
                $varFreq = $n_numb/$total_numb;
                 if ($varFreq < 0.5){
                        print OUTFILE ("$varN\t$varFreq\t");
                }
        }
        else {
                $varFreq = $n_numb/$total_numb;
                if ($varFreq < 0.5){
                        print OUTFILE ("$varN\t$varFreq\t");
                }
        }
 }
$varN = $array2[6];$n_numb = $array2[7];
if (($n_numb > 3)&&($varN ne $n_B)){
```

```
		$tmpAA = &getAA($n_A,$varN,$n_C);
		if ($tmpAA eq $consensusAA){
			$varFreq = $n_numb/$total_numb;
			if ($varFreq < 0.5){
				print OUTFILE ("$varN\t$varFreq\t");
			}
		}
		else {
			$varFreq = $n_numb/$total_numb;
			 if ($varFreq < 0.5){
				print OUTFILE ("$varN\t$varFreq\t");
			}
		}
}
 $varN = $array2[8];$n_numb = $array2[9];
if (($n_numb > 3)&&($varN ne $n_B)){
		$tmpAA = &getAA($n_A,$varN,$n_C);
		if ($tmpAA eq $consensusAA){
			$varFreq = $n_numb/$total_numb;
			if ($varFreq < 0.5){
				print OUTFILE ("$varN\t$varFreq\t");
			}
		 }
		else {
			$varFreq = $n_numb/$total_numb;
			if ($varFreq < 0.5){
				print OUTFILE ("$varN\t$varFreq\t");
			}
		}
}
print OUTFILE ("\n");@array2 = ();

#== nucleotide 3 start ================================
$total_numb = $array3[1];$varN = $array3[2];$n_numb = $array3[3];
print OUTFILE ("$n_C\t$total_numb\t");

 if (($n_numb > 3)&&($varN ne $n_C)){
		 $tmpAA = &getAA($n_A,$n_B,$varN);
		if ($tmpAA eq $consensusAA){
			$varFreq = $n_numb/$total_numb;
			if ($varFreq < 0.5){
				print OUTFILE ("$varN\t$varFreq\t");
			}
		}
		else {
			$varFreq = $n_numb/$total_numb;
			if ($varFreq < 0.5){
				print OUTFILE ("$varN\t$varFreq\t");
			}
		}
}
$varN = $array3[4];$n_numb = $array3[5];
if (($n_numb > 3)&&($varN ne $n_C)){
		$tmpAA = &getAA($n_A,$n_B,$varN);
		if ($tmpAA eq $consensusAA){
```

```
                                        $varFreq = $n_numb/$total_numb;
                                        if ($varFreq < 0.5){
                                                print OUTFILE ("$varN\t$varFreq\t");
                                        }
                                }
                                else {
                                        $varFreq = $n_numb/$total_numb;
                                        if ($varFreq < 0.5){
                                                print OUTFILE ("$varN\t$varFreq\t");
                                        }
                                }
                        }
                        $varN = $array3[6];$n_numb = $array3[7];
                        if (($n_numb > 3)&&($varN ne $n_C)){
                                $tmpAA = &getAA($n_A,$n_B,$varN);
                                if ($tmpAA eq $consensusAA){
                                        $varFreq = $n_numb/$total_numb;
                                        if ($varFreq < 0.5){
                                                print OUTFILE ("$varN\t$varFreq\t");
                                        }
                                }
                                else {
                                        $varFreq = $n_numb/$total_numb;
                                        if ($varFreq < 0.5){
                                                print OUTFILE ("$varN\t$varFreq\t");
                                        }
                                }
                        }
                        $varN = $array3[8];$n_numb = $array3[9];
                        if (($n_numb > 3)&&($varN ne $n_C)){
                                $tmpAA = &getAA($n_A,$n_B,$varN);
                                if ($tmpAA eq $consensusAA){
                                        $varFreq = $n_numb/$total_numb;
                                        if ($varFreq < 0.5){
                                                print OUTFILE ("$varN\t$varFreq\t");
                                        }
                                }
                                else {
                                        $varFreq = $n_numb/$total_numb;
                                        if ($varFreq < 0.5){
                                                print OUTFILE ("$varN\t$varFreq\t");
                                        }
                                }
                        }
                        print OUTFILE ("\n");
                        @array3 = ();
                }#  end of else (real work)
        }#  end of while loop
        close(INFILE);close(OUTFILE);
}#  end of for loop

# == end of script 9 ==========================================================
```

Perl Script 10

```
#================================================================================
#  This program is used to determine the 454 minor variants (synonimous/or non-synonimous) based on the
#  nt_freq.txt file and print out variants and their frequencies; developed by Binhua Liang; 2009
# ================================================================================

# Variables
$n_A = ""; $n_B = ""; $n_C = "";                    #  the nucleotide 1,2,3
$n_numb = 0; $total_numb = 0;                       #  the number of variant occurrences and
                                                    #  total number of reads
$error_rate1 = 0;$error_rate2 = 0;$error_rate3 = 0; $p = 0;  #  the error rate and p values
$lamda = 0;$sum = 0;$tmp = 0;$token = 1;            #  used for calculating p values and fractional

$seqNum = 1;                                        #  the nucleotide indicator (1 or 2 or 3)
$prevN = "F";                                       #  the previous nucleotide (determine if the next
                                                    #  nucleotide is homopolymer or not)
$consensusAA = ""; $tmpAA = "";$varN = "";          # 454 consensus AA, variant AA, and variant
                                                    #  nucleotide
$true = 0;                                          #  the true indicator (if it is true variant)
$varFreq = 0;                                       #  the variant frequency

$patient_list =
"p04:p06:p08:p09:p11:p13:p14:p19:p20:p21:p22:p23:p24:p25:p26:p27:p28:p29:p30:p31:p32:p33:p34:p35:
p36:p37:p38:p39:p40:p41:p43:p45:p46:p47:p49:p50:p51:p52:p53:p54:p55:p56:p57:p60:p61:p63:p64:p65:p
66:p67:p68:p69:p70:p71:p72:p73:p74:p75:p76:p77:p78:p79:p80:p81:p82:p83:p84:p85:p86:p87:p89:p90:p9
1:p92:p94:p95:p96";

@patientID = (); $patient_id = "";
@array1 = ();@array2 = ();@array3 = ();              #  hold each line of nucleotide 1,2,and 3

# calculate its fractional by giving a number and return the result
sub getFractional($){
        $tmp = $_[0];$token = 1;
        for ($k=1;$k<$tmp+1;++$k){
                $token = $token*$k;
        }
        return $token;
 }# end of subroutine getFractional

# get an amino acid by giving 3 nucleotides and return it
sub getAA {
        $A = $_[0];$B = $_[1];$C = $_[2];

        if ((($A eq "A") && ($B eq "T") && ($C eq "T")) or (($A eq "A") && ($B eq "T") && ($C eq
"C"))or (($A eq "A") && ($B eq "T") && ($C eq "A"))){
                return "I";
        }
        elsif ((($A eq "C") && ($B eq "T") && ($C eq "T")) or (($A eq "C") && ($B eq "T") && ($C eq
"C"))or (($A eq "C") && ($B eq "T") && ($C eq "A")) or (($A eq "C") && ($B eq "T") && ($C eq "G"))
or(($A eq "T") && ($B eq "T") && ($C eq "A")) or (($A eq "T") && ($B eq "T") && ($C eq "G"))){
                return "L";
        }
```

```
		elsif (((($A eq "G") && ($B eq "T") && ($C eq "T")) or (($A eq "G") && ($B eq "T") && ($C eq "C"))or (($A eq "G") && ($B eq "T") && ($C eq "A")) or (($A eq "G") && ($B eq "T") && ($C eq "G"))){
			return "V";
		 }
		elsif (((($A eq "T") && ($B eq "T") && ($C eq "T")) or (($A eq "T") && ($B eq "T") && ($C eq "C"))){
			return "F";
		}
		 elsif (($A eq "A") && ($B eq "T") && ($C eq "G")){
			return "M";
		}
		elsif (((($A eq "T") && ($B eq "G") && ($C eq "T")) or (($A eq "T") && ($B eq "G") && ($C eq "C"))){
			return "C";
		}
		elsif (((($A eq "G") && ($B eq "C") && ($C eq "T")) or (($A eq "G") && ($B eq "C") && ($C eq "C"))or (($A eq "G") && ($B eq "C") && ($C eq "A")) or (($A eq "G") && ($B eq "C") && ($C eq "G"))){
			return "A";
		}
		elsif (((($A eq "G") && ($B eq "G") && ($C eq "T")) or (($A eq "G") && ($B eq "G") && ($C eq "C"))or (($A eq "G") && ($B eq "G") && ($C eq "A")) or (($A eq "G") && ($B eq "G") && ($C eq "G"))){
			return "G";
		}
		elsif (((($A eq "C") && ($B eq "C") && ($C eq "T")) or (($A eq "C") && ($B eq "C") && ($C eq "C"))or (($A eq "C") && ($B eq "C") && ($C eq "A")) or (($A eq "C") && ($B eq "C") && ($C eq "G"))){
			return "P";
		 }
		elsif (((($A eq "A") && ($B eq "C") && ($C eq "T")) or (($A eq "A") && ($B eq "C") && ($C eq "C"))or (($A eq "A") && ($B eq "C") && ($C eq "A")) or (($A eq "A") && ($B eq "C") && ($C eq "G"))){
			return "T";
		}
		elsif (((($A eq "T") && ($B eq "C") && ($C eq "T")) or (($A eq "T") && ($B eq "C") && ($C eq "C"))or (($A eq "T") && ($B eq "C") && ($C eq "A")) or (($A eq "T") && ($B eq "C") && ($C eq "G")) or (($A eq "A") && ($B eq "G") && ($C eq "T")) or (($A eq "A") && ($B eq "G") && ($C eq "C"))){
			return "S";
		}
		elsif (((($A eq "T") && ($B eq "A") && ($C eq "T")) or (($A eq "T") && ($B eq "A") && ($C eq "C"))){
			return "Y";
		}
		elsif (($A eq "T") && ($B eq "G") && ($C eq "G")){
			return "W";
		}
		elsif (((($A eq "C") && ($B eq "A") && ($C eq "A")) or (($A eq "C") && ($B eq "A") && ($C eq "G"))){
			return "Q";
		}
		elsif (((($A eq "A") && ($B eq "A") && ($C eq "T")) or (($A eq "A") && ($B eq "A") && ($C eq "C"))){
			return "N";
```

```
		}
		lsif (((($A eq "C") && ($B eq "A") && ($C eq "T")) or ((($A eq "C") && ($B eq "A") && ($C eq
"C"))){
			return "H";
		 }
		elsif (((($A eq "G") && ($B eq "A") && ($C eq "A")) or ((($A eq "G") && ($B eq "A") && ($C
eq "G"))){
			return "E";
		}
		elsif (((($A eq "G") && ($B eq "A") && ($C eq "T")) or ((($A eq "G") && ($B eq "A") && ($C
eq "C"))){
			 return "D";
		}
		elsif (((($A eq "A") && ($B eq "A") && ($C eq "A")) or ((($A eq "A") && ($B eq "A") && ($C
eq "G"))){
			return "K";
		}
		elsif (((($A eq "C") && ($B eq "G") && ($C eq "T")) or ((($A eq "C") && ($B eq "G") && ($C eq
"C"))or ((($A eq "C") && ($B eq "G") && ($C eq "A")) or ((($A eq "C") && ($B eq "G") && ($C eq "G"))
or((($A eq "A") && ($B eq "G") && ($C eq "A")) or ((($A eq "A") && ($B eq "G") && ($C eq "G"))){
			return "R";
		}
		elsif (((($A eq "T") && ($B eq "A") && ($C eq "A")) or ((($A eq "T") && ($B eq "A") && ($C eq
"G")) or ((($A eq "T") && ($B eq "G") && ($C eq "A"))){
			return "stop";
		}
		else {
			return "X";
		}
 }# end of subroutine getAA

# test if it is true variant by giving the number of variant occurrences, total number of reads, and error rate.
# by return 1 (true) or 0 (not true)
sub iftrue {
		$n_numb = $_[0];$total_numb = $_[1];$error_rate = $_[2];$lamda = $total_numb*$error_rate;
		$sum = (2.71828)**(-$lamda);

		# calculate p values
		for ($i=1;$i<$n_numb;$i++){
			$tmp = &getFractional($i);$sum = (((2.718)**(-$lamda))*($lamda**$i))/$tmp + $sum;
		}#end of for loop
		 $p = 1-$sum;

		# determine if true or not
		if ($p < 0.001){
			return 1;
		}
		else {
			return 0;
		}
 }# end of subroutine "iftrue"

# determine minor variants and print them out
@patientID = split(/:/,$patient_list);                    #  put patientID into @patientID array
for ($j = 0;$j<scalar(@patientID);$j++){
```

```
$patient_id = $patientID[$j];                    #  get each patientID

# Files
$infile_name = "G:/".$patient_id."_Frequency.txt";    # input file format: 454_consensus(0);
                                                      # 454_coverage(1); A(2); #A(3); C(4);
                                                      # C(5); G(6); #G(7); T(8); #T(9)
$outfile_name="G:/".$patient_id."_MinorVar.txt";

# Open files
open (INFILE, $infile_name) || die ("$infile_name did not open");
open OUTFILE, ">$outfile_name" or die "$outfile_name didn't open";
while (<INFILE>){
        $line = $_;chop $line;
        #== deal with nucleotide 1 ==============================
        if ($seqNum == 1){
                @array1 = split (/\t/,$line);$n_A = $array1[0];

                if ($prevN eq $n_A){
                        $error_rate1 = 0.0044;
                }
                else {
                        $error_rate1 = 0.0007;
                }
                $seqNum++;
        }
        #== deal with nucleotide 2 ==============================
        elsif ($seqNum == 2){
                @array2 = split (/\t/,$line);$n_B = $array2[0];

                if ($n_A eq $n_B){
                        $error_rate2 = 0.0044;
                }
                else {
                        $error_rate2 = 0.0007;
                }
                $seqNum++;
        }
        #== deal with nucleotide 1,2,3 together(generate an AA) ==========
        else {
                @array3 = split (/\t/,$line);$n_C = $array3[0];
                if ($n_C eq $n_B){
                        $error_rate3 = 0.0044;
                }
                else {
                        $error_rate3 = 0.0007;
                }
                $seqNum = 1;$prevN = $n_C;$consensusAA = &getAA($n_A,$n_B,$n_C);

                #== nucleotide 1 start  ============================
                $total_numb = $array1[1];$varN = $array1[2];$n_numb = $array1[3];
                print OUTFILE ("$n_A\t$total_numb\t");
                if (($n_numb != 0)&&($varN ne $n_A)){
                        $true = &iftrue($n_numb,$total_numb,$error_rate1);
                        if ($true == 1){
                                $tmpAA = &getAA($varN,$n_B,$n_C);
```

```
                        if ($tmpAA eq $consensusAA){
                                $varFreq = $n_numb/$total_numb;
                                if ($varFreq < 0.5){
                                        print OUTFILE ("$varN\t$varFreq\t");
                                }
                        }
                        else {
                                $varFreq = $n_numb/$total_numb;
                                if ($varFreq < 0.5){
                                        print OUTFILE ("$varN\t$varFreq\t");
                                }
                        }
                }
                $true = 0;
        }
        $varN = $array1[4];$n_numb = $array1[5];
        if (($n_numb != 0)&&($varN ne $n_A)){
                $true = &iftrue($n_numb,$total_numb,$error_rate1);
                if ($true == 1){
                        $tmpAA = &getAA($varN,$n_B,$n_C);
                        if ($tmpAA eq $consensusAA){
                                $varFreq = $n_numb/$total_numb;
                                if ($varFreq < 0.5){
                                        print OUTFILE ("$varN\t$varFreq\t");
                                }
                        }
                        else {
                                $varFreq = $n_numb/$total_numb;
                                if ($varFreq < 0.5){
                                        print OUTFILE ("$varN\t$varFreq\t");
                                }
                        }
                }
                $true = 0;
        }
        $varN = $array1[6];$n_numb = $array1[7];
        if (($n_numb != 0)&&($varN ne $n_A)){
                $true = &iftrue($n_numb,$total_numb,$error_rate1);
                if ($true == 1){
                        $tmpAA = &getAA($varN,$n_B,$n_C);
                        if ($tmpAA eq $consensusAA){
                                $varFreq = $n_numb/$total_numb;
                                if ($varFreq < 0.5){
                                        print OUTFILE ("$varN\t$varFreq\t");
                                }
                        }
                        else {
                                $varFreq = $n_numb/$total_numb;
                                if ($varFreq < 0.5){
                                        print OUTFILE ("$varN\t$varFreq\t");
                                }
                        }
                }
                $true = 0;
        }
```

```
$varN = $array1[8];$n_numb = $array1[9];
if (($n_numb != 0)&&($varN ne $n_A)){
        $true = &iftrue($n_numb,$total_numb,$error_rate1);
        if ($true == 1){
                $tmpAA = &getAA($varN,$n_B,$n_C);
                if ($tmpAA eq $consensusAA){
                        $varFreq = $n_numb/$total_numb;
                        if ($varFreq < 0.5){
                                print OUTFILE ("$varN\t$varFreq\t");
                        }
                }
                else {
                        $varFreq = $n_numb/$total_numb;
                        if ($varFreq < 0.5){
                                print OUTFILE ("$varN\t$varFreq\t");
                        }
                }
        }
        $true = 0;
}
print OUTFILE ("\n");@array1 = ();

#== nucleotide 2 start ==========================================
$total_numb = $array2[1];$varN = $array2[2];$n_numb = $array2[3];
print OUTFILE ("$n_B\t$total_numb\t");

if (($n_numb != 0)&&($varN ne $n_B)){
        $true = &iftrue($n_numb,$total_numb,$error_rate2);
        if ($true == 1){
                $tmpAA = &getAA($n_A,$varN,$n_C);
                if ($tmpAA eq $consensusAA){
                        $varFreq = $n_numb/$total_numb;
                        if ($varFreq < 0.5){
                                print OUTFILE ("$varN\t$varFreq\t");
                        }
                }
                else {
                        $varFreq = $n_numb/$total_numb;
                        if ($varFreq < 0.5){
                                print OUTFILE ("$varN\t$varFreq\t");
                        }
                }
        }
        $true = 0;
}
$varN = $array2[4];$n_numb = $array2[5];
if (($n_numb != 0)&&($varN ne $n_B)){
        $true = &iftrue($n_numb,$total_numb,$error_rate2);
        if ($true == 1){
                $tmpAA = &getAA($n_A,$varN,$n_C);
                if ($tmpAA eq $consensusAA){
                        $varFreq = $n_numb/$total_numb;
                        if ($varFreq < 0.5){
                                print OUTFILE ("$varN\t$varFreq\t");
                        }
```

```
                    }
                    else {
                        $varFreq = $n_numb/$total_numb;
                        if ($varFreq < 0.5){
                            print OUTFILE ("$varN\t$varFreq\t");
                        }
                    }
                }
                $true = 0;
        }
        $varN = $array2[6];$n_numb = $array2[7];
        if (($n_numb != 0)&&($varN ne $n_B)){
                $true = &iftrue($n_numb,$total_numb,$error_rate2);
                if ($true == 1){
                    $tmpAA = &getAA($n_A,$varN,$n_C);
                    if ($tmpAA eq $consensusAA){
                        $varFreq = $n_numb/$total_numb;
                        if ($varFreq < 0.5){
                            print OUTFILE ("$varN\t$varFreq\t");
                        }
                    }
                    else {
                        $varFreq = $n_numb/$total_numb;
                        if ($varFreq < 0.5){
                            print OUTFILE ("$varN\t$varFreq\t");
                        }
                    }
                }
                $true = 0;
        }
        $varN = $array2[8];$n_numb = $array2[9];
        if (($n_numb != 0)&&($varN ne $n_B)){
                $true = &iftrue($n_numb,$total_numb,$error_rate2);
                if ($true == 1){
                    $tmpAA = &getAA($n_A,$varN,$n_C);
                    if ($tmpAA eq $consensusAA){
                        $varFreq = $n_numb/$total_numb;
                        if ($varFreq < 0.5){
                            print OUTFILE ("$varN\t$varFreq\t");
                        }
                    }
                    else {
                        $varFreq = $n_numb/$total_numb;
                        if ($varFreq < 0.5){
                            print OUTFILE ("$varN\t$varFreq\t");
                        }
                    }
                }
                $true = 0;
        }
        print OUTFILE ("\n");@array2 = ();

        #== nucleotide 3 start ==========================================
        $total_numb = $array3[1];$varN = $array3[2];$n_numb = $array3[3];
        print OUTFILE ("$n_C\t$total_numb\t");
```

```
if (($n_numb != 0)&&($varN ne $n_C)){
        $true = &iftrue($n_numb,$total_numb,$error_rate3);
        if ($true == 1){
                $tmpAA = &getAA($n_A,$n_B,$varN);
                if ($tmpAA eq $consensusAA){
                        $varFreq = $n_numb/$total_numb;
                        if ($varFreq < 0.5){
                                print OUTFILE ("$varN\t$varFreq\t");
                        }
                }
                else {
                        $varFreq = $n_numb/$total_numb;
                        if ($varFreq < 0.5){
                                print OUTFILE ("$varN\t$varFreq\t");
                        }
                }
        }
        $true = 0;
 }
$varN = $array3[4];$n_numb = $array3[5];
if (($n_numb != 0)&&($varN ne $n_C)){
        $true = &iftrue($n_numb,$total_numb,$error_rate3);
         if ($true == 1){
                $tmpAA = &getAA($n_A,$n_B,$varN);
                if ($tmpAA eq $consensusAA){
                        $varFreq = $n_numb/$total_numb;
                        if ($varFreq < 0.5){
                                print OUTFILE ("$varN\t$varFreq\t");
                         }
                }
                else {
                        $varFreq = $n_numb/$total_numb;
                        if ($varFreq < 0.5){
                                print OUTFILE ("$varN\t$varFreq\t");
                        }
                }
        }
        $true = 0;
}
$varN = $array3[6];$n_numb = $array3[7];
 if (($n_numb != 0)&&($varN ne $n_C)){
        $true = &iftrue($n_numb,$total_numb,$error_rate3);
         if ($true == 1){
                $tmpAA = &getAA($n_A,$n_B,$varN);
                if ($tmpAA eq $consensusAA){
                        $varFreq = $n_numb/$total_numb;
                         if ($varFreq < 0.5){
                                print OUTFILE ("$varN\t$varFreq\t");
                        }
                }
                else {
                        $varFreq = $n_numb/$total_numb;
                        if ($varFreq < 0.5){
                                print OUTFILE ("$varN\t$varFreq\t");
                        }
```

```
                                        }
                                }
                                $true = 0;
                        }
                        $varN = $array3[8];$n_numb = $array3[9];
                        if (($n_numb != 0)&&($varN ne $n_C)){
                                $true = &iftrue($n_numb,$total_numb,$error_rate3);
                                if ($true == 1){
                                        $tmpAA = &getAA($n_A,$n_B,$varN);
                                        if ($tmpAA eq $consensusAA){
                                                $varFreq = $n_numb/$total_numb;
                                                if ($varFreq < 0.5){
                                                        print OUTFILE ("$varN\t$varFreq\t");
                                                 }
                                        }
                                        else {
                                                $varFreq = $n_numb/$total_numb;
                                                if ($varFreq < 0.5){
                                                        print OUTFILE ("$varN\t$varFreq\t");
                                                }
                                        }
                                }
                                $true = 0;
                        }
                        print OUTFILE ("\n");
                        @array3 = ();
                }# else (real work)
        }# end of while loop
        Close(INFILE);close(OUTFILE);
}#  end of for loop

#== end of script 10 ============================================================
```

Perl Script 11

```
#================================================================================
#   This Perl Script is used to determine the 454 minor variants (synonimous/or non-synonimous) based on
#   nt_freq.txt file and print out variants and their frequencies; developed by Binhua Liang; 2009
#================================================================================

# Variables
$n_A = ""; $n_B = ""; $n_C = "";                # the nucleotide 1,2,3
$n_numb = 0; $total_numb = 0;                   # the number of variant occurrences and total
                                                # number of reads
$seqNum = 1;                                    # the nucleotide indicator (1 or 2 or 3)
$prevN = "F";                                   # the previous nucleotide (determine if the next
                                                # nucleotide is
                                                # homopolymer or not)
$consensusAA = ""; $tmpAA = "";$varN = "";      # 454 consensus AA, variant AA, and variant
                                                # nucleotide
$varFreq = 0;                                   # the variant frequency
```

```
 $patient_list =
"p04:p06:p08:p09:p11:p13:p14:p19:p20:p21:p22:p23:p24:p25:p26:p27:p28:p29:p30:p31:p32:p33:p34:p35:
p36:p37:p38:p39:p40:p41:p43:p45:p46:p47:p49:p50:p51:p52:p53:p54:p55:p56:p57:p60:p61:p63:p64:p65:p
66:p67:p68:p69:p70:p71:p72:p73:p74:p75:p76:p77:p78:p79:p80:p81:p82:p83:p84:p85:p86:p87:p89:p90:p9
1:p92:p94:p95:p96";

@patientID = (); $patient_id = "";
@array1 = (); @array2 = (); @array3 = ();              #  hold each line of nucleotide 1,2,and 3


# get an amino acid by giving 3 nucleotides and return it
sub getAA {
        $A = $_[0];$B = $_[1];$C = $_[2];

        if (((($A eq "A") && ($B eq "T") && ($C eq "T")) or (($A eq "A") && ($B eq "T") && ($C eq
       "C"))or   (($A eq "A") && ($B eq "T") && ($C eq "A"))){
                return "I";
       }
       elsif ((($A eq "C") && ($B eq "T") && ($C eq "T")) or (($A eq "C") && ($B eq "T") && ($C eq
       "C"))or (($A eq "C") && ($B eq "T") && ($C eq "A")) or (($A eq "C") && ($B eq "T") && ($C
       eq "G")) or((($A eq "T") && ($B eq "T") && ($C eq "A")) or (($A eq "T") && ($B eq "T") &&
       ($C eq "G"))){
                return "L";
        }
        elsif ((($A eq "G") && ($B eq "T") && ($C eq "T")) or (($A eq "G") && ($B eq "T") && ($C
        eq  "C"))or (($A eq "G") && ($B eq "T") && ($C eq "A")) or (($A eq "G") && ($B eq "T") &&
        ($C eq "G"))){
                return "V";
       }
       elsif ((($A eq "T") && ($B eq "T") && ($C eq "T")) or (($A eq "T") && ($B eq "T") && ($C eq
       "C"))){
                return "F";
       }
        elsif (($A eq "A") && ($B eq "T") && ($C eq "G")){
                 return "M";
       }
        elsif ((($A eq "T") && ($B eq "G") && ($C eq "T")) or (($A eq "T") && ($B eq "G") && ($C
        eq "C"))){
                return "C";
        }
       elsif ((($A eq "G") && ($B eq "C") && ($C eq "T")) or (($A eq "G") && ($B eq "C") && ($C eq
       "C"))or (($A eq "G") && ($B eq "C") && ($C eq "A")) or (($A eq "G") && ($B eq "C") && ($C
       eq "G"))){
                 return "A";
        }
       elsif ((($A eq "G") && ($B eq "G") && ($C eq "T")) or (($A eq "G") && ($B eq "G") && ($C
        eq  "C"))or (($A eq "G") && ($B eq "G") && ($C eq "A")) or (($A eq "G") && ($B eq "G") &&
        ($C eq "G"))){
                return "G";
       }
       elsif ((($A eq "C") && ($B eq "C") && ($C eq "T")) or (($A eq "C") && ($B eq "C") && ($C eq
       "C"))or (($A eq "C") && ($B eq "C") && ($C eq "A")) or (($A eq "C") && ($B eq "C") && ($C
        eq "G"))){
                return "P";
       }
```

```
elsif (((($A eq "A") && ($B eq "C") && ($C eq "T")) or (($A eq "A") && ($B eq "C") && ($C eq
"C"))or (($A eq "A") && ($B eq "C") && ($C eq "A")) or (($A eq "A") && ($B eq "C") && ($C
eq "G"))){
        return "T";
}
elsif (((($A eq "T") && ($B eq "C") && ($C eq "T")) or (($A eq "T") && ($B eq "C") && ($C eq
"C"))or (($A eq "T") && ($B eq "C") && ($C eq "A")) or (($A eq "T") && ($B eq "C") && ($C
eq "G")) or (($A eq "A") && ($B eq "G") && ($C eq "T")) or (($A eq "A") && ($B eq "G") &&
($C eq "C"))){
        return "S";
}
elsif (((($A eq "T") && ($B eq "A") && ($C eq "T")) or (($A eq "T") && ($B eq "A") && ($C
eq "C"))){
        return "Y";
}
elsif (($A eq "T") && ($B eq "G") && ($C eq "G")){
        return "W";
}
elsif (((($A eq "C") && ($B eq "A") && ($C eq "A")) or (($A eq "C") && ($B eq "A") && ($C
eq"G"))){
        return "Q";
}
elsif (((($A eq "A") && ($B eq "A") && ($C eq "T")) or (($A eq "A") && ($B eq "A") && ($C
eq "C"))){
        return "N";
}
elsif (((($A eq "C") && ($B eq "A") && ($C eq "T")) or (($A eq "C") && ($B eq "A") && ($C eq
"C"))){
        return "H";
}
elsif (((($A eq "G") && ($B eq "A") && ($C eq "A")) or (($A eq "G") && ($B eq "A") && ($C
eq "G"))){
        return "E";
}
elsif (((($A eq "G") && ($B eq "A") && ($C eq "T")) or (($A eq "G") && ($B eq "A") && ($C
eq  "C"))){
        return "D";
}
elsif (((($A eq "A") && ($B eq "A") && ($C eq "A")) or (($A eq "A") && ($B eq "A") && ($C
eq  "G"))){
        return "K";
}
elsif (((($A eq "C") && ($B eq "G") && ($C eq "T")) or (($A eq "C") && ($B eq "G") && ($C eq
"C"))or (($A eq "C") && ($B eq "G") && ($C eq "A")) or (($A eq "C") && ($B eq "G") && ($C
eq "G")) or(($A eq "A") && ($B eq "G") && ($C eq "A")) or (($A eq "A") && ($B eq "G") &&
($C eq "G"))){
        return "R";
}
elsif (((($A eq "T") && ($B eq "A") && ($C eq "A")) or (($A eq "T") && ($B eq "A") && ($C eq
"G"))or (($A eq "T") && ($B eq "G") && ($C eq "A"))){
        return "stop";
}
else {
        return "X";
}
```

```
}#  end of subroutine getAA

#  determine minor variants and print them out
@patientID = split(/:/,$patient_list);                    #  put patientID into @patientID array
for ($j = 0;$j<scalar(@patientID);$j++){

        $patient_id = $patientID[$j];                     #  get each patientID

        # Files
        $infile_name = "G:/".$patient_id."_Frequency.txt";  # input file format:  454_consensus (0); 454_
                                                            # _coverage(1); A(2); #A(3); C(4); #C(5);G
                                                            #  (6); #G(7); T(8); T(9)
        $outfile_name="G:/".$patient_id."_MinorVar.txt";

        # Open files
        open (INFILE, $infile_name) || die ("$infile_name did not open");
        open OUTFILE, ">$outfile_name" or die "$outfile_name didn't open";

        while (<INFILE>){
                $line = $_;chop $line;

                #== deal with nucleotide 1 ================================
                if ($seqNum == 1){
                        @array1 = split (/\t/,$line);$n_A = $array1[0];$seqNum++;
                }
                #== deal with nucleotide 2 ================================
                elsif ($seqNum == 2){
                        @array2 = split (/\t/,$line);$n_B = $array2[0];$seqNum++;
                }
                #== deal with nucleotide 1,2,3 together(generate an AA)==========
                else {
                        @array3 = split (/\t/,$line);$n_C = $array3[0];$seqNum = 1;$prevN = $n_C;
                        $consensusAA = &getAA($n_A,$n_B,$n_C);

                         #== nucleotide 1 start ============================
                        $total_numb = $array1[1];$varN = $array1[2];$n_numb = $array1[3];
                        print OUTFILE ("$n_A\t$total_numb\t");
                        if (($n_numb > 3)&&($varN ne $n_A)){
                                $tmpAA = &getAA($varN,$n_B,$n_C);
                                if ($tmpAA eq $consensusAA){
                                        $varFreq = $n_numb/$total_numb;
                                        if ($varFreq < 0.5){
                                                print OUTFILE ("$varN\t$varFreq\t");
                                        }
                                }
                                else {
                                        $varFreq = $n_numb/$total_numb;
                                        if ($varFreq < 0.5){
                                                print OUTFILE ("$varN\t$varFreq\t");
                                        }
                                }
                         }
                        $varN = $array1[4];$n_numb = $array1[5];
                        if (($n_numb > 3)&&($varN ne $n_A)){
                                $tmpAA = &getAA($varN,$n_B,$n_C);
```

```
		if ($tmpAA eq $consensusAA){
			$varFreq = $n_numb/$total_numb;
			if ($varFreq < 0.5){
				print OUTFILE ("$varN\t$varFreq\t");
			 }
		 }
		else {
			$varFreq = $n_numb/$total_numb;
			if ($varFreq < 0.5){
				print OUTFILE ("$varN\t$varFreq\t");
			}
		 }
}
 $varN = $array1[6];$n_numb = $array1[7];
if (($n_numb > 3)&&($varN ne $n_A)){
		$tmpAA = &getAA($varN,$n_B,$n_C);
		if ($tmpAA eq $consensusAA){
			$varFreq = $n_numb/$total_numb;
			if ($varFreq < 0.5){
				print OUTFILE ("$varN\t$varFreq\t");
			}
		}
		else {
			$varFreq = $n_numb/$total_numb;
			if ($varFreq < 0.5){
				print OUTFILE ("$varN\t$varFreq\t");
			 }
		}
}
$varN = $array1[8];$n_numb = $array1[9];
if (($n_numb > 3)&&($varN ne $n_A)){
		$tmpAA = &getAA($varN,$n_B,$n_C);
		if ($tmpAA eq $consensusAA){
			$varFreq = $n_numb/$total_numb;
			 if ($varFreq < 0.5){
				print OUTFILE ("$varN\t$varFreq\t");
			}
		}
		else {
			$varFreq = $n_numb/$total_numb;
			if ($varFreq < 0.5){
				print OUTFILE ("$varN\t$varFreq\t");
			 }
		}
 }
 print OUTFILE ("\n");@array1 = ();

#== nucleotide 2 start ==================================
$total_numb = $array2[1];$varN = $array2[2];$n_numb = $array2[3];
print OUTFILE ("$n_B\t$total_numb\t");

 if (($n_numb > 3)&&($varN ne $n_B)){
		$tmpAA = &getAA($n_A,$varN,$n_C);
		if ($tmpAA eq $consensusAA){
			$varFreq = $n_numb/$total_numb;
```

```
                        if ($varFreq < 0.5){
                                print OUTFILE ("$varN\t$varFreq\t");
                        }
                }
                else {
                        $varFreq = $n_numb/$total_numb;
                        if ($varFreq < 0.5){
                                print OUTFILE ("$varN\t$varFreq\t");
                        }
                }
        }
        $varN = $array2[4];$n_numb = $array2[5];
        if (($n_numb > 3)&&($varN ne $n_B)){
                $tmpAA = &getAA($n_A,$varN,$n_C);
                if ($tmpAA eq $consensusAA){
                        $varFreq = $n_numb/$total_numb;
                        if ($varFreq < 0.5){
                                print OUTFILE ("$varN\t$varFreq\t");
                        }
                }
                else {
                        $varFreq = $n_numb/$total_numb;
                        if ($varFreq < 0.5){
                                print OUTFILE ("$varN\t$varFreq\t");
                        }
                }
        }
        $varN = $array2[6];$n_numb = $array2[7];
        if (($n_numb > 3)&&($varN ne $n_B)){
                $tmpAA = &getAA($n_A,$varN,$n_C);
                if ($tmpAA eq $consensusAA){
                        $varFreq = $n_numb/$total_numb;
                        if ($varFreq < 0.5){
                                print OUTFILE ("$varN\t$varFreq\t");
                        }
                }
                else {
                        $varFreq = $n_numb/$total_numb;
                         if ($varFreq < 0.5){
                                print OUTFILE ("$varN\t$varFreq\t");
                        }
                }
        }
        $varN = $array2[8];$n_numb = $array2[9];
        if (($n_numb > 3)&&($varN ne $n_B)){
                $tmpAA = &getAA($n_A,$varN,$n_C);
                if ($tmpAA eq $consensusAA){
                        $varFreq = $n_numb/$total_numb;
                        if ($varFreq < 0.5){
                                print OUTFILE ("$varN\t$varFreq\t");
                        }
                }
                else {
                        $varFreq = $n_numb/$total_numb;
                        if ($varFreq < 0.5){
```

```
                                print OUTFILE ("$varN\t$varFreq\t");
                        }
                }
}
print OUTFILE ("\n");@array2 = ();

 #=== nucleotide 3 start ================================
$total_numb = $array3[1];$varN = $array3[2];$n_numb = $array3[3];
print OUTFILE ("$n_C\t$total_numb\t");
 if (($n_numb > 3)&&($varN ne $n_C)){
        $tmpAA = &getAA($n_A,$n_B,$varN);
        if ($tmpAA eq $consensusAA){
                $varFreq = $n_numb/$total_numb;
                if ($varFreq < 0.5){
                        print OUTFILE ("$varN\t$varFreq\t");
                }
         }
        else {
                $varFreq = $n_numb/$total_numb;
                if ($varFreq < 0.5){
                        print OUTFILE ("$varN\t$varFreq\t");
                }
        }
}
$varN = $array3[4];$n_numb = $array3[5];
if (($n_numb > 3)&&($varN ne $n_C)){
        $tmpAA = &getAA($n_A,$n_B,$varN);
        if ($tmpAA eq $consensusAA){
                $varFreq = $n_numb/$total_numb;
                 if ($varFreq < 0.5){
                        print OUTFILE ("$varN\t$varFreq\t");
                }
        }
        else {
                 $varFreq = $n_numb/$total_numb;
                 if ($varFreq < 0.5){
                        print OUTFILE ("$varN\t$varFreq\t");
                 }
         }
 }
 $varN = $array3[6];$n_numb = $array3[7];
 if (($n_numb > 3)&&($varN ne $n_C)){
        $tmpAA = &getAA($n_A,$n_B,$varN);
        if ($tmpAA eq $consensusAA){
                $varFreq = $n_numb/$total_numb;
                if ($varFreq < 0.5){
                        print OUTFILE ("$varN\t$varFreq\t");
                }
        }
        else {
                $varFreq = $n_numb/$total_numb;
                if ($varFreq < 0.5){
                        print OUTFILE ("$varN\t$varFreq\t");
                }
        }
```

```
                        }
                        $varN = $array3[8];$n_numb = $array3[9];
                        if (($n_numb > 3)&&($varN ne $n_C)){
                                $tmpAA = &getAA($n_A,$n_B,$varN);
                                if ($tmpAA eq $consensusAA){
                                        $varFreq = $n_numb/$total_numb;
                                        if ($varFreq < 0.5){
                                                print OUTFILE ("$varN\t$varFreq\t");
                                        }
                                }
                                else {
                                        $varFreq = $n_numb/$total_numb;
                                        if ($varFreq < 0.5){
                                                print OUTFILE ("$varN\t$varFreq\t");
                                        }
                                }
                        }
                        print OUTFILE ("\n");
                        @array3 = ();
                }# else (real work)
        }# while loop
        close(INFILE);close(OUTFILE);
}#  end of for loop

# == end of script 11 ============================================================
```

Perl Script 12

```
#================================================================================
#  This script is used to determine the unique variants only detected by 454 based on 454 or Clone AllMin
#  orVar files; developed by Binhua Liang;2009
# ================================================================================

#  variables
@minorVar454 = (); @eachline_454 = (); @minorVarClone = (); @eachline_clone = ();$num1_454 = 0;
$num2_454 = 0; $num3_454 = 0;$num1_clone = 0; $num2_clone = 0; $num3_clone = 0;$count_454 = 0;
$count_clone = 0;

#  files
$infile_name1 = "G:/All454_MinorVar_new.txt";      # 454 minor variants file; format: A;163;A;0;0;0;
                                                   # C;-1;0;0;G;-1;0;0;T;0;0;0
$infile_name2 = "G:/AllClone_MinorVar_new.txt";    #clone minor variants file; format: A;163;A;0;0;0;
                                                   # C;-1;0;0;G;-1;0;0;T;0;0;0

#  open files
open (INFILE1, $infile_name1) || die ("$infile_name1 did not open");
open (INFILE2, $infile_name2) || die ("$infile_name2 did not open");

while (<INFILE1>){
        $line = $_;chop $line; push (@minorVar454,$line);
}# end of while loop
close (INFILE1);
```

```
while (<INFILE2>){
        $line = $_;chop $line; push (@minorVarClone,$line);
}# end of while loop
close (INFILE2);

for ($i = 0; $i<scalar(@minorVarClone);$i++){

        @eachline_clone = split (/\t/,$minorVarClone[$i]);@eachline_454 = split (/\t/,$minorVar454[$i]);

        $tmp1_454 = $eachline_454[3]; $tmp2_454 = $eachline_454[4];  $tmp3_454 = $eachline_454[5];
        $tmp1_clone = $eachline_clone[3]; $tmp2_clone = $eachline_clone[4];
        $tmp3_clone = $eachline_clone[5];

        if  ($tmp1_454 > $tmp1_clone){
                $num1_454 = $tmp1_454 - $tmp1_clone + $num1_454;
         }
         else {
                $num1_clone = $tmp1_clone - $tmp1_454 + $num1_clone;
        }
        if ($tmp2_454 > $tmp2_clone){
                $num2_454 = $tmp2_454 - $tmp2_clone + $num2_454;
        }
        else {
                $num2_clone = $tmp2_clone - $tmp2_454 + $num2_clone;
         }
         if ($tmp3_454 > $tmp3_clone){
                $num3_454 = $tmp3_454 - $tmp3_clone + $num3_454;
         }
        else {
                $num3_clone = $tmp3_clone - $tmp3_454 + $num3_clone;
        }

         $tmp1_454 = $eachline_454[7];$tmp2_454 = $eachline_454[8];  $tmp3_454 = $eachline_454[9];
         $tmp1_clone = $eachline_clone[7]; $tmp2_clone = $eachline_clone[8];
         $tmp3_clone = $eachline_clone[9];

        if ($tmp1_454 > $tmp1_clone){
                $num1_454 = $tmp1_454 - $tmp1_clone + $num1_454;
         }
        else {
                $num1_clone = $tmp1_clone - $tmp1_454 + $num1_clone;
         }
         if ($tmp2_454 > $tmp2_clone){
                $num2_454 = $tmp2_454 - $tmp2_clone + $num2_454;
         }
        else {
                $num2_clone = $tmp2_clone - $tmp2_454 + $num2_clone;
        }
        if ($tmp3_454 > $tmp3_clone){
                $num3_454 = $tmp3_454 - $tmp3_clone + $num3_454;
         }
        else {
                $num3_clone = $tmp3_clone - $tmp3_454 + $num3_clone;
        }
```

```
	$tmp1_454 = $eachline_454[11];$tmp2_454 = $eachline_454[12];  $tmp3_454 =
	$eachline_454[13]; $tmp1_clone = $eachline_clone[11]; $tmp2_clone = $eachline_clone[12];
	$tmp3_clone = $eachline_clone[13];

	 if ($tmp1_454 > $tmp1_clone){
		$num1_454 = $tmp1_454 - $tmp1_clone + $num1_454;
	}
	else {
		$num1_clone = $tmp1_clone - $tmp1_454 + $num1_clone;
	}
	if ($tmp2_454 > $tmp2_clone){
		 $num2_454 = $tmp2_454 - $tmp2_clone + $num2_454;
	}
	else {
		 $num2_clone = $tmp2_clone - $tmp2_454 + $num2_clone;
	}
	if  ($tmp3_454 > $tmp3_clone){
		$num3_454 = $tmp3_454 - $tmp3_clone + $num3_454;
	}
	else {
		$num3_clone = $tmp3_clone - $tmp3_454 + $num3_clone;
	 }

	$tmp1_454 = $eachline_454[15]; $tmp2_454 = $eachline_454[16]; $tmp3_454 =
	$eachline_454[17]; $tmp1_clone = $eachline_clone[15]; $tmp2_clone = $eachline_clone[16];
	$tmp3_clone = $eachline_clone[17];

	if ($tmp1_454 > $tmp1_clone){
		$num1_454 = $tmp1_454 - $tmp1_clone + $num1_454;
	 }
	else {
		$num1_clone = $tmp1_clone - $tmp1_454 + $num1_clone;
	}
	if  ($tmp2_454 > $tmp2_clone){
	 $num2_454 = $tmp2_454 - $tmp2_clone + $num2_454;
	}
	else {
		$num2_clone = $tmp2_clone - $tmp2_454 + $num2_clone;
	}
	if  ($tmp3_454 > $tmp3_clone){
		$num3_454 = $tmp3_454 - $tmp3_clone + $num3_454;
	}
	else {
		$num3_clone = $tmp3_clone - $tmp3_454 + $num3_clone;
	}

	@eachline_clone = (); @eachline_454 = ();
}#  end for loop

$count_454 = $num1_454 + $num2_454 + $num3_454;
$count_clone = $num1_clone + $num2_clone + $num3_clone;
print ("Total 454: $count_454\t>10%\t$num1_454\t2-10%\t$num2_454\t<2%\t$num3_454\n");
print ("Total clone: $count_clone\t>10%\t$num1_clone\t2-10%\t$num2_clone\t<2%\t$num3_clone\n");

# == end of script 12 ============================================================
```