In silico discovery of novel cytotoxic T-lymphocyte epitopes in the HIV-1 Pol region in response to antiretroviral resistance mutations

by

Werner Smidt

Submitted in partial fulfillment of the requirements for the degree

Philosophiae Doctor in Bioinformatics
in the Faculty of Natural and Agricultural Sciences,
University of Pretoria
August 4, 2014

Contents

List of Figures

List of Tables

List of Abbreviations

AIDS	Acquired Immunodeficiency Syndrome
3TC	Lamuvidine
ART	Antiretroviral Therapy
ARV	Antiretroviral
ATV	Atazanavir
AUC	Area Under the Curve
AZT	Azidothymidine
CD4	Cluster of Differentiation 4
CD8	Cluster of Differentiation 8
CTL	Cytotoxic T-lymhocyte
DRM	Drug resistance mutations
DRV	Duranavir
EFV	Efavirenz
FET	Fisher's Exact Test
FPV	Fosamprenavir
FTC	Emtricitabine
HLA	Human Leukocyte Antigen
IC	Inhibitor Concentration
MHC	Major Histocompatibility Complex
NFV	Nelfinavir
NNRTI	Non-nucleoside Reverse Transcriptase Inhibitor
NRTI	Nucleoside Reverse Transcriptase Inhibitor
NVP	Nevirapine
PI	Protease Inhibitor
PR	Protease
ROC	Receiver Operator Characteristic
RT	Reverse Transcriptase
RTI	Reverse Transcriptase Inihibitor
RTV	Ritonavir
RTV	Ritonavir
SDS	Significant Differentiable Substitutions
SQV	Saquinavir
Th	T-helper lymphocyte
TPV	Tipranavir
ZDV	Zidovudine (AZT)

Declaration of Originality

I, Werner Smidt, declare that the thesis/dissertation, which I hereby submit for the degree *Philosophiae Doctor* with specialisation in *Bioinformatics* at the University of Pretoria has not been previously submitted by me for degree purposes at any other University and I take note that, if the thesis is approved, I have to submit the additional copies, as stipulated by the relevant regulations at least six weeks before the following graduation takes place and if I do not comply with the stipulations, the degree will not be conferred upon me.

. . . .

SIGNATURE OF AUTHOR DATE

Acknowledgments

I would like to thank each and every person that supported me through the course of my postgraduate career. Special mention to the National Research Foundation for their continued financial support. To my supervisor, Fourie Joubert, I would like to extend my gratitude in the faith he has shown by allowing me the freedom to pursue my academic goals without impedance. His counseling, academic support and facilitation of bursary funds are highly appreciated. To my lab colleagues, those that have come and gone, I am in your debt for the knowing and unknowing support you have provided. To Peter Hraber and the staff at the Theoretical Biology and Biophysics group at the Los Alamos National Laboratory, thank you for allowing me to spend valuable time and sharing priceless advice with me. To Zabrina Brumme Mark Brockman and Jonathan Carlson, I'd like to extend my gratitude in providing advanced access to data. To my family, mother, father, brother and sister, thank you for all your encouragement and keeping me motivated. Without any of you, this study would not have been possible. To my in-laws, thank you for all your love and support. Lastly, to my wife Alisa, whom I met at the beginning of my studies, your support and love throughout this degree means the world to me - a special thank you for your patience and feedback. The debt I owe you is limitless and I dedicate this thesis to you in its entirety.

Abstract

The Acquired Immunodeficiency Syndrome pandemic continues to have a large social impact. Many advances in the treatment of infection by the causative agent, Human Immunodeficiency Virus, have been made in the last three decades. However, this treatment often means a life-long rigorous adherence to treatment and acquisition of resistance mutations to antiretrovirals. Thus far, the efficacy of promising vaccines has been disappointing. In the last decade, interest has grown concerning the interaction between mutations conferring resistance to antiretrovirals and the effect this has on epitopes recognized by cytotoxic-T-lymphocytes (CTL). Investigating this is a difficult task, owing to both the extreme polymorphism of HIV and the polymorphism of the Human Leukocyte Antigen (HLA) molecules that present peptides to the CTLs. A large amount of HLA-associated CTL escape mutations have been discovered. Together with this, computational approaches in CTL epitope discovery is becoming increasingly accurate. Here, a method of imputing HLA type from patients together with predicting the influence of anitretroviral mutations was used to discover potential epitopes for the HLA B*15 and B*48 types in the HIV-1 Subtype B *pol* region.

1 Introduction

The Human Immunodeficiency Virus (HIV) is the primary agent contributing to the condition known as Acquired Immunodeficiency Syndrome (AIDS). This condition is characterised by compromised immunity in late stages of HIV infection and infected individuals and high risk of succumbing to especially opportunistic infections and neoplasms. According to the WHO report of 2012, there were 35.3 million people infected with HIV worldwide with 2.3 million new infections that year. Additionally, 1.6 million world-wide deaths were estimated to be due to AIDS-related illnesses. Infection of other hosts is typically facilitated through sexual intercourse and blood-to-blood contact.

HIV is a single-stranded RNA lentivirus of the Retroviridae family. Being a retrovirus, the virus contains a reverse-transcriptase enzyme that facilitates reverse transcription of viral RNA to DNA and subsequent integration of the viral genome into the infected cell's genome. This integration into the host genome allows mutual propagation when the infected cell replicates. It is for this very reason that the virus becomes entrenched in the infected host. Progression of HIV-infection is particularly slow, lasting 8-12 years from the time of infection to onset of AIDS. The genome of HIV is particularly polymorphic owing to (among other factors) an error-prone HIV reverse transcriptase. This allows relatively quick escape of immune responses and development of resistance to antiretrovirals, the primary drugs used to combat HIV infection. In this chapter, HIV as a pathogen is discussed. Its infectivity and host immune responses as well as treatment options will be covered. The focus is on protection from the virus by the host's adaptive immune response, specifically cell-mediated immunity, which reacts against endogenous viral peptides. Importantly, it will be discussed how antiretroviral resistance mutations and immunological responses may interact.

1.1 Human Immunodeficiency Virus

The Human Immunodeficiency Virus (HIV) is a lentivirus of the *Retroviridae* family according to the *International Committee on Taxonomy of Viruses* (`http://ictvonline.org/`). The virus occurs throughout the world and is extremely diverse [Rambaut *et al.*, 2004]. The virus is divided into two main types, HIV-1 and HIV-2. The main differences between the two types are virulence and origin. HIV-1 has a higher virulence and infectivity factor and also causes the progression to AIDS faster (8-10 years) than HIV-2 (15-20 years). The origin of HIV-1 is thought to be a complicated series of zoonotic events involving the Simian Immunodeficiency Virus (SIV) from the chimpanzee and Western gorillas in central regions of Africa [Sharp and Hahn, 2011]. HIV-2 is thought to have originated from the sooty mangabey, from West Africa [Sharp *et al.*, 1995]. This is also reflected in the distribution of HIV-1 and HIV-2, with HIV-2 infection occurring more commonly in West Africa. The focus of this study is on HIV-1. HIV-1 is further subdivided into groups and each group contains a set of subtypes [Hemelaar *et al.*, 2006]. The most common group of HIV is group M. This group contains the subtypes A, B, C, D, F, G and H. Distribution of the major subtypes is shown in Table 1.1. Subtype C is the most common subtype, claiming approximately 47.2% of all worldwide infections and particularly concentrated in sub-Saharan Africa. Subtype A makes up approximately 26.7% and subtype B 12.3%.

Table 1.1: The table shows the proportion and geographical distribution of common subtypes of HIV-1. The most common subtype, is subtype C, followed by subtype A and subtype B [Hemelaar *et al.*, 2011].

Subtype	Proportion	Common regions
A	12%	Asia
B	11%	Americas, Europe, Australia
C	48%	Sub-Saharan Africa

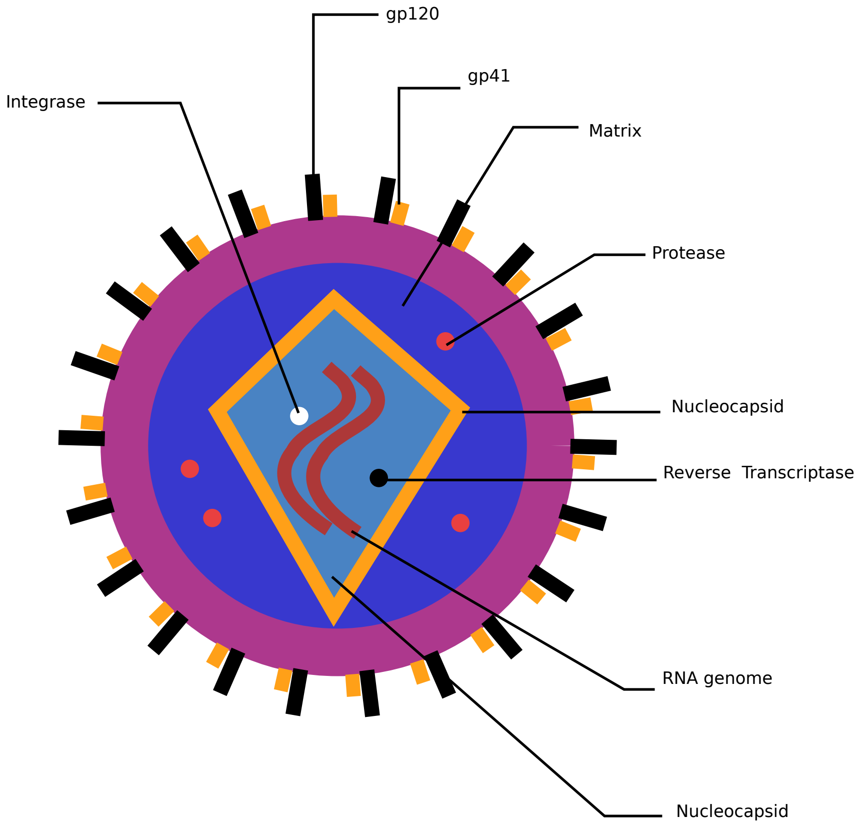


Figure 1.1: A schematic representation of the HIV-1 viron. Two copies of the +ss-RNA genome are bound to a nucleocapsid protein. The surrounding capsid is made up of capsid protein units and also house HIV reverse transcriptase and integrase. This core particle is surrounded by a matrix protein. Finally, a lipid envelope surrounds the virus, with embedded proteins *gp120* and gp41 that are essential for attachment and entry into a target cell. Source: Adapted from: `http://web.archive.org/web/20050531012945/http://www.niaid.nih.gov/factsheets/howhiv.htm`.

1.1.1 Structure

The fine structure of HIV-1 was determined not long after the initial discovery [Gelderblom *et al.*, 1987]. A schematic representation of the structure of HIV is shown in Figure 1.1. Its genome consists of two copies of two single stranded, positive sense RNA molecules. The genetic material is bound to a protein p7 surrounded by the capsid protein, p24. This core particle is surrounded by the matrix protein, *gp120* and the whole core particle is surrounded by a phospholipid layer containing the attachment and fusion glycoproteins, *gp120* and gp41. Non-structural proteins include serine protease, reverse transcriptase and integrase enzymes which are pivotal to complete post-entry processes.

1.1.2 HIV genome features

The RNA genome of HIV is approximately 9.7k nucleotides in length [Wain-Hobson *et al.*, 1985, 1991]. It contains 9 genes including structural and non-structural proteins. As is typical of viruses of the Lentivirinae family, the genome is exceedingly complex and in HIV the sum of the length of the transcripts produced by all the genes is greater than the size of the actual genome. This is achieved through a combination of alternative splicing and alternative reading frames [Purcell and Martin, 1993].

Layout of the HIV-1 genome is depicted in Figure 1.2 on the following page. Structural proteins making up the core particle are encoded by the *gag* gene, containing the proteins *p17*, *p24*, *p2*, *p7*, *p1* and *p6*. The *pol* gene codes for enzymes necessary in integration of the HIV genome in the host cell genome as well as maturation of the virion. The *env* gene codes for the protein *gp160*, which contains *gp120* and *gp41* as well as accessory proteins *rev*, *tat* and *vpu*. The *vif* gene lies between *pol* and *env*. The *nef* gene overlaps the `3'LTR` region [Frankel and Young, 1998]. During initial transcription of the HIV genome, RNA splicing in the nucleus results in only two gene products, i.e. *tat* and *rev*. The function of *rev* is to ensure export of the entire HIV transcript to the cytosol where appropriate splicing can occur to produce the other gene products. Transcription is enhanced by *tat* and thus works in synergy with *rev* [Romani *et al.*, 2010]. The other gene products, *Nef*, *Vif*, *Vpr* and *Vpu* will be discussed in the subsequent section.

The *gag* and *pol* genes overlap and due to a stem-loop structure, there is sometimes a "read-through" of the gag termination codon by the ribosomes [Wills *et al.*, 1994]. The ribosome reads back and pairs up with the -1 nucleotide reading frame of the *pol* gene. This allows for the expression of a *gag-pol* complex. A comparatively higher amount of *gag*-derived protein products are needed to make up the core viral particle than the need for enzymatic *pol*-derived products, and thus this read-through happens in a 1:20 ratio. This ensures that there is a preference for the expression of structural proteins over enzymatic proteins.

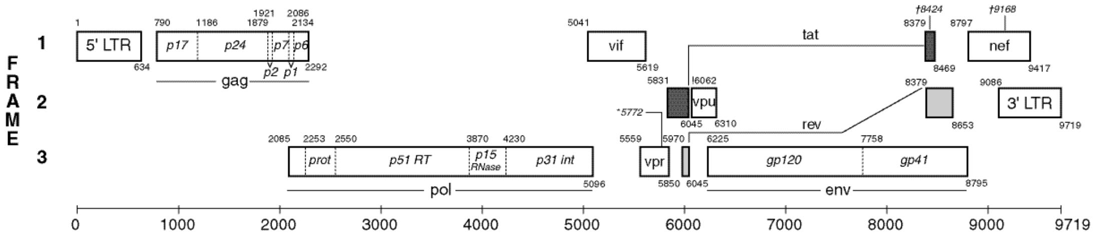


Figure 1.2: This figure depicts genes and other elements of the HIV-1 genome. At both the 5' and 3' end there are long terminal repeats (LTR). The main gene products are *gag*, *pol*, and *env*. The genes *vif*, *vpr*, *vpu* and *nef* are expressed via alternative ORFs, with *nef* residing in the 3' LTR region. The *gag*, *env* and *pol* gene products are post-translationally cleaved to yield their respective products. Source: `http://www.hiv.lanl.gov/content/sequence/HIV/MAP/landmark.html`.

1.1.3 Infection and replication

The membrane bound protein, CD4, acts as a primary ligand for the HIV glycoprotein *gp120*[Kwong *et al.*, 1998]. Attachment of *gp120* is followed by attachment of the fusion protein, gp41 to either CCR5 or CCRX4. This binding induces fusion of the viral envelope with the target cell membrane, and is followed by internalisation of the viral particle [Wyatt and Sodroski, 1998]. After entry, the viral core particle and nucleocapsid undergo proteolysis, releasing the RNA copies of the HIV genome as well as accessory proteins, i.e. integrase, protease and reverse transcriptase [Auewarakul *et al.*, 2005]. The RNA genome is reverse-transcribed by RT and the product can be inserted into the host cell's genome by integrase [Chiu and Davies, 2004]. This integration into the host genome allows perpetuation of the virus along with the T-cell upon cell replication.

The expression of the whole HIV genome usually occurs in conjunction with the activation of the infected cell, notably under the influence of the transcription factor NF-κB, which is expressed at high levels after T-cell activation [Duh *et al.*, 1989]. Initially, the mRNA transcript that consists of the entire HIV RNA genome is spliced, producing transcripts coding for HIV proteins [Felber *et al.*, 1989]. This includes *Rev*, which allows the transport of the entire HIV RNA transcript to the cytosol intact [Fischer *et al.*, 1995]. The *gag* and *env* protein products are produced.

The protein *pol* is cleaved into four different components: protease (PR), reverse transcriptase (RT), RNAseH and integrase (IN) [Haseltine, 1991]

Envelope proteins are transported as *gp160*, before cleavage by Furin in in the Golgi apparatus and finally, the products *gp120* and *gp41* are delivered to the cell membrane [Decroly *et al.*, 1994]. In close proximity are gag proteins and other viral enzymes. Although the budding starts to occur at this point, the structural proteins still need to be cleaved by HIV protease. The protein *gag* contains structural proteins and is cleaved into matrix, spacer, capsid and nucleocapsid components [Haseltine, 1991]. The nucleocapsid proteins associate with the HIV genome copies and the mature virion is released by completion of budding.

1.1.4 HIV-1 pathogenesis

The pathogenesis of HIV-1 is a complex issue involving, especially, apoptotic (programmed cell death) signals. This is further complicated by the dichotomous effect viral products have over apoptosis. Some protein products, such as *Nef*, *Vpr* and *Tat* have both proapoptotic and anti-apoptotic activity. Apoptosis remains the major factor for CD4+ T-cell depletion [Cummins and Badley, 2010]. Since expression of HIV-1 is dependent on the NF-κB transcription factor, activated T-cells express large quantities of HIV-1 particles. This makes non-infected activated Th cells susceptible to infection. Other non-infected cells, that include cytotoxic T-lymphocytes (CTL) and B-cells are also susceptible to the apoptotic effects of the HIV-1 protein products that can exist outside of the infected cells in the form of exosomes [Lenassi *et al.*, 2010]. Endocytosis of these protein products make the non-infected cells more prone to spontaneous apoptosis. These mechanisms partly explain why HIV has such a deleterious effect on the Th population when, on average, only 1 in 1000 to 1 in 100 cells are infected with HIV-1 [Coffin and Swanstrom, 2013].

1.1.5 HIV polymorphism

Polymorphism in the HIV-1 genome is extensive. This is in part due to the error prone HIV-1 reverse transcriptase, which induces a mis-transcription of the HIV-1 RNA genome on average once per 2000 ribonucleotides transcribed [Takeuchi *et al.*, 1988]. Figure 1.4

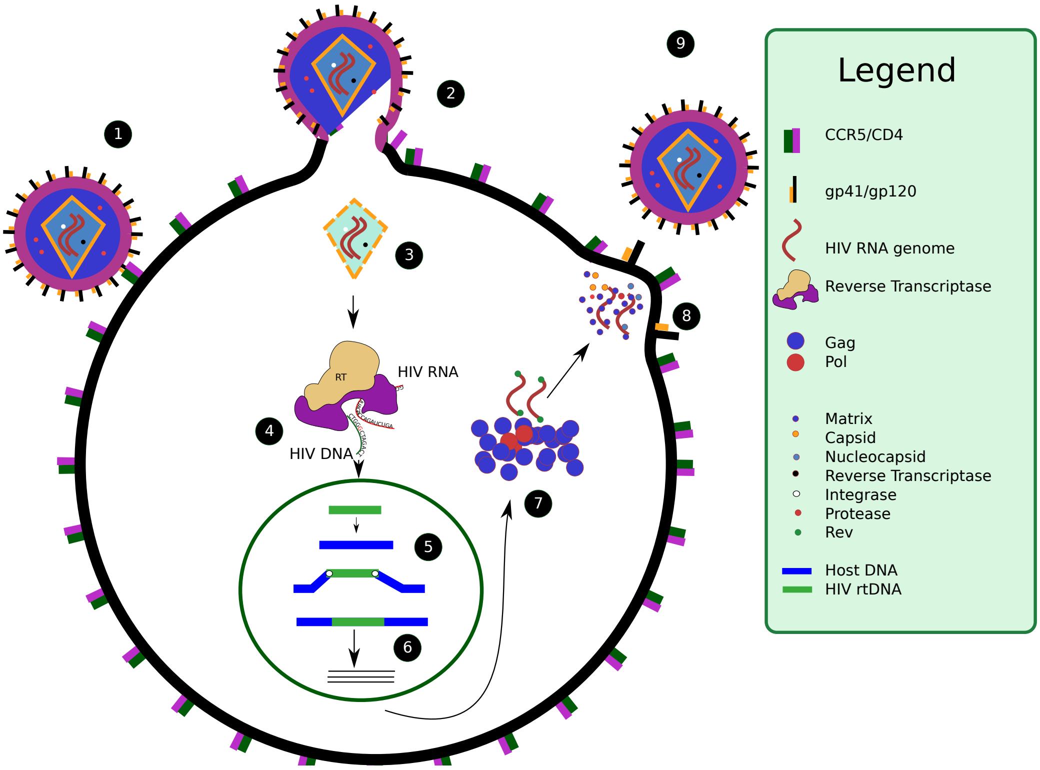


Figure 1.3: This figure illustrates HIV infection and replication in an HIV target cell, which is typically a T-helper cell. 1) *gp120*binds to the CD4 receptor. 2) Fusion is mediated by gp41 and the core viral particle enters the cell. 3) The core particle is broken apart by cellular proteases and releases two copies of the HIV RNA genome as well as reverse transcriptase (RT) and Integrase (INT). 4) The RNA genome is reverse-transcribed by RT and the DNA copy of the genome is transported to the nucleus. 5) Integrase facilitates integration of the HIV rtDNA into the host cell genome. 6) After activation by NF-κB, transcripts are expressed and the full HIV RNA transcript transported to the cytosol and protected by Rev. 7) Precursor proteins are produced and the products, which include the HIV genome, assemble at the cell membrane. 8) The immature virion starts to bud and HIV protease cleaves the precursor proteins so the mature virion is formed. The *gp120*and gp41 proteins are already embedded in the cell membrane. 9) A new virus is produced and buds from the cell. Source: Adapted from `http://commons.wikimedia.org/wiki/File:Hiv_gross_german.png`.

illustrates an example. Furthermore, the infected cell, in an attempt to protect itself from retroviral infections, has the enzyme group APOBEC, which causes G to A substitutions in single stranded DNA transcripts [Wood *et al.*, 2009]. Furthermore, if the cell is infected by multiple strains of the virus, recombination may occur due to RT's ability to "jump" from one transcript to another (see Figure 1.5 on the following page) [Luo and Taylor, 1990]. Taken together, this allows the genome to evolve rapidly to accommodate for selection pressure. Figure 1.6 on page 10 shows the level of substitution at various positions of HIV PR in sequences obtained from patients undergoing drug therapy for HIV-1 infection.

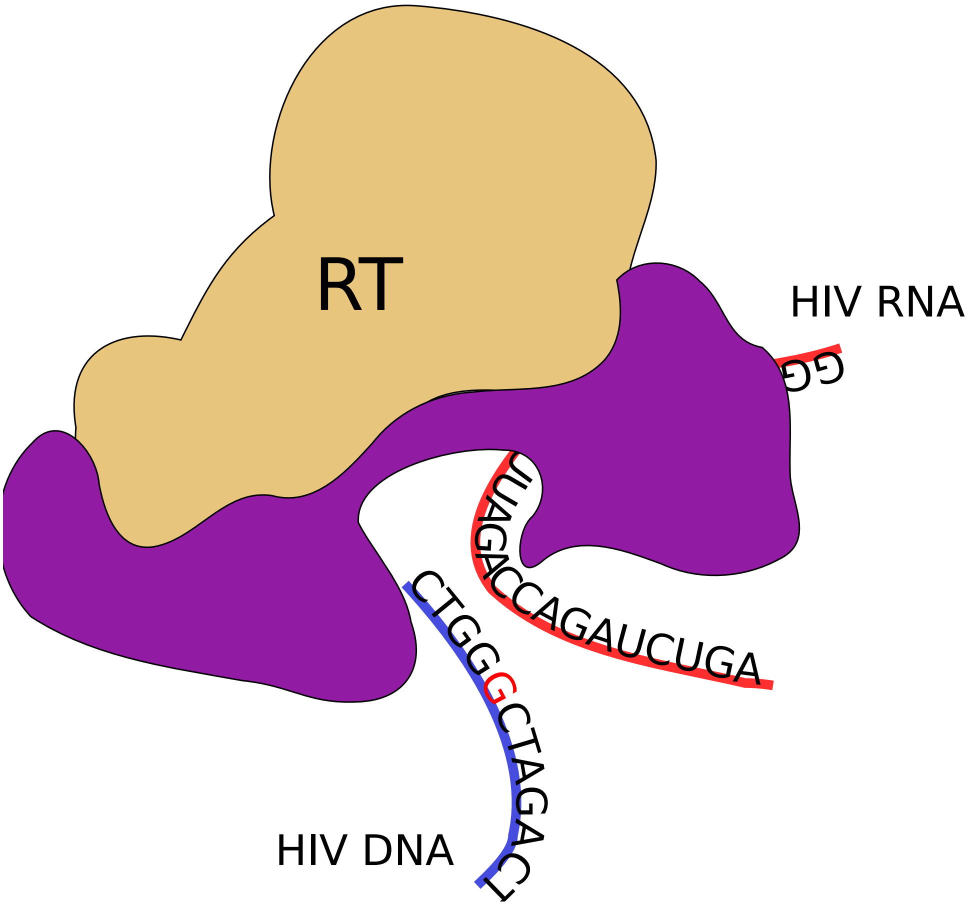


Figure 1.4: This figure illustrates the errors that HIV RT can induce in the rtDNA of the HIV genome. The RNA is threaded through the RT and a complimentary DNA strand is constructed. The red nucleotide indicates a erroneously placed nucleotide. Source: adapted from the PDB structure *3HVT*.

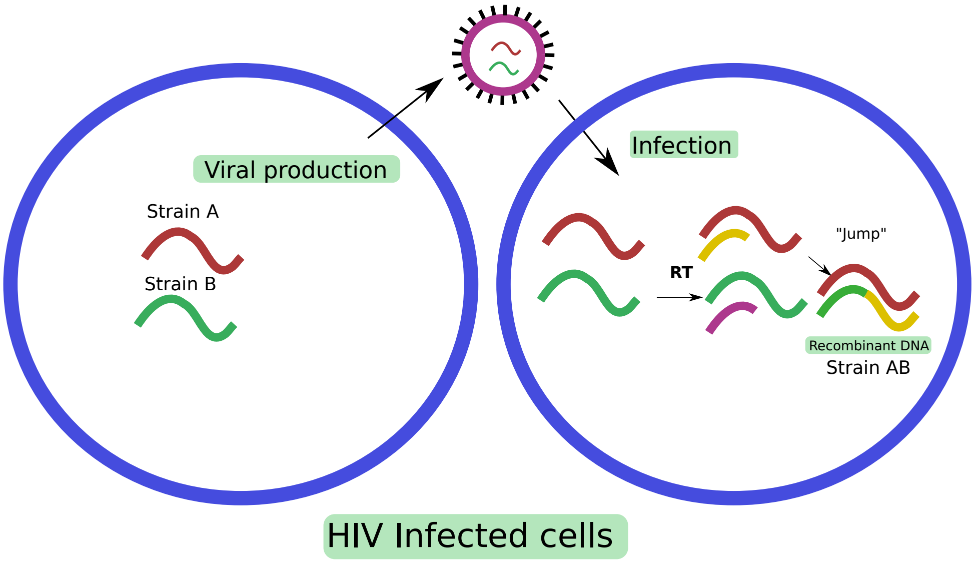


Figure 1.5: Cells infected by multiple strains of HIV can produce virions containing copies of each strain (A and B). Upon infection of a new cell, during reverse transcription of strain A, the RT-DNA complex can jump to strain B producing a recombinant strain, AB. Source: Adapted from the literature [Robertson *et al.*, 1995].

1.1.6 Progression of disease

HIV preferentially infects cells that express the CD4 co-stimulatory protein on the cell membrane. This includes primarily T-helper lymphocytes, but other cells like folicular dendritic cells and macrophages can also be infected. Soon after initial infection, there is a massive decline in the T-helper (CD4+) in conjunction with a high rise in HIV particles [Vergis and Mellors, 2000]. This is followed by a sharp decline in the amount of HIV particles and partial recovery of the T-helper cell population. As the disease progresses, a more steady overall decline in the T-helper lymphocyte population is observed. When the level drops below 100-300 T-helper cells/ml blood, immunity is severely compromised and at this point the infected individual has AIDS. The T-helper cells are pivotal to effective adaptive immune responses [Moir and Fauci, 2009]. The adaptive immune response is involved in immunological responses to specific pathogen targets. This progression is illustrated in Figure 1.7 on page 11.

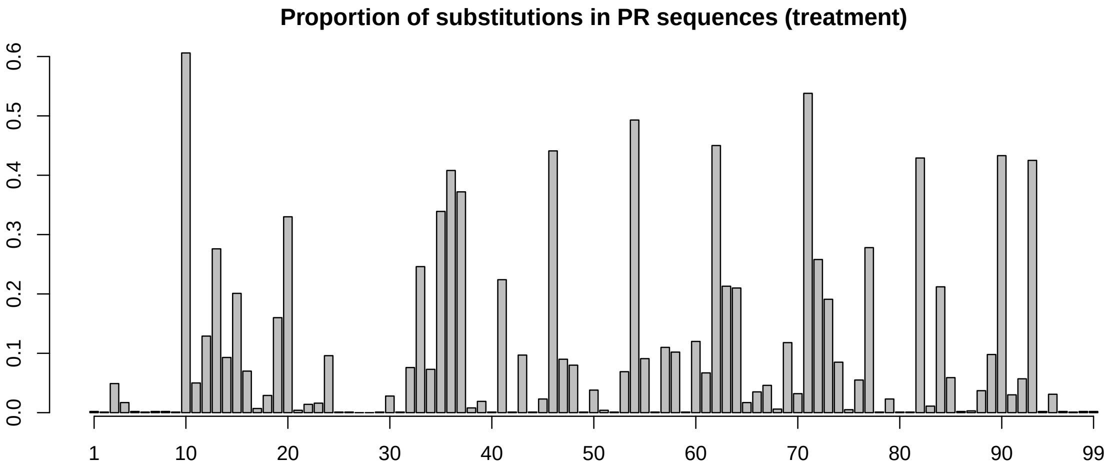


Figure 1.6: This figure illustrates the mutation frequencies in HIV-1 PR spanning positions 20-80. The bars show the $1 - p_{cons}$ frequency, where p_{cons} is the proportion of the consensus residue at a particular position. Source: Figure produced by the author by using HIV Subtype B protease data from treatment experienced patients and determining the proportion of sequences that do not contain the consensus protease (for subtype B) residue at that position.

1.2 The adaptive immune response and the influence of HIV infection

To understand why prolonged HIV infection eventually leads to AIDS, a basic understanding of the function of the immune system is needed. The immune system is a collection of cells and organs that protect the body from foreign entities such as viruses, bacteria, protein toxins as well as cancerous cells [Flajnik and Kasahara, 2010]. It is an exceptionally efficient and well-coordinated response that has the ability to distinguish self from non-self. The immune system can be divided into two main parts, the innate- and adaptive immune system. The innate immune system is a rapid response that is non-specfic, i.e. most foreign entities are dealt with in the same way [Janeway and Medzhitov, 2002]. For example, phagocytosis by macrophages is due to the innate affinity of macrophages for the cell wall of bacteria. However, since pathogens have evolved mechanisms to evade this response, a second response is sometimes required, namely the adaptive immune response [Pancer and Cooper, 2006]. The adaptive immune response differs from the innate response in that it has the ability to home in on specific targets, named antigens that can be of extra- or intracellular origin. The response also has the ability to learn from previous

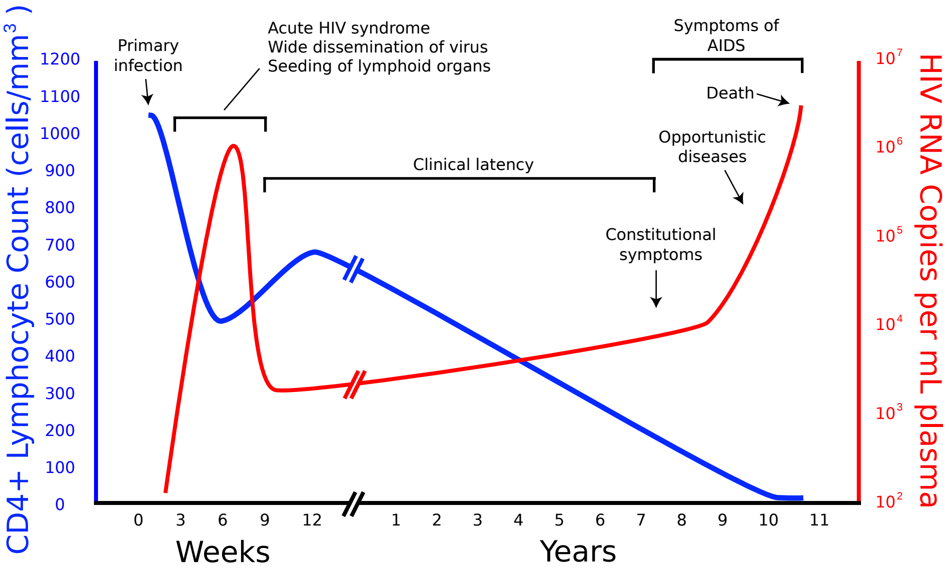


Figure 1.7: This figure illustrates the progression of HIV infection to AIDS. The blue and red lines represent CD4+ T-cell (Th cell) count and HIV plasma load respectively. Initially, there is a sharp spike in HIV levels, correlated with a decline in the Th cell population. The levels of HIV drop sharply and levels of Th increase to sub-normal levels. This the latent stage of the disease which is marked by a progressive loss of Th cells. After a period of 8 years, symptoms of the disease become apparent in the form of an increase in opportunistic infections. After a few more years, the Th cell level is low enough for the immune system to be completely compromised and manifests itself as AIDS. Source: `http://upload.wikimedia.org/wikipedia/commons/0/0e/Hiv-timecourse_copy.svg`.

exposures and remember past antigens. First exposure response to an antigen is slower than the innate response, but subsequent exposures are faster [Dempsey *et al.*, 2003]. Results of a loss of adaptive immunity is demonstrated by later stages in HIV infection and AIDS [Clerici *et al.*, 1989]. Loss of this response makes sufferers of AIDS extremely susceptible to opportunistic infections that would otherwise be controlled by adaptive immunity. The adaptive immune system is further divided into two main parts, namely cell-mediated immunity, which deals with intracellular antigens and humoral immunity, which deals with extracellular antigens. In this section, an overview of mechanism and function of the adaptive immune system will be given as well as the importance of adaptive immune responses in controlling HIV infection and is a reason certain HIV-infected individuals have a tendency to have a very long latent period before the onset of AIDS.

1.2.1 Cell-mediated immunity

Cell-mediated immunity (CMI) is an immune response directed towards mainly intracellular pathogens, such as viruses, and cancerous cells. Effectors of the CMI are cytotoxic T-lymphocytes (CTL) [Lieberman, 2010]. The function of CTLs is to scrutinize the intracellular inventory of proteins of somatic cells. This is achieved by examining small peptide fragments presented by Major Histocompatibility Complex (MHC) Class I, which is present on the cell surface of all nucleated cells. The antigen processing and presentation pathway within cells is responsible for extracting and presentation of peptide fragments [Pamer and Cresswell, 1998; Cresswell *et al.*, 2005]. This process is best illustrated in Figure 1.8 on the next page where a typical virus will be used as an example. During production of viral protein products, some proteins are marked by the protein ubiquitin for recycling. The cell's proteasomes degrade the protein products into small fragments. The fragments are transported to the endoplasmic reticulum (ER), appropriate peptides are loaded onto MHC Class I and the resultant fragment is presented on the cell surface. If an activated (see the subsequent section) CTL recognizes the peptide (via the T-cell Receptor), the CTL will cause a signal cascade to occur triggering cell apoptosis [York and Rock, 1996]. The advantage is that a source of virus-producing cell is eliminated, reducing the overall virus load.

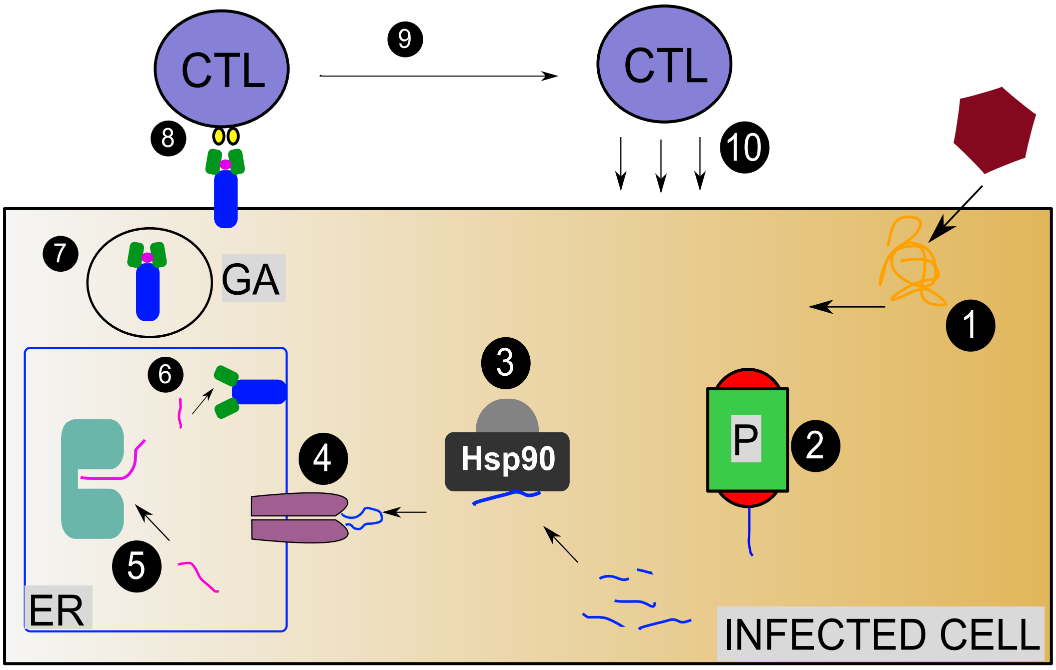


Figure 1.8: 1) A hypothetical virus enters the cell and production of the protein products take place. 2) Some of the viral peptides are digested by the cell's proteasomes. 3) Fragments are carried by chaperones to the Transporter associated with Antigen Presentation (TAP). 4) TAP transports the peptide into the ER lumen. 5) The peptide gets trimmed on its N-terminal by ERAP. 6) The peptide binds with MHC. 7) The peptide-MHC complex is transported by the Golgi apparatus to the cell membrane. 8) A cytotoxic T-lymphocyte (CTL) with a complimentary receptor binds to the MHC-peptide complex. 9) Proliferation and activation of the CTL occurs. 10) The CTL signals the cell's destruction. Adapted from [Pamer and Cresswell, 1998; Cresswell *et al.*, 2005].

Fragment generation by the proteasome

The human proteasome is a 20S barrel-shaped structure (see Figure 1.9 on the following page) with the main purpose of specifically degrading proteins in the cytosol [Unno *et al.*, 2002]. It has affinity for proteins marked for degradation, e.g. by ubiquitination. The proteasome's core cleaving regions are isolated from the cytosol, which protects normal cellular proteins from random degradation. Two types of proteasomes can exist within a cell, the constitutive and the IFN-γ induced immunoproteasome [Pamer and Cresswell, 1998]. The constitutive proteasome is more attuned to cutting after acidic residues while the immunoproteasome prefers to cut after hydrophobic residues [Gaczynska *et al.*, 1996]. The advantages of altering the specificity of the proteasome has to do with the diversity

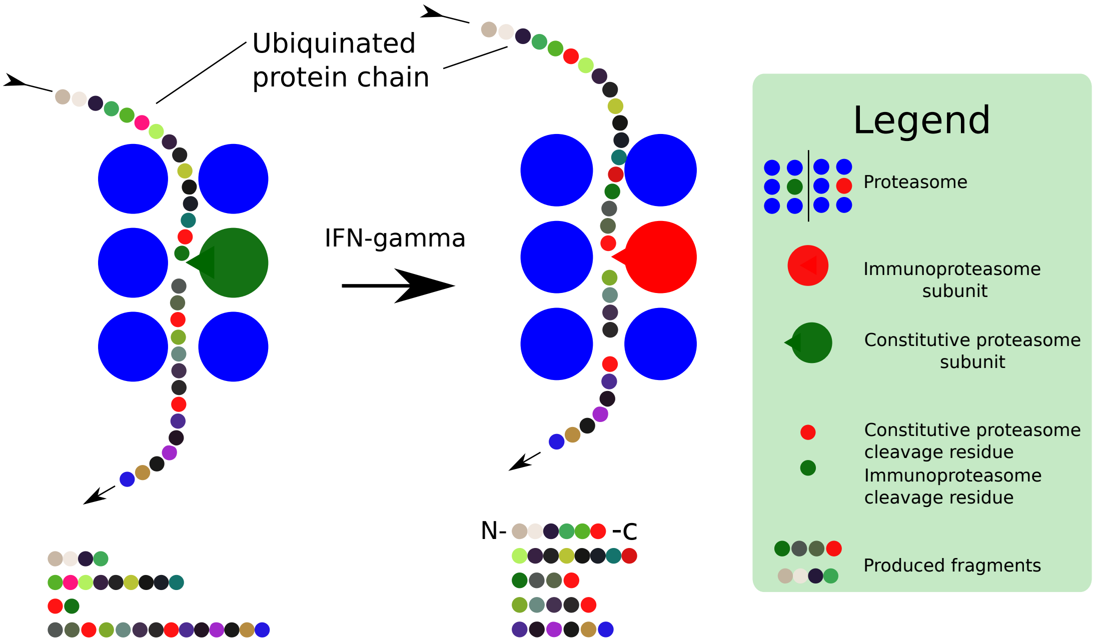


Figure 1.9: The proteasome is a multimeric protein with a barrel-like structure. The figure depicts the cleavage of a protein chain into smaller fragments by the consitutive and immunoproteasomes. The proteasome on the left is the constitutive proteasome. It produces a cleavage site at circles marked dark green along the protein chain (long chain of coloured dots). The green subunit is a subunit in the proteasome when the cell is not under the influence of IFN-γ. With stimulation by IFN-γ, various subunits are replaced and the cleavage preference of the proteasome changes. The immunoproteasome in this figure cleaves at residues marked by a red circle. Below each proteasome are the fragments produced. It is clear from this that the immunoproteasome generates more fragments. Source: adapted from the literature and generated by the author [Unno *et al.*, 2002].

of peptides generated.

Fragments produced by the immunoproteasome usually contain hydrophobic residues at the C-terminal side, which seems preferable for MHC Class I binding (see Section 1.2.1 on the next page). If many of the same protein is degraded by the proteasome, the resultant fragments will overlap. A cleavage site typically has a probability of being cleaved. This ensures the generation of a diverse set of peptides.

Transport into the endoplasmic reticulum

The MHC complex needs to associate with a potential epitope in the ER. Transport of the proteasomal fragments to the ER is thus an essential step in antigen presentation. Two

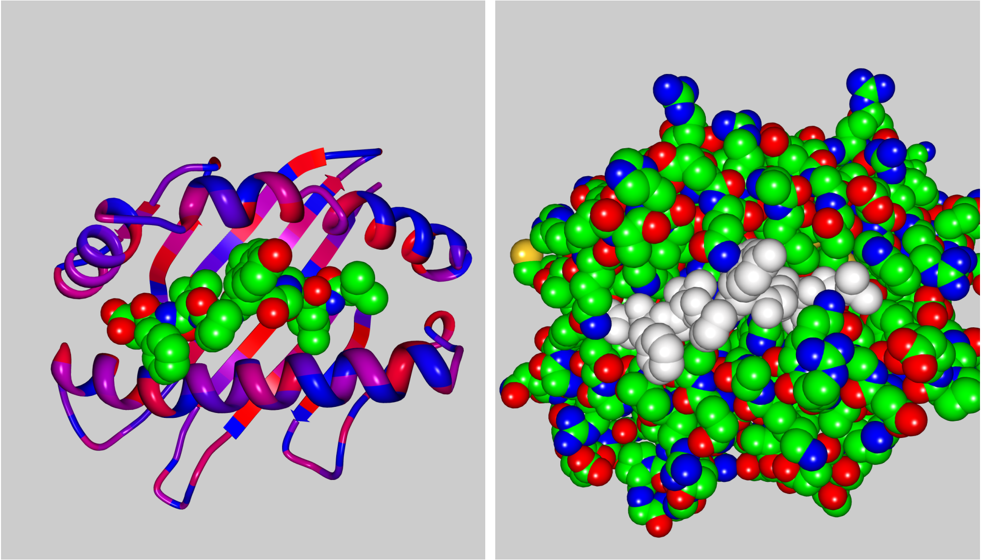


Figure 1.10: This is a structural representation of the MHC binding groove for HLA A*0201 binding the peptide LLGFYPVYV. The picture on the left shows a ribbon model of the MHC binding groove with the peptide between the two alpha helices. The figure on the right shows the space filling model for the MHC binding groove and how only part of the peptide is exposed while the residues at the terminal ends are buried. Source: PDB structure *1DUZ*.

questions need to be addressed here: i) How peptides are transported across the membrane of the ER and ii) How peptides are transported to this cross-membrane transporter. The Transporter Associated with antigen Presentation (hereafter, TAP) facilitates transport of peptides from the cytosol to the lumen of the ER and various chaperones transport peptides to TAP from the proteasome (Wright et al., 2004). TAP also has a binding motif, although a little more relaxed. Especially the three N-terminal amino acids and the first C-terminal amino acid are important in determining transport efficiency.

The peptide transported into the ER is generally not of the appropriate length to be loaded onto the MHC molecule. The endopeptidase, ERAP, ensures that the peptide is trimmed to 9-11 (9 nominal) to fit in the MHC binding groove [Chang *et al.*, 2005]. Note that this trimming only occurs from the N-terminal end.

Loading of peptides onto major mistocompatibility class I

The MHC Class I molecule is the final step before the peptide is presented on the cell membrane. The binding site for the peptide consists of a bed of anti-parallel β-strands, overlaid and flanked by two α-helices [Li and Raghavan, 2010]. The peptide ligand binds in the groove formed by the alpha helices, as shown in Figure 1.10 on the preceding page. There are six MHC class I genes, termed Human Leukocyte Antigen (HLA) in humans, HLAs A, B, C, D, E and F. Of importance here, are the classes A, B and C, which bind peptides from intracellular proteins. HLA molecules are extremely polymorphic and thousands of HLA allotypes are known, the most polymorphic being HLA A, B and to a certain extent, HLA C [Jin and Wang, 2003]. The polymorphisms determine the binding motif for each HLA type and thus the composition of the peptide determines the affinity of it to MHC. The MHC binding groove contains position-specific binding pockets that have strong affinities for a specific range of residues. It preferentially binds peptides that are nine amino acids in length, but peptides with lengths ranging from 8-11 are not uncommon. During motif discovery experiments, it was revealed that there exist 2-3 pockets that can bind a limited range of amino acids [Rammensee *et al.*, 1995]. These positions are called anchor positions and it is essential that the correct amino acids exist in the correct position in order for the potential ligand to bind strongly to the groove. The binding groove of the HLA A*0201 molecule with bound peptide `LLFGYPVYV` is shown in Figure 1.10 on the previous page. From the clefts on either side of the binding groove, it is easy to see that the length of the binding peptide is limited. The peptide needs to span a majority of the groove to form interactions with crucial binding pockets which exist near the clefts. It has been demonstrated that when the P1 residue of the peptide is removed, binding of this peptide to the groove still occurred, but at a severely lowered affinity [Khan *et al.*, 2000]. Empty MHC Class I molecules do not form stable complexes and are not presented on the cell membrane, ensuring that there is no wastage. Binding of an MHC ligand changes the conformation of the groove from an open to a closed state. This allows the MHC-peptide complex to have a long half-life, sometimes even tens of hours [Khan *et al.*, 2000]. The long exposure grants a higher probability that a circulating CTL with a complementary TCR encounters the peptide. Figure 1.11 on the following page illustrates binding motifs for HLA allotypes A*0201 and B*5801. The figure depicts positions of interest along a typical nonamer peptide. For both the allotypes, the primary anchor positions exist in positions N-2 and C-1. For the anchor residues, preferred and tolerated (i.e. does not necessarily abrogate binding, but is lesser of a contributor) amino

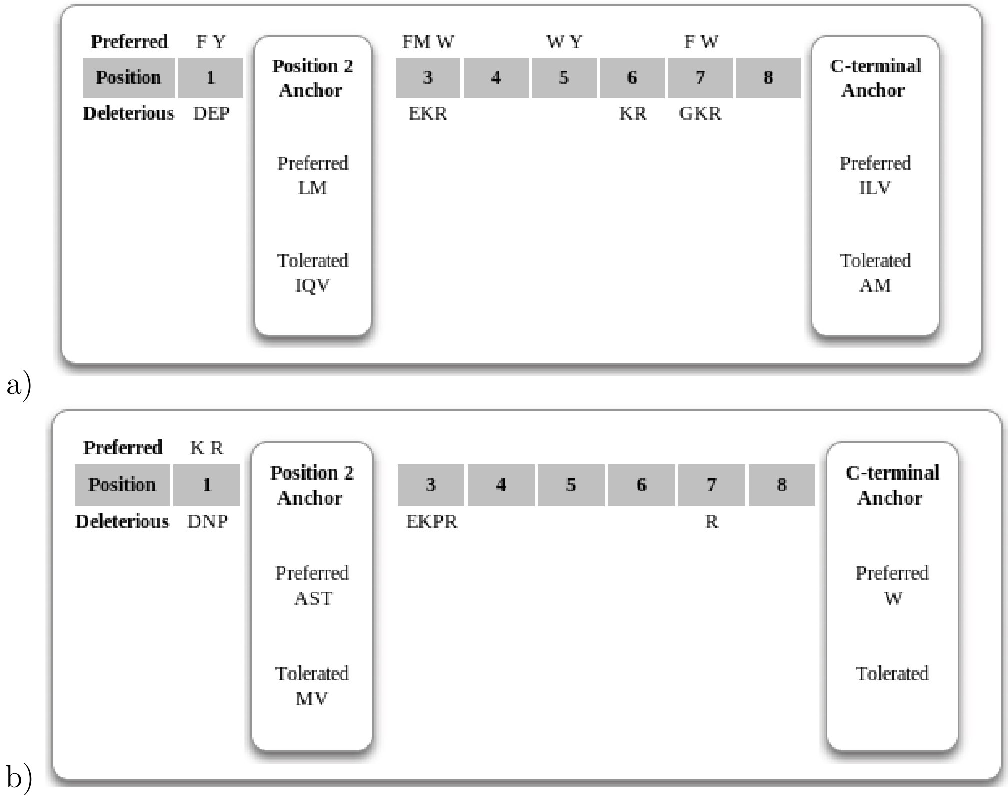


Figure 1.11: Above are the binding motifs for HLA A*0201 (a) and HLA B*5801 (b). It illustrates the preferred residues at positions along a nonamer peptide for binding to the MHC binding groove. The anchor positions show the residues essential for binding of the peptide. Deleterious residues are also listed, for example for both allotypes, glutamic acid, arginine and lysine at position N-3 are considered deleterious for a peptide to bind to these motives. Conversely, they differ by preferred residues in position N-2. HLA A*0201 prefers a leucine or methionine at position N-2 while B*5801 prefers an alanine, serine or threonine at the same position. Source: Immune epitope database [Vita *et al.*, 2010].

acids are listed. Other positions show preferred and deleterious residues. Deleterious residues tend to, by themselves, abrogate binding. It is worth noting that B*5801 has a more restrictive binding motif than A*0201. The difference in binding motif results in HLA restriction, which means that presented epitopes will often differ between individuals. However, this does not necessarily mean that it confers weaker immunity by virtue of a smaller binding repertoire size, as will be discussed in the next section.

HLA Class I nomenclature and classification HLA nomenclature and classification is important in the investigation of CTL epitopes. Given the extreme polymorphic nature of the HLA genes, classification can be difficult. Initially, HLA nomenclature was invented to determine histocompatibility between individuals for transplant purposes. Individuals with closely related HLA allotypes are less likely to develop rejection against transplants and depended on serological (i.e. tagging by specific antibodies) methods. To further enhance the accuracy of classification, genetic methods were utilized. The basic nomenclature for HLA allotypes is in the format `HLA-X:YY:ZZ`, where X is the HLA gene, YY is the serotype, ZZ is variations that do not result in serotype change. However, in terms of binding motif differences between HLA allotypes, both the serological and genetic method are limited. Although it is generally true that genetically similar HLA gene alleles will probably bind a similar peptide profile, this is sometimes not accurate. Therefore, HLA supertypes were invented to group HLA gene alleles based on their binding motifs [Lund *et al.*, 2004; Sidney *et al.*, 2008]. For example, the HLA-B allotypes B*15:01, B*15:02 and B*32:01 belong to the supertype B62, while B*15:03, B*2705 and B*4802 belong to the supertype B*2705. This classification is achieved through similarities in predicted MHC Class I binders for the specific HLA allotypes (see Section 1.5.1 on page 33 for a brief overview of MHC ligand prediction).

Presentation to cytotoxic T-lymphocytes

Interaction between a presented peptide and a CTL is determined by both the peptide and the T-cell receptor of the CTL [Iversen *et al.*, 2006]. An extensive set of T-cell receptors exists and comes about by rearrangement of variable portions forming the α and β chains (for peptide-recognizing T-cells) ensuring this diversity [Davis and Bjorkman, 1988]. There exists a level of degeneracy in terms of peptide recognition. A TCR may recognize different peptides and different TCRs may recognize the same presented peptide [Cox *et al.*, 1994; Tallquist *et al.*, 1996]. Binding of a TCR with sufficient affinity to a presented peptide-MHC-complex result in the binding of CD8 to the MHC molecule to anchor the CTL and subsequent release of perforins and granyzymes by the CTL into the target cell [Russell and Ley, 2002]. The granyzymes are a class of serine proteases and cause a cascade of events to occur that trigger apoptosis. The ability of a presented peptide to elicit a CTL response is deemed its immunogenicity. Various factors influence the immunogenicity of a presented peptide. Commonly, the size of the repertoire the peptide entices is a good indication factor of immunogenicity. By interacting with a larger part of the repertoire, the chances of an appropriate CTL encountering the epitope is increased. The peptide-MHC complex does not stay on the cell membrane indefinitely and peptide-MHC disassociation does occur in a time-based fashion [Cerundolo *et al.*, 1991]. The composition of the peptide influences the stability of the complex.

1.2.2 The importance of HLA allotypes in HIV infection

The impact of HLA haplotype on the progression of HIV-1 to AIDS is striking. The existence of long-term non-progressor (LTNP) patients is evidence of this. Researchers have provided insight to the impact HLA alltoypes have on chronic HIV infection progression to AIDS. Examples of HLA associated with rapid onset of AIDS include HLA-B8 and a variant of HLA-B*3501, namely HLA-B*3501Px (hereafter, Px), which differs by a single amino acid from the normal type [Kaslow *et al.*, 1990; Gao *et al.*, 2001]. Primarily, it seems that the Px variant has a more flexible binding motif. It may seem paradoxical then that Px would have a negative influence on survival of HIV-1 infected individuals bearing the allele. This can be, in part, explained by the generation of the CTL TCR repertoire. During early development of CTLs, the immature CTLs are exposed to self-epitopes in the thymus. Briefly, CTLs are selected in a positive and negative fashion. Positive selection occurs when a CTL is moderately auto-reactive. If the CTL is too reactive towards a self-epitope, it is deleted from the repertoire. The promiscuity of Px may cause a reduction in the CTL TCR repertoire [Kosmrlj *et al.*, 2010]. A decrease in the size of the repertoire may limit the immune system to recognize variants of CTL epitopes and thus makes it easier for HIV-1 to escape the CTL response. Conversely, the HLA B58 supertype has a known protective function in HIV-1 infected patients [Migueles *et al.*, 2000; Altfeld *et al.*, 2003]. It is thought that its more restrictive binding motif allows for the existence of a more extensive TCR repertoire. However, the reason may not be as rudimentary as this. Many of the B*58 as well as B*27-restricted CTL epitopes occur in various regions across the HIV-1 genome that are highly conserved [Wang *et al.*, 2009]. Therefore, CTL escape is more difficult to attain due to the negative impact on fitness of these mutations.

Escape from cytotoxic T-lymphocyte responses

The efficiency of CTL responses can be demonstrated by the selection pressure it applies to viruses, especially HIV-1. The polymorphic nature of the HIV genome allows for quick accumulation of mutations that confer immunological escape [Allen *et al.*, 2005]. The HIV-infected cells that present immunogenic epitopes are deleted from the infected cell population over time. However, the total population of infected cells is not deleted. This gives rise to strains of HIV that harbour CTL escape mutations and are quickly positively selected in the infected host. Mechanisms conferring escape involve all parts of the antigen processing and presentation pathway as well as recognition by a CTL. They can be summarised as:

1. Processing by proteasome [Cardinaud *et al.*, 2011]:
   - Abrogation of an appropriate cleavage site forming the C-terminal end of a CTL epitope.
   - Creation of a cleavage site internal to the epitope, thus lowering the amount of epitope that is available for presentation on MHC.
2. TAP [Tenzer *et al.*, 2005, 2009]:
   - Substitutions near the terminal end of peptide fragments produced by the proteasome may hinder transport into the ER.
3. MHC/HLA affinity [Yokomaku *et al.*, 2004]:
   - Substitutions may reduce the affinity of the peptide to the MHC binding groove that either lower the amount of presented peptide or abrogate it completely.
   - Substitutions may reduce the stability of the peptide-MHC complex on the cell membrane.
4. TCR interaction [Iversen *et al.*, 2006]:
   - Substitutions may lower the affinity of the peptide to the TCR, thus not triggering CTL effector function (apoptosis induction).

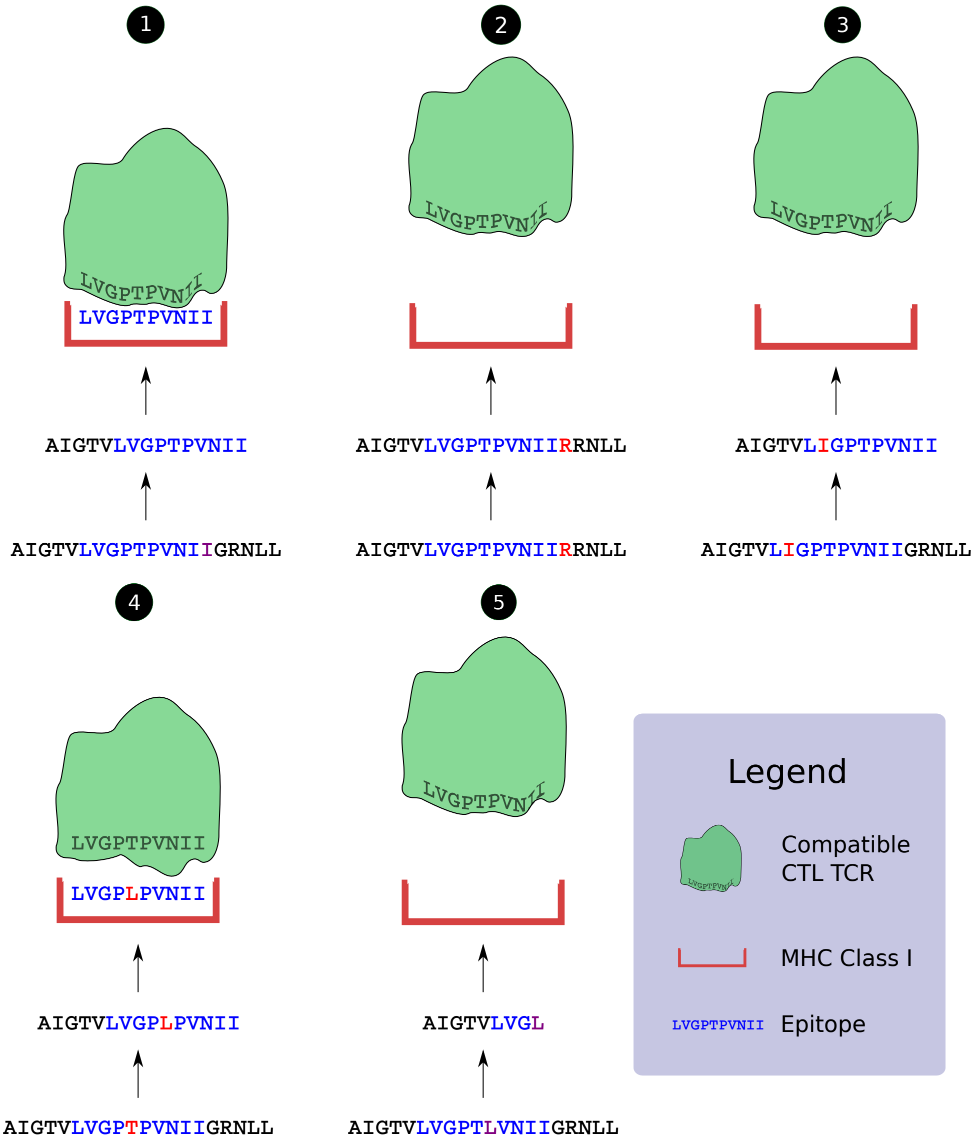


Figure 1.12: This figure illustrates the mechanisms of CTL escape. 1) The epitope here, LVGPTPVNII is extracted from the original protein via a cleavage site (marked with the purple I). The peptide is loaded onto MHC Class I and a compatible TCR recognizes the peptide, and the CTL effector functions commence. 2) The G to R substitution abrogates the normal cleavage site and the protein is not cleaved at the appropriate site. Although ERAP still trims the peptide, it no longer contains the epitope and the peptide is not loaded onto MHC. 3) The V to I substitution removes the anchor residue at N-2 necessary for binding to the MHC groove and the peptide never gets presented. 4) The T to L substitution does not affect processing and presentation of the peptide, but lowers the affinity of the TCR to it, thereby causing a drop in CTL activity. 5) The P to L substitution creates a proteasomal cleavage site internal to the epitope and ”destroys” the epitope or severely lowers amount available for presentation. Source: adapted from the literature and generated by the author [Iversen *et al.*, 2006; Tenzer *et al.*, 2009; Yokomaku *et al.*, 2004; Cardinaud *et al.*, 2011].

1.2.3 Humoral immunity

Humoral immunity is an adaptive response with effector cells that mainly include B-cells and effector proteins known as antibodies. Antibodies have the task of binding to particular proteins, e.g. surface accessible viral proteins or bacterial toxins and mark them for opsonisation (or absorption) by macrophages. Here, the process will only be discussed briefly. Naïve B-cells have membrane bound antibodies that can bind to a particular site in an antigen. This site is known as an epitope. Although a single B-cell has a single membrane bound antibody variant, the diversity of the membrane bound antibodies is diverse. This diversity arises from the random recombination of three gene gene segments, Variable, Joining and Diverse (VDJ) to form a complete gene during the early stages of B-cell development. The combination of segments determines the epitope for which the antibody would have affinity for. Due to the randomness of this recombination, a very wide range of epitopes can be recognized [Tonegawa, 1983]. When an antigen with an appropriate epitope is encountered by a B-cell, the protein is internalized and processed intracellularly. This leads to partial activation of the B-cell. The resultant fragments are then presented on MHC Class II to be scrutinized by a Th cell. A complimentary Th T-cell Receptor (TCR) will cause a cascade of events leading to production of cytokines that activate the B-cell. The B-cell then undergoes proliferation and eventually differentiates into memory B-cells that can respond to subsequent exposures to the same antigen and plasma cells, which produce antibodies that target the same antigen. Figure 1.13 on the following page depicts this process.

1.2.4 The importance of T-helper lymphocytes in immune function

In the previous section, it was demonstrated how the adaptive immune system effectively deals with pathogens. However, because this response is so effective and because the host organism is always at risk of an autoimmune response, various activation steps are required for the effector cells, i.e. B-cell and CTLs to function [Mizel, 1982]. It is the task

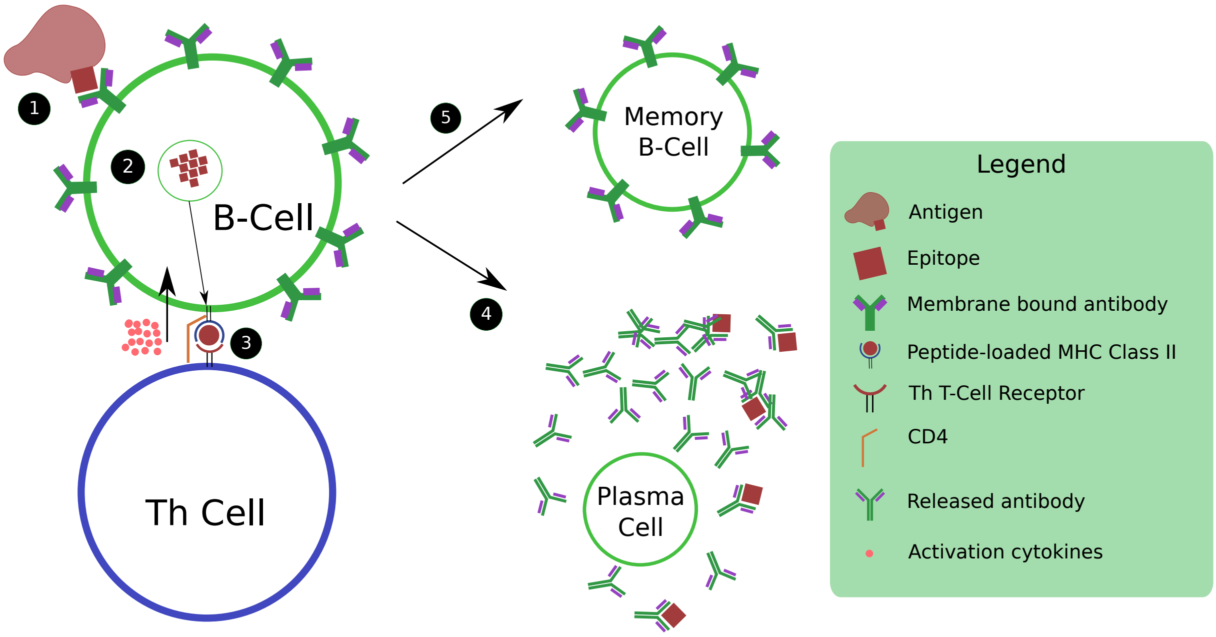


Figure 1.13: This figure illustrates the activation of a B-cell and subsequent antibody production. 1) A B-cell encounters and antigen with an epitope complementary to its membrane-bound antibody. 2) The protein is internalized and digested 3) The resultant fragments are presented on MHC Class II and a Th cell with a complimentary TCR to the presented peptide-MHC complex will cause CD4 to anchor it to the B-cell and activation cytokines are produced by the Th cell. 4) The B-cell then differentiates into memory B-cells and 5) antibody-producing plasma cells that bind any remaining antigen. Source: adapted from the literature and generated by the author [Amanna *et al.*, 2007].

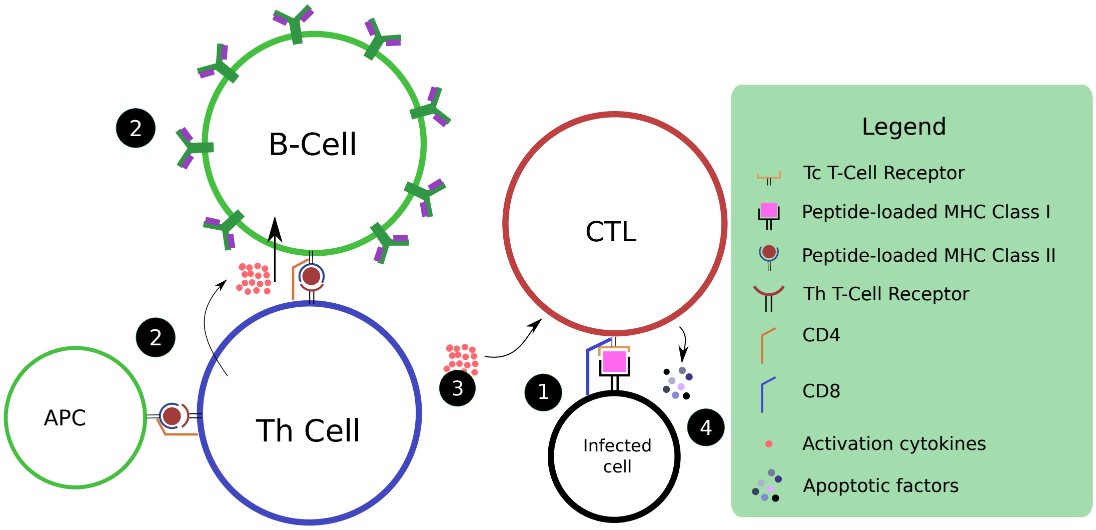


Figure 1.14: This figure illustrates the importance of Th in the activation of B-cells and CTLs. 1) The CTL is activated by an infected cell and awaits further signalling from a Th cell. 2) The Th cell is activated by recognizing peptide fragments presented by professional antigen presenting cells (APC), which can include B-cells. 3) The Th cell releases cytokines that secures the activation of the second signal in the CTL. 4) The CTL proliferates and kills cells presenting the appropriate peptide on MHC Class I. Source: adapted from the literature and generated by the author [Chan and Kim, 1998; Amanna *et al.*, 2007].

of the T-helper cells to regulate immune responses. T-helper cells are similar to CTLs in that they also recognize MHC-peptide complexes by a T-cell receptor. The difference is that the presented peptides are of exogenous origin and the presenting molecule is MHC Class II [Weaver *et al.*, 1988]. The MHC Class II molecule also preferentially binds a nonamer peptide, but the groove has an open cleft, meaning that a nonamer region in a peptide longer than 9 amino acids can be bound. Peptides are usually presented by professional antigen presenting cells (APC), namely dendritic cells, macrophages and B-cells in the lymph nodes. If the TCR of the Th binds with sufficient affinity, CD4 anchors on the MHC Class II molecule. The process of Th activation is illustrated in Figure 1.14. Taking CTLs as an example, the CTL is initially stimulated by a peptide-MHC Class I complex. A second signal in the form of the cytokine interleukin-2 and binding of the co-stimulatory receptor CD28 is needed to fully activate the CTL. If a proximal Th is activated, it releases IL-2 and binding of CD28 fully activates the CTL. Failure of this second signal can cause the CTL to become anergic. Anergy is the state of an effector cell being inactivated and no longer able to respond to a presented epitope. This is a safeguard to prevent CTLs from attacking otherwise normally functioning cells in the body [Schwartz, 1989]. A sub-population of activated T-helper cells retain memory of the event and subsequent exposures to the same epitope results in a far more rapid response, since the second ”verification signal” is not needed [Siekevitz *et al.*, 1987].

1.2.5 The impact of HIV-1 on T-helper function

From the previous section it is clear that a lower-functioning T-helper cell population can severely impair the adaptive immune responses. Although typically severe loss of immune function is attributed to full-blown AIDS, the latent phase of HIV infection can cause a progressive loss or impairment of immunological memory by the destruction of memory T-helper cells [Clerici *et al.*, 1989; Meyaard *et al.*, 1994]. Thus, subsequent exposure to a previously encountered antigen will result in a required reactivation slowing the response. Since the total T-helper lymphocyte population is lower, this step may take longer. Paradoxically, for the production of HIV particles to occur, the infected T-helper cell needs to be in an activated state. This further enhances the anti-immunity response of HIV, by infecting T-helper cells primed to fight infection. Ultimately, at the stages where AIDS manifests itself, the body can no longer maintain a high enough level of T-helper lymphocyte to mount an adequate immune response against pathogens that are vulnerable to the adaptive immune system. Because immunological responses are prone to be evaded by HIV, due to acquisition of escape mutations by the hyper-polymorphic HIV genome, other strategies are needed to either control HIV infection or delay the onset of AIDS.

1.3 Antiretroviral therapies

Over the course of HIV infection, immune responses become less efficient with the accumulation of mutations that confer immune escape. Research over the last few decades have resulted in a multitude of antiretrovirals to help manage HIV infection. Antiretrovirals have been regarded as essential in long term survival of HIV-1 infected individuals.

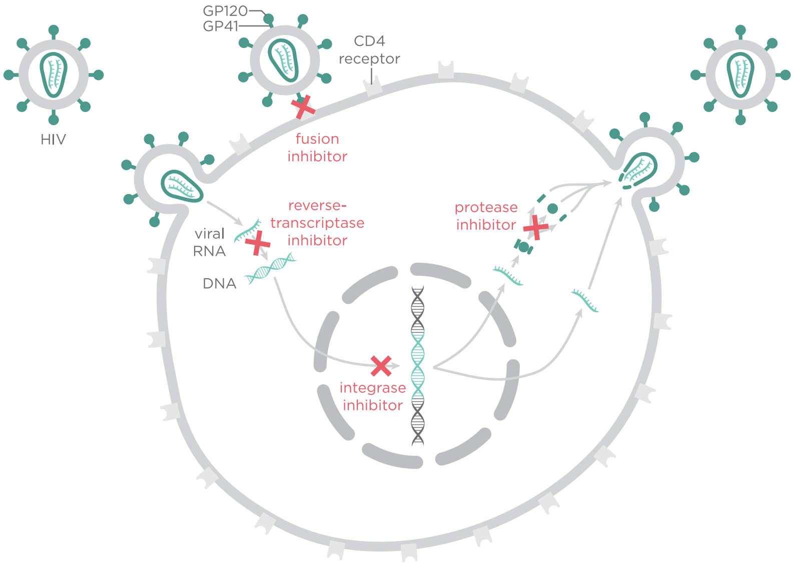


Figure 1.15: The figure illustrates the various targets for antiretrovirals in the HIV-1 replication cycle. Fusion inhibitors prevent entry of the virus into the cell. Reverse transcriptase inhibitors prevent RNA to cDNA transcription. Integrase inhibitors prevent the integration of the cDNA into the cell genome. Protease inhibitors prevent the maturation of the HIV-1 virion. Source: Illustration by Thomas Splettstoesser (`http://upload.wikimedia.org/wikipedia/commons/d/d5/HIV-drug-classes.svg`).

Furthermore, treatment can act as a prophylaxis shortly after exposure to HIV-1. Referring to Figure 1.15, inhibitors are targeted towards steps in the HIV replication cycle. In summary, they are:

1. Entry inhibitors
   - Binds to the CCR5 co-receptor and prevents attachments of the virus to CD4.
2. Fusion inhibitors
   - Inhibits fusion and internalization of virus.
   - Prevents infection of the cell.
3. Reverse transcriptase inhibitors:
   - Nucleoside and non-nucleoside analogs that interfere with the reverse transcription of the HIV RNA genome upon infection.
   - Prevents generation of DNA transcribed copy of HIV.

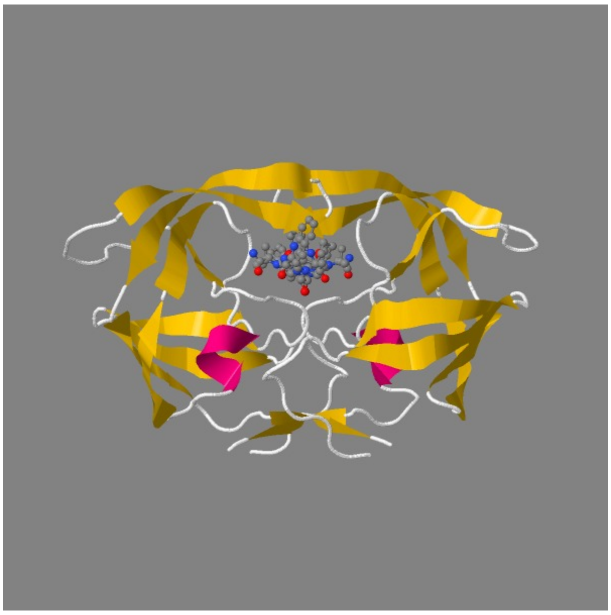


Figure 1.16: This figure illustrates the dimeric structure of HIV (each subunit can be marked by the pink alpha-helices). Saquinavir is displayed as a space-filler model in the centre of the structure, i.e. the active site. Binding of saquinavir inhibits HIV protease (PDB: 1HXB).

4. Integrase inhibitors

   - Interferes with HIV integrase and disrupts incorporation of the DNA transcribed HIV genome into the genome of the cell.

5. Protease inhibitors

   - Interferes with maturation of the immature HIV virion by inhibiting HIV protease.
   - Prevents or reduces the release of new viral particles.

In this section, specific attention will be given to reverse transcriptase inhibitors (RTI) and protease inhibitors (PI). These two classes of drugs are often used in antiretroviral therapy [De Clercq, 2009].

1.3.1 Mechanism of action

Targets of the RTI and PI antiretrovirals are sites on the enzymes RT and PR respectively that disrupt their mechanism of action. PR is a dimeric protein of two identical subunits arranged in a mirrored fashion. By using an analogue of an intermediate product of PR enzymatic activity, the antiretroviral competes with normal substrate for binding to PR [Mammano *et al.*, 1998; Vidal *et al.*, 2011]. This lowers the throughput of product produced by PR and severely cripples the maturation step of the virion. Figure 1.16 on the previous page illustrates HIV protease complexed with the first approved protease inhibitor, Saquinavir (SQV).

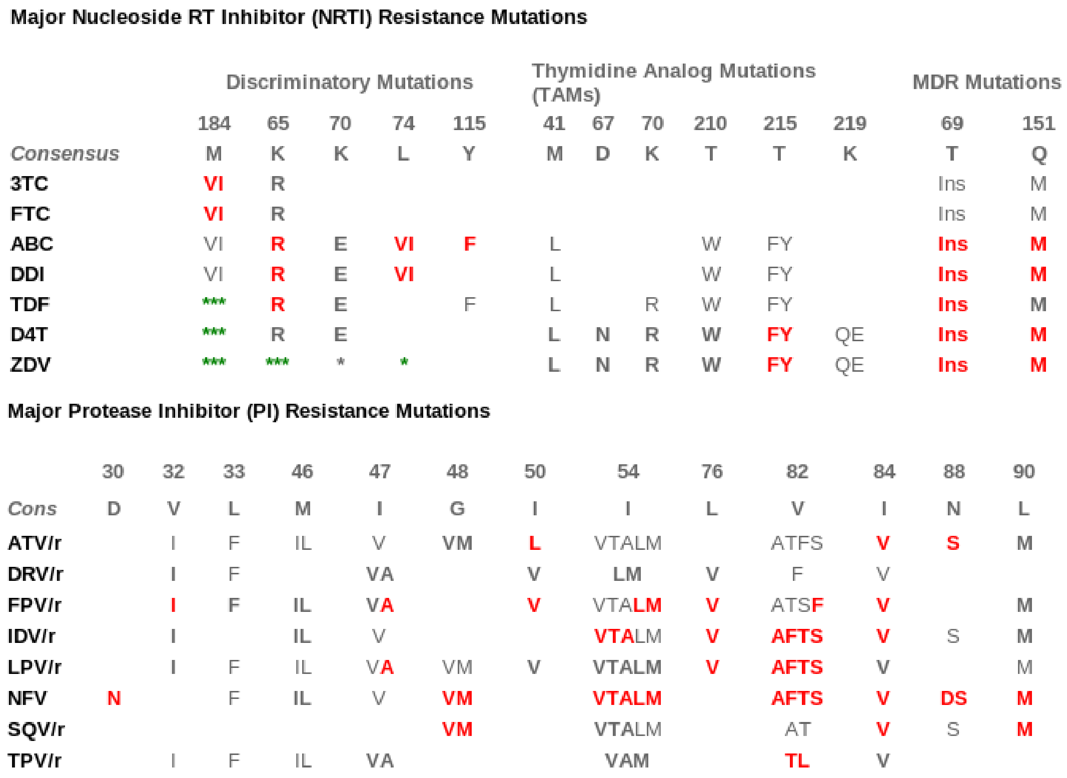
Reverse transcriptase inhibitors come in two classes, nucleoside inhibitors and non-nucleoside inhibitors. The nucleoside inhibitors like Ziduvodine (AZT, ZDV), provide a terminating nucleoside substrate to RT. Incorporation of this nucleoside analog terminates the reverse transcription process. Non-nucleoside inhibitors target a site in RT, which causes a structural change and subsequent loss or diminished RT activity.

1.3.2 Acquisition of resistance

As is the case of acquiring CTL escape mutations due to the hyper-polymorphic nature of the HIV-1 genome, so do resistance mutations to antiretrovirals accumulate [Ammaranond and Sanguansittianan, 2012]. Common mutations for various nucleoside RTIs and PIs are listed in Table 1.2 on the following page. Resistance mutations provide varying degrees of resistance towards ARVs by lowering affinity of the drug to its target site. Although there are sometimes shared resistance mutations between different drugs, this is not always the case. The mutation G48V in PR provides resistance to the PIs nelfinavir (NFV) and saquinavir (SQV), but provides no resistance to tapinavir (TPV), duranavir (DRV), fosamprenavir (FPV) and indinavir (IDV) [Shafer, 2006b; Rhee *et al.*, 2003]. Antiretroviral resistance mutations may have a negative impact on overall fitness of HIV and thus compensatory mutations are often found to emerge that mitigate the fitness impact. Alternatively, other mutations may occur in conjunction with the primary resistance mutation to further reduce the affinity of the ARV to its target. An example of this would be L33F in PR for increased resistance to all but IDV and SQV [Baxter *et al.*, 2006].

Mutations in RT can also include an insertion that provides resistance to most of the commonly used RTIs. Interestingly, mutations that confer resistance to some RTIs,

Table 1.2: The two tables list common mutations providing resistance to antiretrovirals. Resistance mutations in HIV RT (a) and HIV PR (b) conferring resistance to drugs are listed in the left column, and abbreviated as three letters. The red bold mutations are mutations conferring the highest level resistance to the listed drug. "Ins" is an insertion mutation. The bold letters represent mutations that compliment the red mutations and provide high level of resistance. Non-bold letters represent amino acids that augment the resistance of other mutations. The green asterisks in (a) indicate that mutations in that position increase the susceptibility of HIV RT to the respective drugs. Source: The HIVdb hosted at `http://hivdb.stanford.edu` [Rhee *et al.*, 2003; Shafer, 2006a].



Major Nucleoside RT Inhibitor (NRTI) Resistance Mutations

	Discriminatory Mutations					Thymidine Analog Mutations (TAMs)						MDR Mutations	
	184	65	70	74	115	41	67	70	210	215	219	69	151
Consensus	M	K	K	L	Y	M	D	K	T	T	K	T	Q
3TC	**VI**	**R**										Ins	M
FTC	**VI**	**R**										Ins	M
ABC	VI	**R**	**E**	**VI**	**F**	L			W	FY		**Ins**	**M**
DDI	VI	**R**	**E**	**VI**		L			W	FY		**Ins**	**M**
TDF	***	**R**	**E**		F	L		R	W	FY		**Ins**	**M**
D4T	***	**R**	**E**			**L**	**N**	**R**	**W**	**FY**	QE	**Ins**	**M**
ZDV	***	***	*	*		**L**	**N**	**R**	**W**	**FY**	QE	**Ins**	**M**

a.

Major Protease Inhibitor (PI) Resistance Mutations

	30	32	33	46	47	48	50	54	76	82	84	88	90
Cons	D	V	L	M	I	G	I	I	L	V	I	N	L
ATV/r		I	F	IL	V	**VM**	**L**	VTALM		ATFS	**V**	**S**	**M**
DRV/r		I	F		**VA**		**V**	**LM**	**V**	F	V		
FPV/r		**I**	**F**	**IL**	**VA**		**V**	VTA**LM**	**V**	ATS**F**	**V**		**M**
IDV/r		**I**		**IL**	V			**VTA**LM	**V**	**AFTS**	**V**	S	**M**
LPV/r		I	F	IL	V**A**	VM	V	**VTALM**	**V**	**AFTS**	V		M
NFV	**N**		F	**IL**	V	**VM**		**VTALM**		**AFTS**	**V**	**DS**	**M**
SQV/r						**VM**		VTALM		AT	**V**	S	**M**
TPV/r		I	F	IL	**VA**			**VAM**		**TL**	V		

b.

increase the susceptibility of RT to other RTIs. The major resistance mutations to lamivudine (3TC) and emtricitabine (FTC), increase susceptibility of RT to tenofovir (TDF), stavudine (D4T) and AZT [Shafer, 2006b; Rhee *et al.*, 2003].

1.3.3 Highly active antiretroviral therapy (HAART)

Currently, the use of a combination of ARVs is recommended in the treatment of HIV infection [Clifford *et al.*, 1999]. By targeting multiple steps in the HIV replication cycle at the same time, viral blood titres can be reduced to sometimes undetectable levels, i.e less than 50 copies of viral RNA/ml blood. This treatment is referred to as Highly Active Antiretroviral Therapy (HAART). The treatment needs to be continued for the entire duration of the patient's life, due to the drugs being unable to target latent virus, i.e. those that have integrated into infected cell's genomes, but are not expressed. Currently approved combination therapies include a combination of NRTI, NNRTIs and a PI and sometimes an integrase inhibitor. The use of combination therapies puts a huge selection pressure on the HIV genome and slows the acquisition of resistance mutations. Still, resistance may develop and drug substitutions are necessary that are not, or less, affected by already acquired resistance.

1.4 Vaccines for HIV

Vaccination against HIV is a very active research field. Many vaccine candidates have been developed with varying efficacy. Yet, despite the amount of research input, an effective vaccine has yet to be developed. Vaccine targets include a combination of humoral and cell-mediated immunity. The inhibition of HIV entry into target cells can be achieved by eliciting antibodies that bind to the CD4+ ligand, *gp120* [Burton *et al.*, 2004]. Through priming the immune system for recognition of CTL epitopes generated from HIV proteins, infected cells can also theoretically be removed from the immune system and thus curb infection. However, due to the highly variable nature of the HIV-1 genome, especially the envelope proteins, gp120 and gp41, evasion of immune responses is often quickly acquired and thus render the vaccines largely ineffective [Wei *et al.*, 2003; Garber *et al.*, 2004]. The antibodies generally target the variable portions of the envelope proteins of HIV. It is possible for the host to elicit new antibodies for these mutated regions, but eventually through persistence and progressive destruction of the T-helper cell population, the virus wins this arms race. Structural and functional proteins of HIV containing CTL epitopes represent attractive targets, since these proteins can be more conserved and have to cross bigger evolutionary hurdles for some escape mutations to accumulate [Troyer *et al.*, 2009]. The biggest need for a vaccine is to clear HIV before it gets sequestered in cells that would act as reservoirs. It has been shown that even at very low viral levels, mutations conferring immune escape as well as mutations providing antiretroviral resistance can accumulate [Bangsberg *et al.*, 2003].

1.5 Pharmacological and immunological interaction

It has been demonstrated that resistance mutations and immunity escape mutations are acquired during the course of HIV infection. Due to the fact that drug targets of HIV also harbour CTL epitopes, it stands to reason that ARV resistance mutations may sometimes overlap with CTL epitopes. The positions in PR commonly associated with ARV resistance is overlaid with CTL epitopes in Figure 1.17 on the next page. The effect of an ARV mutation within a CTL epitope region can be variable. The RTI resistance mutation M184V enhances the immunogenicity of A2-restricted epitope spanning RT 181-189 [Karlsson *et al.*, 2007]. The common PI resistance mutation, L90M, has a profound effect on the immune hierarchy in patients with an HLA of the A2 supertype. Ordinarily, the Gag epitope SLYNTVATL is the immunodominant epitope, meaning that most CTL efforts are directed towards it. The epitope spanning PR 76-84 becomes dominant when L90M occurs. Karlsson *et. al.* proposed that L9OM, which was predicted by NetChop to produce a novel proteasomal cleavage site at PR 89, affects the processing of the epitope.

The B62 supertype-restricted epitope in PR 44-52 does not ordinarily contain CTL escape mutations in drug naïve HIV infected individuals. However, under drug therapy,

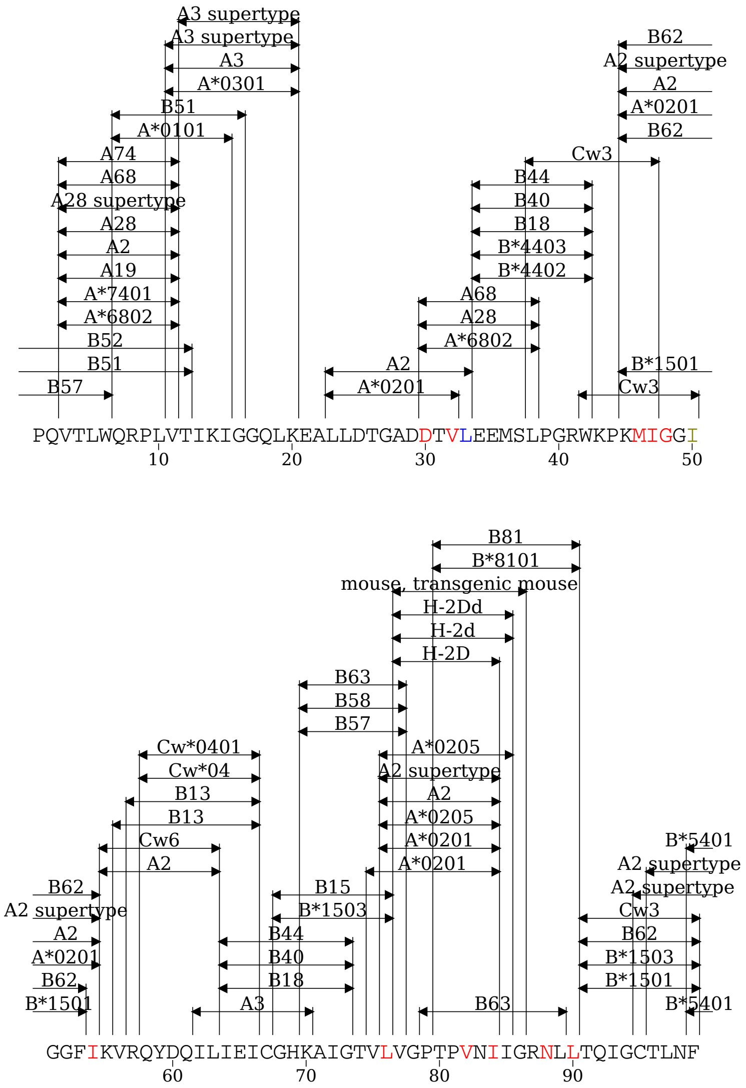


Figure 1.17: This figure shows a map of CTL epitopes (both confirmed and putative) along HIV PR. Consensus residues marked red indicate positions that are of interest in the acquisition of major HIV drug resistance mutations. It is interesting to note that the mutations and CTL epitope regions overlap. Indeed, mutations at positions marked in blue are confirmed to interact with CTL immunity [Mason *et al.*, 2004]. Figure source: HIV Immunology DB hosted at `http://www.hiv.lanl.gov/content/immunology`.

this region is prone to contain ARV resistance mutations. Some ARV resistance mutations attenuate the response to this epitope [Mueller *et al.*, 2011]. The reason for maintaining ARV mutations even in the event that it enhances CTL epitopes can be explained by the higher selective pressure applied by the ARVs on the HIV genome than the CTL epitope alone. Efforts to understand this interaction may further augment treatment strategies for individuals whose HLA genotype is known. This section covers possible strategies to detect CTL epitope changes due to ARV resistance mutations. A summary of ARV resistance mutations and CTL epitope overlap is shown in Figure 1.17 on the preceding page.

1.5.1 Prediction of potential CTL epitopes

Experimental determination of CTL epitopes is a laborious task involving many man hours and funding. This is further complicated by the extreme polymorphic nature of HLA allotypes. Taken together with the large array of HIV mutations known, pursuit of CTL epitope discovery has been augmented with non-experimental approaches. With the advent of computational approximation methods, it has become possible to predict potential CTL epitopes [Tenzer *et al.*, 2005; Korber *et al.*, 2006]. The proteasomal cleavage, TAP affinity, MHC affinity and MHC peptide stability tools have been developed to varying degrees of accuracy [Parker *et al.*, 1994; Karosiene *et al.*, 2012; Nielsen *et al.*, 2007; Lin *et al.*, 2008]. The accuracies of the tools are very dependent on data availability [Lundegaard *et al.*, 2010].

Proteasomal cleavage prediction is marred by the lack of available proteasomal cleavage data [Nielsen *et al.*, 2005; Ginodi *et al.*, 2008]. The tools demonstrate a modest ability to predict cleavage sites. Thus, the use of these tools in predicting a CTL epitopes is fairly limited.

For MHC prediction, resources such as the Immune Epitope Database (`http://www.iedb.org`,) exist providing MHC binding affinity data of many different peptides [Vita *et al.*, 2010]. Because the binding of a peptide to the MHC groove is motif driven, given enough data for a HLA allotype, predictions can become fairly accurate. Recently, methods have been developed where prediction of peptide affinity to a HLA allotype with little data availability is made possible by a form of extrapolation from prediction results of other HLA allotypes. The NetMHCPan tool is specifically designed to undertake this task [Nielsen *et al.*, 2007]. Prediction of MHC affinity, although not perfectly accurate, can give insight to the generation of and attenuation of a CTL epitope.

Another important factor that must be considered beyond MHC affinity, is peptide-MHC complex stability. The longer a peptide-MHC complex is stable, the higher the chance is of encountering a reactive CTL. To this end, NetMHCStab is a tool developed recently to predict peptide-MHC stability for a limted number of HLA allotypes [Jørgensen *et al.*, 2014]. Affinity combined with stability can be a good indicator of whether a peptide is a CTL epitope.

To the author's knowledge, there exists only one tool for predicting immunogenicity of a peptide, namely POPISK. The limitation of this tool is that it only predicts the immunogenicity of epitopes restricted to the HLA A2 supertype and limits the discovery of CTL epitopes. Very recently, another method has been developed that aims to measure the favourability of residues for immunogenicity in a peptide sequence [Calis *et al.*, 2013]. The researchers highlighted the importance of especially aromatic residues in position C-4 to C-6 of the epitope as contributors to immunogenicity.

Construction of MHC binding predictors

Most prediction tools rely on either a simple mathematical model or use of machine learning techniques to predict peptide-MHC binding affinity data [Parker *et al.*, 1994; Tenzer *et al.*, 2005; Nielsen *et al.*, 2007]. The power of machine learning techniques lies in their ability to abstract input information, e.g. an amino acid sequence information, to an expected result [Nielsen *et al.*, 2004; Lundegaard *et al.*, 2008]. Affinity of a peptide to MHC is usually given as an IC50 value. The IC50 value is the concentration in nM needed for the peptide whose affinity is being measured, to displace 50% of a standard peptide bound to an MHC molecule in solution. The lower the amount of peptide needed to displace 50% of the standard peptide, the higher the affinity of the peptide for the MHC molecule. It is an accurate way of determining affinity. Where quantitative data is not available, discreet values are assigned to a peptide designating whether it binds to MHC. Next, the sequence of the measured peptide needs to be encoded to a numerical value (usually discrete). Given that a large enough set of peptides is available, typically a representative set of 300, the predictor can be trained to abstract the input peptide to the output value. For instance, part of NetMHCPan prediction uses sparse encoding of a peptide. The standard amino acids are represented by vector of length 20. For each amino acid, a value of 1 is assigned to one of the 20 values, with the rest remaining 0. Each amino acid has a unique vector associated with it. Alternatively, where there lacks data for a certain position in the peptide sequences, BLOSUM50 encoding may be used. Each amino acid is assigned a vector containing BLOSUM50 values of the amino acid to all 20 amino acids, including itself [Nielsen *et al.*, 2003]. This allows amino acids to be approximated to other amino acids in a quantitative fashion. This is very useful if the peptide set used to train the predictor has little information for amino acids at certain positions. With sparse encoding, there is no way for the predictor to know how to predict the contribution of an amino acid at a specific position if it was not observed in the input data, therefore the BLOSUM encoding allows the predictor to extrapolate from other amino acids. Higher accuracy is obtained by training sparse and BLOSUM-encoded predictors separately and combining the results as a weighted average [Nielsen *et al.*, 2003; Lundegaard *et al.*, 2008].

The encodings are demonstrated in Table 1.3 on page 38. The encoded amino acids represented the input vectors of the artificial neural network (ANN). The ANN consists of three layers. An input layer that contain nodes with the encoded amino acid vectors. A second layer (hidden layer) contains two nodes, each representing a sigmoidal function that uses the inputs of the first layer as coefficients (weights). Finally there is an output layer, usually a simple linear function that uses the output weights of the hidden layer as weights. The output layer produces the IC50 values. Training of the ANN means optimizing the weights between the input and hidden as well as the hidden and output nodes to minimize the error between the predicted and actual values.

Over-fitting, meaning predictions of the network are biased towards the training set and not new data, occurs when too many ANN training cycles are performed. Too little training cycles, and the network tends to under-fit the data. This is avoided by performing a cross-validation on the ANN. For NetMHCPan, the training data is split into five sets. Five neural networks are trained, with each removing a fifth of the data in order to use this as validation of the ANN performance. From this data, the rate and amount of cycles are optimized in such a way that the difference in error in the training and testing sets are minimized. The weights of the final network are an average of the weights for the five ANN constructed during cross-validation [Nielsen *et al.*, 2003]. A separate set of data not used in any of the training is finally used to measure the performance of the predictor.

An example of the error rates of an ANN trained on arbitrary data is shown in Figure 1.18 on the following page. This figure illustrates the error rate of the ANN versus training cycles for the training set and validation set. At each training cycle, the errors for the training and testing set are calculated. It is clear that at a certain point, around 380 training cycles, the error in the testing set is minimized. Further training reduces the error rate in the training set, while increasing the error rate in the testing set. It would therefore not be advisable to train the ANN beyond this point. The performance of the ANN is shown in Figure 1.19 on the next page. The figure shows the false positive rate versus the sensitivity of the predictor, i.e. the error of the positive predictions versus the coverage of the true positives. The two curves represent an ANN trained with 380 cycles (red) and 2000 cycles (blue). It is clear that for lower sensitivity levels the ANNs have very similar performance, while the 380 cycle ANN outperforms the 2000 cycle ANN at higher sensitivity values.

1.5.2 Detection of potential HIV CTL epitopes by measuring HLA associated mutations

Another bioinformatic approach to detect potential CTL epitopes is by analysing mutation frequencies in sets of HLA annotated sequences [Brumme *et al.*, 2009]. If enough sequence data is available, methods such as Fisher's exact test can be used to detect significant

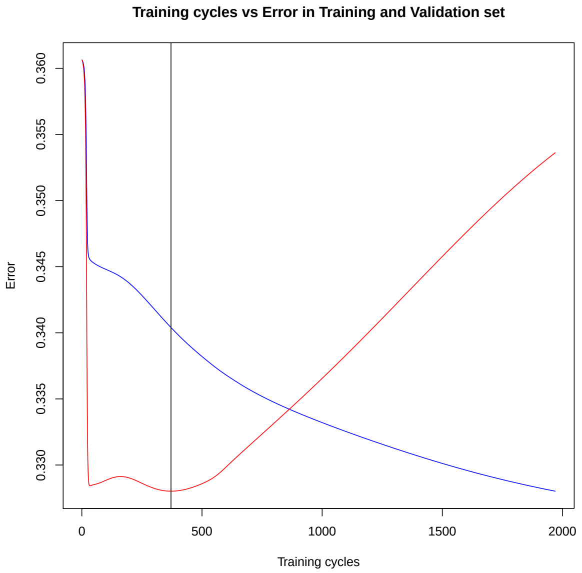


Figure 1.18: This figure depicts the error rates of neural network that is trained over many cycles on arbitrary data. The blue line represents the error over the training data while the red line represents the error over the validation data. The vertical black line indicates the point where the error in the validation data is minimized. Although training beyond this point further reduces the error in the training data, the error increases in the testing data. Therefore, the artificial neural network should not be trained beyond approximately 380 cycles.

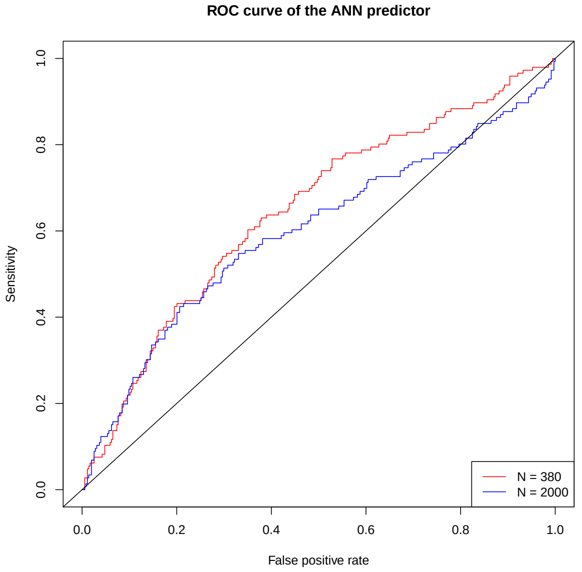


Figure 1.19: This figure depicts the Receiver Operating Characteristic (ROC) curve for the ANN predictor trained with 380 cycles (red curve) and 2000 cycles (blue curve). While initially, the two predictors seem to have an increase in sensitivity and false positive rate, divergence between the two predictors is observed at a false positive rate of approximately 0.35. The ANN network trained with 380 cycles appears to have a higher sensitivity to false positive rate ratio.

Table 1.3: This table illustrates the process of encoding the peptide, `SLYNTVATL` to sparse and BLOSUM50 encoding. For sparse encoding, each amino acid position is a vector of length 20. A value "1" is assigned to the position representing the corresponding amino acid. For instance, the first amino acid, serine, is encoded as a vector, $V1_{sparse} = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0)$. For sparse encoding, a vector of length 20 is assigned 20 values corresponding to the values of the amino acid in question compared with all amino acids in the standard amino acid table, in this case, serine is encoded to $V1_{blosum} = (1, -1, 0, -1, -3, 0, -1, -3, 0, -3, -2, 1, -1, 0, -1, 5, 2, -2, -4, -2)$. Each of the values represent an input node for the artificial neural network [Nielsen *et al.*, 2004].

Sparse encoding

AA	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0	0
L	0	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	0	0
Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**
N	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0
T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0
V	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0
A	**1**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	0	0
L	0	0	0	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	0	0

BLOSUM 50 encoding

AA	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
S	1	-1	0	-1	-3	0	-1	-3	0	-3	-2	1	-1	0	-1	5	2	-2	-4	-2
L	-2	-2	-4	-3	1	-4	-3	2	-3	5	3	-4	-4	-2	-3	-3	-1	1	-2	-1
Y	-2	-3	-3	-2	4	-3	2	-1	-2	-1	0	-2	-3	-1	-1	-2	-2	-1	2	8
N	-1	-2	2	0	-4	0	1	-3	0	-4	-2	7	-2	0	-1	1	0	-3	-4	-2
T	0	-1	-1	-1	-2	-2	-2	-1	-1	-1	-1	0	-1	-1	-1	2	5	0	-3	-2
V	0	-1	-4	-3	-1	-4	-4	4	-3	1	1	-3	-3	-3	-3	-2	0	5	-3	-1
A	5	-1	-2	-1	-3	0	-2	-1	-1	-2	-1	-1	-1	-1	-2	1	0	0	-3	-2
T	0	-1	-1	-1	-2	-2	-2	-1	-1	-1	-1	0	-1	-1	-1	2	5	0	-3	-2
L	-2	-2	-4	-3	1	-4	-3	2	-3	5	3	-4	-4	-2	-3	-3	-1	1	-2	-1

frequency differences in amino acids. A good example is the HLA B62 supertype CTL escape mutation, PR I93L. This mutation is a polymorphism observed in HIV-1 Subtype B. However, it occurs from measured sequence data at a much higher rate in sequences obtained from patients with an HLA allotype belonging to the B62 supertype. Conversely, certain mutations that may enhance CTL activity occur at lower frequencies in certain individuals. It should be noted that these mutations are sometimes not directly involved in CTL epitope escape and may occur as auxiliary, compensatory mutations of primary CTL escape mutations that may have a negative fitness impact on HIV [Crawford *et al.*, 2007].

The correlation between substitutions in HIV-1 proteins and HLA is an exceptionally important discovery. The author notes that the HIV sequence data for HIV-1 protease and reverse transcriptase from the Los Alamos National Laboratory annotated as being obtained from HIV infected individuals undergoing ARV treatment, lacks HLA annotations. This annotation could be useful in identifying ARV resistance mutations that are negatively correlated with certain HLA allotypes. A negative correlation may indicate the presence of a novel CTL epitope and selection against an ARV resistance mutation may prolong the efficacy of antiretroviral treatment. A combination of substitution-HLA correlation together with tools that aid in the prediction of CTL epitopes could be useful in discovering novel or cryptic epitopes that are induced by antiretroviral resistance mutations.

1.6 Summary

The Human Immunodeficiency Virus (HIV) is the primary cause of Acquired Immunodeficiency Syndrome (AIDS). The pandemic has claimed the lives of millions of people. Through the destructive effect this virus has on the immune system, infected persons become susceptible to opportunistic infections. The majority of patients do not have natural immunity against the virus and given enough time, invariably develop AIDS. Current treatment options are hampered by the development of resistance by the highly mutable HIV genome. This applies to vaccines. As of date, no effective cure has been discovered. The interplay between antiretroviral resistance mutations and immune responses remain an active area of research. Computational tools to aid in the discovery of Cytotoxic-T-lymphocyte epitopes provide a new perspective on the interactions between drug resistance mutations and immunological responses.

2 Problem statement and Aims

2.1 Problem statement

With the interaction between ARV resistance mutations and CTL epitopes becoming more evident, it seems prudent to identify potential novel CTL epitopes due to ARV resistance. Lack of HLA annotation of patients could hamper the computational discovery of negative correlations between HLA type and substitutions associated with antiretroviral resistance. Insights into this interaction would be beneficial when considering treatment for HIV infected individuals.

Aims

1. Assign HLA types to patients from which HIV-1 *pol* sequences were obtained.
2. Using this assignment, detect significantly diminished substitutions that are commonly associated with antiretroviral resistance in relation to assigned HLA types.
3. Predict changes in MHC affinity and stability due to these mutations to provide a putative immunological explanation why these residues are diminished.
4. Detect potential escape mutations in these putative epitopes.

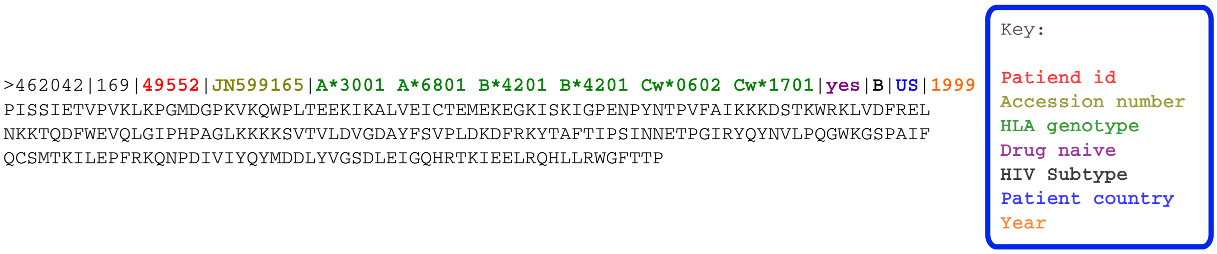
3

Methods

This chapter acts as a reference for the methods used to produce the results in Chapter 4. Although the descriptions of the methods are adequate in themselves, it may be prudent to back reference these methods as they are mentioned in Chapter 4. This chapter covers how HIV *pol* sequences were obtained and clustered according to patients. The assignment of HLA to unannotated patients is described as well as the measurement of substitution frequency discrepancies between assigned HLA types. Also included in this chapter is the predictors used that are related to the prediction of CTL epitopes. Lastly, methods for balancing and or accounting for correlated substitutions are described.

3.1 Acquisition and processing of HIV sequence data

HIV amino acid sequences were obtained from the LANL HIV database (`http://www.hiv.lanl.gov`). All HIV protease (PR) and reverse transcriptase (RT) sequences of subtype B were annotated with patient information including, where available, ARV exposure, HLA information and patient IDs. A total of 26,870 PR sequences from 10,049 patients and 25,342 RT sequences from 10,657 were used. Sequences from patients without patient ID annotations were discarded. Sequences that span the complete PR region and first 217 amino acids of RT (PR-RT) totalled 19,857 from 9,186 patients. For PR, 19% of patients were drug naïve, 25% were treatment experienced and the remainder were not



```
>462042|169|49552|JN599165|A*3001 A*6801 B*4201 B*4201 Cw*0602 Cw*1701|yes|B|US|1999
PISSIETVPVKLKPGMDGPKVKQWPLTEEKIKALVEICTEMEKEGKISKIGPENPYNTPVFAIKKKDSTKWRKLVDFREL
NKKTQDFWEVQLGIPHPAGLKKKKSVTVLDVGDAYFSVPLDKDFRKYTAFTIPSINNETPGIRYQYNVLPQGWKGSPAIF
QCSMTKILEPFRKQNPDIVIYQYMDDLYVGSDLEIGQHRTKIEELRQHLLRWGFTTP
```

```
Key:

Patiend id
Accession number
HLA genotype
Drug naive
HIV Subtype
Patient country
Year
```

Figure 3.1: This figure illustrates a typical entry in the FASTA file obtained from the Los Alamos National Laboratory HIV sequence data base. The annotations of interest are in bold. The colours correspond to the description of the field in the key. The first two annotations are the `sequence ID` and `patient code` fields that were not used. This information can then be used to group sequences according to subtype as well as `patient ID`.

annotated with treatment information. The RT and PR-RT sets showed similar frequencies, with 18% being drug naïve, and 24% of patients treatment experienced. Sequences were grouped according to the respective `patient id` field to which they belong. A total of 10,972 PR-RT sequences annotated as non-drug naïve were assigned to 2,532 patient IDs and 11,315 sequences annotated as drug naïve were assigned to 1,709 patient IDS. The sequences were obtained in September 2012. A simple description of the annotations in the FASTA headers are show in 3.1.

3.2 Imputation of HLA types

Previous researchers have provided data linking certain amino acid substitutions in HIV-1 proteins to CTL escape for certain HLA types [Brumme *et al.*, 2009]. Provided with this is a complete list of HLA-based substitution data. The data was used here to aid in HLA annotation of patient data. The HIV-1 sequences from patients were scanned for HLA-associated mutations and if found, an HLA type was assigned. At least one mutation is necessary for an HLA type to be assigned. There are some additional criteria that need to be mentioned:

- This study focuses on RT and PR proteins, thus only HIV-1 sequences containing both these proteins are used. This limits the amino acids used for HLA assignment to PR and RT.

- Mutations in PR are highly correlated and many mutations in PR associated with CTL escape also confer resistance to many protease inhibitors [John *et al.*, 2005; Mueller *et al.*, 2007; Rhee *et al.*, 2007].
- The majority of RT sequences have at least the first 217 amino acids, thus only substitutions within this region were used.
- Only HLA associated mutations with a q-value (false-discovery rate) of less than 0.20 as well as substitutions that do not occur too frequently (more than 10% of patients) were considered to reduce false-positives [Storey, 2002].

The assignment of HLA type B*15 to patient sequence sets will be demonstrated. Substitution data associated with HLA type B*15, restricted to PR and RT include the following substitutions:

- PR - `L9I A71V, Q92K, I93L`
- RT - `I135I Q174H Q207E Q207H Q207R R211Q R211G`

First, the PR mutations are stripped, leaving only the RT set, i.e. $HLA_{B*15} = (I135I, Q174H, Q207E, Q207H, Q207R, R211Q, R211G)_{RT}$. Next `Q207E` is removed because it is inherently increased in sequences from patients undergoing treatment. The I135I is also removed because it occurs in 51% of patients' RT sequences. The final set is thus $HLA_{B*15} = (I135, IQ174H, Q207H, Q207R, R211Q, R211G)_{RT}$. When using these amino acids as a filter, it produced an HLA_{B*15+} set of 137 patients and an HLA_{B*15-} set containing 2037 patients. Note that this partitioning is not absolute, it merely indicates that HLA_{B*15-} has a *higher* probability of containing HIV-1 sequences from HLA-B*15 patients.

3.2.1 Calcualtion of HLA imputation performance

To evaluate the performance of the HLA imputation procedure, the mutations were used to assign HLA types to other HIV *pol* data [Carlson *et al.*, 2012]. The evaluation performed included *sensitivity* and *positive predictive value* (PPV) measurements. The *sensitivity* measures the proportion HLA of a specific type discovered by the HLA associated mutations and the *PPV* was used to measure what proportion of the predicted HLA types were correctly assigned. The formulae for sensitivity and PPV measurements are given in equations 3.1 and 3.2 respectively. In the equations, TP represents ”true positives”, i.e. the number of sequences that were assigned to the correct HLA type. The ”FN” (false negatives) is the number of sequences that were from patients with the HLA type in question, but could not be detected using the HLA type associated mutations. The ”FP” (false positives) is the number of sequences falsely assigned as being from a patient with the HLA type in question. From these results, it can be estimated when assigning an HLA type, what proportion of the HLA types were detected as well as the accuracy of the positively assigned HLA types.

$$Sens = \frac{TP}{TP + FN} \tag{3.1}$$

$$Sens = \frac{TP}{TP + FP} \tag{3.2}$$

3.3 Using Fisher's exact test to calculate substitution discrepancies

The Fisher's exact test is a statistical method to measure the significance of the odds-ratio of a factor between two data sets. Here, the data sets are sets of sequences and the factor is the frequency of an amino acid substitution along an HIV-1 protein product. As an example, the frequency of the PR amino acid substitution, `G48V` (i.e. the guanine at

Table 3.1: The table shows an example of the Fisher's exact test used to measure the odds of the PR G48V mutation between drug naïve and treatment experienced patients. Each row shows the frequency of 48G and 48V (as shown in the columns) for the drug naïve and drug experienced group respectively. The row and column totals are also indicated. The significance value for this test can be calculated with Equation 3.3.

	48V	48G	Row total
Drug naïve	5 (a)	1631 (b)	1636
ARV treated	113 (c)	1825 (d)	1938
Column Total	118	3456	n=3574

position 48 substituted for a valine), between HIV PR sequences obtained from drug naïve individuals and those undergoing ARV treatment. The results of the frequency counts of `48G` and `48V` (calculated) are shown in Table 3.1.

These values can now be used to calculate a p value that indicate the probability of the predicted, or higher, odds ratio of G48V between treatment experienced and ARV naïve persons if the data were randomly sampled in each set. In this case, the p-value is calculated to be well below 0.001, indicating strongly that the probability of the observed odds-ratio of 19.000 due to a chance event is very low. It can thus be concluded that the PR G48V mutation is significantly associated with treatment.

$$p = \frac{(a+d)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!} \tag{3.3}$$

Equation 3.3 is used to calculate the p statistic for Fisher's exact test. The variables a, b, c and d represent the corresponding cells listed in Table 3.1 with n being the total size of the sample. This results in a very small value ($p << 10^{-16}$) and can thus be deemed significant. Indeed, this mutation is associated with resistance to the protease inhibitor,saquinavir (SQV). The substitutions are counted according to the method described in 3.3.1.

3.3.1 Substitution counting method

The aforementioned example does have a few prerequisites to the analysis. Because the amount of sequence data for each patient can be variable, the test can produce biases, e.g.

one patient with 50 sequences with the same mutation at a given position will artificially increase the substitution count. Therefore, binary counts were assigned to patient sequence mutation data so that a mutation was only counted once per patient. In addition to this, a mutation matching the consensus residue will only be counted if the patient had no other mutations at that point. This is to prevent substitutions being obscured by the consensus residue. Therefore the substitution frequencies equate to the number of patients that have HIV with the mutation in question. If more than one substitution occurred at a position in the sequence set of a patient, it was counted separately. For example, considering position 215 in RT. Common drug related substitutions are T215Y/F [Shafer, 2006a]. Taking a sample of 50 patients, with ten patients having HIV-1 RT sequences that containing only have the T215Y substitution (A), seven that only have the T215F (B) substitution and eleven that have either the T215Y or T215F substitution (C) in the HIV-1 RT sequences. This would mean that a total of $A + C = 10 + 11 = 22$ patients have the T215Y mutation while $B + C = 7 + 11 = 18$ patients have the T214F mutation.

3.3.2 Determining substitution frequency discrepancies in HLA assigned patient sequence sets

Partitioning the sequence sets based on HLA substitutions allowed for measurement of frequency discrepancies between sets imputed to be HLA X positive and those deemed to be HLA X negative, where X is any assigned HLA type. Using the HLA B*15 example earlier, at each position along PR-RT, the substitution counts are tallied and compared using a Fisher's exact test. An example showing the contingency table obtained by using the substitution counting method is show in Table 3.2. The cells in the previous example labelled as a, b, c and d are now S1, S1', S2 and S2'. The number in Sx indicate the group, i.e. "1" being the HLA B*15-positive group and "2" being the HLA B*15-negative group. The accents indicate the number of patients whose PR sequences did not contain the I93L mutation. The significance and odds-ratio of the frequencies between "1" and "2" are calculated with the `fisher.test` function in the R-programming language, yielding

Table 3.2: This contingency table shows the counts for the PR I93L mutation observed in HLA B*15-positive and HLA B*15-negative set. S1 and S1' are the number of patients in the HLA B*15 set that have an HIV-1 sequence containing the I93L mutation and the number of patients that do not. The same applies for S2 and S2', the sequence set assigned as HLA B*15-negative. The letters a, b, c and d correspond to the contingency table in Table 3.1.

	93L	**not 93L**
HLA B*15-positive	79 (S1, a)	59 (S1', b)
HLA B*15-negative	757 (S2, c)	1280 (S2', d)

a p-value of 4.85×10^{-6} and a log_2 odds-ratio of 1.18, indicating that I93L is significantly enriched in the B*15 assigned set. This procedure is repeated over all positions for all substitutions and the results are tallied.

3.4 Calculating epistatic interactions between substitutions within PR and RT

Co-occurrence of substitutions in PR and RT were calculated using the phi-correlation coefficient and a Fisher's exact test. Co-occurrence is defined as a substitution that differs from the consensus sequence at one position co-occurring with a substitution that differs from the consensus sequence at another position. Conversely, the absence of substitution co-occurrence is defined as the occurrence of one substitution at one position with the occurrence of the consensus residue at the other position. A table is generated (needed for both the Fisher's exact test and phi-correlation coefficient) as shown in Table 3.3. This table shows calculation of co-occurrence frequencies of the PR mutations 46I and 90M. Equation 3.3 is used to calculate the significance value of the Fisher's exact test, however this calculation was performed in the R-programming language (`http://www.r-project.org`) and thus odds-ratios were also calculated from the results. The phi-correlation coefficient was calculated using Equation 3.4. The equation produces a value ranging from -1 to 1 and measures the rate of correlation between different substitutions, 1 indicating perfect correlation, -1 indicating perfect negative correlation and 0 indicating no correlation.

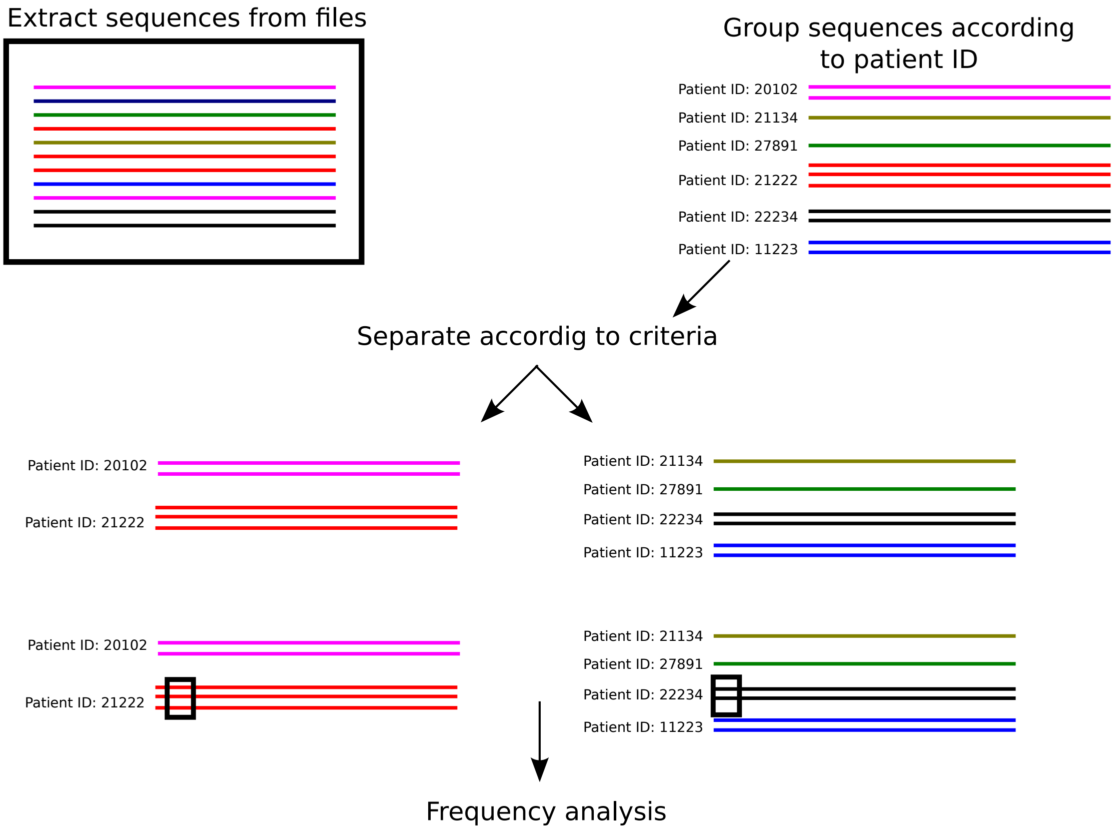


Figure 3.2: This figure demonstrates the process of sequence assignment to the patient ID in the FASTA header line. At first, sequences are divided based on being annotated as drug naïve or drug experienced. Subsequently, sequences are assigned to groups based on the patient ID field in the FASTA header. When subsequent criteria are used to further assign groups, it is done in a patient ID fashion, i.e. the patient ID sets from two or more categorized sequence groups are mutually exclusive. The black rectangle over the sequences indicate the procedure of amino acid frequency count. Therefore, the amino acid count of a position specific residue is assigned to the patient as a binary value.

$$\phi = \frac{ad - bc}{\sqrt{(a+c)(b+d)(a+b)(c+d)}} \tag{3.4}$$

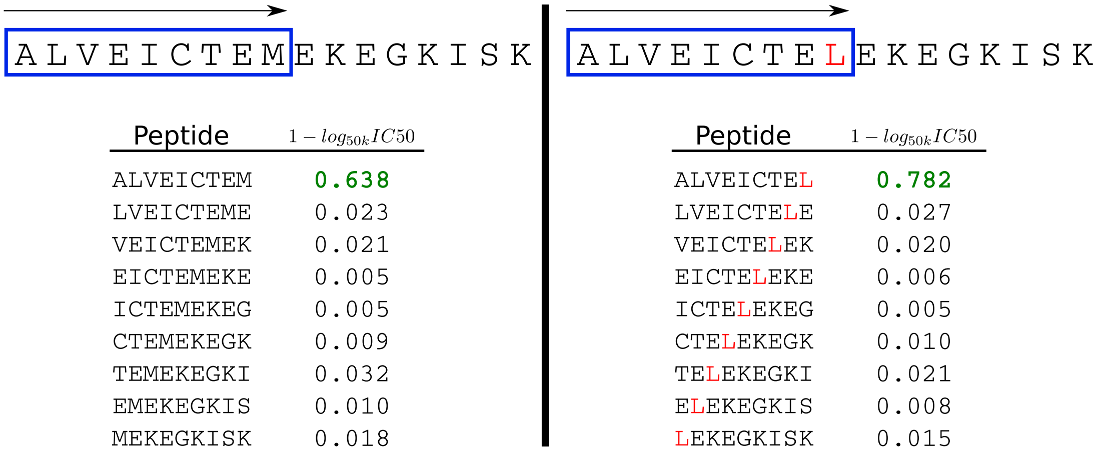
Table 3.3: This table show the values used to calculate the phi-coefficient The rows indicate the frequency of two residues, 46I and 46M and the columns indicate the number of times each residue, 90L and 90M co-occur with each of the 46I or 46M mutations. The values in the cells (a,b,c,d) are used as inputs in Equation 3.4.

	Substitution 90M	Consensus 90L	Row total
Substitution 46I	418 (a)	188 (b)	608
Consensus Pos 46M	436 (c)	1132 (d)	1568
Column Total	854	1250	n=2176

3.5 Prediction of peptide CTL epitope eligibility

3.5.1 Prediction of peptide-MHC affinity and stability

Prediction of MHC affinity and stability of peptides within the PR and RT regions of HIV-1 subtype B were calculated with NetMHCPan and NetMHCStab respectively [Nielsen *et al.*, 2007; Jørgensen *et al.*, 2014]. NetMHCPan allows the prediction of MHC affinity for a large variety of HLA allotypes. The power of this tool lies within its ability to make affinity predictions for HLA allotypes for which no or very little affinity data is available by extrapolating from closely related HLA allotypes for which sufficient data is available [Nielsen *et al.*, 2007]. Values predicted range between 0 and 1 and represent the $1 - log_{50000}IC50$ values for the input peptides, with higher values equating to higher affinity. Note that this is the inverse of the typical $IC50$ values where higher values indicate a *lower* affinity. NetMHCStab predictions are limited to fewer HLA allotypes. A key feature of this predictor is producing half-life time of peptides presented by respective HLA allotypes on the cell membrane [Jørgensen *et al.*, 2014]. A higher half-life indicates longer existence of the peptide-MHC complex on the cell membrane. The output score is produced by the Equation 3.5, although only the NetMHCStab (the actual stabilization value) is used. The output value of NetMHCStab is multiplied by a coefficient, α, which in turn is added to the MHC prediction result of NetMHCStab, another MHC affinity predictor that has greater performance per HLA allotype than NetMHCPan, but does not have the extrapolatory power for other HLA allotypes of NetMHCPan [Karosiene *et al.*, 2012].



Peptide	$1 - log_{50k}IC50$
ALVEICTEM	0.638
LVEICTEME	0.023
VEICTEMEK	0.021
EICTEMEKE	0.005
ICTEMEKEG	0.005
CTEMEKEGK	0.009
TEMEKEGKI	0.032
EMEKEGKIS	0.010
MEKEGKISK	0.018

Peptide	$1 - log_{50k}IC50$
ALVEICTEL	0.782
LVEICTELE	0.027
VEICTELEK	0.020
EICTELEKE	0.006
ICTELEKEG	0.005
CTELEKEGK	0.010
TELEKEGKI	0.021
ELEKEGKIS	0.008
LEKEGKISK	0.015

Figure 3.3: The sequence is scanned from N to C and at each iteration, according to window size, a word is extracted and predictions are performed. The same procedure is repeated with the sequence containing the substitution (marked in red). Score differences are logged.

$$x = \alpha \times NetMHCStab + (1 - \alpha) \times NetMHCcons \tag{3.5}$$

In order to make predictions for both tools, PR and RT protein sequences are scanned and a peptide list is generated by using a fixed-size sliding window over the protein sequence, the window length usually falling in the range of 9-10 amino acids. This peptide list, together with an HLA allotype is used as input for the NetMHCStab and NetMHCPan predictors. The threshold for MHC affinity is set at $500nM$ or 0.42 when using the $1 - log_{50k}IC50$ score. The threshold for NetMHCPan is set at 3 hours. Measuring changes in prediction results for peptide MHC affinity and stability is a straightforward task involving the comparison of predictions produced by the peptide list produced from a region where the mutation occurs with the scores produced from the consensus sequence. The output scores are subtracted from each other. An example of peptide list generation and NetMHCPan affinity prediction for HLA-A*02:01 is shown in Figure 3.3.

3.5.2 Prediction of proteasomal cleavage sites

Proteasomal cleavage prediction is accomplished by using the NetChop 3.1 method [Keşmir *et al.*, 2002; Nielsen *et al.*, 2005]. The prediction tool makes proteasomal cleavage predictions from a 17-mer input peptide and results at the C-terminal end of the amino acid at a particular position given. Word lists are generated the same way as for MHC peptide generation. It should be noted that the accuracy for proteasomal cleavage prediction is significantly lower than the predictions for MHC affinity. Therefore, the results are treated with caution. The prediction score ranges between 0 and 1 and generally scores nearer to the extremes are considered. An offline version of the tool was used.

3.5.3 Clustering HLA allotypes based on binding motifs and affinity

As mentioned in Chapter 1, HLA nomenclature provided valuable insights into the classification and thus relatedness of different HLA allotypes. However, this nomenclature is limited to serological and genetic classifications and may not always provide insight into the binding nature of the HLA allotypes in question. For this reason, various methods have been developed that aim to classify HLA allotypes based on binding affinity [Sidney *et al.*, 2008]. Another method, MHCClust was developed to cluster an arbitrary list of HLA allotypes [Martin Thomsen, 2014]. This program is available via a web-interface and was used to classify and visualize the relatedness of a set of HLA allotypes. As input, the method accepts a list of HLA allotypes to be classified and produces a neighbour-joining tree from by examining the sets of peptides produced by two or more HLA allotypes and using the shared binder list between the two.

3.6 Accounting for Epistatic Interactions to Compensate for Bias

There are cases where epistatic interactions between mutations can account for significant observed frequency changes between two sequence sets. If the change in a mutation, X, is dependent on the change of the other substitutions, collectively called Y, a change in sequence set 1, S_1 having significantly different frequencies of Y will also have significant frequency differences of X in sequences set 2, S_2. Two methods are used here to account for these possible biases. The first is a linear model to calculate an expected value of the frequency of X, i.e. X_E given the frequencies of Y and then calculating whether there is still significant frequency differences between X and X_E. A significant odds-ratio between X and X_E will indicate that the differences of the frequency of X in S_1 and S_2 are not solely due to the influence of Y.

3.6.1 Linear model to predict expected L90M frequency

Linear models for the estimation of L90M frequency given a set of frequencies for co-occurring mutations was constructed. The influence of every mutation was measured individually. Sequence sets were divided into groups depending on the existence or absence of the mutation in question. The sequences in the negative set were increasingly replaced by sequences from the mutation positive set and frequencies of all the mutations recorded at each step. With each mutation, a data set with 5000 frequency data points for each mutation were recorded. For each mutation, other co-occurring mutations were included as additional terms for the equation. For each model, the frequency of L90M, i.e. P(L90M) was calculated by a linear addition of the intercept together with the values produced by the other terms. The terms were chosen by observing frequently occurring combinations of PI resistance mutations listed in the hivDB together with L90M [Shafer, 2006a]. The high R^2 values indicate a high confidence in the equations for estimating L90M frequency. Equations 3.6, 3.7, 3.8, 3.9 and 3.10 represent L90M frequency estimators trained from the V82A, I54V, A71V, M46I and I84V sets respectively. Where $P(L90M_{XXX})$ is the expected frequency of L90M and XXX denote the mutation under which the model was trained. The x_y variables correspond to the frequency of the mutation denoted by the y subscript. The R^2 values are the calulated adjusted R^2 values for the model. A graphical representation of construction of the linear model is shown in Figure 3.4.

$$P(L90M_{V82A}) = 0.0712x_{V82A} + 0.152x_{I54V} + 0.335x_{M46I} + 0.265; R^2 = 0.852 \qquad (3.6)$$

$$P(L90M_{I54V}) = 0.289x_{I54V} + 0.216x_{M46I} + 0.027x_{I84V} + 0.275; R^2 = 0.938 \tag{3.7}$$

$$P(L90M_{A71V}) = 0.364x_{A71V} + 0.224x_{V82A} + 0.228; R^2 = 0.978 \tag{3.8}$$

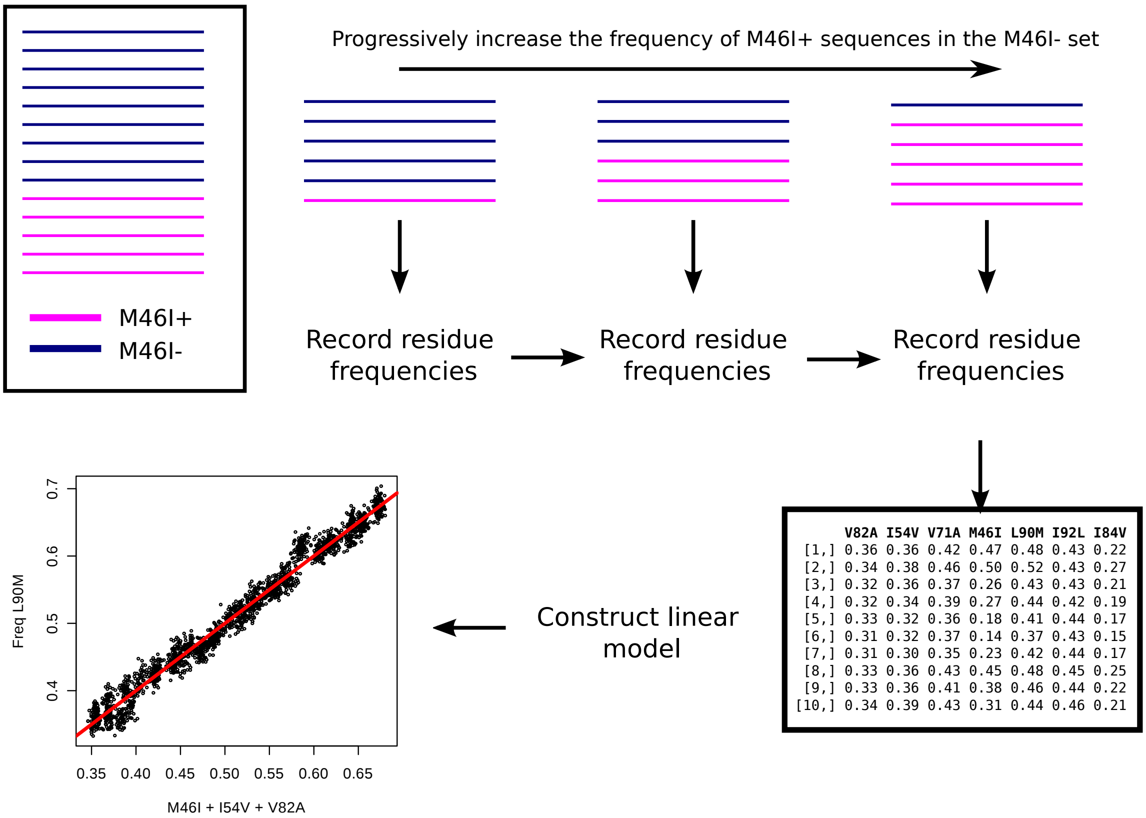
$$P(L90M_{M46I}) = 0.353x_{M46I} + 0.183x_{I54V} - 0.253x_{V82A} + 0.32, R^2 = 0.977 \tag{3.9}$$

$$P(L90M_{I84V}) = 0.356_{I84V} + 0.080_{M46I} + 0.347, R^2 = 0.977 \tag{3.10}$$

3.6.2 Balancing substitution frequencies between sets with a unique sampling procedure

In order to balance the frequencies of substitutions between two sequence sets, a sampling procedure was devised. As input, the procedure needs the substitutions and their target frequencies. For example, consider two sequence sets, S1 and S2 with sample sizes 200 and 1200 respectively. The frequency of the mutations A,B and C in S1 is 0.25, 0.12 and 0.20 respectively, while the frequencies for the same mutations in S2 are 0.40, 0.20, 0.30. In order to balance the frequencies of A, B and C in S2 with those in S1, the following procedure was followed:

1. Split the S2 sets into two parts, $S2_{neg}$ that does not contain sequences with A,B or C and $S2_{pos}$ that contains any of A,B or C. The sizes in this example for $S2_{neg}$ and $S2_{pos}$ are 400 and 800 respectively.


M46I+
M46I-
Progressively increase the frequency of M46I+ sequences in the M46I- set
Record residue frequencies
Record residue frequencies
Record residue frequencies
Construct linear model
Freq L90M
0.4
0.5
0.6
0.7
0.35
0.40
0.45
0.50
0.55
0.60
0.65
M46I + I54V + V82A

```
      V82A I54V V71A M46I L90M I92L I84V
 [1,] 0.36 0.36 0.42 0.47 0.48 0.43 0.22
 [2,] 0.34 0.38 0.46 0.50 0.52 0.43 0.27
 [3,] 0.32 0.36 0.37 0.26 0.43 0.43 0.21
 [4,] 0.32 0.34 0.39 0.27 0.44 0.42 0.19
 [5,] 0.33 0.32 0.36 0.18 0.41 0.44 0.17
 [6,] 0.31 0.32 0.37 0.14 0.37 0.43 0.15
 [7,] 0.31 0.30 0.35 0.23 0.42 0.44 0.17
 [8,] 0.33 0.36 0.43 0.45 0.48 0.45 0.25
 [9,] 0.33 0.36 0.41 0.38 0.46 0.44 0.22
[10,] 0.34 0.39 0.43 0.31 0.44 0.46 0.21
```

Figure 3.4: The figure demonstrates the procedure to construct the data sets used to predict the expected frequency of L90M considering the frequency of co-occurring mutations. First, the sequence set is split into two groups; one devoid of the mutation of interest, in this case PR M46I, and the other that contains the mutation. Note that each line represents the entire sequence set obtained from a single patient. The patient ID in the mutation negative-set are increasingly replaced with patient sequence sets from the M46I group. At each step the frequencies for V82A, I54V, V71A, M46I, L90M and I84V were recorded. From this data, a linear model specific to a particular mutation, M46I in this case was constructed.

2. Calculate the maximum size of the S2 set after adjustment. This can be done with the following formula:

$$n_{max} = \frac{n_{S2_{neg}}}{(1 - f_{max})} \tag{3.11}$$

Where n_{max} is the maximum size of the $S2_{adj}$ set, $n_{S2_{neg}}$ the size of the native S2 set and f_{max} the highest target frequency, in this case, A with a frequency of 0.25. Thus, the maximum size is $n_{max} = \frac{400}{1-0.25} = 533$. Thus the maximum sample sizes for $S2_{neg}$ and $S2_{pos}$ are 400 and 133 respectively.

3. Calculate the relative frequencies of the target frequencies by dividing the target frequencies by f_{max} with:

$$t_{f(ABC)} = (\frac{0.25}{0.25}, \frac{0.12}{0.25}, \frac{0.20}{0.25}) = (1.00, 0.48, 0.80) \tag{3.12}$$

   In this example, there should thus be $1.00 * 133 = 133$ sequences with A, $0.48 * 133 = 64$ sequences with B and $0.80 * 133 = 106$ sequences with C.

4. The substitutions are sorted in an ascending way according to their target frequencies and this order will determined which residue is sampled for first, thus the order is B, C, A according to the target frequencies.

5. Sequences sets are constructed according to the occurrence of the mutations A,B and C within them. Thus, S_B exclusively contains sequences with the B mutations.

6. The sequences are sampled in the following fashion

   (a) A total of 64 sequences of the lowest target frequency substitution, B, is sampled from the $S2_{pos}$ set.

   (b) The amount of the B sequences containing A and C are subtracted from the counts of A and C respectively, in this case 20 and 30 of A and C were found in the B sequences, leaving 103 and 76 sequences to be sampled for each. This ensures that no more sequences containing B are sampled and keeping the count at 64.

   (c) The procedure is repeated for A and C.

7. Finally, the final S2 set is produced by combining the $S2_{neg}$ and A,B,C samples from $S2_{pos}$.

The S1 and S2 sets are again compared in order to determine if other correlated mutations still differ significantly in frequency, indicating a selective pressure on the latter. Due to the nature of the sampling, this procedure can be repeated and subsequent Fisher's exact tests performed to determine confidence intervals of the odds-ratio between S1 and the adjusted S2. An example of this is shown in Figure 4.3 on page 99, where the mutations

M46I, I54V and V82A are balanced and the rest of the substitutions are measured for frequency differences.

3.7 Reconstruction of ancestral states

The influence of phylogeny in the appearance of drug resistance mutations needed to be taken into account. Appearance of inherited drug mutations could confound the results of the Fisher's analysis test in the sense that substitutions in PR and RT in patients may already exist natively in the HIV population, i.e. the transmitted strain [Bhattacharya *et al.*, 2007]. This, in turn, may cause the calculation of a frequency discrepancy of a substitution in an HLA assigned sequence set that may be due to a general accrual of a substitution in the HLA-negative set rather than the HLA type itself.

To determine whether the calculated associations between HLA and drug resistance mutations were merely due to this founder effect, ancestral reconstruction was performed for all the drug resistance mutations. As mentioned before, the majority of patients whose HIV sequences were annotated as "drug experienced", did not have sequences beyond the PR-RT regions of the HIV genome. Thus, the PR (99 amino acids) and RT (first 217 amino acids) were used to construct a phylogenetic tree. Since the drug resistance mutations would have a potential effect on the construction of a phylogenetic tree, i.e. clustering of tips in the tree due to drug resistance mutations, positions in the PR-RT regions involved in drug resistance were masked in the sequences, i.e. 50 positions were masked. The sequences used in the construction of the tree were not limited to the patients that were annotated as treatment experienced. The rationale behind this was to not limit sequences from treatment experienced individuals where drug resistance mutations known to occur and thus bias the construction of ancestral states. Only one sequence per patient was used. The chosen sequence was the one that differed the most from the HIV subtype consensus sequence after masking the drug resistance mutation positions.

The FastTree2 program was used to construct phylogenetic trees. This method allows the quick construction of maximum likelihood phylogenetic trees with large amounts of sequence data. Its accuracy is comparable to that of PhyML [Price *et al.*, 2010]. The `slow` parameter was used to improve the accuracy of the tree, as suggested by program's manual.

3.7.1 Reconstructing ancestral states of drug resistance mutations in HIV *pol*

Other studies have shown the importance of accounting for lineage effects that may be a contributor toe the emergence of mutations in HIV that would otherwise be associated with host factor, such as immune escape or another factor, such as antiretroviral treatment. The `ace` package allows for the construction of joint ancestral states [Paradis *et al.*, 2004] [1].

To construct the ancestral states of drug resistance mutations, a simple procedure was employed. The `ape` R package provides a method, `ace`, to construct ancestral states of tips and nodes within a tree [Paradis *et al.*, 2004]. This method uses the joint likelihood approach to calculate the state of the ancestral nodes in a tree. The ancestral state of a tip in a tree is given as the calculated ancestral state of its parent node. The procedure is as follows:

1. Assign the absence or presence (X and Y respectively) of an amino acid at a specific position in a patient sequence set to the tips of the tree corresponding to the patient ID. Each tip now has a value X or Y associated with it.

2. Use the `ace` function with the model parameter ”ARD” (all rates different) to calculate the ancestral states in the internal nodes of the tree.

3. The ancestral state of the tip is determined to be the state with the highest scaled likelihood in the immediate parent node.

4. Do this for all relevant (e.g. drug resistance) mutations at all relevant positions.

[1] Using version 3.1-2 of the package `ape`.

The model parameter for ancestral state calculation was the ”All rates different” (ARD) model as opposed to the ”Equal rates” model (EqR). This model estimate the transition rate from X → Y as well as from Y → X and thus allows for asymmetry in the transition rates. The ARD model also yielded a significantly higher log-likelihood value in most cases as determined using an ANOVA test to compare the ”EqR” and ”ARD” models. An example of the calculation of the ancestral states is provided in Figure 3.5. In this example, a tree constructed with FastTree using *pol* sequences from subtype B and the tips annotated with two states, ”Y” or ”X”, representing residues at position 93 in protease, namely isoleucine (X) and leucine (Y), represented by the colours blue and red. The likelihood of the ancestral states of these tips are shown in the immediate parental nodes and are represented as pie charts. The most likely ancestral state is represented by the colour showing the highest value in the pie chart.

Using the calculated ancestral states, it can be determined whether the appearance of a substitution is due to, for instance, drug resistance or whether it exists as a common substitution in a sub-population of HIV. The counts for substitutions deemed to be possibly inherited were removed from the Fisher's exact test calculations and a second set of p- and q-values were calculated to remove artificial enrichment or decrease in a substitution between two sets. A q-value of less than 0.2 after correcting for phylogeny was considered significant. An example of the the frequency of RT D67N in HLA A*23 assigned sequence sets is shown in Table 3.4. The first test does not take phylogeny into account and assumes that every occurrence of D67N is novel in response to reverse transcriptase inhibitors. It shows that the substitution is significantly decreased in the HLA A*23 set, yielding a log_2-odds ratio of of -1.03 and a p-value of 0.05. In the second table, the occurrences of D67N are adjusted for phylogeny and the new contingency test yields a log_2-odds ratio of -0.75 with a p-value of 0.21. Thus, it can be suggested that the lower occurrences of D67N in the HLA A*23 set is not conclusively due to the HLA type itself.

Table 3.4: This table illustrates the frequency discrepancy measurement of the drug resistance mutation RT D67N between HLA A*23-positive and A*23-negative assigned patient sequence sets. The first contingency table demonstrates a significant decrease in D67N for the HLA A*23-positive set. The second table shows the contingency test after adjusting the counts of D67N for phylogeny. The significance is lost yielding a new p-value of 0.21, indicating that the increased accumulation in the HLA A*23-negative set of D67N was probably not due to the absence of HLA A*23.

(a) Non-phylogenetic adjusted Fisher's exact test

	N	X
A*23-positive	12	27
A*23-negative	915	1011
p-value	p-value = 0.05	

(b) Phylogenetic adjusted Fisher's exact test

	X → N	X → X
A*23-positive	9	24
A*23-negative	640	1011
p-value	p-value = 0.21	

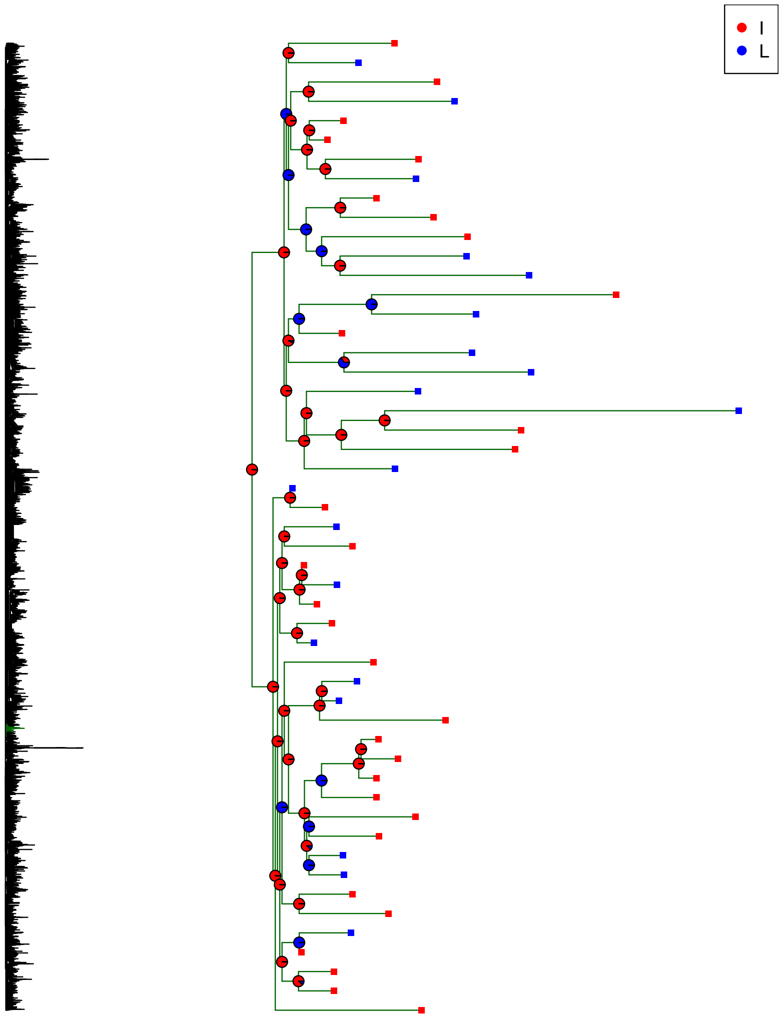


Figure 3.5: This figure shows the tip states and joint likelihood of the ancestral states of position 93 of PR-RT. The tee on the left is the maximum likelihood phylogenetic tree constructed for PR-RT sequences of HIV subtype B. The dark green portion in the tree is the portion of the tree represented on the right. The red and blue squares are the Ile and Leu states for the tips. The pie charts at the internal nodes represent the likelihoods of observing states Ile (blue) or Leu (red) in the descendants. Knowing that for subtype B, the ancestral state of position 93 in PR-RT is Ile, the focus would be on the appearance of Leu. If a tip state is blue and the parent node is red, it means that it is more likely for the blue state to have arisen as a new mutation in the tip. Conversely, some tips that are Leu do show the same state (blue) in the parental node, meaning that the mutation was likely inherited. The 93L mutation is a common polymorphism in HIV subtype B [Pupko *et al.*, 2000; Paradis *et al.*, 2004].

3.8 Summary

This chapter provided a clear description of the methods used to generate results in Chapter 4. The following chapter discusses the use of the methods mentioned herein to provide evidence for potential CTL epitopes that may be induced by antiretroviral resistance mutations. This include the detection of diminished substitutions in HLA type assigned groups, prediction of MHC affinity and stability of peptides as well as prediction of potential proteasomal cleavage sites.

4 Results

Detection of novel CTL epitopes elucidated by antiretroviral resistance mutations is a time consuming and expensive task. Usually, experimental evidence is crucial in the detection of these epitopes. Because of the extreme *pol* ymorphism of the HLA types, these procedures often become laborious and thus experiments are often only limited to a few HLA types/allotypes. The recent developments in detection of HLA-associated substitutions in HIV protein sequences has made it possible, at least in theory, to assign potential HLA types to the patients by examining regions for HLA-associated substitutions [Brumme *et al.*, 2009]. Furthermore, the performance of MHC affinity and the introduction of MHC stability predictors recently, has made the computational task of detecting potential novel epitopes more reliable, given the higher accuracies of the tools [Lundegaard *et al.*, 2010].

In this chapter, the results of investigating interactions of HLA types with amino acid substitutions in HIV RT and PR implicated in drug resistance are investigated. The first analysis measured significant frequency differences in substitutions between PR and RT sequences from patients assigned to particular HLA type and those deemed to be negative for the same HLA type. A few HLA types did indeed show discrepancies in substitution selection. Further investigated, was the influence of epistatic interactions on the significant results produced. The results show that certain mutation discrepancies still existed even after accounting for epistatic interactions. Investigating a possible causal relationship for these discrepancies, the prediction results of peptides in the consensus

PR and RT sequences by NetMHCPan and NetMHCStab were compared with peptides generated by amino acid substitutions [Nielsen *et al.*, 2007; Jørgensen *et al.*, 2014]. These results indicated immunological responses are plausible contributors to some diminished antiretroviral resistance mutations. Particular focus was placed on the PI resistance mutation, L90M and the reverse transcriptase inhibitor resistance mutations RT T215Y/F. These mutations and their relationship with HLA types B*48 and B*15 show strong evidence of negative selection of these substitutions. The potential CTL epitopes resulting from antiretroviral mutations ties in with evidence of HLA type B*15 being beneficial during antiretroviral therapy in HIV-1 infected individuals [Mason *et al.*, 2004; Mahnke and Clifford, 2006; Mueller *et al.*, 2011]. The detection of B*48 selected potential epitopes also renders this HLA type suitable for further investigation of CTL epitopes generated by the presence of antiretroviral mutations. The majority of the chapter focuses on ARV resistance mutations in HIV subtype B, with a similar analysis done for HIV subtype C. The main reason for the focus on subtype B, as will be demonstrated, is a lack of data that includes both imputation of HLA types in subtype C as well as lack of accrual of resistance mutations in available HIV subtype C sets.

4.1 Interaction between HLA type on antiretroviral resistance mutation frequencies in HIV subtype B

The potential influence of HLA type on ARV resistance-related substitutions was measured for various HLA types. Sequence acquisition and assignment to patient IDs is discussed in Section 3.1 on page 42. Assigning HLA types to patients was performed according to the protocol mentioned in Section 3.2 on page 43. Briefly, if an HLA type-associated mutation was found in any of the HIV PR-RT sequences of a patient, that patient was assigned the HLA type, for example if PR-RT sequences from a patient contained a E104D mutation, the HLA type B*40 was assigned to that patient. The patient sample size of the assigned HLA types are shown in Table 4.1 on page 67. Only sequences annotated as from patients that are treatment experienced were used.

The list of ARV resistance-related substitutions was limited to mutations occurring at significantly higher frequencies in treatment-experienced versus treatment-naïve individuals. From this, a list of positions was obtained and occurrences of any substitutions with a frequency of ≥ 0.01 were considered. Frequency discrepancies were measured with the Fisher's Exact Test (FET) (See Section 3.3 on page 45 and Section 3.3.1 on page 46). A modest significance value of 0.05 was chosen as a cut-off for the test. Together with the significance values, Storey's q-value that estimates false discovery rate was calculated to account for multiple tests. Only tests with q-values ≤ 0.10 were considered. A total of 6/18 HLA-A types available could be used for testing, while 14/26 HLA-B types could be used for testing. After filtering, 24 HLA types were assigned to the list of available patients. HLA types that couldn't be assigned due to their related substitutions not meeting criteria mentioned in Section 3.2, were A*01, A*02, A*24, A*25, A*26, A*30, A*31, A*33, A*34, A*66, A*69, A*74, B*08, B*13, B*14, B*38, B*42, B*45, B*50, B*51, B*53, B*55, B*56, and B*58.

4.1.1 Estimated accuracy of HLA type assignment

In order to estimate the accuracy of the HLA imputation method, PR-RT sequences with patient HLA annotations were obtained from the literature [Carlson *et al.*, 2012]. The mutations listed in Table 4.1 were used to impute HLA types. From this data, a confusion matrix was constructed measuring the True Positives (correctly assigned HLA types), false positives (FP, falsely assigned HLA types), true negatives (TN, the amount of sequences correctly assigned as not being the HLA type tested) and false negatives (FN, the number of patients for which the HLA type in question was not inferred by the substitutions). Using this data, the sensitivity, or the fraction of discovered HLA type and the Positive Prediction Value (PPV) or the fraction of correct predictions in the total set were calculated. The results are shown in Table 4.2 on page 68. Generally, it was observed that the PPV values are moderate with sensitivity being low. The low sensitivity can be attributed to two factors. Substitutions were chosen on a basis to maximize the PPV, i.e. only HLA associated substitutions with low odds ratios as well as substitutions that do not occur to frequently and thus may be *pol* ymorphisms, e.g. the PR I93L mutation, which is very strongly associated with B*15, but this mutation occurs in approximately 43% of all HIV subtype B protease sequences. The poorest performers in terms of PPV were B*41, B*44 and B*57 with PPV values below 0.20. The best performers were A*03, B*07, B*15 and B*48 with PPV values above 0.40. It is also noted that B*48 had the highest sensitivity at 0.60. This table also shows the global distribution of the HLA types that were assigned as well as the estimated frequency of the HLA types, given the assignment count, PPV and sensitivity. The estimated frequency of patients bearing the HLA type in each row is calculated with Equation 4.1. It is expected that the estimated HLA counts should be similar to the global frequency of the HLA type. Indeed, taking HLA B*41 as an example, the estimated percentage of patients with this HLA is 0.40% with the global percentage at 0.70%. Some discrepancies do occur with certain HLA types in terms of estimated frequency versus global frequencies, e.g. B*15 and B*40, but the rank of the frequencies are by and large the same.

$$HLA_{est} = \frac{HLA_{count} \times PPV}{Sensitivity} \tag{4.1}$$

4.1.2 Reconstruction of the ancestral states of mutations associated with drug resistance

To mitigate the potential influence of founder effects on the FET results, the ancestral state of each mutation associated with drug resistance was calculated with the method explained in section 3.7.1 on page 58. First, a phylogenetic tree was created using all the subtype B *pol* amino acid sequences. Positions associated with drug resistance as well as some correlated mutations were masked to ensure drug resistance mutations does not influence the construction of the tree. The reason for this was to exclude the possibility of merely the presence of drug resistance influencing the construction of the ancestral states.

Table 4.1: This table represents the breakdown of the sample sizes of HLA-assigned patients from which HIV-1 subtype B *pol* sequences were obtained. All the sequences were annotated as from patients that are treatment-experienced. The first column illustrates the HLA type, the second column indicates the number of assigned patients to that HLA type and the last column indicates the substitutions used for HLA type classification. Approximately half of the patients could be assigned an HLA type by using the available *pol* data. The total represents the overlapping total, meaning that some patients could be assigned multiple HLA types.

HLA Subtype	Number of assigned patients	Filter Residues
A03	166	K265R
A11	231	V205I, K265R
A23	39	K272A, K272R
A29	15	K272T, E303Q
A32	42	R310G
A68	30	K203R
B07	86	A257S, T264I
B15	137	Q273H, Q306H, Q306R, R310Q, R310G
B18	131	E237A, I241T, Q306D
B35	16	K221R
B37	13	Q306G
B39	15	D185E
B40	305	E105D, D222S, K265R, T299I
B41	74	T299I
B44	10	K272T
B48	112	K201R, K202R
B49	156	S147T, K148R
B52	159	I234V
B57	56	V107I, I241T
B81	48	P103S
C04	134	I277L
C07	141	V134M, D220N
C15	45	I234M
C16	16	D222G
Total (Overlapping)	1108 (out of 1965)	

A total of 50 positions were masked. For PR, 25 positions (10, 20, 24, 32, 33, 34, 36, 46, 47, 48, 50, 53, 54, 58, 60, 62, 63, 64, 71, 73, 76, 77, 82, 84 and 90) were masked. For RT, 25 positions (41, 62, 65, 67, 70, 74, 75, 77, 90, 98, 100, 101, 103, 106, 108, 115, 116, 138, 179, 181, 184, 190, 206, 210, 215) were masked. For patients with multiple HIV sequences, the sequence most divergent from the HIV consensus B was chosen. A total of 8,851 sequences were used adding ancestral and consensus sequences from HIV subtype B. A maximum likelihood tree was created with FastTree2. FastTree2 allowed for construction of a maximum likelihood tree in a feasible time while retaining accuracy close to PhyML. The tips of the tree were annotated with the patient ID of the patient from which they originated. For each DRM, a state was assigned to a tip, either "X" or "Y" representing the absence or presence of a mutation in that patient's HIV sequence set. Using the state information, the `ace` function was used with the ARD model parameters and choosing the joint ancestral reconstruction. The likelihood of the states, "X" and "Y" were tallied

Table 4.2: This table depicts imputed HLA types for patients from HIV subtype B PR-RT data. The HLA type is listed with the corresponding sensitivity and positive predictive value (PPV) as described in Section 3.2.1. Together with the performance data, the amount of patients assigned an HLA type is shown. The total estimated amount of patients with the HLA specified in the row is shown in the next column followed by the percentage of patients assigned the HLA type as well as percentage of the world wide representation of the HLA type [Carlson *et al.*, 2012]. HLA data source: `http://www.ncbi.nlm.nih.gov/projects/gv/mhc/`.

HLA	PPV	Sensitivity	HLA count	Adjusted Count	% HLA in set	% Global HLA
A03	0.43	0.23	117	219	11.10	6.10
A11	0.33	0.27	231	282	14.40	11.50
A23	0.32	0.16	39	78	4.00	2.70
A29	0.25	0.15	15	25	1.30	2.60
A32	0.37	0.19	42	82	4.20	1.60
B07	0.32	0.46	86	60	3.10	5.50
B15	0.39	0.16	137	334	17.00	11.50
B18	0.31	0.22	131	185	9.40	2.80
B35	0.27	0.07	16	62	3.20	7.30
B39	0.31	0.16	15	29	1.50	4.00
B40	0.3	0.25	305	366	18.60	12.70
B41	0.07	0.57	74	9	0.50	0.70
B44	0.4	0.04	10	100	5.10	6.00
B48	0.5	0.6	112	93	4.70	2.50
B49	0.12	0.3	156	62	3.20	1.00
B52	0.07	0.52	159	21	1.10	1.40
B57	0.31	0.13	56	134	6.80	2.00

for all the immediate ancestral nodes. The scaled likelihood ranges from 0 to 1 and the ancestral node was assigned the state with the higher likelihood, i.e. a scaled likelihood of greater than 0.50. If the tip and its immediate ancestor were "Y", it was assumed that the "Y" state was acquired rather than newly developed. These inherited states were excluded from the contingency tests performed in the following section. The fractions of remaining "Y" states for each DRM after phylogenetic correction are listed in Table 4.3.

Table 4.3: This table represents the pre- and post-adjusted counts of drug resistance mutations (DRM) in the subtype B data set after accounting for phylogeny. The first, second and third columns represent the position along PR-RT, the reference residue and drug resistance mutation respectively. The fourth column represents the total number of DRMs. The fifth column shows the number of drug resistance mutations remaining after filtering out the sequences deemed have inherited the specific DRM. The last column shows the percentage of the remaining DRM after filtering.

Pos	Res	Mut	Pre-adjusted n	Pos-adjusted n	% of Pre-adjusted
3	I	V	106	84	0.79
10	L	F	270	233	0.86
10	L	I	791	519	0.66
10	L	V	161	152	0.94
11	V	I	78	67	0.86
13	I	V	544	419	0.77

Continued on the following page...

16	G	A	79	54	0.68
19	L	I	202	176	0.87
20	K	R	348	297	0.85
20	K	I	169	151	0.89
20	K	T	68	66	0.97
24	L	I	180	164	0.91
30	D	N	68	37	0.54
32	V	I	160	122	0.76
33	L	F	388	304	0.78
33	L	I	75	64	0.85
34	E	Q	117	89	0.76
35	E	D	662	458	0.69
36	M	L	58	51	0.88
36	M	I	747	549	0.73
37	N	D	406	341	0.84
43	K	T	161	76	0.47
45	K	R	55	49	0.89
46	M	L	263	239	0.91
46	M	I	631	454	0.72
47	I	V	184	162	0.88
48	G	V	117	36	0.31
50	I	V	83	67	0.81
53	F	L	140	106	0.76
54	I	M	92	80	0.87
54	I	V	711	505	0.71
55	K	R	187	154	0.82
58	Q	E	206	184	0.89
60	D	E	252	223	0.88
61	Q	E	80	60	0.75
62	I	V	911	607	0.67
64	I	V	381	309	0.81
69	H	R	52	41	0.79
69	H	K	87	50	0.57
71	A	I	79	78	0.99
71	A	V	809	535	0.66
72	I	V	242	185	0.76
72	I	T	111	64	0.58
72	I	L	79	69	0.87
73	G	S	250	217	0.87
73	G	T	84	74	0.88
74	T	S	101	74	0.73
76	L	V	119	95	0.80
82	V	A	648	462	0.71
82	V	T	86	80	0.93
84	I	V	437	344	0.79

Continued on the following page...

85	I	V	128	107	0.84
88	N	D	75	45	0.60
89	L	M	54	52	0.96
89	L	V	125	87	0.70
90	L	M	894	541	0.61
92	Q	K	91	62	0.68
105	E	K	71	68	0.96
119	K	R	416	358	0.86
134	V	M	139	130	0.94
134	V	L	87	83	0.95
138	T	A	342	305	0.89
140	M	L	1045	672	0.64
142	K	N	65	65	1.00
142	K	E	265	201	0.76
142	K	Q	130	121	0.93
143	E	D	311	282	0.91
143	E	A	86	78	0.91
148	K	R	127	117	0.92
161	A	V	101	34	0.34
163	K	R	70	52	0.74
166	D	G	77	68	0.88
166	D	N	922	648	0.70
168	T	N	142	136	0.96
168	T	D	253	228	0.90
169	K	R	455	402	0.88
171	R	K	62	25	0.40
173	L	I	153	128	0.84
173	L	V	274	240	0.88
174	V	M	118	87	0.74
174	V	I	70	27	0.39
189	V	I	87	84	0.97
197	A	S	154	145	0.94
197	A	G	158	154	0.97
199	L	I	109	99	0.91
200	K	E	139	112	0.81
200	K	Q	116	97	0.84
201	K	Q	88	58	0.66
202	K	N	584	515	0.88
207	V	I	193	178	0.92
217	V	I	682	534	0.78
221	K	P	63	63	1.00
221	K	E	995	589	0.59
222	D	N	109	105	0.96
222	D	E	604	461	0.76
234	I	T	611	475	0.78

Continued on the following page...

241	I	V	221	204	0.92
257	A	S	89	30	0.34
261	S	Y	65	61	0.94
265	K	R	166	125	0.75
273	Q	E	75	48	0.64
273	Q	K	76	55	0.72
273	Q	R	92	46	0.50
276	D	N	84	37	0.44
278	V	I	223	199	0.89
278	V	D	71	28	0.39
280	Y	C	309	267	0.86
283	M	V	1029	718	0.70
287	Y	L	118	91	0.77
289	G	A	262	212	0.81
294	I	L	78	33	0.42
301	I	V	204	148	0.73
302	E	K	252	192	0.76
302	E	D	108	103	0.95
306	Q	E	402	346	0.86
307	H	Y	316	281	0.89
309	L	W	810	603	0.74
310	R	K	992	774	0.78
314	T	Y	1006	710	0.71
314	T	F	262	229	0.87

4.1.3 Diminished antiretroviral resistance mutations are associated with some HLA types

It was tested whether lower frequencies of specific ARV resistance mutations could be observed in HLA-type imputed patient sequence sets. Using the Fisher's Exact Test (FET) results produced for all the HLA type sets, the substitutions with diminished frequencies were extracted. These mutations were cross-referenced with substitutions listed in Table 4.4 on page 79 to determine the ARV for which the listed substitution provides resistance. Because PR and the first 217 amino acids of RT were joined, any positional reference above 99 falls within the RT region. Table 4.4 provides adjustments for the listed mutation positions. Detailed results given in Table 4.5 on page 73. Each sub-table represents the results for the respective HLA type. Note that these results do not reflect substitution discrepancies that occur in positions that are generally not associated with ARV resistance. While significance values are important, the tables also show the direction of selection, i.e. whether a substitution occurs at a significantly higher or lower frequency. The log_2-`Odds` column shows the log_2-Odds of the frequency difference of a substitution between the HLA type-positive set and the HLA type-negative sets. A positive value indicates enrichment, while a negative value indicates a diminished frequency of the substitution in the HLA-positive set. The cut-off for the p-value was set at 0.05 and the q-value at 0.10. Consider the first result for HLA A*03 in Table 4.5. First, the position in PR-RT is listed. Next, the consensus residue (for HIV-1 subtype B) is listed under `Org` with the substituted amino acid listed under `Mut`. There are two groups of three columns, for the native FET where phylogeny was not taken into account, i.e. the *unadjusted* set, followed by the FET corrected for phylogeny, i.e. the *adjusted* set. The last column shows whether a the FET test for a mutation produced significantly different results, i.e. either loss of significance (indicated by l) or a gain (indicated by g) of significance in the measured odds-ratio. For A*03, the mutation T168D is initially non-significant under for the uncorrected test, while it appears to be significantly enriched after the correction. Conversely, for A*23, the mutaiton D166N appeared to be significantly diminished in the uncorrected test, while this significance was lost when corrected for phylogeny. The results of HLA types linked with diminished major ARV resistance mutations are listed in Table 4.6 on page 80 with the results of the FET on the assigned HLA types listed in Table 4.1 presented in Table 4.5.

Referring to Table 4.6, it is clear that for almost all the HLA types listed, there are substitutions that appear to be significantly diminished ARV resistance mutations. For all the residues, the FET resulted in a significance value (p-value) of < 0.05. The underlined residues had corresponding q-values of < 0.10, indicating low probabilities of false-discovery. Diminished mutations that were frequently diminished included PR: M46I, I54V, V82A, I84V and L90M and RT: M41L (M140L), D67N (D166N), L210W (L309W), and T215Y (T314Y). These mutations often display correlation and may interact epistatically [Rhee *et al.*, 2007]. It was important to investigate the level of covariance of multiple substitutions to determine which epistatic interactions needed to be accounted for if, for a given HLA type, substitutions were calculated to be diminished. To this end, the HIVDB was consulted in order to compare frequencies of the aforementioned substitutions occurring individually and together with other substitutions. A summary of the interactions is shown in Table 4.7 on page 81. For each substitution, common covarying substitutions are listed. As a measure of independence, i.e. a proportional value of a substitution occurring independently of other substitutions, the ratio of the number of HIV-1 protein sequences (PR and RT) containing only the lone substitution to the number of sequences containing the most frequently occurring pattern containing the same substitution was measured. This value acted merely as an illustrative value of independence rather than a statistical measurement. Interestingly, L90M was the only substitution to show a high degree of independence. However, it was always correlated with other substitutions. Indeed, consulting Table 4.4, it becomes evident that this mutation is involved in PI resistance at varying levels for all but two PIs, namely tipranavir (TPV) and darunavir (DRV). It is common for the PR substitutions I54V, V82A, I84V and M46I to show correlation, with the exception of V82A and I84V, which shows modest negative correlation, yielding a phi-correlation value of -0.082 (see Section 3.6 on page 52 for an explanation of this calculation).

Table 4.5: The tables represent significant changes in substitutions within the PR-RT region of HIV-1 subtype B *pol*. The *Pos* column shows the position within the concatenated PR-RT sequences, *Org* is the consensus residue at that position, *Mut* is the amino acid substitution. The next three columns show the log_2-Odds of the tested mutation, the p- and q-values of the FET between the HLA-positive and HLA-negative sets *unadjusted* for phylogeny. This is followed by a second group of three columns showing the same odds-ratio, p- and q-values of the FET between HLA-positive and HLA-negative sets *adjusted* for phylogeny. The final column shows a gain (g) or loss (l) of significance for the FET in the *adjusted* results.

Fisher's exact test results

A03

			Unadjusted			Adjusted			
Pos	Res	Mut	*log*-Odds	p	q	*log*-Odds$_{adj}$	p_{adj}	q_{adj}	Change
54	I	V	-0.60	0.02198	0.142	-0.67	0.02723	0.145	

A11

			Unadjusted			Adjusted			
Pos	Res	Mut	*log*-Odds	p	q	*log*-Odds$_{adj}$	p_{adj}	q_{adj}	Change
168	T	D	0.57	0.04494	0.168	0.47	0.11534	0.336	l

Continued on the following page...

A23

			Unadjusted			Adjusted			
Pos	Res	Mut	*log*-Odds	p	q	*log*-Odds$_{adj}$	p_{adj}	q_{adj}	Change
90	L	M	-0.99	0.03457	0.112	-1.17	0.05634	0.167	l
166	D	N	-1.03	0.05064	0.147	-0.75	0.2085	0.450	l
168	T	D	-2.69	0.03675	0.117	-2.49	0.05405	0.161	l
202	K	N	1.39	0.00228	0.013	1.33	0.00588	0.026	
278	V	I	1.42	0.01274	0.049	0.71	0.27655	0.574	l
309	L	W	-1.26	0.01339	0.051	-1.30	0.02926	0.098	

A29

			Unadjusted			Adjusted			
Pos	Res	Mut	*log*-Odds	p	q	*log*-Odds$_{adj}$	p_{adj}	q_{adj}	Change
168	T	D	2.51	0.00078	0.027	2.72	0.00033	0.015	
169	K	R	1.47	0.04193	0.180	1.50	0.05916	0.220	l
314	T	F	1.76	0.02175	0.119	1.75	0.03374	0.159	

A32

			Unadjusted			Adjusted			
Pos	Res	Mut	*log*-Odds	p	q	*log*-Odds$_{adj}$	p_{adj}	q_{adj}	Change
46	M	I	-1.25	0.0197	0.051	-0.69	0.23855	0.439	l
54	I	V	-3.11	0.0	0.000	-2.54	0.00043	0.002	
82	V	A	-1.10	0.04307	0.099	-0.53	0.39973	0.645	l
84	I	V	-1.67	0.02099	0.054	-1.26	0.12298	0.255	l
90	L	M	-2.07	3e-05	0.000	-1.54	0.00721	0.021	
140	M	L	-2.52	0.0	0.000	-2.00	0.00015	0.001	
166	D	N	-2.39	0.0	0.000	-2.05	0.00035	0.002	
169	K	R	0.89	0.05513	0.123	1.03	0.02709	0.067	g
202	K	N	0.90	0.03762	0.089	1.13	0.01211	0.033	
309	L	W	-5.07	0.0	0.000	-4.59	0.0	0.000	
314	T	Y	-2.63	0.0	0.000	-2.29	3e-05	0.000	

A68

			Unadjusted			Adjusted			
Pos	Res	Mut	*log*-Odds	p	q	*log*-Odds$_{adj}$	p_{adj}	q_{adj}	Change
None									

B07

			Unadjusted			Adjusted			
Pos	Res	Mut	*log*-Odds	p	q	*log*-Odds$_{adj}$	p_{adj}	q_{adj}	Change
82	V	A	-0.74	0.05746	0.380	-0.97	0.03407	0.320	g
85	I	V	1.07	0.03889	0.354	1.28	0.02329	0.296	
173	L	V	-1.78	0.0098	0.160	-2.01	0.00672	0.153	

Continued on the following page...

B15

			Unadjusted			Adjusted			
Pos	Res	Mut	*log*-Odds	p	q	*log*-Odds$_{adj}$	p_{adj}	q_{adj}	Change
46	M	I	-0.64	0.02824	0.069	-0.45	0.18285	0.273	l
54	I	V	-1.41	0.0	0.000	-1.35	7e-05	0.000	
84	I	V	-1.41	0.00013	0.001	-1.21	0.00316	0.010	
90	L	M	-1.08	4e-05	0.000	-0.92	0.0023	0.008	
140	M	L	-1.30	0.0	0.000	-1.18	5e-05	0.000	
166	D	N	-1.02	8e-05	0.000	-1.03	0.00055	0.002	
168	T	D	-0.93	0.04099	0.087	-0.71	0.13801	0.224	l
169	K	R	0.54	0.05408	0.099	0.59	0.04077	0.088	g
202	K	N	0.53	0.04024	0.087	0.72	0.00642	0.019	
280	Y	C	-1.43	0.00136	0.005	-1.12	0.01569	0.041	
289	G	A	-1.14	0.01607	0.042	-0.74	0.16143	0.254	l
309	L	W	-1.67	0.0	0.000	-1.34	1e-05	0.000	
314	T	Y	-1.38	0.0	0.000	-1.22	2e-05	0.000	

B18

			Unadjusted			Adjusted			
Pos	Res	Mut	*log*-Odds	p	q	*log*-Odds$_{adj}$	p_{adj}	q_{adj}	Change
32	V	I	0.85	0.0434	0.544	0.76	0.11881	0.736	l
200	K	E	1.50	0.00022	0.022	1.53	0.00067	0.050	
278	V	I	0.83	0.02123	0.345	0.56	0.16252	0.736	l

B35

			Unadjusted			Adjusted			
Pos	Res	Mut	*log*-Odds	p	q	*log*-Odds$_{adj}$	p_{adj}	q_{adj}	Change
53	F	L	1.54	0.10923	0.391	2.05	0.04993	0.266	g
54	I	V	-Inf	0.0012	0.041	-Inf	0.01005	0.114	
90	L	M	-3.69	0.00088	0.034	-2.98	0.0173	0.140	
202	K	N	1.82	0.01244	0.111	1.89	0.0094	0.113	
278	V	I	2.53	0.00136	0.042	2.73	0.00065	0.031	
309	L	W	-Inf	0.00021	0.022	-Inf	0.00142	0.046	
314	T	Y	-2.92	0.00226	0.054	-2.36	0.02226	0.163	

B37

			Unadjusted			Adjusted			
Pos	Res	Mut	*log*-Odds	p	q	*log*-Odds$_{adj}$	p_{adj}	q_{adj}	Change
54	I	V	-2.72	0.04092	0.285	-2.15	0.20517	0.622	l
202	K	N	2.48	0.00302	0.077	2.70	0.00127	0.047	

B39

			Unadjusted			Adjusted			
Pos	Res	Mut	*log*-Odds	p	q	*log*-Odds$_{adj}$	p_{adj}	q_{adj}	Change

Continued on the following page...

82	V	A	-2.84	0.02883	0.746	-2.29	0.13951	1.000	l
84	I	V	-Inf	0.03097	0.746	-Inf	0.09111	1.000	l
90	L	M	-2.50	0.01024	0.560	-Inf	0.00713	0.528	
140	M	L	-2.93	0.00167	0.168	-3.23	0.00695	0.528	
202	K	N	-1.52	0.17533	1.000	-Inf	0.0274	0.771	g
309	L	W	-3.35	0.004	0.268	-2.88	0.02865	0.771	

B40

			Unadjusted			Adjusted			
Pos	Res	Mut	log-Odds	p	q	log-Odds$_{adj}$	p_{adj}	q_{adj}	Change
90	L	M	-0.36	0.04725	0.458	-0.39	0.07302	0.576	l
309	L	W	-0.37	0.04433	0.444	-0.47	0.02552	0.361	

B41

			Unadjusted			Adjusted			
Pos	Res	Mut	log-Odds	p	q	log-Odds$_{adj}$	p_{adj}	q_{adj}	Change
24	L	I	0.98	0.04018	0.911	1.17	0.01664	0.702	
168	T	N	1.22	0.01889	0.669	1.01	0.05801	0.959	l
169	K	R	0.78	0.03534	0.886	0.89	0.02496	0.776	
289	G	A	-1.91	0.01375	0.625	-2.11	0.01976	0.732	
309	L	W	-1.38	0.00047	0.057	-1.32	0.00291	0.287	

B44

			Unadjusted			Adjusted			
Pos	Res	Mut	log-Odds	p	q	log-Odds$_{adj}$	p_{adj}	q_{adj}	Change
168	T	N	2.16	0.04713	0.174	2.25	0.04034	0.151	
168	T	D	2.83	0.00177	0.037	3.03	0.00091	0.028	
169	K	R	1.79	0.03398	0.146	1.75	0.05001	0.173	l
314	T	F	2.28	0.01307	0.098	2.20	0.02678	0.132	

B48

			Unadjusted			Adjusted			
Pos	Res	Mut	log-Odds	p	q	log-Odds$_{adj}$	p_{adj}	q_{adj}	Change
54	I	V	-1.23	0.00014	0.001	-1.19	0.00183	0.007	
82	V	A	-1.02	0.00219	0.008	-0.73	0.05481	0.127	l
84	I	V	-1.35	0.00116	0.005	-1.21	0.00782	0.024	
90	L	M	-1.30	1e-05	0.000	-1.24	0.00058	0.003	
140	M	L	-1.31	0.0	0.000	-0.95	0.00245	0.009	
166	D	N	-0.94	0.00116	0.005	-0.57	0.06872	0.154	l
168	T	N	0.89	0.04162	0.099	0.88	0.05414	0.127	l
200	K	E	1.06	0.02216	0.059	1.42	0.00397	0.014	
309	L	W	-1.47	0.0	0.000	-1.33	0.00016	0.001	
314	T	Y	-1.08	0.00019	0.001	-0.99	0.00213	0.008	

Continued on the following page...

B49

			Unadjusted			Adjusted			
Pos	Res	Mut	log-Odds	p	q	log-Odds$_{adj}$	p_{adj}	q_{adj}	Change
32	V	I	-1.21	0.04564	0.433	-1.37	0.05466	0.465	l
54	I	V	-0.65	0.01532	0.242	-0.62	0.05106	0.463	l
314	T	F	-1.58	0.00085	0.041	-1.76	0.00133	0.060	

B52

			Unadjusted			Adjusted			
Pos	Res	Mut	log-Odds	p	q	log-Odds$_{adj}$	p_{adj}	q_{adj}	Change
54	I	V	-0.63	0.02001	0.556	-0.63	0.04149	0.567	
90	L	M	-0.55	0.03129	0.556	-0.56	0.05822	0.567	l
314	T	F	0.47	0.14296	0.556	0.69	0.04846	0.567	g

B57

			Unadjusted			Adjusted			
Pos	Res	Mut	log-Odds	p	q	log-Odds$_{adj}$	p_{adj}	q_{adj}	Change
202	K	N	1.01	0.01109	0.555	0.75	0.09755	0.961	l
289	G	A	-2.05	0.02675	0.799	-1.66	0.11598	0.961	l

B81

			Unadjusted			Adjusted			
Pos	Res	Mut	log-Odds	p	q	log-Odds$_{adj}$	p_{adj}	q_{adj}	Change
54	I	V	1.46	0.00041	0.021	1.10	0.02675	0.078	
82	V	A	1.90	0.0	0.000	1.82	8e-05	0.008	

C04

			Unadjusted			Adjusted			
Pos	Res	Mut	log-Odds	p	q	log-Odds$_{adj}$	p_{adj}	q_{adj}	Change
280	Y	C	0.63	0.049	0.303	0.72	0.04227	0.308	

C07

			Unadjusted			Adjusted			
Pos	Res	Mut	log-Odds	p	q	log-Odds$_{adj}$	p_{adj}	q_{adj}	Change
47	I	V	0.84	0.02395	0.315	0.88	0.02463	0.349	
84	I	V	0.67	0.01588	0.252	0.59	0.05709	0.545	l
85	I	V	1.08	0.01145	0.203	1.12	0.01572	0.274	
140	M	L	0.77	0.00315	0.081	0.84	0.00351	0.089	
166	D	N	1.19	0.0	0.000	1.29	0.0	0.000	
168	T	N	0.90	0.02735	0.336	0.91	0.03417	0.413	
168	T	D	0.75	0.02593	0.325	0.64	0.06946	0.636	l
173	L	I	1.35	0.00043	0.026	1.25	0.00317	0.089	
280	Y	C	0.65	0.04062	0.408	0.57	0.11097	0.733	l
309	L	W	0.92	0.00028	0.019	0.92	0.00089	0.057	

Continued on the following page...

314	T	Y	0.75	0.00324	0.081	0.74	0.00788	0.167	

C15

			Unadjusted			Adjusted			
Pos	Res	Mut	*log*-Odds	p	q	*log*-Odds$_{adj}$	p_{adj}	q_{adj}	Change
53	F	L	1.35	0.03289	0.214	1.64	0.02053	0.170	
82	V	A	0.90	0.05123	0.273	1.16	0.01557	0.142	g
90	L	M	-1.33	0.00617	0.098	-1.26	0.02919	0.200	
202	K	N	1.79	8e-05	0.011	1.96	1e-05	0.002	
283	M	V	1.43	0.00244	0.056	1.53	0.00306	0.058	

C16

			Unadjusted			Adjusted			
Pos	Res	Mut	*log*-Odds	p	q	*log*-Odds$_{adj}$	p_{adj}	q_{adj}	Change
202	K	N	2.05	0.00936	0.079	2.12	0.00532	0.064	

4.2 Exploring causal relationships between HLA types and diminished ARV resistance-related substitutions

Prior to examining the diminished residues, scenarios were considered that offered possible explanations of frequency discrepancies:

1. The HLA type-positive group contains a high number of individuals that have not undergone drug treatment long enough for ARV resistance-related mutations to accumulate.

2. The type of ARVs used may not include those where an observed diminished residue contributes to resistance.

3. The substitutions may result in the generation of immunologically active CTL epitopes, or alternatively, there are negative correlations between mutations conferring CTL escape and ARV drug resistance.

Table 4.4: These are the lists of major ARV resistance mutations with associated protease inhibitors and reverse transcriptase inhibitors. Table a) represents mutations providing resistance to protease inhibitors (PI). Table b) represents mutations providing resistance to nucleoside analogue reverse transcriptase inhibitors. Table c) represents mutations providing resistance against non-nucleoside reverse transcriptase inhibitors. Obtained from `http://hivdb.stanford.edu` [Rhee *et al.*, 2003]

.

(a) PI Resistance Mutations

PR Pos	30	32	33	46	47	48	50	54	76	82	84	88	90
Cons		**V**	**L**	**M**	**I**	**G**	**I**	**I**	**L**	**V**	**I**	**N**	**L**
ATV/r		I	F	IL	V	VM	L	VTALM		ATFS	V	S	M
DRV/r		I	F		VA		V	LM	V	F	V		
FPV/r		I	F	IL	VA		V	VTALM	V	ATSF	V		M
IDV/r		I		IL	V			VTALM	V	AFTS	V	S	M
LPV/r		I	F	IL	VA	VM	V	VTALM	V	AFTS	V		M
NFV	N		F	IL	V	VM		VTALM		AFTS	V	DS	M
SQV/r						VM		VTALM		AT	V	S	M
TPV/r		I	F	IL	VA			VAM		TL	V		

(b) Nucleoside Analog Resistance Mutations

	Discriminatory Mutations					Thymidine Analog Mutations (TAMs)						
RT Pos	**184**	**65**	**70**	**74**	**115**	**41**	**67**	**70**	**210**	**215**	**219**	**151**
Adj Pos	**283**	**164**	**169**	**173**	**214**	**140**	**166**	**169**	**309**	**314**	**318**	**250**
Consensus	**M**	**K**	**K**	**L**	**Y**	**M**	**D**	**K**	**T**	**T**	**K**	**Q**
3TC	VI	R										M
FTC	VI	R										M
ABC	VI	R	E	VI	F	L			W	FY		M
DDI	VI	R	E	VI		L			W	FY		M
TDF	***	R	E		F	L		R	W	FY		M
D4T	***	R	E			L	N	R	W	FY	QE	M
ZDV	***	***	*	*		L	N	R	W	FY	QE	M

(c) Non-Nucleoside Reverse Transcriptase Inhibitors

RT Pos	100	101	103	106	138	181	188	190	230
Adj Pos	**199**	**200**	**202**	**205**	**237**	**280**	**287**	**289**	**329**
Cons	**L**	**K**	**K**	**V**	**E**	**Y**	**Y**	**G**	**M**
NVP	I	EP	NS	AM		CIV	LCH	ASE	L
EFV	I	EP	NS	AM		CIV	LCH	ASE	L
ETR	I	EP			AGKQ	CIV	L	ASE	L
RPV	I	EP			AGKQ	CIV	L	ASE	L

If the first scenario holds, the natural consequence would be that PR and RT sequences associated with an HLA type will have lower ARV resistance mutations. This could directly confound the results that appear to produce FET results with high significance (both low p-values and q-values) and the causal relationship might just be due to the low acquisition of ARV resistance mutations.

Scenario two would entail that the observation of a diminished residue could just be due to low exposure of ARVs associated with the diminished mutation in the assigned group. Considering the residues of interest for both PR and RT, there is enough ARV resistance

Table 4.6: This table represents substitutions within the PR-RT region of *pol* that are diminished in HLA type positive sets. The left column represents the HLA types set and the right column represents the diminished substitutions. The underlined substitutions are those that yielded a significance value of $p < 0.05$ and a q-value of < 0.100 in both the *unadjusted* and *adjusted* FET results. Substitutions that are not underlined indicate significance in the *unadjusted* but not *adjusted* FET results. A double asterisk (**) indicates a q-value above 0.10 with a p-value of less than 0.05. for both the adjusted and unadjusted results, while a positive (+) indicates a gain of significance, i.e. p-value of less than 0.05 but with a q-value of greater than 0.10, whereas a double positive (++) indicates a q-value of less than 0.10. A single asterisk indicates significance for both *unadjusted* and *adjusted* sets, but with a q-value greater than 0.10. A double asterisk indicate a gain of p-value significance, but with a q-value of greater than 0.10. *None* indicates no discrepancies in DRM for the specific HLA type.

HLA Subtypes	Diminished Substitutions
A*03	54V
A*11	None
A*23	90M 166N
A*29	None
A*32	46I 54V, 82A, 84V 90M, 140L, 166N, 309W, 314Y
A*68	None
B*07	46I, 82A++ 166N, 309W, 314F, 314Y
B*15	46I, 48V, 54V, 84V, 90M, 140L, 166N, 280C, 309W, 314Y
B*18	None
B*35	54V, 90M, 309W, 314Y
B*37	54V
B*39	90M*, 140L*, 309W
B*40	90M**, 289A+, 309W**
B*41	309W
B*44	None
B*48	54V, 82A, 84V, 90M, 140L, 166N, 309W, 314Y
B*49	54V 314F
B*52	54V** 90M**
B*57	289A
B*81	None

mutation redundancy, i.e. all the diminished mutations mentioned earlier were considered drug resistance mutations for a majority of the PR inhibitors and RT inhibitors.

The third scenario posed an interesting prospect. Mentioned in Chapter 1, was the motif-like nature of immunologically active CTL epitopes. A mutations within a CTL epitope can change its viability, either by processing and presentation of the epitope, stability of the peptide-MHC complex on the cell membrane and recognition of the peptide by the appropriate T-cell receptors. Since CTL escape is possible through these mutations, so too can epitopes be generated or enhanced through the same mechanism [Mason *et al.*, 2004]. An immunological mechanism as a reason for the observed diminished residues was investigated. To this end, a comparative analysis was performed between the prediction results of *NetMHCPan* and *NetMHCStab* for the consensus HIV-1 subtype B *pol* (PR and

Table 4.7: In this table, major mutations and their positively correlated counterparts are listed. Columns from the left represent the substitution, the frequency of this substitution in the sequences from HIV-1 subtype B, the common correlated mutations, the most frequent mutation pairs that occur simultaneously with the substitution and the independence value, which is the number of sequences that have only the substitution in question divided by the highest occurring substitution pattern. Only PR L90M has an independence value of more than 1.00.

Substitution	Subtype Frequency	**Common Interacting Substitutions**	Most frequent Pattern	Independence value
Protease				
M46I	0.26	54V, 82A, 84V, 90M	46I, 90M	0.58
I54V	0.28	46I, 82A, 84V, 90M	54V,82A,90M	0.15
V82A	0.26	46I, 54V, 90M	54V,82A,90M	0.52
I84V	0.18	46I, 54V, 90M	46I,84V,90M	0.21
L90M	0.39	46I, 54V, 82A, 84V	46I, 90M	3.12
Reverse Transcriptase				
M41L	0.26	67N, 69D, 184V, 210W, 215Y	41L,184V,215Y	0.83
D67N	0.23	67N, 69D, 70R, 184V, 210W, 215Y	67N,70R,184V	0.22
L210W	0.26	67N, 69D, 74V, 184V, 210W, 215Y	41L,184V,210W,215Y	0.16
T315Y	0.32	67N, 69D, 70R, 184V, 210W, 215Y	41L,184V,215Y	0.38

RT) sequences, versus the prediction results obtained by including the substitutions in these sequences. This section explores the results from these immunological analyses and potential effect of this on ARV resistance substitution selection in PR and RT sequences from treatment-experienced patients.

4.2.1 Substitutions in PR and RT have effects on predicted peptide MHC affinity and stability

The results of the changes in MHC binding affinity and stability predictions are shown in Table 4.9 on page 84. These results represent changes caused by substitutions occurring in the PR and RT regions (up to position 217) of the *pol* protein. The results were filtered to only include those that induced significant changes, i.e. a *NetMHCPan* change of 0.10 and a NetMHCStab change of 0.50. The table is further annotated with markers indicating a change that breaks the binding threshold of *NetMHCPan*. Positive changes indicated increased binding affinity and stability respectively. The allotypes tested belonged to all the known HLA types, even those excluded from FET. This was done to demonstrate possible CTL epitope generation or escape for the HLA type in the absence of a FET. It should be noted that due to the limited HLA allotypes available for MHC stability prediction by NetMHCStab, the results were extrapolated from the closest neighbour.

Although it is impossible to confirm with the available data that there is a large correlation of MHC stability between related HLA allotypes, these results were still considered, albeit interpreted with caution. The MHCClust tool was used to cluster HLA allotypes not available in NetMHCStab with those that are available in NetMHCStab [Martin Thomsen, 2014]. See Section 3.5.3 on page 52 for a description of MHCClust. The results obtained from MHCClust are shown in Figure 4.1. The HLA allotypes used in NetMHCPan and NetMHCStab are shown in Table 4.8. Briefly, the MHCClust tool uses the top predicted NetMHCPan results of chosen HLA allotypes as a distance metric for clustering. The tree is bootstrapped over 100 iterations and the bootstrap values are displayed on the edges. This tree also illustrates that although by definition the HLA allotype nomenclature shows genetic similarity, small changes also affects the binding motif of the allotype and thus can affect the true relatedness of allotypes beyond their nomenclature. An example of this, is the HLA allotype B*4802, which is clustered together with B*1503 rather than B*4801.

Table 4.8: This table represents the nearest neighbour for HLA allotypes that do not have NetMHCStab predictions available. The left column represents NetMHCStab allotype and the right column the HLA allotypes to which they are extrapolated.

HLA Allotype in NetMHCStab	Allotype List
A*0101	A*0101, A*2901
A*0201	A*0203, A*0201, A*3201
A*0301	A*0301
A*1101	A*1101
A*2402	A*2301, A*2402
A*2601	A*2601
B*0702	B*0702
B*1501	B*1501, B*1503, B*4802
B*2705	B*2705
B*3501	B*3501
B*3901	B*3901
B*4001	B*4002, B*4402, B*4001
B*5801	B*5801

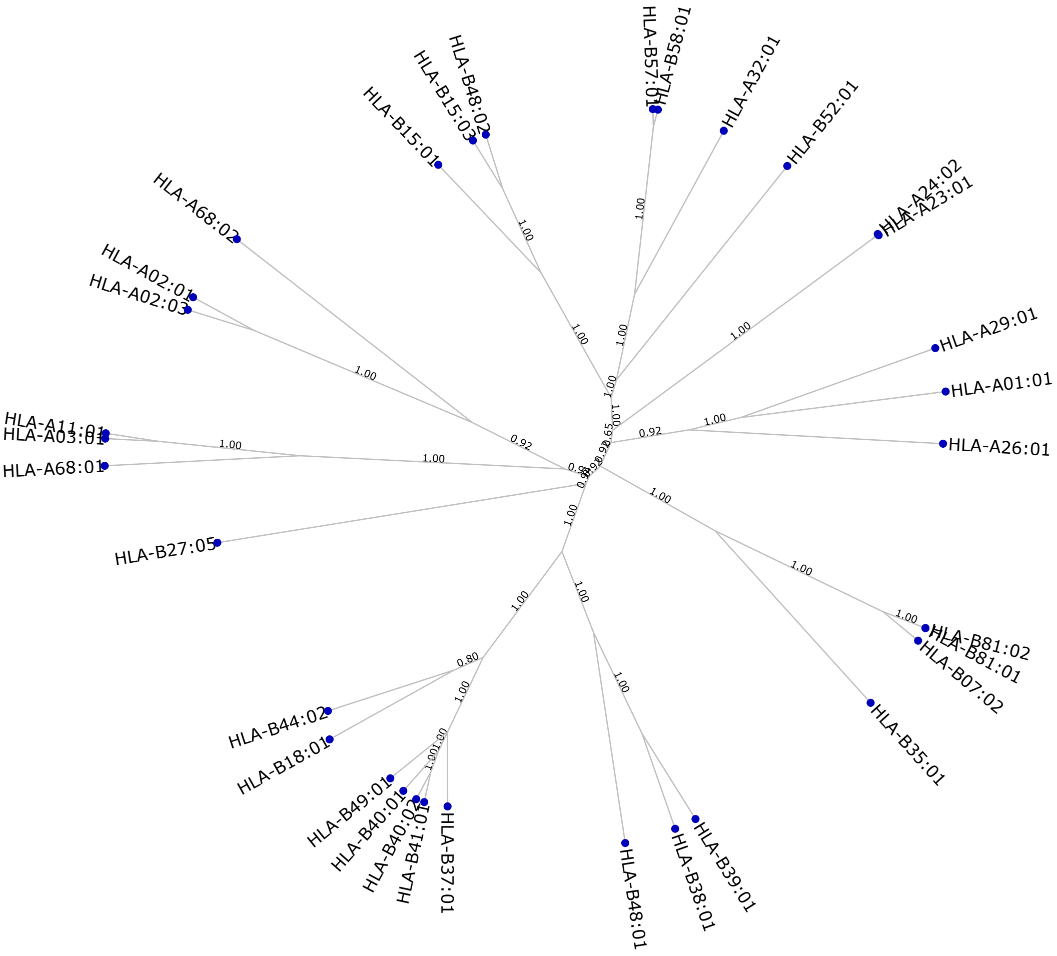


Figure 4.1: This figure is the bootstrapped tree obtained from MHCClust. The nodes show the HLA allotypes. The values on the edges are the bootstrapped values. HLA allotypes placed closer together are predicted to have larger set of common presented peptides.

4.2.2 Comparison of prediction results of mutated CTL epitopes to results from previous studies

It was interesting to note that some of the prediction results could be corroborated by literature studies. The A*02 epitope, VLVGPTPVNI (PR 76-84) has been shown to exhibit a slight increase in avidity by the PR I83V ARV resistance mutation [Singh and Barry, 2004]. The prediction results indicate that there was an increase in MHC binding affinity for A*0201 and A*0203 as well as an increased stability for A*0201 of 1.95 hours and could lead to longer availability of the epitope to encounter an appropriate TCR. The same applies to the PR mutation I54V, which produced slightly higher affinity and stability scores and corresponding to a slight increase in avidity as reported elsewhere [Mueller *et al.*, 2007]. The RT mutation, M41L resulted in a large predicted increase in stability of 4.06 hours of the A*02 epitope `ALVEICTEL` versus the wild type `ALVEICTEM` and has been reported as inducing a significantly stronger CTL response [Mason *et al.*, 2004]. The RT V36M (V134M) mutation also had a largely positive impact on MHC stability, with a 6.21 hour increase in stability prediction, but this mutation was found to quickly revert to the wild type in another study [Karlsson *et al.*, 2007]. It was observed that the same mutation had a profoundly negative impact on binding to B*48:02 and B*1501/3, but no previous studies were found to confirm this nor whether the wild type form is indeed an epitope of the aforementioned HLA allotypes. Interestingly, the prediction results also revealed the CTL escape mechanisms of some substitutions. The RT K166R mutation is predicted as an escape mutation for both A*03 and A*11 types. The mutated peptide, `AIFQSSMTR` had a 12.47 and 6.34 hour predicted half life decrease for A*1101 and A*0301 respectively, although the predicted affinity remained above the threshold for both allotypes [Koibuchi *et al.*, 2005].

Table 4.9: The following table represents the intial predicted MHC affinity and stability predictions by NetMHCPan and NetMHCstab respectively for HIV subtype B PR-RT. The HLA, Mut and Freq columns represent the HLA allotype for which the prediction was made, the substitution and the frequency of the substitution. The following two columns represent the consensus and mutated peptide, with the mutated residue in orange and underlined. The $IC50_{pep}$ and $IC50_{mut}$ columns show the predicted $1-log_{50000}IC50$ value obtained from prediction results for the consensus and mutated peptide, while the $IC50_{\triangle}$ column shows the log difference between the two predicted scores. The o_{pep} and o_{mut} columns show the predicted stability (measured as half-rate in hours) for the consensus and mutated peptide. The $o_{\triangle}$ value shows the difference between the two scores. The "T" column shows when a change in score resulted in the crossing of the $1-log_{50000}IC50$ threshold of 0.42, with "-T" meaning the affinity dropped to below the threshold and "+T" meaning the affinity increased above the threshold.

HLA	Mut	Freq	Pep	Mut	$IC50_{pep}$	$IC50_{mut}$	$IC50_{\triangle}$	T	o_{pep}	o_{mut}	$o_{\triangle}$
B*1501	L10F	0.131	PQITLWQRPL	PQITLWQRPF	0.17	0.38	0.21		0.87	2.06	1.19
B*1503	L10F	0.131	PQITLWQRPL	PQITLWQRPF	0.53	0.73	0.20		0.87	2.06	1.19
B*4802	L10F	0.131	PQITLWQRPL	PQITLWQRPF	0.47	0.62	0.15		0.87	2.06	1.19
A*0203	L10I	0.388	TLWQRPLVTI	TLWQRPIVTI	0.48	0.47	-0.01		4.45	5.02	0.57
A*0201	L10V	0.08	TLWQRPLVTI	TLWQRPVVTI	0.56	0.53	-0.02		4.45	5.01	0.57

Continued on the following page...

HLA	Mut	Freq	Pep	Mut	$IC50_{pep}$	$IC50_{mut}$	$IC50_{\Delta}$	T	o_{pep}	o_{mut}	o_{Δ}
A*0203	L10V	0.08	TLWQRPLVTI	TLWQRPVVTI	0.48	0.47	-0.01		4.45	5.01	0.57
A*0201	I13V	0.264	TLWQRPLVTI	TLWQRPLVTV	0.56	0.67	0.11		4.45	8.55	4.11
A*0203	I13V	0.264	TLWQRPLVTI	TLWQRPLVTV	0.48	0.61	0.13		4.45	8.55	4.11
A*1101	I13V	0.264	VTIKIGGQLK	VTVKIGGQLK	0.67	0.63	-0.04		3.33	2.69	-0.65
A*1101	L19I	0.096	VTIKIGGQLK	VTIKIGGQIK	0.67	0.56	-0.10		3.33	2.44	-0.90
A*0301	K20R	0.17	VTIKIGGQLK	VTIKIGGQLR	0.46	0.21	-0.24	-T	1.56	0.53	-1.03
A*1101	K20R	0.17	VTIKIGGQLK	VTIKIGGQLR	0.67	0.44	-0.23		3.33	0.94	-2.39
A*0201	L24I	0.082	LLDTGADDTV	LIDTGADDTV	0.60	0.39	-0.21	-T	2.19	0.93	-1.25
A*0203	L24I	0.082	LLDTGADDTV	LIDTGADDTV	0.50	0.33	-0.17	-T	2.19	0.93	-1.25
B*1501	M46I	0.302	KMIGGIGGF	KIIGGIGGF	0.67	0.49	-0.18		9.48	3.77	-5.71
B*1503	M46I	0.302	KMIGGIGGF	KIIGGIGGF	0.81	0.56	-0.26		9.48	3.77	-5.71
B*4802	M46I	0.302	KMIGGIGGF	KIIGGIGGF	0.70	0.39	-0.31	-T	9.48	3.77	-5.71
B*1501	M46L	0.129	KMIGGIGGF	KLIGGIGGF	0.67	0.57	-0.10		9.48	5.18	-4.30
B*1503	M46L	0.129	KMIGGIGGF	KLIGGIGGF	0.81	0.66	-0.15		9.48	5.18	-4.30
B*4802	M46L	0.129	KMIGGIGGF	KLIGGIGGF	0.70	0.49	-0.21		9.48	5.18	-4.30
A*1101	I47V	0.079	MIGGIGGFIK	MVGGIGGFIK	0.62	0.66	0.03		0.97	1.60	0.63
B*1501	I47V	0.079	KMIGGIGGF	KMVGGIGGF	0.67	0.62	-0.05		9.48	8.42	-1.06
B*1503	I47V	0.079	KMIGGIGGF	KMVGGIGGF	0.81	0.78	-0.03		9.48	8.42	-1.06
B*4802	I47V	0.079	KMIGGIGGF	KMVGGIGGF	0.70	0.64	-0.06		9.48	8.42	-1.06
A*0201	I54V	0.349	KMIGGIGGFI	KMIGGIGGFV	0.64	0.76	0.11		1.40	2.51	1.11
A*0203	I54V	0.349	KMIGGIGGFI	KMIGGIGGFV	0.74	0.84	0.10		1.40	2.51	1.11
B*1503	K55R	0.085	KVRQYDQIPI	RVRQYDQIPI	0.51	0.58	0.07		1.28	1.76	0.48
A*0201	Q58E	0.098	RQYDQIPIEI	REYDQIPIEI	0.64	0.19	-0.45	-T	2.86	1.15	-1.71
A*0203	Q58E	0.098	RQYDQIPIEI	REYDQIPIEI	0.58	0.17	-0.41	-T	2.86	1.15	-1.71
B*1501	Q58E	0.098	RQYDQIPIEI	REYDQIPIEI	0.42	0.21	-0.20		3.78	0.44	-3.34
B*1503	Q58E	0.098	RQYDQIPIEI	REYDQIPIEI	0.87	0.75	-0.12		3.78	0.44	-3.34
B*4802	Q58E	0.098	RQYDQIPIEI	REYDQIPIEI	0.84	0.79	-0.05		3.78	0.44	-3.34
B*1501	D60E	0.116	RQYDQIPIEI	RQYEQIPIEI	0.42	0.39	-0.03		3.78	4.24	0.47
B*1503	D60E	0.116	RQYDQIPIEI	RQYEQIPIEI	0.87	0.88	0.01		3.78	4.24	0.47
B*4802	D60E	0.116	RQYDQIPIEI	RQYEQIPIEI	0.84	0.85	0.02		3.78	4.24	0.47
B*1501	I64V	0.18	RQYDQIPIEI	RQYDQIPVEI	0.42	0.43	0.02	+T	3.78	4.63	0.85
B*1503	I64V	0.18	RQYDQIPIEI	RQYDQIPVEI	0.87	0.87	-0.00		3.78	4.63	0.85
B*4802	I64V	0.18	RQYDQIPIEI	RQYDQIPVEI	0.84	0.83	-0.01		3.78	4.63	0.85
B*3901	A71V	0.395	HKAIGTVLV	HKVIGTVLV	0.47	0.39	-0.07	-T	1.88	1.38	-0.49
B*3901	I72T	0.052	GHKAIGTVL	GHKATGTVL	0.48	0.43	-0.06		1.13	1.61	0.48
B*3901	I72T	0.052	HKAIGTVLV	HKATGTVLV	0.47	0.54	0.07		1.88	2.43	0.55
B*0702	V82A	0.305	TPVNIIGRNL	TPANIIGRNL	0.53	0.62	0.09		1.91	2.85	0.94
A*0201	I84V	0.206	VLVGPTPVNI	VLVGPTPVNV	0.54	0.67	0.13		2.16	4.10	1.95
A*0203	I84V	0.206	VLVGPTPVNI	VLVGPTPVNV	0.58	0.71	0.14		2.16	4.10	1.95
A*0201	L90M	0.423	LLTQIGCTL	LMTQIGCTL	0.46	0.50	0.04		4.58	3.35	-1.23
A*0201	L90M	0.423	NLLTQIGCTL	NLMTQIGCTL	0.43	0.56	0.13		1.87	2.38	0.51
A*0203	L90M	0.423	LLTQIGCTL	LMTQIGCTL	0.55	0.58	0.03		4.58	3.35	-1.23
A*0203	L90M	0.423	NLLTQIGCTL	NLMTQIGCTL	0.40	0.65	0.24	+T	1.87	2.38	0.51
A*3201	L90M	0.423	LTQIGCTLNF	MTQIGCTLNF	0.37	0.53	0.16	+T	2.00	2.49	0.49
B*1501	L90M	0.423	LLTQIGCTL	LMTQIGCTL	0.34	0.45	0.11	+T	1.59	3.13	1.53

Continued on the following page...

HLA	Mut	Freq	Pep	Mut	$IC50_{pep}$	$IC50_{mut}$	$IC50_{\Delta}$	T	o_{pep}	o_{mut}	o_{Δ}
B*1503	L90M	0.423	LLTQIGCTL	LMTQIGCTL	0.57	0.75	0.18		1.59	3.13	1.53
B*2705	L90M	0.423	GRNLLTQIGC	GRNLMTQIGC	0.31	0.29	-0.02		2.42	1.70	-0.72
B*4802	L90M	0.423	LLTQIGCTL	LMTQIGCTL	0.49	0.70	0.21		1.59	3.13	1.53
B*5801	L90M	0.423	LTQIGCTLNF	MTQIGCTLNF	0.46	0.60	0.14		2.00	2.49	0.49
A*0301	K119R	0.194	KLKPGMDGPK	KLKPGMDGPR	0.55	0.36	-0.19	-T	3.33	0.81	-2.52
B*0702	K119R	0.194	GPKVKQWPL	GPRVKQWPL	0.54	0.74	0.20		1.59	3.10	1.51
A*0201	V134M	0.067	ALVEICTEM	ALMEICTEM	0.64	0.83	0.19		5.28	11.49	6.21
A*0201	V134M	0.067	ALVEICTEME	ALMEICTEME	0.12	0.27	0.15		0.86	1.64	0.78
A*0203	V134M	0.067	ALVEICTEM	ALMEICTEM	0.73	0.85	0.12		5.28	11.49	6.21
A*0203	V134M	0.067	ALVEICTEME	ALMEICTEME	0.18	0.31	0.13		0.86	1.64	0.78
A*1101	V134M	0.067	LVEICTEMEK	LMEICTEMEK	0.48	0.48	-0.01		2.67	1.03	-1.65
B*1501	V134M	0.067	ALVEICTEM	ALMEICTEM	0.45	0.54	0.08		2.66	4.56	1.90
B*1503	V134M	0.067	ALVEICTEM	ALMEICTEM	0.61	0.74	0.13		2.66	4.56	1.90
B*4802	V134M	0.067	ALVEICTEM	ALMEICTEM	0.48	0.62	0.14		2.66	4.56	1.90
A*0201	T138A	0.166	ALVEICTEM	ALVEICAEM	0.64	0.67	0.03		5.28	6.85	1.57
A*0203	T138A	0.166	ALVEICTEM	ALVEICAEM	0.73	0.75	0.02		5.28	6.85	1.57
A*1101	T138A	0.166	LVEICTEMEK	LVEICAEMEK	0.48	0.45	-0.03		2.67	1.69	-0.98
A*0201	M140L	0.527	ALVEICTEM	ALVEICTEL	0.64	0.78	0.14		5.28	9.34	4.06
A*0203	M140L	0.527	ALVEICTEM	ALVEICTEL	0.73	0.82	0.09		5.28	9.34	4.06
A*3201	M140L	0.527	KALVEICTEM	KALVEICTEL	0.35	0.40	0.04		5.49	4.05	-1.44
B*1501	M140L	0.527	ALVEICTEM	ALVEICTEL	0.45	0.28	-0.17	-T	2.66	1.82	-0.84
B*1503	M140L	0.527	ALVEICTEM	ALVEICTEL	0.61	0.44	-0.17		2.66	1.82	-0.84
B*4802	M140L	0.527	ALVEICTEM	ALVEICTEL	0.48	0.35	-0.13	-T	2.66	1.82	-0.84
B*5801	M140L	0.527	KALVEICTEM	KALVEICTEL	0.47	0.42	-0.05	-T	5.49	4.05	-1.44
A*1101	K142E	0.125	LVEICTEMEK	LVEICTEMEE	0.48	0.04	-0.44	-T	2.67	0.46	-2.21
B*2705	K169R	0.218	TKWRKLVDF	TRWRKLVDF	0.17	0.49	0.32	+T	0.45	1.61	1.16
A*0301	L173V	0.135	KLVDFRELNK	KVVDFRELNK	0.65	0.60	-0.05		4.03	3.29	-0.74
A*1101	L173V	0.135	KLVDFRELNK	KVVDFRELNK	0.67	0.75	0.07		2.99	7.58	4.59
A*0203	V207I	0.075	GLKKKKSVTV	GLKKKKSVTI	0.47	0.33	-0.14	-T	2.04	1.20	-0.84
B*1501	K221E	0.511	FSVPLDKDF	FSVPLDEDF	0.34	0.37	0.03		1.57	2.76	1.19
B*1503	K221E	0.511	FSVPLDKDF	FSVPLDEDF	0.52	0.54	0.02		1.57	2.76	1.19
B*3501	K221E	0.511	VPLDKDFRKY	VPLDEDFRKY	0.57	0.65	0.08		2.03	2.57	0.55
B*4802	K221E	0.511	FSVPLDKDF	FSVPLDEDF	0.41	0.47	0.06	+T	1.57	2.76	1.19
B*1501	D222E	0.307	FSVPLDKDF	FSVPLDKEF	0.34	0.45	0.10	+T	1.57	2.40	0.82
B*1503	D222E	0.307	FSVPLDKDF	FSVPLDKEF	0.52	0.63	0.11		1.57	2.40	0.82
B*3501	D222E	0.307	FSVPLDKDF	FSVPLDKEF	0.30	0.43	0.13	+T	1.20	1.58	0.38
B*4802	D222E	0.307	FSVPLDKDF	FSVPLDKEF	0.41	0.51	0.10	+T	1.57	2.40	0.82
A*0203	I234T	0.303	SINNETPGI	STNNETPGI	0.54	0.29	-0.25	-T	1.61	0.87	-0.73
A*1101	I234T	0.303	SINNETPGIR	STNNETPGIR	0.36	0.44	0.08	+T	1.20	2.22	1.01
A*2301	I234T	0.303	KYTAFTIPSI	KYTAFTIPST	0.73	0.26	-0.47	-T	4.91	0.91	-4.00
A*2402	I234T	0.303	KYTAFTIPSI	KYTAFTIPST	0.67	0.18	-0.48	-T	4.91	0.91	-4.00
B*1503	I234T	0.303	INNETPGIRY	TNNETPGIRY	0.53	0.38	-0.14	-T	1.82	0.90	-0.92
B*4802	I234T	0.303	INNETPGIRY	TNNETPGIRY	0.44	0.32	-0.12	-T	1.82	0.90	-0.92
A*0201	I241V	0.098	SINNETPGI	SINNETPGV	0.31	0.47	0.17	+T	1.61	3.01	1.40
A*0203	I241V	0.098	SINNETPGI	SINNETPGV	0.54	0.70	0.16		1.61	3.01	1.40

Continued on the following page...

HLA	Mut	Freq	Pep	Mut	$IC50_{pep}$	$IC50_{mut}$	$IC50_{\Delta}$	T	o_{pep}	o_{mut}	o_{Δ}
A*0301	K265R	0.072	AIFQSSMTK	AIFQSSMTR	0.75	0.59	-0.16		8.02	1.68	-6.34
A*1101	K265R	0.072	AIFQSSMTK	AIFQSSMTR	0.81	0.71	-0.11		16.29	3.81	-12.47
A*1101	K265R	0.072	MTKILEPFRK	MTRILEPFRK	0.67	0.64	-0.03		1.57	0.85	-0.72
B*1501	K265R	0.072	FQSSMTKIL	FQSSMTRIL	0.34	0.41	0.07		2.07	3.19	1.13
B*1503	K265R	0.072	FQSSMTKIL	FQSSMTRIL	0.76	0.82	0.06		2.07	3.19	1.13
B*2705	K265R	0.072	TKILEPFRK	TRILEPFRK	0.18	0.50	0.32	+T	0.63	3.54	2.91
B*4802	K265R	0.072	FQSSMTKIL	FQSSMTRIL	0.70	0.76	0.06		2.07	3.19	1.13
B*1501	V278I	0.102	KQNPDIVIY	KQNPDIIIY	0.54	0.51	-0.03		20.99	19.71	-1.28
B*1501	V278I	0.102	RKQNPDIVIY	RKQNPDIIIY	0.23	0.21	-0.01		2.35	1.93	-0.42
B*1501	V278I	0.102	VIYQYMDDLY	IIYQYMDDLY	0.41	0.46	0.05	+T	2.68	3.34	0.66
B*1503	V278I	0.102	KQNPDIVIY	KQNPDIIIY	0.83	0.82	-0.01		20.99	19.71	-1.28
B*1503	V278I	0.102	RKQNPDIVIY	RKQNPDIIIY	0.77	0.77	-0.01		2.35	1.93	-0.42
B*1503	V278I	0.102	VIYQYMDDLY	IIYQYMDDLY	0.56	0.62	0.06		2.68	3.34	0.66
B*4802	V278I	0.102	KQNPDIVIY	KQNPDIIIY	0.81	0.82	0.00		20.99	19.71	-1.28
B*4802	V278I	0.102	RKQNPDIVIY	RKQNPDIIIY	0.66	0.67	0.00		2.35	1.93	-0.42
B*4802	V278I	0.102	VIYQYMDDLY	IIYQYMDDLY	0.44	0.53	0.08		2.68	3.34	0.66
A*0101	Y280C	0.153	VIYQYMDDLY	VICQYMDDLY	0.38	0.35	-0.04		1.15	2.29	1.15
A*0201	Y280C	0.153	YQYMDDLYV	CQYMDDLYV	0.79	0.53	-0.26		2.73	0.95	-1.77
A*0203	Y280C	0.153	YQYMDDLYV	CQYMDDLYV	0.74	0.48	-0.26		2.73	0.95	-1.77
A*2901	Y280C	0.153	VIYQYMDDLY	VICQYMDDLY	0.74	0.63	-0.11		1.15	2.29	1.15
B*1501	Y280C	0.153	KQNPDIVIY	KQNPDIVIC	0.54	0.15	-0.39	-T	20.99	1.78	-19.21
B*1501	Y280C	0.153	RKQNPDIVIY	RKQNPDIVIC	0.23	0.03	-0.19		2.35	0.304	-2.05
B*1501	Y280C	0.153	VIYQYMDDLY	VICQYMDDLY	0.41	0.29	-0.12		2.68	2.43	-0.25
B*1503	Y280C	0.153	KQNPDIVIY	KQNPDIVIC	0.83	0.45	-0.38		20.99	1.78	-19.21
B*1503	Y280C	0.153	RKQNPDIVIY	RKQNPDIVIC	0.77	0.36	-0.41	-T	2.35	0.304	-2.05
B*1503	Y280C	0.153	KQNPDIVIYQ	KQNPDIVICQ	0.33	0.31	-0.03		1.53	2.51	0.98
B*1503	Y280C	0.153	VIYQYMDDLY	VICQYMDDLY	0.56	0.35	-0.21	-T	2.68	2.43	-0.25
B*4802	Y280C	0.153	KQNPDIVIY	KQNPDIVIC	0.81	0.45	-0.37		20.99	1.78	-19.21
B*4802	Y280C	0.153	RKQNPDIVIY	RKQNPDIVIC	0.66	0.30	-0.36	-T	2.35	0.304	-2.05
B*4802	Y280C	0.153	KQNPDIVIYQ	KQNPDIVICQ	0.34	0.32	-0.02		1.53	2.51	0.98
B*4802	Y280C	0.153	VIYQYMDDLY	VICQYMDDLY	0.44	0.30	-0.14	-T	2.68	2.43	-0.25
B*3501	M283V	0.496	NPDIVIYQYM	NPDIVIYQYV	0.58	0.30	-0.28	-T	1.71	0.68	-1.03
B*1501	E302K	0.095	GQHRTKIEEL	GQHRTKIKEL	0.25	0.20	-0.04		3.30	2.05	-1.24
B*1503	E302K	0.095	GQHRTKIEEL	GQHRTKIKEL	0.68	0.63	-0.05		3.30	2.05	-1.24
B*4802	E302K	0.095	GQHRTKIEEL	GQHRTKIKEL	0.60	0.51	-0.10		3.30	2.05	-1.24
B*5801	E302K	0.095	EELRQHLLRW	KELRQHLLRW	0.12	0.25	0.14		0.53	4.08	3.54
B*1501	Q306E	0.196	RQHLLRWGF	REHLLRWGF	0.44	0.21	-0.23	-T	4.27	0.42	-3.85
B*1503	Q306E	0.196	RQHLLRWGF	REHLLRWGF	0.87	0.70	-0.17		4.27	0.42	-3.85
B*2705	Q306E	0.196	RQHLLRWGF	REHLLRWGF	0.60	0.36	-0.24	-T	2.62	0.52	-2.10
B*4802	Q306E	0.196	RQHLLRWGF	REHLLRWGF	0.80	0.66	-0.14		4.27	0.42	-3.85
A*0201	H307Y	0.158	HLLRWGFTT	YLLRWGFTT	0.48	0.76	0.28		1.00	3.52	2.51
A*0201	H307Y	0.158	HLLRWGFTTP	YLLRWGFTTP	0.18	0.41	0.23		0.71	2.00	1.29
A*0203	H307Y	0.158	HLLRWGFTT	YLLRWGFTT	0.33	0.51	0.17	+T	1.00	3.52	2.51
A*0203	H307Y	0.158	HLLRWGFTTP	YLLRWGFTTP	0.18	0.32	0.14		0.71	2.00	1.29
B*1501	H307Y	0.158	RTKIEELRQH	RTKIEELRQY	0.10	0.38	0.28		1.06	4.91	3.85

Continued on the following page...

HLA	Mut	Freq	Pep	Mut	$IC50_{pep}$	$IC50_{mut}$	$IC50_{\Delta}$	T	o_{pep}	o_{mut}	o_{Δ}
B*1503	H307Y	0.158	RTKIEELRQH	RTKIEELRQY	0.15	0.41	0.26		1.06	4.91	3.85
B*2705	H307Y	0.158	RQHLLRWGF	RQYLLRWGF	0.60	0.67	0.07		2.62	2.09	-0.53
B*4802	H307Y	0.158	RTKIEELRQH	RTKIEELRQY	0.09	0.27	0.18		1.06	4.91	3.85
A*3201	L309W	0.416	KIEELRQHLL	KIEELRQHLW	0.31	0.48	0.17	+T	0.403	3.29	2.89
B*1501	L309W	0.416	RQHLLRWGF	RQHLWRWGF	0.44	0.45	0.01		4.27	5.74	1.47
B*1503	L309W	0.416	RQHLLRWGF	RQHLWRWGF	0.87	0.89	0.02		4.27	5.74	1.47
B*2705	L309W	0.416	RQHLLRWGF	RQHLWRWGF	0.60	0.62	0.02		2.62	3.13	0.50
B*4802	L309W	0.416	RQHLLRWGF	RQHLWRWGF	0.80	0.81	0.01		4.27	5.74	1.47
B*5801	L309W	0.416	KIEELRQHLL	KIEELRQHLW	0.16	0.60	0.44	+T	0.403	3.29	2.89
B*2705	R310K	0.503	RQHLLRWGF	RQHLLKWGF	0.60	0.50	-0.10		2.62	3.16	0.53
B*1501	T314F	0.129	RQHLLRWGFT	RQHLLRWGFF	0.12	0.46	0.33	+T	0.393	3.25	2.85
B*1503	T314F	0.129	RQHLLRWGFT	RQHLLRWGFF	0.48	0.84	0.36		0.393	3.25	2.85
B*2705	T314F	0.129	RQHLLRWGFT	RQHLLRWGFF	0.34	0.50	0.16	+T	1.02	2.71	1.69
B*4802	T314F	0.129	RQHLLRWGFT	RQHLLRWGFF	0.38	0.70	0.32	+T	0.393	3.25	2.85
B*1501	T314Y	0.505	RQHLLRWGFT	RQHLLRWGFY	0.12	0.45	0.32	+T	0.393	6.16	5.77
B*1503	T314Y	0.505	RQHLLRWGFT	RQHLLRWGFY	0.48	0.77	0.29		0.393	6.16	5.77
B*2705	T314Y	0.505	RQHLLRWGFT	RQHLLRWGFY	0.34	0.47	0.12	+T	1.02	1.94	0.91
B*4802	T314Y	0.505	RQHLLRWGFT	RQHLLRWGFY	0.38	0.68	0.30	+T	0.393	6.16	5.77

4.2.3 ARV resistance mutations that demonstrate higher MHC affinity and stability are diminished in certain HLA types

To establish a causal relationship between the diminished substitutions listed in Table 4.6 and HLA types, Table 4.9 was consulted and the prediction changes for these residues, where significant, were scrutinized. It was investigated which of the diminished substitutions listed in Table 4.6 had an impact on the prediction scores. The substitution M46I resulted in reduced affinity of a known epitope for B*15 spanning the PR region 45-53. This held true for the HLA allotype B*4802. The mutation I54V has a predicted decrease in affinity. The PR V82A mutation showed a marginal increase in predicted affinity and stability for B*0702. The I84V mutation resulted in a predicted increase in affinity and a corresponding increase in affinity and stability for A*0201/3. The mutation L90M that is very frequently associated with PI resistance (observed in 42.3% of patients) had an increase in affinity and stability for various allotypes spanning the PR 89-97 region [Rhee *et al.*, 2003; Shafer, 2006b; Rhee *et al.*, 2007]. Increases in affinity were predicted for B*1501, B*1503, B*4802 as well as A*3201 and B*5801 when considering 10-mer peptides spanning PR 90-99. The RT M41L mutation (M140L) increased the affinity of the peptide spanning regions RT 33-41 for A*0201, A*3201 while decreasing affinity and stability for the same peptide presented on B*1501, B*1503 and B*4802. No significant changes in affinity were reported for either RT N67D (D165N) nor RT N70D (T168D).

Interestingly, the Y181C that was calculated to be diminished in the B*15 and B*48 HLA type sets was predicted to have a profoundly negative impact on both affinity and stability of the known B*1501 epitope spanning RT 171-179 (hereafter referred to as KQ9), including predictions for B*1503, and B*4802 . This result seemed paradoxical, since the Y181C mutation could also naturally provide a mechanism of CTL escape. The escape mutation for B*1501 is listed as RT D177E [Frahm *et al.*, 2006]. It is unclear whether this mutation interferes epistatically with Y181C, explaining possible lower levels. It was calculated that the mutations RT V174D and V174E were found to be enriched in the B*15 and B*48 sets. This result is not shown in Table 4.5, due to the fact that the mutations listed there were limited to positions where major ARV mutations occur. Both these mutations were predicted to abrogate binding and significantly reduce stability, suggesting that these two mutations may also be considered escape mutations for the KQ9 epitope.

The RT L210W (L309W) mutation had a moderate increase in affinity and stability predictions for B*1501, B*1503, B*4802 and A*3201 for a peptide spanning RT 206-214 (hereafter, RQ9). It is not known whether this is a true epitope of the aforementioned peptides, but it is known that mutations at RT Q207 are associated with HLA types B*48 and B*15. With all of these mutations predicted to severely lower the affinity of the peptides to B*1501, B*1503, B*4802, particularly the substitution RT Q207E (Q306E). Lastly, both mutations T215F and T215Y (RT T215F/Y) had increased prediction scores for both affinity and stability for the HLA allotypes B*1501, B*1503, B*4802 and B*2705. It was observed that the T215Y mutation had a higher impact than the T215F for all of the aforementioned HLA allotypes except B*2705 and B*4802. Taken together, these results may provide evidence of a mechanism for negative selection of some major ARV resistance mutations in patients that are positive for allotypes belonging to some of the listed HLA types.

4.3 Evidence of multiple diminished antiretroviral related *pol* substitutions in HLA type B*15 and B*48 assigned sequence sets

The MHC affinity and stability reductions provided some evidence of DRM-induced CTL epitopes that may result in selection against these DRMs. Short of experimental evidence further analysis is needed to support the case of HLA type-associated diminishing of certain DRMs. The two HLA types, B*48 and B*15, were considered for further analysis. The associated MHC affinity and stability increases due to mutations in PR and RT that also appeared to be diminished in both B*15 and B*48 sets warranted this. Interestingly, many of the prediction results for HLA allotypes belonging to each of these HLA types showed similar changes in MHC binding affinity due to DRMs. In particular, the diminished accounts of the substitutions L90M, L309W (RT L210W) and T314X (RT T314X, where ”X” indicates multiple substitutions) were worthy of investigation.

However, other mutations related to ARV resistance were also observed to be diminished. It is known that the PR L90M mutation correlates with various other mutations in PR. The same applies to RT T314Y (T215Y). Thus, it needed to be proved that these diminished substitutions were not either due to lower accumulation of DRMs in the B*15/B*48 population or that substitutions used for assignment of these HLA type groups did not interact negatively in an epistatic manner with other ARV-related substitutions, making it appear as if the DRMs were diminished. This section shows the results of FET on the mutations L90M, L309W and T314X after compensating for diminished correlated mutations. These results might provide further evidence of novel, ARV resistance DRM generated epitopes. In a recent publication by the author of this manuscript, evidence was provided for the diminished frequency of the PR mutation L90M in HLA types B*15 and B*48 [Smidt, 2013]. This mutation is often associated with higher HIV-1 viral loads in patients undergoing ARV therapy [Mackie *et al.*, 2010].

4.3.1 Diminished substitutions in B*48 and B*15 sets remain significant after compensating for diminished correlated mutations

It is noted in Table 4.6 that in addition to the mutations L90M, L309W, and T313X (abbreviated from here on as SDS, or Significant Diminished Substitutions), other mutations are also diminished in the B*15 and B*48 sets. However, because these other mutations did not result in potential increased immunogenicity through MHC affinity and/or stability predictions, they were not considered as a direct driving force behind the diminished listed mutations. Both the HIVdb and analyses performed for this study were used to find highly correlated mutations.

Two approaches were used to accomplish this task:

1. *Quantify the contribution of epistatically-linked substitutions to SDS frequencies by the use of a linear model* - this approach attempted to predict the expected frequency of an SDS residue, given frequencies of correlated mutations as input.

2. *A non-parametric method involving the stabilization of the SDS correlated mutation frequencies between a HLA type-positive and HLA type-negative patient sequence sets* - this approach ensured that SDS-correlated mutations do not differ significantly in frequency between the HLA type-pos-tive and HLA type-negative sequence sets. Variation in SDS after this ”coercion” may strongly indicate alternative selection pressures that may be due to the imputed HLA type.

First, it was investigated whether any of the substitutions used for assigning HLA types to patients were diminished in sequences from treatment-experienced patients. These results are presented in Table 4.10. It is noted that the R310G (RT R211G) is diminished in treatment-experienced patients and interacts negatively with RT L210W. This mutation is used in assignment of B*15 and A*32. In subsequent analyses, this mutation was sometimes omitted to account for its negative correlation with RT L210W. It was not excluded when measuring levels of L90M or other PR mutations.

Table 4.10: This table represents residues that are both associated with HLA types and diminished in PR-RT sequences obtained from treatment-experienced patients. The results are represented in a similar fashion to the other FET results, with the difference that q-values are omitted and the final column representing published data of substitution-HLA type association. The substitution information is read as HLA type/Adapted or Nonadapted /Direct or Indirect/$log_1 0$ p-value/$log_1 0$ q-value. Adapted means that the substitution provides CTL escape, whereas *Non-adapted* means the substitution is diminished in this HLA type. *Direct* means a direct adaptation to the HLA type's epitope, whereas *Indirect* means that this mutation is strongly correlated with a direct mutation. The p and q-values are in $log_1 0$ format to make it easier to read. Typically, p-values of < -2 and q-values < -1 are considered strong indicators of association.

Pos	Org	AA	S1	$S1'$	S2	$S2'$	Odds	p-value	OddsP	OddsN	dOdds	q-value	HLA type Association
11	T	A	36	1927	73	1509	-2.59	0.0	0.02	0.05	-0.03		B52/A/D/-10/-8 C01/A/D/-3/-1
11	T	S	57	1906	85	1497	-1.90	0.00027	0.03	0.05	-0.02		**B51**/A/I/-3/-1
13	K	R	210	1694	224	1353	-1.34	0.00535	0.11	0.14	-0.03		**A31**/N/D/-3/-1 **B51**/A/D/-13/-10
14	I	V	418	1510	425	1172	-1.31	0.00065	0.22	0.27	-0.05		**B51**/A/D/-6/-4 **C06**/N/I/-3/-1
36	N	S	212	1724	317	1307	-1.97	0.0	0.11	0.20	-0.09		**B44**/A/D/-4/-2
38	P	S	37	1926	62	1561	-2.07	0.00047	0.02	0.04	-0.02		**C14**/A/D/-3/-1
62	P	A	57	1894	211	1390	-5.04	0.0	0.03	0.13	-0.10		**C04**/A/D/-3/-1
62	P	S	71	1880	103	1498	-1.82	0.00016	0.04	0.06	-0.03		**B13**/A/D/-7/-4
62	P	T	66	1885	144	1457	-2.82	0.0	0.03	0.09	-0.06		**B13**/A/I/-6/-3
104	E	D	79	1847	107	1486	-1.68	0.00063	0.04	0.07	-0.03		**B40**/A/D/-6/-3
106	V	I	10	1937	27	1596	-3.28	0.00078	0.01	0.02	-0.01		**B57**/A/D/-3/-1
202	K	R	31	1908	56	1557	-2.21	0.00042	0.02	0.03	-0.02		**A68**/A/D/-4/-2
233	I	V	162	1771	185	1415	-1.43	0.00176	0.08	0.12	-0.03		**B52**/A/D/-3/-1
260	S	C	254	1674	427	1198	-2.35	0.0	0.13	0.26	-0.13		**B07**/A/D/-13/-10
298	T	I	75	1818	170	1433	-2.87	0.0	0.04	0.11	-0.07		**B40**/A/D/-5/-3 **B41**/A/D/-6/-4
309	R	G	43	1869	119	1459	-3.54	0.0	0.02	0.08	-0.05		**A32**/A/I/-3/-1 **B15**/A/I/-4/-2

4.3.2 Compensation for diminished epistatic interactions of mutations to L90M by use of a linear model

The major epistatically-linked mutations to PR L90M were found to be PR M46I, I54V, V71A, V82A and I84V. The impact of the frequency of each mutation on the frequency of L90M was measured by using a patient sequence set devoid of each mutation and steadily populating that set with the excluded mutation. Taking M46I as an example, patients that contained HIV-1 protease sequences with the M46I mutation were filtered out. The patients in this ”clean slate” set were increasingly replaced by patients that have HIV-1 protease sequences with the M46I mutation using a 5% replacement per iteration. At each level of M46I sequence inclusion, the frequencies of each of the other mutations, I54V, V71A, V82A, I84V and L90M were recorded. Finally, M46I, V71A, V82A and I84V were used as variables measuring the L90M frequencies. Thus, a linear model was constructed measuring the expected frequency of L90M in a sequence set, given the frequencies of the other mutations. The model was constructed from resampled sequences, ensuring the model remained robust. Refer to Section 3.6.1 on page 53 for further details. The purpose of linear model construction was also to determine whether there is a linear relationship between the frequencies of these mutations to L90M. The produced linear models showed high R^2 values, indicating the validity of using a linear model for expected L90M frequency prediction.

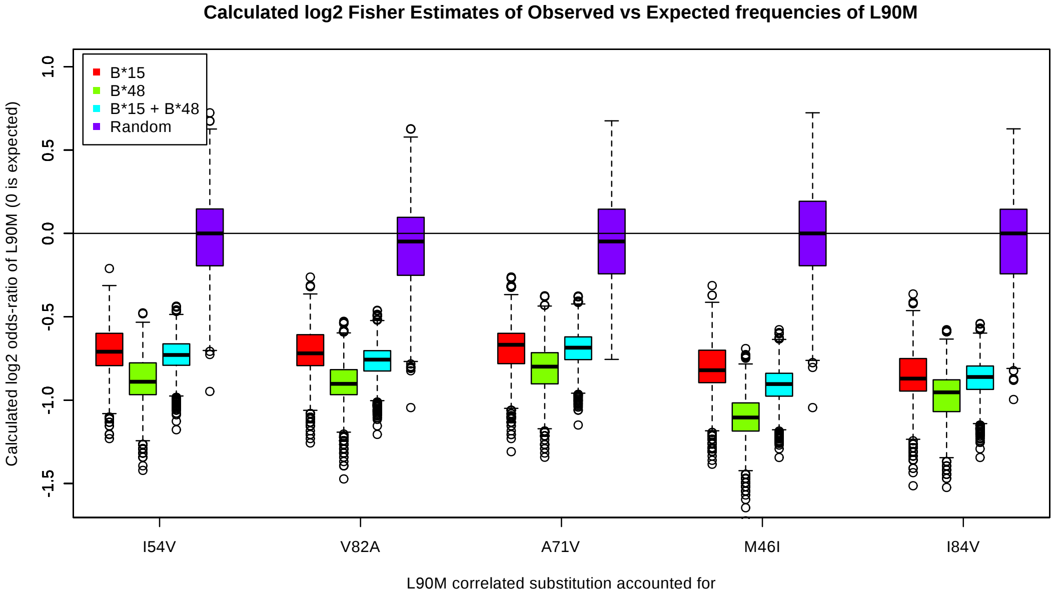


Figure 4.2: This figure is taken from an article by the author of this manuscript [Smidt, 2013]. It represents box-and-whiskers plots of the log_2 odds-ratios of the observed frequencies of L90M versus the expected frequencies of L90M in HLA types B*15, B*48 and a combination of the sets. Values below zero indicate a diminished observed frequency, while values above indicate enrichement. These values were calculated with the linear model mention in the text. The mutations listed on the x-axis are the mutations varied to train the linear model. All but the purple boxplot showed diminished residues. The purple boxes represent the log_2 odds-ratio of sequence sets obtained by random sampling.

The linear models were used to calculate an expected frequency of L90M, given the frequencies of L90M-correlated mutations. Because each model was constructed by varying the frequencies of a single substitution in the training set, all models were used to determine if any one of them resulted in rendering the FET results for L90M between the HLA type-postive and HLA type-negative sets non-significant. Figure 4.2 depicts the log_2-odds of FET for L90M between the HLA type-postive and HLA-type negative patient sequence sets for B*15, B*48 and B*15/B*48 combined as well as data from a random sequence set. In all cases, the L90M mutation still remained significantly diminished. To ensure a robust analysis, the FET was performed on a resampled 0.8 fraction of the patients for both the HLA type-postive and HLA type-negative sets. The result is a range of calculated odds-ratios. From Figure 4.2, it is clear that for all models that used different substitutions as basis for generating model data, the L90M values were siginificantly decreased in B*15, B*48 and B*15/B*48 sets, with the exception of the random control, which was expected to have a log_2 odds-ratio for L90M centered around zero.

The linear model was not reconstructed for detection of consistent diminishing mutations in RT. This was partly due to the complex nature of RT interactions involving T215X with L210W and M41L. For this, a more robust method was needed.

4.3.3 Compensation for diminished epistatic interactions of mutations to L90M by use of a robust sampling model

The non-parametric procedure focusing on the stabilization of SDS-correlated mutations involved coercing both the HLA type-positive and HLA type-negative sets to contain an equal proportion of correlated mutations, except the SDS residues. Because the HLA type-postive set is significantly smaller than the HLA type-negative sets, the SDS correlated mutations' frequencies were adjusted to match the frequencies of the SDS residues in the B*48 and B*15 type-positive sets. This ensured lower variance when resampling procedures were applied.

The results are represented graphically in Figure 4.3 on page 99. For type B*15 (Figure 4.3a) the odds-ratio of L90M and T314Y (RT T215F) versus the B*15 negative set still remained negative. The mutation T314F was enriched. In Figure 4.3b, the mutation R310G (RT R211G) was removed from classification to compensate for possible epistatic effects. Although the L90M and T314Y mutations were less significantly diminished in this case, the T314 (RT T215F) mutation was further enriched. Similar results are observed for the B*48 set, with L90M and T314Y (RT T215Y) being slightly more diminished, with T314F only being slightly enriched to the point where it can be considered non-significant.

Figure 4.3d demonstrates the effect of combining the B*48 and B*15-positive sets and resulted in an expected average of the results obtained for each HLA type individually. The results of the same test on two multiply sampled random sets are shown in Figure 4.4. As is shown, the mean odds ratios are centered around zero. Although the variances are high, it is clear that the results shown in Figure 4.3 differ significantly from the random case.

It should be noted that while initially, L90M and T314Y were diminished in sequence sets for HLA types A*29, A*32, B*35 and B*39, it did not remain significant after compensating for SDS correlated mutations.

4.3.4 Enriched substitutions in B*15 and B*48 sets may reveal mechanisms of escape for predicted epitopes

To further investigate if more evidence could be found for the existence of epitopes generated by the mutations L90M, T314F and T314Y, it was investigated whether there are mutations enriched in the B*15 and B*48 sets that could possibly indicate mechanisms of CTL epitope escape. Mentioned in Section 1.2.2 on page 20, CTL escape can be achieved by substitutions affecting the processing, presentation and recognition of a CTL epitope or a combination of these. Thus, the effects of mutations in and around these CTL epitope regions on proteasomal cleavage and MHC stability and affinity were analyzed. For proteasomal cleavage predictions, creation of a proteasomal cleavage site internal to the CTL epitope could severely impact how many of these epitopes finally exist. The same applies to negative impacts on MHC affinity, which can lower the amount of presented epitope, if at all. Stability affects the longevity of the presented peptide and naturally, a lower stability will make time window for recognition of the epitope smaller and decrease the encounter with an appropriate T-cell receptor. Immunogenicity is difficult to predict and the most recent development of a tool did not provide enough discriminatory power to be included as a predictor of immunogenicity in this study [Calis *et al.*, 2013].

First, it was determined which residues were enriched in B*15/B*48 sets. In order to improve the power of detecting enriched residues, the B*15 and B*48 sets were combined and only patients with L90M, the diminished residue mentioned earlier were included for comparison. This mutation is generally a good indicator of acquisition of drug resistance, owing to the fact that it is one of the most enriched mutations when comparing HIV PR and RT sequences from drug naïve with treatment-experienced patients. The other mutation that could also be considered, is D165N (RT D67N). The results are shown in Figure 4.5 on page 100. The results are limited to residues occurring within the predicted epitope regions. For PR, there was an increase in substitutions at position 91. The Q91K mutation is a mutation linked with escape of the B*15 CTL epitope spanning 90-98, `TQIGCTLNF` (TQ9). It is unknown whether Q91R is also an escape mutation. In RT, there is an increased substitution rate of Q306 (RT Q207) with the substitutions Q306K and Q306A being higher. It is noted that Q306H is also higher, but this mutation was used in imputation of B*15 patients and thus not considered here. Naturally, the conservation of Q306, illustrated as Q306Q in Figure 4.5, was lower, due to the higher amount of substitution at this position. It is interesting to note that the mutation, T314F remained higher as well as the mutation T314I, which wasn't observed before.

Effects of enriched residues on proteasomal cleavage prediction

First, the effects of enriched residues in the PR 89-97 region on predicted proteasomal cleavage scores were measured. The results for the PR region spanning 89-99 are shown in Table 4.11a on page 101. NetChop 3.0 was used as proteasomal predictor, using 0.500 as the threshold for a positive cleavage site (see Section 3.5.2 on page 51. For the consensus sequence (The Q92I93L) column, probable cleavage sites are at positions 93, 97 and 99. In particular the very high score obtained for 97 would allude to the proper cleavage of the predicted epitope spanning PR 89-97. The I93L mutation is a common *pol* ymorphism and also strongly associated with escape of the epitope TQ9. Reports of the mechanism behind this escape are conflicting in the literature [Mueller *et al.*, 2007; Bhattacharya *et al.*, 2007] [1]. A report by Mueller *et al.* states that the mechanism could be a decrease in recognition of the epitope after the I93L substitution [Mueller *et al.*, 2007]. However, in another report by Bhattacharya *et al.*, it seems that in the majority of patients, CTL

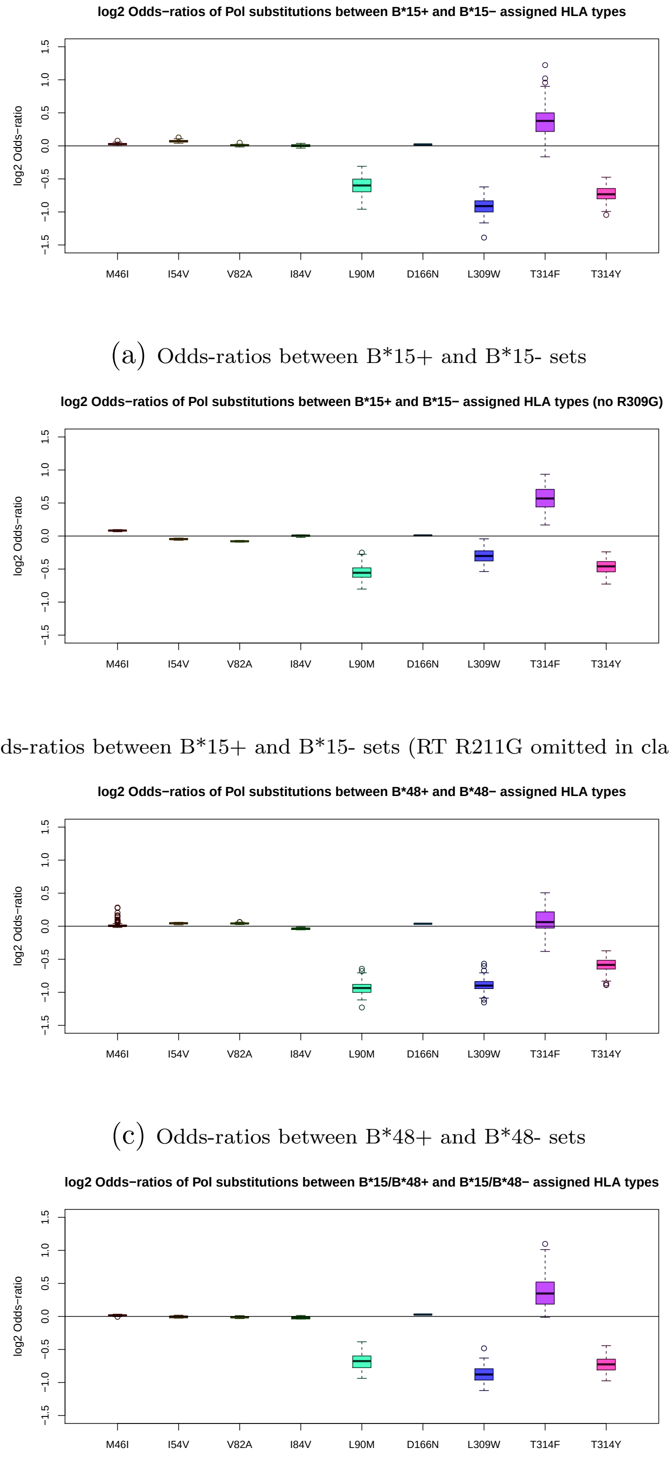
[1] See the supplementary section of [Bhattacharya *et al.*, 2007] describing the immunogenicity of TQ9.

recognition is enhanced by this mutation [Bhattacharya *et al.*, 2007]. When examining the results in Table 4.11a, an alternative explanation might be an increased probability of a proteasomal cleavage site within the epitope, since the prediction scores of position 93 with isoleucine has a score of 0.611, while the I93L substitution increases this score to 0.894. The Q92K mutation also appears to increase the probability of proteasomal cleavage with the K92 mutation increasing the predicted score of a cleavage site at position 92 from 0.090 (Q92) to 0.549. A further increase is predicted for the coexistence of the substitutions Q92K and I93L. However, it was calculated that Q92K is negatively correlated with the substitution Q93L. It is unknown whether a lower fitness is the result of these two mutations coexisting.

The mutations in Q306 (RT Q207) did not have a measurable effect on proteasomal cleavage sites within the predicted epitope spanning 305-314 (RT 206-215, RQ10), but the mutations around 314 (RT 215) all had a positive impact on predicted proteasomal cleavage score. The mutations T314I, T314F and T314Y increased the proteasomal prediction from 0.14 to 0.45, 0.76 and 0.75. As mentioned in a previous section, both the T314Y/F mutations resulted in increased MHC affinity and stability predictions for HLA B*1501, B*1503 and B*4802, and the potential creation of an appropriate proteasomal cleavage site does seem appropriate for an epitope to occur in this RT region. Furthermore, the T314I mutation does improve the proteasomal cleavage prediction score, but fails to breach the threshold. It is uncertain if this is a mechanism of escape.

Effect of enriched residues on MHC affinity and stability predictions

The effect of the enriched substitutions within 89-97 and 305-314 (RT 206-315) on the MHC affinity and stability of potential novel epitopes spanning these regions were calculated the same way as demonstrated in Section 4.2.3 on page 88. The results of which are presented in Table 4.12. The results are limited to significant changes observed. No significant changes were observed for the PR Q91K/R substitutions for the potential novel epitope `LMTQIGCTL` (LM9), lest for a minor decrease in predicted stability due to the Q91K substitution (not shown). For the RT 206-315 potential epitope, `RQHLLLRWGF[Y/F]` all the substitutions involved (including those not explicitly stated as enriched) were used to determine the MHC affinity and stability changes due to these mutations. For all these mutations, there were significant predicted decreases in MHC affinity and stability. Again, the stability predictions for B*1503 and B*4802 were extrapolated from B*1501. All the substitutions listed caused dramatic drops in predicted affinity and stability, supporting the notion that these mutations may all be regarded as potential escape mutations.



(a) Odds-ratios between B*15+ and B*15- sets

(b) Odds-ratios between B*15+ and B*15- sets (RT R211G omitted in classification)

(c) Odds-ratios between B*48+ and B*48- sets

(d) Odds-ratios between B*15/B*48+ and B*15/B*48- combined sets.

Figure 4.3: The above figures represent the log_2 odds-ratios of the listed mutations between the HLA type-positive and HLA type-negative sets after compensation for correlated mutations. The correlated mutations, M46I, I54V, V82A and I84V are shown together with the tested mutations L90M, L309W, T314F and T314Y. In figure a) the native B*15-positive set is compared with the B*15-negative set. In figure b), the same is done, but with the assigned B*15 type not including R310G as a selection factor. Figure c) show the results obtained for the B*48-type positive set and Figure d) shows the result by combining the B*15-positive and B*48-positive sets. Again, values above 0 indicate enriched residues, while values below 0 indicate diminished residues.

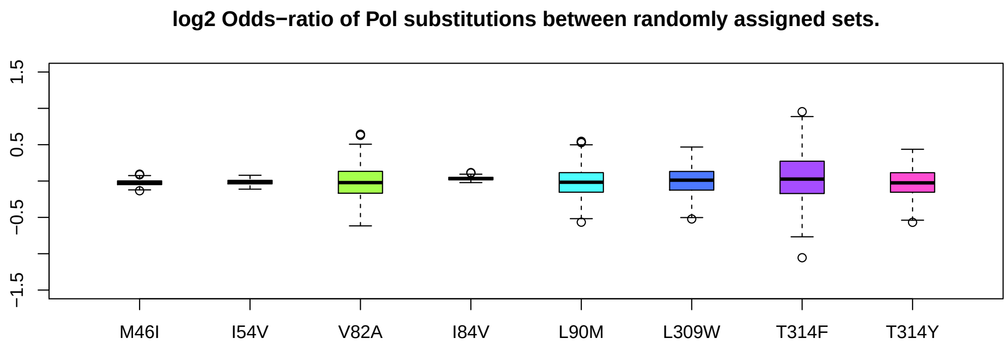


Figure 4.4: This figure depicts the odds-ratio of the residue frequencies between two random sets. The descriptions are the same as in Figure 4.3.

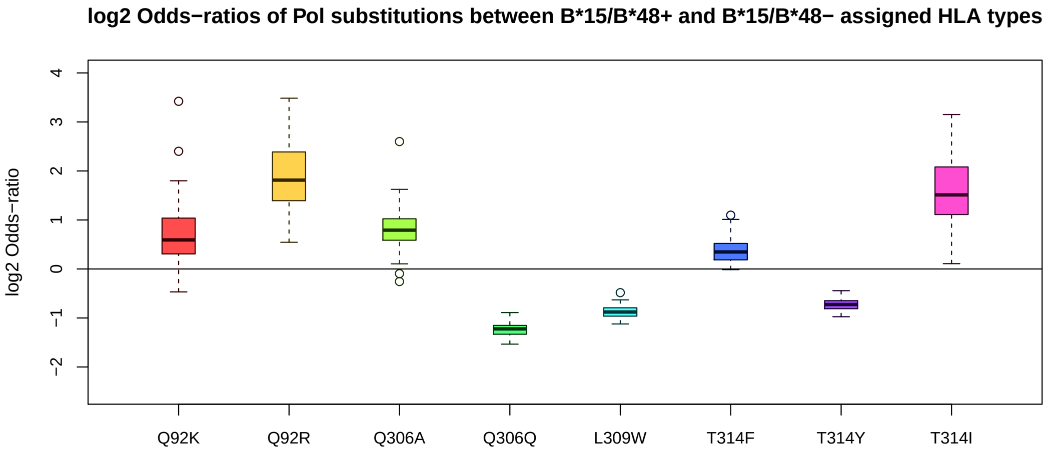


Figure 4.5: This figure represents the odds-ratios of substitutions within predicted epitope regions, PR 89-97 and 305-314 (RT 206-315). The ratios were measured between the B*15/B*48-positive set and B*15/B*48-negative set, with each set having sequences that contain the mutation L90M. The substitutions Q92K/R, Q306A/K and T314F/I were all enriched, while T314Y remained diminished.

Table 4.11: Changes in NetChop results due to DRM in PR and RT.

(a) The Q92I93 represents the results obtained from the consensus PR sequence of HIV-1 subtype B. Other columns show the scores obtained by substitutions at position 92 and 93. Using default parameters, above 0.500 indicate a predicted cleavage site.

Position	Q92I93	Q92L93	K92I93	K92L93	R92I93	R92L93
89	0.406	0.629	0.423	0.646	0.454	0.695
90	0.240	0.404	0.252	0.427	0.343	0.570
91	0.036	0.045	0.031	0.036	0.031	0.037
92	0.062	0.090	0.549	0.735	0.251	0.374
93	0.611	0.894	0.657	0.893	0.592	0.888
94	0.115	0.148	0.074	0.094	0.057	0.071
95	0.030	0.030	0.026	0.026	0.026	0.026
96	0.083	0.059	0.087	0.062	0.081	0.059
97	0.975	0.968	0.975	0.968	0.973	0.965
98	0.218	0.219	0.247	0.249	0.228	0.229
99	0.623	0.689	0.581	0.641	0.910	0.930

(b) This table depicts the proteasomal prediction scores obtained in the RT 215 region, comparing the scores obtained as a result of the RT T215Y/F substitutions versus scores obtained from the consensus sequence.

Pos	RT Pos	314T	314Y	314F	314I
304	203	0.93	0.93	0.93	0.93
305	204	0.93	0.93	0.93	0.93
306	205	0.80	0.80	0.80	0.80
307	206	0.81	0.81	0.81	0.81
308	207	0.22	0.22	0.22	0.22
309	208	0.97	0.97	0.97	0.97
310	209	0.83	0.83	0.83	0.83
311	210	0.95	0.98	0.96	0.97
312	211	0.12	0.14	0.24	0.12
313	212	0.96	0.95	0.89	0.91
314	213	0.14	0.76	0.75	0.45
315	214	0.45	0.42	0.35	0.34
316	215	0.15	0.16	0.15	0.13

Table 4.12: This table depicts the MHC stability and affinity score changes for the RQ10 peptide alternating between the RT T215F/Y/I mutations as well as the RT Q207E/H/K/A mutations. The results are represented in the same fashion as was in other tables depicting changes in MHC affinity and stability predictions.

HLA	Mut	Freq	Pep	Mut	$IC50_{pep}$	$IC50_{mut}$	$IC50_{\Delta}$	T	o_{pep}	o_{mut}	o_{Δ}
B*1501	**Q305H**	**0.008**	RQHLLRWGFF	RHHLLRWGFF	**0.46**	**0.13**	**-0.32**	**-T**	**3.25**	**0.303**	**-2.94**
B*1503	Q305H	0.008	RQHLLRWGFF	RHHLLRWGFF	0.84	0.60	**-0.24**		3.25	0.303	**-2.94**
B*4802	Q305H	0.008	RQHLLRWGFF	RHHLLRWGFF	0.70	0.42	**-0.28**		3.25	0.303	**-2.94**
B*1501	Q305A	0.017	RQHLLRWGFF	RAHLLRWGFF	0.46	0.30	**-0.15**	-T	3.25	0.62	**-2.63**
B*1503	Q305A	0.017	RQHLLRWGFF	RAHLLRWGFF	0.84	0.70	**-0.14**		3.25	0.62	**-2.63**
B*4802	Q305A	0.017	RQHLLRWGFF	RAHLLRWGFF	0.70	0.54	**-0.17**		3.25	0.62	**-2.63**
B*1501	Q305E	0.196	RQHLLRWGFF	REHLLRWGFF	0.46	0.22	**-0.24**	-T	3.25	0.409	**-2.84**
B*1503	Q305E	0.196	RQHLLRWGFF	REHLLRWGFF	0.84	0.64	**-0.20**		3.25	0.409	**-2.84**
B*4802	Q305E	0.196	RQHLLRWGFF	REHLLRWGFF	0.70	0.56	**-0.14**		3.25	0.409	**-2.84**
B*1501	Q305K	0.016	RQHLLRWGFF	RKHLLRWGFF	0.46	0.17	**-0.29**	-T	3.25	0.383	**-2.86**
B*1503	Q305K	0.016	RQHLLRWGFF	RKHLLRWGFF	0.84	0.71	**-0.13**		3.25	0.383	**-2.86**
B*4802	Q305K	0.016	RQHLLRWGFF	RKHLLRWGFF	0.70	0.51	**-0.20**		3.25	0.383	**-2.86**
B*1501	Q305H	0.008	RQHLLRWGFY	RHHLLRWGFY	0.45	0.17	**-0.28**	-T	6.16	0.368	**-5.79**
B*1503	Q305H	0.008	RQHLLRWGFY	RHHLLRWGFY	0.77	0.58	**-0.19**		6.16	0.368	**-5.79**
B*4802	Q305H	0.008	RQHLLRWGFY	RHHLLRWGFY	0.68	0.43	**-0.24**		6.16	0.368	**-5.79**
B*1501	Q305A	0.017	RQHLLRWGFY	RAHLLRWGFY	0.45	0.35	-0.09	-T	6.16	1.05	**-5.11**
B*1503	Q305A	0.017	RQHLLRWGFY	RAHLLRWGFY	0.77	0.66	**-0.11**		6.16	1.05	**-5.11**
B*4802	Q305A	0.017	RQHLLRWGFY	RAHLLRWGFY	0.68	0.54	**-0.14**		6.16	1.05	**-5.11**
B*1501	Q305E	0.196	RQHLLRWGFY	REHLLRWGFY	0.45	0.24	**-0.20**	-T	6.16	0.58	**-5.58**
B*1503	Q305E	0.196	RQHLLRWGFY	REHLLRWGFY	0.77	0.59	**-0.18**		6.16	0.58	**-5.58**
B*4802	Q305E	0.196	RQHLLRWGFY	REHLLRWGFY	0.68	0.56	**-0.12**		6.16	0.58	**-5.58**
B*1501	Q305K	0.016	RQHLLRWGFY	RKHLLRWGFY	0.45	0.17	**-0.27**	-T	6.16	0.51	**-5.66**
B*1503	Q305K	0.016	RQHLLRWGFY	RKHLLRWGFY	0.77	0.62	**-0.14**		6.16	0.51	**-5.66**
B*4802	Q305K	0.016	RQHLLRWGFY	RKHLLRWGFY	0.68	0.49	**-0.19**		6.16	0.51	**-5.66**
B*1503	Q305H	0.008	RQHLLRWGFI	RHHLLRWGFI	0.67	0.37	**-0.30**	-T	0.55	0.222	-0.33
B*4802	Q305H	0.008	RQHLLRWGFI	RHHLLRWGFI	0.54	0.28	**-0.27**	-T	0.55	0.222	-0.33
B*1503	Q305A	0.017	RQHLLRWGFI	RAHLLRWGFI	0.67	0.44	**-0.23**		0.55	0.269	-0.28
B*4802	Q305A	0.017	RQHLLRWGFI	RAHLLRWGFI	0.54	0.32	**-0.23**	-T	0.55	0.269	-0.28
B*1503	Q305E	0.196	RQHLLRWGFI	REHLLRWGFI	0.67	0.45	**-0.21**		0.55	0.241	-0.31
B*4802	Q305E	0.196	RQHLLRWGFI	REHLLRWGFI	0.54	0.42	**-0.12**		0.55	0.241	-0.31
B*1503	Q305K	0.016	RQHLLRWGFI	RKHLLRWGFI	0.67	0.45	**-0.21**		0.55	0.239	-0.31
B*4802	Q305K	0.016	RQHLLRWGFI	RKHLLRWGFI	0.54	0.32	**-0.22**	-T	0.55	0.239	-0.31

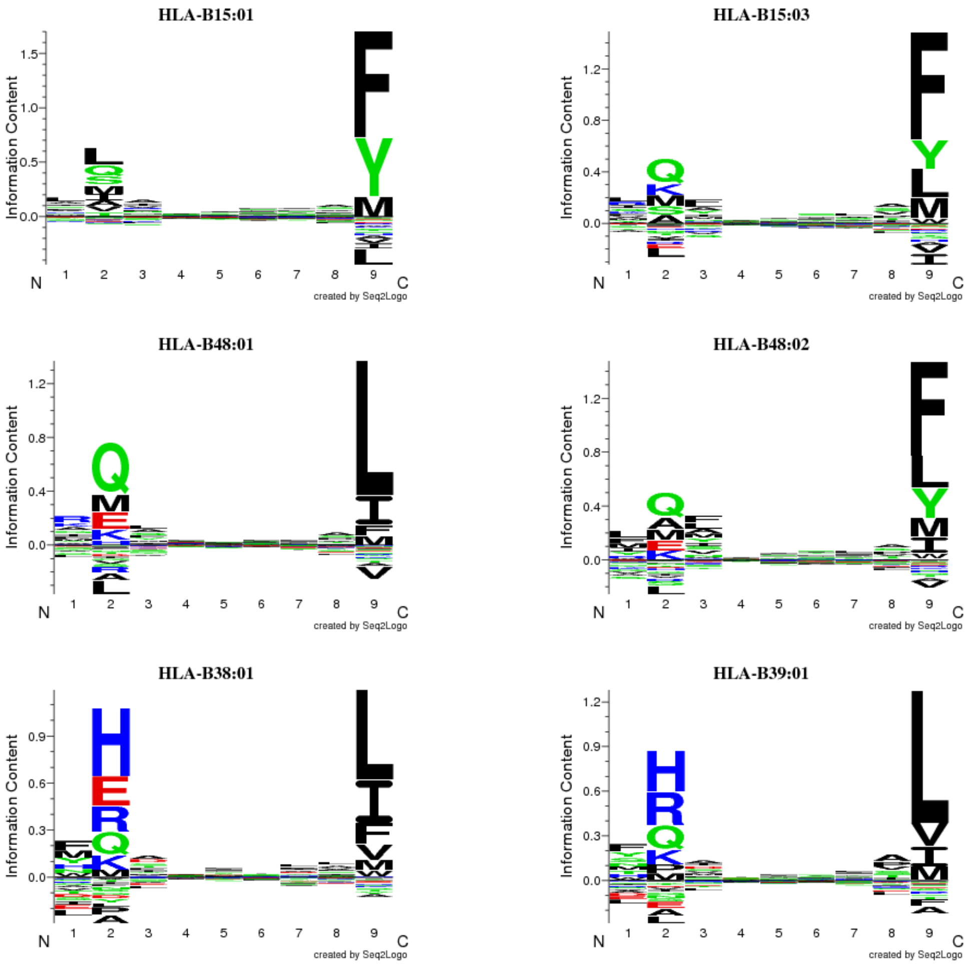
4.4 Binding motifs may reveal a mechanism for similar CTL epitope presentation of HLA types B*15 and B*48

Large similarities in predicted potential CTL epitopes existed between B*15 and B*48 related allotypes. The clustering performed for the allotypes B*1501, B*1503, B*4802 and B*4801 revealed a clustering pattern, with B*4802 and B*1503 being clustered closely together, although B*4801 was clustered with B*3801 and B*3901. Referring to Figure 4.6, the reason for the clustering becomes clearer. The SeqLogo plots (generated as part of the results of MHCClust) show residues that affect how well a peptide binds to the listed HLA allotype. The most important positions contributing to binding are 2 and 9. It becomes apparent why B*1503 is considered more similar to B*4802 than B*1501 in terms of the set of peptides that can be presented when examining position 2 and 9. Although Q, A and M are shared between all three allotypes as beneficial residues for peptide binding, a Leucine at position two is deemed unfavourable for both B*4802 and B*1503, while a Leucine at position 9 is not considered an anchor residue for B*1501. Considering the potential PR L90M induced epitope, `LMTQIGCTL`, a few factors need to be considered. Because the 9L residue is unfavourable, it would mean that this epitope is less likely to be presented by B*1501, however it is still very favourable for B*1503 and B*4802 and even B*4801. The LM9 epitope was run through another MHC affinity predictor, NetMHCCons [Karosiene *et al.*, 2012], using all the available HLA allotypes that fall under HLA type B*15 (n=189). Surprisingly, only 59/189 of the B*15 group were predicted to be non-binders of the LM9 epitope, while 102/189 were weak binders and 28/189 were predicted to be strong binders. For HLA allotypes of the B*48 suptypes, only 2/23 tested allotypes resulted in non-binding of LM9, while 13/23 were strong binders and 8/23 were weak binders of LM9. When examining the proposed escape mutations for the LM9 epitope, the Q92K mutation was thought to act as an escape mutation via the creation of a proteasomal cleavage site internal to the epitope (mentioned in the previous Chapter). It is of interest to note that this mutation is still a favourable anchor residue for B*1503, B*4802 and B*4801. Further investigation revealed that the mutation Q92R was enriched in the B*48 HLA type group and that this mutation is unfavourable for MHC binding in B*1503, B*4802 and B*4801, indicating a possible stronger selection pressure for this substitution. It should also be noted that the B*15 related escape mutation, I93L associated with the epitope spanning the 91-99 region, `TQIGCTLNF` is also enriched in the B*48 region. This further suggests similarity in the presented peptide repertoire between B*15 and B*48.

For the RT T215F/Y (T314F/Y) induced potential epitope, `RQHLLRWGF[F/Y]`, more similar increases in affinity for RQ10 was observed. From the SeqLogo data in Figure 4.6, it is clear that for all the HLA allotypes, except B*4801, the most preferred anchor residue at position 9, is Phe. Note that the calculation of 10-mer MHC binding is done through extrapolation, and position 10 would equate to the anchor residue at position 9 in the binding motif. The consensus residues have favourable residues both at position 2 and 10 (9). It was observed that Ile was enriched in the B*15/B*48 set. From the SeqLogos, it can be deduced that for the HLA allotypes B*1501 and B*1503, Ile is not a favourable position 10 (9) residue, Ile being below the 0 bit score. Interestingly, the T314I mutation is a more favourable residue for the B*4801 allotype than either F or Y. While the T314I mutation does improve binding affinity for all the HLA allotypes mentioned, MHC stability predictions for B*1501 are the lowest (0.55 hours) with this substitution, with T313F being higher (3.24 hours) and T314Y being the highest $t_{\frac{1}{2}}$ (5.26 hours). The T314F substitution is enriched over the T314Y substitution in the B*15 HLA type set, while T314Y/F are both diminished in the B*48 subset set. Future MHC experimental stability measurements might provide solid reasons for the selection of the respective residues and provide insights to the enrichment of the T314I and T314F substitutions.

4.5 The viability of the peptides LM9 and RQ10 as epitopes.

Although the results presented in this study do provide evidence for the existence of the LM9 and RQ10 epitopes, it is insufficient to absolutely confirm that they generate a CTL


HLA-B15:01

HLA-B15:03

HLA-B48:01

HLA-B48:02

HLA-B38:01

HLA-B39:01

Figure 4.6: The SeqLogos represent position specific residues that significantly contribute to affinity of a peptide to the HLA allotype described above each logo. The major contributions are clustered around positions 1-3 and 9. Resides that have a bitscore above 0 are generally associated with increasing the binding affinity of a peptide while those below severely decrease binding affinity and are considered non-favourable.

response. No information was found in the literature confirming the validity of LM9 as an epitope. For RQ10, there were conflicting reports concerning the T314F/Y induced epitope. In one study, it did appear that the T314Y mutation did not induce an immune response [Samri *et al.*, 2000]. The study mentioned that the epitope was tested for the serotype B62, to which many of the allotype members of B*15 belong. However, although the results do appear to reduce confidence in RQ10 as a CTL epitope, a few things need to be considered. First, even if the appropriate HLA allotype exists in a patient and even if an epitope is considered highly immunogenic, it may not always elicit a response. In the study by Samri, only one patient with B62 was tested. The possibility exists that this epitope was not recognized by the patient. Furthermore, it is unknown if the RT sequence obtained from the patient contained mutations that may alter the immunogenicity of RQ10. In the previous section, it was mentioned that not all members of B*15 were predicted to be binders of RQ10 [Frahm *et al.*, 2006]. It could very well be that the B62 member of the patient in the study could not present RQ10 and thus not elicit an immune response. Lastly, the study by Brumme *et al.* suggests that there are CTL escape substitutions in position 306 (RT 207) providing both direct and indirect escape mutations [Brumme *et al.*, 2009]. Indeed, the Q306E mutation is enriched in HIV RT sequences from treatment experienced patients, but there is no listed knowledge of its direct involvement in drug resistance. Unless it is as yet an unidentified functional covarying mutation, it suggests that another mechanism may be driving the selection of this mutation. The same applies to other mutations found at the same position, Q306H/R/H/D/A/S/T (RT 207) that are also enriched in the B*48 and B*15 HLA type sets. Other researchers have confirmed a shorter epitope, `RQHLLRWGF` and `RQHLLRWGL` (hereafter RQ9) elucidating a CTL response [Li *et al.*, 2011; Larke *et al.*, 2007]. The RQ10 epitope is one amino acid longer than the reference RQ9, but is still predicted to be a strong binder along with higher predicted MHC stability. Other known epitopes, such as the HIV-1 Gag A2 restricted epitope, `SLYNTVATL` has a variant `SLYNTVATLY` that is also highly immunogenic as well as the `LLDTGADDTV` and `LLDTGADDTVL` variants [Lorin *et al.*, 2005; Wang *et al.*, 2007].

Recently, a predictor was developed that examines a peptide for residues often associated with high immunogenicity [Calis *et al.*, 2013]. Although the tool has a moderate prediction accuracy (yielding an AUC value of 0.65), the predicted immunogenicity of the RQ10 and LM9 was calculated with this tool. The tool produces a score ranging from -1 to 1, with 0 being the threshold. The LM9 epitope produced a score slightly above the threshold of 0.03. Interestingly, the proposed CTL escape substitution, Q92K produced a score of -0.07, while an increase in immunogenicity was predicted for the Q92R substitution of 0.20. Granted that these scores are correct, it would mean that the likely mechanism of escape is still the creation of a proteasomal cleavage site internal to the epitope. It is noted that the TQ9 epitope that is known to be immunogenic, only produced a score of 0.05.

The RQ10 epitope produced a score of 0.28. The immunogenicity scores are in this tool unaffected by residues that are likely buried within the MHC binding groove and thus possibly don't interfere with the interaction of a complementary CTL T-cell Receptor. Interestingly, it was observed that the L309W mutation dramatically increased the immunogenicity prediction score to 0.50. By examining the frequency of L309W in the B*48/B*15 sets, this mutation was still diminished even after including only patient sequence sets containing the highly correlated T314Y mutation [Rhee *et al.*, 2007], resulting in a log_2 odds-ratio of -1.40 ($p < 0.003, q = 0.038$). This suggests that the L309W mutation could also be selected against, owing to the positive impact it has on predicted immunogenicity. The results are to be interpreted with caution, since the tool only has a moderate accuracy, but it is generally true of classification predictors that the results tend to be more accurate the further the score is from the threshold, which is 0.00 in this case [Eng, 2005].

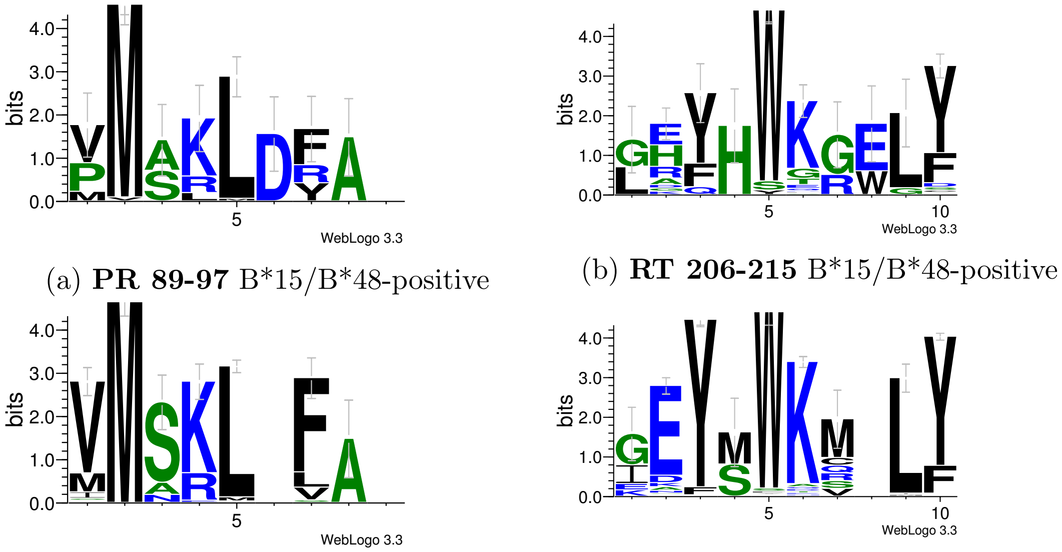
4.6 Evidence of higher selection pressure in the predicted epitope regions

Epitopes generated by antiretroviral mutations might also have an increase in sequence variability or additional substitutions within these epitope regions [Karlsson *et al.*, 2007]. The sequence variability needed to be measured in the context of the substitution that generates this response or use a marker that acts as a indicator of general antiretroviral resistance. Sequences from a total of 51 patients for the B*48/B*15-positive set and 803 patients from the B*48/B*15-negative set were used, i.e. patients harbouring the L90M and T314Y/F sequences respectively. Substitutions used for assignment of the HLA types as well as commonly associated mutations were excluded to reduce a positive-bias for the B*15/B*48-positive sets. Surprisingly, in the LM9 region, a total of 25 variants in the B*15/B*48 positive were found versus 61 Variants in the B*15/B*48-negative set. A total of 88 variants were found in the B*15/B*48-positive set of 88 patients for RQ10 versus 165 variants in the B*15/B*48-negative set of 1165 patients. This result was quite extraordinary, considering that the total pool of patients in the B*15/B*48-negative set were 15 times more. Using a unique sequence set per patient, SeqLogos were generated with WebLogo 3.0 (`http://weblogo.threepointone.org`). The logos are shown in Figure 4.7 on the next page. These logos represent the proportion of substitutions found in the LM9 and RQ10 regions respectively. In order to clearly identify substitutions, the consensus sequence residues spanning LM9 and RQ10 were excluded from the logo creation.

Interestingly, from the SeqLogos it can be deduced that regions from the LM9 and RQ10 regions in B*15/B*48-positive sets have more substitutions associated with them. Indeed, for LM9, three mutations, PR 89P, 92L, 94D and 95R/Y (posititions 1, 4, 6 and 7 in the Figure) are seen, although it should be mentioned that some substitutions seen in the B*15/B*48-negative group are not observed. The difference in substitutions in the RQ10 between the B*15/B*48-positive and -negative sets is more extreme. Indeed, substitutions 209H/Q (B*15/B*48-positive) versus 209M/S (B*15/B*48-negative, position 4 in the Figure) are in complete disagreement between the two sets. There are also a additional observed mutations 213E/W (position 8 in the Figure) that are not seen in the B*15/B*48-negative set. All this evidence points to selective pressures in these regions other than those caused by antiretroviral therapy. Therefore, although Samri *et al.* did not find an epitope induced by the mutation RT215Y/F, the evidence provided here does warrant further investigation of the existence of the predicted epitopes.

4.7 Concerning HIV subtype C

It was of interest whether this analysis can be applied to another HIV subtype. HIV subtype C contributes to the majority of HIV infections and is concentrated in sub-Saharan Africa. Although the available data from drug naïve patients infected with subtype C is comparable to data from subtype B infected individuals, subtype C data from treatment experienced patients is limited. Indeed, where there were 1,965 sequences available for subtype B from treatment experienced individuals, only 375 sequences were available for subtype C. Furthermore, many DRMs were found at lower levels, the frequencies of



(a) **PR 89-97** B*15/B*48-positive

(b) **RT 206-215** B*15/B*48-positive

(c) **PR 89-97** *15/B*48-*negative*

(d) **RT 206-215** B*15/B*48-*negative*

Figure 4.7: These figures represent SeqLogos generated for the regions PR 89-97 and RT 206-215. Figure a and c are the SeqLogos for the regions spanning LM9 and TQ10 in the B*15/B*48-positive sequence set, whereas the figures b and d are for the B*15/B*48 negative set. The scores are given as bit scores (Shannon Entropy).

L90M, T314F and T314Y were 9.7%, 9.2% and 1.2% in the subtype C set compared to the 13.0%, 51.1% and 43.5% in the subtype B set. As demonstrated, the HLA imputation method used here demonstrated low sensitivity. Combined with the lower frequency of major DRMs, using FET to infer HLA-associated diminished residues was challenging. The HLA-associated mutation set by Brumme *et al.* was constructed for HIV subtype B. Due to the nature of HIV subtype specific CTLs and that HLA associated mutations may be HIV subtype specific, another set was acquired to impute HLA types for subtype C data [Frahm *et al.*, 2006; Carlson *et al.*, 2014] [2].

HLA types were imputed for patients of the subtype C set using the same method as for subtype B, but with the different HLA associated mutations provided by Carlson *et al.*. Performance measurements are shown in Table 4.2. As with subtype B, the performance measurements in this case show a moderate PPV value with generally low sensitivity. After imputing the HLA types for patients in the treatment experienced HIV subtype B set, an FET was performed again also compensating for phylogenetic effects on the drug resistance mutations. The tree was constructed with FastTree using PR-RT sequences

[2] Please refer to the supplementary information in the article by Carlson *et al.* [Carlson *et al.*, 2014].

of HIV subtype C and masking positions along PR-RT where DRMs occur. A total of 1,767 PR-RT sequences that included sequences from treatment experienced patients, drug naïve patients and patients not annotated as either treatment experienced or drug naïve were used for the construction of the tree.

The breakdown of the imputed HLA types is shown in table 4.13. The performance values are shown in Table 4.14. As before, the HLA-assignments were used to compare substitution frequencies HLA-positive and HLA-negative groups. The results are shown in Table 4.15. It is evident from the results, considerably fewer significant results were obtained for HIV subtype C than HIV subtype B. This may be due to the lack of adequate sequence information as well as the general lack of accrual of certain DRMs, rather than an artifact of HIV subtype C itself. Still, certain residues were indeed observed to be diminished. Again, MHC affinity and stability were predicted for the consensus subtype C sequence as well as changes caused by DRMs. The results are shown in Table 4.16 on page 118.

Because of the smaller sample size, it was difficult to ascertain the significance of MHC affinity and stability changes in the selection of DRMs. The mutation, L90M noted earlier as diminished in the HLA B*15 assigned sets again showed an increase in binding affinity for HIV subtype C. The difference being that the consensus residue for HIV subtype C has a methionine at position 89 and a leucine at position 93 PR and yielding a predicted epitope with the sequence `MMTQLGCTL` rather than `LMTQIGCTL`. The FET results for subtype C showed that the counts were 0/36 for B*15-positive set versus 10/302 for the B*15-negative set, yielding a p-value of 0.60 and therefore could not be shown to be significantly diminished. However, a slight enrichment of the mutation M89L was observed, yielding a p-value of 0.14. While in itself not significant, it should be remembered that there is a correlation between the development of the mutation at PR 89 and emergence of L90M [Abecasis *et al.*, 2005]. Acquisition of more sequence data from subtype C patients may yield some clarity in the selection of L90M in B*15 patients infected with HIV subtype C undergoing treatment that include protease inhibitors. It is noted that the RT Q174R (Q273R) substitution is favoured for HLA types A*03 and B*15. The common *pol* ymorphism at this position is lysine. This *pol* ymorphism is predicted to create a novel CTL epitope for A*03, KILEPFRAK, while the ”R” mutation at the same position does not increase binding affinity and stability to above the prediction thresholds. While both the R and K mutations severely diminished the stability of another peptide presented by B*15, AQNPEIVIY, the ”R” mutation has a more dramatic negative impact on the predicted binding affinity of AQNPEIVIY. The importance of the Q174K mutation is that it commonly associated with the M184V mutation that provides antiretroviral resistance against lamivudine (3TC) and emtricitabine (FTC) [Doualla-Bell *et al.*, 2004]. It is unknown what the impact is of the ariginine mutation over lysine. Another mutation that is negatively associated with A*03 is V106M, which is associated with resistance against efavirenz (EFV) and nevirapine (NVP). However, this mutation did not yield any significant changes in MHC affinity or stability for A*03. Interestingly, it did yield an increase in predicted affinity for B*15. Since A*03 and B*15 share a substitution for HLA assignment, it will be worthy of investigating the effect of this mutation on MHC affinity and stability in HLA B*15. Lastly, the mutation G190A (G289A), which provides resistance against efavirenz (EFV) and nevirapine (NVP) was diminished in B*18 [Uhlmann *et al.*, 2004]. However, no immunological association in terms of MHC affinity or stability could be found. The mutation M184V (M283V) was observed to be decreased in the HLA B*07 and B*35 set. No explanation could be provided for the B*35 set, but the mutation T165I (T264I) was used for B*07 assignment. It is observed that this mutation is negatively correlated with M184V (M283V), yielding a phi-correlation value of -0.157. While in itself, this mutation is almost inconsequential in terms of providing resistance to emtricitabine (FTC) and tenofovir (TDF), the mutation does decrease the overall fitness of HIV [Kulkarni *et al.*, 2012].

While the analysis of HLA-driven selection of the DRMs in HIV subtype C PR-RT region do not provide as convincing results as obtained for HIV subtype C as for subtype B, largely due to the smaller sample size and lower accumulation of DRMs in the sequence set, some of the results do indicate a possible influence of HLA type on DRM selection. Further data may provide better insight to the HLA-DRM interactions in HIV subtype C.

Table 4.13: This table represents the breakdown of the sample sizes of HLA-assigned patients from which HIV-1 subtype C *pol* sequences were obtained. All the sequences were annotated as from patients that are treatment-experienced. The first column illustrates the HLA type, the second column indicates the number of assigned patients to that HLA type and the last column indicates the substitutions used for HLA type classification. Approximately half of the patients could be assigned a subtype by using the available *pol* data. The total represents the overlapping total, meaning that some patients could be assigned multiple HLA types.

HLA Type	**Number of assigned patients**	**Filter Residues**
A02	48	T134I, E138K, E139D, E237A
A03	55	K265R, Q273R
A26	18	K131R, I234R
A29	19	I234V
A68	33	E139D, E152D, K163R
B07	17	T264I
B15	36	T238A, Q273R, Q273E
B18	46	E237A, E306D
B35	55	D220H, E221K, D222E
B44	25	E105K, E303D, E306K, E306N
Total (Overlapping)	230	

Table 4.15: The tables represent significant changes in substitutions within the PR-RT region of HIV-1 subtype B *pol*. The *Pos* column shows the position within the concatenated PR-RT sequences, *Org* is the consensus residue at that position, *Mut* is the amino acid substitution. The next three columns show the log_2-Odds of the tested mutation, the p- and q-values of the FET between the HLA-positive and HLA-negative sets *unadjusted* for phylogeny. This is followed by a second group of three columns showing the same odds-ratio, p- and q-values of the FET between HLA-positive and HLA-negative sets *adjusted* for phylogeny. The final column shows a gain (g) or loss (l) of significance for the FET in the *adjusted* results.

A02

			Unadjusted			Adjusted			
Pos	Res	Mut	*log*-Odds	p	q	*log*-Odds$_{adj}$	p_{adj}	q_{adj}	Change
13	I	V	2.76	0.00602	0.109	2.76	0.00602	0.109	
36	I	M	1.69	0.00987	0.161	1.69	0.00987	0.160	
37	N	K	2.16	0.00596	0.109	2.16	0.00596	0.109	
41	K	N	1.91	0.01161	0.180	1.91	0.01161	0.179	
60	D	E	1.21	0.03595	0.366	1.21	0.03595	0.371	
89	M	I	2.27	0.0152	0.215	2.27	0.0152	0.213	
127	E	K	2.15	0.03576	0.366	2.41	0.0241	0.297	
134	T	I	inf	0.0	0.000	inf	0.0	0.000	
138	E	D	-2.13	0.00657	0.113	-2.13	0.00657	0.112	
138	E	K	inf	0.0	0.000	inf	0.0	0.000	
139	E	D	inf	0.0	0.000	inf	0.0	0.000	
217	V	I	1.90	0.03088	0.360	1.90	0.03088	0.352	
220	D	H	-Inf	0.03583	0.366	-Inf	0.05469	0.490	l

Continued on the following page...

222	D	S	1.39	0.00291	0.079	1.58	0.00102	0.035	
234	I	T	0.83	0.09353	0.678	1.05	0.04487	0.432	g
234	I	L	1.92	0.05054	0.471	2.14	0.0361	0.371	g
237	E	A	inf	0.0	0.000	inf	0.0	0.000	
261	S	A	-2.85	0.02693	0.325	-2.85	0.02693	0.319	
272	A	T	1.36	0.00587	0.109	1.36	0.00587	0.109	
276	E	G	2.74	0.01475	0.215	2.74	0.01475	0.213	
276	E	D	1.01	0.04217	0.404	1.07	0.03544	0.371	
313	F	L	1.53	0.00581	0.109	1.53	0.00581	0.109	

A03

			Unadjusted			Adjusted			
Pos	Res	Mut	*log*-Odds	p	q	*log*-Odds$_{adj}$	p_{adj}	q_{adj}	Change
14	K	R	1.45	0.01987	0.463	1.51	0.02527	0.540	
36	I	V	2.17	0.03824	0.526	2.17	0.03824	0.542	
138	E	D	1.29	0.00966	0.315	1.29	0.00966	0.316	
205	V	M	-2.22	0.01753	0.455	-2.09	0.04063	0.544	
220	D	Y	1.66	0.00387	0.154	3.07	0.00996	0.316	
220	D	H	1.84	0.00426	0.154	1.91	0.00797	0.316	
221	E	K	-Inf	0.03246	0.526	-Inf	0.03246	0.542	
222	D	G	-3.40	0.00243	0.132	-3.10	0.01026	0.316	
265	K	R	inf	0.0	0.000	inf	0.0	0.000	
273	Q	K	-2.00	8e-05	0.005	-2.00	8e-05	0.005	
273	Q	R	inf	0.0	0.000	inf	0.0	0.000	
295	G	E	1.79	0.02663	0.526	1.79	0.02663	0.540	

A26

			Unadjusted			Adjusted			
Pos	Res	Mut	*log*-Odds	p	q	*log*-Odds$_{adj}$	p_{adj}	q_{adj}	Change
159	I	V	-1.40	0.05491	1.000	-1.85	0.03674	1.000	g
197	A	S	2.52	0.02962	1.000	2.81	0.01943	1.000	
220	D	Y	1.74	0.03902	1.000	-Inf	1.0	1.000	l

A29

			Unadjusted			Adjusted			
Pos	Res	Mut	*log*-Odds	p	q	*log*-Odds$_{adj}$	p_{adj}	q_{adj}	Change
19	I	V	2.24	0.00696	0.543	2.40	0.00451	0.446	
35	E	D	1.45	0.04397	1.000	1.45	0.04397	1.000	
138	E	D	-Inf	0.01759	0.802	-Inf	0.01759	1.000	
234	I	T	-Inf	0.018	0.802	-Inf	0.03048	1.000	
234	I	V	inf	0.0	0.000	inf	0.0	0.000	
283	M	I	3.58	0.00653	0.543	3.58	0.00653	0.485	

A33

			Unadjusted			Adjusted			

Continued on the following page...

Pos	Res	Mut	*log*-Odds	p	q	*log*-Odds$_{adj}$	p_{adj}	q_{adj}	Change
60	D	E	1.77	0.02457	0.596	1.77	0.02457	0.741	
89	M	I	2.72	0.02227	0.596	2.72	0.02227	0.741	
139	E	D	2.30	0.04142	0.638	2.30	0.04142	0.741	
200	K	E	1.82	0.04971	0.692	0.89	0.3344	1.000	l
217	V	I	2.42	0.03433	0.638	2.42	0.03433	0.741	
222	D	S	1.34	0.061	0.697	1.50	0.03083	0.741	g
234	I	T	1.48	0.04188	0.638	1.56	0.0549	0.801	l
237	E	A	inf	0.0	0.000	inf	0.0	0.000	
257	A	S	2.57	0.01198	0.463	2.57	0.01198	0.520	
276	E	G	4.40	0.00039	0.042	4.40	0.00039	0.040	
278	V	I	2.30	0.04142	0.638	2.30	0.04142	0.741	
313	F	L	2.42	0.00147	0.078	2.42	0.00147	0.074	

A36

				Unadjusted			Adjusted		
Pos	Res	Mut	*log*-Odds	p	q	*log*-Odds$_{adj}$	p_{adj}	q_{adj}	Change
19	I	V	4.84	0.0035	0.283	4.98	0.00274	0.209	
89	M	I	5.00	0.00631	0.291	5.00	0.00631	0.275	
93	L	I	5.15	0.00528	0.284	5.15	0.00528	0.268	
134	T	M	5.71	0.00272	0.283	5.71	0.00272	0.209	
161	A	V	5.71	0.00272	0.283	5.71	0.00272	0.209	
234	I	T	3.46	0.03397	1.000	3.74	0.02211	0.674	

A68

				Unadjusted			Adjusted		
Pos	Res	Mut	*log*-Odds	p	q	*log*-Odds$_{adj}$	p_{adj}	q_{adj}	Change
12	S	P	2.35	0.04403	0.897	2.35	0.04403	0.954	
12	S	T	-1.72	0.01209	0.493	-1.42	0.04644	0.954	
60	D	E	1.79	0.00461	0.250	1.79	0.00461	0.237	
63	L	T	1.49	0.04381	0.897	1.49	0.04381	0.954	
139	E	D	inf	0.0	0.000	inf	0.0	0.000	
152	E	D	inf	0.0	0.000	inf	0.0	0.000	
207	V	I	2.04	0.01132	0.493	1.46	0.09201	1.000	l
222	D	S	1.09	0.06222	1.000	1.26	0.03291	0.954	g
222	D	G	1.36	0.03281	0.897	1.01	0.14427	1.000	l

B07

				Unadjusted			Adjusted		
Pos	Res	Mut	*log*-Odds	p	q	*log*-Odds$_{adj}$	p_{adj}	q_{adj}	Change
220	D	Y	2.29	0.0066	0.337	3.41	0.02925	0.723	
222	D	E	2.62	0.02529	0.658	2.62	0.02529	0.723	
234	I	T	1.78	0.01606	0.546	1.33	0.15168	1.000	l
264	T	I	inf	0.0	0.000	inf	0.0	0.000	
273	Q	R	2.17	0.02613	0.658	2.17	0.02613	0.723	

Continued on the following page...

277	I	L	2.42	0.00273	0.231	1.74	0.06233	0.945	l
283	M	V	-2.45	0.00475	0.291	-2.92	0.00183	0.176	

B15

			Unadjusted			Adjusted			
Pos	Res	Mut	*log*-Odds	p	q	*log*-Odds$_{adj}$	p_{adj}	q_{adj}	Change
63	L	T	1.66	0.02017	0.506	1.66	0.02017	0.495	
220	D	Y	1.64	0.01085	0.393	2.95	0.02089	0.495	
220	D	H	2.13	0.00315	0.147	2.06	0.01122	0.432	
222	D	G	-2.67	0.04117	0.685	-2.37	0.09885	0.949	l
222	D	E	1.91	0.0455	0.685	1.91	0.0455	0.681	
273	Q	K	-3.06	2e-05	0.002	-3.06	2e-05	0.002	
273	Q	R	inf	0.0	0.000	inf	0.0	0.000	
306	E	A	-Inf	0.03553	0.685	-Inf	0.03553	0.681	

B18

			Unadjusted			Adjusted			
Pos	Res	Mut	*log*-Odds	p	q	*log*-Odds$_{adj}$	p_{adj}	q_{adj}	Change
119	K	R	2.06	0.00104	0.068	2.06	0.00104	0.064	
139	E	D	1.82	0.03506	0.544	1.82	0.03506	0.600	
202	K	N	1.22	0.01222	0.455	1.20	0.01537	0.473	
205	V	M	-0.76	0.48258	1.000	-2.72	0.04243	0.622	g
234	I	T	1.10	0.03354	0.544	1.20	0.02316	0.549	
237	E	A	inf	0.0	0.000	inf	0.0	0.000	
276	E	G	2.81	0.01263	0.455	2.81	0.01263	0.473	
277	I	M	1.50	0.03312	0.544	1.27	0.15623	0.895	l
278	V	I	1.82	0.03506	0.544	1.82	0.03506	0.600	
289	G	A	-1.09	0.32915	1.000	-2.63	0.0415	0.622	g
306	E	D	inf	0.0	0.000	inf	0.0	0.000	

B35

			Unadjusted			Adjusted			
Pos	Res	Mut	*log*-Odds	p	q	*log*-Odds$_{adj}$	p_{adj}	q_{adj}	Change
20	K	R	-1.36	0.02638	0.552	-1.75	0.00774	0.183	
35	E	D	-1.38	0.03281	0.594	-1.38	0.03281	0.532	
63	L	A	3.26	0.0032	0.104	3.26	0.0032	0.099	
220	D	H	inf	0.0	0.000	inf	0.0	0.000	
221	E	K	inf	0.0	0.000	inf	0.0	0.000	
222	D	N	-3.57	0.00088	0.032	-3.40	0.00239	0.082	
222	D	S	-4.67	0.0	0.000	-Inf	0.0	0.000	
222	D	G	-Inf	0.00019	0.008	-Inf	0.00092	0.035	
222	D	E	inf	0.0	0.000	inf	0.0	0.000	
272	A	E	2.17	0.03824	0.656	2.17	0.03824	0.589	
273	Q	R	1.75	0.00855	0.214	1.75	0.00855	0.188	
283	M	V	-0.98	0.02709	0.552	-1.09	0.01599	0.308	

Continued on the following page...

310	K	R	-1.05	0.10678	0.758	-1.72	0.02123	0.385	g

B42

			Unadjusted			Adjusted			
Pos	Res	Mut	log-Odds	p	q	log-Odds$_{adj}$	p_{adj}	q_{adj}	Change
89	M	I	3.69	0.02083	0.611	3.69	0.02083	0.642	
93	L	I	3.85	0.01753	0.611	3.85	0.01753	0.642	
127	E	K	4.02	0.01448	0.611	4.20	0.01176	0.604	
278	V	I	3.31	0.03218	0.749	3.31	0.03218	0.826	
295	G	E	4.20	0.00251	0.205	4.20	0.00251	0.193	

B44

			Unadjusted			Adjusted			
Pos	Res	Mut	log-Odds	p	q	log-Odds$_{adj}$	p_{adj}	q_{adj}	Change
35	E	D	2.38	0.00013	0.008	2.38	0.00013	0.010	
37	N	S	1.97	0.01354	0.441	1.97	0.01354	0.596	
257	A	S	2.56	0.00574	0.234	2.56	0.00574	0.354	
306	E	N	inf	0.0	0.000	inf	0.0	0.000	

B45

			Unadjusted			Adjusted			
Pos	Res	Mut	log-Odds	p	q	log-Odds$_{adj}$	p_{adj}	q_{adj}	Change
93	L	I	3.36	0.02901	1.000	3.36	0.02901	1.000	
238	T	K	4.21	0.0117	1.000	4.21	0.0117	1.000	
272	A	K	3.53	0.02407	1.000	3.53	0.02407	1.000	
306	E	Q	inf	0.0	0.000	inf	0.0	0.000	

B81

			Unadjusted			Adjusted			
Pos	Res	Mut	log-Odds	p	q	log-Odds$_{adj}$	p_{adj}	q_{adj}	Change
103	P	S	inf	0.0	0.000	inf	0.0	0.000	
135	A	E	2.93	0.04839	1.000	2.84	0.08597	1.000	l
217	V	I	3.43	0.02816	1.000	3.43	0.02816	1.000	
265	K	R	2.94	0.02096	1.000	3.04	0.01787	1.000	
289	G	A	3.39	0.00494	0.403	3.06	0.02107	1.000	

C02

			Unadjusted			Adjusted			
Pos	Res	Mut	log-Odds	p	q	log-Odds$_{adj}$	p_{adj}	q_{adj}	Change
12	S	P	2.83	0.02069	0.320	2.83	0.02069	0.334	
63	L	T	1.63	0.04566	0.544	1.63	0.04566	0.575	
89	M	L	1.58	0.01187	0.294	1.45	0.03011	0.440	
138	E	D	1.78	0.00741	0.241	1.78	0.00741	0.227	
220	D	Y	2.45	0.00061	0.033	3.75	0.00557	0.214	
220	D	H	2.90	0.0002	0.013	2.82	0.00131	0.080	

Continued on the following page...

222	D	G	-Inf	0.03361	0.455	-Inf	0.05616	0.575	1
222	D	E	2.56	0.01277	0.294	2.56	0.01277	0.318	
264	T	I	2.17	0.02613	0.369	1.42	0.21293	0.887	1
272	A	E	2.83	0.02069	0.320	2.83	0.02069	0.334	
273	Q	K	-Inf	0.0	0.000	-Inf	0.0	0.000	
273	Q	R	inf	0.0	0.000	inf	0.0	0.000	
277	I	L	1.86	0.00827	0.244	1.61	0.03456	0.482	

C04

			Unadjusted			Adjusted			
Pos	Res	Mut	log-Odds	p	q	log-Odds$_{adj}$	p_{adj}	q_{adj}	Change
20	K	R	-1.57	0.02551	0.416	-1.44	0.03781	0.485	
35	E	D	-1.70	0.02071	0.375	-1.70	0.02071	0.399	
36	I	L	2.26	0.00235	0.109	2.26	0.00235	0.110	
63	L	A	2.85	0.01165	0.271	2.85	0.01165	0.276	
127	E	K	2.88	0.00447	0.146	3.20	0.00251	0.110	
134	T	K	1.86	0.03208	0.455	1.86	0.03208	0.462	
138	E	D	1.14	0.02852	0.428	1.14	0.02852	0.445	
197	A	G	1.72	0.04218	0.509	2.01	0.02422	0.437	
220	D	H	inf	0.0	0.000	inf	0.0	0.000	
273	Q	R	1.86	0.00827	0.225	1.86	0.00827	0.232	
283	M	V	-0.97	0.03892	0.508	-1.04	0.03453	0.462	
299	A	E	inf	0.0	0.000	inf	0.0	0.000	
303	E	K	2.04	0.0408	0.509	2.04	0.0408	0.503	

C07

			Unadjusted			Adjusted			
Pos	Res	Mut	log-Odds	p	q	log-Odds$_{adj}$	p_{adj}	q_{adj}	Change
10	L	V	5.00	0.00502	0.722	5.00	0.00502	0.682	
306	E	N	4.74	0.00664	0.722	4.74	0.00664	0.682	

C14

			Unadjusted			Adjusted			
Pos	Res	Mut	log-Odds	p	q	log-Odds$_{adj}$	p_{adj}	q_{adj}	Change
138	E	K	3.29	0.03182	1.000	3.29	0.03182	1.000	
222	D	N	2.52	0.02254	1.000	2.71	0.01569	0.825	
306	E	Q	3.95	0.01539	1.000	3.95	0.01539	0.825	

C18

			Unadjusted			Adjusted			
Pos	Res	Mut	log-Odds	p	q	log-Odds$_{adj}$	p_{adj}	q_{adj}	Change
205	V	M	3.37	0.01792	1.000	3.51	0.01427	1.000	

4.7.1 MHC change

Table 4.14: This table depicts imputed HLA types for patients from HIV subtype C PR-RT data. The HLA type is listed with the corresponding sensitivity and positive predictive value (PPV) as described in Section 3.2.1 [Carlson *et al.*, 2014].

HLA	PPV	Sensitivity
A02	0.29	0.21
A03	0.28	0.33
A26	0.12	0.19
A29	0.21	0.11
A33	0.07	0.20
A36	0.19	0.09
A68	0.19	0.26
B07	0.37	0.19
B15	0.35	0.18
B18	0.19	0.28
B35	0.20	0.41
B42	0.32	0.21
B44	0.49	0.29
B45	0.19	0.07
B81	0.67	0.36
C02	0.27	0.10
C04	0.39	0.09
C07	0.64	0.06
C14	0.12	0.20
C18	0.44	0.11

Table 4.16: The following table represents the initial predicted MHC affinity and stability predictions by NetMHCPan and NetMHCstab respectively for HIV subtype C PR-RT. The HLA, Mut and Freq columns represent the HLA allotype for which the prediction was made, the substitution and the frequency of the substitution. The following two columns represent the consensus and mutated peptide, with the mutated residue in orange and underlined. The $IC50_{pep}$ and $IC50_{mut}$ columns show the predicted $1 - log_{50000}IC50$ value obtained from prediction results for the consensus and mutated peptide, while the $IC50_{\Delta}$ column shows the log difference between the two predicted scores. The o_{pep} and o_{mut} columns show the predicted stability (measured as half-rate in hours) for the consensus and mutated peptide. The o_{Δ} value shows the difference between the two scores. The "T" column shows when a change in score resulted in the crossing of the $1 - log_{50000}IC50$ threshold of 0.42, with "-T" meaning the affinity dropped to below the threshold and "+T" meaning the affinity increased above the threshold.

HLA	Mut	Freq	Pep	Mut	$IC50_{pep}$	$IC50_{mut}$	$IC50_{\Delta}$	T	o_{pep}	o_{mut}	o_{Δ}
A*1101	S12P	0.026	VSIKVGGQIK	VPIKVGGQIK	0.50	0.09	-0.41	-T	1.51	0.47	-1.04
B*0702	S12P	0.026	VSIKVGGQI	VPIKVGGQI	0.12	0.56*	0.44	+T	0.34	1.97	1.63
A*0201	S12T	0.351	TLWQRPLVSI	TLWQRPLVTI	0.59	0.56	-0.03		3.30	4.45	1.14
A*1101	S12T	0.351	VSIKVGGQIK	VTIKVGGQIK	0.50	0.56	0.06		1.51	2.82	1.31

Continued on the following page...

HLA	Mut	Freq	Pep	Mut	$IC50_{pep}$	$IC50_{mut}$	$IC50_{\Delta}$	T	o_{pep}	o_{mut}	o_{Δ}
A*0201	I13V	0.029	TLWQRPLVSI	TLWQRPLVSV	0.59	0.70	0.11		3.30	6.61	3.30
B*5801	G16E	0.066	VSIKVGGQI	VSIKVEGQI	0.40	0.46*	0.06	+T	1.33	0.97	-0.36
A*1101	I19L	0.086	VSIKVGGQIK	VSIKVGGQLK	0.50	0.63	0.13		1.51	1.99	0.47
A*1101	K20R	0.25	VSIKVGGQIK	VSIKVGGQIR	0.50	0.28	-0.22	-T	1.51	0.59	-0.92
B*1503	L33F	0.029	LDTGADDTVL	LDTGADDTVF	0.31	0.47*	0.16	+T	0.46	1.03	0.57
A*1101	E35D	0.221	VLEEINLPGK	VLEDINLPGK	0.40	0.42*	0.02	+T	1.02	1.25	0.24
A*1101	I36M	0.083	INLPGKWKPK	MNLPGKWKPK	0.23	0.34	0.11		1.20	1.28	0.07
A*0301	N37S	0.086	VLEEINLPGK	VLEEISLPGK	0.39	0.44*	0.05	+T	1.74	1.94	0.20
A*0301	N37S	0.086	INLPGKWKPK	ISLPGKWKPK	0.15	0.38	0.23		0.65	1.03	0.39
A*1101	N37S	0.086	VLEEINLPGK	VLEEISLPGK	0.40	0.43*	0.03	+T	1.02	1.16	0.14
A*1101	N37S	0.086	INLPGKWKPK	ISLPGKWKPK	0.23	0.57*	0.34	+T	1.20	2.80	1.60
B*0702	K45R	0.057	KPKMIGGIG	KPRMIGGIG	0.23	0.40	0.17		1.06	2.02	0.96
B*0702	K45R	0.057	KPKMIGGIGG	KPRMIGGIGG	0.19	0.35	0.16		0.90	1.73	0.83
B*1503	K45R	0.057	KMIGGIGGF	RMIGGIGGF	0.81	0.85	0.04		9.48	11.99	2.50
A*2601	M46I	0.029	KMIGGIGGF	KIIGGIGGF	0.29	0.41	0.12		0.54	1.07	0.53
B*0702	M46I	0.029	LPGKWKPKM	LPGKWKPKI	0.42	0.29	-0.13	-T	1.16	0.80	-0.36
B*1503	M46I	0.029	KMIGGIGGFI	KIIGGIGGFI	0.60	0.28	-0.32	-T	1.01	0.49	-0.52
A*0301	L63A	0.023	ILIEICGKK	IAIEICGKK	0.43	0.27	-0.16	-T	1.09	0.53	-0.56
A*1101	L63A	0.023	ILIEICGKK	IAIEICGKK	0.41	0.44*	0.03	+T	3.31	2.95	-0.36
A*1101	L63A	0.023	QILIEICGKK	QIAIEICGKK	0.41	0.42*	0.02	+T	1.72	1.35	-0.38
B*0702	L63A	0.023	KVRQYDQIL	KVRQYDQIA	0.45	0.22	-0.23	-T	1.58	0.66	-0.92
B*0702	L63A	0.023	KVRQYDQILI	KVRQYDQIAI	0.25	0.47*	0.23	+T	0.79	1.33	0.55
B*1503	L63A	0.023	KVRQYDQIL	KVRQYDQIA	0.55	0.34	-0.20	-T	1.42	0.44	-0.97
A*0301	L63P	0.405	ILIEICGKK	IPIEICGKK	0.43	0.07	-0.36	-T	1.09	0.363	-0.73
B*0702	L63P	0.405	KVRQYDQIL	KVRQYDQIP	0.45	0.07	-0.38	-T	1.58	0.416	-1.16
B*0702	L63P	0.405	KVRQYDQILI	KVRQYDQIPI	0.25	0.43*	0.18	+T	0.79	1.22	0.43
B*1503	L63P	0.405	KVRQYDQIL	KVRQYDQIP	0.55	0.21	-0.34	-T	1.42	0.50	-0.91
B*1503	L63P	0.405	KVRQYDQILI	KVRQYDQIPI	0.40	0.51*	0.11	+T	0.62	1.28	0.66
A*0301	L63S	0.063	ILIEICGKK	ISIEICGKK	0.43	0.34	-0.09	-T	1.09	0.62	-0.47
A*1101	L63S	0.063	ILIEICGKK	ISIEICGKK	0.41	0.57*	0.16	+T	3.31	5.74	2.43
B*0702	L63S	0.063	KVRQYDQIL	KVRQYDQIS	0.45	0.10	-0.35	-T	1.58	0.51	-1.06
B*0702	L63S	0.063	KVRQYDQILI	KVRQYDQISI	0.25	0.43*	0.18	+T	0.79	1.29	0.50
B*1503	L63S	0.063	KVRQYDQIL	KVRQYDQIS	0.55	0.23	-0.32	-T	1.42	0.46	-0.95
B*1503	L63S	0.063	RQYDQILIEI	RQYDQISIEI	0.88	0.88	0.01		5.22	6.53	1.31
A*0301	L63T	0.083	ILIEICGKK	ITIEICGKK	0.43	0.39	-0.04	-T	1.09	1.01	-0.08
A*1101	L63T	0.083	ILIEICGKK	ITIEICGKK	0.41	0.62*	0.21	+T	3.31	10.71	7.41
B*0702	L63T	0.083	KVRQYDQIL	KVRQYDQIT	0.45	0.14	-0.31	-T	1.58	0.63	-0.95
B*1503	L63T	0.083	KVRQYDQIL	KVRQYDQIT	0.55	0.27	-0.28	-T	1.42	0.43	-0.98
B*1503	L63T	0.083	RQYDQILIEI	RQYDQITIEI	0.88	0.88	0.00		5.22	7.08	1.86
A*0201	V77I	0.103	VLVGPTPVNI	VLIGPTPVNI	0.54	0.62	0.08		2.16	3.55	1.40
B*0702	V82A	0.057	TPVNIIGRNM	TPANIIGRNM	0.49	0.59	0.11		1.70	2.53	0.83
A*0201	M89L	0.224	MLTQLGCTL	LLTQLGCTL	0.49	0.42	-0.07	-T	2.54	3.06	0.51
B*3501	M89L	0.224	TPVNIIGRNM	TPVNIIGRNL	0.54	0.40	-0.14	-T	2.00	1.07	-0.93
A*0201	L90M	0.029	NMLTQLGCTL	NMMTQLGCTL	0.46	0.60	0.14		1.35	1.69	0.34
B*1503	L90M	0.029	MLTQLGCTL	MMTQLGCTL	0.68	0.82	0.14		1.34	2.50	1.16

Continued on the following page...

HLA	Mut	Freq	Pep	Mut	$IC50_{pep}$	$IC50_{mut}$	$IC50_{\Delta}$	T	o_{pep}	o_{mut}	o_{Δ}
B*5801	L90M	0.029	LTQLGCTLNF	MTQLGCTLNF	0.45	0.59	0.14		1.49	1.82	0.33
A*0201	L93I	0.032	MLTQLGCTL	MLTQIGCTL	0.49	0.53	0.04		2.54	3.78	1.24
B*1503	L93I	0.032	TQLGCTLNF	TQIGCTLNF	0.85	0.86	0.01		3.27	6.79	3.52
A*0201	P103S	0.02	TLNFPISPI	TLNFPISSI	0.52	0.42	-0.10	-T	1.29	1.04	-0.25
A*1101	P103S	0.02	SPIETVPVK	SSIETVPVK	0.14	0.62*	0.48	+T	1.55	10.27	8.72
A*1101	P103S	0.02	ISPIETVPVK	ISSIETVPVK	0.37	0.57*	0.20	+T	1.00	1.53	0.53
A*1101	P103S	0.02	PIETVPVKLK	SIETVPVKLK	0.24	0.58*	0.35	+T	0.94	2.03	1.09
B*0702	P103S	0.02	SPIETVPVKL	SSIETVPVKL	0.54	0.12	-0.42	-T	2.28	0.347	-1.93
B*3501	P103S	0.02	SPIETVPVKL	SSIETVPVKL	0.55	0.19	-0.35	-T	1.61	0.54	-1.07
A*0301	K119R	0.092	KLKPGMDGPK	KLKPGMDGPR	0.55	0.36	-0.19	-T	3.33	0.81	-2.52
A*1101	K119R	0.092	KLKPGMDGPK	KLKPGMDGPR	0.46	0.30	-0.16	-T	1.32	0.57	-0.74
B*0702	K119R	0.092	GPKVKQWPL	GPRVKQWPL	0.54	0.74	0.20		1.59	3.10	1.51
A*0201	E127K	0.029	KQWPLTEEKI	KQWPLTKEKI	0.46	0.37	-0.09	-T	1.07	0.93	-0.14
A*0301	E127K	0.029	KVKQWPLTE	KVKQWPLTK	0.13	0.62*	0.49	+T	0.93	10.11	9.18
A*1101	E127K	0.029	KVKQWPLTE	KVKQWPLTK	0.13	0.68*	0.55	+T	0.42	2.67	2.25
A*0201	T134I	0.023	ALTAICEEM	ALIAICEEM	0.45	0.60	0.15		4.17	6.37	2.21
B*4402	T134I	0.023	TEEKIKALT	TEEKIKALI	0.17	0.28	0.11		0.351	1.08	0.73
A*0201	T134M	0.023	ALTAICEEM	ALMAICEEM	0.45	0.73	0.28		4.17	8.69	4.52
A*0201	T134M	0.023	ALTAICEEME	ALMAICEEME	0.06	0.16	0.10		0.68	1.06	0.38
B*1503	T134M	0.023	ALTAICEEM	ALMAICEEM	0.53	0.70	0.18		0.99	2.89	1.90
B*1801	T134M	0.023	TEEKIKALT	TEEKIKALM	0.15	0.45*	0.30	+T	0.351	1.04	0.69
B*4402	T134M	0.023	TEEKIKALT	TEEKIKALM	0.17	0.32	0.15		0.351	1.04	0.69
A*0201	A135E	0.448	ALTAICEEM	ALTEICEEM	0.45	0.57	0.12		4.17	5.90	1.73
A*1101	A135E	0.448	TAICEEMEK	TEICEEMEK	0.45	0.13	-0.32	-T	1.65	0.90	-0.75
A*0201	M140L	0.092	ALTAICEEM	ALTAICEEL	0.45	0.61	0.16		4.17	7.41	3.25
B*1503	T147S	0.19	ITKIGPENPY	ISKIGPENPY	0.46	0.58	0.12		5.56	6.03	0.46
A*1101	A161V	0.023	FAIKKKDSTK	FVIKKKDSTK	0.32	0.43*	0.10	+T	0.98	2.07	1.09
B*2705	D166N	0.147	KKDSTKWRK	KKNSTKWRK	0.13	0.18	0.05		0.99	4.27	3.28
B*2705	K169R	0.083	TKWRKLVDF	TRWRKLVDF	0.17	0.49*	0.32	+T	0.45	1.61	1.16
B*2705	K169R	0.083	TKWRKLVDFR	TRWRKLVDFR	0.20	0.51*	0.31	+T	0.48	1.36	0.89
A*1101	L173V	0.02	KLVDFRELNK	KVVDFRELNK	0.67	0.75	0.07		2.99	7.58	4.59
A*0201	A197G	0.046	VQLGIPHPA	VQLGIPHPG	0.57	0.17	-0.40	-T	1.46	0.63	-0.82
A*1101	A197G	0.046	GIPHPAGLKK	GIPHPGGLKK	0.45	0.41	-0.04	-T	1.10	1.03	-0.07
A*0201	A197S	0.04	VQLGIPHPA	VQLGIPHPS	0.57	0.24	-0.32	-T	1.46	0.62	-0.84
B*1801	K200E	0.078	LKKKKSVTVL	LEKKKSVTVL	0.04	0.32	0.28		0.158	1.08	0.92
B*4402	K200E	0.078	LKKKKSVTVL	LEKKKSVTVL	0.04	0.24	0.21		0.158	1.08	0.92
A*0201	K200Q	0.023	LKKKKSVTV	LQKKKSVTV	0.02	0.16	0.14		0.65	1.18	0.53
A*0201	K200Q	0.023	GLKKKKSVTV	GLQKKKSVTV	0.18	0.31	0.14		2.04	3.39	1.35
B*1503	K200Q	0.023	LKKKKSVTV	LQKKKSVTV	0.38	0.54*	0.17	+T	0.36	3.82	3.46
B*1503	K200Q	0.023	LKKKKSVTVL	LQKKKSVTVL	0.60	0.73	0.13		0.67	9.83	9.16
A*1101	K201Q	0.02	GIPHPAGLKK	GIPHPAGLKQ	0.45	0.04	-0.41	-T	1.10	0.321	-0.78
B*1503	K201Q	0.02	KKKKSVTVL	KQKKSVTVL	0.71	0.82	0.11		0.88	13.23	12.35
B*1503	K201Q	0.02	KKKKSVTVLD	KQKKSVTVLD	0.13	0.28	0.15		0.347	1.67	1.32
B*0702	K202R	0.02	KKKKSVTVL	KKRKSVTVL	0.16	0.29	0.13		0.63	1.22	0.59
B*2705	K202R	0.02	KKKSVTVLD	KRKSVTVLD	0.06	0.31	0.25		0.405	1.24	0.84

Continued on the following page...

HLA	Mut	Freq	Pep	Mut	$IC50_{pep}$	$IC50_{mut}$	$IC50_{\Delta}$	T	o_{pep}	o_{mut}	o_{Δ}
B*2705	K202R	0.02	KKKSVTVLDV	KRKSVTVLDV	0.13	0.45*	0.32	+T	0.49	1.66	1.17
A*2601	V205M	0.132	VTVLDVGDAY	MTVLDVGDAY	0.38	0.67*	0.29	+T	0.81	1.24	0.44
B*1503	V205M	0.132	VTVLDVGDAY	MTVLDVGDAY	0.45	0.59	0.14		2.62	2.44	-0.18
A*0101	D220H	0.078	PLDEDFRKY	PLHEDFRKY	0.45	0.11	-0.34	-T	1.06	0.399	-0.66
A*1101	D220H	0.078	SVPLDEDFRK	SVPLHEDFRK	0.45	0.52	0.07		1.33	2.52	1.19
B*1503	D220H	0.078	PLDEDFRKY	PLHEDFRKY	0.17	0.38	0.21		0.338	1.36	1.02
B*3501	D220H	0.078	DAYFSVPLD	DAYFSVPLH	0.08	0.27	0.19		0.98	1.81	0.83
A*0101	D220Y	0.112	PLDEDFRKY	PLYEDFRKY	0.45	0.14	-0.30	-T	1.06	0.375	-0.68
A*1101	D220Y	0.112	SVPLDEDFRK	SVPLYEDFRK	0.45	0.56	0.12		1.33	2.00	0.67
A*2301	D220Y	0.112	YFSVPLDEDF	YFSVPLYEDF	0.63	0.77	0.14		0.64	1.24	0.60
A*2402	D220Y	0.112	YFSVPLDEDF	YFSVPLYEDF	0.55	0.70	0.15		0.64	1.24	0.60
A*2601	D220Y	0.112	DAYFSVPLD	DAYFSVPLY	0.04	0.62	0.58	+T	0.34	0.88	0.55
A*2901	D220Y	0.112	PLDEDFRKY	PLYEDFRKY	0.32	0.51*	0.19	+T	1.06	0.375	-0.68
B*1503	D220Y	0.112	PLDEDFRKY	PLYEDFRKY	0.17	0.46*	0.28	+T	0.338	1.50	1.16
B*3501	D220Y	0.112	DAYFSVPLD	DAYFSVPLY	0.08	0.61*	0.53	+T	0.98	4.42	3.44
A*0101	E221K	0.063	PLDEDFRKY	PLDKDFRKY	0.45	0.38	-0.07	-T	1.06	1.00	-0.06
A*1101	E221K	0.063	AYFSVPLDE	AYFSVPLDK	0.06	0.52*	0.46	+T	0.361	1.24	0.87
A*1101	E221K	0.063	DAYFSVPLDE	DAYFSVPLDK	0.04	0.43*	0.39	+T	0.315	1.01	0.70
B*1503	D222E	0.04	FSVPLDEDF	FSVPLDEEF	0.54	0.65	0.11		2.76	4.30	1.54
B*3501	D222E	0.04	FSVPLDEDF	FSVPLDEEF	0.35	0.49*	0.14	+T	1.28	1.69	0.41
B*5801	D222E	0.04	FSVPLDEDF	FSVPLDEEF	0.59	0.69	0.11		1.05	1.64	0.59
A*0201	I234L	0.032	SINNETPGI	SLNNETPGI	0.31	0.58*	0.28	+T	1.61	4.55	2.95
A*0201	I234M	0.02	YTAFTIPSI	YTAFTIPSM	0.61	0.41	-0.20	-T	1.24	0.70	-0.54
A*0201	I234M	0.02	SINNETPGI	SMNNETPGI	0.31	0.57*	0.27	+T	1.61	3.33	1.72
A*2601	I234M	0.02	YTAFTIPSI	YTAFTIPSM	0.46	0.80	0.34		0.74	1.65	0.91
B*1503	I234M	0.02	YTAFTIPSI	YTAFTIPSM	0.48	0.71	0.23		0.47	1.26	0.79
B*3501	I234M	0.02	YTAFTIPSI	YTAFTIPSM	0.28	0.62*	0.34	+T	0.67	1.74	1.07
A*0201	I234T	0.218	YTAFTIPSI	YTAFTIPST	0.61	0.33	-0.29	-T	1.24	0.58	-0.66
A*1101	I234T	0.218	SINNETPGIR	STNNETPGIR	0.36	0.44*	0.08	+T	1.20	2.22	1.01
A*2301	I234T	0.218	KYTAFTIPSI	KYTAFTIPST	0.73	0.26	-0.47	-T	4.91	0.91	-4.00
A*2402	I234T	0.218	KYTAFTIPSI	KYTAFTIPST	0.67	0.18	-0.48	-T	4.91	0.91	-4.00
B*1503	I234T	0.218	INNETPGIRY	TNNETPGIRY	0.53	0.38	-0.14	-T	1.82	0.90	-0.92
B*5801	I234T	0.218	SINNETPGI	STNNETPGI	0.12	0.26	0.14		0.57	1.23	0.67
A*0201	I234V	0.055	YTAFTIPSI	YTAFTIPSV	0.61	0.76	0.15		1.24	2.31	1.08
A*2601	E237A	0.055	ETPGIRYQY	ATPGIRYQY	0.77	0.34	-0.43	-T	2.08	0.52	-1.56
B*3501	E237A	0.055	NETPGIRYQY	NATPGIRYQY	0.23	0.60*	0.36	+T	0.60	2.40	1.80
A*0201	I241V	0.04	SINNETPGI	SINNETPGV	0.31	0.47*	0.17	+T	1.61	3.01	1.40
A*1101	A257S	0.052	PAIFQSSMTK	PSIFQSSMTK	0.40	0.49*	0.09	+T	0.89	1.28	0.39
B*0702	A257S	0.052	LPQGWKGSPA	LPQGWKGSPS	0.59	0.40	-0.19	-T	1.67	1.29	-0.38
B*1503	A257S	0.052	GWKGSPAIF	GWKGSPSIF	0.41	0.43*	0.02	+T	1.39	1.25	-0.14
A*0201	S261A	0.118	AIFQSSMTKI	AIFQASMTKI	0.40	0.42*	0.02	+T	1.65	1.97	0.32
A*1101	S261A	0.118	SSMTKILEP	ASMTKILEP	0.10	0.09	-0.02		2.25	3.46	1.21
B*1503	S261A	0.118	SSMTKILEPF	ASMTKILEPF	0.81	0.79	-0.02		2.25	3.33	1.08
A*0201	S261C	0.118	AIFQSSMTKI	AIFQCSMTKI	0.40	0.46*	0.06	+T	1.65	2.03	0.38
B*5801	S261C	0.118	SSMTKILEPF	CSMTKILEPF	0.53	0.56	0.02		1.38	2.98	1.60

Continued on the following page...

HLA	Mut	Freq	Pep	Mut	$IC50_{pep}$	$IC50_{mut}$	$IC50_{\Delta}$	T	o_{pep}	o_{mut}	o_{Δ}
B*0702	T264I	0.049	SPAIFQSSMT	SPAIFQSSMI	0.39	0.63*	0.24	+T	1.18	2.03	0.85
B*1503	T264I	0.049	SMTKILEPF	SMIKILEPF	0.86	0.90	0.04		3.32	6.30	2.99
B*1503	K265R	0.095	FQSSMTKIL	FQSSMTRIL	0.76	0.82	0.06		2.07	3.19	1.13
A*1101	A272K	0.029	MTKILEPFRA	MTKILEPFRK	0.13	0.67*	0.54	+T	0.284	1.57	1.28
B*1503	A272K	0.029	AQNPEIVIY	KQNPEIVIY	0.80	0.82	0.01		19.34	22.29	2.95
B*5801	A272K	0.029	RAQNPEIVI	RKQNPEIVI	0.46	0.06	-0.40	-T	1.02	0.405	-0.62
B*5801	A272K	0.029	RAQNPEIVIY	RKQNPEIVIY	0.50	0.05	-0.44	-T	1.79	0.46	-1.33
A*0301	Q273K	0.397	KILEPFRAQ	KILEPFRAK	0.12	0.61*	0.49	+T	0.76	5.17	4.42
A*1101	Q273K	0.397	KILEPFRAQ	KILEPFRAK	0.10	0.61*	0.50	+T	0.60	10.03	9.43
B*1503	Q273K	0.397	RAQNPEIVIY	RAKNPEIVIY	0.74	0.71	-0.03		6.45	10.59	4.15
B*5801	Q273K	0.397	RAQNPEIVI	RAKNPEIVI	0.46	0.34	-0.12	-T	1.02	0.66	-0.37
B*5801	Q273K	0.397	RAQNPEIVIY	RAKNPEIVIY	0.50	0.37	-0.12	-T	1.79	1.03	-0.77
A*0301	Q273R	0.072	KILEPFRAQ	KILEPFRAR	0.12	0.38	0.27		0.76	1.01	0.25
A*1101	Q273R	0.072	KILEPFRAQ	KILEPFRAR	0.10	0.41	0.31		0.60	2.31	1.70
B*0702	Q273R	0.072	RAQNPEIVI	RARNPEIVI	0.24	0.45*	0.20	+T	0.62	1.00	0.38
B*1503	Q273R	0.072	RAQNPEIVIY	RARNPEIVIY	0.74	0.77	0.03		6.45	10.27	3.82
B*5801	Q273R	0.072	RAQNPEIVI	RARNPEIVI	0.46	0.38	-0.09	-T	1.02	0.59	-0.44
A*0101	E276D	0.233	NPEIVIYQY	NPDIVIYQY	0.12	0.24	0.12		0.55	1.07	0.52
B*0702	I277L	0.129	EPFRAQNPEI	EPFRAQNPEL	0.44	0.54	0.10		0.85	1.49	0.63
B*3501	I277L	0.129	EPFRAQNPEI	EPFRAQNPEL	0.42	0.56	0.14		0.92	1.20	0.29
B*1503	I277M	0.072	AQNPEIVIY	AQNPEMVIY	0.80	0.80	-0.00		19.34	20.77	1.43
B*3501	I277M	0.072	EPFRAQNPEI	EPFRAQNPEM	0.42	0.70	0.28		0.92	2.45	1.53
A*0101	Y280C	0.124	VIYQYMDDLY	VICQYMDDLY	0.38	0.35	-0.04		1.15	2.29	1.15
A*2301	Y280C	0.124	IYQYMDDLYV	ICQYMDDLYV	0.61	0.07	-0.54	-T	1.07	0.294	-0.77
A*2402	Y280C	0.124	IYQYMDDLYV	ICQYMDDLYV	0.56	0.05	-0.51	-T	1.07	0.294	-0.77
A*2901	Y280C	0.124	IYQYMDDLY	ICQYMDDLY	0.60	0.37	-0.23	-T	0.56	1.12	0.56
A*2901	Y280C	0.124	VIYQYMDDLY	VICQYMDDLY	0.74	0.63	-0.11		1.15	2.29	1.15
B*1503	Y280C	0.124	AQNPEIVIY	AQNPEIVIC	0.80	0.38	-0.42	-T	19.34	1.38	-17.96
B*1503	Y280C	0.124	RAQNPEIVIY	RAQNPEIVIC	0.74	0.27	-0.48	-T	6.45	0.44	-6.01
B*1503	Y280C	0.124	VIYQYMDDLY	VICQYMDDLY	0.56	0.35	-0.21	-T	2.68	2.43	-0.25
B*5801	Y280C	0.124	RAQNPEIVIY	RAQNPEIVIC	0.50	0.12	-0.38	-T	1.79	0.64	-1.16
B*3501	M283I	0.023	NPEIVIYQYM	NPEIVIYQYI	0.51	0.26	-0.25	-T	1.74	0.74	-1.00
B*3501	M283V	0.523	NPEIVIYQYM	NPEIVIYQYV	0.51	0.24	-0.27	-T	1.74	0.68	-1.06
A*0201	Y287L	0.02	LYVGSDLEI	LLVGSDLEI	0.09	0.49*	0.41	+T	0.364	1.82	1.45
A*2301	Y287L	0.02	IYQYMDDLY	IYQYMDDLL	0.54	0.68	0.14		0.90	2.00	1.10
A*2301	Y287L	0.02	LYVGSDLEI	LLVGSDLEI	0.63	0.10	-0.53	-T	1.36	0.319	-1.04
A*2402	Y287L	0.02	IYQYMDDLY	IYQYMDDLL	0.51	0.66	0.15		0.90	2.00	1.10
A*2402	Y287L	0.02	LYVGSDLEI	LLVGSDLEI	0.53	0.07	-0.47	-T	1.36	0.319	-1.04
A*2901	Y287L	0.02	VIYQYMDDLY	VIYQYMDDLL	0.74	0.13	-0.61	-T	1.15	0.34	-0.81
B*1503	Y287L	0.02	LYVGSDLEI	LLVGSDLEI	0.43	0.51	0.08		0.322	1.37	1.05
B*1503	Y287L	0.02	VIYQYMDDLY	VIYQYMDDLL	0.56	0.42	-0.14	-T	2.68	0.56	-2.12
A*0201	G289A	0.121	YQYMDDLYVG	YQYMDDLYVA	0.22	0.64*	0.41	+T	0.58	1.28	0.70
A*2301	G289A	0.121	QYMDDLYVG	QYMDDLYVA	0.36	0.45*	0.09	+T	1.03	0.87	-0.16
B*1503	G289A	0.121	YQYMDDLYVG	YQYMDDLYVA	0.71	0.83	0.11		1.17	1.24	0.07
A*0301	E302K	0.037	EELREHLLK	KELREHLLK	0.08	0.15	0.07		0.45	3.00	2.55

Continued on the following page...

HLA	Mut	Freq	Pep	Mut	$IC50_{pep}$	$IC50_{mut}$	$IC50_{\Delta}$	T	o_{pep}	o_{mut}	o_{Δ}
A*1101	E302K	0.037	EELREHLLK	KELREHLLK	0.12	0.19	0.07		0.51	1.70	1.19
B*5801	E302K	0.037	EELREHLLKW	KELREHLLKW	0.11	0.24	0.13		0.54	3.95	3.42
B*1503	E306Q	0.026	REHLLKWGF	RQHLLKWGF	0.65	0.85	0.20		0.43	4.32	3.89
B*1801	E306Q	0.026	REHLLKWGF	RQHLLKWGF	0.56	0.25	-0.31	-T	1.30	0.53	-0.76
B*2705	E306Q	0.026	REHLLKWGF	RQHLLKWGF	0.28	0.50*	0.22	+T	0.61	3.16	2.55
B*2705	E306Q	0.026	REHLLKWGFT	RQHLLKWGFT	0.13	0.27	0.13		0.355	1.04	0.69
B*4402	E306Q	0.026	REHLLKWGF	RQHLLKWGF	0.44	0.21	-0.23	-T	1.30	0.53	-0.76
A*0201	H307Y	0.023	HLLKWGFTT	YLLKWGFTT	0.49	0.76	0.28		1.19	4.37	3.18
A*0201	H307Y	0.023	HLLKWGFTTP	YLLKWGFTTP	0.18	0.41	0.23		0.79	2.35	1.56
B*1503	H307Y	0.023	AKIEELREH	AKIEELREY	0.44	0.76	0.31		0.53	2.13	1.60
B*1503	H307Y	0.023	RAKIEELREH	RAKIEELREY	0.24	0.58*	0.34	+T	1.50	6.75	5.25
B*1801	H307Y	0.023	REHLLKWGF	REYLLKWGF	0.56	0.66	0.10		1.30	1.22	-0.08
B*5801	L309W	0.037	KIEELREHLL	KIEELREHLW	0.15	0.58*	0.43	+T	0.388	2.64	2.25
A*0201	F313L	0.135	HLLKWGFTT	HLLKWGLTT	0.49	0.32	-0.17	-T	1.19	0.76	-0.43
B*1801	F313L	0.135	REHLLKWGF	REHLLKWGL	0.56	0.40	-0.15	-T	1.30	1.92	0.62
B*4402	F313L	0.135	REHLLKWGF	REHLLKWGL	0.44	0.36	-0.07	-T	1.30	1.92	0.62
B*1801	T314F	0.034	REHLLKWGFT	REHLLKWGFF	0.16	0.46*	0.29	+T	0.47	1.15	0.68
B*4402	T314F	0.034	REHLLKWGFT	REHLLKWGFF	0.20	0.38	0.18		0.47	1.15	0.68
A*0201	T314Y	0.106	HLLKWGFTT	HLLKWGFYT	0.49	0.60	0.12		1.19	1.30	0.11

4.8 Summary

In this chapter, it was discovered that some mutations strongly associated with antiretroviral resistance were diminished in sequence sets with assigned HLA types. Of particular interest were the mutations PR L90M and RT 215Y/F which are strongly associated with resistance of HIV protease inhibitors and HIV reverse transcriptase inhibitors. These mutations remained significantly diminished even after compensating for diminished correlated mutations. To investigate a possible causal relationship between HLA types B*15 and B*48, and diminished antiretroviral resistance residues, NetMHCPan, NetMHCStab and NetChop 3.0 were utilized to predict changes in MHC affinity, stability and proteasomal cleavage prediction scores. Interestingly, the diminished substitutions did result in increased probability of peptide MHC affinity and stability predictions within the regions of the diminished substitutions. The similarities in the diminished substitutions observed in B*15 and B*48 can be attributed to the similarities of binding motif of HLA allotypes within these HLA types, particularly the similarity between HLA B*48:02 and B*15:03.

Further investigation lead to the discovery of enriched residues within the PR 89-97 and RT 206-215 regions. These substitutions resulted in decreased MHC affinity and stability predictions in RT as well as an increased chance in proteasomal cleavage within the LM9 epitope in PR. The patient assignments for the two HLA types were corrected for phylogenetic effects and the results still remained significant. Taken together, these results strongly indicate the presence of immunological pressure exerted by the acquisition of the mentioned antiretroviral resistance mutations. Although results for subtype C were not conclusive, additional data from patients undergoing ARV treatment may provide the necessary data to determine negative correlation between DRMs and HLA in HIV subtype C.

5

Conclusionary Discussion

In the previous chapter, evidence was provided that allude to the existence of CTL epitopes generated by antiretroviral resistance mutations. It was shown that certain mutations commonly associated with antiretroviral resistance were diminished in patient sequence sets to which HLA type B*15 and B*48 were assigned. For both these HLA types, it appeared that the PR mutation L90M as well as the RT mutations RT M41L, L210W and T215Y (M140L, L309W, T314Y) were diminished. In an attempt to explain these diminished residues, computational methods were employed to predict the changes in MHC affinity and stability. The prediction results yielded higher affinity and stability scores and thus it can thought that the mutations may result in the generation of CTL epitopes. Furthermore, the B*15 and B*48 sets also indicated enriched residues that in turn allude to possible escape of these mutations. Phylogenetic correction identified founder effects in some of the HLA assigned patients, but could not account for the differences in frequencies of the mutations. In this chapter, the results obtained will be discussed in light of the positive impact of CTL epitope generating ARV resistance mutations in terms of lowering HIV viral load as discovered by other researchers and the implications for future treatment strategies to combat HIV infection.

5.1 Fitness and genetic hurdles of ARV resistance acquisition under induction of CTL response

Other research has shown that the transition from antiretroviral susceptible state to a resistance state can sometimes be challenging if there is pressure from CTL responses [Mueller *et al.*, 2011]. In that study, it was demonstrated how the acquisition of a PR mutation PR I47A, which provides resistance to LPV, was challenged by the presence of HLA B*1501. It is known that the PR 45-53 epitope, `KMIGGIGGF` (KM9), lies within this region. For Ile to mutate to Ala, requires at least two nucleotide substitutions and Val is a transition residue in this process. However, the I48V mutation still renders the KM9 epitope viable and emergent HIV strains within a B*1501+ host with this mutation can be quickly selected against [Mueller *et al.*, 2011].

A similar mechanism could be the reason for the diminished residues associated with the predicted LM9 and RQ10 epitopes. Evidence was provided here that allude to escape mutations, but it is unknown which of these mutations need to be acquired first in order to elude the immune response. Although the mutation PR I93L provides an escape to the PR 91-99 TQ9 epitope, coexistence of this mutation with the proposed LM9 escape mutations, Q92K/R seems to be diminished. For the RQ10 epitope, the acquisition of ARV resistance first needs to be discussed. According to the research by Rhee *et. al.*, the mutation RT T215Y precedes the mutations L210W and M41L [Rhee *et al.*, 2007]. The transition from Thr to Tyr requires two substitutions, the minimum cost being a $T \rightarrow S \rightarrow Y$ series of mutations, with an alternative being $T \rightarrow I \rightarrow F \rightarrow Y$. It was noted that the mutations 215F/I were enriched. The 215F mutation is a known ARV resistance mutation, the 215I mutation often only develops after the acquisition of 215F/Y. It was shown that the 215I mutation could provide a mechanism of CTL escape for the RQ10 epitope. It may also be that the 215I mutation acts as a transitional mutation to 215F/Y and enriched due to the selection against the 215F/Y variants.

It would be interesting to know what the fitness burden would be of the enriched and novel substitutions that occur in the LM9 and RQ10 regions would be on HIV-1. Further experimental analysis would be needed to assess this. It can be assumed that some fitness impact is imposed by the mutations Q92K/R, since they occur at very low frequencies. In a similar fashion, the Saquinavir protease inhibitor is subjected to a resistance mutation, PR G48V, which has been shown to negatively impact HIV fitness [Mammano *et al.*, 1998]. Unfortunately, access to replicative fitness data for HIV HIV subtype B determined by other researchers could not be obtained and thus this hypothesis could not be theoretically calculated [Hinkley *et al.*, 2011].

5.2 The importance of the substitutions PR 90M and RT 215Y/F

There are reasons why emphasis is placed on the three mutations, PR L90M and RT 215F/Y. Referring to Table 1.2, is is seen that PR L90M confers resistance to all but two of the nine listed HIV protease inhibitors. Similarly, RT T215F/Y provides resistance to all but two of the seven reverse transcriptase inhibitors. The observed frequencies of these mutations in the sequence set used here, were 0.42, 0.51 and 0.13. It also appears that these mutations are favourable precursors to other mutations that, in combination with them provide high levels of ARV resistance. However, even if immunological responses may slow the accumulation of these mutations, it is unlikely that they would completely prevent them, as can be demonstrated by the RT K103N mutation that is a ARV resistance mutation against efavirenz (EFV) and nevirapine (NVP) [Mahnke and Clifford, 2006]. Researchers have demonstrated that this mutation induces a CTL response, but it appears that there is no selection against this mutation in patients exhibiting a CTL response. Selection of a mutation could also be due to the level of selection pressure applied. For instance, the KF9 CTL epitope that lies in a region prone to accumulate ARV resistance mutations, very rarely does so in the absence of ARV pressure. Indeed, there appears to be a slight negative correlation in the acquisition of ARV resistance in regions where potent CTL epitopes overlap [Karlsson *et al.*, 2003; Brumme *et al.*, 2009]. The rationale is that alternative substitutions could be needed to either provide resistance to the ARVs or escape from the novel epitopes. An illustration of the ideal scenario is provided in

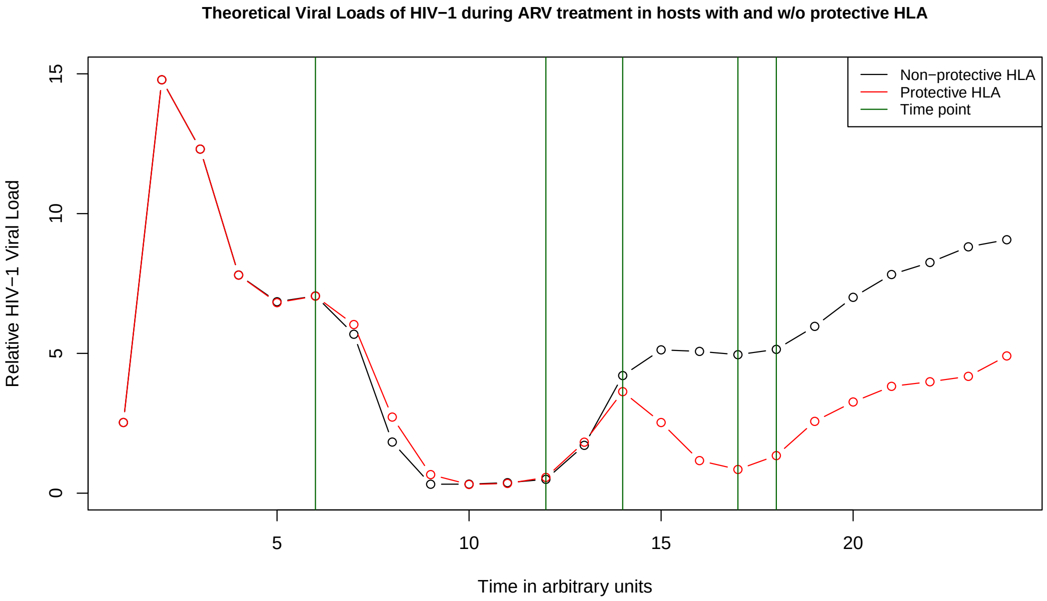


Figure 5.1: This figure depicts a theoretical model showing the HIV viral load in infected patients over time, both in arbitrary units. The black line represents viral load of a patient with a non-protective HLA allotype (patient #1) while the red line represents the HIV viral load in a patient with a protective HLA allotype (patient #2). The five green lines represent crucial time points, that represent the following: Time point 6 - the administration of of an ARV. Time 12 - the acquisition of a primary antiretroviral resistance mutation. Time point 14 - generation of a CTL response in patient #2 and subsequent drop in viral load. Time point 17 - acquisition of a mutation diminishing the CTL response in patient #2. Time point 18 - acquisition of a secondary ARV resistance mutation. Source: created by the author.

Figure 5.1. This figure illustrates the hypothetical time course of an HIV infection showing the viral loads of two patients at different time points. The first time point is associated with the administration of an ARV. A dramatic subsequent drop in viral load is observed. A primary ARV resistance mutation cause the viral load to increase. At time point 14, a novel CTL epitope in patient #2 (red line) developed, driving the viral load down. At the following time point a mutation resulting in a diminished CTL response is acquired in patient #2 and at the following time point another ARV resistance mutation is acquired. In the later time points, although the viral load in both patients #1 and #2 have increased, patient #2 retained a lower viral load.

5.3 Interaction between HLA and antiretroviral resistance mutations in light of vaccine design

Several studies have been performed arguing for the inclusion of antiretroviral variants of HIV proteins in vaccines and this has been extensively reviewed [Boberg and Isaguliants, 2008]. The theory is that immune responses can be augmented by priming the CTL responses against an antigen that may develop during the course of ARV treatment due to resistance mutations. One striking example is the emergence of a novel CTL epitope due to the acquisition of the Lamuvidine (3TC) mutation RT M184V. This mutation initially provides both resistance and escape to the 3TC drug, but eventually CTL clones reactive towards the new epitope expand and provide a measure of immune control [Schmitt *et al.*, 2000]. Other researchers have demonstrated that a generalized response against drug resistance RT proteins could be achieved in immunized mice [Isaguliants *et al.*, 2004]. Indeed, the drug resistant variants used for CTL stimulation included mutations within the RT 210-220 region, implying the inclusion of RT T215F/Y, although the specific CTL stimulatory epitopes were not determined, so results here could not be confirmed. The enhancement of response against the HLA A2 epitope PR 76-84 via a plausible processing mutation induced by L90M is worthy of consideration in HIV vaccine against drug resistant variants, especially considering that the I84V substitution commonly associated with drug resistance in all the protease inhibitors is also immunogenic. The I84V mutation is also highly correlated with L90M [Rhee *et al.*, 2007]. It may be argued that given these responses against these induced epitopes are elicited in any case in an individual, a vaccine or rather, a pre-exposure to these antigens may not augment the response. However, it should be remembered that secondary exposure to CTL epitopes result in rapid responses and that emergent epitopes in ARV treated individuals could make acquisition of CTL escape mutations to these epitopes difficult [Koup and Douek, 2011]. The other point to consider is that if a drug resistant variant is transmitted, the accepting host could already have acquired protection against these strains through a vaccine.

Still, the extremely polymorphic nature of the HIV genome is problematic when constructing vaccines. Other factors such as distribution of HLA allotypes and HIV subtypes need to be taken into account [Kawashima *et al.*, 2009; Chaudhari *et al.*, 2013]. For the results here, an interesting observation is made. The proposed epitope RQ10 spans the region RT 206-215. The consensus sequence of this region, is RQHLLKWGFT for HIV subtype B (analysed here), whereas HIV subtype A and C respectively show the consensus sequences RAHLLKWGFT and REHLLKWGFT. The mutation RT Q207E is a known escape mutation for subtype B*15, while Q207A was shown here to abolish or severely attenuate the binding capacity of the peptide to B*1501, B*1503, B*4801 and B*4802. It is interesting to note that HIV-1 HIV subtype A is prevalent in Asia together with the HLA allotype B*4801. Whether the RT 207A variant was accumulated partly due to B*4801 needs to be tested. The reason for the 207A variant instead of the 207E variant, is that Glu (E) is still a favourable residue for the binding pocket as position 2 for B*4801. The Glu residue at position two is an unfavourable binding residue for the HLA allotypes B*4802 and B*1503. The HLA allotype, B*1503 is found in populations of sub-Saharan African descent and HIV subtype C is by far the most prevalent HIV subtype in this region. The B*4802 HLA allotype, however, is found in native-American populations and HIV subtype B is main subtype found in this region. The LM9 epitope spanning PR 89-97 spans a region for which differences in the subtypes mainly occur at PR 93 and alternating between Ile and Leu. It should be noted, however, that in the subtype A and C sequences, the occurrence of L90M is far lower and thus although this substitution may elicit an immune response, the low accumulation of the mutation in subtype A and C may reduce it's candidacy as a vaccine target.

The sequences used here for HLA subtype assignment were only limited to HIV-1 subtype B. Due to the nature of the HLA subtype assignment procedure, which seems to be limited to subtype B as well as drug resistance data, the results could not be extrapolated to other HIV subtypes. Another issue with HLA subtype assignment is the diversity of the HLA allotypes belonging to the same subtype and seemingly agreement of MHC binding motif across different HLA types.

Further investigations of potential ARV resistance induced protective CTL epitopes is needed to elucidate a large enough set for treatment-augmented vaccines. Indeed, evidence provided here and by other researchers strongly imply that the HLA genotype of an infected individual should be taken into consideration prior to the design of a treatment regimen, by virtue of the expected mutations that can accumulate as a result of prolonged ARV treatment.

5.4 Further improvements of CTL epitope detection in HIV strains with acquired antiretroviral resistance

Although a lot of evidence was provided in this study for the existence of the LM9 and RQ10 epitopes induced by PR L90M and RT T215F/Y, one of the major drawbacks of the approach is that it might miss a lot of potential epitopes. Heavy reliance was placed on the assignment of HLA types according to substitutions observed in the *Pol* sequences due to the lack of HLA annotation in especially sequences from treatment-experienced patients. Added to this is that very little sequences were available from treatment-experienced patients that included other protein products of the HIV genome. However, the substitutions used to assign HLA types were deemed appropriate according to the p and q-values obtained for their HLA association from the study by Brumme *et. al.* [Brumme *et al.*, 2009]. Furthermore, with the exception of the mutation RT R211G, none of the mutations used to assign the HLA types were diminished in sequences from patients that are treatment-experienced. The effects of false HLA type assignment should also be evident by non-discovery of substitutions that differ significantly in frequency. The protocol used here, i.e. detection of diminished frequency of a substitution followed by providing evidence of CTL epitopes in the form of high MHC affinity and stability predictions is limited in various aspects, e.g. the mutational pressure exerted by a drug may override the immunological pressure encountered by the generation of novel epitopes [Mahnke and Clifford, 2006; Karlsson *et al.*, 2007]. Although a tool was recently created to predict to some degree the potential of an epitope to be immunogenic, this tool had a modest accuracy and thus the results are interpreted with caution. Lastly, it has to be taken into account that the whole HLA genotype of a patient can play a crucial role in determining the immunogenicity of a CTL epitope [Kosmrlj *et al.*, 2010].

Experimental assignment of HLA types for the patients would still be the best way to analyse frequency discrepancies between HLA types. Using tools such as MHCClust, HLA allotypes can be grouped into clusters that can be analysed as a whole.

The current methods for predicting potential epitopes are moderately good, but acquisition of additional experimental data in the form of proteasomal cleavage sites, MHC binding data for multiple HLA allotypes, MHC stability predictions and immunogenicity predictions could vastly improve the rate at which epitopes can be discovered and the mechanism of their escape be inferred [Groot and Berzofsky, 2004; Groot, 2006; Korber *et al.*, 2006]. It is very possible, given an improvement in accuracy of these tools that drug induced vaccine targets can be discovered for HIV-1. Perhaps future treatments will include HLA genotype of a patient into account to augment ARV treatment regiments.

5.5 Conclusion

This study provided novel methods for the detection of potential CTL epitopes elucidated by the mutations PR L90M and RT T215Y/F that are both implicated in resistance to many HIV protease and reverse transcriptase inhibitors respectively. In this endeavour, unique methods were used in the assignment of HLA types to patients that are otherwise unannotated. To compensate for the possibility of lower occuring epistatically-interacting ARV mutations in the assigned groups as a whole, both a linear model and robust method were created to compensate for the diminished frequencies of L90M and T215F/Y in the HLA type B*48 assigned sequence sets. Even after compensation, these mutations still differed significantly in frequency. Comparing the prediction results of MHC affinity and stability in the subtype B consensus HIV-1 Pol protein, it was determined that these mutations may have impact on the generation of novel CTL epitopes. Two epitopes, LM9, `LMTQIGCTL` and `RQHLLRWGF(Y/F)` were discovered. Subsequent analyses revealed the enrichment of residues in these sequence groups. It is proposed that these mutations could act as CTL escape mutations. Taking into account the importance of members of the B*15 HLA type in the management of HIV-1 infection, it is hoped that this new data may be helpful in determining treatment augmentation of HIV-1 infected individuals harbouring the appropriate HLA allotypes. The author encourages other researchers to continue the detection of potential CTL epitopes generated by antiretroviral resistance mutations. Depending on the nature of these epitopes, i.e. the immunogenicity, the conservative nature of the region where it occurs and the HLA allotype coverage, it could have a profoundly positive impact on future vaccine and treatment designs that do take antiretroviral resistance mutations into account.

Bibliography

[Abecasis *et al.* 2005] Abecasis, A. B., Deforche, K., Snoeck, J., Bacheler, L. T., McKenna, P., Carvalho, A. P., Gomes, P., Camacho, R. J., and Vandamme, A.-M. [2005] Protease mutation M89I/V is linked to therapy failure in patients infected with the HIV-1 non-B subtypes C, F or G. *AIDS*, **19** [16], 1799–1806.

[Allen *et al.* 2005] Allen, T. M., Altfeld, M., Geer, S. C., Kalife, E. T., Moore, C., O'sullivan, K. M., Desouza, I., Feeney, M. E., Eldridge, R. L., Maier, E. L., Kaufmann, D. E., Lahaie, M. P., Reyor, L., Tanzi, G., Johnston, M. N., Brander, C., Draenert, R., Rockstroh, J. K., Jessen, H., Rosenberg, E. S., Mallal, S. A., and Walker, B. D. [2005] Selective escape from CD8+ T-cell responses represents a major driving force of human immunodeficiency virus type 1 (HIV-1) sequence diversity and reveals constraints on HIV-1 evolution. *J Virol*, **79** [21], 13239–13249.

[Altfeld *et al.* 2003] Altfeld, M., Addo, M. M., Rosenberg, E. S., Hecht, F. M., Lee, P. K., Vogel, M., Yu, X. G., Draenert, R., Johnston, M. N., Strick, D., Allen, T. M., Feeney, M. E., Kahn, J. O., Sekaly, R. P., Levy, J. A., Rockstroh, J. K., Goulder, P. J., and Walker, B. D. [2003] Influence of HLA-B57 on clinical presentation and viral control during acute HIV-1 infection. *AIDS*, **17** [18], 2581–2591.

[Amanna *et al.* 2007] Amanna, I. J., Carlson, N. E., and Slifka, M. K. [2007] Duration of humoral immunity to common viral and vaccine antigens. *N Engl J Med*, **357** [19], 1903–1915.

[Ammaranond and Sanguansittianan 2012] Ammaranond, P. and Sanguansittianan, S. [2012] Mechanism of HIV antiretroviral drugs progress toward drug resistance. *Fundam Clin Pharmacol*, **26** [1], 146–161.

[Auewarakul *et al.* 2005] Auewarakul, P., Wacharapornin, P., Srichatrapimuk, S., Chutipongtanate, S., and Puthavathana, P. [2005] Uncoating of HIV-1 requires cellular activation. *Virology*, **337** [1], 93–101.

[Bangsberg *et al.* 2003] Bangsberg, D. R., Charlebois, E. D., Grant, R. M., Holodniy, M., Deeks, S. G., Perry, S., Conroy, K. N., Clark, R., Guzman, D., Zolopa, A., and Moss, A. [2003] High levels of adherence do not prevent accumulation of HIV drug resistance mutations. *AIDS*, **17** [13], 1925–1932.

[Baxter *et al.* 2006] Baxter, J. D., Schapiro, J. M., Boucher, C. A. B., Kohlbrenner, V. M., Hall, D. B., Scherer, J. R., and Mayers, D. L. [2006] Genotypic changes in human immunodeficiency virus type 1 protease associated with reduced susceptibility and virologic response to the protease inhibitor tipranavir. *J Virol*, **80** [21], 10794–10801.

[Bhattacharya *et al.* 2007] Bhattacharya, T., Daniels, M., Heckerman, D., Foley, B., Frahm, N., Kadie, C., Carlson, J., Yusim, K., McMahon, B., Gaschen, B., Mallal, S., Mullins, J. I., Nickle, D. C., Herbeck, J., Rousseau, C., Learn, G. H., Miura, T., Brander, C., Walker, B., and Korber, B. [2007] Founder effects in the assessment of HIV polymorphisms and HLA allele associations. *Science*, **315** [5818], 1583–1586.

[Boberg and Isaguliants 2008] Boberg, A. and Isaguliants, M. [2008] Vaccination against drug resistance in HIV infection. *Expert Rev Vaccines*, **7** [1], 131–145.

[Brumme *et al.* 2009] Brumme, Z. L., John, M., Carlson, J. M., Brumme, C. J., Chan, D., Brockman, M. A., Swenson, L. C., Tao, I., Szeto, S., Rosato, P., Sela, J., Kadie, C. M., Frahm, N., Brander, C., Haas, D. W., Riddler, S. A., Haubrich, R., Walker, B. D., Harrigan, P. R., Heckerman, D., and Mallal, S. [2009] HLA-associated immune escape pathways in HIV-1 subtype B Gag, Pol and Nef proteins. *PLoS One*, **4** [8], e6687.

[Burton *et al.* 2004] Burton, D. R., Desrosiers, R. C., Doms, R. W., Koff, W. C., Kwong, P. D., Moore, J. P., Nabel, G. J., Sodroski, J., Wilson, I. A., and Wyatt, R. T. [2004] HIV vaccine design and the neutralizing antibody problem. *Nat Immunol*, **5** [3], 233–236.

[Calis *et al.* 2013] Calis, J. J. A., Maybeno, M., Greenbaum, J. A., Weiskopf, D., De Silva, A. D., Sette, A., Keşmir, C., and Peters, B. [2013] Properties of MHC class I presented peptides that enhance immunogenicity. *PLoS Comput Biol*, **9** [10], e1003266.

[Cardinaud *et al.* 2011] Cardinaud, S., Consiglieri, G., Bouziat, R., Urrutia, A., Graff-Dubois, S., Fourati, S., Malet, I., Guergnon, J., Guihot, A., Katlama, C., Autran, B., van Endert, P., Lemonnier, F. A., Appay, V., Schwartz, O., Kloetzel, P. M., and Moris, A. [2011] CTL Escape Mediated by Proteasomal Destruction of an HIV-1 Cryptic Epitope. *PLoS Pathog*, **7** [5], e1002049.

[Carlson *et al.* 2012] Carlson, J. M., Brumme, C. J., Martin, E., Listgarten, J., Brockman, M. A., Le, A. Q., Chui, C. K. S., Cotton, L. A., Knapp, D. J. H. F., Riddler, S. A., Haubrich, R., Nelson, G., Pfeifer, N., Deziel, C. E., Heckerman, D., Apps, R., Carrington, M., Mallal, S., Harrigan, P. R., John, M., Brumme, Z. L., and , I. H. I. V. A. C. [2012] Correlates of protective cellular immunity revealed by analysis of population-level immune escape pathways in HIV-1. *J Virol*, **86** [24], 13202–13216.

[Carlson *et al.* 2014] Carlson, J. M., Schaefer, M., Monaco, D. C., Batorsky, R., Claiborne, D. T., Prince, J., Deymier, M. J., Ende, Z. S., Klatt, N. R., DeZiel, C. E., Lin, T.-H., Peng, J., Seese, A. M., Shapiro, R., Frater, J., Ndung'u, T., Tang, J., Goepfert, P., Gilmour, J., Price, M. A., Kilembe, W., Heckerman, D., Goulder, P. J. R., Allen, T. M., Allen, S., and Hunter, E. [2014] HIV transmission. Selection bias at the heterosexual HIV-1 transmission bottleneck. *Science*, **345** [6193], 1254031.

[Cerundolo *et al.* 1991] Cerundolo, V., Elliott, T., Elvin, J., Bastin, J., Rammensee, H. G., and Townsend, A. [1991] The binding affinity and dissociation rates of peptides for class I major histocompatibility complex molecules. *Eur J Immunol*, **21** [9], 2069–2075.

[Chan and Kim 1998] Chan, D. C. and Kim, P. S. [1998] HIV entry and its inhibition. *Cell*, **93** [5], 681–684.

[Chang *et al.* 2005] Chang, S.-C., Momburg, F., Bhutani, N., and Goldberg, A. L. [2005] The ER aminopeptidase, ERAP1, trims precursors to lengths of MHC class I peptides by a ”molecular ruler” mechanism. *Proc Natl Acad Sci U S A*, **102** [47], 17107–17112.

[Chaudhari *et al.* 2013] Chaudhari, D. V., Chavan, V. R., Ahir, S. P., Kerkar, S. C., Mehta, P. R., and Mania-Pramanik, J. [2013] Human leukocyte antigen B distribution in HIV discordant cohort from India. *Immunol Lett*, **156** [1-2], 1–6.

[Chiu and Davies 2004] Chiu, T. K. and Davies, D. R. [2004] Structure and function of HIV-1 integrase. *Curr Top Med Chem*, **4** [9], 965–977.

[Clerici *et al.* 1989] Clerici, M., Stocks, N. I., Zajac, R. A., Boswell, R. N., Lucey, D. R., Via, C. S., and Shearer, G. M. [1989] Detection of three distinct patterns of T helper cell dysfunction in asymptomatic, human immunodeficiency virus-seropositive patients. Independence of CD4+ cell numbers and clinical staging. *J Clin Invest*, **84** [6], 1892–1899.

[Clifford *et al.* 1999] Clifford, D. B., Yiannoutsos, C., Glicksman, M., Simpson, D. M., Singer, E. J., Piliero, P. J., Marra, C. M., Francis, G. S., McArthur, J. C., Tyler, K. L., Tselis, A. C., and Hyslop, N. E. [1999] HAART improves prognosis in HIV-associated progressive multifocal leukoencephalopathy. *Neurology*, **52** [3], 623–625.

[Coffin and Swanstrom 2013] Coffin, J. and Swanstrom, R. [2013] HIV pathogenesis: dynamics and genetics of viral populations and infected cells. *Cold Spring Harb Perspect Med*, **3** [1], a012526.

[Cox *et al.* 1994] Cox, A. L., Skipper, J., Chen, Y., Henderson, R. A., Darrow, T. L., Shabanowitz, J., Engelhard, V. H., Hunt, D. F., and Slingluff, Jr, C. [1994] Identification of a peptide recognized by five melanoma-specific human cytotoxic T cell lines. *Science*, **264** [5159], 716–719.

[Crawford *et al.* 2007] Crawford, H., Prado, J. G., Leslie, A., Hué, S., Honeyborne, I., Reddy, S., van der Stok, M., Mncube, Z., Brander, C., Rousseau, C., Mullins, J. I., Kaslow, R., Goepfert, P., Allen, S., Hunter, E., Mulenga, J., Kiepiela, P., Walker, B. D., and Goulder, P. J. R. [2007] Compensatory mutation partially restores fitness and delays reversion of escape mutation within the immunodominant HLA-B*5703-restricted Gag epitope in chronic human immunodeficiency virus type 1 infection. *J Virol*, **81** [15], 8346–8351.

[Cresswell *et al.* 2005] Cresswell, P., Ackerman, A. L., Giodini, A., Peaper, D. R., and Wearsch, P. A. [2005] Mechanisms of MHC class I-restricted antigen processing and cross-presentation. *Immunol Rev*, **207**, 145–157.

[Cummins and Badley 2010] Cummins, N. W. and Badley, A. D. [2010] Mechanisms of HIV-associated lymphocyte apoptosis: 2010. *Cell Death Dis*, **1**, e99.

[Davis and Bjorkman 1988] Davis, M. M. and Bjorkman, P. J. [1988] T-cell antigen receptor genes and T-cell recognition. *Nature*, **334** [6181], 395–402.

[De Clercq 2009] De Clercq, E. [2009] Anti-HIV drugs: 25 compounds approved within 25 years after the discovery of HIV. *Int J Antimicrob Agents*, **33** [4], 307–320.

[Decroly *et al.* 1994] Decroly, E., Vandenbranden, M., Ruysschaert, J. M., Cogniaux, J., Jacob, G. S., Howard, S. C., Marshall, G., Kompelli, A., Basak, A., and Jean, F. [1994] The convertases furin and PC1 can both cleave the human immunodeficiency virus (HIV)-1 envelope glycoprotein gp160 into gp120 (HIV-1 SU) and gp41 (HIV-I TM). *J Biol Chem*, **269** [16], 12240–12247.

[Dempsey *et al.* 2003] Dempsey, P. W., Vaidya, S. A., and Cheng, G. [2003] The art of war: Innate and adaptive immune responses. *Cell Mol Life Sci*, **60** [12], 2604–2621.

[Doualla-Bell *et al.* 2004] Doualla-Bell, F., Gaseitsiwe, S., Ndungú, T., Modukanele, M., Peter, T., Novitsky, V., Ndwapi, N., Tendani, G., Avalos, A., Wester, W., Bussmann, H., Cardiello, P., Marlink, R., Moffat, H., Thior, I., Wainberg, M. A., and Essex, M. [2004] Mutations and polymorphisms associated with antiretroviral drugs in HIV-1C-infected African patients. *Antivir Chem Chemother*, **15** [4], 189–200.

[Duh *et al.* 1989] Duh, E. J., Maury, W. J., Folks, T. M., Fauci, A. S., and Rabson, A. B. [1989] Tumor necrosis factor alpha activates human immunodeficiency virus type 1 through induction of nuclear factor binding to the NF-kappa B sites in the long terminal repeat. *Proc Natl Acad Sci U S A*, **86** [15], 5974–5978.

[Eng 2005] Eng, J. [2005] Receiver operating characteristic analysis: a primer. *Acad Radiol*, **12** [7], 909–916.

[Felber *et al.* 1989] Felber, B. K., Hadzopoulou-Cladaras, M., Cladaras, C., Copeland, T., and Pavlakis, G. N. [1989] rev protein of human immunodeficiency virus type 1 affects the stability and transport of the viral mRNA. *Proc Natl Acad Sci U S A*, **86** [5], 1495–1499.

[Fischer *et al.* 1995] Fischer, U., Huber, J., Boelens, W. C., Mattaj, I. W., and Lührmann, R. [1995] The HIV-1 Rev activation domain is a nuclear export signal that accesses an export pathway used by specific cellular RNAs. *Cell*, **82** [3], 475–483.

[Flajnik and Kasahara 2010] Flajnik, M. F. and Kasahara, M. [2010] Origin and evolution of the adaptive immune system: genetic events and selective pressures. *Nat Rev Genet*, **11** [1], 47–59.

[Frahm *et al.* 2006] Frahm, N., Kiepiela, P., Adams, S., Linde, C. H., Hewitt, H. S., Sango, K., Feeney, M. E., Addo, M. M., Lichterfeld, M., Lahaie, M. P., Pae, E., Wurcel, A. G., Roach, T., St John, M. A., Altfeld, M., Marincola, F. M., Moore, C., Mallal, S., Carrington, M., Heckerman, D., Allen, T. M., Mullins, J. I., Korber, B. T., Goulder, P. J. R., Walker, B. D., and Brander, C. [2006] Control of human immunodeficiency virus replication by cytotoxic T lymphocytes targeting subdominant epitopes. *Nat Immunol*, **7** [2], 173–178.

[Frankel and Young 1998] Frankel, A. D. and Young, J. A. [1998] HIV-1: fifteen proteins and an RNA. *Annu Rev Biochem*, **67**, 1–25.

[Gaczynska *et al.* 1996] Gaczynska, M., Goldberg, A. L., Tanaka, K., Hendil, K. B., and Rock, K. L. [1996] Proteasome subunits X and Y alter peptidase activities in opposite ways to the interferon-gamma-induced subunits LMP2 and LMP7. *J Biol Chem*, **271** [29], 17275–17280.

[Gao *et al.* 2001] Gao, X., Nelson, G. W., Karacki, P., Martin, M. P., Phair, J., Kaslow, R., Goedert, J. J., Buchbinder, S., Hoots, K., Vlahov, D., O'Brien, S. J., and Carrington, M. [2001] Effect of a single amino acid change in MHC class I molecules on the rate of progression to AIDS. *N Engl J Med*, **344** [22], 1668–1675.

[Garber *et al.* 2004] Garber, D. A., Silvestri, G., and Feinberg, M. B. [2004] Prospects for an AIDS vaccine: three big questions, no easy answers. *Lancet Infect Dis*, **4** [7], 397–413.

[Gelderblom *et al.* 1987] Gelderblom, H. R., Hausmann, E. H., Ozel, M., Pauli, G., and Koch, M. A. [1987] Fine structure of human immunodeficiency virus (HIV) and immunolocalization of structural proteins. *Virology*, **156** [1], 171–176.

[Ginodi *et al.* 2008] Ginodi, I., Vider-Shalit, T., Tsaban, L., and Louzoun, Y. [2008] Precise score for the prediction of peptides cleaved by the proteasome. *Bioinformatics*, **24** [4], 477–483.

[Groot 2006] Groot, A. S. D. [2006] Immunomics: discovering new targets for vaccines and therapeutics. *Drug Discov Today*, **11** [5-6], 203–209.

[Groot and Berzofsky 2004] Groot, A. S. D. and Berzofsky, J. A. [2004] From genome to vaccine–new immunoinformatics tools for vaccine design. *Methods*, **34** [4], 425–428.

[Haseltine 1991] Haseltine, W. A. [1991] Molecular biology of the human immunodeficiency virus type 1. *FASEB J*, **5** [10], 2349–2360.

[Hemelaar *et al.* 2011] Hemelaar, J., Gouws, E., Ghys, P. D., Osmanov, S., , W. H. O.-U. N. A. I. D. S. N. f. H. I. V. I., and Characterisation [2011] Global trends in molecular epidemiology of HIV-1 during 2000-2007. *AIDS*, **25** [5], 679–689.

[Hemelaar *et al.* 2006] Hemelaar, J., Gouws, E., Ghys, P. D., and Osmanov, S. [2006] Global and regional distribution of HIV-1 genetic subtypes and recombinants in 2004. *AIDS*, **20** [16], W13–W23.

[Hinkley *et al.* 2011] Hinkley, T., Martins, J., Chappey, C., Haddad, M., Stawiski, E., Whitcomb, J. M., Petropoulos, C. J., and Bonhoeffer, S. [2011] A systems analysis of mutational effects in HIV-1 protease and reverse transcriptase. *Nat Genet*, **43** [5], 487–489.

[Isaguliants *et al.* 2004] Isaguliants, M. G., Zuber, B., Boberg, A., Sjöstrand, D., Belikov, S. V., Rollman, E., Zuber, A. K., Rechinsky, V. O., Rytting, A.-S., Källander, C. F. R., Hinkula, J., Kochetkov, S. N., Liu, M., and Wahren, B. [2004] Reverse transcriptase-based DNA vaccines against drug-resistant HIV-1 tested in a mouse model. *Vaccine*, **22** [13-14], 1810–1819.

[Iversen *et al.* 2006] Iversen, A. K. N., Stewart-Jones, G., Learn, G. H., Christie, N., Sylvester-Hviid, C., Armitage, A. E., Kaul, R., Beattie, T., Lee, J. K., Li, Y., Chotiyarnwong, P., Dong, T., Xu, X., Luscher, M. A., MacDonald, K., Ullum, H., Klarlund-Pedersen, B., Skinhøj, P., Fugger, L., Buus, S., Mullins, J. I., Jones, E. Y., van der Merwe, P. A., and McMichael, A. J. [2006] Conflicting selective forces affect T cell receptor contacts in an immunodominant human immunodeficiency virus epitope. *Nat Immunol*, **7** [2], 179–189.

[Janeway and Medzhitov 2002] Janeway, Jr, C. A. and Medzhitov, R. [2002] Innate immune recognition. *Annu Rev Immunol*, **20**, 197–216.

[Jin and Wang 2003] Jin, P. and Wang, E. [2003] Polymorphism in clinical immunology - From HLA typing to immunogenetic profiling. *J Transl Med*, **1** [1], 8.

[John *et al.* 2005] John, M., Moore, C. B., James, I. R., and Mallal, S. A. [2005] Interactive selective pressures of HLA-restricted immune responses and antiretroviral drugs on HIV-1. *Antivir Ther*, **10** [4], 551–555.

[Jørgensen *et al.* 2014] Jørgensen, K. W., Rasmussen, M., Buus, S., and Nielsen, M. [2014] NetMHCstab - predicting stability of peptide-MHC-I complexes; impacts for cytotoxic T lymphocyte epitope discovery. *Immunology*, **141** [1], 18–26.

[Karlsson *et al.* 2007] Karlsson, A. C., Chapman, J. M., Heiken, B. D., Hoh, R., Kallas, E. G., Martin, J. N., Hecht, F. M., Deeks, S. G., and Nixon, D. F. [2007] Antiretroviral drug therapy alters the profile of human immunodeficiency virus type 1-specific T-cell responses and shifts the immunodominant cytotoxic T-lymphocyte response from Gag to Pol. *J Virol*, **81** [20], 11543–11548.

[Karlsson *et al.* 2003] Karlsson, A. C., Deeks, S. G., Barbour, J. D., Heiken, B. D., Younger, S. R., Hoh, R., Lane, M., Sällberg, M., Ortiz, G. M., Demarest, J. F., Liegler, T., Grant, R. M., Martin, J. N., and Nixon, D. F. [2003] Dual pressure from antiretroviral therapy and cell-mediated immune response on the human immunodeficiency virus type 1 protease gene. *J Virol*, **77** [12], 6743–6752.

[Karosiene *et al.* 2012] Karosiene, E., Lundegaard, C., Lund, O., and Nielsen, M. [2012] NetMHCcons: a consensus method for the major histocompatibility complex class I predictions. *Immunogenetics*, **64** [3], 177–186.

[Kaslow *et al.* 1990] Kaslow, R. A., Duquesnoy, R., VanRaden, M., Kingsley, L., Marrari, M., Friedman, H., Su, S., Saah, A. J., Detels, R., and Phair, J. [1990] A1, Cw7, B8, DR3 HLA antigen combination associated with rapid decline of T-helper lymphocytes in HIV-1 infection. A report from the Multicenter AIDS Cohort Study. *Lancet*, **335** [8695], 927–930.

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

[Keşmir *et al.* 2002] Keşmir, C., Nussbaum, A. K., Schild, H., Detours, V., and Brunak, S. [2002] Prediction of proteasome cleavage motifs by neural networks. *Protein Eng*, **15** [4], 287–296.

[Khan *et al.* 2000] Khan, A. R., Baker, B. M., Ghosh, P., Biddison, W. E., and Wiley, D. C. [2000] The structure and stability of an HLA-A*0201/octameric tax peptide complex with an empty conserved peptide-N-terminal binding site. *J Immunol*, **164** [12], 6398–6405.

[Koibuchi *et al.* 2005] Koibuchi, T., Allen, T. M., Lichterfeld, M., Mui, S. K., O'Sullivan, K. M., Trocha, A., Kalams, S. A., Johnson, R. P., and Walker, B. D. [2005] Limited sequence evolution within persistently targeted CD8 epitopes in chronic human immunodeficiency virus type 1 infection. *J Virol*, **79** [13], 8171–8181.

[Korber *et al.* 2006] Korber, B., LaBute, M., and Yusim, K. [2006] Immunoinformatics comes of age. *PLoS Comput Biol*, **2** [6], e71.

[Kosmrlj *et al.* 2010] Kosmrlj, A., Read, E. L., Qi, Y., Allen, T. M., Altfeld, M., Deeks, S. G., Pereyra, F., Carrington, M., Walker, B. D., and Chakraborty, A. K. [2010] Effects of thymic selection of the T-cell repertoire on HLA class I-associated control of HIV infection. *Nature*, **465** [7296], 350–354.

[Koup and Douek 2011] Koup, R. A. and Douek, D. C. [2011] Vaccine design for CD8 T lymphocyte responses. *Cold Spring Harb Perspect Med*, **1** [1], a007252.

[Kulkarni *et al.* 2012] Kulkarni, R., Babaoglu, K., Lansdon, E. B., Rimsky, L., Van Eygen, V., Picchio, G., Svarovskaia, E., Miller, M. D., and White, K. L. [2012] The HIV-1 reverse transcriptase M184I mutation enhances the E138K-associated resistance to rilpivirine and decreases viral fitness. *J Acquir Immune Defic Syndr*, **59** [1], 47–54.

[Kwong *et al.* 1998] Kwong, P. D., Wyatt, R., Robinson, J., Sweet, R. W., Sodroski, J., and Hendrickson, W. A. [1998] Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. *Nature*, **393** [6686], 648–659.

[Larke *et al.* 2007] Larke, N., Im, E.-J., Wagner, R., Williamson, C., Williamson, A.-L., McMichael, A. J., and Hanke, T. [2007] Combined single-clade candidate HIV-1 vaccines induce T cell responses limited by multiple forms of in vivo immune interference. *Eur J Immunol*, **37** [2], 566–577.

[Lenassi *et al.* 2010] Lenassi, M., Cagney, G., Liao, M., Vaupotic, T., Bartholomeeusen, K., Cheng, Y., Krogan, N. J., Plemenitas, A., and Peterlin, B. M. [2010] HIV Nef is secreted in exosomes and triggers apoptosis in bystander CD4+ T cells. *Traffic*, **11** [1], 110–122.

[Li *et al.* 2011] Li, F., Finnefrock, A. C., Dubey, S. A., Korber, B. T. M., Szinger, J., Cole, S., McElrath, M. J., Shiver, J. W., Casimiro, D. R., Corey, L., and Self, S. G. [2011] Mapping HIV-1 vaccine induced T-cell responses: bias towards less-conserved regions and potential impact on vaccine efficacy in the Step study. *PLoS One*, **6** [6], e20479.

[Li and Raghavan 2010] Li, X. C. and Raghavan, M. [2010] Structure and function of major histocompatibility complex class I antigens. *Curr Opin Organ Transplant*, **15** [4], 499–504.

[Lieberman 2010] Lieberman, J. [2010] Anatomy of a murder: how cytotoxic T cells and NK cells are activated, develop, and eliminate their targets. *Immunol Rev*, **235** [1], 5–9.

[Lin *et al.* 2008] Lin, H. H., Ray, S., Tongchusak, S., Reinherz, E. L., and Brusic, V. [2008] Evaluation of MHC class I peptide binding prediction servers: applications for vaccine research. *BMC Immunol*, **9**, 8.

[Lorin *et al.* 2005] Lorin, C., Delebecque, F., Labrousse, V., Da Silva, L., Lemonnier, F., Brahic, M., and Tangy, F. [2005] A recombinant live attenuated measles vaccine vector primes effective HLA-A0201-restricted cytotoxic T lymphocytes and broadly neutralizing antibodies against HIV-1 conserved epitopes. *Vaccine*, **23** [36], 4463–4472.

[Lund *et al.* 2004] Lund, O., Nielsen, M., Kesmir, C., Petersen, A. G., Lundegaard, C., Worning, P., Sylvester-Hvid, C., Lamberth, K., Røder, G., Justesen, S., Buus, S., and Brunak, S. [2004] Definition of supertypes for HLA molecules using clustering of specificity matrices. *Immunogenetics*, **55** [12], 797–810.

[Lundegaard *et al.* 2010] Lundegaard, C., Lund, O., Buus, S., and Nielsen, M. [2010] Major histocompatibility complex class I binding predictions as a tool in epitope discovery. *Immunology*, **130** [3], 309–318.

[Lundegaard *et al.* 2008] Lundegaard, C., Lund, O., and Nielsen, M. [2008] Accurate approximation method for prediction of class I MHC affinities for peptides of length 8, 10 and 11 using prediction tools trained on 9mers. *Bioinformatics*, **24** [11], 1397–1398.

[Luo and Taylor 1990] Luo, G. X. and Taylor, J. [1990] Template switching by reverse transcriptase during DNA synthesis. *J Virol*, **64** [9], 4321–4328.

[Mackie *et al.* 2010] Mackie, N. E., Phillips, A. N., Kaye, S., Booth, C., and Geretti, A.-M. [2010] Antiretroviral drug resistance in HIV-1-infected patients with low-level viremia. *J Infect Dis*, **201** [9], 1303–1307.

[Mahnke and Clifford 2006] Mahnke, L. and Clifford, D. [2006] Cytotoxic T cell recognition of an HIV-1 reverse transcriptase variant peptide incorporating the K103N drug resistance mutation. *AIDS Res Ther*, **3**, 21.

[Mammano *et al.* 1998] Mammano, F., Petit, C., and Clavel, F. [1998] Resistance-associated loss of viral fitness in human immunodeficiency virus type 1: phenotypic analysis of protease and gag coevolution in protease inhibitor-treated patients. *J Virol*, **72** [9], 7632–7637.

[Martin Thomsen 2014] Martin Thomsen, Claus Lundegaard, S. B. [2014] MHCcluster, a method for functional clustering of MHC molecules. *Unpublished*.

[Mason *et al.* 2004] Mason, R. D., Bowmer, M. I., Howley, C. M., Gallant, M., Myers, J. C. E., and Grant, M. D. [2004] Antiretroviral drug resistance mutations sustain or enhance CTL recognition of common HIV-1 Pol epitopes. *J Immunol*, **172** [11], 7212–7219.

[Meyaard *et al.* 1994] Meyaard, L., Otto, S. A., Hooibrink, B., and Miedema, F. [1994] Quantitative analysis of CD4+ T cell function in the course of human immunodeficiency virus infection. Gradual decline of both naive and memory alloreactive T cells. *J Clin Invest*, **94** [5], 1947–1952.

[Migueles *et al.* 2000] Migueles, S. A., Sabbaghian, M. S., Shupert, W. L., Bettinotti, M. P., Marincola, F. M., Martino, L., Hallahan, C. W., Selig, S. M., Schwartz, D., Sullivan, J., and Connors, M. [2000] HLA B*5701 is highly associated with restriction of virus replication in a subgroup of HIV-infected long term nonprogressors. *Proc Natl Acad Sci U S A*, **97** [6], 2709–2714.

[Mizel 1982] Mizel, S. B. [1982] Interleukin l and T Cell Activation. *Immunological Reviews*, **63** [1], 51–72.

[Moir and Fauci 2009] Moir, S. and Fauci, A. S. [2009] B cells in HIV infection and disease. *Nat Rev Immunol*, **9** [4], 235–245.

[Mueller *et al.* 2007] Mueller, S. M., Schaetz, B., Eismann, K., Bergmann, S., Bauerle, M., Schmitt-Haendle, M., Walter, H., Schmidt, B., Korn, K., Sticht, H., Spriewald, B., Harrer, E. G., and Harrer, T. [2007] Dual selection pressure by drugs and HLA class I-restricted immune responses on human immunodeficiency virus type 1 protease. *J Virol*, **81** [6], 2887–2898.

[Mueller *et al.* 2011] Mueller, S. M., Spriewald, B. M., Bergmann, S., Eismann, K., Leykauf, M., Korn, K., Walter, H., Schmidt, B., Arnold, M.-L., Harrer, E. G., Kaiser, R., Schweitzer, F., Braun, P., Reuter, S., Jaeger, H., Wolf, E., Brockmeyer, N. H., Jansen, K., Michalik, C., Harrer, T., and , G. C. N. f. H. I. V. I. D. S. [2011] Influence of major HIV-1 protease inhibitor resistance mutations on CTL recognition. *J Acquir Immune Defic Syndr*, **56** [2], 109–117.

[Nielsen *et al.* 2007] Nielsen, M., Lundegaard, C., Blicher, T., Lamberth, K., Harndahl, M., Justesen, S., Røder, G., Peters, B., Sette, A., Lund, O., and Buus, S. [2007] NetMHC-pan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. *PLoS One*, **2** [8], e796.

[Nielsen *et al.* 2005] Nielsen, M., Lundegaard, C., Lund, O., and Keşmir, C. [2005] The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage. *Immunogenetics*, **57** [1-2], 33–41.

[Nielsen *et al.* 2004] Nielsen, M., Lundegaard, C., Worning, P., Hvid, C. S., Lamberth, K., Buus, S., Brunak, S., and Lund, O. [2004] Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach. *Bioinformatics*, **20** [9], 1388–1397.

[Nielsen *et al.* 2003] Nielsen, M., Lundegaard, C., Worning, P., Lauemøller, S. L., Lamberth, K., Buus, S., Brunak, S., and Lund, O. [2003] Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci*, **12** [5], 1007–1017.

[Pamer and Cresswell 1998] Pamer, E. and Cresswell, P. [1998] Mechanisms of MHC class I–restricted antigen processing. *Annu Rev Immunol*, **16**, 323–358.

[Pancer and Cooper 2006] Pancer, Z. and Cooper, M. D. [2006] The evolution of adaptive immunity. *Annu Rev Immunol*, **24**, 497–518.

[Paradis *et al.* 2004] Paradis, E., Claude, J., and Strimmer, K. [2004] APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*, **20** [2], 289–290.

[Parker *et al.* 1994] Parker, K. C., Bednarek, M. A., and Coligan, J. E. [1994] Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. *J Immunol*, **152** [1], 163–175.

[Price *et al.* 2010] Price, M. N., Dehal, P. S., and Arkin, A. P. [2010] FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One*, **5** [3], e9490.

[Pupko *et al.* 2000] Pupko, T., Pe'er, I., Shamir, R., and Graur, D. [2000] A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Mol Biol Evol*, **17** [6], 890–896.

[Purcell and Martin 1993] Purcell, D. F. and Martin, M. A. [1993] Alternative splicing of human immunodeficiency virus type 1 mRNA modulates viral protein expression, replication, and infectivity. *J Virol*, **67** [11], 6365–6378.

[Rambaut *et al.* 2004] Rambaut, A., Posada, D., Crandall, K. A., and Holmes, E. C. [2004] The causes and consequences of HIV evolution. *Nat Rev Genet*, **5** [1], 52–61.

[Rammensee *et al.* 1995] Rammensee, H. G., Friede, T., and Stevanovi?c, S. [1995] MHC ligands and peptide motifs: first listing. *Immunogenetics*, **41** [4], 178–228.

[Rhee *et al.* 2003] Rhee, S.-Y., Gonzales, M. J., Kantor, R., Betts, B. J., Ravela, J., and Shafer, R. W. [2003] Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res*, **31** [1], 298–303.

[Rhee *et al.* 2007] Rhee, S.-Y., Liu, T. F., Holmes, S. P., and Shafer, R. W. [2007] HIV-1 subtype B protease and reverse transcriptase amino acid covariation. *PLoS Comput Biol*, **3** [5], e87.

[Robertson *et al.* 1995] Robertson, D. L., Sharp, P. M., McCutchan, F. E., and Hahn, B. H. [1995] Recombination in HIV-1. *Nature*, **374** [6518], 124–126.

[Romani *et al.* 2010] Romani, B., Engelbrecht, S., and Glashoff, R. H. [2010] Functions of Tat: the versatile protein of human immunodeficiency virus type 1. *J Gen Virol*, **91** [Pt 1], 1–12.

[Russell and Ley 2002] Russell, J. H. and Ley, T. J. [2002] Lymphocyte-mediated cytotoxicity. *Annu Rev Immunol*, **20**, 323–370.

[Samri *et al.* 2000] Samri, A., Haas, G., Duntze, J., Bouley, J. M., Calvez, V., Katlama, C., and Autran, B. [2000] Immunogenicity of mutations induced by nucleoside reverse transcriptase inhibitors for human immunodeficiency virus type 1-specific cytotoxic T cells. *J Virol*, **74** [19], 9306–9312.

[Schmitt *et al.* 2000] Schmitt, M., Harrer, E., Goldwich, A., Bäuerle, M., Graedner, I., Kalden, J. R., and Harrer, T. [2000] Specific recognition of lamivudine-resistant HIV-1 by cytotoxic T lymphocytes. *AIDS*, **14** [6], 653–658.

[Schwartz 1989] Schwartz, R. H. [1989] Acquisition of immunologic self-tolerance. *Cell*, **57** [7], 1073–1081.

[Shafer 2006a] Shafer, R. W. [2006a] Rationale and uses of a public HIV drug-resistance database. *J Infect Dis*, **194 Suppl 1**, S51–S58.

[Shafer 2006b] Shafer, R. W. [2006b] Rationale and uses of a public HIV drug-resistance database. *J Infect Dis*, **194 Suppl 1**, S51–S58.

[Sharp and Hahn 2011] Sharp, P. M. and Hahn, B. H. [2011] Origins of HIV and the AIDS pandemic. *Cold Spring Harb Perspect Med*, **1** [1], a006841.

[Sharp *et al.* 1995] Sharp, P. M., Robertson, D. L., and Hahn, B. H. [1995] Cross-species transmission and recombination of 'AIDS' viruses. *Philos Trans R Soc Lond B Biol Sci*, **349** [1327], 41–47.

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

[Siekevitz *et al.* 1987] Siekevitz, M., Kocks, C., Rajewsky, K., and Dildrop, R. [1987] Analysis of somatic mutation and class switching in naive and memory B cells generating adoptive primary and secondary responses. *Cell*, **48** [5], 757 – 770.

[Singh and Barry 2004] Singh, R. A. K. and Barry, M. A. [2004] Repertoire and immunofocusing of CD8 T cell responses generated by HIV-1 gag-pol and expression library immunization vaccines. *J Immunol*, **173** [7], 4387–4393.

[Smidt 2013] Smidt, W. [2013] Potential elucidation of a novel CTL epitope in HIV-1 protease by the protease inhibitor resistance mutation L90M. *PLoS One*, **8** [8], e71888.

[Storey 2002] Storey, J. D. [2002] A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B Statistical Methodology*, **64** [3], 479–498.

[Takeuchi *et al.* 1988] Takeuchi, Y., Nagumo, T., and Hoshino, H. [1988] Low fidelity of cell-free DNA synthesis by reverse transcriptase of human immunodeficiency virus. *J Virol*, **62** [10], 3900–3902.

[Tallquist *et al.* 1996] Tallquist, M. D., Yun, T. J., and Pease, L. R. [1996] A single T cell receptor recognizes structurally distinct MHC/peptide complexes with high specificity. *J Exp Med*, **184** [3], 1017–1026.

[Tenzer *et al.* 2005] Tenzer, S., Peters, B., Bulik, S., Schoor, O., Lemmel, C., Schatz, M. M., Kloetzel, P.-M., Rammensee, H.-G., Schild, H., and Holzhütter, H.-G. [2005] Modeling the MHC class I pathway by combining predictions of proteasomal cleavage, TAP transport and MHC class I binding. *Cell Mol Life Sci*, **62** [9], 1025–1037.

[Tenzer *et al.* 2009] Tenzer, S., Wee, E., Burgevin, A., Stewart-Jones, G., Friis, L., Lamberth, K., Chang, C.-h., Harndahl, M., Weimershaus, M., Gerstoft, J., Akkad, N., Klenerman, P., Fugger, L., Jones, E. Y., McMichael, A. J., Buus, S., Schild, H., van Endert, P., and Iversen, A. K. N. [2009] Antigen processing influences HIV-specific cytotoxic T lymphocyte immunodominance. *Nat Immunol*, **10** [6], 636–646.

[Tonegawa 1983] Tonegawa, S. [1983] Somatic generation of antibody diversity. *Nature*, **302** [5909], 575–581.

[Troyer *et al.* 2009] Troyer, R. M., McNevin, J., Liu, Y., Zhang, S. C., Krizan, R. W., Abraha, A., Tebit, D. M., Zhao, H., Avila, S., Lobritz, M. A., McElrath, M. J., Le Gall, S., Mullins, J. I., and Arts, E. J. [2009] Variable fitness impact of HIV-1 escape mutations to cytotoxic T lymphocyte (CTL) response. *PLoS Pathog*, **5** [4], e1000365.

[Uhlmann *et al.* 2004] Uhlmann, E. J., Tebas, P., Storch, G. A., Powderly, W. G., Lie, Y. S., Whitcomb, J. M., Hellmann, N. S., and Arens, M. Q. [2004] Effects of the G190A substitution of HIV reverse transcriptase on phenotypic susceptibility of patient isolates to delavirdine. *J Clin Virol*, **31** [3], 198–203.

[Unno *et al.* 2002] Unno, M., Mizushima, T., Morimoto, Y., Tomisugi, Y., Tanaka, K., Yasuoka, N., and Tsukihara, T. [2002] The structure of the mammalian 20S proteasome at 2.75 A resolution. *Structure*, **10** [5], 609–618.

[Vergis and Mellors 2000] Vergis, E. N. and Mellors, J. W. [2000] Natural history of HIV-1 infection. *Infect Dis Clin North Am*, **14** [4], 809–25, v–vi.

[Vidal *et al.* 2011] Vidal, J. E., Freitas, A. C., Song, A. T., Campos, S. V., Dalben, M., and Hernandez, A. V. [2011] Prevalence and factors associated with darunavir resistance mutations in multi-experienced HIV-1-infected patients failing other protease inhibitors in a referral teaching center in Brazil. *Braz J Infect Dis*, **15** [3], 245–248.

[Vita *et al.* 2010] Vita, R., Zarebski, L., Greenbaum, J. A., Emami, H., Hoof, I., Salimi, N., Damle, R., Sette, A., and Peters, B. [2010] The immune epitope database 2.0. *Nucleic Acids Res*, **38** [Database issue], D854–D862.

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

[Wain-Hobson *et al.* 1991] Wain-Hobson, S., Vartanian, J. P., Henry, M., Chenciner, N., Cheynier, R., Delassus, S., Martins, L. P., Sala, M., Nugeyre, M. T., and Guétard, D. [1991] LAV revisited: origins of the early HIV-1 isolates from Institut Pasteur. *Science*, **252** [5008], 961–965.

[Wang *et al.* 2007] Wang, S., Sun, Y., Zhai, S., Zhuang, Y., Zhao, S., Kang, W., Li, X., Huang, D., Yu, X. G., Walker, B. D., and Altfeld, M. A. [2007] Identification of HLA-A11-restricted HIV-1-specific cytotoxic T-lymphocyte epitopes in China. *Curr HIV Res*, **5** [1], 119–128.

[Wang *et al.* 2009] Wang, Y. E., Li, B., Carlson, J. M., Streeck, H., Gladden, A. D., Goodman, R., Schneidewind, A., Power, K. A., Toth, I., Frahm, N., Alter, G., Brander, C., Carrington, M., Walker, B. D., Altfeld, M., Heckerman, D., and Allen, T. M. [2009] Protective HLA class I alleles that restrict acute-phase CD8+ T-cell responses are associated with viral escape mutations located in highly conserved regions of human immunodeficiency virus type 1. *J Virol*, **83** [4], 1845–1855.

[Weaver *et al.* 1988] Weaver, C. T., Hawrylowicz, C. M., and Unanue, E. R. [1988] T helper cell subsets require the expression of distinct costimulatory signals by antigen-presenting cells. *Proc Natl Acad Sci U S A*, **85** [21], 8181–8185.

[Wei *et al.* 2003] Wei, X., Decker, J. M., Wang, S., Hui, H., Kappes, J. C., Wu, X., Salazar-Gonzalez, J. F., Salazar, M. G., Kilby, J. M., Saag, M. S., Komarova, N. L., Nowak, M. A., Hahn, B. H., Kwong, P. D., and Shaw, G. M. [2003] Antibody neutralization and escape by HIV-1. *Nature*, **422** [6929], 307–312.

[Wills *et al.* 1994] Wills, N. M., Gesteland, R. F., and Atkins, J. F. [1994] Pseudoknot-dependent read-through of retroviral gag termination codons: importance of sequences in the spacer and loop 2. *EMBO J*, **13** [17], 4137–4144.

[Wood *et al.* 2009] Wood, N., Bhattacharya, T., Keele, B. F., Giorgi, E., Liu, M., Gaschen, B., Daniels, M., Ferrari, G., Haynes, B. F., McMichael, A., Shaw, G. M., Hahn, B. H., Korber, B., and Seoighe, C. [2009] HIV evolution in early infection: selection pressures, patterns of insertion and deletion, and the impact of APOBEC. *PLoS Pathog*, **5** [5], e1000414.

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

[Yokomaku *et al.* 2004] Yokomaku, Y., Miura, H., Tomiyama, H., Kawana-Tachikawa, A., Takiguchi, M., Kojima, A., Nagai, Y., Iwamoto, A., Matsuda, Z., and Ariyoshi, K. [2004] Impaired processing and presentation of cytotoxic-T-lymphocyte (CTL) epitopes are major escape mechanisms from CTL immune pressure in human immunodeficiency virus type 1 infection. *J Virol*, **78** [3], 1324–1332.

[York and Rock 1996] York, I. A. and Rock, K. L. [1996] Antigen processing and presentation by the class I major histocompatibility complex. *Annu Rev Immunol*, **14**, 369–396.