Impact of p2/NC Cleavage Site Polymorphisms on HIV-1 Subtype C Viral Fitness

by

Serron Wilson

Submitted in fulfilment of the requirements for the degree of Master of Medical Science in Molecular Virology

School of Laboratory Medicine and Medical Sciences

University of KwaZulu Natal

2012

Preface

The experimental work described in this dissertation was carried out in the Hasso Plattner Research Laboratory of the HIV Pathogenesis Programme at the Doris Duke Medical Research Institute, Nelson R. Mandela School of Medicine, University of KwaZulu-Natal, Durban, from February 2010 to March 2012 under the supervision of Dr Michelle Lucille Gordon.

These studies represent original work by the author and have not otherwise been submitted in any form for any degree or diploma to any other University. Where use has been made of the work of others, it is duly acknowledged in the text.

Signed: _____ Date: _____

Serron Wilson (Candidate)

Signed: _____ Date: _____

Dr Michelle Lucille Gordon (Supervisor)

Plagiarism declaration:

I, Serron Wilson, declare that:

i. The research reported in this dissertation, except where otherwise indicated, is my original work.

ii. This dissertation has not been submitted for any degree or examination at any other university.

iii. This dissertation does not contain other persons' data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons.

iv. This dissertation does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:

   a. their words have been re-written but the general information attributed to them has been referenced;

   b. Where their exact words have been used, their writing has been placed inside quotation marks, and referenced.

v. Where I have reproduced a publication of which I am an author, co-author or editor, I have indicated in detail which part of the publication was actually written by myself alone and have fully referenced such publications.

vi. This dissertation does not contain text, graphics or tables copied and pasted from the Internet, unless specifically acknowledged, and the source being detailed in the dissertation and in the Reference sections.

Signed: _____ Date: _____

Serron Wilson (Candidate)

Signed: _____ Date: _____

Dr Michelle Lucille Gordon (Supervisor)

Ethical Approval

Full ethical approval, was obtained for this study from the Biomedical Research Ethics Committee of the Nelson R. Mandela School of Medicine, University of KwaZulu-Natal (ref: BF 068/08).

Presentations

Part of this work was presented as a poster at the 5th SA AIDS Conference, held in Durban, South Africa from the 7th-10th of June 2011. The poster was titled *'3D modelling of HIV-1 subtype C Gag p2/NC cleavage site polymorphisms'.*

This work was also presented at Keystone Symposia's 2012 Meeting on HIV Vaccines, held in Keystone, Colorado, USA from the 21st-26th March, 2012. The poster was titled *'Variation at p2/NC cleavage site has no effect on subtype C viral fitness'.*

Acknowledgements

I would like to thank and acknowledge Dr Michelle Gordon for her contribution to this study, for her assistance in the conception and design of this project and the writing of this dissertation, none of which would have been possible without her.

I would like to thank Professor Thumbi Ndung'u for his wise advice, guidance and knowledge, which have been gratefully received.

I would like to thank Dr Julie Prado for her contribution to this work, specifically with regard to the western blotting assays.

I am extremely grateful to Dr Jaclyn Wright for her assistance in the technical matters of this project, specifically the replication capacity assays, and also for her contributions of plasmids, primers and numerous other materials used this study.

I would like to thank the HIV Pathogenesis Programme and the National Research Foundation for financial assistance.

My sincerest heartfelt gratitude goes to all my family.

Abstract

Subtype C accounts for the majority of HIV infections and in South Africa, is the dominant subtype. The Gag cleavage sites of subtype C viruses show a high degree of natural variation compared to subtype B and group M sequences, with the p2/NC site having the highest degree of variation among all cleavage sites and between all subtypes. This study therefore aimed to determine the functional effect of this variation on viral fitness. A library of drug naïve subtype C sequences were screened using computational analysis to predict binding affinity between HIV protease and the Gag substrate at the p2/NC site. Ligands with high predicted affinity had hydrophobic cleavage sites with substantial diversity at positions P5-P3. Lower ranking ligands were mostly similar to the consensus subtype C. Three ligands were selected for fitness assays from each the high ranking and low ranking groups. Chimeric viruses expressing selected cleavage sites were generated by site directed mutagenesis. Replication capacity assays of these viruses showed moderate differences in fitness but failed to demonstrate a correlation with computational estimates of binding affinity. Enzymes assays were performed to further investigate substrate preferences and the binding mechanism of protease. To this end, recombinantly expressed HIV-1 protease was tested against a range of substrates the matching the p2/NC cleavage sites used in the replication capacity assay. Results of the enzyme assay did not correlate with either the computation studies or the replication capacity assay results, suggesting a sequence independent binding and recognition mechanism of HIV-1 protease. Taken together the results suggest that processing of Gag is determined by tertiary folding of the polyprotein and not amino acid sequence at the cleavage site.

Table of Contents

List of Figures

List of Tables

Abbreviations and Acronyms

CS	-	Cleavage site
DMSO	-	Dimethyl sulfoxide
DNA	-	Deoxyribonucleic acid
DTT	-	Dithiothreitol
E. coil	-	*Escherichia coli*
EDTA	-	Ethylenediaminetetraacetic acid
GFP	-	Green Fluorescent Protein
gp120	-	Glycoprotein 120
gp41	-	Glycoprotein 41
HIV	-	Human Immunodeficiency Virus
IPTG	-	Isopropyl β-D-1-thiogalactopyranoside
NDOH	-	National Department of Health
NFV	–	Nelfinavir
PBMC	-	Peripheral blood mononuclear cell
PCR	-	Polymerase chain reaction
PI	-	Protease Inhibitor
PR	-	Protease
RNA	-	Ribonucleic Acid
RNAse H	-	Ribonuclease H
RPM	-	Revolutions per minute
RT	-	Reverse Transcriptase
SDM	-	Site directed mutagenesis
SDS PAGE	-	Sodium dodecyl sulphate polyacrylamide gel electrophoresis

TBE	-	Tris-Borate-EDTA
TBS	-	Tris Buffered Saline
UV	-	Ultraviolet
X-Gal	-	5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside

CHAPTER 1

LITERATURE REVIEW

1.1 INTRODUCTION

Acquired Immune Deficiency Syndrome (AIDS) is the defining public health crisis of our time. Since its discovery in the early 1980's AIDS has contributed to the deaths of approximately 30 million people, with about 33.3 million people worldwide living with HIV (UNAIDS, 2011). The causative agent of AIDS, the Human Immunodeficiency Virus (HIV), is a retrovirus which is passed between individuals by sexual transmission and contaminated bodily fluids such as blood products. Currently, most infections are due to heterosexual transmission; however this varies with geographical location. The sharing of used needles between intravenous drug users (IDUs) is also a major source of transmission, especially in Eastern Europe and central Asia (Simon et al., 2006). While the disease was first identified in North America, today the vast majority of infections are found in sub-Saharan Africa, with approximately 22.5 million infected individuals (UNAIDS, 2011). This accounts for 68% of all infections worldwide. This places a heavy burden on some of the world's weakest economies and health systems.

The development of antiretroviral (ARV) therapy to treat HIV/AIDS has been the only effective approach to HIV. ARV has been able to slow disease progression significantly and decrease associated morbidities and mortality (Simon et al., 2006). However, while all efforts are being made yet no cure or vaccine is available at present. This greatly hampers efforts to prevent the further spread of the disease

1.1.1 History of HIV and AIDS

A report published by the Atlanta based Centre for Disease Control (CDC) in 1981 described the incidence of rare opportunistic infections amongst a small number of homosexual men. The main symptom of the disease was a severely compromised immune system, leading to death from otherwise uncommon infections. Initially it was thought this disease was isolated to homosexuals. This was proven false when similar symptoms began to appear in non-homosexuals. High risk groups included intravenous drug users, haemophiliacs and Haitian immigrants (Goedert and Gallo, 1985).

In 1983 Dr Luc Montagnier and Dr Francois Barre-Sinoussi from the Pasteur Institute in France isolated a retrovirus which they suggested was the cause of AIDS. The virus was named lymphadenopathy-associated virus (LAV) (Barre-Sinoussi et al., 1983). Soon after, Dr Robert Gallo of the National Cancer Institute isolated a virus named HTLV-III (Gallo et al., 1984). It soon became apparent that LAV and HTLV-III were the same virus. Therefore in 1986 the virus was renamed Human Immunodeficiency Virus (HIV) by the International Committee on Taxonomy of Viruses (Case, 1986). Evidence emerged over the course of the next few years to suggest transmission of the virus was due to sexual contact and contaminated blood products (Goedert and Gallo, 1985).

1.1.2 Current figures and statistics

In 2010 there was an estimated 34 million people in the world living with HIV, approximately 2.7 million new infections occurred and 1.8 million AIDS related deaths were reported (UNAIDS, 2011). Roughly a quarter of all new HIV-1 infections are found in individuals under 25 years old, with females having 3 to 6 times higher infection rates than males in the same age group (Simon et al., 2006). Approximately 85% of HIV-1 infections

are as a result of heterosexual transmission, however, outside sub-saharan Africa, an estimated third of all infections are a result of injecting drug use, most of which are in eastern Europe and central and southeast Asia (Simon et al., 2006, WHO, 2009).

There is a disproportionately high HIV burden in developing nations, such as South American and African countries, India and other eastern nations. First world countries such as the USA, the United Kingdom and other European nations have much lower HIV incidence. Global distribution of HIV infection at the end of 2009 can be seen in Figure 1.1 (UNAIDS, 2011).

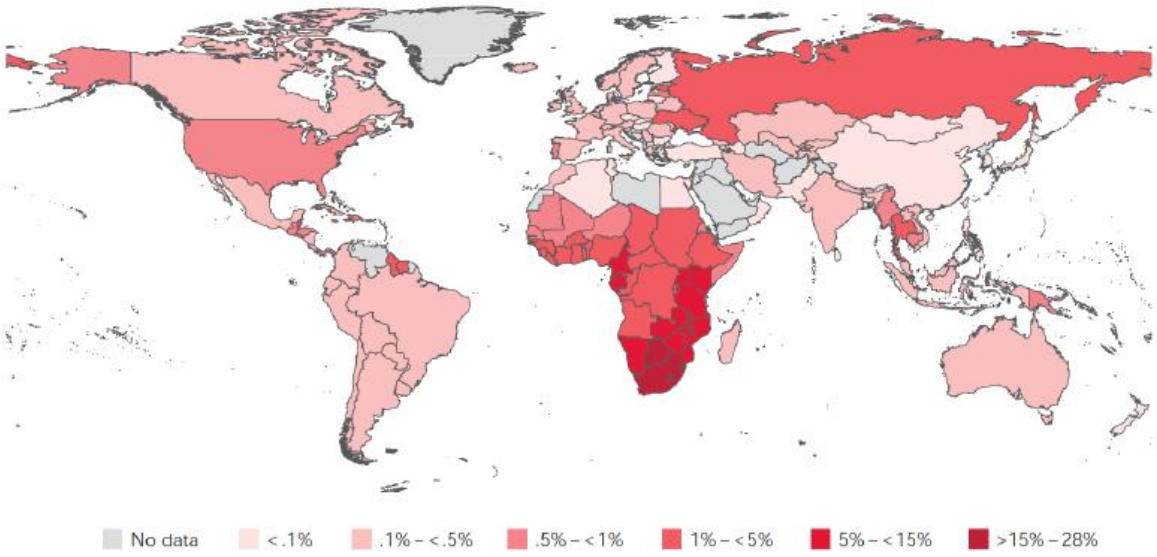


Figure 1.1 Global HIV prevalence and distribution. Sub-Saharan Africa reports the highest prevalence of infected individuals. The adult population of South Africa (15-49 years of age) has between 15 -28% prevalence rate. Taken from WHO/UNAIDS 2010 Global report (2011).

The HIV burden is greatest in the southern regions of Africa, with South Africa having the highest prevalence in the world. Current estimates are that 5.5 million people in South

Africa are living with HIV (NDOH, 2012), which equates to roughly 11% of the population. Since 2004 the government of South Africa has employed a policy of free HIV treatment for those who meet the criteria (NDOH, 2004), which has substantially decreased mortality and morbidity associated with the disease. Consequently, the health care system and the economy of South African have been placed under an intense burden.

ARV therapy for such large numbers is enormously expensive thus is not the ideal solution. In 2011 the Global Fund spent US$ 1.9 billion supporting 3.5 million patients on ARV therapy in low and middle income countries (Stover et al., 2011). Due to an ever increasing number of HIV infections, this cost will grow substantially; and as such the prevention of new infections is of the utmost importance. However, despite well established and effective methods of HIV prevention, new infections have not substantially reduced (UNAIDS, 2011). For these reasons, the ideal strategy to tackle this disease is the development of a safe and effective vaccine. However, attempts to design such a vaccine have been not yet been successful

1.2 HUMAN IMMUNODEFICIENCY VIRUS

1.2.1 Origin of the virus

HIV is from the *Lentivirus* genus of the *Retroviridae* family. It is believed to have originated from simian immunodeficiency virus (SIV) during a cross-species transmission event in West Africa early in the 20th century (Hemelaar et al., 2011). SIV is a family of over 40 lentiviruses with high sequence homology to HIV and can infect a range of non-human primates, including the African green monkeys, mandrills, sooty mangabeys, red-capped mangabeys, and numerous others (Silvestri et al., 2007). Non-natural hosts of SIV (i.e. those infected via a cross-species transmission events) include the rhesus macaque, pig-

tailed macaques and chimpanzees (Pandrea and Apetrei, 2010). Natural hosts experience a non-pathogenic, non-progressive infection. In contrast, infection of non-natural hosts results in disease outcomes similar to HIV infection in humans (Clements and Zink, 1996).

Transmission from primates to humans is thought to have occurred through the hunting and butchering of infected primates (Hemelaar, 2011). This is common practice in certain parts of Africa has to date permitted transmission of at least 2 other simian viruses to human hosts, namely the foamy virus and Primate /Human T-lymphotropic virus, making transmission of SIV to humans via this mechanism quite probable (McCutchan, 2006). Multiple transmission events have occurred, each with different outcomes and each giving rise to different strains of HIV (VandeWoude and Apetrei, 2006). The virus which infects the common chimpanzee, SIVcpz, is considered the origin of HIV-1 and while HIV-2 is thought to have originated from SIVsmm, which infects the sooty mangabeys (McCutchan, 2006, Silvestri et al., 2007). The phylogenetic relationship between SIV and HIV can be seen in Figure 1.2.

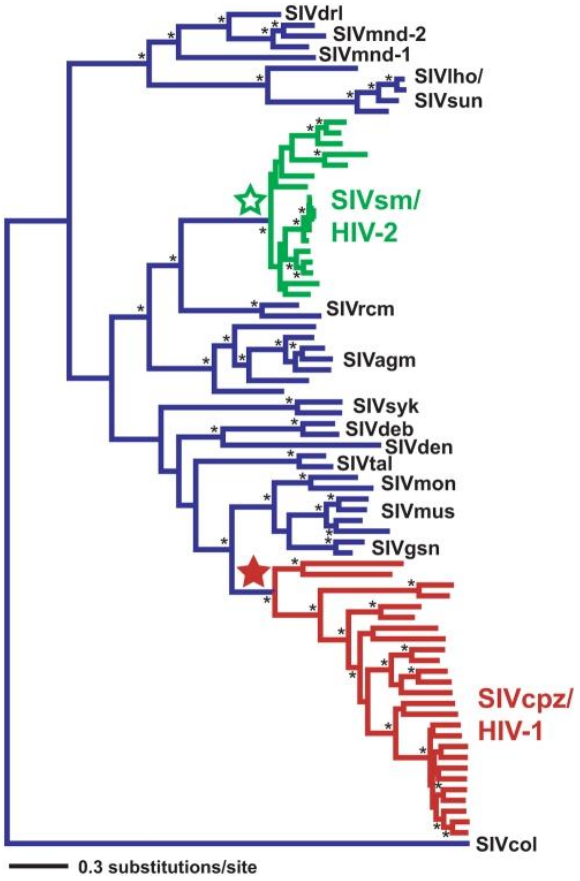


Figure 1.2 Phylogeny of the SIV and HIV *env* locus. Green indicates lineages SIVsm/HIV-2 lineages and red indicates SIVcpz/HIV-1 lineages. Stars represent most recent common ancestor. All other SIV lineages are blue. Taken from Wertheim and Worobey (2009).

1.2.2 Clinical course of infection

Primary HIV infection, also known as the acute phase of infection, lasts roughly 2- 4 weeks (Pope and Haase, 2003). During this phase rapid uncontrolled viral replication results in very high plasma viral load, in the range of 10^7 – 10^8 RNA copies per ml of plasma, and is accompanied by an initial loss of $CD4^+$ T cells in the peripheral blood and a mass depletion of $CD4^+$ T cells in the gut-associated lymphoid tissue (GALT) (Simon et al., 2006).

Usually, after about 4 weeks of infection, cellular immune responses to HIV are generated and are partially able to control the viral replication (Pope and Haase, 2003). This results in a rise of the $CD4^+$ T cell count accompanied by a reduction of plasma viral load. The viral 'set point' is classified as the level at which the viral load stabilises. Set points vary greatly among different individuals and are predictive of disease progression (Dykes and Demeter, 2007).This period is known as the chronic phase and may last several years, especially if the infected individual is on ARV therapy.

Therapy is able to reduce the viral load to undetectable levels (below 50 copies per mL plasma) and rescue the $CD4^+$ T cell plasma count. However despite years of therapy, $CD4^+$ cell levels in the GALT never recover (Simon et al., 2006). A gradual increase of plasma viral load together with decline in $CD4^+$ T cell count is inevitable resulting in the final stage of the disease, AIDS (Figure 1.3).

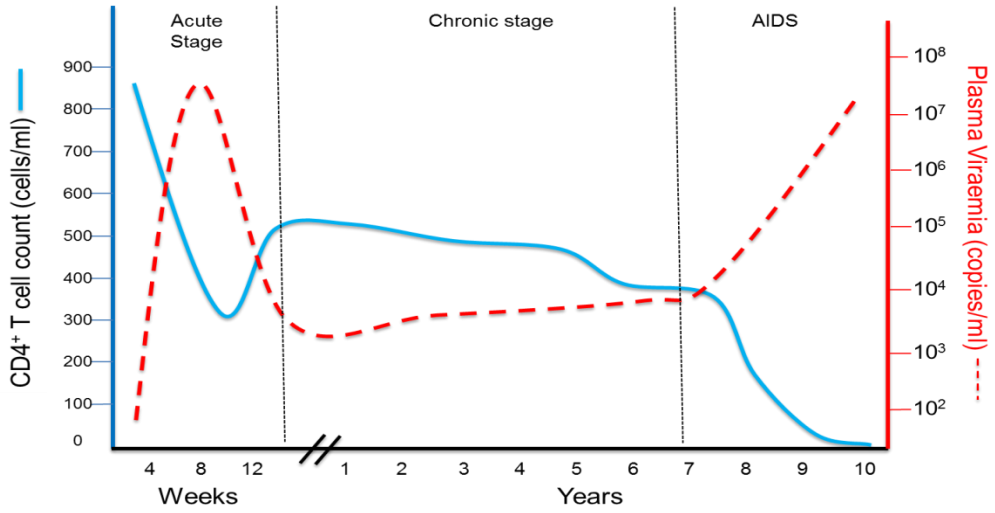


Figure 1.3 Clinical course of HIV Infection. Blue line = $CD4^+$ T cell count, dotted red line = viral load. During acute infection plasma viral load rises sharply while $CD4^+$ T cell count drops. During chronic infection, plasma viral load drops and peripheral blood $CD4^+$ T cells rise. The final stage of the disease, AIDS, control of viral load is lost and $CD4^+$ T cell counts fall sharply. Adapted from Simon et al (2006).

1.2.3 Genome arrangement

HIV is expressed as 9 genes, flanked by 5' and 3' long terminal repeats (LTR) (Figure 1.4). All known lentiviruses are exogenous and contain *gag, pol* and *env* as the structural and enzymatic genes. Between lentiviruses the number of accessory genes can vary, but *vif* (virus infectivity factor) and *rev* (regulator of virus gene expression) are constant inclusions in the lentiviral genome (VandeWoude and Apetrei, 2006). Addition accessory genes in HIV are *vpr, vpu, tat, and nef.* The main structural proteins capsid, matrix and nucleocapsid are encoded in the *gag* gene. The enzymatic components of the virus such as reverse transcriptase, integrase and protease are expressed in the *pol* gene, while the *env* gene codes for the surface envelope glycoprotein. Regulatory genes *rev* and *tat* control transcription and translation of viral genes and RNA transport. Production of infectious virus particles is regulated by *vif* and *vpu*, while *vpr* and *nef* are involved in disease manifestations (Clements and Zink, 1996).

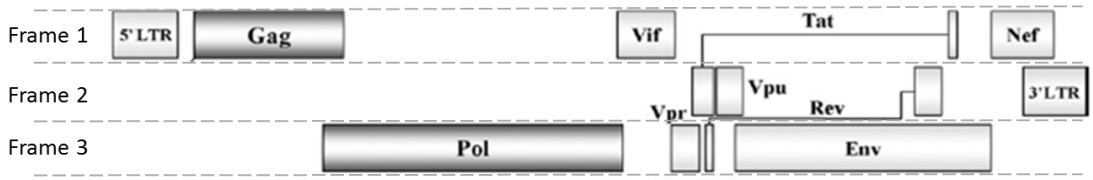


Figure 1.4 Genetic organisation of HIV-1. Long terminal repeats (LTR) flank the 9 HIV genes at both the 5' and 3' ends. *Gag, pol* and e*nv* are the main structural genes. Accessory genes *rev and tat* regulate transcription and translation. *Vpu* and v*if* ensure virion production while *nef* and *vpr* are involved in disease manifestation. Taken from Wensing et al (2010).

1.2.4 Physical structure

Retroviruses are unique among virus families as they possess a diploid RNA genome (Kieken et al., 2002). This RNA is found inside the conical core of the virion (Figure 1.5a).

The core is composed of the Gag proteins p24 (capsid), p17 (matrix), and p7 (nucleocapsid). The 55kDa (or p55) Gag polyprotein is processed via cleavage to form these products. The viral core is surrounded by a viral envelope which is derived from the host cell membrane. On the surface of the virion is the highly glycosylated envelope spike. Three heterodimers composed of the trans-membrane gp41 glycoprotein and the surface glycoprotein gp120 make up the envelope molecule (Figure 1.5b) (White et al., 2011). The envelope glycoprotein interacts with receptors on the surface of the host cell, allowing viral entry. The CD4 cell surface marker is the main target of the envelope glycoprotein, with the chemokine receptors CCR5 or CXCR4 acting as co-receptors. Cells not expressing the CD4 receptor cannot be infected by the virus. The mature virus particle is approximately 100 nm in diameter (Clements and Zink, 1996).

(a) (b) (c)

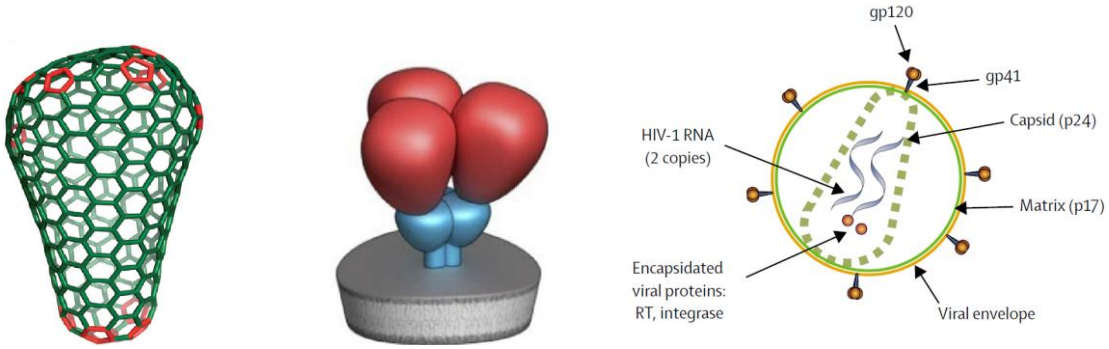


Figure 1.5 Physical structure of HIV Virion (a). Conical nature of virion core, taken from Ganser-Pornillos et al. (2008). (b) Schematic view of trimeric HIV envelope glycoprotein, taken from White et al. (2011). (c) diagrammatic view of single HIV virion, taken from Simon et al (2006) .

1.2.5 Life cycle

The HIV life cycle (Figure 1.6) begins with the binding of the envelope glycoprotein to the primary receptor, the CD4 molecule on the surface of the host cell. This molecule is found

on cells of the immune system, such as helper T cells, monocytes/macrophages and dendritic cells, to name a few. Binding of the envelope spike to the CD4 receptor triggers a conformational change in gp120, exposing the co-receptor binding site in the glycoprotein. The cell surface co-receptors used by the virus may be either CCR5 or CXCR4, both members of the chemokine receptor family. The CCR5 receptor is found on macrophages and dendritic cells, and CXCR4 is found on T cells. Depending on the co-receptor usage, the virus stain is known as either M-tropic or T-tropic.

Binding of the envelope glycoprotein to the co-receptor is followed by fusion of the viral and host cell membrane. The viral core enters the cytoplasm and uncoats, releasing the diploid viral RNA genome. The virion associated reverse transcriptase begins transcribing the viral RNA into double stranded complementary DNA (cDNA). The cDNA is then transported to the nucleus where it is integrated into the host genome by HIV integrase. Genomic (unspliced) and messenger (spliced) RNA is generated by transcription of this cDNA. The mRNA is transported to the cytoplasm where translation of the viral proteins occurs. The *gag* and *pol* genes are translated as immature poly-protein precursors, which require further processing. The *gag* gene is expressed as Gag (p55). A (-1) ribosomal frame shift mechanism occurs near the C terminus of the *gag* gene causing *pol* to be expressed as a GagPol 'fusion' protein (Pettit et al., 1994, Clements and Zink, 1996).

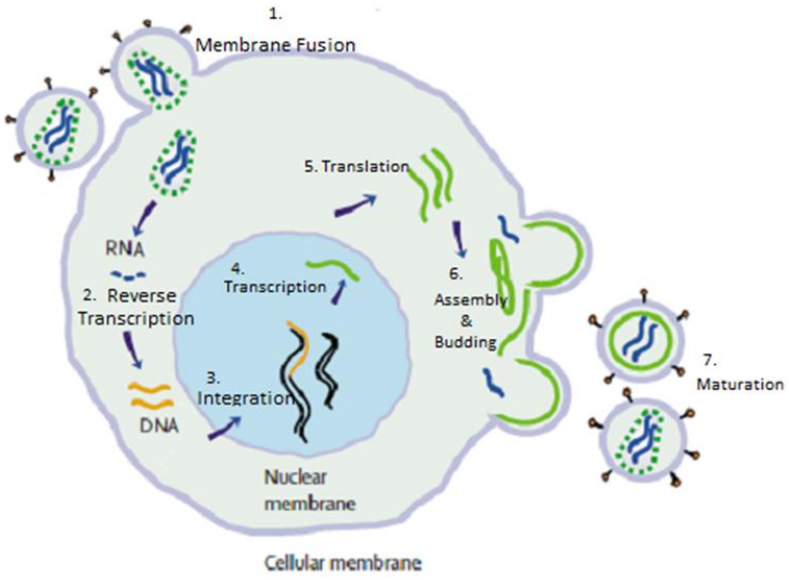


Figure 1.6 Life cycle of Human immunodeficiency virus. (1) Envelope surface glycoprotein binds to the CD4 host cell receptor. Virus and host membrane fuse, the viral genome and enzymes enter the cell. (2) Viral RNA is transcribed into DNA by HIV Reverse Transcriptase. (3) This is integrated into the host genome by HIV Integrase. (4 & 5) The viral genes are transcribed and translated by the host cell machinery. (6) Viral proteins accumulate at the cell surface, where assembly begins and immature virions bud off. (7) Maturation occurs when HIV Protease processes the Gag polyprotein precursor into its functional forms. Adapted from Simon et al (2006).

The Gag and Gag-Pol gene products collect at the cell surface, where viral assembling begins and immature virus particles bud off from the infected cell. The immature virion is essentially a layer of unprocessed Gag molecules coating the inside of the virion membrane. These particles are not infectious until the characteristic viral core has formed (Figure 1.7). This is mediated by the HIV aspartyl protease, which cleaves the Gag and Gag-Pol precursors into their functional forms (Monini et al., 2004).

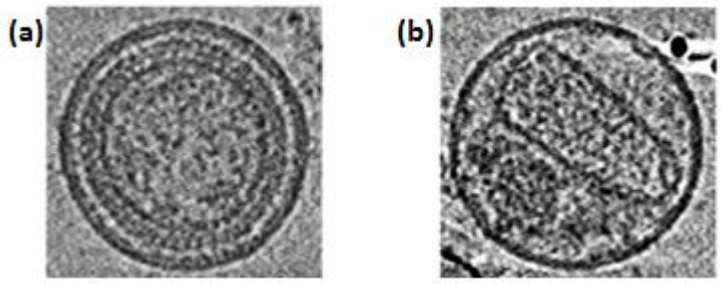


Figure 1.7 Physical maturation of HIV virion. Post budding processing of the Gag polyprotein precursor by HIV Protease induces an irreversible conformation change resulting in a morphologically distinct fully mature infectious virion. **(a)** The immature virion before cleavage by PR**. (b)** The mature virion post cleavage .Taken from (Ganser-Pornillos et al., 2008).

1.2.6 Proteolytic cleavage of HIV-1 Gag

Gag and Gag-Pol are cleaved in a strictly sequential manner at well-defined sites by the viral protease. Processing of the Gag-Pol precursor yields the enzymes of the virus, namely HIV protease (PR), reverse transcriptase (RT) and integrase, as well as the Gag structural proteins (Pettit et al., 1994). The Gag precursor (p55) is cleaved by PR into 6 peptides, namely matrix (MA or p17), capsid (CA or p24), nucleocapsid (NC or p7), and three smaller peptides, p6, p1 and p2 (Figure 1.8) (Holguin et al., 2005, Ganser-Pornillos et al., 2008).

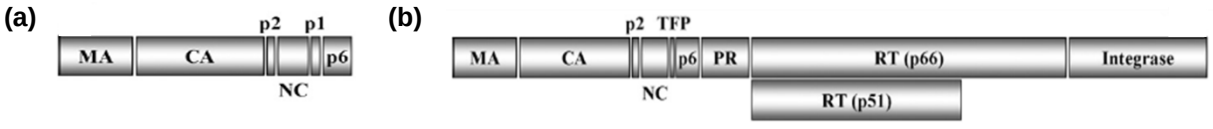


Figure 1.8 Gag and Gag-Pol polyprotein precursors. (a) The Gag polypeptide precursor, comprising of matrix (MA), capsid (CA), p2, nucleocapsid (NC), p1 and p6. **(b)** The Gag-Pol polyprotein, comprising of Gag proteins plus transframe protein (TFP), protease (PR), reverse transcriptase (RTp51), RNase H (RTp66) and integrase. Translation of Gag-Pol occurs as a function of a (-1) frame shift mechanism at the C term of Gag. Taken from (Wensing et al., 2010).

MA forms the inner layer of the virus, between the plasma layer derived from the host cell and the capsid of viral material. CA protein forms the conical capsid which contains the genetic material and enzymes. This structure is the portion of the virus which is injected into the host cell upon infection (Ganser-Pornillos et al., 2008). NC associates with the viral RNA inside the capsid. This protein is responsible for the recognition and binding of RNA. Its chemical composition is generally basic and highly conserved. NC also has a characteristic Zinc Finger motif, and acts largely as a facilitator of RT and integrase function (Ho et al., 2008, Thomas and Gorelick, 2008). The small peptides p2, p1 and p6, all co-ordinate membrane binding and Gag-Gag lattice formation.

There are three stages of cleavage during Gag processing. First, the site between the p2 and NC peptides (p2/NC) is cleaved. This is followed by essentially simultaneous cleavage at the MA/CA and p1/p6 site. The final step in processing of Gag is the removal of the small spacer peptides p2 and p1 from CA and NC, respectively (De Oliveira et al., 2003) (Figure 1.9). This final step has the slowest rate and hence is known as the rate determining step. The highest cleavage rate is at the p2/NC cleavage site (Wensing et al., 2010).

In addition to the 5 site in Gag, PR also has to cleave at 6 sites in Gag-Pol and 1 in Nef, giving a total of 12 sites. These sites show very little sequence homology, hence there is a great deal of uncertainty regarding the mechanism by which PR recognises the cleavage sites. Pettit and colleagues (1994) suggested three possible factors which may determine the order of cleavage of the precursor: (1) amino acid sequence (2) three dimensional shape of the peptide at the cleavage site or (3) the accessibility of the cleavage site to protease.

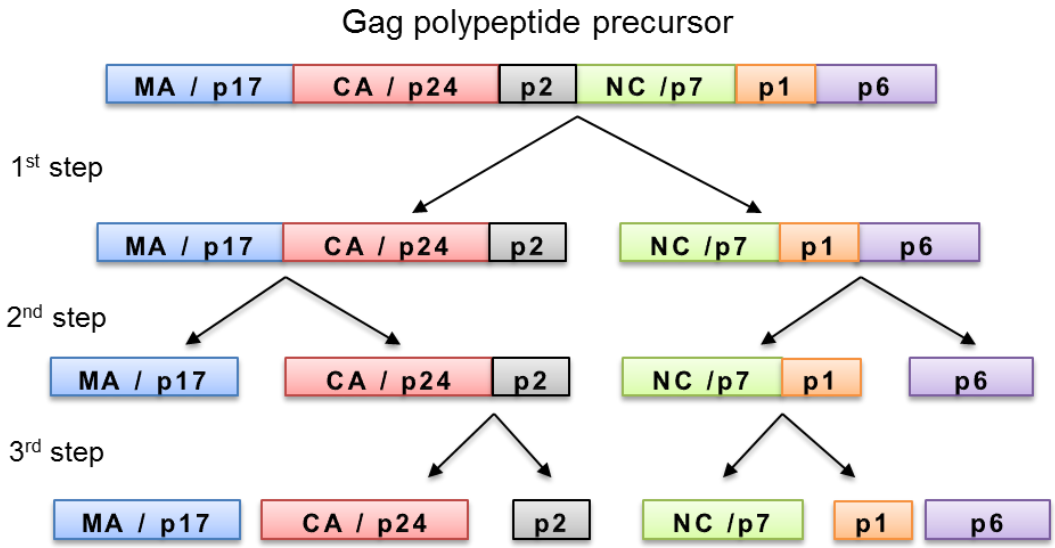


Figure 1.9 Order of Gag cleavage. The first site of cleavage in Gag is between p2 and NC. This is followed by cleavage at the MA/CA site and the p1/p6 site. The final cleavage step is the removal of the small spacer peptides p2 from CA and p1 from NC.

While amino acid sequence is considered to play a role in determining proteolytic cleavage, due to the limited cleavage site sequence identity between PR substrates this seems unlikely to be the case for Gag cleavage. A few studies suggest that the 3 dimensional shape of the site determines cleavage (Nalam et al., 2010, Özen et al., 2012). Other studies have found the accessibility of the cleavage site to protease plays the more significant role in determining the location and succession of cleavage (Perez et al., 2010).

HIV Protease

PR is a member of the aspartic protease family. This enzyme is a symmetrical homodimer consisting of 2 identical subunits of 99 amino acids. Each subunit contributes one aspartic acid residue to the active site, which is found at the dimer interface, at the centre of the

enzyme. The catalytic triad of the enzyme consists of the residues Asp-Thr-Gly, conserved across all aspartic proteases (Brik and Wong, 2002), indicating a conserved mechanism for this group of proteases. Pepstatin is a natural inhibitor which selectively inhibits aspartic proteases; and is able to inhibit PR. Each subunit contains an extended beta sheet glycine rich loop, known as the flap (Figure 1.10), which constitutes part of the substrate binding cavity. During substrate binding the flaps open slightly to allow the substrate into the active site region and subsequently close, tightening around the substrate (Pietrucci et al., 2009). The substrate is held in the cavity through hydrogen bonding and Van der Wals interactions (Perez et al., 2010).

The physical parameters and governing principles which control cleavage site specificity are poorly understood. In total there are 5 cleavage sites in Gag (Table 1.1). These sites share little sequence similarity and are seemingly unrelated but are cleaved with high fidelity with regard to location and sequence of cleavage. Despite the fact that the enzyme is symmetrical, the substrates it recognises are asymmetrical around the position of cleavage both in terms of the size and charge of the residues (Prabu-Jeyabalan et al., 2002).

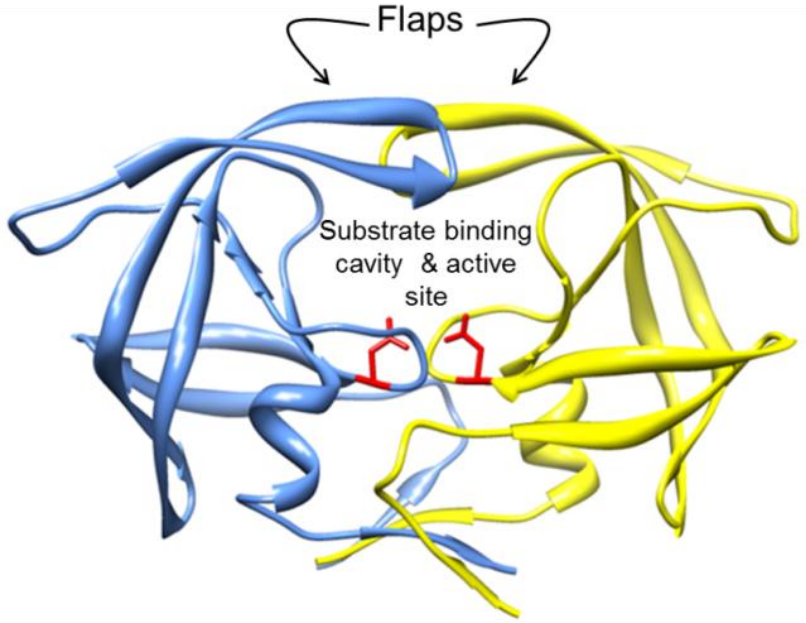


Figure 1.10 Crystal structure of HIV-1 subtype C protease (protein database ID: 2R5Q). Domain A = blue, domain b = yellow, active sites aspartic acid residues are coloured in red and rendered as sticks. Molecular graphics generated with the UCSF Chimera package (http://www.cgl.ucsf.edu/chimera) (Pettersen et al., 2004).

Table 1.1 Consensus sequences for 5 Gag cleavage sites of HIV-1 subtype C.

Cleavage site	Substrate Sequence
MA/CA	VSQNY * PIVQN
CA/p2	KARVL * AEAMS
p2/NC	NTNIM * MQKSN
NC/p1	ERQAN * FLGKI
p1/p6^{gag}	RPGNF * LQSRP

* denotes precise position of cleavage. Adapted from De Oliveira et al (2003).

Given that PR activity is essential for the production of infectious virions and continued replication, this enzyme is a prime target for therapy. Protease inhibitors have proven an effective method of treatment, yet resistance to these drugs may arise rapidly (Wensing et

al., 2010). This is in part due to the high viral replication rate and error prone nature of RT. These characteristics of HIV also prevent the development of an effective vaccine and are one of the main causes of extensive HIV diversity.

1.3 HIV DIVERSITY

HIV is characterised by extensive sequence diversity. Since the cross-species transmission events, HIV has diversified in the human population resulting in significant divergence among HIV strains (Wertheim and Worobey, 2009). Multiple cross-species transmission events are however only responsible for a small degree of the HIV diversity since these transmission events have been rare despite frequent contact between humans and SIV infected monkeys. Of 10 documented cases, only one has given rise to the current pandemic seen today (VandeWoude and Apetrei, 2006).

In an infected individual the virus exists as a swarm of related but non-identical variants. The group of viruses is known as the 'quasi-species'. Up to 10% viral diversity can exist in this quasi-species. Within a particular subtype 20% difference can occur. Roughly 25% difference in amino acid sequence of the envelope protein exists between the various subtypes of group M (Perrin et al., 2003).

Three main factors are responsible for the extensive HIV diversity; firstly, the replication rate of the virus is exceptionally high; up to 10^{10} virions may be produced daily in an infected individual. Second, reverse transcriptase is highly error prone, incorporating about one error per genome per replication round. Lastly, recombination occurs at a frequency of 7 to 30 cross-over events per replication round and is a major force behind

diversity (McCutchan, 2006). Host immune pressure and antiretroviral drugs also contribute to viral diversity (Perrin et al., 2003).

1.3.1 Strains, groups, subtypes, and circulation recombinant forms

There are many different strains of HIV. The first distinction is between HIV-1 and HIV-2, each from a separate cross species transmission event from primates to humans. These two strains are genetically distinct and have largely differing disease outcomes. HIV-2 is localised to West Africa (the Democratic Republic of Congo) and has not spread much outside this region. This strain appears to have a lower transmission capacity and is less pathogenic than HIV-1 (McCutchan, 2006). Two groups of HIV-2 exist – groups A and B, supposedly each from a different transmission event. HIV-1 is pathogenic and found on every continent. This strain can be further classified into three groups; group M (for Main), group N (for Non-M group) and group O (for Outliers). The current pandemic is caused by group M viruses. Group N and O infections have a reduced pathogenesis compared to group M viruses. Group O is rare and accounts for a very small portion of infections in Cameroon, appearing infrequently beyond the borders. Only a very small number of group N infections have been identified (McCutchan, 2006).

Group M viruses can be further divided into 9 subtypes (or clades); A- D, F-H, J and K. Circulating recombinant forms (CRF) develop as a results of recombination of two different strains within one infected individual (Perrin et al., 2003). CRF01_AE is an example of a recombinant between subtypes A and E. Subtypes tend to be geographically oriented (Figure 1.11). In the USA 96% of infections are subtype B infections while in South Africa 98% of infections are subtype C (Hemelaar et al., 2011).

Most strains, subtypes and CRFs are present in central Africa, while on other continents there are generally only one or two predominant subtypes. This lends support to the hypothesis that HIV first arouse in Africa and was subsequently spread to other continents by a few individuals (Perrin et al., 2003).

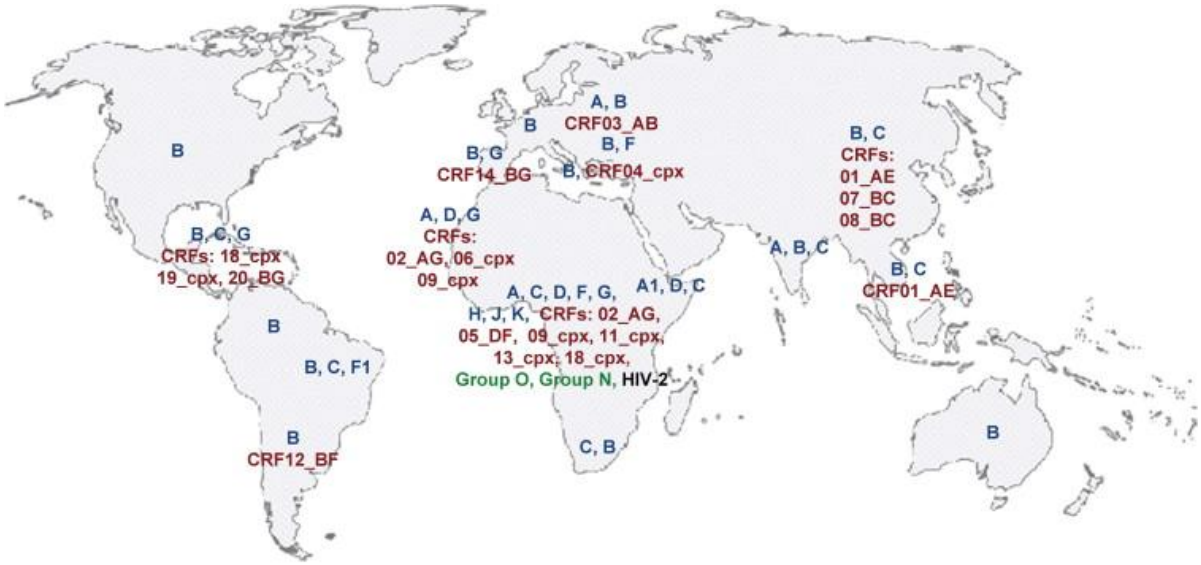


Figure 1.11 Geographical distribution of HIV subtypes. Most strains, subtypes and circulating recombinant forms (CRF) are found in central Africa, the geographic origin of the disease. Subtype B is widespread but subtype C is responsible for the highest total number of infections. Key: Blue= subtypes of group M. Red= Circulation recombinant forms (CRFs), cpx = complex CRF of more than 2 subtypes. Taken from Ramirez et al (2008).

Geographical Distribution of Subtypes

Subtype A is found mainly in East Africa and former Soviet Republics. This is the second most common subtype, accounting for 12 % of infections globally (McCutchan, 2006, Hemelaar et al., 2006, WHO, 2009). Subtype B is widespread but dominates in Australia, Western Europe and North America and accounts for approximately 11% of infections worldwide (Holguín et al., 2006). This clade is also found in other locations but generally is present as a minority. Subtype C is the dominant HIV subtype in the world, accounting for about 48% of all infections with especially high incidence rates in Southern African and

India (McCutchan, 2006, Hemelaar et al., 2006, WHO, 2009, Walter et al., 2009, Lihana et al., 2012). Subtype D is concentrated in East Africa and accounts for only 2% of infections worldwide (McCutchan, 2006, Hemelaar et al., 2006). Subtype E exists only as part of CRF01_AE. Most likely the original subtype E was present early in the epidemic. A recombination event then occurred with subtype A forming CRF01_AE. Subtype E subsequently became extinct. A pure strain E has not been isolated. Subtype F is found mainly in South America (Brazil) and also central Africa (Ramirez et al., 2008). Subtype G accounts for 5% of infections and is found in Nigeria (Charurat et al., 2012). Subtypes H, J and K are found almost exclusively in west and central Africa (Hemelaar et al., 2011). Together with subtype F, these subtypes account for less than 1 % of infections (Hemelaar et al., 2006). Circulating recombinant forms account for about 20% of infections (Hemelaar et al., 2006). In South-East Asia CRF01_AE is the major strain (McCutchan, 2006, Bandaranayake et al., 2010). This strain is mostly subtype A with a subtype E envelope sequence. CRF02_AG accounts for 8 % of infections worldwide and is found in West Africa (McCutchan, 2006, WHO, 2009, Esbjornsson et al., 2011, Charurat et al., 2012).

There are several proposed reasons for the unequal distribution of subtypes. These include social behaviour and genetic factors of the affected population. For example, immune responses are related to Human Leukocyte Antigen (HLA), which is the most polymorphic of all human genes and varies across populations (McCutchan, 2006). This may contribute to geographical concentration of a particular subtype (Perrin et al., 2003). There is also evidence that subtypes can behave differently when challenged immunologically and therapeutically. An example of this is found with subtype C.

1.3.2 Subtype C

Since its discovery during the late 1980's in Ethiopia (Bessong, 2008), subtype C has spread rapidly to sub-Saharan Africa, Brazil, China and India, mainly through heterosexual contact. A possible explanation may be particular social and sexual practises or founder effects of the population. However, this cannot be true for certain regions of China and Brazil, where this subtype has over-taken previously dominant subtypes (De Oliveira et al., 2003, Gordon et al., 2003). There is also strong evidence to suggest that subtype C was introduced into southern Africa after clades A and D (Walter et al., 2009). Furthermore, data from India indicates that the introduction of HIV to the population occurred at one or very few events. Therefore the overwhelming dominance of subtype C in India (90-95%) compared to the other subtypes present in the country (A and B together account for 5-10%) cannot be attributed to repeated re-introduction of this subtype (Rodriguez et al., 2009). This data suggests that subtype C has viral characteristics which provide it with enhanced capacity to proliferate, increased viral fitness or higher transmission capacity.

Subtype C has several distinct genomic features, including (1) a prematurely truncated *rev* open reading frame, (2) an extra NF kappa B binding site in the long terminal repeat, and (3) an enlarged Vpu protein (Ndung'u et al., 2001, Gordon et al., 2003). Subtype C PR also displays higher enzyme activity than protease from other subtypes and distinct signature sequences that distinguishes it other subtypes (Velazquez-Campoy et al., 2001). In addition, this subtype displays a fairly high level of variability. Several of the naturally occurring polymorphisms in subtype C are considered drug resistance mutations in subtype B and have been found to have a role in reduced susceptibility to therapy (Malet et al., 2007, Nijhuis et al., 2007b, Bessong, 2008, Ho et al., 2008, Jinnopat et al., 2009)

1.3.3 Variability in subtype C

The degree of variability differs depending on the subtype and the region of the genome. Certain regions of the genome are conserved across the subtypes as they are essential for function, for example, the CD4 binding site (Alexandre et al., 2011). Sequence variation of the CD4 binding site may destroy its function. Conversely, other regions can be extremely variable across the clades. This may exist to aid the virus in evading the immune response. For example the *env* gene is especially variable, as loops on the surface of the protein mutate rapidly to evade antibody response (Walker et al., 2005). Another factor known to induce sequence variation is antiretroviral therapy. Drug resistance mutations which arise during ARV therapy allow the virus to continue replicating despite the presence of the drug (Wensing et al., 2010).

The *gag* gene is relatively conserved in the structural domains despite limited sequence conservation at the primary amino acid level (Ganser-Pornillos et al., 2008). However, Gag cleavage sites are diverse within subtypes and between subtypes in drug naïve individuals, representing natural polymorphism (De Oliveira et al., 2003, Nijhuis et al., 2008). Subtype C Gag cleavage sites display significantly more diversity than cleavage sites of other subtypes. This variation could arise as a result of mutations, selected for under ARV pressure, or could be present as natural polymorphisms (Bally et al., 2000, Verheyen et al., 2009).

A 2003 study found that the degree of amino acid sequence diversity at Gag cleavage sites depended on the HIV subtype and the particular cleavage site (De Oliveira et al., 2003). In general, polymorphisms were more common in subtype C than B. At all

cleavage sites the variability of subtype B was lower than subtype C viruses. Subtype C cleavage site diversity was similar to that observed for the entire group M, which encompassed the remaining 8 subtypes. The authors found that each subtype has distinct signature sequences at the cleavage sites. Three cleavage sites (MA/CA, CA/p2, and NC/p1) were relatively well conserved over time within subtypes B, C and group M, while the p1/p6^{gag} site showed moderate variation. The p2/NC cleavage site exhibited the most variability overall, across the entire M group with a mean percentage distance of 39.2%. The same site in subtype B was considerably less variable with a mean percentage distance of 18.7 %. In subtype C this site had a mean percentage difference of 42.4%, the highest of all cleavage sites within all subtypes (De Oliveira et al., 2003). Other studies have also found this cleavage site extensively variable in therapy naïve individuals (Barrie et al., 1996, Goodenow et al., 2002, Ho et al., 2009).

Cleavage at the p2/NC site occurs between amino acid positions 377 and 378 of the Gag polyprotein in the model strain HXB2. As mentioned before, this is the first site of cleavage in Gag. This site also has the highest cleavage rate. In the De Oliveira study, positions P5, P4 and P3 of the cleavage site were found to be highly variable (De Oliveira et al., 2003) (see appendix A for nomenclature of peptide cleavage sites).

1.4 EFFECT OF VARIATION IN GAG

1.4.1 Viral fitness

Genetic variability contributes to differences in viral fitness (Goodenow et al., 2002). Fitness is an evolutionary term used to describe the ability of an organism to reproduce and adapt to its environment and may be defined as a variants ability to contribute to successive generations by producing infectious progeny. The ability to escape from drug

or host immune pressure therefore are factors which contribute to fitness (Dykes and Demeter, 2007, Rodriguez et al., 2009). Replication capacity (RC) is also part of the fitness dynamic. This term is used to describe the rate of viral replication. A viral strain with a high RC will produce more virions in a set time than a strain with a lower RC. Therefore RC is often used as a surrogate marker for viral fitness. It has been suggested that variants with lower replication capacity may be less pathogenic (Dykes and Demeter, 2007).

While genetic diversity is advantageous for continued viral evolution, some variation can elicit a fitness cost. This occurs when essential protein interactions are abrogated. Drug resistance mutations improve fitness by allowing the virus to grow in the presence of the drug but generally coincide with a decrease in RC. The decrease in RC is generally modest (2 to 10 fold) compared to the increase in resistance (>100 fold). Hence resistant viruses will almost always be selected for even at the cost of a decrease in RC (Clavel and Mammano, 2010).

An example of this can be found with Gag. Variation in this protein has been shown to have a significant effect on RC (Nijhuis et al., 2007a, Ho et al., 2008, Parry et al., 2009, Wright et al., 2010). Certain mutations in Gag CS has been shown to improve RC (Cote et al., 2001, Dykes and Demeter, 2007, Dam et al., 2009, Clavel and Mammano, 2010, van Maarseveen et al., 2012). Generally these are considered compensatory mutations which rescue RC after resistance mutations in PR. Mutations at the p2/NC cleavage site have been shown to improve RC after PI resistance has developed (Koch et al., 2006 , Ho et al., 2008). Differences in fitness have been attributed to changes in cleavage efficiency. One proposed mechanism is the altered amino acid creates an alternative contact with the

enzyme in a domain which is not usually occupied by the substrate (Clavel and Mammano, 2010).

Fitness assays

Several varieties of fitness assays have been developed in order to measure viral RC. Generally though, fitness is determined by an *in vitro* tissue culture assay, where either the percentage of infected cells or a product of viral metabolism is measured. The degree to which this measure correlates with viral fitness in an infected individual is not known, but it has been suggested that viral replication capacity may be predictive of progression to AIDS (Miura et al., 2009, Rodriguez et al., 2009). Such assays are referred to as replication fitness assays, however, it is recognised that these assays do not fully represent selective forces impacting viral fitness. An excellent and thorough review by Dykes and Demeter (2007) discussed common features, advantages and disadvantages of various fitness assays, and is summarised below.

A general feature of all RC assays is the comparison of a reference strain to a test strain, either a laboratory generated site directed mutant, or a viral strain isolated from an infected individual (Padiglione et al., 2010). Additional features of RC assays may be further classified based on 5 aspects: (1) parallel infections versus competitive growth assays (2) single or multiple cycles (3) clinical isolates versus recombinant virus assays (4) direct or indirect methods of detection and finally (5) the use of T-cell lines or primary human cells.

Parallel infections versus competitive growth assays

The first aspect is parallel infections versus competition growth assay. Competitive growth assays have the advantage of being able to detect subtle differences between viral variants. An additional benefit is the elimination of differences in culture conditions. However, such assays can be technically challenging and require prior knowledge of viral sequences. Viral recombination is another concern with this type of assay, and results may be difficult to quantify (Brockman et al., 2006).

Single or multiple cycles

Single cycle assays typically involve the deletion of the envelope gene from the viral genome. Such viruses cannot produce infectious progeny therefore infection is limited to a single round. The advantage of this type of assay is shorter time frames. Multiple cycle assays are more sensitive as differences in replication fitness can be amplified over several rounds of infection (Brockman et al., 2006).

Clinical isolates versus recombinant virus assays

Assays using whole virus or clinical specimens are performed using intact virus culture derived directly from patient samples. These assays have the advantage of being a more accurate measure of viral fitness in a patient, but detection methods are limited to measuring viral genes or gene products. A recombinant virus assay is usually performed by cloning a gene of interest into a HIV vector encoding the remaining viral genes. This type of assay is suitable for determining the effect of a particular gene or gene segment or individual mutation, since background variation is controlled (Brockman et al., 2006). Additional advantages are that modification of the recombinant virus to express a reporter

gene is less complicated and that the isolation of infectious virus from patient samples is not required. The disadvantage of a recombinant virus assay is that possible effects of other viral genes are not accounted for (Dykes and Demeter, 2007).

Detection Methods

Direct methods of measuring viral growth include p24 ELISA (Martinez-Picado et al., 2006) or reverse transcriptase (RT) activity (Brockman et al., 2006). For growth competition assays, allele-specific real time PCR, heteroduplex tracking assays (Quinones-Mateu et al., 2000) and sequencing of bulk PCR product can be used to determine the dominant strain. While these measurements provide a quantitative measure of HIV particles, a qualitative determination of viral infectivity is not assessed (Ball et al., 2003). Indirect measurements of viral growth are performed using a reporter gene, such as luciferase (Habu et al., 2005) or green fluorescent protein (GFP) (Gervaix et al., 1997). These methods offer high sensitivity and the ability to quantify infection, however variants cannot be distinguished, and therefore growth competition assays are not possible.

Target cells

Target cells of the infection may vary between T-cell lines and primary human cells (referred to as peripheral blood mononuclear cells- PBMCs). The type of cells used in the assay may influence the quantified value of fitness for a particular strain. Differences in viral RC between T-cell lines and PBMCs could be due to differences in the number of cell surface receptors (i.e. CD4, CXCR4 or CCR5), cell growth and metabolism, or cell activation status (Brockman et al., 2006). PMBCs have been used for fitness assays (Padiglione et al., 2010) yet have a few issues related to their use. For example, PBMCs have lower concentrations of nucleotides than T-cell lines and experience difficulty in

remaining in culture for extended periods of time. PMBCs also require stimulation to become susceptible to infection, therefore permissiveness to HIV infection may change over time following activation (Campbell et al., 2003). Cell lines have the advantage of less cell to cell variability than primary human cells and can be engineered to contain a reported gene.

1.4.2 Protease inhibitors

PR inhibitors (PIs) are some of the most important drugs used for ARV therapy. Understanding the effect of CS variation on proteolytic processing is therefore important for the design of new PIs. This class of drugs prevents viral maturation by inhibiting PR from processing the immature polyproteins. PIs mimic the natural substrate, binding in the active site of the enzyme. The drug becomes fixed in the substrate binding cleft as it cannot be cleaved by the enzyme (Nalam et al., 2010, Wensing et al., 2010). This prevents the enzyme from acting on any other substrate.

Mutations in PR allow the enzyme to continue functioning in the presence of the drug. These mutations alter the shape of the substrate binding cleft, resulting in the enzyme having a higher affinity for the natural substrate than for the inhibitor. Interestingly, mutations in Gag CS are also able to contribute strongly and directly to PI resistance (Doyon et al., 1996, Zhang et al., 1997, Feher et al., 2002, Nijhuis et al., 2007b, Dam et al., 2009, Clavel and Mammano, 2010). While these mutations are relatively common, on their own they are only able to cause low level resistance (Verheyen et al., 2009). Maximisation of the Van De Waals contacts is the proposed mechanism by which CS mutations contribute to PI resistance.

Computational Methods

HIV-1 PR is an attractive target for computer based drug design due to an abundance of structural information and as such, computational methods have been used extensively to study PR-inhibitor interactions (Jenwitheesuk and Samudrala, 2003, Pèpe et al., 2008, Verkhivker, 2009, Khedkar et al., 2010, Genoni et al., 2010, Nalam et al., 2010). These methods have become powerful tools to investigate interactions on a molecular scale. Researchers have used several methods for computational analysis, including molecular dynamics (Genoni et al., 2010, Perez et al., 2010, Soares et al., 2010), computational proteomics (Verkhivker, 2009) and molecular docking (Khedkar et al., 2010, Jayakanthan et al., 2010).

Molecular dynamics may be described as computational simulations, and are used to predict the physical motions of atoms and molecules. Molecules are allowed to interact for a given time; trajectories of the molecules are then determined using Newton's equations of motion. This method is often used in material science and modelling of biological molecules (Kirchmair et al., 2011). Computational proteomics may be defined as the study of protein behaviour, for example where a protein may be found in a biological sample, relative levels of expression, post-translational modification and interactions with other proteins (Colinge and Bennett, 2007). Molecular docking aims to predict the structure of a protein receptor-ligand complex and the resulting binding energies based on the 3D atomic structure of the two molecules. Consequently, molecular docking is often used in the pharmaceutical industry for purposes of drug discovery and design (Sousa et al., 2006).

In general, the purpose of drug discovery is to identify and develop small drug molecules which bind more strongly to the target protein than the natural substrate, thereby inhibiting the target biochemical reaction. Usually, drugs are discovered by *in vitro* high-throughput screening methods which test many compounds against a given target protein. This approach is expensive and time consuming. Simulated molecular docking studies can be used when the 3-dimensional structure of the target protein is known. Large databases of potential drugs /ligands can be screened virtually. Promising candidates are then tested by *in vitro* laboratory methods. This approach is much faster and cheaper.

During computational docking, structures can either be treated as rigid or flexible bodies. The main advantage of treating the structures as flexible is that is enables a search without the bias introduced by the initial model (Jenwitheesuk and Samudrala, 2003). Ligand conformation may change significantly; therefore incorporating flexibility into the ligand model is important. However the number of possible conformations increases exponentially when flexibility is taken into account, and therefore a large amount of computing power is required to complete such calculations. Complex algorithms have therefore been designed to tackle these issues, usually incorporated into software packages. There are many commercial and non-commercial docking software packages. Examples include AutoDock, Dock 6.0, GOLD and Molegro.

AutoDock was designed in collaboration with UCLA institute for Genomics and Proteomics. This is a script driven Linux program used to dock flexible ligands into protein structures and then predict the binding energy of the bound conformations of ligands with macromolecule targets. This program used a grid-based method to evaluate the binding energy of a particular ligand conformation. A probe atom is placed at every point in a grid embedded in the protein, and an energy value is calculated. This set of energy values

allows for rapid evaluation of the energy of a conformation for which an atom in the ligand coincides with the identical atom in the grid. The interaction energy between the probe atom and the target macromolecule are then calculated. Scoring is based on thermodynamics models where intramolecular energies are evaluated (Morris et al., 2009). A similar program is DOCK 6.0, which was designed in collaboration with the University of California San Francisco. It is also a script driven program for use in a Linux based system that uses a grid based scoring function (Ewing et al., 2001).The advantage of these types of program is that they allow the users freedom to determine every parameter. However, a limitation with these programs is that the user must have an appropriate knowledge of the use of script based programs.

For users lacking experience in command line driven docking, software packaged with a user friendly Graphical User Interface are available. GOLD is an example of docking program which uses a search interface. It is a product of collaborations between the University of Sheffield and GlaxoSmithKline. This program is used for calculating the docking modes of small molecules in protein binding sites. The program is provided as part of the GOLD *Suite,* a package of programs for the visualisation and manipulation of structures. The parameters of the genetic algorithmm used by GOLD are optimised for virtual screening applications (Jones et al., 1995). Molegro Virtual Docker is yet another docking program with a user friendly interface. This program uses the MolDock search algorithm to find the most likely conformation of a ligand in a macromolecular target and then to predict the binding energies (Thomsen and Christensen, 2006). This software provides a docking wizard to assist the user in the preparation of the molecules and docking calculations. While many programs and software packages are available, it is ultimately up to the user to determine which will best suit the desired purpose of the docking.

1.5 PROJECT RATIONALE

Subtype C is the dominant form of HIV infection. Several distinct qualities of this clade could be responsible for the apparent improved viral fitness at the population level. This subtype has in particular a high degree of polymorphism in Gag at the sites of PR cleavage. The p2/NC cleavage site in clade C has the highest degree of polymorphism of all cleavage sites. Therefore variation at this site could have an effect on viral replication capacity.

1.5.1 Aims and Objectives

The aim of this study was to determine the impact of Gag p2/p7 cleavage site polymorphisms on subtype C viral fitness by assaying replication capacity and cleavage of Gag.

As it was not practical to assay all possible p2/NC cleavage site variants, a computational approach was employed to select the sequences which would be used in the laboratory assays. These calculations were used to predict binding affinity between PR and the p2/NC cleavage site. The results predicted by the computational studies were compared to the experimentally measured parameters

The first hypothesis was that ligands with a higher binding affinity would result in better ligand-protein interactions and hence display enhances cleavage efficiency of Gag. This was tested in an enzyme assay. The second hypothesis was that the ligands with higher

cleavage efficiency would result in a higher replication capacity. This was tested in a replication capacity assay.

To achieve this, the overall study objectives were:

1. To use molecular dynamics simulations to predict impact of p2/NC site polymorphisms on binding affinity between PR and Gag.
2. To preform replication capacity assays to determine the impact of selected cleavage site polymorphisms.
3. To preform enzyme assays to determine the effect of these same polymorphisms on Gag cleavage by PR.

CHAPTER 2

COMPUTATIONAL STUDIES

2.1 INTRODUCTION

Computational studies are often used to screen large libraries of small drug molecules in an attempt to predict those with desirable qualities, for example molecules which will bind with high affinity to a particular enzyme, thereby inhibiting it. This process is known as docking.

Molegro Virtual Docker, a software package, was used for the computational studies in this project. This software uses a search algorithm to identify energetically favourable ligand conformations. The general hypothesis for docking is that low energy interactions are more stable hence represent more favourable ligand-protein binding.

Molegro uses the MolDock search algorithm (Thomsen and Christensen, 2006). This algorithm has been shown to have a high accuracy rating. During development and validation MolDock was able to correctly predict the binding conformation of 87% of protein-ligand complexes. This was higher than all competitors, including Glide, Surflex, GOLD, FlexX. The closest competitor, Glide, was able to correctly predict binding conformations of 82% of protein-ligand complexes (Thomsen and Christensen, 2006).

The MolDock algorithm uses a process known as guided differential evolution. This combines a differential evolution technique with a cavity prediction algorithm. The cavity

prediction algorithm was designed specifically for MolDock. This function of the search algorithm is used to focus the search during docking. Ligand orientations outside the cavity are excluded from the results. The differential evolution (DE) algorithm, created by Storn and Price (Storn and Price, 1995), evaluates protein-ligand conformations as candidate solutions according to a fitness function, calculated using a docking scoring function.

The fitness of a candidate solution is determined by the $E_{score.}$ This is the sum of the intermolecular ligand-protein interaction energy (E_{inter}) and intramolecular energy of the ligand (E_{intra}):

$$E_{score} = E_{inter} + E_{intra}$$

E_{inter} is the ligand-protein interaction energy. This term is based on piecewise linear potential (PLP) function originally created by Gehlhaar et al in 1995 (Gehlhaar et al., 1995). This function takes into account 2 sets of parameters: the first approximates steric or van der Waals forces, the other approximates hydrogen bonds. In MolDock, this scoring function is extended to account for hydrogen bond directionality and charge based interactions.

E_{intra} is the internal energy of the ligand. This term uses the same PLP function and accounts for interactions between each atom pair in the ligand (excluding pairs connected via 2 bonds or less). This term also accounts for torsion energy of bonds and energy of atom clashes. A clash is defined as distance of less than 2.0 Å between two atoms. Non-feasible conformations are therefore penalised with this term.

Scoring of each protein-ligand interaction allows the ligands to be ranked in terms of binding stability. The general hypothesis is the ligands with low energy score have a better stability. During drug discovery ligands with lower energy score are taken forward for further *in vitro* testing. Computational methods have similarly been used here to predict the affinity of PR for peptide ligands representing the p2/NC cleavage site.

2.1.1 Aims and objectives for computational studies

The primary aim of the computational studies was to screen a library of p2/NC cleavage site peptide sequences against PR in an effort to predict binding affinity between the enzyme and ligand. The objectives for this aim of the study included:

1. Retrieve all available subtype C drug naïve Gag sequences from the public database.
2. Generate 3 dimensional models of the peptide sequence at the p2/NC cleavage site to be used as ligands for docking.
3. Predict binding affinity between the peptide ligands and PR via molecular docking simulations.

2.2 METHODS AND MATERIALS

2.2.1 Sequence data

All available HIV-1 subtype C drug naïve Gag sequences were retrieved from a public database, Los Alamos HIV Sequence Database (URL: http://www.hiv.lanl.gov/). This database contains HIV sequences isolated from infected individuals. As a result all sequences were from replication competent viruses.

The library initially contained 609 nucleotide sequence entries. Nucleotide sequences were translated to amino acid sequences. Sequences were trimmed to include only the p2/NC cleavage site. The p2/NC site was considered to begin after a conserved SQA sequence immediately N-terminal to the P5 position and end on the C-terminus of a conserved SNF sequence. The sequence length varied from 10 to 15 amino acids long, depending on the number of insertions and deletions. Duplicate sequences were removed to avoid redundancy, leaving 387 unique sequences.

2.2.2 Generation of peptide ligand structures

For the docking process 3 dimensional structures of both the enzyme and ligand were required. However, crystal structures of the cleavage site peptide sequences were not available. Therefore models were manually generated using UCSF Chimera [1] (Pettersen et al., 2004). Swiss-Model was used to predict the secondary structure of all peptide ligands (Zdobnov and Apweiler, 2001, Arnold et al., 2006).

2.2.3 Protease structure

The atomic coordinates of the HIV-1 Subtype C PR in complex with Nelfinavir (NFV) was retrieved from the Protein Data Bank (PDB ID: 2R5Q). This structure was resolved by X-Ray crystallography to 2.30 Å resolution (Coman et al., 2008). The protein sequence for 2R5Q contains no drug resistance mutations (Stanford HIV drug resistance database, URL: http://hivdb.stanford.edu) (Shafer, 2006), and closely resembles the consensus sequence of subtype C PR (Figure 2.1). The NFV ligand was removed from the substrate binding cavity of the enzyme before the docking simulation.

[1] Chimera is developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from the National Institutes of Health [National Center for Research Resources grant 2P41RR001081, National Institute of General Medical Sciences grant 9P41GM103311]

```
                    5          15         25         35         45
Consensus M     PQVTLWQRPL VTIKIGGQLK EALLDTGADD TVLEEMSLPG RWKPKMIGGI
Consensus C     ..I..... .S..V...I. .......... .....IN... K.........
2R5Q            ..I...K... .S..V...I. .......... ..I..IA... ..........

                    55         65         75         85         95
Consensus M     GGFIKVRQYD QILIEICGHK AIGTVLVGPT PVNIIGRNLL TQIGCTLNF
Consensus C     .......... .........K. .......... .........M. ..L......
2R5Q            .......... ..I.....K. .......... .........M. ..L......
```

Figure 2.1 Amino acid sequence alignment of 2R5Q protease used in docking experiments. Consensus M sequence downloaded from Los Alamos HIV sequence database. Consensus subtype C protease sequence was calculated using all subtype C sequences available on Los Alamos HIV Sequence Database in the Consensus Maker tool, also available on the Los Alamos HIV sequence Database website (http://www.hiv.lanl.gov/content). Enzyme structure used for docking has the Protein Database ID 2R5Q. Dots represent amino acid identical to the consensus M sequence.

2.2.4 Docking

The graphical user interface software program Molegro Virtual Docker (MVD) was used to prepare, run and analyse molecular docking simulations. Flexible ligand docking was used. This ensured that optimal ligand geometry was always determined during docking. Ligand flexibility was defined automatically by MVD. Ligand preparation was also performed by MVD. This process included assigning bonds, bond orders and hybridisation, creating explicit hydrogens, assigning charges and detecting flexible torsions in ligands. The enzyme was regarded as rigid during docking.

All water molecules were removed from protein during preparation. Although certain water molecules are needed for the reaction mechanism, these molecules have very little or no influence on the docking simulation results, as shown by Thomsen and Christensen

during validation of MolDock (2006). Therefore the water molecules were removed from docking simulation for simplicity. All default search parameters were used for docking. Regardless of parameters used, the best overall scoring pose will always be returned for each ligand.

A maximum of 25 ligands could be analysed per docking run. Therefore the peptide ligands were analysed in batches. The NFV ligand was included for analysis in each batch. The score returned for NFV was used to normalise the scores for the rest of the batch. This was done to minimise variability between batches. All 387 ligands were then ranked based on this score.

2.3 RESULTS

2.3.1 Sequence Data Characteristics

A total of 609 sequences were initially retrieved from the public database. The sequences were from viruses isolated between 1992 and 2008. The majority of the viruses were isolated in Southern Africa; 53% from Zimbabwe and 43% from South Africa. The remaining 4% were from Argentina, Brazil, Canada, China, Cyprus, Spain and India (Table 2.1)

Table 2.1 Countries of sequence origin

Country of Origin	Number of sequences	Percentage
Argentina	1	0.16
Brazil	1	0.16
Canada	7	1.15
China	1	0.16
Cyprus	5	0.82

Spain	3	0.49
India	7	1.15
South Africa	260	42.70
Zimbabwe	324	53.20
Total	**609**	**100**

Of the 609 sequences, there were 387 unique sequences. The most common sequence (NTNIMMQKSN) appeared 22 times and was representative of subtype C p2/NC consensus sequence (Table 2.2).

Table 2.2 Frequency of Subtype C p2/NC cleavage site sequences in Los Alamos public database, showing the 6 most common sequences.

Frequency	Cleavage site sequence
22*	NTNIMMQKSN
19	.A.....R..
15	.S.....R..
14	.T.....R..
12	.S..L..R..
9	.I.....RN.

* Consensus C sequence

The N terminal region of the cleavage site was highly variable, with many insertions, deletions and substitutions/polymorphisms. This causes uncertainty in determining the exact location of P5'. In contrast to this, positions P1 and P1' of the cleavage site sequence were well conserved, consistently represented by MM. Amino acid polymorphisms observed in the sequence dataset and their respective frequencies are listed in Table 2.3.

Table 2.3 Amino acid polymorphisms observed at Gag p2/NC cleavage site in HIV-1 subtype C

Amino acid position	Cleavage position		Consensus	Polymorphisms observed and relevant frequency												
373	P5	a.a	**N**	**S**	**H**	**Q**	**G**	**K**	**T**	**A**	**R**	**C**				
		n	413	105	26	19	16	15	7	6	1	1				
		%	67.8	17.2	4.3	3.1	2.6	2.5	1.1	1.0	0.2	0.2				
374	P4	a.a.	**T**	**A**	**S**	**I**	**N**	**V**	**G**	**P**	**M**	**K**	**Q**	**H**	**L**	**C**
		n	191	133	79	55	37	34	31	23	11	5	4	3	2	1
		%	31.4	21.8	13.0	9.0	6.1	5.6	5.1	3.8	1.8	0.8	0.7	0.5	0.3	0.2
375	P3	a.a.	**N**	**S**	**H**	**A**	**T**	**Q**	**G**	**M**	**K**	**D**	**V**	**I**		
		n	476	47	22	21	16	9	6	5	3	2	1	1		
		%	78.2	7.7	3.6	3.4	2.6	1.5	1.0	0.8	0.5	0.3	0.2	0.2		
376	P2	a.a.	**I**	**V**	**M**	**N**	**T**									
		n	530	62	10	4	3									
		%	87.0	10.2	1.6	0.7	0.5									
377	P1	a.a.	**M**	**L**	**K**											
		n	518	90	1											
		%	85.1	14.8	0.2											
378	P1’	a.a.	**M**	**I**	**V**	**L**	**T**									
		n	580	21	5	2	1									
		%	95.2	3.4	0.8	0.3	0.2									
379	P2’	a.a.	**Q**	**R**	**L**											
		n	606	2	1											
		%	99.5	0.3	0.2											
380	P3’	a.a.	**R**	**K**	**G**	**N**										
		n	467	137	4	1										
		%	76.7	22.5	0.7	0.2										
381	P4’	a.a.	**S**	**G**	**N**											
		n	377	136	96											
		%	61.9	22.3	15.8											
382	P5’	a.a.	**N**	**S**	**K**	**T**	**Q**	**D**	**A**							
		n	599	3	2	2	1	1	1							
		%	98.4	0.5	0.3	0.3	0.2	0.2	0.2							

Consensus refers to the most commonly observed amino acid. The cleavage site consists of 5 amino acid residues on either side of the cleavage site. Cleavage occurs between positions 377 and 378. Letters refer to amino acid substitutions (a.a.=amino acid), n refers to the number of times the substitutions was seen, percentage is shown below..

2.3.2 Docking

As mentioned previously there were 387 ligands used in the docking procedure. All NFV ligands returned a negative score while most peptides ligands returned a positive score. The resulting score represented energy; therefore lower scores represented more stable ligand conformation. The scoring function had no units. All ligand scores were then normalised with the score for NFV in that particular batch. A normalised value close to the score of 1 represented ligands with binding almost comparable to the inhibitor NFV. A score above 1 would imply a higher binding affinity for the peptide ligand than for NFV. Sixteen batches were required to complete all docking simulations.

2.3.2.1 Scoring results

Out of a possible 387 only 11 ligands returned positive normalised scores, ranging from 7.05×10^{-1} to 2.22×10^{-3} units (Table 2.4). These 11 peptide ligand sequences were considered the top group. For reasons of symmetry and practicality the 11 ligands with the lowest scores were denoted as the bottom group. Scores from the bottom group ranged from -40.9 to -77.1 units. The median score for all 387 ligands was -12.4 units.

2.3.2.2 Peptide sequences patterns

A consistent pattern was noticed in the highest and lowest scoring ligands. The P5-P2 positions of highest scoring ligands tended to have small, hydrophobic amino acids such as Glycine and Alanine. Insertions were also common. Sequences from the bottom group closely resembled the subtype C consensus with only a few substitutions and no insertions or deletions. Positions P3'-P5' contained either identical or similar residues to those found in the consensus sequence (Table 2.5).

Table 2.4 Representative docking results for peptide ligands

Ligand Name	Molegro Docking Score (Units)	Normalised scores (Units)	Rank	Frequency
NFV*	-172.9	1.00	N/A	N/A
Ligand # 44	-128.00	0.705	1	2
Ligand # 281	-115.72	0.635	2	2
Ligand #10	-100.88	0.609	3	2
Ligand #17	-78.49	0.474	4	1
Ligand #16	-59.95	0.362	5	1
Ligand # 217	-59.39	0.308	6	1
Ligand # 39	-50.76	0.296	7	1
Ligand # 241	-43.98	0.269	8	2
Ligand # 387	-38.90	0.225	9	12
Ligand # 78	-4.34	0.028	10	1
Ligand # 202	-0.43	0.002	11	1
Ligand # 354	7076.55	-40.931	377	1
Ligand # 156	7573.73	-41.336	378	6
Ligand # 149	6602.69	-42.167	379	2
Ligand # 64	7270.63	-44.075	380	1
Ligand #14	7359.91	-44.444	381	1
Ligand # 271	7468.03	-44.485	382	1
Ligand # 80	7316.27	-46.764	383	2
Ligand # 115	8056.84	-49.947	384	1
Ligand # 164	10332.00	-56.391	385	14
Ligand # 145	9531.56	-60.871	386	4
Ligand #250	12587.10	-77.062	387	1

Ligand numbers were assigned alphabetically according to amino acid sequence. Molegro Docking score was calculated using Molegro Virtual Docker. Docking was completed in 16 batches of maximum 25 ligands each. The nelfinavir ligand was included in each batch as a control. Normalised scores were calculated by dividing the Molegro Docking Score for each ligand by the score received for nelfinavir ligand in the same batch. NFV –Nelfinavir. * median score calculated for nelfinavir, as a total of 16 different scores were obtained. Ligands above the dark line represent the top scoring ligands; below the line are the ligands with the lowest predicted binding affinity.

Table 2.5 Amino acid peptide sequences of the p2/NC cleavage from representative sequences for top and bottom ranked ligand groups

Rank	Ligand name		P5		P4	P3		P2	P1	P1`	P2`	P3`	P4`	P5`
	Consensus Group M		N		T	T		I	M	M	Q	R	G	N
	Consensus Subtype B		S		A	T		I	M	M	Q	R	G	N
	Consensus Subtype C		N		T	N		I	M	M	Q	R	S	N
1	44		G	H	P	N		V	M	M	Q	R	G	N
2	281		K			A		I	M	I	Q	R	G	N
3	10		N			A		V	M	M	Q	K	S	N
4	17		G		A	A	G	I	M	M	Q	R	S	N
5	16		G		A	A	G	I	M	M	Q	K	S	N
6	217		S		T	K		I	M	I	Q	N	S	N
7	39		G	G	H	A	G	I	M	M	Q	R	S	N
8	241	S	G	A	A	A	A	I	M	M	Q	K	S	N
9	387		S			N		I	L	M	Q	R	S	N
10	78		G		T	N		I	M	I	Q	R	N	N
11	202		S			N		I	L	V	Q	R	S	S
377	354		N		T	H		I	M	M	Q	K	N	N
378	156		N		T	N		I	L	M	Q	R	S	N
379	149		N		S	N		I	M	M	Q	R	N	N
380	64		S		A	N		I	M	M	Q	R	S	N
381	14		S		I	N		I	M	M	Q	R	S	N
382	271		N		V	N		I	M	M	Q	R	S	N
383	80		G		T	N		I	M	M	Q	R	G	N
384	115		N		G	A		I	M	M	Q	R	G	N
384	164		N		T	N		I	M	M	Q	R	S	N
386	145		N		P	N		I	M	M	Q	K	S	N
387	250		N		T	N		I	M	M	Q	K	S	N

Amino acids are colour-coded based on their physio-chemical properties. Key: Yellow= small hydrophobic; green=large hydrophobic; light blue= polar hydroxyl group; dark blue= polar group negative; purple= polar group positive.

2.4 DISCUSSION

Docking simulations were used in this study to predict PR:Gag interactions, specifically at the p2/NC cleavage site. Computational methods have been used previously to study HIV PR. These studies have used molecular dynamics (Soares et al., 2010), computation proteomics (Verkhivker, 2009) or thermodynamics simulations (Altman et al., 2008) to investigate PR: inhibitor binding, and how resistance mutations abrogate this interaction (Genoni et al., 2010). Multiple studies have used docking to assess the affinity between HIV PR and inhibitors (Olson and Goodsell, 1998, Jenwitheesuk and Samudrala, 2003, Khedkar et al., 2010, Makatini et al., 2012). One study (Pietrucci et al., 2009) has used molecular dynamics simulations to determine the substrate binding mechanism of PR,

using the p2/NC CS as the substrate. No other studies (to the author's knowledge) have used molecular docking to determine affinities between PR and the natural substrate.

Analysis of the sequence library found that the p2/NC cleavage site exhibited extensive variation, more so on the N terminal than the C terminal side of the CS. Others have also found this site to be variable (Barrie et al., 1996, Cote et al., 2001, Feher et al., 2002, De Oliveira et al., 2003, Malet et al., 2007, Ho et al., 2009, Ho et al., 2008). Polymorphisms found in the sequence data for this study included I376V and a R380K. The I376V (Malet et al., 2007, Ho et al., 2008) and the R380K (Cote et al., 2001, Myint et al., 2004, Malet et al., 2007) polymorphisms have been seen previously in viruses from both PI naïve and experienced patients.

The polymorphisms S373Q, A374P /S, and T375S have only been seen in PI experienced patients (Malet et al., 2007), but were commonly observed in this data set. Three possible explanations exist for this apparent paradox. Firstly, these sequences were incorrectly labelled as PI naïve (or the patients mistakenly identified themselves as PI naïve). Secondly, these sequences represented transmitted PI associated mutations. Thirdly, these sequences could represent naturally occurring polymorphisms. Subtype C sequences often contain polymorphisms which are considered resistance mutations in subtype B (Velazquez-Campoy et al., 2001, Martinez-Cajas et al., 2008).

The results of the docking procedure found that ligands with better docking scores were mostly composed of small, hydrophobic amino acids. This was to be expected given that the substrate binding cavity of HIV PR is lined with hydrophobic residues (Brik and Wong, 2002, Perez et al., 2010). Therefore hydrophobic amino acids would theoretically have

more favourable interactions in that environment. These interactions (hydrophilic/hydrophobic and negative/positive charge) are taken into account during scoring. Compact amino acid side chains could have an advantage, as these residues were less likely to clash sterically with protein residues inside the substrate binding cavity.

Other studies have classified the CS as 4 amino acids on either side of the location of cleavage (Perez et al., 2010, Özen et al., 2011). For this study, 5 flanking amino acid were rather used, which has also been supported (Malet et al., 2007, Ho et al., 2009, Verheyen et al., 2009). Longer peptides may allow for the determination of the effect for 3D protein folding. However this docking software may not be to correctly identify the cleavage site if the peptide were longer. In addition, full length Gag precursor has not yet been crystallised.

In addition, the uncertainty regarding the recognition and binding mechanism of PR in is in part related to the lack of knowledge of the tertiary protein structure of Gag. A 2010 study (Perez et al.) used molecular dynamics simulations to predict free energy of several PR: peptide complexes. As with docking, the implication of a lower free energy is a more stable enzyme-ligand complex, equating to higher affinity. That study found that several non-CS sequences had a lower free energy than well-defined CS. This contradicts the traditional active site binding recognition approach, suggesting an alternative recognition mechanism of PR. The authors of that paper suggested that the tertiary structure of Gag determines the location of PR cleavage by virtue of whether a site was recessed in the core of the protein or exposed on the surface. Among the exposed sites, recognition of cleavage sites is determined by which residues lie in an accessible conformation for PR to bind and cleave. With regard to the result of this study, hydrophobic residues such as those in the higher scoring ligands are less likely to be exposed to the external surface of

the protein, and therefore, according to this model are less likely to be experience optimal cleavage by PR.

With regard to possible limitations of this approach, the docking software has not been specifically designed for enzyme-peptide interactions, but mainly for studies using small drug molecules. Nevertheless, this method was chosen as it was able to score and rank a library of peptide ligands based on predicted affinity.

In summary, computational studies were used here as a tool for the objective selection of p2/NC cleavage site sequences for use in the fitness assay. Analysis of PI naïve subtype C sequences retrieved from the Los Alamos HIV Sequence Database has revealed the p2/NC CS as extensively polymorphic, with the N terminal portion of the cleavage site substantially more variable than the C terminal region. Results from the docking procedure predicted that sequences similar to the consensus sequence for subtype C would have a lower binding affinity than those with small hydrophobic amino acids in positions P5' to P2'. Three sequences with predicted high binding affinity and three with predicted low affinity binding were selected for further characterisation studies.

CHAPTER 3

FITNESS ASSAYS

3.1 INTRODUCTION

Proteolytic processing of Gag is essential for viral replication. After budding from the cell, viral maturation is achieved when Gag is cleavage by HIV protease (PR) into its 6 respective peptides. For this to take place, HIV protease (PR) must recognise and bind to specific cleavage sites in Gag. Naturally occurring variation in Gag cleavage sites (CS) is frequently observed; in particular the region between the p2 and NC peptides is considerably variable (De Oliveira et al., 2003). Given that sequence variation in Gag has been shown to influence viral replication capacity (RC) at both cleavage sites (Goodenow et al., 2002, van Maarseveen et al., 2012) and non- cleavage sites (Schneidewind et al., 2007, Wright et al., 2010), the focus of this chapter was to determine the functional consequences of natural variation at the p2/NC CS on viral fitness.

Viral variants were assayed for fitness via tissue culture and immunoblot based assays. These variants were selected from a library of 609 drug naïve subtype C sequences. Selection was based on the results of the previous chapter (*Chapter 2, Computational studies*), where docking studies were used to score predicted binding affinity between p2/NC CS sequences and PR. Three ligands from each the high and low scoring groups respectively were taken forward to these fitness assays.

Variants expressing the selected sequences were generated by site directed mutagenesis. The rate of replication (RC) was investigated by an *in vitro* tissue culture

method developed by Brockman and colleagues (2006). Cleavage of Gag was investigated using a western blotting technique previously described by Prado et al (2009).

3.2 MATERIALS AND METHODS

The aim of the computational studies described in chapter 2 was to select p2/NC peptide sequences for further studies. As mentioned previously, the group of top scoring ligands exhibited mostly small hydrophobic amino acids while bottom group ligands were mostly similar to the consensus sequence for subtype C. Three amino acid sequences representative of both top and bottom scoring groups were selected as such:

```
                    5    10
Top group:          ....|....|...
Ligand_017          GAA~~GIMMQRSN
Ligand_016          GAA~~GIMMQKSN
Ligand_241          SGAAAAIMMQKSN
Bottom group:
Ligand_164          NNT~~NIMMQRSN
Ligand_064          GSA~~NIMMQRSN
Ligand_250          TNT~~NIMMQKSN
```

3.2.1 Site directed mutagenesis

Site directed mutagenesis (SDM) was performed to generate virus which expressed Gag as per the selected peptide sequences. Mutagenesis was performed on a TOPO 2.1 cloning vector containing the HIV Gag-Protease region as an insert. The insert was isolated from patient SK254 from the Sinikithemba study. The TOPO 2.1 vector containing the Gag-protease region was a gift from Dr. Jaclyn Wright.

Site directed mutagenesis was performed using the Quikchange® Lightning kit (Agilent Technologies, previously Stratagene, California) according to manufacturer's instructions.

This kit allows for the introduction of mutations into most double stranded plasmids. Single or multiple mutations, insertions and deletions are possible with this kit.

Briefly, this method employs polymerase chain reaction (PCR) to amplify a double stranded DNA vector containing an insert with the gene of interest. Two oligonucleotide primers are used which contain the desired mutation. These primers are each complementary to the region of interest on opposite strands of the plasmid. During temperature cycling, polymerase extends the primers to generate a plasmid containing the desired mutation. Parental plasmid DNA is then digested with *Dpn I,* an endonuclease which specifically digests methylated and hemi-methylated DNA. Given that the parental plasmid DNA is isolated from bacteria, it is susceptible to this enzyme. The action of this enzyme therefore selects for the mutated plasmid. The resulting plasmid product is used to transform ultra-competent *E. coli*.

3.2.1.1 Primer design

Mutagenic primers were designed using Stratagene's web-based QuikChange® primer design program available online at http://www.stratagene.com/qcprimerdesign.

Primer design was based on the following principles:

1. Both the mutagenic primer sequences contained the desired mutations and anneal to the same sequence on the opposite strands of the plasmid.
2. Primers were between 25 and 45 bases long with a melting temperature ≥78°C.
3. The desired mutation was in the middle of the primer sequence with 10-15 bases of correct sequence on either side.
4. A minimum GC content of 40%, terminating in at least one C or G.

The primers were designed for the SK254 TOPO clone, a subtype C drug naïve sequence, lacking resistance associated mutations in Gag and protease. Briefly, this clone consisted of a 1733 base pair region of Gag-Protease generated from patient 254 of the Sinikithemba cohort. The pCR®2.1-TOPO® vector from the TOPO TA Cloning® Kit (Invitrogen, USA) was 3931bp long. The Gag_Protease region was flanked by primer binding site of 100 nucleotides in length, necessary for generation of chimeric viruses via co-transfection (described fully in *3.2.2 Generation of Chimeric Viruses)*. Molecular features of the SK254 TOPO clone can be seen in Figure 3.1

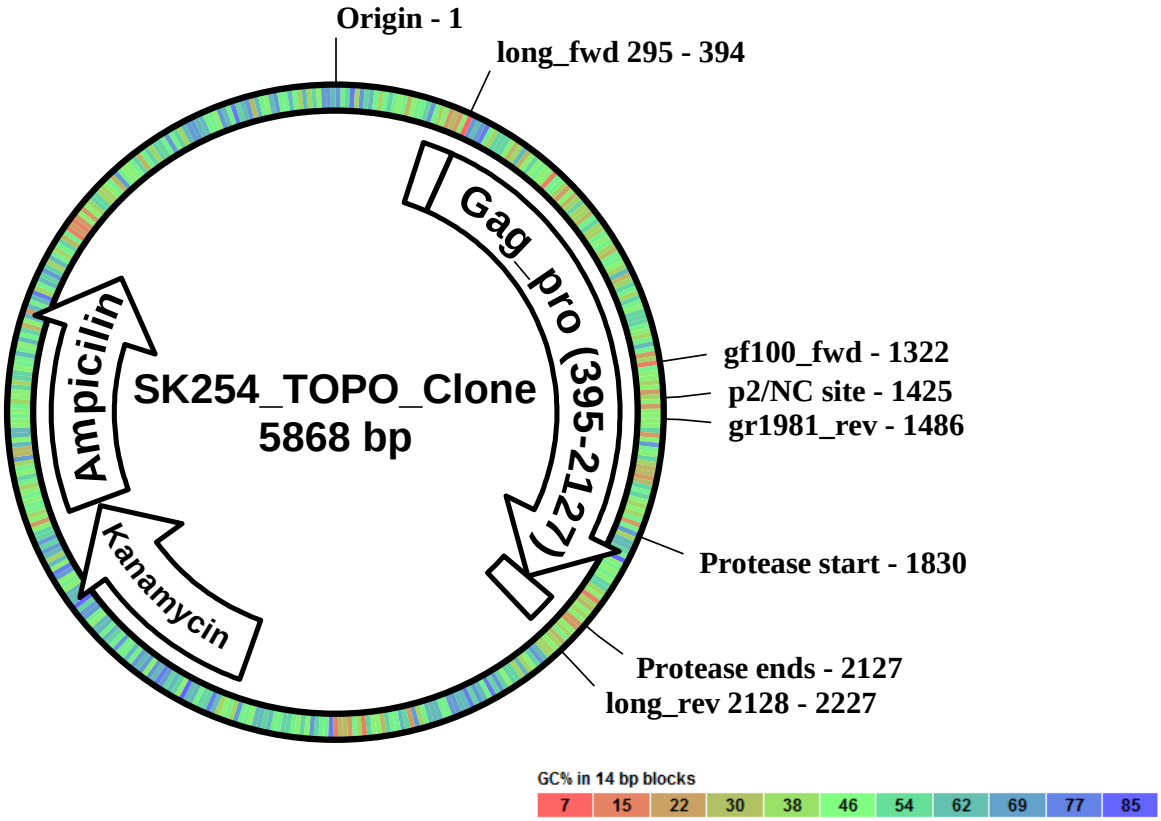


Figure 3.1 Molecular features of SK254 TOPO clone. Gag-Protease was amplified from patient SK254 of the Sinikithemba cohort and cloned into pCR®2.1-TOPO® vector. The SK254 clone was used as a template to generate all mutants. Labels *gf100_fwd* and *gr1981_rev* indicate binding location of sequencing primers, described in section *3.2.1.4 Screening of mutants.* Labels *long_fwd* and *long_rev* indicate binding position of PCR primers. Image generated with pDRAW (AcaClone Software).

This method had the limitation of only being able to introduce a maximum of six nucleotide mutations at once. Therefore sequential rounds of SDM were required to introduce all mutations. Mutations which represented insertions were introduced first, followed by nucleotide substitutions. Table 3.1 shows the mutations required to generate the desired amino acid sequences of the selected variants (see appendix B for primer information). HPLC purified oligonucleotide primers were ordered from Roche Diagnostics, South Africa.

Generation of virus 17

A glycine (GGT) insertion between amino acid positions 372-373 was introduced first. This was followed by mutating the asparagine in position 373 to alanine (AAC to GCT). Position 374 was then changed from serine to alanine (AGT to GCA). Lastly position 375 was changed from an asparagine to a glycine (AAC to GGC).

Generation of virus 16

The completed ligand 17 was used as a template to generate virus 16. Therefore only one unique primer was required. In position 380, the arginine was changed to lysine (AGA to AAA, i.e. only a single nucleotide mutation was required).

Generation of virus 241

An alanine (GCA) insertion between amino acid positions 372-373 was introduced first. Between this insertion and amino acid position 372 two additional amino acids were inserted, namely Serine-Glycine (TCA GGT). This was followed by mutating the

asparagine in position 373 to an alanine (AAC to GCA). The fourth mutation introduced was at position 374, where serine was changed to alanine (AGT to GCA). Position 375 was changed from asparagine to alanine (AAC to GCC). Finally, the arginine in position 380 was changed to lysine (AGA to AAA).

Generation of virus 64

A glycine (GGC) insertion between amino acid positions 372-373 was introduced first. The next mutation introduced was at position 374, where serine was changed to alanine (AGT to GCA). This was followed by mutating the asparagine in position 373 to serine (AAC to AGT).

Generation of virus 164

First, asparagine (AAT) was inserted between amino acid positions 372-373. This was followed by mutation at position 374, where serine was changed to threonine (AGT to ACA).

Generation of virus 250

A threonine (ACC) insertion between amino acid positions 372-373 was introduced first. The next mutation introduced was at position 374, where serine was changed to threonine (AGT to ACA). Finally, the arginine in position 380 was changed to lysine (AGA to AAA).

3.2.1.2 Mutagenesis reaction

The mutagenesis reaction was similar to a standard Polymerase Chain Reaction (PCR). The content of each reaction mixture was as follows:

Reaction component	Volume per rxn (ul)	Final concentration
10X reaction buffer	5	1X
Template DNA[a] (10 ng/ul)	1	10 ng
Primer #1 (125ng/ul)	1	125 ng
Primer #2 (125ng/ul)	1	125 ng
dNTPs mix[b]	1	N/A
QuikSolution[b]	1.5	N/A
QuikChange® Lightning Enzyme (2.5U/µl)	1	2.5 U
Sterile dH_2O	38.5	N/A
Total	50	N/A

a. For sample reactions, 10 ng was found to be the optimal amount of DNA. For control reactions pWhitescript 4.5-kb control plasmid (5 ng/µl) was used.

b. Information not available; contents of solutions trademarked by Stratagene.

c. All primers were all a final mass of 125 ng

The cycling parameters for QuikChange ® Lightning Site Directed Mutagenesis method were as follows:

Segment	Cycles	Temperature (°C)	Time	Process
1	1	95	2 minutes	Initial Denaturing
2	18	95	20 seconds	Denaturing
		60	10 seconds	Annealing
		68	30 seconds/kb plasmid length*	Extension
3	1	68	5 minutes	Final Extension

*TOPO clone-254 template DNA was 5.8 kb in length, therefore 2. 50 seconds was used for this step.

Immediately after temperature cycling a restriction enzyme *(Dpn I*, 2 µl) was added to each reaction mixture and incubated for 5 minutes at 37°C.

Table 3.1 Mutagenesis strategy for generation of 6 mutant viruses

HXB2 DNA numbering	**1898-1906**				**1907-1909**	**1910-1912**	**1913-1915**	**1916-1918**	**1919-1921**	**1922-1924**	**1925-1927**	**1928-1930**	**1931-1933**	**1934-1936**
HXB2 amino acid numbering	**370-372**				**373**	**374**	**375**	**376**	**377**	**378**	**379**	**380**	**381**	**382**
DNA. SK254 clone-Template	**AGCCAAGCA**	**~**	**~**	**~**	**AAC**	**AGT**	**AAC**	**ATA**	**ATG**	**ATG**	**CAG**	**AGA**	**AGC**	**AAT**
A.A.	**SQA**	**~**	**~**	**~**	**N**	**S**	**N**	**I**	**M**	**M**	**Q**	**R**	**S**	**N**
DNA. 17	CAACAAGCA	**~**	**~**	***GGT***	***GCT***	***GCA***	***GGC***	ATT	ATG	ATG	CAG	AGA	AGC	AAT
A.A. 17	QQA	**~**	**~**	***G***	***A***	***A***	***G***	I	M	M	Q	R	S	N
mutation introduced				1st	2nd	3rd	4th							
DNA. 16	CAACAAGCA	**~**	**~**	***GGC***	***GCT***	***GCA***	***GGC***	ATT	ATG	ATG	CAG	A***A***A	AGC	AAT
A.A. 16	QQA	**~**	**~**	***G***	***A***	***A***	***G***	I	M	M	Q	**K**	S	N
mutation introduced				1st	2nd	3rd	4th					5th		
DNA. 241	AGCCAGGCA	***TCA***	***GGT***	***GCA***	***GCA***	***GCA***	***GCC***	ATA	ATG	ATG	CAG	A***A***A	AGC	AAC
A.A. 241	SQA	***S***	***G***	***A***	***A***	***A***	***A***	I	M	M	Q	**K**	S	N
mutation introduced		2nd	2nd	1st	3rd	4th	5th					6th		
DNA. 64	AGCCAAGCA	**~**	**~**	***GGC***	A***GT***	***GCA***	AAC	ATA	ATG	ATG	CAG	AGA	AGC	AAT
A.A. 64	SQA	**~**	**~**	***G***	***S***	***A***	N	I	M	M	Q	R	S	N
mutation introduced				1st	3rd	2nd								
DNA. 164	AGCCAAGCC	**~**	**~**	***AAT***	AAT	A***CA***	AAC	ATA	ATG	ATG	CAG	AGA	AGC	AAT
A.A. 164	SQA	**~**	**~**	***N***	N	***T***	N	I	M	M	Q	R	S	N
mutation introduced				1st		2nd								
DNA. 250	AGCCAAGCA	**~**	**~**	***ACC***	AAT	A***CA***	AAC	ATA	ATG	ATG	CAG	A***A***A	AGC	AAT
A.A. 250	SQA	**~**	**~**	***T***	N	***T***	N	I	M	M	Q	**K**	S	N
mutation introduced				1st		2nd						3rd		

SK254 clone was the template used to generate mutants. Key: DNA = final nucleotide sequence after SDM; A.A. = final amino acid sequence after SDM. Underlined nucleotide sequence indicates mutations required to achieve desired amino acid sequence. Boxed cells indicate amino acid changes.

3.2.1.3 Transformation of the XL-Gold ultra-competent cells

Transformation was carried out as per manufactures instructions. Briefly, cells were removed from -80°C freezer and placed on ice to thaw for 5 minutes. *Dpn I* treated DNA from the mutagenesis reaction (2 µl) was added to 45 µl of ultra-competent cells, and incubated on ice for 30 minutes. Each reaction mixture was subjected to a 30 second heat pulse at 42°C, and immediately incubated on ice for a further 2 minutes. To each reaction, 0.5 ml preheated S.O.C media was added and incubated at 37°C with shaking at 230 rpm for 1 hour. Each reaction mixture was plated onto agar plates containing ampicillin at 50µg/ml. Ampicillin was included to select for cells containing the TOPO vector, as this vector contains the ampicillin resistance gene. The plates were incubated overnight for at least 16 hours.

3.2.1.4 Screening of mutants

Five colonies from each transformation reaction were randomly selected for screening PCR. Using a sterile pipette tip, colonies were first touched to a master plate then added to 10 µl of sterile dH_2O, and boiled for 2 minutes at 95°C. Two µl of this solution was used as the template DNA for the screening PCR. *TaKaRa Ex Taq* HS version PCR kit (Takara Biotechnology, Japan) was used for the PCR. Components for the screening PCR were as follows:

Reaction component	Volume per rxn (ul)	Final Concentration
Template DNA	1.0	-
10X *Ex Taq* buffer[a]	2.5	1X
Forward primer (10µM)	0.5	0.2 µM
Reverse primer (10µM)	0.5	0.2 µM
dNTP mixture (2.5 mM each)	2.0	200 µM
Sterile dH_2O	18.4	-
TaKaRa Ex Taq (5 U/ul)	0.125	0.625 U
Total Volume	25	-

a. Ex Taq buffer contained 20mM $MgCl_2$, resulting in a 2 mM final concentration.

The primers used for this reaction are given below:

Primer name	Sequence	HXB2 position
Long_fwd:	5′-GACTCGGCTTGCTGAAGCGCGCACGGCAAGAGGCGAGGGGCG GCGACTGGTGAGTACGCCAAAAATTTTGACTAGCGGAGGCTAGAA GGAGAGAGATGGG-3′	695→794
Long_rev	5′-GGCCCAATTTTTGAAATTTTTCCTTCCTTTTCCATTTCTGT ACAAATTTCTACTAATGCTTTTATTTTTTCTTCTGTCAATGGCCA TTGTTTAACTTTTG-3′	2706←2805

The cycling parameters for the screening PCR were as follows:

Segment	Cycles	Temperature (°C)	Time	Process
1	1	94	2 minutes	Initial Denaturing
2	40	94	30 seconds	Denaturing
		60	30 seconds	Annealing
		72	2 minutes	Extending
3	1	72	7 minutes	Final extend

A 1% agarose gel was used to confirm the presence of a PCR product. The gel was prepared by adding 0.5 g of agarose to 50ml of 1X TBE buffer (Sigma-Aldrich, USA) and heating until the agarose completely dissolved. The solution was poured into a casting tray. Once set, the gel was placed in an electrophoresis tank and covered with 1X TBE running buffer. Two µl of DNA Molecular Weight Marker X (Roche diagnostics, Germany) and 5µl of each sample was mixed with 2 µl of gel loading dye (Sigma-Aldrich, USA) containing 1 in 10,000 dilution of GelRed™ (Biotium, USA) before being loaded onto the gel. The gel was run at 100 V for 30 minutes on an Electrophoresis Power Supply- EPS 301 (Amersham Biosciences, Sweden). The gel was viewed under UV light using the GelVue UV Transilluminator (SynGene, London).

DNA was extracted from colonies positive for a PCR product via the Fermentas GeneJet Plasmid Miniprep Kit (Thermo Fisher Scientific, USA) according to manufacturer's instructions. The purified plasmid DNA was sequenced to confirm the presence of the desired mutation. Sequencing primers used are listed below:

Primer Name	Primer sequence (5'-3')	HXB2 binding region
gf100	5′-TAGAAGAAATGATGACAG-3′	1817→1834
gr1981	5′-CCTTGCCACAGTTGAAACATTT-3′	1960←1981

Sequencing was performed with the BigDye® Terminator Ready Reaction Mix V3 (Applied Biosystems, USA). Sequencing reaction mixture was prepared as follows:

Reagent	Volume per reaction (µl)	Final concentration (per rxn)
5X sequencing buffer	2	1X
Primer (2 µM)	2.6	0.52 µM
Template DNA(150 ng)	1	15ng/µl
Water	4	-
BigDye terminator mix	0.4	-
Total Volume	10	-

The sequencing reaction was performed in an optical 96-well plate (Applied Biosystems, California) under the following conditions:

Segment	Cycles	Temperature (°C)	Time	Process
1	1	96	1 minute	Initial Denaturing
2	25	96	10 seconds	Denaturing
		50	5 seconds	Annealing
		60	4 minutes	Extending

The sequencing products were purified immediately following temperature cycling. To each well, 1µl of EDTA was added, followed by 26 µl of a 1 in 26 solution of 3M sodium acetate in 100% ethanol. The plate was covered by foil, briefly vortexed and centrifuged at 3000 x g for 20 minutes. The plate was carefully removed from the centrifuge, inverted onto folded paper towel and centrifuged at 150 x g for 1 minute to remove the liquid. Immediately following this step, 35 µl of ice cold ethanol was added to each well, the plate was then centrifuged at 3000 x g for 5 minutes. Once again, the plate was carefully removed from the centrifuge, inverted onto folded paper towel and centrifuged at 150 x g for 1 minute. The plate was dried at 50°c for 5 minutes and stored at -20°C. Prior to sequencing the dried pellets were resuspended in 10µl formamide and denatured in at 95°C for 3 minutes. Samples were loaded onto the ABI 3130 Genetic Analyzer (Applied Biosystems, California). Resulting chromatograms were analysed with the Sequencher™ 4.10.1 software package (Gene Codes.).

3.2.2 Generation of chimeric viruses

Chimeric viruses were generated by inserting the mutant Gag-Protease region of the SK254 clone into a NL4-3 backbone via a co-transfection method (Wright et al., 2010). Briefly, PCR was performed on the SK254 TOPO clone with a set of 100-mer primers each complementary to the pNL4-3 clone on either side of the Gag protease region (HXB2 positions 695→794 and 2706←2805). The resulting product was co-transfected via electroporation into CEM-GXR25 cells (Brockman et al., 2006) with a Gag-Protease deleted pNL4-3 plasmid (pNL4-3ΔGag-Pro). The long primers allow sufficient overlap for recombination to occur, generating a complete HIV genome.

3.2.2.1 Gag-Protease amplification by PCR

Mutated Gag-Protease region of the SK254 TOPO clone was amplified using the set of 100 nucleotide primers (previously described in section *3.2.1.4. Screening of mutants*). Two PCR mixtures of 50 µl were prepared for each sample. The content of each reaction was as follows:

Reaction component	Volume per rxn (ul)	Final Concentration
Template DNA (50 ng/µl)	2.0	2 ng/µl
10 X *Ex Taq* buffer[a]	5	1 X
Forward primer (10µM)	0.8	0.16 µM
Reverse primer (10µM)	0.8	0.16 µM
dNTP mixture (2.5 mM each)	4.0	200 µM
Sterile dH_2O	37.15	-
TaKaRa Ex Taq (5U/ul)	0.25	0.025 U/µl
Total Volume	**50**	**-**

a. Ex Taq buffer contained 20mM $MgCl_2$, resulting in a 2 mM final concentration.

The following temperature cycling parameters were used to amplify the Gag-Protease region:

Segment	Cycles	Temperature (°C)	Time	Process
1	1	94	2 minutes	Initial Denaturation
2	40	94	30 seconds	Denaturation
		60	30 seconds	Primer Annealing
		72	2 minutes	Extension
3	1	72	7 minutes	Final extension

Agarose gel (1%) electrophoresis was used as described in section *3.2.1.4 Screening of mutants* to visualize the resulting PCR fragment of 1.9 kb.

3.2.2.2 pNL4-3ΔGag-Pro plasmid digestion

The pNL4-3ΔGag-Protease plasmid was generated by deleting the Gag-Protease region and inserting a *BstEII* restriction endonuclease site (GGTnAC_C). The plasmid was successfully digested with *BstEII* (New England Biolabs, USA*)* to obtain a linear DNA fragment. Digestion was performed on the day of co-transfection to reduce re-ligation of the restriction site. The digestion reaction was prepared as follows and incubated at 60°C for 2 hours:

Reaction component	**Volume**
Plasmid	10 µg/sample
BstEII enzyme	2 units/ µg plasmid
Buffer (10X concentration)	1/10 of total reaction volume
BSA (100X concentration)	1/100 of total reaction volume
Sterile H_2O	Fill to final desired volume

3.2.2.3 Co-transfection by electroporation

As mentioned above, the Gag-Protease PCR amplified fragment was co-transfected with the linear pNL4-3ΔGag-Protease plasmid into CEM-GXR25 cells. This cell line, first described by Gervaix et al. (1997), is a stable human T cell line containing a Tat-inducible GFP reporter plasmid. Following infection GFP is expressed in a Tat-dependent manner. Initially these cells expressed CD4 and CXCR4 but were later modified by Brockman and co-workers (2006) to also express CCR5. Cells were maintained in R10 medium (RPMI-1640 supplemented with 2mM L-glutamine [both from Sigma-Aldrich, USA], 50 U/ml penicillin-streptomycin, 10% foetal calf serum, 10mM HEPES [all from Gibco, New York]).

To achieve the co-transfection, two million CEM-GXR25 cells was co-transfected by electroporation with 2 x 50 µl PCR reaction mixtures and 10 µg digested plasmid for each mutant. Samples were electroporated at 250 volts and 950 µF. The cells were then rested

in the electroporation cuvette at room temperature for 5 minutes before being transferred to 10 ml of warm media (R10 as described above, with 4 µg/ml polybrene) containing 1 million GXR cells. Cultures were incubated at 37°C, 5% CO_2 for 4 days, after which an additional 5 ml of fresh media was added to the culture.

Every second day fresh media was added to the cultures and the percentage of GFP positive cell was determined by flow cytometry on a FACSCalibur (BD Biosciences, San Jose). When cells reached ~30% infection, virus in the culture supernatant was harvested by centrifugation at 1700 x g for 10 minutes, then stored at -80°C.

3.2.2.4 Flow Cytometry

To determine percentage infection, 1 ml of culture was centrifuged at 1700 x g for 10 minutes to pellet cells. Cells were fixed in 200 µl phosphate-buffered saline (PBS) containing 2% paraformaldehyde. GFP expression was measured using flow cytometry. Cultures above the threshold of 0.05% GFP+ cells were considered positive for infection.

3.2.3 Viral Replication assay

3.2.3.1 Infectivity calculations

Viral infectivity were measured as previously described (Brockman et al., 2007, Schneidewind et al., 2007, Miura et al., 2009) to achieve 0.3% infection cells on day 2 of the assay. Briefly, the measurement was performed by adding 400 µl of virus stock to 1 million GXR cells in 100 µl R10 media. To generate the negative control reactions, virus stock was replaced with media. The culture was incubated for 2 hours at 37° C and

agitated every 30 minutes. The cells were pelleted by centrifugation at 1700 x g for 10 minutes followed by washing with PBS. The cell pellet was re-suspended in 500 µl of pre-warmed R10 media, plated out into a 24 well cell culture plate and incubated overnight at 37°C. The following day 1 ml of fresh pre-warmed media was added to each culture. On day 2 of the assay the percentage of infected cells was determined by flow cytometry. The percentage infection results were used to proportionally scale down the volume of viral stock required to achieve 0.3% infection on day 2 of the assay. This was calculated as follows:

$$\text{Volume of virus stock to add} = \frac{0.3\ \%}{\text{(Percentage of cells infected on day 2)}} \times 400\ \mu l$$

3.2.3.2 Replication Capacity Assay

Each experiment was scaled up 6 times from the infectivity calculation to a total of 9 ml. This was done to ensure enough material for both the flow cytometry and the western blotting assay. Six million cells were incubated with the virus for 2 hours at 37°C with agitation every 30 minutes before being washed with PBS. Based on the results from the viral infectivity calculation, different amounts of each virus stock were used. After washing, cells were plated out into 3 ml of fresh warm R10 media in a 6 well culture plate and incubated at 37°C. On day 1 of the assay, 6 ml of media was added to each well. Each experiment was performed in duplicate. From day 2 until day 6 of the experiment, 3 ml was removed from each culture for analysis, and replaced with 3 ml of fresh media. 500 µl was analysed by the FACSCalibur to measure percentage infection. The remaining 2.5 ml was reserved for the western blotting cleavage assay, as described in section *3.2.4 Western blotting cleavage assay*.

3.2.3.3 Analysis

The mean slope of exponential growth from days 2 to 6 was determined using the semi-log method in MS Excel. This value was normalised to that of the NL4-3 virus control, which was included on each plate to give a measure of RC for each virus relative to the NL4-3 control.

3.2.4 Western blotting cleavage assay

3.2.4.1 Sample preparation

Western blotting was used to assess the pattern of Gag cleavage during the RC assay (section *3.2.3.2)*. Cytoplasmic extracts of samples taken from day 2-6 of the RC assay were analysed. Samples were centrifuged at 1700 x g for 10 minutes and cell pellets were washed with cold PBS. Cells were lysed with Cytobuster^{TM} Protein Extraction Reagent (Novagen, California). Protease inhibitor cocktail (Sigma-Aldrich, USA) was added to the reagent at a 1 in 200 dilution to reduced proteolysis and degradation. Cell pellets were resuspended in 150 µl Cytobuster lysis reagent per 10^6 cells and incubated at 4°C for 30 minutes with intermittent agitation. Insoluble cell debris was pelleted by centrifugation at 4600 x g for 30 minutes at 4°C and discarded. Protein concentrations of supernatants were determined by the Bradford assay (Bradford, 1976). The Bradford reagent was purchased from Fermentas (Thermo Fisher Scientific, USA). Absorbance values at 590 nm were read on a NanoDrop® spectrophotometer (NanoDrop Technologies Inc, USA).

Samples were prepared for electrophoresis by adding an equal volume of Laemmli Sample Buffer with β-merceptoethanol (BioRad Laboratories, Inc. California), and boiling for 10 minutes at 95°C.

3.2.4.2 Western Blotting

Prior to western blot analysis, proteins were separated via sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS PAGE). Briefly,10 µg of total protein from each sample was loaded onto 12% Mini Protean® TGX™ pre-cast SDS polyacrylamide gel (BioRad Laboratories, Inc.,). Five µl of Precision Plus Protein™ Standards Kaleidoscope (BioRad Laboratories, Inc.) was also loaded onto each gel as a protein molecular mass marker. Gels were run at 18 mA per gel for 45 minutes until the dye front reached the end of the gel.

Proteins from cell lysates were transferred onto nitrocellulose membrane (BioRad Laboratories, Inc,) in a Trans-Blot® Semi-Dry blotting apparatus (BioRad) at 15 V for 1 hour. Unoccupied protein binding sites on the nitrocellulose were blocked by overnight incubation with shaking at room temperature in 5% (m/v) low fat milk powder, 0.5% (v/v) Tween in Tris Buffered Saline (TBS).

Blots were probed with primary mouse monoclonal anti-human beta actin or anti-HIV1 p24 antibodies. This was followed by incubation with secondary rabbit polyclonal anti-mouse IgG conjugated to horse radish peroxidase (HRP). All antibodies were purchased from Abcam (Cambridge, The United Kingdom) and used at a 1 in 4000 dilution in Calbiochem® SignalBoost™ Immunoreaction Enhancer solution (Merck, Germany). Incubations were carried out at room temperature for 1 hour with shaking. After both primary and secondary antibody incubations, blots were washed 5 times for 10 minutes each with 0.5% (v/v) Tween-TBS.

Antigen-antibody complexes were visualised via chemiluminescence by incubation with LumiGLO® Chemiluminescent substrate (Kirkegaard & Perry Laboratories, USA) for 1 minute. Blots were viewed in a Bio-Rad ChemiDoc MP Imaging System using an exposure time of 1 minute or longer if required.

3.2.4.3 Analysis of blots

Density of p55 and p24 bands were quantified using the image analysis software ImageJ (Rasband, 2009).

3.3 RESULTS

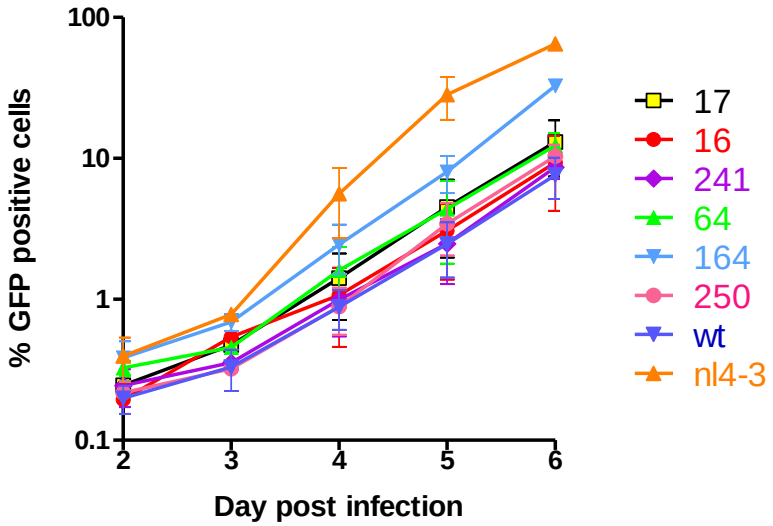
3.3.1 Replication capacity of chimeric HIV variants

The RC assay measures the ability of a virus to replicate, determined by the percentage of cells which become infected as a function of time. Flow cytometry was used to measure percentage infection. RC was calculated as the slope of the exponential curve of each variant.

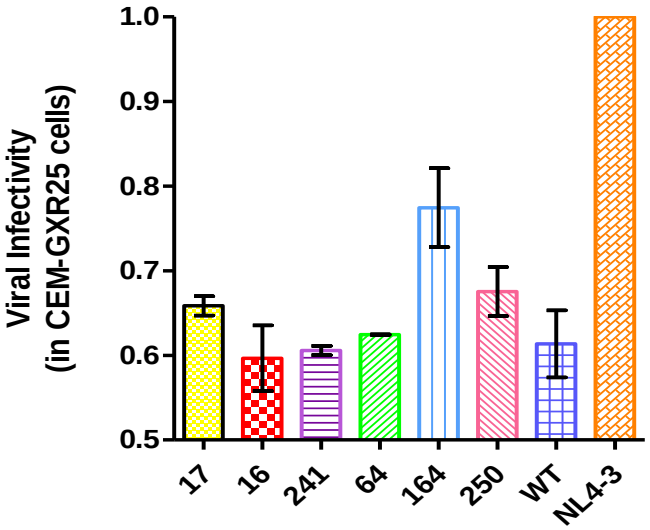
Apart from the NL4-3 control, there were no significant differences in the replication kinetics (Figure 3.2a). Viral infectivity was normalised to NL4-3 for all variants, and ranged from 0.60 to 0.77. Virus 164 had the highest infectivity and virus 16 had the lowest (Figure 3.2b). Viruses were grouped into high or low scoring ligands based on scores from the docking procedure. No significant difference (p=0.1320) was found between these two groups (Figure 3.2c). However, ligands which scored higher in the docking analysis

tended to have a lower viral infectivity. Viral RC and matching amino acid sequences can be seen in Table 3.2.

(a)



(b)



(c)

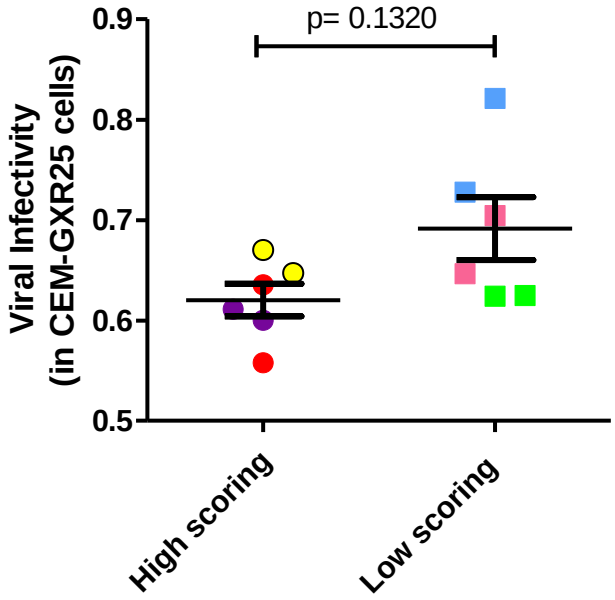


Figure 3.2 Replication kinetics of mutant viruses. CEM-GXR25 cells were infected with 8 viral variants and the course of infection was followed daily over a 6 day period. (a) Viral replication kinetics as measured by daily FACS analysis. (b) Viral infectivity, calculated as the slope of the exponential growth. (c) Associations between viral infectivity and docking score as calculated in Chapter 2.

Table 3.2 Replication capacities and associated p2/NC amino acid sequences

Sample	Mean RC		P5		P4	P3		P2	P1	P1'	P2'	P3'	P4'	P5'
Consensus Subtype C	-		N		T	N		I	M	M	Q	R	S	N
nl4-3	1		N		P	A	T	I	M	I	Q	K	G	N
164	0.774	N	N		T	N		I	M	M	Q	R	S	N
250	0.675	T	N		T	N		I	M	M	Q	K	S	N
17	0.659		G		A	A	G	I	M	M	Q	R	S	N
64	0.624	G	S		A	N		I	M	M	Q	R	S	N
Wt	0.614		N		S	N		I	M	M	Q	R	S	N
241	0.606	S	G	A	A	A	A	I	M	M	Q	K	S	N
16	0.597		G		A	A	G	I	M	M	Q	K	S	N

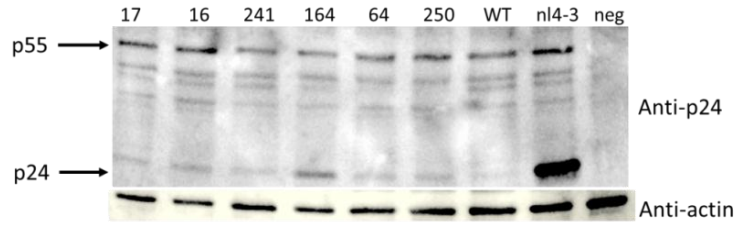
Viruses are ranked from highest to lowest replication capacity values. Amino acids are colour coded on the same basis as Table 2.5.

3.3.2 Gag Cleavage

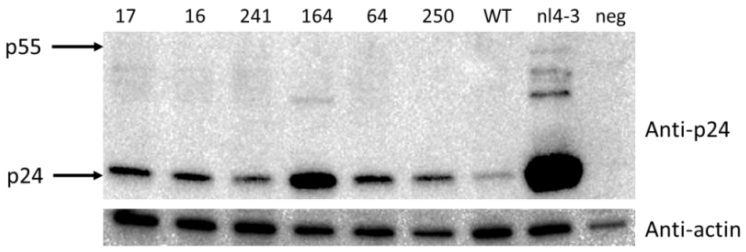
Cleavage of Gag by protease was assessed by western blotting as previously described by Prado et al. (2009). Samples for analysis were collected from HIV infected cell culture as described in section 3.2.3.2 *Replication Capacity Assay.* The relative levels of processed and unprocessed Gag molecules, namely p24 and p55 were studied in order to determine any differences between the wild type (unmutated SK254) and mutant viruses. Changes in this p55/p24 ratio would indicate altered polyprotein cleavage resulting from the p2/NC CS mutations.

Initially (on day 2 blots) only the p55 was visible, albeit faintly. By days 3 and 4 the p24 band was more evident, and by day 6 it was the dominant band with the p55 band barely visible (Figure 3.3).

(a) Day 4



(b) Day 5



(c) Day 6

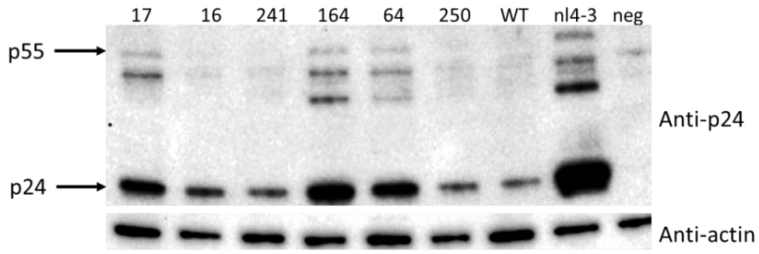


Figure 3.3 Representative western blot analyses of cell lysates. Samples were taken from (a) day 4 (b) day 5 (c) day 6 post infections of CEM-GXR-25 cells. Total protein was extracted from cell pellets, separated on SDS PAGE, blotted onto nitrocellulose membranes and probed with anti-p24 and anti-actin antibodies. Mature p24 protein and unprocessed p55 Gag precursor are indicated by arrows. Actin is used as a loading control. WT- wild type refers to the unmutated SK254 Gag sequence.

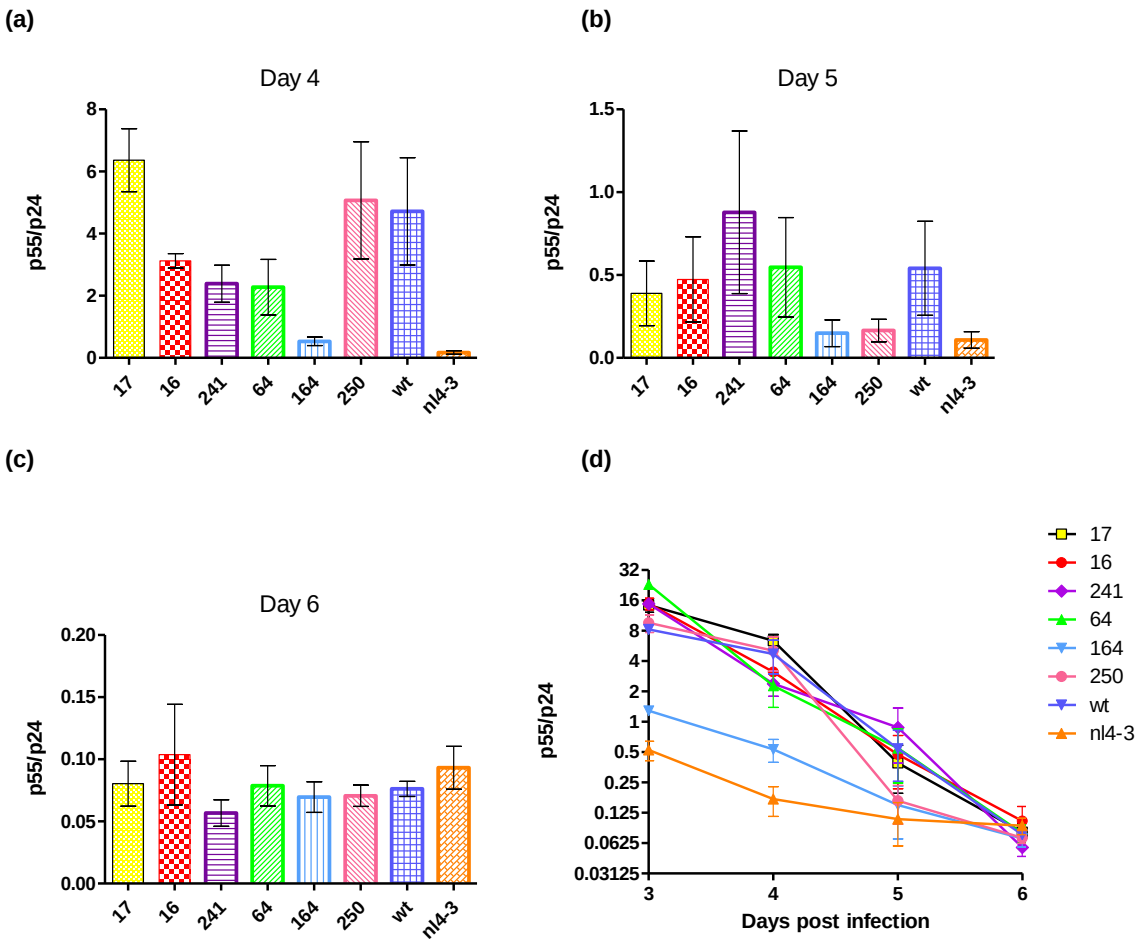
In an attempt to quantify the processing of the Gag polyprotein precursor, the p55/p24 ratio was calculated for each virus over days 3 to 6 of the RC assay. A low ratio would indicate more advanced processing of the precursor. Besides the nl4-3 control, virus 164 consistently displayed efficient processing of the Gag p55 precursor to the mature p24 peptide. The p55/p24 ratio dropped for all viruses between days 4 and 6, but only slight differences between viruses were observed. The ratios are shown in Figure 3.4. The results for each day are briefly described below.

On day 4 (Figure 3.4a), virus 164 had the lowest p55/p24 ratio of 0.53; and at 6.4, virus 17 had the highest. Virus 250 and WT also displayed poor p55 processing with p55/p24 ratios of 5.07 and 4.72 respectively. Moderate p55 cleavage was seen in viruses 16, 241 and 64, with p55/p24 ratios of 3.12, 2.39 and 2.28, respectively.

On day 5 (Figure 3.4b), virus 164 continued to have the most efficient p55 processing (besides for the nl4-3 control). Conversely, virus 17 no longer had the highest p55/p24 ratio, but displayed improved Gag processing. Virus 250 had a lower p55/p24 ratio of

0.165 from the previous day's sample. Viruses with moderate Gag processing were 16, 64 and the wild type. Virus 241 experienced poor p55 processing in day 5 samples and had the least decline in p55/p24 ratio from the previous day.

By day 6 (Figure 3.4c), all the viruses displayed roughly equivalent states of Gag processing. The p55/p24 ratio for all viruses, including the nl4-3 positive control virus, ranged from 0.057 to 0.103. Interestingly, the lowest p55/p24 ratio was not from virus nl4-3 but from virus 241.



(e)

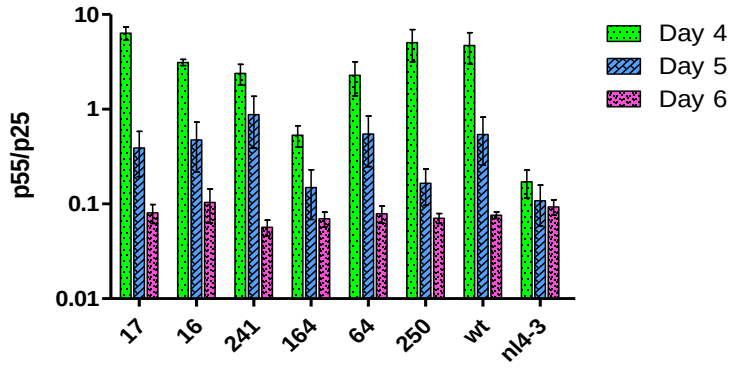


Figure 3.4 Immunoblot analysis of p55/p24 viral protein ratio in CEM-GXR25 cell lysate. p55/p24 ratios were calculated by band densitometry using ImageJ software from (a) day 4, (b) day 5 and (c) day 6 blots. (d) and (e) Change in p55/p24 ratios over course of 6 days of infection. Graphs represent mean quantification of two blots each measured twice. wt= Wild type, refers to the unmutated SK254 Gag sequence. Scales have been adjusted to allow for direct comparison between viruses.

3.4 DISCUSSION

As previously mentioned, studies have shown that elements in Gag can have a dominant effect on viral replication fitness, independent of drug sensitivity (Wright et al., 2010). Specifically, polymorphisms at the p2/NC CS has been shown to have impact on replicative fitness of the virus (Goodenow et al., 2002, Ho et al., 2008, van Maarseveen et al., 2012). This study was performed to gain deeper understanding of the functional consequences of variation at this site.

From this study, it was found that viruses resembling the subtype C consensus sequence had improved replication kinetics than those with more variable sequences. These results are unsurprising given that, in terms of Darwinian evolution, fitness of a viral variant is directly proportional to the frequency with which it is observed in the population. By

definition, the consensus sequence is the most frequently observed. However it must be noted that subtype C performs poorly in replicative fitness assays (Abraha et al., 2009).

Differences in replication capacities of the 6 variants were not substantial, suggesting that the variation at p2/NC CS has only a modest effect on RC. Given that others have shown a dominant effect of variation at cleavage sites and non-cleavage sites of Gag, an alternative explanation is that the particular cleavage site variations used in this study were not those capable of causing large differences in RC.

Besides the nl4-3 control virus, the highest RC was found with virus 164, identical to the consensus sequence for subtype C. This variant differed from virus 250 at position 380 (in the HXB2 reference strain). Virus 250 displayed a lysine residue while virus 164 had an arginine residue in that position. Similarly, viruses 17 and 16 differed at the same position; with 17 containing an arginine and16 instead had a lysine residue. In both instances, viruses with the arginine residue had a higher RC. Analysis of the sequence data revealed that the polymorphism was found frequently in the population. Out of a possible 609, arginine appeared 467 times and lysine appeared 137 times. The fact that arginine is more common in this position implies a fitness advantage.

A possible explanation for the replication advantage of arginine over lysine may be found when the structures of these amino acids are considered. While both lysine and arginine have polar, positive side chains, arginine potentially has 2 additional hydrogen bonds donors due to its three nitrogen atoms as opposed to the one found in lysine (Figure 3.5). Given that, up to nine hydrogen bonds hold the substrate in the active site (Das et al., 2010), this may provide CSs containing an arginine residue with additional contacts during

PR during cleavage. The consensus sequences for subtype C, B, group M and the HXB2 reference sequence all contain an arginine residue in position 380, while the infectious molecular clone pNL4-3 has a lysine. Presence of a lysine in position 380 has been linked to PI exposure (Malet et al., 2007).

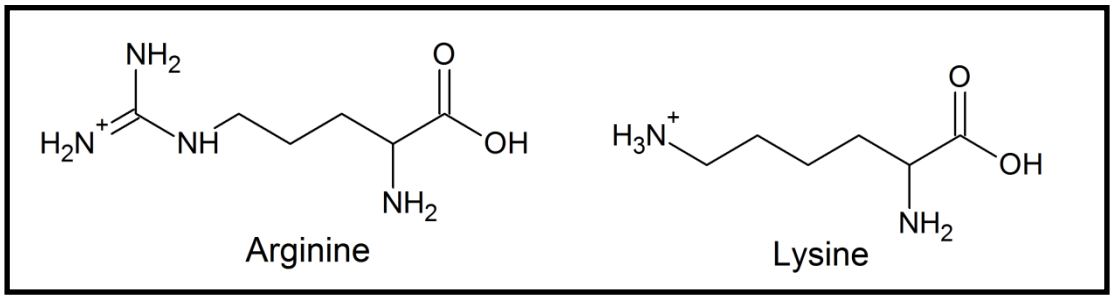


Figure 3.5 Structural comparison of amino acids arginine and lysine

Immunoblot analysis revealed that the processing rate of p55 showed some correlation with replication rate. Viruses which performed well in both assays were the nl4-3 control and virus 164. The p55 processing of the remaining viruses was not substantially different from each other. A similar result was found in a study by Goodenow and co-workers (2002), which showed higher RC variants had faster rates of p24 production. These authors also found that the optimal processing was found when Gag and Pol regions were derived from the same allele. Another study also found that residues in NC of Gag and PR displayed significant co-variation. (Ho et al., 2009). This suggests specific functional interactions between Gag and PR which may enhance cleavage *in vivo*. Given that Gag and PR were not matched for this study, maximum cleavage efficiency of each p2/NC variant cannot be assured.

In summary, these findings suggest that variation at the p2/NC cleavage site has only a limited impact on viral fitness of HIV-1 subtype C. To further elucidate interactions

between PR and these variant CS, synthetic substrate assays were performed, as described in chapter 4.

CHAPTER 4

ENZYME ASSAYS

4.1 INTRODUCTION

The order of Gag cleavage during proteolytic processing is strictly regulated. However, the precise factors which control the processing are currently not well defined. For many proteases such as chymosin, chymotrypsin, trypsin (Voet and Voet, 2004), proteolysis is determine by amino acid sequence. However, due to the minimal CS sequence homology of PR substrates this appears unlikely to be the case with HIV-1 PR. There have been many studies proposing different determinants of cleavage; however, the main factors which control PR cleavage remain poorly defined (Perez et al., 2010). An understanding of the elements which regulate PR:Gag interactions are relevant in determining factors which influence viral fitness.

Enzyme assays can be used to investigate the substrate preference of an enzyme against a range of substrates. Such assays were used in this study to determine the effect of CS variation on PR cleavage. Comparison between enzyme assay and replication capacity assay result could also provide a deeper understanding of the recognition mechanism of PR.

For the purpose of the enzyme assays, the HIV protease gene from SK254 TOPO clone was recombinantly expressed via the pMAL protein fusion and expression system. The PR gene from SK254 was specifically used (as opposed to commercially available PR) to ensure consistency between the enzyme assay and the replication capacity assay

(described in chapter 3). The Protease-Glo™ Assay kit (Promega, USA) was used to examine enzyme preference for a range of substrates representing the p2/NC CS of Gag.

4.2 MATERIALS AND METHODS

4.2.1 Recombinant expression of HIV protease from SK254 TOPO clone

According to the literature, functional PR can be difficult to express, mainly because the enzyme can be mis-folded and toxic (Stebbins and Debouck, 1994, Komai et al., 1997, Vickrey et al., 2003, Volonte et al., 2011). Incorrect protein folding results in an insoluble, non-functional enzyme. PR is often expressed as insoluble inclusion bodies. This is accompanied with the necessity to refold the protein and uncertainty of whether the refolded protein retained any biological activity; a reduction in protein yield also accompanies the refolding procedure. Expression of toxic proteins can be problematic because the bacterial host cells are killed as expression begins; as a result expression levels are usually very low.

The pMAL expression system was chosen to express this enzyme, for several reasons (discussed below). This system provides a method for expression and purification of proteins produced from a cloned gene. The cloned gene is ligated into the pMAL vector downstream of the *malE* gene of *E. coli*, which encodes maltose binding protein (MBP) (Figure 4.1). The result is the expression of a fusion protein of MBP and the protein of interest.

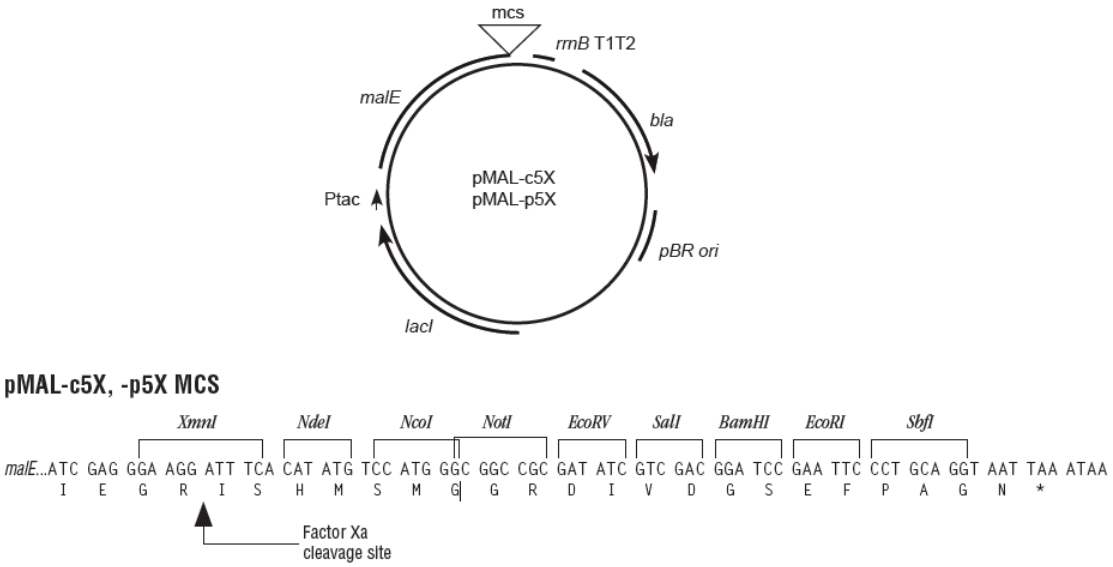


Figure 4.1 Molecular features of pMAL-c5x and –p5x. pMAL-p5x includes the *malE* signal sequence, directing expression of the fusion protein to the periplasm. Arrows indicate direction of transcription. MCS- multiple cloning site. Taken from the pMAL Protein Fusion and Purification System Instruction Manual (New England Biolabs, USA).

Expression as a fusion protein has been shown to improve folding, solubilisation (LaVallie et al., 1993, Samuelsson et al., 1994, Wilkinson et al., 1995) and protein yields (Butt et al., 1989, Koken et al., 1993). In general, proteins smaller than 20 kDa benefit from association with a stable fusion partner due to improved folding or protection from proteolysis (Makrides, 1996, Baneyx, 1999). Since HIV PR is an 11kDa monomer, there is an obvious advantage in expression of this enzyme as a fusion protein. PR has previously been successfully expressed as a fusion protein with, among others, β-galactosidase (Gehringer et al., 2003), glutathione-S-transferase and bacterial periplasmic protein dithiol oxidase (DsbA) (Volonte et al., 2011), as well as MBP (Wan et al., 1995). MBP has been shown to be superior to both glutathione –S- transferase and thioredoxin as a solubilizing partner and is proposed to act as a intramolecular chaperone to aid correct 3 dimensional folding (Kapust and Waugh, 1999).

The pMAL system uses the lac operon and *'tac'* promoter to control expression of the fusion protein. pMAL vectors carry the *lacI* gene, which codes for the Lac repressor. The repressor binds to the *tac* promoter thereby preventing expression of the fusion protein. Upon addition of lactose (or IPTG, its chemical analogue), the expression of the Lac repressor is inhibited, the *tac* promoter is liberated and expression of the fusion protein is initiated.

The pMAL system also allows the freedom to express the protein either in the cytoplasm or periplasm. The optimal expression mode would be dependent on the desired product. It does this by giving the choice of two vectors; pMAL-c5x (cytoplasm) and pMAL-p5x(periplasm), which determine the compartment of expression. In general, cytoplasmic expression results in higher protein yields. However, the expressed protein is more susceptible to degradation as the cytoplasm contains more host proteases than other compartments. An addition challenge is the need to purify the protein of interest from a large pool of intracellular proteins. The pMAL-p5x vector includes a signal peptide which directs the fusion protein through the cytoplasmic membrane into the periplasm. Advantages of periplasmic expression include (1) improved protein folding and disulphide bond formation due to the oxidizing environment (2) the protein of interest is concentrated and purification is simpler as only 4% of total *E.coli* proteins are expressed in this compartment and (3) protein degradation is less extensive in this compartment (Makrides, 1996).

An additional advantage of pMAL expression is ease of purification using maltose-bound affinity chromatography, which binds to the maltose-binding protein. The fusion partners may then be separated via protease specific cleavage between MBP and the protein of interest. A choice of factor Xa, enterokinase or genease I protease cleavage sites is available.

4.2.1.1 Amplification of PR gene

Oligonucleotide primers were designed for the SK245 TOPO clone (described previously in *Chapter 3*). Both primers contained a restriction endonuclease site to facilitate ligation into the pMAL vector. The primers were designed to ensure the PR gene was in the correct reading frame for functional expression. The forward primer overlapped the template by 21 base pairs. The reverse primer overlapped by 20 base pairs. A stop codon was introduced into the reverse primer by mismatch; TGC in the template became TGA in the product. The desired product was a 340 base pair amplicon flanked by Not1 and *EcoR1* endonuclease sites (Figure 4.2). Primer information is as follows:

Primer name	Nucleotide sequence	Melting Temp (°C)	Endonuclease restriction site	Binding region on SK254 TOPO clone
pMAL_fwd	5'-(GC_GGCCGC)GGGAAG AAAGACAGGGAACC -3'	69.5	Not1	1800→1820
pMAL_rev	5'-(G_AATTC)GGATATCT TCAGAATTCGCC -3'	56.5	*EcoR1*	2131←2150

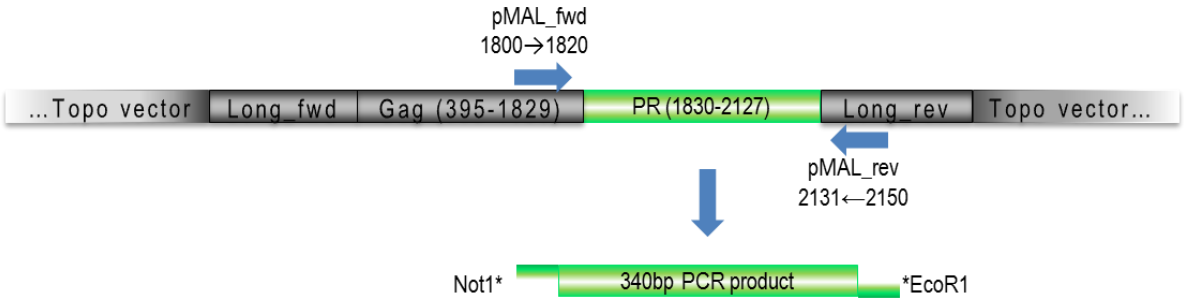


Figure 4.2 Arrangements of primers used to amplify of HIV protease gene from SK254 TOPO clone. Blue arrows indicate binding regions for the pMAL_fwd and pMAL_rev primers, which were designed to bind to the SK254 clone order to amplify the PR gene. Calculated size of the PCR product was 340 base pairs, containing a Not1 site and an EcoR1 site as indicated.

Amplification was carried out with Expand High Fidelity PCR system (Roche Applied Science) from the preciously described SK254 TOPO clone. Reaction components were as follows:

Reaction component	Volume per rxn (ul)	Final Concentration
H2O	17.3	-
10 X Buffer (15 mM $MgCl_2$)	2.5	1.5mM
dNTPs (2.5 mM)	2	0.2mM
Fwd primer (10 µM)	0.4	0.16µM
Rev Primer (10 µM)	0.4	0.16µM
Taq (3.5 U/µl)	0.375	1.3 U
Template DNA (10 ng/µl)	2	20 ng
Total	25	-

The cycling parameters for the reaction were as follows:

Segment	Cycles	Temperature (°C)	Time	Process
1	1	94	2 minutes	Initial Denaturing
2	30	94	15 seconds	Denaturing
		58	30 seconds	Primer Annealing
		72	45 seconds	Extension
3	1	72	7 minutes	Final Extension

Agarose gel electrophoresis was used to separate and visualise the PCR product. A 2% agarose gel was prepared by adding 1.0 g of agarose to 50ml of 1X TBE buffer and heating until the agarose completely dissolved. The solution was poured into a casting tray. Once set, the gel was placed in an electrophoresis tank and covered with 1X TBE buffer. Two µl of DNA Molecular Weight Marker X (Roche diagnostics, Mannheim, Germany,) and 5µl of each sample was mixed with 2 µl of gel loading dye (Sigma-Aldrich, USA) containing 1 in 10,000 dilution of GelRed^{TM} (Biotium, USA) before being loaded onto the gel. The gel was run at 100 V for 30 minutes on an Electrophoresis Power Supply-EPS 301 (Amersham Biosciences, Sweden). The gel was viewed under UV light using the GelVue UV Transilluminator (SynGene, London).

4.2.1.2 Cloning into pCR® TOPO 2.1 vector

This PCR amplicon was ligated into pCR® TOPO 2.1 vector from the TOPO TA Cloning kit (Invitrogen, USA) for maintenance of the PCR product. Ligation was achieved following manufacturer's directions. Briefly, the PCR product was purified from the agarose gel using the Illustra PCR DNA & Gel band purification kit (GE Healthcare, United Kingdom). From this purified product, 10ng was used added to a ligation mixture, along with 10ng vector. The reaction was incubated at room temperature for 30 minutes.

Transformation

Top10 competent cells from the TOPO TA cloning KIT (Invitrogen, USA) were transformed with the ligation mixture, as per manufacturer's instructions. Briefly, 3 µl of ligation reaction was added to one vial of TOP10 competent cells (50µl) and incubated on ice for 30 minutes. The mixture was exposed to 30 seconds of heat shock at 42° C and placed immediately on ice for 2 minutes. Two hundred and fifty µl of S.O.C media

(supplied with kit) was added to each reaction mixture. After 1 hour incubation at 37°C with shaking at 230 rpm, the ligation mixtures were plated out onto pre-warmed agar containing ampicillin (50 µg/ml). The plates had previously been plated with 40µl of X-Gal (50 mg/ml solution) to allow for blue/white selection.

Blue/white selection was used to select clones containing the insert. The site of ligation of the PCR product is found in the coding sequence of the *lacZα* gene of the pCR® TOPO 2.1 vector. This gene expressed the enzyme β-galatosidase. The activity of this enzyme on X-Gal generates blue colonies by metabolism of X-Gal into colourless galactose and 4-chloro-3-brom-indigo, an intense blue precipitate. Successful ligation of an insert into the vector interrupts this gene, preventing production of the blue precipitate, producing white colonies. For this study, 12 white clones were selected at random for screening by PCR and restriction endonuclease digestion.

Screening PCR

Colony PCR was used to screen the 12 clones selected clones. Colonies were first touched to a master plate then added to 10 µl of sterile dH_2O, and boiled for 2 minutes at 95°C. Two µl of this solution was used as the template DNA for the screening PCR. The reaction components and cycling parameters were identical as described above (section 4.2.1.*2 Amplification of PR gene)*. Agarose gel electrophoresis was again used to visualise the PCR amplicon as described above. Colonies containing the desired amplicon were taken forward for further screening via endonuclease restriction digest.

Digestion

The pCR® TOPO 2.1 vector contains *EcoR1* sites flanking the region of the insert. Therefore, *EcoR1* digestion of a correctly ligated vector with an insert would yield two DNA fragments, the PCR product (340 base pairs) and the remnants of the TOPO vector (3.9 kb). The Fermentas GeneJet Plasmid Miniprep Kit (Thermo Fisher Scientific, USA) was used according to manufacturer's instruction to extract DNA from colonies containing the desired plasmid. Resulting plasmid DNA was digested with Fermentas FastDigest *EcoR1* (Thermo Fischer Scientific, USA) for 5 minutes at 37°C. For each sample, the digestion reaction was prepared as follows:

Reaction component	**Volume (µl)**
Plasmid DNA	0.5*
EcoR1 enzyme	0.5
Buffer (10X concentration)	1.0
Sterile H_2O	8.0
Total	10.0

*concentration of samples ranged from 122.4 – 151.5 ng/µl

Agarose gel electrophoresis was used as described in section 4.2.1.2 *(Amplification of PR gene)* to visualise successful digestion.

4.2.1.3 Subcloning into pMAL-p5x and -c5x

The gene of interest (PR) was sub-cloned into the pMAL vectors (-c5x and –p5x).To achieve this, the pMAL vectors and the TOPO vectors containing the PR gene were digested with Fermentas FastDigest enzymes *Not1* and *EcoR1* (Thermo Fischer Scientific, USA). *EcoR1* was added after 25 minutes of *Not1* digestion and the reaction incubated for a further 5 minutes. The reaction components were as follows:

Reaction component	Volume (µl)
Plasmid DNA	2.0
Not1 enzyme	1.0
EcoR1 enzyme	1.0
FastDigest Buffer (10X concentration)	2.0
Sterile H_2O	14.0
Total	20.0

Again, agarose Gel electrophoresis was used to visualise the successful restriction digest as described above (4.2.1.2 *Amplification of PR gene*). The Illustra PCR DNA & Gel band purification kit (GE Healthcare, United Kingdom) was used to purify the relevant DNA bands from the gel. These bands were the PR gene from digestion of the TOPO vector and the digested pMAL vectors. The DNA fragment coding for PR was ligated into the pMAL–c5x and –p5x vectors using Quick T4 DNA Ligase (New England Biolabs, USA). The components of the ligation reaction were as follows:

Reaction component	Volume (µl)
Plasmid DNA(pMAL)	3.4 (40ng)
Insert DNA	2.0 (20ng)
Ligase enzyme	1.0
2 X ligase buffer	10
Sterile H_2O	3.6
Total	20.0

Top10 competent cells from the TOPO TA cloning KIT (Invitrogen, USA) were transformed with the ligation mixture, as described above. X-Gal was not used however as the pMAL vector does not contain the *lacZ* gene. The plates were incubated overnight and the following day 10 colonies were randomly selected. Colony PCR and agarose gel electrophoresis was used as described above to confirm the presence of the desired insert.

The Fermentas GeneJet Plasmid Miniprep Kit (Thermo Fisher Scientific, USA) was used according to manufacturer's instruction to extract DNA from colonies containing the desired plasmid. The resulting plasmid DNA was sequenced as described in Chapter 3, section 3.2.1.4 *Screening of mutants.* Sequencing primer information is listed below:

Primer Name	Primer sequence (5'-3')	pMAL-c5x binding region
NEB#S1237	5'-GGTCGTCAGACTGTCGATGAAGCC-3´	2587→2610
NEB#S1288	5'-TGTCCTACTCAGGAGAGCGTTCAC-3´	2823←2846

Plasmid DNA from pMAL clones containing the insert in the correct reading frame were used to transform *E. coli* ER2523, known as NEB Express (supplied with pMAL kit). These cells are optimised for expression with the pMAL system, however they are not competent. The Fermentas TransformAid ^{TM} Bacterial Transformation Kit was used as per manufacturer's instructions to convert the NEB Express to chemically competent cells. Colony PCR and agarose gel electrophoresis were once again used as described previously to confirm the presence of the plasmid in the chemically transformed NEB Express cells. NEB Express clones containing the desired plasmids were used for expression of HIV PR.

4.2.1.4 Expression of PR

Expression of the PR-MBP fusion protein was carried out following manufacturer's guidelines. 80 ml of rich both + glucose and ampicillin (1% m/v tryptone; 0.5% m/v yeast extract; 0.5% m/v NaCl; 0.2% m/v glucose; 100 µg/ml ampicillin) was inoculated with 0.8 ml of overnight culture of cells containing the fusion plasmid. This culture was incubated at 37°C with shaking at 230 rpm until an optical density (OD) of A_{600} =~0.5 was reached. This OD value is indicative of the culture reaching the log phase of growth, which is optimal for protein expression (Volonte et al., 2011).

IPTG was added to a final concentration of 0.3 mM to induce the expression of the fusion protein. Incubation at 37°C was continued for a further 2 hours. Cells were harvested by centrifugation at 4000 x g for 10 minutes. Cell pellet from –c5x and –p5x constructs were processed differently. The MBP open reading frame of both the pMAL-c5x and –p5x vectors lacking the PR gene insert was also expressed for use as background controls during the enzyme assay, hereafter referred to as the 'no-insert' controls.

For pMAL-c5x constructs, cell pellets were resuspended in 10 ml column buffer (20 mM Tris-HCl; 200 mM NaCl; 1 mM EDTA; pH 7.0) then frozen overnight at -20°C. Solutions were thawed in cold water and exposed to a further 3 freeze-thaw cycles in an effort to lyse cells. Solutions were sonicated in an ice-water bath for 2 minutes, in a series of 15 seconds pulses. Insoluble material was pelleted by centrifugation at 6900 x g for 1 hour. The supernatant; hereafter known as crude extract, was saved on ice.

In order to further purify the PR enzyme, the crude extract was subjected to affinity chromatography. Amylose bound resin (supplied with the pMAL kit) was washed twice in column buffer in a microcentrifuge tube. 50 µl of crude extract was added to 50 µl of the amylose resin slurry, and incubated on ice for 15 minutes. The amylose resin and bound protein was separated from unbound protein fraction via centrifugation for 1 minute at 12000 x g. The supernatant containing the unbound proteins was discarded and the amylose resin was washed with 1 ml of column buffer. The resin was centrifuged and the supernatant removed. The resin was resuspended in 50µl SDS PAGE Laemmli sample buffer (previously described in chapter 3, section *3.2.4.1 Sample preparation*)

Affinity purification was attempted only during the pilot expression, as subsequently the PR enzyme was not found in the bound protein fraction, but rather as a soluble protein in the crude extract. For use in the enzyme assay, the crude extract was used as the enzyme solution without further purification.

For pMAL-p5x constructs, cell pellets were resuspended in 20 ml osmostic shock buffer (30 mM Tris-HCl, 20% m/v sucrose, pH 7.0). EDTA was added to a final concentration of 1 mM and incubated at room temperature for 10 minutes with shaking. The solution was centrifuged at 6900 x g, 4°C, for 30 minutes and the pellet was resuspended in 10 ml of ice cold 5 mM $MgSO_4$. The solution was incubated for 10 minutes in an ice water bath with shaking and then centrifuged again at 6900 x g, 4°C, for 30 minutes. The resulting supernatant, known as the cold osmostic shock fluid, contained the periplasmic fraction and the fusion protein.

Western blotting was used to confirm the presence or absence of PR in the crude extract and the cold osmotic shock fluid. Samples were taken from each step of the expression and purification for both constructs. Western blotting was carried out as described previously in chapter 3, section 3.2.4.2 *Western Blotting*, with the following exceptions. The protein concentration of the osmotic shock fluid from –p5x expression was relatively dilute, hence requiring SDS/KCl precipitation (Li et al., 1989) to increase the concentration to allow for adequate visualisation. This was achieved by adding 10 µl 5% SDS then 10 µl of 3M KCl to 100 µl of protein solution. The solution was then mixed by inversion followed by centrifugation at 12000 x g for 2 minutes. The pellet was resuspended in 10 µl SDS PAGE running buffer (BioRad Laboratories Inc., USA).

10 µl of each sample was added to an equal volume of Laemmli Sample Buffer with β-merceptoethanol (BioRad Laboratories Inc., USA), and prepared as describe earlier, before being loaded onto a 12% gel. After protein separation by gel electrophoresis and transfer to nitrocellulose, membranes were probed with primary mouse monoclonal anti-HIV1 PR. This was followed by secondary rabbit polyclonal anti-mouse IgG conjugated to horse radish peroxidase (HRP). Both antibodies were purchased from Abcam (Cambridge, The United Kingdom) and diluted to 1 in 4000 in Calbiochem® SignalBoost™ Immunoreaction Enhancer solution (Merck, Germany). Incubations, washing and visualisation were as described before.

4.2.2 Enzyme assay

The Protease-Glo™ Assay Kit (Promega, USA) was used for this assay. This kit allows for the detection of protease activity from a wide range of enzyme types, including HIV PR. The assay uses a genetically modified firefly (*Photinus pyralis*) luciferase enzyme encoding the protease recognition site as the protease substrate. When the recognition site is intact, the enzyme has very low background activity. However once the recognition is digested with the relevant protease, the luciferase activity is restored, causing the emission of light (Figure 4.3).

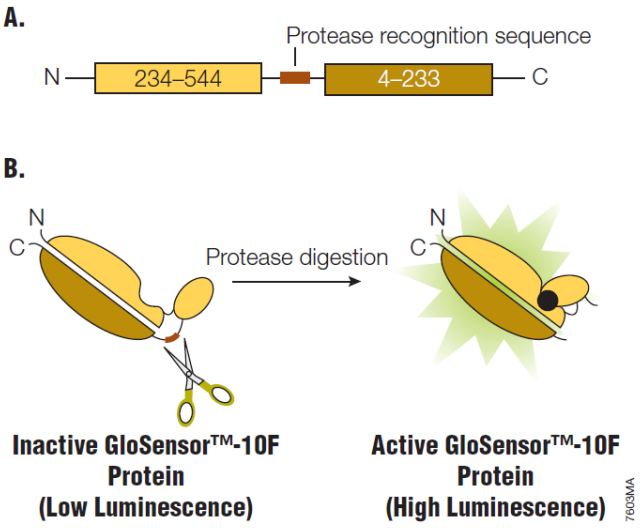


Figure 4.3 Modulation of firefly luciferase with polypeptide linker. (A) A protease recognition sequence of interest is cloned into the pGloSensor™ -10F linear vector **(B)** After translation, the polypeptide linker containing the protease recognition site greatly reduces luciferase activity. Proteolytic cleavage of the recognition site (represented by the scissors) activates the luciferase enzyme, resulting in an increase of luminescence signal in the presence of luciferase substrate. Taken from Protease-Glo™ Assay Technical Manual.

To perform the assay, the vectors encoding the cleavage site of interest are generated. This is done by annealing 2 complementary oligonucleotides encoding the peptide sequence of the cleavage site. The double stranded DNA fragment is then ligated into the pGloSensor™ linear vector (Figure 4.4). A cell free expression system is used to translate the vector into the modified firefly luciferase enzyme, encoding the protease recognition site. This protease recognition site is then digested by the protease of interest, allowing full activity of the luciferase enzyme. The final step of the assay is to measure the luciferase activity with a luminometer. The Protease-Glo™ assay protocol was carried out as per manufacturer's guidelines.

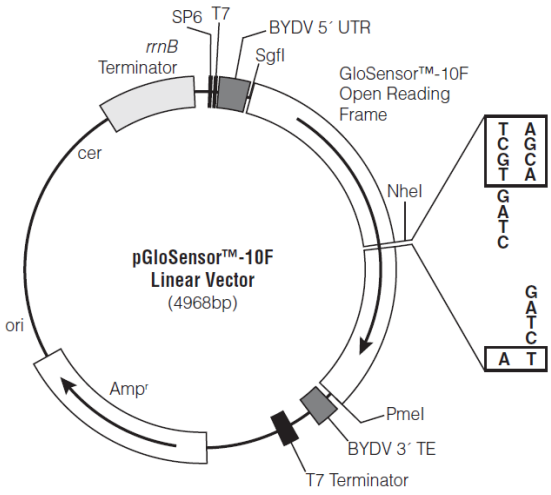


Figure 4.4 pGloSensor™-10F linear vector map and sequence reference points. Oligonucleotides sequences representing selected p2/NC CSs were cloned into the GloSensor™ Open Reading Frame of the linear vector directly after the *NheI* restriction endonuclease site. Taken from the Protease-Glo™ Assay Technical Manual.

4.2.2.1 Oligonucleotide design

Oligonucleotides were designed using the Oligonucleotide Designer tool made available by Promega at their website (www.promega.com/techserv/tools/). Necessary design aspects are listed below:

1. The length of the recognition site was between 4 and 14 amino acids long. Shorter amino acid sequences may increase the chances of the GloSensor™ enzyme being cleaved, thereby reducing activation. A longer amino acid sequence would result in higher baseline luminescence.

2. To ensure correct reading frame and orientation into the linear vector, oligonucleotides must contain a 5' and 3' overhang. The 'A' oligonucleotide must contain a 5' CTAGC and a 3' G overhang flanking the nucleic acid sequence coding for the protease recognition. Oligonucleotide 'B' is complementary to oligonucleotide 'A' at the protease recognition site. Additional sequences required are 5' GATCC and 3' G overhangs. This

creates two flexible amino acid linkers Ala-Ser and Gly-Ser flanking the cleavage site. Annealing of the two oligonucleotide results in the following DNA fragment:

Oligonucleotide A 5´ CTAGC — cleavage site — G 3´

Oligonucleotide B 3´ G — cleavage site — CCTAG 5´

Expression of the oligonucleotides results in the following amino acid sequence:

Ala Ser - (cleavage sequence) - Gly Ser

HPLC purified oligonucleotide primers were ordered from Roche Diagnostics, South Africa. The following protease recognitions sites were designed:

Sequence 17

Gly Ala Ala Gly Ile Met Met Gln Arg Ser Asn

5' CTAGC GGC GCC GCC GGC ATC ATG ATG CAG CGC AGC AAC G 3'

3' G CCG CGG CGG CCG TAG TAC TAC GTC GCG TCG TTG CCTAG 5'

Sequence 16

Gly Ala Ala Gly Ile Met Met Gln Lys Ser Asn

5' CTAGC GGC GCC GCC GGC ATC ATG ATG CAG AAG AGC AAC G 3'

3' G CCG CGG CGG CCG TAG TAC TAC GTC TTC TCG TTG CCTAG 5'

Sequence 241

Ser Gly Ala Ala Ala Ala Ile Met Met Gln Lys Ser Asn

5' CTAGC AGC GGC GCC GCC GCC GCC ATC ATG ATG CAG AAG AGC AAC G 3'

3 'G TCG CCG CGG CGG CGG CGG TAG TAC TAC GTC TTC TCG TTG CCTAG 5'

Sequence 164

Asn Asn Thr Asn Ile Met Met Gln Arg Ser Asn

5' CTAGC AAC AAC ACC AAC ATC ATG ATG CAG CGC AGC AAC G 3'

3' G TTG TTG TGG TTG TAG TAC TAC GTC GCG TCG TTG CCTAG 5'

Sequence 64

Gly Ser Ala Asn Ile Met Met Gln Arg Ser Asn

5' CTAGC GGC AGC GCC AAC ATC ATG ATG CAG CGC AGC AAC G 3'

3' G CCG TCG CGG TTG TAG TAC TAC GTC GCG TCG TTG CCTAG 5'

Sequence 250

Thr Asn Thr Asn Ile Met Met Gln Lys Ser Asn

5' CTAGC ACC AAC ACC AAC ATC ATG ATG CAG AAG AGC AAC G 3'

3' G TGG TTG TGG TTG TAG TAC TAC GTC TTC TCG TTG CCTAG 5'

Sequence WT

Asn Ser Asn Ile Met Met Gln Arg Ser Asn

5' CTAGC AAC AGC AAC ATC ATG ATG CAG CGC AGC AAC G 3'

3' G TTG TCG TTG TAG TAC TAC GTC GCG TCG TTG CCTAG 5'

pNL4-3

Thr Asn Pro Ala Thr Ile Met Ile Gln Lys Gly Asn

5' CTAGC ACC AAC CCA GCC ACC ATC ATG ATC CAG AAG GGC AAC G 3'

3' G TGG TTG GGT CGG TGG TAG TAC TAG GTC TTC CCG TTG CCTAG 5'

4.2.2.2 Cloning Protease recognition sequence inserts into the pGloSensor™-10F Linear vector

Oligonucleotide annealing

Oligonucleotides A and B were resuspended in nuclease-free water to a final concentration of 100 µM. Two µl of each of these solutions was added to 46 µl of Oligo Annealing Buffer (supplied with Protease-Glo™ Assay) to make a total volume of 50µl of the annealing reaction mixture (note: the final concentration of each oligo was 4 µM). The reaction mixture was heated to 90°C for 3 minutes, and then incubated at room temperature for 15 minutes. The annealed oligonucleotides were immediately serially diluted to 10nM and ligated into the linear pGloSensor™ plasmid.

Ligation of insert into pGloSensor™ Plasmid

Generation of the circular pGloSensor™ vector was achieved by ligation of the linear vector with the double stranded annealed oligonucleotide sequences encoding the recognition site of interest. The reaction mixture was incubated at room temperature for 1 hour after assembly, in the following manner:

Component	Standard Reaction
2 X Rapid Ligation Buffer	5µl
pGloSensor™-10F Linear Vector (50ng)	1µl
annealed oligonucleotides A and B (10nM)*	1µl
Nuclease-Free Water	2µl
T4 DNA Ligase	1µl
Total volume	**10µl**

* Negative control reactions did not include annealed oligonucleotides

Transformation of E. coli with pGloSensor™ plasmid

One Shot® TOP10 competent cells (Invitrogen, USA) were transformed with the ligation mixtures. Transformation was carried out as previously described in section 4.2.1.3 *Cloning into pCR® TOPO 2.1 vector*. However, X-Gal was not used as the pGloSensor™ plasmid does not contain the lacZα gene.

Recombinant DNA purification

From each pGloSensor™ [protease site] plasmid generated via the ligation reaction, 5 clones were randomly selected for screening. DNA was isolated from each using the Fermentas GeneJet Plasmid Miniprep Kit (Thermo Fisher Scientific, USA) as per manufacturer's recommendations.

Screening by for Inserts via SgfI/PmeI Digestion

The circular pGloSensor™ vector contains the restriction endonuclease sites *SgfI and PmeI* flanking the open reading frame of the GloSensor™ proteins. These sites have been included to provide a method for detection of an insert. Digestion of a vector containing an insert yields two DNA fragments, a 3.4 kb and a 1.6 kb fragment. Vectors lacking the insert would yield 3 fragments, at 3.4kb, 940 and 700 base pairs. The digestion reaction was performed by adding the follow components together:

Component	Volume (µl)	Final concentration
10 X *SgfI /PmeI* Flexi Enzyme Blend*	2	1 X
Plasmid DNA (~100-250 ng/µl)	1	-
5X Flexi Digest Buffer*	4	1 X
Nuclease-Free water	10	-
Total	**20**	**-**

*Purchased from Promega, USA

The reaction was incubated at 37°C for 2 hours. The resulting DNA fragments were visualised via agarose gel electrophorese as previously described, using a 1% agarose gel.

4.2.2.3 Production of GloSensor™ [protease site] Protein by cell free transcription and translation

The inactive modified firefly luciferase, GloSensor™ Protein, was generated using a cell free expression system, namely, the TNT®SP6 High-Yield Wheat Germ Master Mix Expression System (supplied with Protease-Glo™ Assay Kit). The Wheat Germ Master Mix was removed from -80°C storage and thawed on ice. For each pGloSensor™ [protease site] plasmid, 30 µl of TNT®SP6 High-Yield Wheat Germ Master Mix was added

to 2.5 µg of plasmid DNA, and filled to a final volume of 50 µl with nuclease-free water, as follows:

Plasmid	DNA concentration (ng/µl)	Volume (µl)	Water (µl)	Total (µl)
pGloSensor[seq_17]	457.2	5.5	14.5	20
pGloSensor [seq_16]	590.2	4.0	16	20
pGloSensor [seq_241]	489.0	5.0	15	20
pGloSensor [seq_164]	579.0	4.5	15.5	20
pGloSensor [seq_64]	461.8	5.5	14.5	20
pGloSensor [seq_250]	382.3	6.5	13.5	20
pGloSensor [seq_WT]	572.9	4.5	15.5	20
pGloSensor [seq_NL4-3]	542.5	4.5	15.5	20

For the 'no-DNA' control reactions, plasmid DNA was replaced with nuclease-free water. The reactions were incubated at 25°C for 2 hours. The resulting modified firefly luciferase was use immediately in the Protease digestion reaction.

4.2.2.4 Protease Digestion

Optimal conditions for the PR digestion reaction were determined empirically, using guidelines from the Protease-Glo™ Assay Kit technical manual. The assay buffer supplied with the Protease-Glo™ kit was not used as it was not ideal for HIV PR activity. Instead, an HIV PR specific assay buffer was used. The composition of this buffer was based on the buffer used for the commercially available HIV Protease substrate (Sigma-Aldrich, USA). The buffer used for this assay was a solution of 0.1 M sodium acetate, 0.1 M sodium chloride, 1.0 mM EDTA, 1.0 mM DTT, 10% DMSO pH 5.2. Two features of this buffer were modified to better suit this particular assay. The pH was increased from recommended value of 4.7 to 5.2 and the NaCl concentration was decreased from 1.0 M to 0.1 M. In both instances, the motivation for changing the buffer conditions was the

sensitivity of the firefly luciferase to low pH and high NaCl concentrations. The PR digestion reaction was assembled as follows:

Component	Volume (µl)
HIV assay buffer	15
TNT® Wheat Germ Master Mix plus DNA- containing the GloSensor™ protein	15
HIV PR[b]	5
Total	35

a. Previously expressed in the cell free expression system, described in section *4.2.2.3 Production of GloSensor™ [Protease site] Protein by cell free transcription and translation.* For a negative control, the 'no-DNA' wheat Germ master mix solution was used, which lacked a GloSensor™ protein.

b. Crude extract, of pMAL-c5x expression of HIV-1 PR, as described in section *4.2.1.4 Expression of PR.* For the negative control, the enzyme solution was replaced with nuclease free water to control for background luminescence activity.

The PR digestion reaction mixture was incubated at 30°C for 30 minutes. 37 °C is the optimal temperature for PR. However, a lower temperature was used, as luciferase activity is reduced above temperatures of 30°C. The resulting solution was assayed immediately for luciferase enzyme activity.

Given that PR was not completely purified from the total cell lysate, a 'no insert' control was used to measure background protease activity from the expression host cells. The 'no insert' control was the expression product of the pMAL vector lacking the PR insert.

4.2.2.5 Luminescence detection

For detection of luciferase activity, Bright-Glo™ substrate (supplied with the Protease-Glo™ Assay Kit) was suspended in Bright-Glo™ assay buffer and equilibrated to room temperature. The protease digestion reaction mixture (generated in section 4.2.2.4

Protease Digestion) was diluted 1 in 4 in nuclease-free water. Fifty µl of this solution was added to an equal volume of Bright-Glo™ substrate solution in a white, flat bottom, 96-well plate and incubated at room temperature for 5 minutes. Each sample was measured in duplicate on a Modulus™ Microplate Multimode Reader, in luminometer mode (Turner Biosystems, USA) using a 1-second integration.

4.2.2.6 Enzyme activity analysis

The GloSensor™ luciferase enzyme displayed background luminescence in the absence of PR digestion. It was therefore necessary to calculate the increase in luciferase activity due to the cleavage of the recognition site so as to provide an appropriate measure of PR digestion. Fold activation of the GloSensor™ Protein was calculated as follows:

$$\text{Fold activation} = \frac{(\text{luminescence from tube A}) - (\text{luminescence from tube C})}{(\text{luminescence from tube B}) - (\text{luminescence from tube D})}$$

Where each tube A-D contained the following reaction components:

Tube	Plus- DNA Wheat Germ Master mix	No-DNA Wheat Germ Master Mix	Protease
A	+	-	+
B	+	-	-
C	-	+	+
D	-	+	-

Therefore, fold activation is a measure of luminescence increase due to specific protease digestion of the protease recognition site.

4.3 RESULTS

4.3.1. Expression of HIV Protease

The pMAL protocol allows for the rapid one step purification via affinity chromatography. However, immune blotting revealed that PR was not in the fraction bound to the amylose column, but rather in the soluble fraction, as an 11 kDa subunit (Figure 4.5). The most probably explanation for this is that the PR was able to catalyse its own cleavage from the fusion protein. Other studies have shown this same process of PR auto-maturation (Meek et al., 1989, Krausslicht et al., 1989, Wan et al., 1995, Volonte et al., 2011). This is similar to the behaviour of PR during expression of *pol* - the enzyme dimerises and cleaves itself out of the polyprotein (Wan et al., 1996).

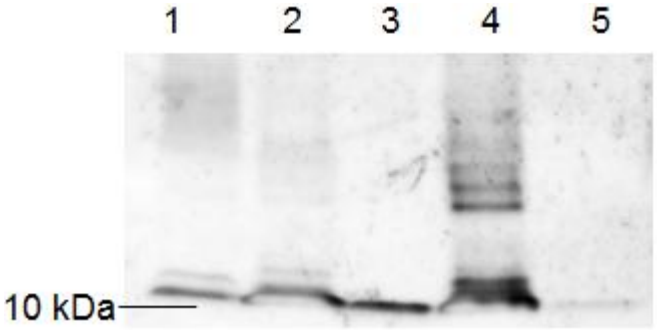


Figure 4.5 Immunoblot analysis of pMAL-c5x expression. Samples were taken from each stage of expression and purification. Lane 1- uninduced cells; lane 2- induced cells; lane 3- crude extract, lane 4-insoluble matter; lane 5: protein bound to amylose resin. Protein was separated via a 12% SDS PAGE gel, transferred to nitrocellulose membrane and probed with mouse monoclonal anti-HIV1 Protease antibody at a 1 in 4000 dilution.

4.3.2 Enzyme assay

The PR enzyme used in the substrate assay was present as a fraction of the crude extract, which contained numerous other *E. coli* proteins. Therefore, contaminating *E. coli* enzymes would almost certainly interfere with the assay. In an attempt to monitor the effect of these enzymes, a ‘no- insert’ control solution was used to evaluate the level of

background enzyme activity from the crude extract. This ‘no-insert’ control solution was prepared as described above (section *4.2.1.4 Expression of PR).* Briefly, this solution was derived from expression of the pMAL-c5x vector lacking the PR gene insert.

Both the -c5x and –p5x PR extracts were assayed to determine which solution would provide the optimal enzyme performance. While the –c5x had a higher yield of PR than –p5x, the potential disadvantage of this solution was the presence of more contaminating enzymes, given that the PR solution was essentially whole cell lysate. However, after testing both cytoplasmic and periplasmic expression products, the crude extract from –c5x expression showed much higher activity than that of –p5x, which was essentially devoid of enzyme activity. Additionally, the background activity from contaminating enzymes of the –c5x expression was roughly equivalent to negative control reactions lacking enzyme (‘no-enzyme’ control). Therefore, the PR expressed in the cytoplasmic (–c5x) was chosen for the remaining assays.

Once assay conditions had been optimised, PR activity was evaluated against the range of 8 substrates. A substantial increase of luminescence was found in reactions containing PR over the controls for ‘no-enzyme’ background control, while there was no significant difference between ‘no-insert’ controls over the ‘no-enzyme’ controls ($p=0.7422$). This data supports the notion that contaminating *E. coli* enzymes were not responsible for the increase in luminescence observed during experiments with PR. For reactions containing PR, substantial differences were observed between the 8 substrate types. Increase in luminescence ranged from 0.5 fold for sequence 250 up to a maximum of 37.5 fold for sequence 64 (Figure 4.6). Sequences 241, 64 and 164 all displayed a substantial increase in luminescence due to PR digestion, while sequence 16 and WT both displayed limited increase in luminescence. Sequence 250 displayed essentially no luminescence

above background levels. Fold activation for substrates 241, 64 and 164 were all statistically significantly different from the fold activation of substrates 16, 250, wt and nl4-3 (Table 4.1). However, while these results may be statistically significant, the fold activation required to be considered significant is unknown.

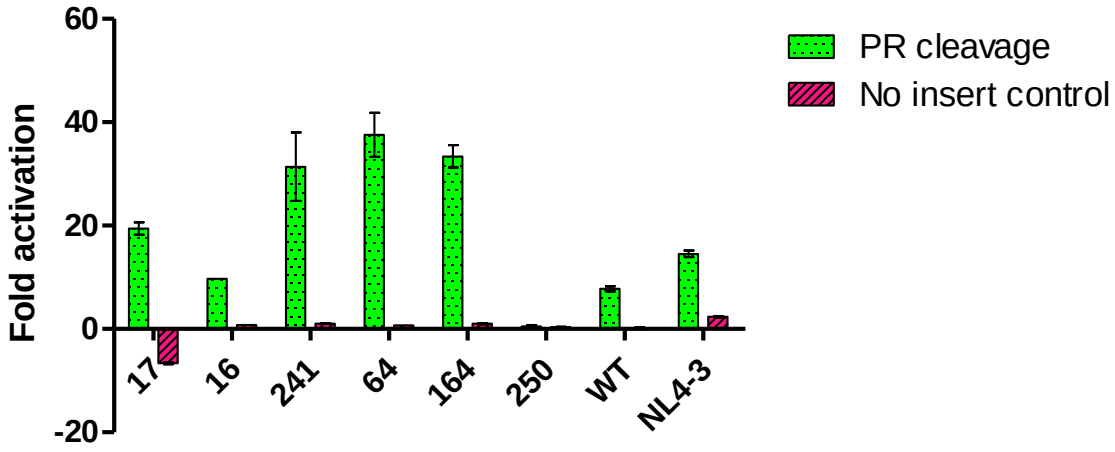


Figure 4.6 Effect of PR activity on GloSensor™ Protein substrates. GloSensor™ [protease recognition site] proteins were incubated for 30 minutes with recombinantly expressed HIV-1 PR. Luminescence was measured after addition of luminol substrate. Fold activation was calculated as increase in luminescence over background controls.

Table 4.1 Statistical differences in fold difference between PR substrates

	17	16	241	64	164	250	wt	nl4-3
17	-							
16	ns	-						
241	ns	***	-					
64	**	***	ns	-				
164	*	***	ns	ns	-			
250	**	ns	***	***	***	-		
wt	ns	ns	***	***	***	ns	-	
nl4-3	ns	ns	**	***	**	*	ns	-

Significance was determined using Tukey's Multiple Comparison Test in GraphPad Prism.
***= ($p<0.001$); ** = ($0.001<p<0.01$); * = ($0.01<p<0.05$); ns = not significant, ($p > 0.05$).

4.4 DISCUSSION

The results of this study revealed no correlation between levels of PR activity and amino acid sequence of the protease recognition sequence. For example, sequences 17 and 16 differed by only 1 amino acid, yet displayed considerable differences in proteolysis. Also, sequences 164 and 250 differed in only 2 positions yet had substantially different enzyme activities; for sequence 164, a 33.5 fold increase in luminescence and while only a 0.5 fold increase for sequence 250. These data imply a sequence independent mechanism for HIV-1 PR.

While many studies have proposed possible determinants of Gag cleavage, most agree that 3 dimensional structural plays a significant role in the regulation of proteolysis. The authors of a 2002 study (Prabu-Jeyabalan et al., 2002) suggested that specificity of HIV-1 PR was determined by the 3 dimensional asymmetric shape conserved across all CSs, rather than a particular amino acid sequence. Subsequently this view has been has been supported by several other studies and is referred to as the 'substrate envelope hypothesis' (Nalam et al., 2010, Özen et al., 2011, Özen et al., 2012). In addition, Ozen et al (2011) suggested that the variation between cleavage sites was responsible for the regulation of the sequential processing of Gag due to slight differences in 3D substrate envelope structure between different CSs. The results of this study, however, do not fully support this hypothesis, given that similar amino acid sequence (ie sequences 17 and 16), did not have experience similar levels of proteolysis.

Alternatively, the model proposed by Perez and colleagues (2010) could be supported by the results of this study. As previously described (chapter 2), these authors suggested that cleavage does not depend solely on the amino acid sequence, but rather on the 3 dimensional protein folding of the Gag precursor. Within this context, the enzyme is only

able to cleave sites which are exposed for a sufficient length of time and are geometrically accessible to the active site of PR. More specifically, non-substrates are not cleaved because they are physically buried in the core of the polyprotein precursor and hence are not accessible to the enzyme. This implies that PR activity on the natural substrate cannot be estimated by assays using either shortened peptides fragments or amino acids sequences representing the CSs in any context other than the Gag precursor (such as the Protease-Glo™ Assay), as 3D folding of Gag is not accounted for. In the context of the work done here, this may explain why the enzyme assay results show no correlation with amino acid sequence. In agreement with this view, a 2012 study (Lee et al.) found that proteolytic processing of certain CSs, including the p2/NC site, is affected by the surrounding context. The authors maintained that long range features of cleavage regulation cannot be addressed with peptide substrates.

In summary, the enzyme assays described in this chapter indicate that amino acid sequence of the CSs is not the principal determinant of proteolysis. The results support the hypothesis that cleavage of the Gag precursor by PR is dependent instead on the contextual setting of the cleavage site.

CHAPTER 5

General discussion and conclusions

Viral fitness is a combination of several factors, including replication capacity and potential to evade immune and drug pressure. High levels of natural variation create more opportunities for viral evolution. Therefore a particular HIV strain may have reduced replication capacity *in vitro* but has greater evolutionary potential, hence greater viral fitness at the population level. This could be one explanation for the global dominance of subtype C. Therefore, the aim of this study was to assess the impact of Gag p2/NC CS polymorphisms on viral fitness. Several studies have shown that variation at Gag CSs can influence viral fitness and may be associated with the response to PI therapy. Therefore, the variation observed at this CS may play a role in certain subtype C characteristics.

For this study, a computational approach was used to predict the effect of p2/NC CS variation on replication capacity. The hypothesis being that higher binding affinity between PR and Gag would lead to increased proteolysis and subsequently higher replication capacity. However, no correlation was observed between results from the docking procedure and the RC assay. Modest differences in RC were observed between the viral variants, suggesting that RC is not significantly influenced by Gag cleavage rate. Alternatively, the results may imply that cleavage of Gag by PR is independent of amino acid sequence, and any observable differences in RC may be due to altered protein function such as delayed PR:Gag dissociation or reduced RNA packaging by NC (Goodenow et al., 2002). In order to determine the basis of PR cleavage, enzyme assays were performed. The results of this assays suggested that PR activity was not controlled by amino acid sequence of the cleavage site.

Taken together, the results of this study are best explained by the model proposed by Perez and co-workers (2010). According to this model, differences in amino acid sequence would have only a slight impact on RC, as CS variation would not greatly affect the overall tertiary folding of Gag. This model is also supported by the results of the enzyme assay, which implied a sequence independent mechanism of HIV-1 PR cleavage.

In summary, the work performed during this study found that p2/NC cleavage site variation does not have a significant impact on subtype C viral fitness. A possible explanation is that cleavage of Gag by PR is determined by 3 dimensional protein folding of Gag, which determines accessibility of the cleavage site to the PR, hence cleavage.

Future investigations into the tertiary structure of the Gag polyprotein via computational studies and crystallography may clarify the effect of cleavage site variation and provide further insights in the complex nature of PR:Gag interactions. Additionally, RC and enzyme assays could be performed with a larger selection of variants, including drug resistant virus sequences, in order to determine if the results observed in this study are applicable for a wider range of variants.

Appendix A

For HIV PR, a cleavage site is considered to be 5 amino acids on either side of the exact position of cleavage. In the nomenclature of Schechter and Berger (1967) the immediate amino acids on either side are known as P1 and P1' (Figure 1.12).

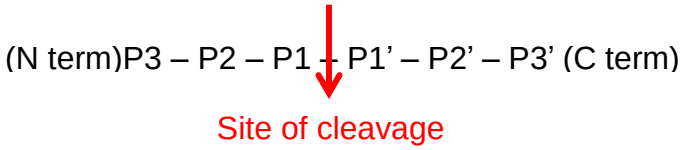


Figure A. Cleavage site position terms of nomenclature of Schechter and Berger (1967). P3 to P3' represent amino acid residues. The cleavage site is between residues P1 and P1'.

Appendix B. Site directed mutagenesis primers

Primer name	Primer Sequence (5' To 3')	Length	Melting Temp (°C)	HXB2 position
Primer 17-1				
ins_3nt_after_99	5'-ctgaggcaatgagccaagcaggtaacagtaacataatgataca-3'	43	78.53	1884→1926
ins_3nt_after_99_antisense	5'-tgtatcattatgttactgttacctgcttggctcattgcctcag-3'	43	78.53	1884←1926
Primer 17-2				
a103g_a104c_c105t	5'-gctgaggcaatgagccaagcaggtgctagtaacataatgatacagag-3'	47	78.2	1883→1929
a103g_a104c_c105t_antisense	5'-ctctgtatcattatgttactagcacctgcttggctcattgcctcagc-3'	47	78.2	1883←1929
Primer 17-3				
a106g_g107c_t108a	5'-gcaatgagccaagcaggtgctgcaaacataatgatacagagaagc-3'	45	78.06	1889→1933
a106g_g107c_t108a_antisense	5'-gcttctctgtatcattatgtttgcagcacctgcttggctcattgc-3'	45	78.06	1889←1933
Primer 17-4				
a109g_a110g	5'-gagccaagcaggtgctgcaggcataatgatacagagaagc-3'	46	79.1	1894→1933
a109g_a110g_antisense	5'-gcttctctgtatcattatgcctgcagcacctgcttggctc-3'	46	79.1	1894←1933
Primer 16-5				
g125a	5'-cgctgcaggcataatgatacagaaaagcaattttaaaggatctaaaa-3'	47	78.97	1906→1952
g125a_antisense	5'-ttttagatcctttaaaattgcttttctgtatcattatgcctgcagcg-3'	47	78.97	1906←1952
Primer 241-1				
ins_3nt_after_99	5'-ctgaggcaatgagccaagcagcaaacagtaacataatgataca-3'	43	78.53	1884→1926
ins_3nt_after_99_antisense	5'-tgtatcattatgttactgtttgctgcttggctcattgcctcag-3'	43	78.53	1884←1926
Primer 241-2				
ins_6nt_after_99	5'-ctgaggcaatgagccaagcatcaggtgcaaacagtaacataa-3'	42	78.19	1878→1919
ins_6nt_after_99_antisense	5'-ttatgttactgtttgcacctgatgcttggctcattgcctcag-3'	42	78.19	1878→1919
Primer 241-3				
a109g_a110c_c111a	5'-ggcaatgagccaagcatcaggtgcagcaagtaacataatgatacagaga-3'	49	78.34	1882→1930
a109g_a110c_c111a_antisense	5'-tctctgtatcattatgttacttgctgcacctgatgcttggctcattgcc-3'	49	78.34	1882←1930
Primer 241-4				
a112g_g113c_t114a	5'-agccaagcatcaggtgcagcagcaaacataatgatacagagaagc-3'	45	78.06	1889→1933
a112g_g113c_t114a_antisense	5'-gcttctctgtatcattatgtttgctgctgcacctgatgcttggct-3'	45	78.06	1889←1933
Primer 241-5				
a115g_a116c	5'-agcatcaggtgcagcagcagccataatgatacagagaagc-3'	40	78.8	1894→1933

a115g_a116c_antisense	5'-gcttctctgtatcattatggctgctgctgcacctgatgct-3'	40	78.8	1894←1933
Primer 241-6				
g131a	5'-agcagcagccataatgatacagaaaagcaatttaaaggatctaaaa-3'	47	78.1	1906→1952
g131a_antisense	5'-ttttagatcctttaaaattgcttttctgtatcattatggctgctgct-3'	47	78.1	1906←1952
Primer 64-1				
ins_3nt_after_99	5'-ctgaggcaatgagccaagcaggcaacagtaacataatgataca-3'	43	78.53	1884→1926
ins_3nt_after_99_antisense	5'-tgtatcattatgttactgttgcctgcttggctcattgcctcag-3'	43	78.53	1884←1926
Primer 64-2				
a106g_g107c_t108a	5'-gcaatgagccaagcaggcaacgcaaacataatgatacagagaagc-3'	45	78.06	1889→1933
a106g_g107c_t108a_antisense	5'-gcttctctgtatcattatgtttgcgttgcctgcttggctcattgc-3'	45	78.06	1889←1933
Primer 64-3				
a104g_c105t	5'-gaggcaatgagccaagcaggcagtgcaaacataatgatacag-3'	42	79.21	1886→1927
a104g_c105t_antisense	5'-ctgtatcattatgtttgcactgcctgcttggctcattgcctc-3'	42	79.21	1886←1927
Primer 164-1				
ins_3nt_after_99	5'-ctgaggcaatgagccaagcaaataacagtaacataatgataca-3'	43	78.53	1884→1926
ins_3nt_after_99_antisense	5'-tgtatcattatgttactgttatttgcttggctcattgcctcag-3'	43	78.53	1884←1926
Primer 164-2				
g107c_t108a	5'-ggcaatgagccaagcaaataacacaaacataatgatacagagaagcaa-3'	48	78.65	1888→1935
g107c_t108a_antisense	5'-ttgcttctctgtatcattatgtttgtgttatttgcttggctcattgcc-3'	48	78.65	1888←1935
Primer 250-1				
ins_3nt_after_99	5'-ctgaggcaatgagccaagcaaccaacagtaacataatgataca-3'	43	78.53	1884→1926
ins_3nt_after_99_antisense	5'-tgtatcattatgttactgttggttgcttggctcattgcctcag-3'	43	78.53	1884←1926
Primer 250-2				
g107c_t108a	5'-caatgagccaagcaaccaacacaaacataatgatacagagaagc-3'	44	78.39	1890→1933
g107c_t108a_antisense	5'-gcttctctgtatcattatgtttgtgttggttgcttggctcattg-3'	44	78.39	1890←1933
Primer 250-3				
g125a	5'-caaccaacacaaacataatgatacagaaaagcaattttaaaggatctaaaagaat-3'	55	78.59	1902→1956
g125a_antisense	5'-attcttttagatcctttaaaattgcttttctgtatcattatgtttgtgttggttg-3'	55	78.59	1902←1956

REFERENCES

ABRAHA, A., NANKYA, I. L., GIBSON, R., DEMERS, K., TEBIT, D. M., JOHNSTON, E., KATZENSTEIN, D., SIDDIQUI, A., HERRERA, C., FISCHETTI, L., SHATTOCK, R. J. & ARTS, E. J. 2009. CCR5- and CXCR4-tropic subtype C human immunodeficiency virus type 1 isolates have a lower level of pathogenic fitness than other dominant group M subtypes: implications for the epidemic. *J. Virol.,* 83**,** 5592-605.

ACACLONE SOFTWARE pDRAW32. http://www.acaclone.com.

ALEXANDRE, K. B., GRAY, E. S., PANTOPHLET, R., MOORE, P. L., MCMAHON, J. B., CHAKAUYA, E., O'KEEFE, B. R., CHIKWAMBA, R. & MORRIS, L. 2011. Binding of the Mannose-Specific Lectin, Griffithsin, to HIV-1 gp120 Exposes the CD4-Binding Site. *J. Virol.,* 85**,** 9039-9050.

ALTMAN, M. D., NALIVAIKA, E. A., PRABU-JEYABALAN, M., SCHIFFER, C. A. & TIDOR, B. 2008. Computational design and experimental study of tighter binding peptides to an inactivated mutant of HIV-1 protease. *Proteins,* 70**,** 678-694.

ARNOLD, K., BORDOLI, L., KOPP, J. & SCHWEDE, T. 2006. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics,* 22**,** 195-201.

BALL, S. C., ABRAHA, A., COLLINS, K. R., MAROZSAN, A. J., BAIRD, H., QUINONES-MATEU, M. E., PENN-NICHOLSON, A., MURRAY, M., RICHARD, N., LOBRITZ, M., ZIMMERMAN, P. A., KAWAMURA, T., BLAUVELT, A. & ARTS, E. J. 2003. Comparing the ex vivo fitness of CCR5-tropic human immunodeficiency virus type 1 isolates of subtypes B and C. *J. Virol.,* 77**,** 1021–1038.

BALLY, F., MARTINEZ, R., PETERS, S., SUDRE, P. & TELENTI, A. 2000. Polymorphism of HIV type 1 Gag p7/p1 and p1/p6 cleavage sites: Clinical significance and implications for resistance to protease inhibitors. *AIDS Res. Hum. Retroviruses,* 16**,** 1209-1213.

BANDARANAYAKE, R. M., KOLLI, M., KING, N. M., NALIVAIKA, E. A., HEROUX, A., KAKIZAWA, J., SUGIURA, W. & SCHIFFER, C. A. 2010. The Effect of Clade-Specific Sequence Polymorphisms on HIV-1 Protease Activity and Inhibitor Resistance Pathways. *J. Virol.,* 84**,** 9995-10003.

BANEYX, F. 1999. Recombinant protein expression in Escherichia coli. *Curr. Opin. Biotechnol.,* 10**,** 411-421.

BARRE-SINOUSSI, F., CHERMANN, J. C., REY, F., NUGEYRE, M. T., CHAMARET, S., GRUEST, J., DAUGUET, C., C. AXLER-BLIN, VEZINET-BRUN, F., ROUZIOUX, C.,

ROSENBAUM, W., MONTAGNIER., L. & MONTAGNIER, L. 1983. Isolation of a T-lymphocyte retrovirus from a patient at risk for acquired immunodeficiency syndrome (AIDS). *Science,* 220**,** 868–871.

BARRIE, K. A., PEREZ, E. E., LAMERS, S. L., FARMERIE, W. G., DUNN, B. M., SLEASMAN, J. W. & GOODENOW, M. M. 1996. Natural Variation in HIV-1 Protease, Gag p7 and p6, and Protease Cleavage Sites within Gag/Pol Polyproteins: Amino Acid Substitutions in the Absence of Protease Inhibitors in Mothers and Children Infected by Human Immunodeficiency Virus Type 1. *Virology,* 219**,** 407-416.

BESSONG, P. O. 2008. Polymorphisms in HIV-I subtype C proteases and the potential impact on protease inhibitors. *Trop. Med. Int. Health,* 13**,** 144-151.

BRADFORD, M. M. 1976. Rapid and sensitive method for quantitation of microgram quantities of protein utilizing principle of protein-dye binding. *Anal. Biochem.,* 72**,** 248-254.

BRIK, A. & WONG, C.-H. 2002. HIV-1 protease: mechanism and drug discovery. *Org. Biomol. Chem.,* 1.

BROCKMAN, M. A., SCHNEIDEWIND, A., LAHAIE, M., SCHMIDT, A., MIURA, T., DESOUZA, I., RYVKIN, F., DERDEYN, C. A., ALLEN, S., HUNTER, E., MULENGA, J., GOEPFERT, P. A., WALKER, B. D. & ALLEN, T. M. 2007. Escape and Compensation from Early HLA-B57-Mediated Cytotoxic T-Lymphocyte Pressure on Human Immunodeficiency Virus Type 1 Gag Alter Capsid Interactions with Cyclophilin A. *J. Virol.,* 81**,** 12608-12618.

BROCKMAN, M. A., TANZI, G. O., WALKER, B. D. & ALLEN, T. M. 2006. Use of a novel GFP reporter cell line to examine replication capacity of CXCR4- and CCR5-tropic HIV-1 by flow cytometry. *J. Virol. Methods,* 131**,** 134-142.

BUTT, T. R., JONNALAGADDA, S., MONIA, B. P., STERNBERG, E. J., MARSH, J. A., STADEL, J. M., ECKER, D. J. & CROOKE, S. T. 1989. Ubiquitin fusion augments the yield of cloned gene products in *Escherichia coli Proc. Natl. Acad. Sci. USA* 86**,** 2540–2544.

CAMPBELL, T. B., SCHNEIDER, K., WRIN, T., PETROPOULOS, C. J. & CONNICK, E. 2003. Relationship between in vitro human immunodeficiency virus type 1 replication rate and virus load in plasma. *J. Virol. ,* 77 12105–12112.

CASE, K. 1986. Nomenclature: Human Immunodeficiency Virus. *Ann. Intern. Med.,* 105**,** 133-133.

CENTRE FOR DISEASE CONTROL 1981. Kaposi's sarcoma and Pneumocystis pneumonia among homosexual men--New York City and California. *MMWR,* 30**,** 305-8.

CHARURAT, M., NASIDI, A., DELANEY, K., SAIDU, A., CROXTON, T., MONDAL, P., ALIYU, G. G., CONSTANTINE, N., ABIMIKU, A. L., CARR, J. K., VERTEFEUILLE, J. & BLATTNER,

W. 2012. Characterization of Acute HIV-1 Infection in High-Risk Nigerian Populations. *J. Infect. Dis.,* 205**,** 1239-1247.

CLAVEL, F. & MAMMANO, F. 2010. Role of Gag in HIV Resistance to Protease Inhibitors. *Viruses,* 2**,** 1411-1426.

CLEMENTS, J. & ZINK, M. 1996. Molecular biology and pathogenesis of animal lentivirus infections. *Clin. Microbiol. Rev.,* 9**,** 100-117.

COLINGE, J. & BENNETT, K. L. 2007. Introduction to computational proteomics. *PLoS Comput. Biol.,* 3.

COMAN, R. M., ROBBINS, A. H., FERNANDEZ, M. A., GILLILAND, C. T., SOCHET, A. A., GOODENOW, M. M., MCKENNA, R. & DUNN, B. M. 2008. The contribution of naturally occurring polymorphisms in altering the biochemical and structural characteristics of HIV-1 subtype C protease. *Biochemistry,* 47**,** 731-743.

COTE, H. C. F., BRUMME, Z. L. & HARRIGAN, P. R. 2001. Human immunodeficiency virus type 1 protease cleavage site mutations associated with protease inhibitor cross-resistance selected by indinavir, ritonavir, and/or saquinavir. *J. Virol.,* 75**,** 589-594.

DAM, E., QUERCIA, R., GLASS, B., DESCAMPS, D., LAUNAY, O., DUVAL, X., KRAUSSLICH, H. G., HANCE, A. J. & CLAVEL, F. 2009. Gag Mutations Strongly Contribute to HIV-1 Resistance to Protease Inhibitors in Highly Drug-Experienced Patients besides Compensating for Fitness Loss. *PLoS Pathog.,* 5(3): e1000345. doi:10.1371/journal.ppat.1000345.

DAS, A., MAHALE, S., PRASHAR, V., BIHANI, S., FERRER, J. L. & HOSUR, M. V. 2010. X-ray Snapshot of HIV-1 Protease in Action: Observation of Tetrahedral Intermediate and Short Ionic Hydrogen Bond SIHB with Catalytic Aspartate. *J. Am. Chem. Soc.,* 132**,** 6366-6373.

DE OLIVEIRA, T., ENGELBRECHT, S., JANSE VAN RENSBURG, E., GORDON, M., BISHOP, K., ZUR MEGEDE, J., BARNETT, S. W. & CASSOL, S. 2003. Variability at Human Immunodeficiency Virus Type 1 Subtype C Protease Cleavage Sites: an Indication of Viral Fitness? *J. Virol.,* 77**,** 9422-9430.

DOYON, L., CROTEAU, G., THIBEAULT, D., POULIN, F., PILOTE, L. & LAMARRE, D. 1996. Second locus involved in human immunodeficiency virus type 1 resistance to protease inhibitors. *J. Virol.,* 70.

DYKES, C. & DEMETER, L. M. 2007. Clinical significance of human immunodeficiency virus type 1 replication fitness. *Clin. Microbiol. Rev.,* 20**,** 550-578.

ESBJORNSSON, J., MILD, M., MA°NSSON, F., NORRGREN, H. & MEDSTRAND, P. 2011. HIV-1 Molecular Epidemiology in Guinea-Bissau, West Africa: Origin, Demography and Migrations. *PLoS ONE,* 6**,** e17025.

EWING, T. J., MAKINO, S., SKILLMAN, A. G. & KUNTZ, I. D. 2001. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J. Comput. Aided Mol. Des.,* 15**,** 411-428.

FEHER, A., WEBER, I. T., BAGOSSI, P., BOROSS, P., MAHALINGAM, B., LOUIS, J. M., COPELAND, T. D., TORSHIN, I. Y., HARRISON, R. W. & TOZSER, J. 2002. Effect of sequence polymorphism and drug resistance on two HIV-1 Gag processing sites. *Eur. J. Biochem.,* 269**,** 4114-4120.

GALLO, R. C., SALAHUDDIN, S. Z., POPOVIC, M., SHEARER, G. M., KAPLAN, M., HAYNES, B. F., PALKER, T. J., REDFIELD, R., OLESKE, J., SAFAI, B., WHITE, G., FOSTER, P. & MARKHAM, P. D. 1984. Frequent detection and isolation of cytopathic retroviruses (HTLV-III) from patients with AIDS and at risk for AIDS. *Science,* 224**,** 500-503.

GANSER-PORNILLOS, B. K., YEAGER, M. & SUNDQUIST, W. I. 2008. The structural biology of HIV assembly. *Curr. Opin. HIV. AIDS. ,* 18**,** 203-217.

GEHLHAAR, D. K., VERKHIVKER, G., REJTO, P. A., FOGEL, D. B., FOGEL, L. J. & FREER, S. T. 1995. Docking Conformationally Flexible Small Molecules into a Protein Binding Site through Evolutionary Programming. *Proceedings of the Fourth International Conference on Evolutionary Programming***,** 615-627.

GEHRINGER, H., VON DER HELM, K., SEELMEIR, S., WEIßBRICH, B., EBERLE, J. & NITSCHKO, H. 2003. Development and evaluation of a phenotypic assay monitoring resistance formation to protease inhibitors in HIV-1-infected patients. *J. Virol. Methods,* 109**,** 143-152.

GENONI, A., MORRA, G., MERZ, K. M. & COLOMBO, G. 2010. Computational Study of the Resistance Shown by the Subtype B/HIV-1 Protease to Currently Known Inhibitors. *Biochemistry,* 49**,** 4283-4295.

GERVAIX, A., WEST, D., LEONI, L. M., RICHMAN, D. D., WONG-STAAL, F. & CORBEIL, J. 1997. A new reporter cell line to monitor HIV infection and drug susceptibility in vitro. *Proc. Natl. Acad. Sci. U. S. A.,* 94**,** 4653-4658.

GOEDERT, J. J. & GALLO, R. C. 1985. Epidemiological Evidence That HTLV-III Is The AIDS Agent. *Eur. J. Epidemiol.,* 1**,** 155-159.

GOODENOW, M. M., BLOOM, G., ROSE, S. L., POMEROY, S. M., O'BRIEN, P. O., PEREZ, E. E., SLEASMAN, J. W. & DUNN, B. M. 2002. Naturally Occurring Amino Acid Polymorphisms

in Human Immunodeficiency Virus Type 1 (HIV-1) Gag p7NC and the C-Cleavage Site Impact Gag-Pol Processing by HIV-1 Protease. *Virology,* 292**,** 137-149.

GORDON, M., DE OLIVEIRA, T., BISHOP, K., COOVADIA, H. M., MADURAI, L., ENGELBRECHT, S., VAN RENSBURG, E. J., MOSAM, A., SMITH, A. & CASSOL, S. 2003. Molecular characteristics of human immunodeficiency virus type 1 subtype C viruses from KwaZulu-Natal, South Africa: Implications for vaccine and antiretroviral control strategies. *J. Virol.,* 77**,** 2587-2599.

HABU, Y., MIYANO-KUROSAKI, N., KITANO, M., ENDO, Y., YUKITA, M., OHIRA, S., TAKAKU, H., NASHIMOTO, M. & TAKAKU, H. 2005. Inhibition of HIV-1 gene expression by retroviral vector-mediated small-guide RNAs that direct specific RNA cleavage by tRNase ZL *Nucleic Acids Res.,* 33**,** 235–243.

HEMELAAR, J. 2011. The origin and diversity of the HIV-1 pandemic. *Trends Mol. Med.,* 18**,** 182-192.

HEMELAAR, J., GOUWS, E., GHYS, P. D. & OSMANOV, S. 2006. Global and regional distribution of HIV-1 genetic subtypes and recombinants in 2004. *AIDS,* 20**,** W13-W23.

HEMELAAR, J., GOUWS, E., GHYS, P. D., OSMANOV, S. & CHARACTERISATION, W.-U. N. F. H. I. A. 2011. Global trends in molecular epidemiology of HIV-1 during 2000–2007. *AIDS,* 25**,** 679–689.

HO, S. K., COMAN, R. M., BUNGER, J. C., ROSE, S. L., O'BRIEN, P., MUNOZ, I., DUNN, B. M., SLEASMAN, J. W. & GOODENOW, M. M. 2008. Drug-associated changes in amino acid residues in Gag p2, $p7^{NC}$, and $p6^{Gag}$/$p6^{Po}$l in human immunodeficiency virus type 1 (HIV-1) display a dominant effect on replicative fitness and drug response. *Virology,* 378**,** 272-281.

HO, S. K., PEREZ, E. E., ROSE, S. L., COMAN, R. M., LOWE, A. C., HOU, W., MA, C., LAWRENCE, R. M., DUNN, B. M., SLEASMAN, J. W. & GOODENOW, M. M. 2009. Genetic determinants in HIV-1 Gag and Env V3 are related to viral response to combination antiretroviral therapy with a protease inhibitor. *AIDS*.

HOLGUIN, A., ALVAREZ, A. & SORIANO, V. 2005. Differences in the length of Gag proteins among different HIV type 1 subtypes. *AIDS Res. Hum. Retroviruses,* 21**,** 886-893.

HOLGUÍN, A., SUÑE, C., HAMY, F., SORIANO, V. & KLIMKAIT, T. 2006. Natural polymorphisms in the protease gene modulate the replicative capacity of non-B HIV-1 variants in the absence of drug pressure. *J. Clin. Virol.,* 36**,** 264-271.

JAYAKANTHAN, M., CHANDRASEKAR, S., MUTHUKUMARAN, J. & MATHUR, P. P. 2010. Analysis of CYP3A4-HIV-1 protease drugs interactions by computational methods for Highly Active Antiretroviral Therapy in HIV/AIDS. *J. Mol. Graph.,* 28**,** 455-463.

JENWITHEESUK, E. & SAMUDRALA, R. 2003. Improved prediction of HIV-1 protease-inhibitor binding energies by molecular dynamics simulations. *BMC Struct Biol,* 3**,** 2.

JINNOPAT, P., ISARANGKURA-NA-AYUTHAYA, P., UTACHEE, P., KITAGAWA, Y., DE SILVA, U. C., SIRIPANYAPHINYO, U., KAMEOKA, Y., TOKUNAGA, K., SAWANPANYALERT, P., IKUTA, K., AUWANIT, W. & KAMEOKA, M. 2009. Impact of Amino Acid Variations in Gag and Protease of HIV Type 1 CRF01_AE Strains on Drug Susceptibility of Virus to Protease Inhibitors. *J. Acquir. Immune Defic. Syndr.,* 52**,** 320-328.

JONES, G., WILLETT, P. & GLEN, R. C. 1995. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *Journal of Molecular Biology,* 245**,** 43-53.

KAPUST, R. B. & WAUGH, D. S. 1999. *Escherichia coli* maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which is it fused. *Protein Sci.,* 8**,** 1668-1674.

KHEDKAR, V. M., AMBRE, P. K., VERMA, J., SHAIKH, M. S., PISSURLENKAR, R. R. S. & COUTINHO, E. C. 2010. Molecular docking and 3D-QSAR studies of HIV-1 protease inhibitors. *J. Mol. Model.,* 16**,** 1251-1268.

KIEKEN, F., ARNOULT, E., BARBAULT, F., PAQUET, F., HUYNH-DINH, T., PAOLETTI, J., GENEST, D. & LANCELOT, G. 2002. HIV-1(Lai) genomic RNA: combined used of NMR and molecular dynamics simulation for studying the structure and internal dynamics of a mutated SL1 hairpin. *Eur. Biophys. J. Biophy.,* 31**,** 521-531.

KIRCHMAIR, J., DISTINTO, S., LIEDL, K. R., MARKT, P., ROLLINGER, J. M., SCHUSTER, D., SPITZER, G. M. & WOLBER, G. 2011. Development of anti-viral agents using molecular modeling and virtual screening techniques. *Infect. Disord. Drug. Targets,* 11**,** 64-93.

KOCH, S., COMAN, R., MUNOZ, I., DUNN, B. M., SLEASMAN, J. W. & DOODENOW, M. M. 2006 Amino acid changes in Gag that develop in vivo during protease inhibitor therapy have a dominant effect on viral replication and sensitivity to protease inhibitors ex vivo. . *Antivir. Ther.,* 11.

KOKEN, M. H. M., ODIJK, H. H. M., DUIN, M. V., FORNEROD, M. & HOEIJMAKERS, J. H. J. 1993. Augmentation of protein production by a combination of the T7 RNA polymerase system and ubiquitin fusion: overproduction of the human DNA repair protein, ERCC1, as a ubiquitin fusion protein in *Escherichia coli*. *Biochem. Biophys. Res. Commun.,* 195**,** 643–653.

KOMAI, T., ISHIKAWA, Y., YAGI, R., SUZUKISUNAGAWA, H., NISHIGAKI, T. & HANDA, H. 1997. Development of HIV-1 protease expression methods using the T7 phage promoter system. *Appl. Microbiol. Biotechnol.,* 47**,** 241-245.

KRAUSSLICHT, H.-G., INGRAHAMS, R. H., SKOOG, M. T., WIMMERT, E., PALLAIT, P. V. & CARTERT, C. A. 1989. Activity of purified biosynthetic proteinase of human immunodeficiency virus on natural substrates and synthetic peptides. *Proc. Natl. Acad. Sci. U. S. A.,* 86**,** 807-811.

LAVALLIE, E. R., DIBLASIO, E. A., KOVACIC, S., GRANT, K. L., SCHENDEL, P. F. & MCCOY, J. M. 1993. A thioredoxin gene fusion expression system that circumvents inclusion body formation in the E. coli cytoplasm. *Bio/Technology,* 11**,** 187–193.

LEE, S.-K., POTEMPA, M., KOLLI, M., ÖZEN, A., SCHIFFER, C. A. & SWANSTROM, R. 2012. Context Surrounding Processing Sites Is Crucial in Determining Cleavage Rate of a Subset of Processing Sites in HIV-1 Gag and Gag-Pro-Pol Polyprotein Precursors by ViralProtease. *J. Biol. Chem.,* 287**,** 13279-13290.

LI, J. K., JOHNSON, T., YANG, Y. Y. & SHORE, V. 1989. Selective separation of virus proteins and double-stranded RNAs by SDS-KCl precipitation. *J. Virol. Methods,* 26**,** 3-15.

LIHANA, R. W., SSEMWANGA, D., ABIMIKU, A. L. & NDEMBI, N. 2012. Update on HIV-1 Diversity in Africa: A Decade in Review. *AIDS Reviews,* 14**,** 83-100.

MAKATINI, M. M., PETZOLD, K., ARVIDSSON, P. I., HONARPARVAR, B., GOVENDER, T., MAGUIRE, G. E. M., PARBOOSING, R., SAYED, Y., SOLIMAN, M. E. S. & KRUGER, H. G. 2012. Synthesis, screening and computational investigation of pentacycloundecane-peptoids as potent CSA-HIV PR inhibitors. *Eur. J. Med. Chem.*

MAKRIDES, S. C. 1996. Strategies for Achieving High-Level Expression of Genes in *Escherichia coli*. *Microbiol. Rev.,* 60**,** 512-538.

MALET, I., ROQUEBERT, B., DALBAN, C., WIRDEN, M., AMELLAL, B., AGHER, R., SIMON, A., KATLAMA, C., COSTAGLIOLA, D., CALVEZ, V. & MARCELIN, A. G. 2007. Association of Gag cleavage sites to protease mutations and to virological response in HIV-1 treated patients. *J. Infect.,* 54**,** 367-374.

MARTINEZ-CAJAS, J. L., PANT-PAI, N., KLEIN, M. B. & WAINBERG, M. A. 2008. Role of Genetic Diversity amongst HIV-1 Non-B Subtypes in Drug Resistance: A Systematic Review of Virologic and Biochemical Evidence. *AIDS Reviews,* 10**,** 212-223.

MARTINEZ-PICADO, J., PRADO, J. G., FRY, E. E., PFAFFEROTT, K., LESLIE, A., CHETTY, S., THOBAKGALE, C., HONEYBORNE, I., CRAWFORD, H., MATTHEWS, P., PILLAY, T., ROUSSEAU, C., MULLINS, J. I., BRANDER, C., WALKER, B. D., STUART, D. I., KIEPIELA, P. & GOULDER, P. 2006. Fitness Cost of Escape Mutations in p24 Gag in Association with Control of Human Immunodeficiency Virus Type 1. *J. Virol.,* 80**,** 3617-3623.

MCCUTCHAN, F. E. 2006. Global epidemiology of HIV. *J. Med. Virol.,* 78**,** S7-S12.

MEEK, T. D., DAYTON, B. D., METCALF, B. W., DREYER, G. B., STRICKLER, J. E., GORNIAK, J. G., ROSENBERG, M., MOORE, M. L., MAGAARD, V. W. & DEBOUCK, C. 1989. Human immunodeficiency virus 1 protease expressed in Escherichia coli behaves as a dimeric aspartic protease. *Proc. Natl. Acad. Sci. U. S. A.,* 86**,** 1841-1845.

MIURA, T., BROCKMAN, M. A., BRUMME, Z. L., BRUMME, C. J., PEREYRA, F., TROCHA, A., BLOCK, B. L., SCHNEIDEWIND, A., ALLEN, T. M., HECKERMAN, D. & WALKER, B. D. 2009. HLA-Associated Alterations in Replication Capacity of Chimeric NL4-3 Viruses Carrying Gag-protease from Elite Controllers of Human Immunodeficiency Virus Type 1. *J. Virol.,* 83**,** 140-149.

MONINI, P., SGADARI, C., TOSCHI, E., BARILLARI, G. & ENSOLI, B. 2004. Antitumour effects of antiretroviral therapy. *Nat. Rev. Cancer,* 4**,** 861-875.

MORRIS, G. M., HUEY, R., LINDSTROM, W., SANNER, M. F., BELEW, R. K., GOODSELL, D. S. & OLSON, A. J. 2009. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.,* 30**,** 2785-2791.

MYINT, L., MATSUDA, M., MATSUDA, Z., YOKOMAKU, Y., CHIBA, T., OKANO, A., YAMADA, K. & SUGIURA, W. 2004. Gag non-cleavage site mutations contribute to full recovery of viral fitness in protease inhibitor-resistant human immunodeficiency virus type 1. *Antimicrob. Agents Chemother.,* 48**,** 444-452.

NALAM, M. N. L., ALI, A., ALTMAN, M. D., REDDY, G. S. K. K., CHELLAPPAN, S., KAIRYS, V., ZEN, A. E. L. O., CAO, H., GILSON, M. K., TIDOR, B., RANA, T. M. & SCHIFFER, C. A. 2010. Evaluating the Substrate-Envelope Hypothesis: Structural Analysis of Novel HIV-1 Protease Inhibitors Designed To Be Robust against Drug Resistance. *J. Virol.,* 84.

NDUNG'U, T., RENJIFO, B. & ESSEX, M. 2001. Construction and analysis of an infectious human immunodeficiency virus type 1 subtype C molecular clone. *J. Virol.,* 75**,** 4964-4972.

NIJHUIS, M., VAN MAARSEVEEN, N. M. & BOUCHER, C. A. B. 2007a. HIV protease resistance and viral fitness. *Curr. Opin. HIV. AIDS. ,* 2**,** 108-115.

NIJHUIS, M., VAN MAARSEVEEN, N. M., LASTERE, S., SCHIPPER, P., COAKLEY, E., GLASS, B., ROVENSKA, M., DE JONG, D., CHAPPEY, C., GOEDEGEBUURE, I. W., HEILEK-SNYDER, G., DULUDE, D., CAMMACK, N., BRAKIER-GINGRAS, L., KONVALINKA, J., PARKIN, N., KRÄUSSLICH, H.-G., BRUN-VEZINET, F. & BOUCHER, C. A. B. 2007b. A Novel Substrate-Based HIV-1 Protease Inhibitor Drug Resistance Mechanism. *PLoS Medicine,* 4**,** e36.

NIJHUIS, M., VAN MAARSEVEEN, N. M., VERHEYEN, J. & BOUCHER, C. A. B. 2008. Novel mechanisms of HIV protease inhibitor resistance. . *Curr. Opin. HIV. AIDS. ,* 3**,** 627-632.

OLSON, A. J. & GOODSELL, D. S. 1998. Automated docking and the search for HIV protease inhibitors. *SAR QSAR Environ. Res.,* 8**,** 273-85.

ÖZEN, A., HALILOĞL, T. & SCHIFFER, C. A. 2011. Dynamics of Preferential Substrate Recognition in HIV-1 Protease: Redefining the Substrate Envelope. *J. Mol. Biol.,* 410**,** 726-744.

ÖZEN, A., HALILOĞL, T. & SCHIFFER, C. A. 2012. HIV-1 Protease and Substrate Coevolution Validates the Substrate Envelope As the Substrate Recognition Pattern. *J. Chem. Theory Comput.,* 8**,** 703-714.

PADIGLIONE, A., ALEKSIC, E., FRENCH, M., ARNOTT, A., WILSON, K. M., TIPPETT, E., KAYE, M., GRAY, L., ELLETT, A., CRANE, M., LESLIE, D. E., LEWIN, S. R., BRESCHKIN, A., BIRCH, C., P.R., G., MCPHEE, D. A. & CROWE, S. M. 2010. Extremely prolonged HIV seroconversion associated with an MHC haplotype carrying disease susceptibility genes for antibody deficiency disorders. *Clin. Immunol.,* doi:10.1016/j.clim.2010.07.003.

PANDREA, I. & APETREI, C. 2010. Where the Wild Things Are: Pathogenesis of SIV Infection in African Nonhuman Primate Hosts. *Curr. HIV/AIDS Rep.,* 7**,** 28-36.

PARRY, C. M., KOHLI, A., BOINETT, C. J., TOWERS, G. J., MCCORMICK, A. L. & PILLAY, D. 2009. Gag Determinants of Fitness and Drug Susceptibility in Protease Inhibitor-Resistant Human Immunodeficiency Virus Type 1. *J. Virol.,* 83**,** 9094-9101.

PÈPE, G., COURCAMBECK, J., PERBOST, R., JOUANNA, P. & HALFON, P. 2008. Prediction of HIV-1 protease inhibitor resistance by Molecular Modeling Protocols (MMPs) using GenMol(TM) software. *Eur. J. Med. Chem.,* 43**,** 2518-2534.

PEREZ, M. A. S., FERNANDES, P. A. & RAMOS, M. J. 2010. Substrate Recognition in HIV-1 Protease: A Computational Study. *J. Physical. Chem.,* 114**,** 2525-2532.

PERRIN, L., KAISER, L. & YERLY, S. 2003. Travel and the spread of HIV-1 genetic variants. *Lancet Infect. Dis. ,* 3**,** 22-27.

PETTERSEN, E. F., GODDARD, T. D., HUANG, C. C., COUCH, G. S., GREENBLATT, D. M., MENG, E. C. & FERRIN, T. E. 2004. UCSF Chimera—A visualization system for exploratory research and analysis. *J. Comput. Chem.,* 25**,** 1605-1612.

PETTIT, S. C., MOODY, M. D., WEHBIE, R. S., KAPLAN, A. H., NANTERMET, P. V., KLEIN, C. A. & SWANSTROM, R. 1994. The p2 Domain of Human Immunodeficiency Virus Type 1Gag Regulates Sequential Proteolytic Processing and Is Required To Produce Fully Infectious Virions. *J. Virol.,* 68.

PIETRUCCI, F., MARINELLI, F., CARLONI, P. & LAIO, A. 2009. Substrate Binding Mechanism of HIV-1 Protease from Explicit-Solvent Atomistic Simulations. *J. Am. Chem. Soc.,* 131**,** 11811-11818.

POPE, M. & HAASE, A. T. 2003. Transmission, acute HIV-1 infection and the quest for strategies to prevent infection. *Nat. Med.,* 9**,** 847-852.

PRABU-JEYABALAN, M., NALIVAIKA, E. & SCHIFFER, C. A. 2002. Substrate Shape Determines Specificity of Recognition for HIV-1 Protease: Analysis of Crystal Structures of Six Substrate Complexes. *Structure,* 10**,** 369-381.

PRADO, J. G., HONEYBORNE, I., BRIERLEY, I., PUERTAS, M. C., MARTINEZ-PICADO, J. & GOULDER, P. J. R. 2009. Functional Consequences of Human Immunodeficiency Virus Escape from an HLA-B*13-Restricted CD8 T-Cell Epitope in p1 Gag Protein. *J. Virol.,* 83**,** 1018-1025.

QUINONES-MATEU, M. E., BALL, S. C., MAROZSAN, A. J., TORRE, V. S., ALBRIGHT, J. L., VANHAM, G., VAN DER GROEN, G., COLEBUNDERS, R. L. & ARTS, E. J. 2000. A dual infection/competition assay shows a correlation between ex vivo human immunodeficiency virus type 1 fitness and disease progression. *J. Virol.,* 74**,** 9222–9233.

RAMIREZ, B. C., SIMON-LORIERE, E., GALETTO, R. & NEGRONI, M. 2008. Implications of recombination for HIV diversity. *Virus Res.,* 134**,** 64-73.

RASBAND, W. S. 2009. ImageJ. http://rsb.info.nih.gov/ij: U.S. National Institutes of Health, Bethesda, Maryland, USA.

RODRIGUEZ, M. A., DING, M., RATNER, D., CHEN, Y., TRIPATHY, S. P., KULKARNI, S. S., CHATTERJEE, R., TARWATER, P. M. & GUPTA, P. 2009. High replication fitness and transmission efficiency of HIV-1 subtype C from India: Implications for subtype C predominance. *Virology,* 385**,** 416-424.

SAMUELSSON, E., MOKS, T., NILSSON, B. & UHLE´N, M. 1994. Enhanced in vitro refolding of insulin-like growth factor I using a solubilizing fusion partner. *Biochemistry* 33**,** 4207–4211.

SCHNEIDEWIND, A., BROCKMAN, M. A., YANG, R., ADAM, R. I., LI, B., GALL, S. L., RINALDO, C. R., CRAGGS, S. L., ALLGAIER, R. L., POWER, K. A., KUNTZEN, T., TUNG, C.-S., LABUTE, M. X., MUELLER, S. M., HARRER, T., MCMICHAEL, A. J., GOULDER, P. J. R., AIKEN, C., BRANDER, C., KELLEHER, A. D. & ALLEN, T. M. 2007. Escape from the Dominant HLA-B27-Restricted Cytotoxic T-Lymphocyte Response in Gag Is Associated with a Dramatic Reduction in Human Immunodeficiency Virus Type 1 Replication. *J. Virol.,* 81.

SHAFER, R. W. 2006. Rationale and uses of a public HIV drug-resistance database. *J. Infect. Dis.,* 194**,** S51-S58.

SILVESTRI, G., PAIARDINI, M., PANDREA, I., LEDERMAN, M. M. & SODORA, D. L. 2007. Understanding the benign nature of SIV infection in natural hosts. *J. Clin. Invest.,* 117**,** 3148- 3154.

SIMON, V., HO, D. D. & KARIM, Q. A. 2006. HIV/AIDS epidemiology, pathogenesis, prevention, and treatment. *Lancet,* 368**,** 489-504.

SOARES, R. O., BATISTA, P. R., COSTA, M. G. S., DARDENNE, L. E., PASCUTTI, P. G. & SOARES, M. A. 2010. Understanding the HIV-1 protease nelfinavir resistance mutation D30N in subtypes B and C through molecular dynamics simulations. *J. Mol. Graph.,* 29**,** 137-147.

SOUSA, S. F., FERNANDES, P. A. & RAMOS, M. J. 2006. Protein-ligand docking: current status and future challenges. *Proteins,* 65**,** 15-26.

SOUTH AFRICAN NATIONAL DEPARTMENT OF HEALTH 2004. National Antiretroviral Treatment Guidlines.

SOUTH AFRICAN NATIONAL DEPARTMENT OF HEALTH 2012. Global AIDS Response - Progress Report South Africa.

STEBBINS, J. & DEBOUCK, C. 1994. Retroviral proteases. *In:* LAWRENCE C. KUO & SHAFER, J. A. (eds.) *Methods in Enzymology.*

STORN, R. & PRICE, K. 1995. Differential Evolution - A Simple and Efficient Adaptive Scheme for Global Optimization over Continuous Spaces. Technical Report; International Computer Science Institute, Berkley,CA.

STOVER, J., KORENROMP, E. L., BLAKLEY, M., KOMATSU, R., VIISAINEN, K., BOLLINGER, L. & ATUN, R. 2011. Long-Term Costs and Health Impact of Continued Global Fund Support for Antiretroviral Therapy. *PLoS ONE,* 6**,** e21048.

THOMAS, J. A. & GORELICK, R. J. 2008. Nucleocapsid protein function in early infection processes. *Virus Res.,* 134**,** 39-63.

THOMSEN, R. & CHRISTENSEN, M. H. 2006. MolDock: A New Technique for High-Accuracy Molecular Docking. *J. Med. Chem.,* 49**,** 3315-3321.

UNAIDS 2011. World AIDS Day report.

VAN MAARSEVEEN, N. M., ANDERSSON, D., LEPSIK, M., FUN, A., SCHIPPER, P. J., DE JONG, D., BOUCHER, C. A. B. & NIJHUIS, M. 2012. Modulation of HIV-1 Gag NC/p1 cleavage efficiency affects protease inhibitor resistance and viral replicative capacity. *Retrovirology,* 9.

VANDEWOUDE, S. & APETREI, C. 2006. Going Wild: Lessons from Naturally Occurring T-Lymphotropic Lentiviruses. *Clin. Microbiol. Rev.,* 19**,** 728-762.

VELAZQUEZ-CAMPOY, A., TODD, M. J., VEGA, S. & FREIRE, E. 2001. Catalytic efficiency and vitality of HIV-1 proteases from African viral subtypes. *Proc. Natl. Acad. Sci. U. S. A.,* 98**,** 6062-6067.

VERHEYEN, J., KNOPS, E., KUPFER, B., HAMOUDA, O., SOMOGYI, S., SCHULDENZUCKER, U., HOFFMANN, D., KAISER, R., PFISTER, H. & KUCHERER, C. 2009. Prevalence of C-terminal Gag cleavage site mutations in HIV from therapy-naive patients. *J. Infect.,* 58**,** 61-67.

VERKHIVKER, G. 2009. Computational proteomics analysis of binding mechanisms and molecular signatures of the HIV-1 protease drugs. *Artif. Intell. Med.,* 45**,** 197-206.

VICKREY, J. F., LOGSDON, B. C., PROTEASA, G., PALMER, S., WINTERS, M. A., MERIGAN, T. C. & KOVARI, L. C. 2003. HIV-1 protease variants from 100-fold drug resistant clinical isolates: expression, purification, and crystallization. *Protein Expression Purif.,* 28**,** 165-172.

VOET, D. & VOET, J. 2004. *Biochemistry,* USA, John Wiley & Sons, Inc.

VOLONTE, F., PIUBELLI, L. & POLLEGIONI, L. 2011. Optimizing HIV-1 protease production in Escherichia coli as fusion protein. *Microb. Cell. Fact.,* 10**,** 53.

WALKER, P. R., PYBUS, O. G., RAMBAUT, A. & HOLMES, E. C. 2005. Comparative population dynamics of HIV-1 subtypes B and C: subtype-specific differences in patterns of epidemic growth. *Infect. Genet. Evol.,* 5**,** 199-208.

WALTER, B. L., ARMITAGE, A. E., GRAHAM, S. C., DE OLIVEIRA, T., SKINHØJ, P., JONES, E. Y., STUART, D. I., MCMICHAEL, A. J., CHESEBRO, B. & IVERSEN, A. K. N. 2009. Functional characteristics of HIV-1 subtype C compatible with increased heterosexual transmissibility. *AIDS,* 23**,** 1047-1057.

WAN, M., TAKAGI, M., LOH, B. N. & IMANAKA, T. 1995. Comparison of HIV-1 protease expression in different fusion forms. *Biochem. Mol. Biol. Int.,* 36**,** 411-9.

WAN, M., TAKAGII, M., LOH, B.-N., XU, X.-Z. & IMANAKA, T. 1996. Autoprocessing : an essential step for the activation of HIV-1 protease. *Biochem. J.,* 316**,** 569-573.

WENSING, A. M. J., VAN MAARSEVEEN, N. M. & NIJHUIS, M. 2010. Fifteen years of HIV Protease Inhibitors: raising the barrier to resistance. *Antiviral Res.,* 85**,** 59-74.

WERTHEIM, J. O. & WOROBEY, M. 2009. Dating the age of the SIV lineages that gave rise to HIV-1 and HIV-2. *PLoS Comput. Biol.,* 5**,** 1.

WHITE, T. A., BARTESAGHI, A., BORGNIA, M. J., DE LA CRUZ, M. J. V., NANDWANI, R., HOXIE, J. A., BESS, J. W., LIFSON, J. D., MILNE, J. L. S. & SUBRAMANIAM, S. 2011. Three-Dimensional Structures of Soluble CD4-Bound States of Trimeric Simian Immunodeficiency Virus Envelope Glycoproteins Determined by Using Cryo-Electron Tomography. *J. Virol.,* 85**,** 12114-12123.

WHO 2009. AIDS epidemic update.

WILKINSON, D. L., MA, N. T., HAUGHT, C. & HARRISON, R. G. 1995. Purification by immobilized metal affinity chromatography of human atrial natriuretic peptide expressed in a novel thioredoxin fusion protein. *Biotechnol. Prog. ,* 11**,** 265–269.

WRIGHT, J. K., BRUMME, Z. L., CARLSON, J. M., HECKERMAN, D., KADIE, C. M., BRUMME, C. J., WANG, B., LOSINA, E., MIURA, T., CHONCO, F., STOK, M. V. D., MNCUBE, Z., BISHOP, K., GOULDER, P. J. R., WALKER, B. D., BROCKMAN, M. A. & NDUNG'U, T. 2010. Gag-Protease-Mediated Replication Capacity in HIV-1 Subtype C Chronic Infection: Associations with HLA Type and Clinical Parameters. *J. Virol.,* 84**,** 10820-10831.

ZDOBNOV, E. M. & APWEILER, R. 2001. InterProScan – an integration platform for the signature-recognition methods in InterPro. *Bioinformatics,* 17**,** 847-848.

ZHANG, Y. M., IMAMICHI, H., IMAMICHI, T., LANE, H. C., FALLOON, J., VASUDEVACHARI, M. B. & SALZMAN, N. P. 1997. Drug resistance during Indinavir therapy is caused by mutations in the protease gene and in its Gag substrate cleavage sites. *J. Virol.,* 71**,** 6662-6670.