

**THE CHARACTERISATION AND EXPRESSION OF HIV-1
SUBTYPE C GAG**

CANDICE CORENE SAMPSON

**This thesis presented in partial fulfilment of the requirements for the degree of
Master of Science in Medical Sciences (Medical Virology) in the Faculty of
Health Sciences, University of Stellenbosch.**



PROMOTER

Candice Sampson
Signature

Prof. E Janse van Rensburg

CO-PROMOTER

Dr S Engelbrecht

March 2002

SUMMARY

The gag gene of HIV-1 encodes for two to five main structural proteins and contains several conserved epitopes. The conserved epitopes and the immune responses are important in the pathogenesis of HIV-1 infection. The HIV-1 subtype C gag sequence has been published.

DECLARATION

I, the undersigned, hereby declare that the work contained in this thesis is my own original work and that I have not previously in its entirety or in part submitted it at any university for a degree.

Fifteen HIV-1 subtype C isolates selected for this study were from 1997 and 1998 and 1999 from the HIV-1 positive patients attending the infectious diseases department at Tygerberg Hospital. The gag gene of these isolates was amplified, cloned into mammalian expression vectors and sequenced. Restriction digests as well as phylogenetic analyses were performed on the sequencing data. Published mutational analyses and CTL epitopes were compared to the predicted amino acid sequences of the gag clones.

Signature

Date

SUMMARY

The *gag* gene of HIV-1 encodes for one of the major structural proteins, which contains several conserved cytotoxic T cell (CTL) epitopes. Gag specific CTL responses are important in controlling viral load during acute infection and asymptomatic stages of the infection. Currently, only one complete South African HIV-1 subtype C *gag* sequence has been published. The first aim of this study was to characterise the complete *gag* gene of 15 HIV-1 subtype C isolates, to be used as a set of reference sequences in the design of a South African HIV-1 subtype C vaccine.

Fifteen HIV-1 subtype C isolates selected for this study, were isolated during 1998 and 1999 from the HIV-1 positive patients attending the Infectious Disease Clinic at Tygerberg Hospital. The *gag* gene of these isolates was amplified by PCR, cloned into mammalian expression vectors and sequenced. Restriction digest analyses as well as phylogenetic analyses were performed on the sequencing data. Previously published mutational analyses and CTL epitopes were compared to the predicted amino acid sequences of the *gag* clones.

Sequences of 23 complete *gag* genes representing the 15 HIV-1 subtype C isolates as well as one complete sequence of an HIV-1 subtype B isolate were compiled. Subtyping by restriction fragment length polymorphism (RFLP) would have correctly identified 14 of the 15 subtype C isolates as subtype C and one as unidentifiable. The subtype B isolate would have also been correctly identified. Phylogenetic analyses showed that our subtype C isolates clustered with reference subtype C strains from various countries, including Botswana, India, Israel, Tanzania and Zambia. Strains from Ethiopia and Brazil formed a separate subtype C cluster. The diversity between our isolates was comparable to the diversity seen between all the HIV-1 subtype C strains. Comparisons of previously published mutational analyses and CTL epitopes to the predicted amino acid sequences of

the *gag* clones, showed conservation in most of the clones throughout the sequence.

A second aim was to establish transfection and Western Blot techniques in our laboratory for use in future studies. An *in vitro* transcription/ translation assay was performed on the *gag* clones and the protein producing clones were used to transfect mammalian cells using electroporation. A Western blot was then used to screen for Gag protein expression in the transfected cell lysates.

The *in vitro* transcription/ translation assay showed that seven of the 23 clones could produce a protein of ~55 kDa in size. Four out of the seven of these clones gave a weak expression of a ~55 kDa protein after transfection in a mammalian cell line. Since the completion of the experimental work of this study, other cloned HIV-1 genes have successfully been transfected into mammalian cells using the electroporation technique and the proteins produced were screened for by Western blot.

To conclude with; the native form of the *gag* gene does not elicit strong expression of the protein, but studies have shown that expression can be improved by sequence-modification of the *gag* nucleotide sequence. Due to the conservation of *gag*, the sequence of any subtype C strain can be used for the development of a Southern African vaccine.

OPSOMMING

Die HIV-1 *gag* geen kodeer vir een van die hoof strukturele proteïene en bevat verskeie sitotoksiese T-limfosiet epitope. Gag spesifieke sellulêre immuun respons is belangrik vir die beheer van virale lading tydens akute infeksies en tydens asimptomatiese fases van die infeksie. Tans is slegs een volledige Suid Afrikaanse HIV-1 sub tipe C nukleïensuur volgorde gepubliseer. Die eerste doel van hierdie studie was om die volledige *gag* geen van 15 HIV-1 sub tipe C isolate te karakteriseer, om gebruik te word as 'n stel verwysings nukleïensuur volgordes, vir die ontwerp van 'n Suid Afrikaanse HIV-1 sub tipe C entstof.

Die 15 HIV-1 sub tipe C isolate wat vir hierdie studie geselekteer is, is tydens 1998 en 1999 geïsoleer vanaf HIV-1 positiewe pasiënte wat die Infeksiesiekte Kliniek, Tygerberg Hospitaal bygewoon het. Die *gag* geen van hierdie isolate is geamplifiseer deur PCR, gekloneer in soogdier ekspressie vektore en die nukleïensuur volgorde is bepaal. Die nukleïensuur volgorde is gebruik in restriksie ensiem analyses asook filogenetiese analyses. Reeds gepubliseerde mutasie analyses en limfosiet epitope is met die voorspelde aminosuur volgorde van die *gag* klone vergelyk.

Die nukleïensuur volgordes van die 23 volledige *gag* gene wat die 15 HIV-1 sub tipe C isolate verteenwoordig, asook een volledige nukleïensuur volgorde van een HIV-1 sub tipe B isolaat, is saamgestel. Subtipering deur middel van restriksie fragment lengte polimorfisme (RFLP) sou 14 uit die 15 sub tipe C isolate korrek geïdentifiseer het, maar sou een nie kon identifiseer nie. RFLP sou ook die sub tipe B isolaat korrek geïdentifiseer het. Filogenetiese analyses het gewys dat ons sub tipe C isolate met die verwysings sub tipe C stamme van verskeie lande, insluitend Botswana, Indië, Israel, Tanzanië en Zambië groepeer. Stamme van Ethiopië en Brasilië het 'n aparte sub tipe C groep gevorm. Die diversiteit tussen ons isolate was vergelykbaar met die diversiteit tussen al die sub tipe C stamme. Vergelykings van gepubliseerde mutasie analyses en limfosiet epitope met die

voorspelde aminosuur volgorde van die *gag* klone, het konservasie in meeste van die klone, deur die hele nukleïensuur volgorde, getoon.

Die tweede doel was om die metodes van transfeksie en Westerse klad in ons laboratorium tot stand te bring. *In vitro* transkripsie/ translasië toets is gedoen op die *gag* klone en die proteïen produserende klone is gebruik om soogdierselle te transfekteer deur gebruik te maak van elektroporasie. 'n Westerse klad is toe gebruik om vir Gag proteïenuitdrukking in die sellisate te toets.

Die *in vitro* transkripsie/ translasië toets het getoon dat sewe uit 23 klone, 'n proteïen van ~55 kDa kon produseer. Vier uit die sewe van hierdie klone het 'n ~55 kDa proteïen swak uitgedruk na transfektering van soogdier selle. Sedert die voltooiing van die eksperimentele werk van hierdie studie, is ander gekloneerde HIV-1 gene suksesvol in soogdierselle getransfekteer met die gebruik van elektroporasie en die proteïene is met 'n Westerse klad aangetoon.

Ten slotte: die natuurlike vorm van die *gag* geen ontlok nie 'n sterk ekspressie van die proteïen nie, maar ander studies het wel aangetoon dat die ekspressie verbeter kan word met modifikasie van die *gag* nukleïensuur volgorde. As gevolg van die konservasie van *gag*, kan die nukleïensuur volgorde van enige sub tipe C stam gebruik word vir die ontwikkeling van 'n Suider Afrikaanse entstof.

ACKNOWLEDGEMENTS

I wish to extend my sincere thanks and appreciation to:

Prof. E Janse van Rensburg, my promoter, and Dr S Engelbrecht, my co-promoter, for all their advice, assistance and guidance.

The Poliomyelitis Research Foundation, the South African AIDS Vaccine Initiative and the Harry Crossley Foundation for the funding of this study.

My family and friends for their support and encouragement throughout the study period.

LIST OF GENERAL ABBREVIATIONS

bp	basepairs
kb	kilo-basepairs
°C	degree Celsius
nm	nanometre
mm	millimetre
cm	centimetre
pmol	picomole
mM	millimolar
M	molar
N	normal
µg	microgram
mg	milligram
g	gram
µl	microlitre
ml	millilitre
l	litre
µF	microFaraday
kDa	kilodaltons
Da	daltons

CONTENTS

CHAPTER 3:

	Page
Results	
Summary	i
Opsomming	iii
Acknowledgements	v
List of general abbreviations	vi

CHAPTER 1:

Introduction, Aim of Study and Literature Review	1
1.1 Introduction	2
1.2 Aim of Study	3
1.3 Literature Review	4

CHAPTER 2:

Material and Methods	29
2.1 Materials	31
2.2 Methods	34
Appendix Tables	50

CHAPTER ONE

Introduction, Aim of this study and Literature review

CHAPTER 3:

Results 55

3.1 Characterisation Studies 57

3.2 Expression Studies 64

Tables 66

Figures 72

CHAPTER 4:

Discussions and conclusion 103

4.1 Discussion 104

4.2 Conclusion 123

CHAPTER 5:

5. References 125

APPENDICES:

Appendix A Vaccines Strategies 150

Appendix B Immunology 158

CHAPTER ONE

Introduction, Aim of this study and Literature review

	Page
1.1 Introduction	2
1.2 Aim of this Study	3
1.3 Literature Review	4
1.3.1 History of the Human Immunodeficiency virus	4
1.3.2 Diversity of the Human Immunodeficiency virus	4
1.3.3 HIV-1 in South Africa	6
1.3.4 The HIV-1 virion and life cycle	8
1.3.4.1 HIV-1 virion	8
1.3.4.2 The HIV-1 lifecycle	10
1.3.5 The Gag Protein	11
1.3.5.1 Matrix	13
1.3.5.2 Capsid	14
1.3.5.3 Nucleocapsid	15
1.3.5.4 p6 and Spacer peptides	15
1.3.5.5 Techniques used to characterise the <i>gag</i> gene and Gag protein	16
1.3.5.5a Cloning and Sequencing	16
1.3.5.5b Molecular Phylogenetics	19
1.3.5.5c <i>In vitro</i> transcription/ translation	22
1.3.5.5d Transfection	22
1.3.5.5e Western Blot	24
1.3.6 HIV-1 Vaccines Strategies	25
1.3.6.1 Immunology	25
1.3.6.1.1 Humoral Immunity	25
1.3.6.1.2 Cell mediated Immunity	26
1.3.6.2 Types of vaccines	26

CHAPTER ONE

1. Introduction and Literature Review

1.1 Introduction

According to the latest United Nations global AIDS report, there are currently 36.1 million people world-wide living with HIV/AIDS and a total of 16.3 million people have died of AIDS related diseases since the beginning of the epidemic (UNAIDS/WHO, 2000). It is therefore of global importance to develop an effective vaccine to combat further spread of the disease. HIV-1 subtype C constitutes more than 55 % of the circulating viruses world-wide (Esparza & Bhamarapravati, 2000) and is the most prevalent subtype in South Africa (Bredell et al, 1998; Engelbrecht et al, 1999; Van Harmelen et al, 1999a). It is therefore of national importance to concentrate on developing an HIV-1 subtype C vaccine. A safe and effective HIV vaccine should be able to elicit a broad-based humoral and cellular immune response without potentially causing disease.

The *gag* gene is of the more conserved genes in the virus genome and is translated into a polyprotein precursor, which is proteolytically cleaved into p17 matrix, p24 capsid, p7 nucleocapsid and p6 proteins. Gag specific cytotoxic T-cell responses play a meaningful role in the control of disease progression. It is therefore important to look at the Gag polyprotein as a promising candidate in the development of an HIV vaccine.

The first Chapter will give an overview of the history, diversity and morphology of the HIV-1 virus, and then progress to describe Gag, as a protein, as well as to look at the techniques used in this study to characterise the gene. A short overview will then be given of HIV-1 vaccines. The second Chapter will describe the materials and methods used in this study. In Chapter three the results will be presented, and discussed in Chapter four. Large Figures and Figures stretching over several pages were listed in the Appendices. The Figures of materials and methods are

given in Appendix A, while the tables are given at the end of Chapter two. Tables and Figures (photos of gels and blots, phylogenetic trees, examples of similarity plots, graphs, and a helix cartoon) of results are listed at the back of Chapter three, while additional result figures (sequence raw data, restriction analysis and maps, similarity plots, distance matrixes, and alignments) are listed in Appendix B.

1.2 Aim of this Study

Currently, only one complete South African HIV-1 subtype C *gag* sequence has been described (Rodenburg et al, 2001). The first aim of this study was to characterise the *gag* gene of 15 HIV-1 subtype C isolates, to be used as a set of reference sequences in the design of a South African HIV-1 subtype C vaccine. The characterisation of *gag* was achieved by the amplification, cloning and sequencing of the gene. The sequencing data was then used in restriction digest- and phylogenetic analyses, as well as to compile predicted amino acid sequences for the protein. Phylogenetic analyses, and amino acid comparisons to previous mutational analyses and CTL epitopes were done on the predicted amino acid sequence. Conserved CTL epitopes are important for vaccine induced immunity.

A second aim was to establish transfection and Western Blot techniques in our laboratory for use in future studies. Although it was known that *gag* did not elicit good protein expression, the expression of proteins from the cloned *gag* genes was however studied to establish the techniques, by transfecting the clones into mammalian cells and screening for protein production by Western blot. This was done by evaluating three different transfection methods on the production of chloramphenicol acetyltransferase type 1 protein (CAT) in cells transfected with a CAT expression plasmid. The technique, which delivered the highest concentration of the CAT protein, was then chosen to transfect cells with selected *gag* clones, which were selected on their ability to produce a ~55 kDa protein during *in vitro* transcription/ translation.

1.3 Literature Review

1.3.1 History of the Human Immunodeficiency virus

Twenty years ago, in 1981, several cases of acquired immune deficiency were reported in *The New England Journal of Medicine* (Gottlieb et al, 1981a; Masur et al, 1981; Siegal et al, 1981). All patients reported on, mostly of whom were homosexual men, were previously healthy, but presented with *Pneumocystis carinii* pneumonia, a disease caused by an organism, which rarely if ever causes disease in immunologically competent persons (Hughes, 1977). Some of the patients also developed Kaposi's sarcoma. The combination of Kaposi's sarcoma and *Pneumocystis* pneumonia in homosexual men were also previously reported on in the *Morbidity and Mortality Weekly Report* of June, July and August 1981, already suggesting the start of an outbreak (Gottlieb et al, 1981b; Friedman-Kien et al, 1981; Friedman et al, 1981). The cases were from both the west and east coasts of the United States of America in mostly homosexual men, but also in drug abusers. The disease was named: the acquired immune deficiency syndrome (AIDS). Two years later, in 1983, the group of Luc Montagnier isolated the first T-lymphotropic retrovirus in a patient at risk for AIDS (Barré-Sinoussi et al, 1983). Although the exact role that this retrovirus played in AIDS was not known at that time, it was later established to be the etiologic agent of AIDS (Broder & Gallo, 1984; Ratner et al, 1985) and was named the Human Immunodeficiency virus (HIV) (Coffin et al, 1986a, 1986b).

1.3.2 Diversity of the Human Immunodeficiency virus

Since the realisation of HIV/AIDS in 1981, the epidemic has spread through out the world with an estimated 36.1 million people currently infected with HIV, of whom 5.3 million were infected in the year 2000. The global distribution of HIV infected people is given in Figure 1.1. Since the beginning of the epidemic a total of 21.8

million people have died of AIDS related illnesses. Of the 36.1 million HIV infected people, 25.3 million live in sub-Saharan Africa – a staggering 70% (UNAIDS/WHO, 2000).

Adults and children estimated to be living with HIV/AIDS as of end 2000



Total: 36.1 million



Figure 1.1 The global view of the total number of people living with HIV/AIDS. (http://www.unaids.org/epidemic_update/report/index.html)

There are two distinct human immunodeficiency viruses: type one (HIV-1) and type two (HIV-2) (Clavel et al, 1986). The worldwide pandemic is caused predominantly by HIV-1, with HIV-2 primarily found in West Africa (Van der Loeff et al, 1999). HIV-1 is divided into major groups, M (main), N (non-M, non-O) and O (outlier) (Louwagie et al, 1995; De Leys et al, 1990; Gürtler et al, 1994; Simon et al, 1998), while group M is further divided into nine genetically distinct subtypes (A-D, F-H, J, K) (Kuiken et al, 2000). Most strains found worldwide and which is responsible for

the pandemic, belong to the HIV-1 group M. Genetic diversity of group M is further broadened by the occurrence of quasispecies within the same host (Hahn et al, 1986; Sakai et al, 1988), giving rise to recombinant forms (Hu & Temin, 1990; Robertson et al 1995), which are spread and result in circulating recombinant forms (CRFs). The former subtypes E and I are now classified as CRFs (Kuiken et al, 2000). The most commonly transmitted HIV virus, is the HIV-1 subtype C virus, which constitute more than 55% of all circulating viruses worldwide and is the dominant subtype in southern Africa and India (Esparza & Bhamarapravati, 2000).

The high genetic diversity of HIV is the result of multiple introductions of simian viruses into humans (Hahn et al, 2000). Both HIV-1 and HIV-2 are results of zoonotic transmissions from chimpanzees (*Pan troglodytes*) (SIVcpz) (Gao et al, 1999) and sooty mangabeys (*Cercocebus atys*) (SIVsm) (Hirsch et al, 1989; Gao et al, 1992) respectively. Similarly, the three groups of HIV-1 represent multiple introductions of SIVcpz into humans (Hahn et al, 2000). The HIV-1 group M viruses, however appear to have risen from a single cross-species transmission (Gao et al, 1999). The genetic divergence within this group, might have been caused by the high error rate of reverse transcriptase and the high replication rate of the virus in infected individuals (Peeters & Sharp, 2000).

1.3.3 HIV-1 in South Africa

South Africa has one of the fastest growing HIV-1 epidemics (UNAIDS/WHO, 2000) with 4.7 million people out of a population of 40.6 million infected with the virus. There is a geographical variation in severity of the epidemic with KwaZulu-Natal having the highest prevalence rate (36.2%) and the Western Cape the lowest (8.7%) (Department of Health / Directorate Health Systems Research, 2001). The initial HIV-1 epidemic in South Africa occurred mainly in the homosexual population, but then shifted to a predominantly heterosexual epidemic (Williamson et al, 1995). In South Africa, homosexual transmission is associated with HIV-1

subtype B viruses, while heterosexual transmission is associated with the subtype C virus (Williamson et al, 1995; Van Harmelen et al, 1997). HIV-1 subtype C is the dominant subtype in the present South African epidemic (Van Harmelen et al, 1999a).

HIV-1 infected South Africans account for 13% of HIV-1 infections worldwide. Several factors may contribute to the high prevalence in South Africa, which include the migrant labour system, the low socio-economic status of women, high rates of other sexually transmitted diseases, and the limited availability of anti-HIV drug therapies in South Africa (Morris & Williamson, 2001). Studies have also suggested that the period from HIV-1 infection to the progression of AIDS is shortened by about two years in developing countries, where access to health care is limited (Grant et al, 1997). The opportunistic infection, tuberculosis, which is associated with poor communities, low socio-economic status, and malnutrition, is the major cause of death in sub-Saharan African AIDS patients (Shafer et al, 1996; De Cock et al, 1996, Morris & Williamson, 2001).

Although poverty and the lower socio-economic status of sub-Saharan African countries surely play a role in the high prevalence rate of HIV-1 infected people, a multitude of other factors also contribute to this African epidemic. One such a factor might be the viral genotype. HIV-1 subtype C, the dominant subtype in South Africa and the world, differ from other HIV-1 subtypes in several ways. The preferred co-receptor for subtype C viruses isolated from patients at all stages of disease is the CC chemokine type 5 receptor (CCR5) (Morris et al, 2001; Peeters et al, 1999; Treurnicht et al, in press), while other subtypes often switch to CXC chemokine type 4 receptor (CXCR4) usage at later stages of the disease (Tscherning et al, 1998; Connor et al, 1997). Majority of subtype C viruses, including the South African subtype C strains, contains an additional NF- κ B binding site in their LTR regions (Hunt & Tiemessen, 2000; Johansson et al, 1995; Montano et al, 1997). A recent study also identified further additional NF- κ B-like sites in South African HIV-1 subtype C isolates (Scriba et al, in press). Both the

occurrence of additional NF- κ B-binding sites and the dominant CCR5 co-receptor usage might theoretically play a role in improving the replication kinetics and transmission respectively, but must still be proven experimentally (Morris & Williamson, 2001, Scriba et al, in press).

Although the HIV-1 subtype C virus is the most common HIV-1 subtype worldwide, it has been relatively poorly studied. Currently there are 47 full length HIV-1 subtype C *gag* sequences in the Los Alamos sequence database, of which only 1 comes from a South African strain (Rodenburg et al, 2001). While there are 16 South African subtype C p17 sequences available in the Los Alamos database (Becker et al, 1995; Williamson et al, 1995; Rodenburg et al, 2001), there are no South African sequences, except for the full length sequence, available for p24. A study, which looked at the HIV-1 Gag-specific CTL responses in South African adults and children, found the dominant Gag epitope in these subjects to be in the p24 region (Goulder et al, 2000).

1.3.4 The HIV-1 virion and life cycle

1.3.4.1 HIV-1 virion

The near 9 kb viral genome encodes several viral proteins, of which there are structural, regulatory and accessory proteins. The structural proteins, namely, capsid (CA), matrix (MA), nucleocapsid (NC), surface envelope glycoprotein (SU), transmembrane envelope glycoprotein (TM), reverse transcriptase (RT), protease (PR) and integrase (IN), form the virus particle. All mature structural proteins are translated as polyprotein precursors, which are proteolytically cleaved by either viral or host cellular protease. The regulatory proteins, Tat (trans-activator of viral transcription) and Rev (regulator of viral expression) control viral replication by regulating viral transcription, and viral RNA transport and splicing respectively. The accessory proteins, Nef, Vpu, Vif and Vpr, play a variety of roles throughout

the viral life cycle. Figure 1.2 illustrates the structural protein arrangement in an HIV virion.

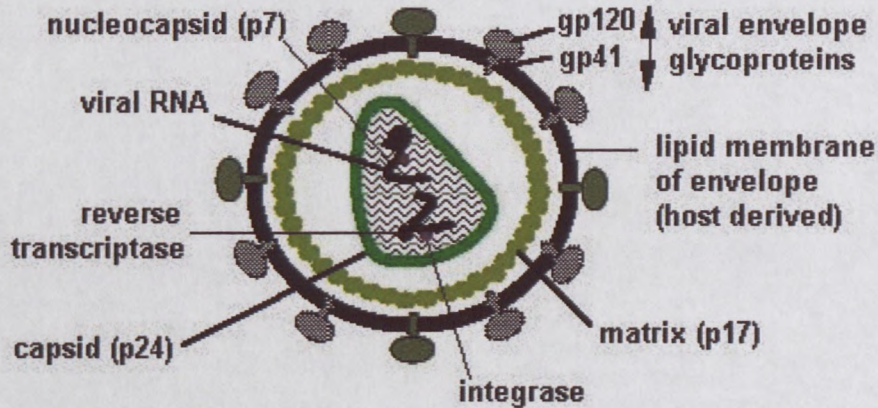


Figure 1.2 A cartoon portraying the HIV virion (<http://www.urmc.rochester.edu/smd/mbi/grad2/hiv99B.html>). The HIV-1 virion consists of several structural proteins. The envelope glycoproteins, gp120 and gp 41 function as adhesion and fusion proteins. The virus is covered by a lipid host derived membrane, which is surrounded in the inside by the p17 matrix proteins. The p24 capsid forms a shell surrounding the nucleocore, consisting of viral genome, p7 nucleocapsid, p6, reverse transcriptase, and integrase.

The HIV-1 virion contains a host derived lipid membrane, with the envelope glycoproteins protruding from it. The matrix (MA) surrounds the inner portion of the membrane. Inside the virion, the capsid proteins together with a cellular protein, cyclophilin A (CypA) form an icosahedral cone encapsulating the viral genome and associated proteins (NC, RT, and IN). Each virion contains two copies of the

single stranded RNA genome, which is surrounded by the nucleocapsid proteins, two copies of the reverse transcriptase, and the integrase protein.

1.3.4.2 The HIV-1 lifecycle

HIV-1 infects CD4⁺ cells like T cells or macrophages (Maddon et al, 1986). The HIV-1 life cycle begins with viral attachment to the host cell by means of the binding of the surface glycoprotein, gp120 to the cellular CD4 molecule. This interaction causes conformational changes in both the envelope glycoproteins, gp120 and gp41, resulting in the exposure of a variable loop in gp120, the V3 loop, which then binds to a chemokine receptor serving as a viral entry co-receptor. The predominant chemokine receptors used by HIV-1 are the CCR5 and the CXCR4. CCR5 usage is predominantly found in early infections, while the CXCR4 usage is mainly found in late-stage infections or after *in vitro* passage (Connor et al, 1997). As stated before, the subtype C viruses are mostly the exception to the co-receptor rule. The interaction of gp120, the CD4 molecule and the co-receptor results in a conformational change in the gp41, exposing a hydrophobic fusion domain, allowing fusion between the viral and cell membranes, releasing the capsid (p24) containing the viral genome into the cellular cytoplasm (Chan & Kim, 1998). CypA facilitates the uncoating of the p24 capsid (Gamble et al, 1996), releasing the nucleocore into the cytoplasm, where the viral RNA is converted into double stranded linear DNA by the reverse transcriptase. The viral DNA, together with the nucleocapsid, matrix, integrase and vpr, form a preintegration complex, which due to the nuclear localisation signals contained in the nucleocapsid and vpr, can be imported into the cellular nucleus via an active uptake mechanism (Bukrinsky et al, 1993; Gallay et al, 1995, Jenkins et al, 1998). Inside the nucleus, the viral integrase causes the linear HIV-1 DNA to be integrated into the host cell chromosome, which then becomes a provirus (Farnet & Bushman, 1996). The provirus contains two copies, one on each side of the HIV-1 genome, of the viral long terminal repeat (LTR), which contains various transcriptional control elements,

including a TATAA box (Okamoto et al, 1990). Viral RNA is synthesised from the provirus and expression is regulated by both cellular and viral factors. The viral protein Tat interacts with cellular kinases to enhance the processivity of RNA polymerase II at the elongation process. In the nucleus, the viral mRNA exists in three forms: doubly spliced mRNA, which codes for Rev, Tat and Nef, singly spliced mRNA, which codes for Env, Vpr, Vpu and Vif, and unspliced mRNA, which codes for Gag-pol. The unspliced and singly spliced mRNA are unstable RNAs, which is not transported to the cytoplasm by cellular factors and would be retained and degraded in the nucleus. Rev however, shuttles the unstable mRNAs through an active process out of the nucleus into the cytoplasm, where cellular proteins translate it into viral proteins and undergo post-translational processes like phosphorylation and glycosylation (Coffin, 1996). The envelope proteins (Env) are translated as a 160 kDa polyprotein, which is cleaved by a cellular protease into a gp120 and gp41. The Env proteins are glycosylated and displayed on the outside of the cellular membrane, with gp120 on the outside and gp41 serving as an anchor within the membrane. The two glycoproteins are non-covalently attached to each other (Chan & Kim, 1998). The Gag protein is a 55 kDa precursor polyprotein, which is cleaved by the viral protease into the p17 matrix, p24 capsid, p7 nucleocapsid, p6 and two spacer peptides. Protease is a product of self-cleavage of a polyprotein, Pol, which also produces reverse transcriptase and integrase. The viral proteins, together with copies of the unspliced mRNA, which will serve as the viral genome, assemble at the cellular membrane, containing the envelope glycoproteins. New HIV-1 virions bud, which can infect new cells and the process repeated. Cleavage of the Gag polyprotein occurs during or after budding (Freed, 1998).

1.3.5 The Gag Protein

The *gag* gene is 1.5 kb in length and is translated into a 55 kDa polyprotein precursor (Pr55^{Gag}), which can produce non-infectious, virus-like particles in the

absence of other viral proteins or packageable viral RNA (Freed, 1998). Pr55^{Gag} plays an important role in virus assembly and budding, while the mature Gag proteins play important roles in disassembly following virus entry into cells. Maturation is a result of proteolytic cleavage of the Pr55^{Gag} by the viral protease into the p17 matrix, p24 capsid, p7 nucleocapsid, p6, p2 and p1 (Figure 1.3).

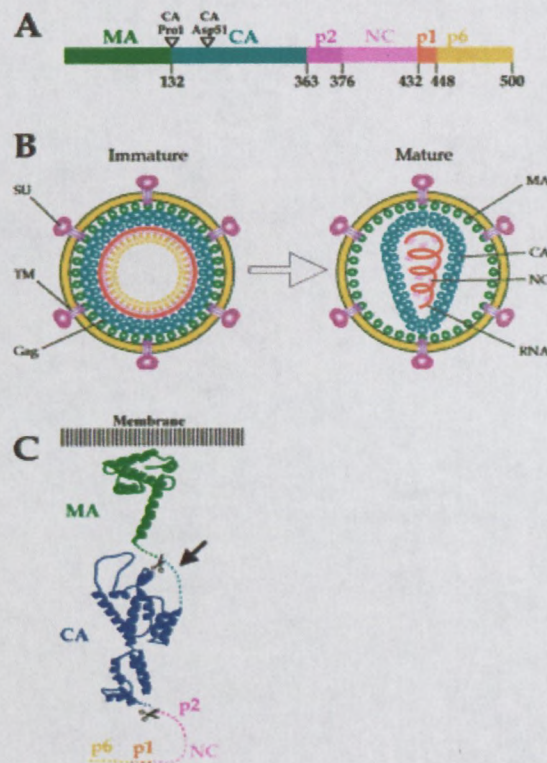


Figure 1.3 An illustration of the genomic (A) and virion layout of the immature and mature Gag proteins. The cleavage sites are indicated in C. (Von Schwedler et al, 1998).

The HIV-1 protease recognises two types of cleavage sites as substrates, namely the class one aromatic*proline cleavage site and the class two hydrophobic*hydrophobic cleavage site (Pettit et al, 1991; Griffiths et al, 1992),

where the cleavage sites are indicated by an asterisk (*). The matrix/ capsid cleavage site falls into the class one type, while the rest of the cleavage sites in the Gag polyprotein fall into the class two type.

1.3.5.1 Matrix

The matrix forms the N-terminal domain of the Pr55^{Gag} and plays a role in targeting of Gag to the plasma membrane, Env glycoprotein incorporation into virions, and early postentry events (Freed, 1998). The mature matrix protein folds into a compact core domain consisting largely of α -helices and three β -sheets and is trimeric. The structure was determined by both nuclear magnetic resonance (NMR) spectroscopy and X-ray crystallography (Massiah et al, 1996). A major function of the immature matrix protein is to direct binding and assembly at the plasma membrane, which is made possible by a myristoylated N-terminal glycine residue (Göttlinger et al, 1989; Bryant & Ratner, 1990) and several domains in the N-terminus as well as in the vicinity of residues 55 and 85 (Freed et al, 1994). The highly basic domain between residue 17 and 31 have also been linked to membrane targeting (Yuan et al, 1993; Zhou et al, 1994, Ono et al, 2000). Upon maturation, the matrix protein undergoes conformational changes, which favour membrane dissociation after viral entry (Scarлата et al, 1998). The matrix protein contains two nuclear localisation signals (NLS), which regulates the nuclear import of the pre-integration complex (PIC) (Haffar et al, 2000). The membrane associated matrix protein plays an important role in the incorporation of Env glycoprotein complex. While single and multiple amino acid mutations can block the incorporation of full-length HIV-1 Env proteins into virions, short cytoplasmic tail Env proteins can still be incorporated (Freed & Martin, 1995, 1996).

1.3.5.2 Capsid

The mature capsid protein forms the capsule around the viral genome and associated proteins. The structure of the capsid has been studied by both NMR spectrophotometry (Gitti et al, 1996) and X-ray crystallography (Gamble et al, 1996; Momany et al, 1996) and have been divided into two domains, namely the core N-terminal domain and the dimerisation C-terminal domain. The core domain functions in the virion maturation and CypA incorporation, while the dimerisation domain contributes to the Gag-Gag interactions and therefore formation of dimers. Both domains are highly helical, with the core domain composed of seven α -helices, two β -sheets and an exposed loop, which serves as a binding site for CypA (Gamble et al, 1996). The core domain also contains eight highly conserved proline residues, which is essential for the assembly of proper functional virion cores (Fitzon et al, 2000). Upon proteolytic cleavage, the core domain refolds into a β -hairpin/ helix structure, which is stabilised by a salt bridge formation between the proline residue at position one (Pro-1) and a highly conserved aspartine residue (Asp-51) (Von Schwedler et al, 1998). A conserved proline rich region within the capsid was also implicated in the binding of the capsid to CypA. Cyclophilin A is a cytoplasmic protein and is a receptor for the immunosuppressant cyclosporine A. It binds Gag in a ratio of one molecule CypA to 10 Gag molecules (Gamble et al, 1996) and may play a role in the destabilisation of the capsid cone after infection (Endrich et al, 1999). Incorporation of CypA requires a glycine-proline motif and is present in the capsid protein in the core domain at position 90 (Luban, 1996) as well as in the dimerisation domain at positions 157 and 224 (Endrich et al, 1999).

The phosphorylation of three serine residues (Ser-109, Ser-149, and Ser-178) within the capsid protein, have also been implicated as being essential for viral uncoating (Cartier et al, 1999). The dimerisation domain contains a region that is highly conserved among different genera of retroviruses called the major homology region (MHR) (Wills & Craven, 1991). Mutations in the MHR showed defects in

virus assembly, maturation and infectivity (Mammano et al, 1994), which might be due to structural hindrance. Looking at the crystal structure of the capsid, the MHR forms a network of hydrogen bonds that stabilise the overall structure of the dimerisation domain (Gamble et al, 1997). The C-terminal region of the MHR form an amphipathic α -helix, which might play a role in the interaction between adjacent Gag monomeres, other viral components or cellular factors to facilitate virus assembly and/ or infection (Clish et al, 1996).

1.3.5.3 Nucleocapsid

The HIV-1 nucleocapsid protein contains two domains with a Cys-X₂-Cys-X₄-His-X₄-Cys (CCHC) motif, where X equals any variable amino acid (Freed, 1998). These two domains are similar to zinc finger domains found in many cellular DNA binding proteins (Berg, 1986) and plays an important role in virion morphogenesis, genomic RNA packaging and viral infectivity, through binding with single-stranded nucleic acid (Freed, 1998). The basic linker domain (²⁹RAPRKKG³⁵) between the two zinc fingers is thought to be important in the spatial proximation of the two domains (Morellet et al, 1994). The nucleocapsid also assists in several steps of reverse transcription, including the stimulation of the binding of the primer for HIV-1 reverse transcription, tRNA^{Lys} to the primer binding site (Freed, 1998).

1.3.5.4 p6 and Spacer peptides

The C-terminal p6 domain of Gag is a proline-rich domain and is involved in late stages of virus maturation, which include the packaging of cleaved Pol proteins in viral particles (Dettenhofer & Yu, 1999). Mutational analyses of the hydrophobic tail of p6 also revealed that mutations within the hydrophobic tail block the incorporation of the Env proteins (Ott et al, 1999). A highly conserved motif (L-X-X-L-F) (Kondo & Göttlinger, 1996) or more specifically, (L-X-S-L-F-G), was found to

play a direct role in the incorporation of Vpr. Point mutations within this motif showed to completely abolish the interaction between Vpr and Pr55^{Gag}, thus preventing the incorporation of Vpr into the virion (Bachand et al, 1999).

The capsid and nucleocapsid, as well as the nucleocapsid and p6 domain are separated by short spacer peptides, p1 and p2 respectively. Although there exists no consensus in the literature to the naming of the spacer peptides (Pettit et al, 1994; Wiegers et al, 1998), this study will refer to the spacer peptide between the capsid and nucleocapsid as p2, while the spacer peptide between the nucleocapsid and p6 domain will be referred to as p1. Assembly of virus particles at the plasma membrane requires several stages of particle development, which involves the multimerisation of the Gag proteins (Tritel & Resh, 2000; Morikawa et al, 2000). It was shown that the p2 domain is essential for ordered assembly as well as for the higher order multimerisation of Gag proteins (Kräusslich et al, 1995; Morikawa et al, 2000). The spacer peptides are also involved in the regulation of the sequential steps of the maturation of the Gag polyprotein (Wiegers et al, 1998; Pettit et al, 1994; Kräusslich et al, 1995).

1.3.5.5 Techniques used to characterise the *gag* gene and Gag protein

1.3.5.5a Cloning and Sequencing

In 1973 Stanley Cohen and his colleagues performed the first cloning experiment (Cohen et al, 1973) and together with Kary Mullis and his colleagues, who discovered the polymerase chain reaction (PCR) (Mullis & Faloona, 1987), empowered us with two of the most powerful techniques in molecular biology. These days, one can isolate a particular gene by using PCR and clone it into vectors, which can be inserted into various organisms to characterise the gene or the protein it expresses.

Cloning involves inserting a foreign gene into a vector that can replicate in the organism to be transformed with the cloned gene. Usually the vector is a plasmid, which is placed into a bacterial strain. Plasmids are circular double-stranded DNA, found in a variety of bacterial species and may carry accessory genes, which can confer resistance to antibiotics, the use of unusual substances as nutrients and/ or production of toxins. They are accessory chromosomes and can replicate autonomously of the host bacteria. Commercially available plasmids contain multiple cloning sites, with various restriction enzyme recognition sites, in which the foreign gene can be inserted. When cloning directional, both the insert and plasmid are cut with a combination of two enzymes; each enzyme with a different recognition site, resulting in only one orientation that the insert can be cloned into. TA cloning is bi-directional and involves the ligation of an insert, usually a PCR product, containing 3' deoxyadenosine (A) residues, into a vector containing 3' deoxythymidine (T) residues resulting in an insert that is inserted in any orientation (Zhou et al, 1995). Once the foreign gene is inserted into the vector, the recombinant vector can be inserted into a bacterial strain. It is important to choose a bacterial strain with a genotype that is complimentary to the plasmid/ vector, as certain enzymes produced by certain bacterial strains can inhibit or interfere with the reproduction of foreign DNA (Hanahan, 1985; Kaiser & Murray, 1985).

Plasmids contain a number of characteristics, important for its existence in the host cell. For replication in bacteria, the plasmid contains an origin of replication site (Ori), which allows the plasmid to replicate independently from the host DNA. For expression of the cloned gene in eukaryotic cells, the plasmid must contain a promoter region and a polyadenylation signal for mRNA stability and effective termination of transcription. It is important that the foreign gene be inserted in the forward orientation in expression vectors, when looking at protein production (Weaver & Philip, 1992).

PCR is the amplification of DNA by using cycles of heat denaturation, primer annealing, incorporation of dNTPs and elongation of DNA strands by a DNA

polymerase (Mullis & Faloona, 1987) and can therefore be seen as an induced form of DNA replication. Depending on the primers chosen, PCR can be very specific and can be used to isolate very specific DNA fragments or even to confirm the identity of DNA fragments.

Sequencing is the ultimate in genetic mapping, whereby the exact base sequence of a DNA fragment can be obtained. The method used for sequencing is a PCR based method, incorporating both dNTPs and dideoxy ribonucleotides (ddNTPs). Where a ddNTP is incorporated, the elongation of the chain is terminated, resulting in various lengths of DNA strands, which can be distinguished by electrophoresis, where the strands are separated according to length (Sambrook et al, 1989). The four different ddNTPs are distinguished from one another by labelling each one with a different dye or separately adding them in four different reactions. The labelled ddNTPs can be done in one reaction and visualised on machines recognising the dyes, such as the ABI Prism 310 Genetic Analyzer (Applied Biosystems, USA), which uses a polymer in an electrophoresis capillary column in which the DNA fragments are separated according to size; the smaller fragments run through faster. A window in the capillary is positioned by a laser, which detects the dye terminators as they pass through and generates an electropherogram, converting the signals into peaks. The ddNTP connected to a particular length strand, correlates to the dNTP at that particular position in the sequence (Swerdlow & Gesteland, 1990). High quality DNA is necessary for sequencing as contaminants, such as RNA, can interfere with the sequencing reactions by acting as random primers, resulting in false peaks on the electropherogram. Contaminating inhibitors of polymerase can also inhibit the sequencing reactions. It is therefore important to use DNA that is free of most contaminants (Sambrook et al, 1989).

1.3.5.5b Molecular Phylogenetics

Molecular phylogenetics can be defined as the study of evolutionary relationships among organisms by using molecular data, such as DNA and protein sequences (Li & Graur, 1991). The comparison of two sequences is referred to as a sequence alignment. Pairs of bases can align as either matched bases, mismatched bases, or as insertions/ deletions. An alignment of two sequences can however not show whether a deletion or insertion had occurred, and the outcome is referred to as a gap or indel. A gap penalty can be assigned to an alignment of sequences, which depends on the frequency of gap occurrence relative to point substitutions. A greater gap penalty will therefore be assigned for sequences with a tendency for insertions or deletions, while a lesser gap penalty will be given to sequences where mismatches are more common.

The process in the evolution of DNA sequences is the change in nucleotides with time. Nucleotide substitutions can be either a transition or a transversion. Transitions are substitutions between the two purines, adenine and guanine, or between the two pyrimidines, thymine and cytosine, while transversions are substitutions between purines and pyrimidines (Weaver & Philip, 1992). There are numerous mathematical models describing the dynamics of nucleotide substitutions, but the two most frequently used, are the Jukes and Cantor's one-parameter model and the Kimura's two-parameter model. The Jukes and Cantor's model makes the assumption that substitutions occur randomly and with no bias among the four nucleotides, which is not realistic in most DNA models (Li & Graur, 1991). Kimura's two-parameter model takes into account that transitions generally occur more frequently than transversions do (Kimura, 1980). Due to the degeneracy of the genetic code, most of the 20 amino acids are encoded by more than one codon, and therefore not all nucleotide substitutions in a codon result in a change in amino acid. A nucleotide substitution that does not change the amino acid is a synonymous substitution, while a substitution resulting in a change in amino acid is a nonsynonymous substitution. Generally, the third position in a

codon is a nonsynonymous site. The rate of nucleotide substitution is defined as the number of substitutions per site per year. For the majority of eukaryotic genes, the synonymous substitution rate is greater than the nonsynonymous substitution rate. The variation in substitution rates is determined by two factors: the rate of mutation and the probability of fixation of a mutation. Both factors indicate a selection process, which depends on whether the mutation is advantageous, neutral, or deleterious. Substitution rates vary between different genes as well as within different regions of genes, depending on the selection pressures the regions are exposed to.

The evolutionary relationships among a group of organisms are illustrated as a phylogenetic tree, which is composed of nodes and branches, in which the nodes represent taxonomic units and the branches the relationship between the taxonomic units in terms of their ancestry. External nodes represent the units to be compared, called operational taxonomic units (OTUs) while the internal nodes represent the ancestral units. Phylogenetic trees can be either rooted, with a common ancestor as a root, or unrooted, which only describes the relationships among the OTUs. There are numerous tree-making methods and most can be divided into two types: the distance matrix methods and maximum parsimony methods. Distance matrix methods first convert aligned sequences into a pairwise distance matrix, then input that matrix into a tree building method, whereas the maximum parsimony methods consider each nucleotide site directly (Page & Holmes, 1998).

The simplest method for tree construction is the unweighted pair group method with arithmetic mean (UPGMA). The UPGMA method assumes that the rates of evolution among the different lineages are approximately constant (Tateno et al, 1982). The number of nucleotide or amino acid substitutions are used to employ a sequential clustering algorithm, whereby the local topological relationships among OTUs depend on similarity and the tree is built in a stepwise manner of pairing and clustering. If the rate of evolution among the lineages are not constant, UPGMA

may give an erroneous topology, but can be corrected by using the transformed distance method, which uses an outgroup as reference and then applies UPGMA to infer the tree topology. An outgroup is an OTU that is known to have diverged from the common ancestor prior to the other OTUs.

One of the most common used tree-making methods is the neighbour-joining method. This method computes a distance matrix and selects neighbours by calculating the distances between pairs of OTUs and selecting the smallest sum. On an unrooted tree, neighbours are connected through a single internal node. The neighbour-joining method sequentially identifies neighbour pairs that minimise the total length of the tree and results in a single, unique tree (Saitou & Nei, 1987).

Another commonly used tree-making method is the maximum parsimony method, which identifies the tree that requires the smallest number of evolutionary changes to explain the differences in the observed OTUs, but often results in more than one tree with the same minimum number of changes found (Fitch, 1971).

The majority of tree-constructing methods result in unrooted trees, which can be rooted with an outgroup. A node is then placed between the outgroup and the rest of the OTUs. Although the outgroup had to diverge prior to the rest of the OTUs from the common ancestor, it should not be too distantly related to the rest of the OTUs, as it might become difficult to obtain reliable estimates of the distances between the outgroup and other OTUs. In the absence of an outgroup, the root can be placed at the midpoint of the longest pathway between two OTUs, which makes the that the rate of evolution has been uniform over all the branches. To evaluate whether the inferred tree topology is reliable, the bootstrap technique is employed. The bootstrap estimates the confidence level by repeatedly resample of data from the original data set (Felsenstein, 1985). Bootstrap values are indicated on internal branches defining clades.

1.3.5.5c *In vitro* transcription/ translation

In vitro transcription/ translation is the synthesis of mRNA and translation thereof into proteins in cell-free extracts. For the synthesis of mRNAs, a bacteriophage DNA-dependant RNA polymerase is used that specifically recognises the promoter, such as the T7 promoter, of the plasmid wherein the gene to be transcribed is cloned. Lysates of rabbit reticulocytes or extracts of wheat germ can be prepared to contain the necessary ribosomes and amino acids needed for translation (Sambrook et al, 1989). Certain commercially available kits exclude certain amino acids, which can be substituted for by labelled amino acids. The labelled amino acids can then be screened for, which simplifies the screening process for the produced protein.

1.3.5.5d Transfection

Most eukaryotic proteins need to undergo post-translational modifications, such as disulphide bond formation, glycosylation, phosphorylation, oligomerisation, or specific proteolytic cleavage, which is not available in the prokaryotic model of bacterial cells. Thus, the expression of proteins from eukaryotic genes is best studied in eukaryotic cells, such as mammalian cells.

The introduction of foreign genes into mammalian cells is termed transfection. Transfection can be stable or transient. Foreign genes are firstly cloned into mammalian expression vectors and then used to transfect the cells. Stable transfections result in clones stably producing the protein coded for by the inserted gene. This is achieved when the vector becomes integrated into the host DNA where it replicates with the host cell. Selection based on negative pressure, like addition of lethal drugs, allows for the identification of cells containing the expression vector conferring protection against the negative pressure. Cells that survive the negative pressure, such as drug treatment, expand into clonal groups,

which can be selected, propagated and analysed. Transient transfection, however does not require the integration of the vector into the host DNA. The expression vectors used contain all the transcription factors necessary for transcription of the foreign gene. All the transient transfected cells are harvested within 72 hours of transfection and screened for protein expression. Expression vectors can also contain genes encoding selectable markers like the chloramphenicol acetyltransferase type 1 protein (CAT), which can be detected and quantified by many commercially available ELISA assays. (Sambrook et al, 1989, Aldovini & Feinberg, 1990)

There are many transfection techniques available, including the calcium phosphate-mediated (CaPO_4) method, electroporation and the use of liposomes. The CaPO_4 method involves the formation of a DNA-calcium precipitate, which is taken up by cells by phagocytosis or endocytosis and transferred to the nucleus (Graham & Van der Eb, 1973). The DNA forms a tight complex with the CaPO_4 and this complex is then resistant to nuclease present in the cell or serum (Loyter et al, 1982). The CaPO_4 method is widely used, because the components used in the method is easily available and the protocol easy to use. Electroporation introduces nucleic acids into cells by using an electric pulse to greatly disturb the cell membrane and form pores through which the nucleic acid can pass (Wong & Neumann, 1982; Shigekawa & Dower, 1988). Cells that are poorly transfected by other methods, or are sensitive to chemicals have been shown to be effectively transfected by electroporation (Aldovini & Feinberg, 1990). The use of liposomes (lipofection) involves the encapsulation of the negatively charged nucleic acid by the cationic lipid molecule (Fraleigh et al, 1980). The lipid-DNA molecule has a positive net charge, which allows it to come into close proximity of the negatively charged cell membrane, followed by internalisation of the DNA complex. Irrespective of the transfection method used, the efficiency is largely dependent on the cell type used, as some cell lines are resistant to transfection by standard methods. It is therefore important to optimise the technique for each cell line.

1.3.5.5e Western Blot

To screen for protein production, the protein preferably has to be in a soluble form. This involves the harvesting of the transfected cells and lysing the cells to release the expressed proteins. It is important that the method of lysis be as gentle on the protein as possible, leaving the protein in an immunoreactive, undegraded and biologically active form. Most cytoplasmic proteins can be effectively solubilised in a lysis buffer containing the nonionic detergent Nonidet P-40 (NP-40) and the addition of protease inhibitors, such as phenylmethylsulfonyl fluoride (PMSF) and ethylenediaminetetra-acetic acid (EDTA) will limit the amount of degradation (Sambrook et al, 1989).

Proteins can then be separated according to size on a polyacrylamide gel. Under denaturing and reducing conditions the proteins are dissociated into their basic polypeptide subunits, allowing more accurate prediction of the protein size. The use of sodium dodecyl sulphate (SDS), β -mercaptoethanol and heat are of the most common substances used for denaturing and reducing conditions (Laemmli, 1970). Under native conditions, the protein exists in its correctly folded, oligomerised structure and migration through a polyacrylamide gel is not entirely dependent on size. The gels can be non-specifically stained to visualise all the proteins in the cell lysate. The electrophoretically separated proteins can also be transferred to a solid support, where specific proteins can be immunologically detected for by using a Western blot assay. The transfer of proteins to a nitrocellulose solid support requires the addition of a current (Towbin et al, 1979). Once the protein is on the membrane, the rest of the membrane is blocked for non-specific binding and then probed with a primary antibody specific for the protein detected for. A secondary antibody directed against the primary antibody, usually species specific, and which is conjugated with an enzyme with an easily detectable biological activity, is then incubated with the membrane. Lastly, a substrate for the

conjugated enzyme is added, resulting in a reaction, which can be visualised as bands (Sambrook et al, 1989).

1.3.6 HIV-1 vaccines strategies

1.3.6.1 Immunology

The immune system responds against infectious agents as a humoral response and a cell-mediated response. The humoral response involves the production of secreted antibodies by the B lymphocytes, which recognises and destroys antigens in the extracellular fluid. Cell-mediated responses induce helper T lymphocytes (CD4+ cells) that stimulates antibody responses and cytotoxic T lymphocytes (CTLs)/ (CD8+ cells). CTLs recognise intracellularly processed antigens displayed on the major histocompatibility complex (MHC) of an infected cell and lyse the infected cell, before new antigens/ virus can be produced within that cell. Both of the humoral and cell-mediated immune responses can be manipulated in the development of vaccines.

1.3.6.1a Humoral Immunity

Shortly after HIV-1 infection, the body produces antibodies against different viral proteins and can be detected by a variety of techniques, including ELISA. Most of the antibodies produced are however not neutralising and do not protect the host from further infection or disease. The neutralising epitopes of the HIV-1 virus are located on the gp120 envelope protein and are protected by the variable loops V1 and V2 (Sanders et al, 2000; Cao et al, 1997), which shows a high percentage of diversity in sequence. Neutralising antibodies directed against the native gp120 protein tend to protect against homologous challenge, but not against heterologous challenge. Recent studies are looking at different constructs and modifications of

the Env protein to direct antibody productions against the more conservative parts of the protein (Stamatatos & Cheng-Mayer, 1998; Barnett et al, 2001).

1.3.6.1.b Cell mediated Immunity

HIV-1 specific CTL responses have been shown to be important in controlling the virus in HIV-1 infected individuals in the acute stage of infection (Borrow et al, 1994, Koup et al, 1994). It also appears that a strong CTL response correlates inversely to the viral load in infected patients (Clerici et al, 1996; Musey et al, 1997). Indirect evidence of CTL control of the HIV-1 virus, comes from highly exposed, HIV-seronegative persons who display CTL responses against HIV (Langlade-Demoyen et al, 1994; Rowland-Jones et al, 1993, 1999), and the late seroconversion of previously HIV-resistant prostitutes after the waning of their HIV-specific CD8+ responses due to reduced antigenic exposure (Kaul et al, 2001). More specifically, there appears to be a relationship between HIV-1 Gag-specific CTL and disease progression (Klein et al, 1995). Gag is one of the more conserved HIV-1 proteins and broad, cross-clade CTL responses have been detected in HIV infected people (Durali et al, 1998; McAdam et al, 1998; Buseyne et al, 2001). According to the latest Los Alamos Immunology compendium, there are currently 187 published Gag CTL epitopes. There are 56 p17 CTL epitopes spanning three areas of the protein, while the 127 epitopes found in the p24 domain, cover the entire region, excluding a small area of 10 amino acids. The nucleocapsid, p6 and spacer peptide domain contains four published CTL epitopes spanning three areas (Korber et al, 2000). It has been shown that HIV-specific CTL responses can be induced in humans by immunisation with a DNA vaccine in the form of a plasmid (Calarota et al, 1998, 1999; MacGregor et al, 1998).

1.3.6.2 Types of Vaccines

Two main types of vaccines have been developed in the past to stimulate protective immune responses against viral pathogens. These two types of

vaccines are firstly, the inactivated whole virus vaccines, which have been successfully used to protect against poliomyelitis and influenza, and secondly, the live attenuated viruses, which are infectious but do not cause disease and have been used as the Sabin oral polio vaccine and the vaccine for mumps, rubella, and yellow fever (Rovinski & Klein, 1994). It is however thought not to be wise to use these classical types of vaccines for HIV-1, as there is the risk of integration of the viral genome into the host genome, or the risk of back mutations resulting in active viruses (Vermund et al, 1994). Current approaches involve genetically engineered HIV-1 vaccines, which include the recombinant envelope-based subunit vaccines, non-infectious HIV-1 like particles, live recombinant vaccines and DNA vaccines. Live recombinant vaccines use for example, the recombinant vaccinia viruses or the *Mycobacterium bovis bacillus* Calmette-Guerin (BCG) as vectors, into which HIV-1 genes are inserted (Rovinski & Klein, 1994). There are several types of DNA vaccines, which mainly differs on the gene inserted and whether the whole gene or epitopes are used.

Considering the dominance of the subtype C virus within South Africa, the potential vaccine for South Africa should preferably be derived from the subtype C strains (Morris et al, 1997; Esparza & Bhamarapavati, 2000). The current vaccine trials around the world are of subtype B or E origin (Esparza & Bhamarapavati, 2000). Table 1.1 summarises the current vaccine trials in developing countries. Only one vaccine candidate has so far progressed to a phase III trial, which tests for efficacy of the vaccine. The outcome of the phase III trial will become known within the year 2002. Although there are studies done on developing subtype C vaccines, none have reached a phase I trial yet.

Table 1.1

HIV-1 preventive vaccine trials in less-developed countries*

Starting date	Candidate vaccine	Subtype	Country	Number of volunteers
Phase I/II				
1993	Synthetic peptide MN-V3	B	China	23
1994	Synthetic peptide MN-V3	B	Thailand	24
	Synthetic peptide MN-V3	B	Brazil	30
1995	Envelope gp120	B	Thailand	30
	Envelope gp120	B	Thailand	52
1996	Recombinant V3 protein	B	Cuba	30
1997	Envelope gp120	B, E, B/E	Thailand	380
1998	Envelope bivalent gp120	B/E	Thailand	90
1999	Canarypox vector	B	Uganda	40
2000	Prime-boost canarypox vector plus gp160 or gp120	E + E	Thailand	130
	Prime-boost canarypox vector plus gp160 or gp120	E + B/E	Thailand	125
Phase III				
1999	Envelope bivalent gp120	B/E	Thailand	2500

* Esparza & Bhamarapravati, 2000

CHAPTER TWO

Materials and Methods

	Page
2.1 Materials	31
2.1.1 HIV-1 Isolates	31
2.1.2 Plasmids	31
2.1.2.1 pCR [®] 3.1 (Invitrogen, USA)	31
2.1.2.2 pcDNA3.1 (Invitrogen, USA)	32
2.1.2.3 pCR [®] 3.1/CAT (Invitrogen, USA)	32
2.1.3 Competent Bacterial strain (Top10F')	33
2.1.4 Cell lines	33
2.1.4.1 Vero cell line	33
2.1.4.2 CV-1 cell line	34
2.2 Methods	34
2.2.1 Characterisation studies	34
2.2.1.1 Amplification of <i>gag</i>	34
2.2.1.1a Templates for the amplification of <i>gag</i>	34
2.2.1.1a(i) Near full-length fragment of the HIV-1 genome	34
2.2.1.1a(ii) RNA	34
2.2.1.1b Optimisation of PCR conditions	35
2.2.1.1c PCR Reactions	37
2.2.1.2 Cloning of <i>gag</i>	37
2.2.1.2a Cloning into pCR [®] 3.1 (Invitrogen, USA)	37
2.2.1.2b Cloning into pcDNA3.1(-) (Invitrogen, USA)	37
2.2.1.2b(i) Restriction Digest	38
2.2.1.2b(ii) Ligation and Transformation Reactions	38
2.2.1.2c Screening for recombinants	39
2.2.1.3 Sequencing of <i>gag</i>	40
2.2.1.3a Preparation of high quality DNA for sequencing	40

2.2.1.3b	Sequencing	41
2.2.1.3c	Sequencing Data Analysis	42
2.2.1.4	DNA Quality Assurance	44
2.2.2	Expression studies	44
2.2.2.1	<i>In vitro</i> transcription/ translation	44
2.2.2.2	Screening for Mycoplasma contamination in cell lines	45
2.2.2.3	Transfections	46
2.2.2.3a	Preparation of endotoxin free plasmid DNA for transfections	46
2.2.2.3b	Transfection techniques	46
2.2.2.3b(i)	Tfx™ Transfection (Promega, USA)	46
2.2.2.3b(ii)	Calcium Phosphate (CaPO ₄) Mediated Transfection	46
2.2.2.3b(iii)	Electroporation	47
2.2.2.3c	Antigen Preparation	47
2.2.2.3d	Evaluation of the Transfection Techniques	48
2.2.2.4	SDS-Polyacrylamide Gel Electrophoresis (SDS-PAGE)	48
2.2.2.5	Western Blot	49
	Tables	50

CHAPTER TWO

2. Materials and Methods

2.1 Materials

2.1.1 HIV-1 Isolates

HIV-1 was previously isolated from positive patients attending the Infectious Disease Clinic at the Tygerberg hospital during the period from 1998 to 1999. The subtype of the virus was determined by serotyping and 18 isolates (15 HIV-1 subtype C and three HIV-1 subtype B) were chosen to be expanded to high concentrations virus stock (Claassen, 1999). Briefly, the peripheral blood mononuclear cells (PBMCs) of the HIV-1 positive patients were isolated, co-cultured with uninfected donor PBMCs and the HIV-1 virus allowed to grow to high concentrations (Gartner & Popovic, 1990). The PBMCs were lysed and high molecular weight DNA isolated (Sambrook et al, 1989). The clinical data and characteristics of the 18 patients are summarised in Table 2.1. The HIV-1 isolates were named according to their respective patients from whom they were isolated. There is no TV011. The isolates TV015 and TV016 were also genotyped as HIV-1 subtype B isolates by sequencing a ~650 bp fragment in the gp120 protein, including the V3 loop (Treurnicht et al, in press).

2.1.2 Plasmids

2.1.2.1 pCR[®]3.1 (Invitrogen, USA)

The vector pCR[®]3.1 is a 5060 bp plasmid and is a mammalian expression vector. The vector is supplied linearised for TA-cloning in the Eukaryotic TA cloning kit (Invitrogen, USA). A map and summary of the features of pCR[®]3.1 is given in Appendix A1. The vector contains a cytomegalovirus (CMV) promoter, ensuring

high-level expression of the cloned gene and a bovine growth hormone (BGH) polyadenylation signal that improves mRNA stability and effective termination. pCR[®]3.1 confers resistance to the antibiotics kanamycin, neomycin as well as ampicillin. For replication and maintenance in *E. coli*, the vector contains a pUC origin. The insert can be sequenced by using the BGH reverse priming site and T7 primer binding site. The T7 promoter can also be used in *in vitro* transcription and translation assays.

2.1.2.2 pcDNA3.1 (Invitrogen, USA)

For expression studies in mammalian cells, the gene to be studied must be in the forward orientation in the expression vector. The isolates of which no positive clones with inserts in the forward orientation in pCR[®]3.1 were obtained, were directionally cloned into pcDNA3.1. The plasmid pcDNA3.1 is a mammalian expression vector of 5428 bp. A map and summary of the features of pcDNA3.1 is given in Appendix A2. pcDNA3.1 has the same basic features as described for the pCR[®]3.1, only without the kanamycin resistance gene. There are two versions of the pcDNA3.1 vector, which differs in the orientation of the multiple cloning site. In this study we used the pcDNA3.1(-) vector.

2.1.2.3 pCR[®]3.1/CAT (Invitrogen, USA)

To evaluate the different transfection techniques, cells were transfected with the plasmid, pCR[®]3.1/CAT (Invitrogen, USA). A map and summary of the vector pCR[®]3.1/CAT is given in Appendix A3. The pCR[®]3.1/CAT vector is a 5809 bp plasmid containing the gene that encodes for chloramphenicol acetyltransferase type 1 (CAT). Except for the CAT insert, the plasmid is identical to the vector pCR[®]3.1 (Invitrogen, USA) described earlier.

2.1.3 Competent Bacterial strain (Top10F')

As recommended by the suppliers of the plasmids, the competent bacterial strain, Top 10F' was used in the transformation reactions using pCR[®]3.1 and pcDNA3.1. A list of genotypes and their corresponding phenotypes is given in Appendix A4.

The genotype of the Top10F' competent cells is:

F'{*lacI^qTn10(Tet^R)*} *mcrA*Δ(*mrr-hsdRMS-mcrBC*) ϕ 80*lacZ*ΔM15 Δ*lacX74*
deoR recA1 araD139 Δ(*ara-leu*)7697 *galU galK rpsL endA1 nupG*

2.1.4 Cell lines

2.1.4.1 Vero cell line

The Vero cell line, an adherent epithelial cell line of the African green monkey kidney was obtained from W. Hann and J.S. Rhim through the American Type Culture Collection, USA (ATCC), ATCC number CCL-81. The Vero cell line was cultured in Dulbecco's modified eagle's medium (DMEM) (Sigma-Aldrich, UK) supplemented with 10% heat inactivated fetal bovine serum (FBS), 100 units/ml penicillin and 100 μg/ml streptomycin (complete medium) and incubated at 37°C. The media was changed every three to four days and the cells subcultured once a week. Subculturing or trypsinisation was done by incubating the cells in a 0.25% active trypsin versine (ATV) solution (135 mM NaCl, 5 mM KCl, 5.6 mM Glucose, 7 mM NaHCO₃, 0.5% Trypsin, 0.5 mM EDTA) for two minutes or until the cells detached from the flask wall. The cells were then centrifuged at 1 000 g for five minutes, the trypsin removed, and the cells seeded at 3 x 10⁵ cells per 75 mm cell culture flask in complete medium (Freshney, 1987).

2.1.4.2 CV-1 cell line

The CV-1 cell line, an adherent fibroblast cell line of the African green monkey kidney, was obtained by F.C. Jensen (Jensen, 1964) through ATCC, number CCL-70. The CV-1 cell line was cultured and subcultured as described for the Vero cell line.

2.2 Methods

2.2.1 Characterisation Studies

2.2.1.1 Amplification of *gag*

2.2.1.1a Templates for the amplification of *gag*

2.2.1.1a(i) Near full-length fragment of the HIV-1 genome

A near full-length ~9 kb fragment of the HIV-1 genome was previously amplified by Dr. S Engelbrecht. The ~9 kb fragment was amplified by PCR from high molecular weight (HMW) DNA with the primer set UP1A and Low2 (Gao et al, 1998) and the Expand Long Template kit (Roche, Germany). The PCR products were separated by agarose gel electrophoresis, the ~9 kb fragment excised and purified using the QIAEX II Gel Extraction kit (Qiagen, Germany).

2.2.1.1a(ii) RNA

RNA was purified from supernatant fluid (SNF) of the last passage of viral culture, before HMW DNA isolation (2.1.1.1a(i)), using the QIAamp[®] Viral RNA mini kit (Qiagen, Germany) according to the manufacturer's instructions. The RNA was converted to cDNA by using the Access RT-PCR system (Promega, USA)

according to the manufacturer's guidelines, using only the AMV reverse transcriptase enzyme, without the Tfl DNA polymerase in a 50 µl reaction. The reaction was incubated at 48°C for a period of 45 minutes and the reverse transcriptase was inactivated by an incubation period of three minutes at 94°C.

2.2.1.1b Optimisation of PCR conditions

A nested PCR was done, using the ~ 9kb fragment as a pre-nested template. The primer pair used for the amplification of the *gag* gene, was GagF (5'GCTAGAAGGTCTAGAATGGGTGCGAGAGCG3'), with a melting temperature (T_m) of 67.21°C, and GagR (5'AGTTGCCCCCGAATTCTTATTGTGACGAGG3'), with a T_m of 66.65°C (Qiu *et al*, 1999). Both primers have restriction enzyme cleavage sites, shown by the underlined sections. GagF has a *Xba*I cleavage site and GagR an *Eco*RI cleavage site.

The Expand High Fidelity PCR System (Roche, Germany) was used for the amplification. The $MgCl_2$ concentration was optimised for the primer pair, using a titration of 1.5 mM, 2 mM and 3 mM $MgCl_2$ in the reactions. The enzyme, primer and nucleotide concentrations were used according to the manufacturer's instructions.

The PCR temperature cycle used was as follows: one cycle of denaturing at 94°C for five minutes, primer annealing at 59°C for one minute and elongation at 72°C for five minutes. This was followed by 30 cycles of denaturing at 94°C for one minute, primer annealing at 59°C for one minute and elongation at 72°C for five minutes. A final elongation was allowed at 72°C for 15 minutes after which the reaction was cooled down to 4 °C for an indefinite period, until the reaction tubes were removed from the PCR machine.

A 3 μ l aliquot of the PCR sample was added to 2 μ l of sample loading buffer (0.25% bromophenol blue, 0.25% xylene cyanol FF, 15% Ficoll) and diluted with water to a final volume of 12 μ l. The mixture was loaded into a slot of a 16 well 1% agarose minigel in TAE buffer (0.04 M Tris acetate, 0.001 M EDTA), containing 0.5 μ g/ml ethidium bromide. The DNA molecular weight marker, Lambda/Hind III (Stratagene, California) was used as a size marker. The dimensions of the minigel electrophoresis chamber were 10 cm by 10cm and was run at a constant voltage of 5 Volts/cm. The DNA bands were visualised under an ultra violet light at a wavelength of 302 nm and photographed with the Syngene GeneGenius (Synoptics Ltd, UK). Non-specific bands were identified as background bands other than the ~ 1.5 kb band. The ratio of the concentration of smaller non-specific bands compared to the concentration of the ~ 1.5 kb bands were determined by using the GeneTools version 2.10.02 (Synoptics Ltd, UK) program.

To confirm that the amplified products were indeed of HIV-1 origin and more specifically of *gag*, the products were hybridised with the 41 bp digoxigenin labeled oligo probe, SK19 (5'-ATCCTGGGATTAATAAAATAGTAAGAATGTATAGCCCTAC-3') specific for HIV-1 *gag* (Ou et al, 1988). The probe was labeled using the DIG oligonucleotide 3'-end labeling kit (Roche, Germany) according to the manufacturer's instructions. The detection was in the form of spot blots. Briefly, 2 μ l of the PCR product was mixed with 8 μ l denaturation buffer (4 N NaOH, 250 mM EDTA) and spotted 2.5 μ l at a time on a Hybond™-N gridded nylon membrane (Amersham Life Science, UK). Eight μ l of the denaturation buffer was used as a negative control. The membrane was air-dried and the nucleic acid fixed with ultra violet light using a Spectrolinker XL-1500 UV crosslinker (Spectronics Corporation, USA) at 120 millijoules per unit area. The hybridisation and detection were performed according to the manufacturer's instructions supplied in the DIG nucleic acid detection kit (Roche, Germany) using a pre-hybridisation and hybridisation temperature of 55°C.

2.2.1.1c PCR Reactions

Two μl of the purified ~ 9 kb fragment and 5 μl of the cDNA were used in subsequent PCR reactions, using the optimised conditions (2.2.1.1b). The PCR products were loaded on a 1% agarose gel and electrophoresis and visualisation performed as described earlier. The concentration ratios for smaller non-specific bands to the ~ 1.5 kb bands were measured with the GeneTools version 2.10.02 (Synoptics Ltd, UK) program. Spot blot hybridisations, with an HIV-1 specific probe as described earlier, were performed on all the PCR products. The products were then purified using the QIAquick PCR purification kit (Qiagen, Germany).

2.2.1.2 Cloning of *gag*

2.2.1.2a Cloning into pCR[®]3.1 (Invitrogen, USA)

The Eukaryotic TA cloning[®] kit (Invitrogen, USA) was used to clone the amplified *gag* PCR products into the plasmid vector pCR[®]3.1. The ligation reactions were done according to the manufacturer's guidelines. Briefly, 4 μl of the purified PCR product was ligated into 2 μl of the pre-cut vector pCR[®]3.1. The ligation reaction was incubated overnight at 14°C. Top 10F' competent cells (Invitrogen, USA) were transformed with the ligation reactions according to the manufacturer's instructions. Fifty μl and 100 μl of the transformed bacterial cells were plated on Luria-Bertani medium (LB) plates (10 g/l bacto-tryptone, 5 g/l bacto-yeast extract, 10 g/l NaCl, 15 g/l bacto agar) containing 50 $\mu\text{g}/\text{ml}$ of the antibiotic, kanamycin. The plates were incubated at 27°C overnight. After the incubation period, the plates were screened for recombinant colonies.

2.2.1.2b Cloning into pcDNA3.1(-) (Invitrogen, USA)

2.2.1.2b(i) Restriction Digest

For directional cloning into the vector pcDNA3.1(-) (Invitrogen, USA), both the enzymes *EcoRI* and *XbaI* were used to digest the PCR products and the vector. Twelve units of each enzyme were used to digest 1 µl of purified PCR product, in the presence of buffer H (0.9 M Tris-Cl, pH 7.5, 0.5 M NaCl, 0.1 M MgCl₂) and 2 µg of acetylated bovine serum albumin (BSA) at 37°C for four hours. A 5 µl aliquot of the digested samples was run on a 1% agarose gel to establish if the chosen restriction enzymes only cut in the primer sequences and that the PCR product did not contain a recognition site for *EcoRI* or *XbaI*. One µl of the PCR products that had recognition sites for *XbaI* or *EcoRI* were digested with 12 units of each enzyme separately and run on a 1% agarose gel as described earlier, to determine which enzyme cleaved the product.

Two µg of pcDNA3.1 was digested with either 12 units each of *EcoRI* and *XbaI* or 12 units each of *EcoRI* and *EcoRV*, under the same conditions as earlier. To remove unwanted smaller digested fragments of DNA, the samples were purified through a Centri-sep sephadex spin column (Princeton Separations, USA).

2.2.1.2b(ii) Ligation and Transformation Reactions

PCR products digested with both *XbaI* and *EcoRI* were ligated into the vector, pcDNA digested with both *XbaI* and *EcoRI*. The PCR product digested with only *EcoRI* was ligated into the vector pcDNA3.1 digested with *EcoRI* and *EcoRV*. Briefly, 2 µl of the digested vector and 4 µl of the digested PCR product was ligated using 4.0 Weiss units of T4 ligase (Invitrogen, USA) in the presence of 1 x ligation buffer (6 mM Tris-Cl, pH 7.5, 6 mM MgCl₂, 5 mM NaCl, 0.1 mg/ml BSA, 7 mM β-mercaptoethanol, 0.1 mM ATP, 2 mM dithiothreitol, 1 mM spermidine). The ligase

reaction was incubated at 14°C overnight. Top 10F' competent cells (Invitrogen, USA) were transformed with the ligation mixture according to the manufacturer's instructions. Fifty µl and 100 µl of the transformed bacterial cells were plated on LB plates (10 g/l bacto-tryptone, 5 g/l bacto-yeast extract, 10 g/l NaCl, 15 g/l bacto agar) containing 50 µg/ml of the antibiotic, ampicillin. The plates were incubated at 27°C overnight. After the incubation period, the plates were screened for recombinant colonies.

2.2.1.2c Screening for recombinants

The transformed bacterial colonies were screened for recombinants with HIV-1 *gag* inserts. The protocol followed was the alkaline lysis method as described in small-scale preparations of plasmid DNA (Sambrook *et al*, 1989). The isolated plasmid DNA (miniprep DNA) was loaded on a 1% agarose gel, and electrophoresis and visualisation performed as described earlier (2.2.1.1b). The clones were evaluated by size, as a plasmid containing a 1.5 kb insert would be significantly larger than the original plasmid. The recombinant plasmids were named according to the plasmid, the isolate number and the number of the miniprep culture, eg. pTV001G15 was cloned into the vector pCR[®]3.1, from the isolate TV001 and was the fifteenth miniprep DNA isolated on the day after cloning, while pcTV004G18 was cloned into pcDNA3.1, from the isolate TV004 and the eighteenth miniprep DNA isolated.

To confirm that the larger plasmids contained a copy of HIV-1 *gag*, denatured plasmid DNA was spot blotted on a Hybond[™]-N Gridded nylon membrane (Amersham Life Science, UK) and hybridised as described earlier. The spot blot method requires an overnight incubation and is time consuming. The confirmation was therefore also done by using PCR incorporating HIV-1 *gag* specific primers. PCR confirmation was performed when a set of 20 or less recombinants were screened, while the spot blot confirmation was performed for the screening of more

than 20 recombinants at a time. The PCR amplified a 160 bp fragment, using *Taq* DNA polymerase (Promega, USA) and the primers HIVgagA (5'-AGAGAACCAAGGGGAAGTGA-3') with a T_m of 56.56°C and HIVgagB (5'-TCTCTAAAGGGTTCCTTTGG-3') with a T_m of 52.59°C (Kemp et al, 1989). The PCR cycle performed was with two minutes of denaturation at 96°C, followed by 30 cycles of one minute at 94°C, one minute at 44°C and one minute at 72°C. A further one minute elongation at 72°C was allowed before ending the cycle at 4°C. The amplified products were loaded on a 2% agarose gel, and electrophoresis and visualisation performed as described earlier (2.2.1.1b).

The orientation of the insert in pCR[®]3.1 was determined by digesting 1 µl of miniprep DNA with 6 units of *EcoRI* in buffer H (0.9 M Tris-Cl, pH 7.5, 0.5 M NaCl, 0.1 M MgCl₂) and in the presence of 2 µg of acetylated bovine serum albumin at 37°C for four hours. Ten µl of the digestion reaction was loaded onto a 1% agarose gel, and electrophoresis and visualisation performed as described earlier (2.2.1.1b).

2.2.1.3 Sequencing of *gag*

2.2.1.3a Preparation of high quality DNA for sequencing

High quality DNA was prepared by using the Qiagen plasmid midi kits (Qiagen, Germany). The manufacturer's instructions, with a few modifications were used. Briefly, 100 ml LB medium (10 g/l bacto-tryptone, 5 g/l bacto-yeast extract, 10 g/l NaCl, 15 g/l) containing 50 µg/ml of either kanamycin or ampicillin (depending on plasmid) was inoculated with 200 µl of an overnight culture and incubated at 27°C overnight. The bacterial cells were pelleted and resuspended in 12 ml Buffer P1 (50 mM Tris-Cl, pH 8.0, 10 mM EDTA, 100 µg/ml Rnase A). The cells were lysed by gentle mixing with 12 ml of Buffer P2 (200 mM NaOH, 1% SDS). The cell lysate was then neutralized with 12 ml of chilled Buffer P3 (3.0 M potassium acetate, pH

5.5) and the genomic DNA, proteins and cell debris precipitated by incubation on ice for 30 minutes. The precipitate was removed by centrifugation at 20 000 x g for 30 minutes and the cleared cell lysate placed in a clean tube. Ten ml of the lysate was applied to an equilibrated Qiagen-tip 100 and allowed to flow through the resin by using vacuum suction. The column was washed twice with 10 ml of Buffer QC (1.0 mM NaCl, 50 mM MOPS, pH 7.0, 15% isopropanol) and the DNA eluted with 5 ml of Buffer QF (1.25 M NaCl, 50 mM Tris, pH 8.5, 15% isopropanol). The column was equilibrated and the remaining cell lysate applied, washed and eluted in two subsequent applications. The DNA was precipitated in 30 ml isopropanol, centrifuged at 20 000 x g for 45 minutes, washed in one ml 70% ethanol and placed in a microfuge tube. The DNA was pelleted at 13 000 g in a microcentrifuge, dried and dissolved in 50 µl to 150 µl of deionised water.

The concentration of the DNA was determined by measuring the optical density of a diluted sample of the plasmid DNA at 260 nm (OD_{260}). The concentration was then calculated by using the formula, $OD_{260} \div 20 \times \text{dilution factor}$. The optical density was measured in a Spectronic[®] Genesys 5 spectrophotometer. To determine the purity of the plasmid DNA, the optical density was measured at 260 nm (OD_{260}) and 280 nm (OD_{280}). The purity of the DNA was calculated as a ratio of optical densities at 260 nm and 280 nm (Sambrook et al, 1989). A value between 1.7 and 1.9 indicated pure DNA, suitable for sequencing.

2.2.1.3b Sequencing

Sequencing was done on the ABI Prism 310 Genetic Analyzer (Applied Biosystems, USA) using the BigDye[™] terminator cycle sequencing ready reaction DNA sequencing kit (Applied Biosystems, UK). The BigDye[™] terminator cycle sequencing is a PCR based system using AmpliTaq DNA Polymerase, FS, and incorporates both dye labeled dideoxynucleoside-triphosphates (ddNTPs) and unlabeled dideoxynucleoside-triphosphates (dNTPs).

The manufacturer's instructions for sequencing plasmids and PCR products were used to prepare the sequencing reactions. Briefly, four μl of the terminator ready reaction mix, 400 ng of plasmid DNA and five pmol of the primer were added to a final volume of 10 μl . Seven primers were used to sequence the whole 1.5 kb *gag*. Table 2.2 summarises the primers used for sequencing. The positions of T7 and BGH reverse are also illustrated on the map of pCR[®]3.1 (Appendix A1). A minimum of four primers were used to sequence the whole gene. The cycle sequencing was performed by repeating 25 cycles of 10 seconds denaturation at 96°C, annealing for five seconds at either 45°C or 59°C (depending on the primer) and followed by a four minute elongation at 60°C. The reaction was cooled down to 4°C until the purification was done. The unincorporated dye terminators interfere with the analysis and must therefore be removed. The sequencing reactions were purified through Centri-Sep[™] spin columns (Princeton Separations, USA) and the manufacturer's instructions were used. The samples were allowed to dry in a vacuum centrifuge and resuspended in 25 μl of Template Suppression reagent (Applied Biosystems, USA). The resuspended samples were denatured at 95°C for five minutes. The denatured sample were then loaded in the ABI Prism 310 Genetic Analyzer (Applied Biosystems, USA) according to the manufacturer's instructions.

2.2.1.3c Sequencing Data Analysis

The ABI Prism 310 Genetic Analyzer sequence data was converted to an electropherogram using the DNA Sequencing Analysis software, version 3.3 (Applied Biosystems, USA). The electropherogram was analysed in Chromas, version 1.43 (McCarthy, Australia) and the nucleotide sequence analysed, assembled and translated in DNAMAN, version 4.0 (Lynnon BioSoft, Canada). To evaluate whether our subtype C sequences could be subtyped using a restriction fragment length polymorphism (RFLP) analysis (Van Harmelen et al, 1999b), the

sequences were analysed for *AluI* and *Accl* recognition sites, which was performed using DNAMAN, version 4.0 (Lynnon BioSoft, Canada). The clone sequences were also analysed for unique, which only cut once, and non-cutting restriction enzyme recognition sites.

The NCBI HIV-1 subtyping tool (<http://www3.ncbi.nlm.nih.gov/retroviruses/subtype/makepage.cgi?page=sub&type=0>) was used to subtype the *gag* sequences, which were compared to the set of reference sequences available in the database. Possible recombinants were analysed by Bootscan analyses using the SimPlot, version 2.5 software package (Ray, 1999). Sequences of clones identified as possible recombinants in the subtyping tool were analysed against the subtype B strains, K03455, M17451, and our subtype B TV016 isolate.

Nucleotide sequences were aligned using the CLUSTAL X (1.8) software program (Thompson et al, 1997). The alignment of the nucleotide and predicted amino acid sequences were viewed and analysed in GeneDoc (Nicholas & Nicholas, 1997). Published mutational analyses and CTL epitopes were compared to the alignment of the predicted amino acid sequences. Phylogenetic trees were constructed using the nucleotide alignments. Kimura-2-parameter distance calculation, neighbour-joining tree construction, and bootstrap analysis were done with TREECON for Windows, version 1.3b (Van de Peer, 1994). Construction of a phylogenetic tree with all the HIV-1 groups were done by comparing the *gag* clones to the reference sequence set in the Los Alamos database and rooted with the HIV-1 group O strains. To construct a subtype C tree, the *gag* clone sequences were compared to the full length HIV-1 subtype C *gag* sequences displayed in Table 2.3 and rooted with the subtype B strains. Hydrophilic and hydrophobic graphs were constructed in DNAMAN, version 4.0 (Lynnon BioSoft, Canada), using the predicted amino acid sequences of the *gag* clones. The graphs were used to establish the maintenance of the hydrophilic or hydrophobic profiles of different areas of the Gag protein domains. The amphipathic α -helix of the MHR in the capsid domain, was reconstructed using the predicted amino acid sequences of the *gag* clones, to

establish whether the helix was maintained in our sequences. The original α -helix was derived from a subtype B sequence (Clish et al, 1996).

2.2.1.4 DNA Quality Assurance

RNA was isolated from the original patient plasma as well as from the earliest SNF available. A ~160 bp fragment was amplified by PCR using the primers GagA and GagB (Table 2.2), and sequenced. Two Neighbour-joining phylogenetic trees were constructed using the programs as described earlier (2.2.1.3c). The first tree was constructed using all three codon positions, while the second tree was constructed using synonymous substitutions. The tree was rooted with the sequences of the subtype B isolates.

2.2.2 Expression Studies

2.2.2.1 *In vitro* transcription/ translation

The TNT[®] Quick Coupled Transcription/ Translation System (Promega, USA) together with the Transcend[™] Non-Radioactive Translation Detection System (Promega, USA) was used to verify whether the *gag* clones could produce proteins.

The transcription/ translation master mix contained RNA polymerase, nucleotides, salts, ribonuclease inhibitors and a reticulocyte lysate solution. The T7 RNA polymerase was used as all the *gag* inserts were cloned into vectors, downstream from a T7 promoter. DNA of the clones was prepared by the standard alkaline lysis method (Sambrook *et al*, 1989), the DNA concentration determined as described earlier and a total of 1 μ g of DNA was used in each reaction. Half of the recommended volumes were used, therefore 20 μ l of TNT[®] Quick Master Mix was

used and the final volume of each reaction was 25 μ l. The reactions were incubated at 30°C for 90 minutes after which the detection was done.

The detection system used, incorporated biotinylated lysine residues into the protein during translation. Five μ l of the translated mixture was added to 20 μ l of SDS sample buffer (125 mM Tris, 4% SDS, 20% Glycerol, 1% 2-Mercaptoethanol) and loaded on a 10% Tris-glycine SDS-polyacrylamide gel. The electrophoresis was performed at a constant voltage of 150 Volts for two hours in a "Mighty Small" slab gel electrophoresis unit SE 250 (Hoefer Scientific Instruments, Ca, USA) containing electrophoresis buffer (0.025 M Tris pH 8.3, 0.192 M glycine, 0.1% SDS) in both the upper and lower chambers. The proteins were transferred to a Hybond ECL nitrocellulose membrane (Amersham Life Science, UK) using the Mighty Small™ tank transfer unit (Hoefer Scientific Instruments, USA) filled with the Towbin transfer buffer (25 mM Tris, 192 mM Glycine, 10% methanol) (Towbin et al, 1979) at 100 Volts for one hour. Non-specific binding sites on the membrane were blocked by incubation with TBS-T (20 mM Tris-Cl, pH 7.5, 150 mM NaCl, 0.5% Tween-20) and the rest of the detection was done according to the manufacturer's instructions. The translated proteins were colorimetricly detected after incubation with a streptavidin-alkaline phosphatase conjugate.

2.2.2.2 Screening for Mycoplasma contamination in cell lines

A nested PCR assay, to detect any mycoplasma contamination in the cell cultures used for transfections, was performed by Ms M. Claasen or Mrs L. Wilsdorf. The Mycoplasma detection kit (ATCC) was used to screen the cell cultures for mycoplasma contamination. In a mycoplasma contaminated cell culture, the detection kit produces a PCR product ranging in size from 236 bp to 365 bp, depending on the *Mycoplasma* species present in the contaminated culture (Harasawa et al, 1993; Tang et al, 2000). The supplier's instructions were followed. Screening for mycoplasma contamination is performed routinely by technologists on all cell lines in use in the Department.

2.2.2.3 Transfections

2.2.2.3a Preparation of Endotoxin free plasmid DNA for transfections

The removal of endotoxins from DNA can improve transfection efficiency into mammalian cells, especially in sensitive cells. The EndoFree plasmid Maxi protocol (Qiagen, Germany) was used according to the manufacturer's instructions. The DNA concentration and purity was determined as described earlier.

2.2.2.3b Transfection Techniques

2.2.2.b(i) Tfx™ Transfection (Promega, USA)

The Tfx™ reagents contain a synthetic cationic lipid molecule [N,N,N',N'-tetramethyl-N,N'-bis(2-hydroxy-ethyl)-2,3-di(oleoyloxy)-1,4-butanediammonium iodide]. There are three different Tfx™ reagents available: Tfx™-10, Tfx™-20 and Tfx™-50 and the manufacturer's instructions were used in the transfection reactions. The Tfx™-10 reagents were used for transfecting Vero cells and Tfx™-50 for CV-1 cells, which was according to the guidelines given in the Tfx™ reagents kit protocol (Promega, USA). A range of DNA concentrations was used, ranging from 0.25 µg to 1.00µg of endotoxin free plasmid DNA. The Tfx™ reagent to DNA ratio was also optimized, using a 4: 1 and 2:1 ratio in reactions.

2.2.2.3b(ii) Calcium Phosphate (CaPO₄) Mediated Transfection

The standard protocol for CaPO₄ transfection of adherent cells as described in (Sambrook *et al*, 1989) was followed. Briefly, cells were trypsinised and plated at 2 x 10⁶ cells per 60 mm culture flask. The cultures were incubated for 24 hours at 37°C in a humidified incubator with 5% CO₂ atmosphere. After 24 hours the cells

were directly exposed to DNA precipitate containing 8.8 µg of endotoxin free plasmid DNA, HEPES-buffered saline (HBS) (140 mM NaCl, 5 mM KCl, 0.75 mM Na₂HPO₄·2H₂O, 6 mM dextrose, 25 mM HEPES) and 124 mM CaCl₂ for 15 minutes at room temperature. Seven ml of pre-warmed complete medium was then added to the cells and incubated for 16 to 18 hours at 37°C in a humidified incubator with a 5% CO₂ atmosphere. After 18 hours the cells were washed once with phosphate-buffered saline (PBS) (80 mM Na₂HPO₄, 20 mM NaH₂PO₄, 100 mM NaCl, pH 7.5) and seven ml of pre-warmed complete medium was added and incubated for a further 60 hours in a humidified incubator with a 5% CO₂ atmosphere at 37°C.

2.2.2.3b(iii) Electroporation

The Gene Pulser II system (Bio-Rad, USA) with the Capacitance Extender PLUS module (Bio-Rad, USA) was used for the electroporation. Cells were trypsinised and seeded at 2 x 10⁶ cells/ml in serum-free DMEM medium (Sigma-Aldrich, UK) containing 100 units/ml penicillin and 100 µg/ml streptomycin. Five hundred µl of the cells were added together with 10 µg of endotoxin free plasmid DNA to a 0.4 cm cuvette and placed into the shocking chamber of the electroporator. An electric pulse of 0.3 Volts was applied to the cuvette at a capacitance of 1000 µF. The electroporated cells were immediately added to 1.5 ml warm complete medium in one well of a six well plate and incubated for 72 hours at 37°C in a humidified incubator with a 5% CO₂ atmosphere.

2.2.2.3c Antigen preparation

The transfected cells were harvested 72 hours post transfection. The supernatant fluid (SNF) was removed and the cells washed twice with ice cold PBS (80 mM Na₂HPO₄, 20 mM NaH₂PO₄, 100 mM NaCl, pH 7.5). One ml or 500 µl of lysis

buffer (150 mM NaCl, 10 mM Tris-Cl, pH 7.4, 1 mM EDTA, 1% NP-40, 1 mM PMSF, 1% Sodium deoxycholate, 0.1% SDS, 0.02% Sodium azide) was added to one 60 mm flask or one well of a 6 well plate respectively and incubated on ice for 30 minutes. The cell lysate was then centrifuged in a microcentrifuge at 15 000 x g for 10 minutes, after which the clear SNF was placed into a clean microfuge tube and stored at -20°C until used in further applications.

2.2.2.3d Evaluation of the Transfection Techniques

The levels of CAT expression in the transfected cells were determined by using a CAT ELISA (Roche, Germany). The manufacturer's instructions were followed.

2.2.2.4 SDS-Polyacrylamide Gel Electrophoresis (SDS-PAGE)

Fifteen µl of cell lysate was added to 15 µl of SDS sample buffer (125 mM Tris, 4% SDS, 20% Glycerol, 1% 2-Mercaptoethanol), denatured and loaded on a 10% Tris-glycine SDS-polyacrylamide gel. The full range Rainbow molecular weight marker (Amersham Life Science, UK) was used as a size standard on the gel and HIV-1_{IIIIB}p55Gag was added as a positive control. The HIV-1_{IIIIB}p55Gag was obtained from Mr Steve Showalter and Ms Maria Garcia-Moll (BioMolecular Technology) through the NIH AIDS Research and Reference Reagent Program, DIADS, NIAID, NIH. The electrophoresis was performed at a constant voltage of 200 Volts for 1.5 hours in a Hoefer miniVE vertical electrophoresis system (Amersham Pharmacia Biotech, USA) containing electrophoresis buffer (0.025 M Tris pH 8.3, 0.192 M glycine, 0.1% SDS) in both the upper buffer chamber and tank.

2.2.2.5 Western Blot

After the SDS-polyacrylamide electrophoresis was complete, the proteins were transferred to a Hybond ECL nitrocellulose membrane (Amersham Life Science, UK) using the Hoefer miniVE Blot module (Amersham Pharmacia Biotech, USA) according to the manufacturer's instructions. Briefly, the module containing the gel and membrane sandwich, was filled with the Towbin transfer buffer (25 mM Tris, 192 mM Glycine, 10% methanol) and the tank filled with chilled distilled water. Transfer was performed at 25 Volts for one hour. Non-specific binding sites were blocked with TBS-T (20 mM Tris-Cl, pH 7.5, 150 mM NaCl, 0.5% Tween-20) containing 5% skim milk for at least one hour at room temperature. The membrane was washed twice in TBS-T (20 mM Tris-Cl, pH 7.5, 150 mM NaCl, 0.5% Tween-20) and incubated in either a 1/100 diluted human HIV-1 positive serum or a 1/1000 diluted sheep antiserum to HIV-1 p24 for at least one hour. The antiserum to HIV-1 p24 was obtained from Dr. Michael Phelan through the AIDS Research and Reference Reagent Program, Division of AIDS, NIAID, NIH (Karacostas et al, 1989). The membrane was washed three times for five minutes each with TBS-T (20 mM Tris-Cl, pH 7.5, 150 mM NaCl, 0.5% Tween-20) and incubated for an hour with a 1/10 000 diluted either anti-sheep or anti-human secondary antibody. The anti-sheep secondary antibodies were horseradish-peroxidase conjugated and the anti-human antibodies were either alkaline phosphatase- or horseradish-peroxidase conjugated. The membrane was washed three times of five minutes each with TBS-T (20 mM Tris-Cl, pH 7.5, 150 mM NaCl, 0.5% Tween-20). For the horseradish-peroxidase detection, the ECL Plus kit (Amersham Pharmacia biotech, USA) was used according to the manufacturer's instructions. The Western Blue[®] stabilized substrate (Promega, USA) was used according to the manufacturer's instructions for the detection of proteins conjugated with alkaline phosphatase.

Table 2.1

Demographics of 18 HIV-1 infected patients

Patient	Age	Gender/ Risk factor	Clinical Symptoms	CD4 count	Serotype	Source of infection
TV001	36	Male/ Het ^a	Dermatitis	196	C	Unkown
TV002	35	Male/ Het ^a	Oral Candidiosis	309	C	Unkown
TV003	38	Male/ Het ^a	Pulmonary TB	117	C	Transkei/ CT ^d
TV004	38	Female/ Het ^a	TB adenitis	435	C	Unkown
TV005	36	Female/ Het ^a	Oral/tracheal Candidiosis	64	C	Kraaifontein
TV006	26	Female/ Het ^a	TB pleuritis	116	C	Bellville
TV007	24	Female/ Het ^a	Dermatitis	363	C	Gauteng
TV008	31	Female/ Het ^a	Bladder infection	125	C	Zimbabwe
TV009	31	Male/ Het ^a	Pulmonary TB	171	C	Mfuleni
TV010	21	Female/ Het ^a	Lymphadenopathy	167	C	Khayelitsha
TV012	32	Male/ Het ^a	Oral/nasopharyngeal Candidiosis, diarrhea	3	C	Unkown
TV013	43	Female/ Het ^a	Pneumonia	765	C	Unkown
TV014	33	Male/ Het ^a	Pulmonary TB	unknown	C	Namibia
TV015	28	Male/ Hom ^b	Wasting, stomach ulcers	420	No reaction	Unkown
TV016	39	Male/ Bi ^c	Loose stools	421	B/D	Unkown
TV017	22	Male/ Het ^a	Glandular fever, Candidiosis, wasting	200	B	Unkown
TV018	56	Male/ Het ^a	Pulmonary TB	201	C	Transkei/ CT ^d
TV019	37	Female/ Het ^a	Asymptomatic	100	C	Unkown

^a Heterosexual, ^b Homosexual, ^c Bisexual, ^d Cape Town

Table 2.2

Sequencing primers

Primer	Sequence (5' – 3')	Orientation	Reference
GagF	GCTAGAAGGTCTAGAATGGGT GCGAGAGCG	Forward	Qiu et al, 1999
GagR	AGTTGCCCCCGAATTCTTATT GTGACGAGG	Reverse	Qiu et al, 1999
GagA	AGAGAACCAAGGGGAAGTGA	Forward	Kemp et al, 1989
GagB	TCTCTAAAGGGTTCCTTTGG	Reverse	Kemp et al, 1989
G40	GACACCAAGGAAGCTTTAGA	Forward	Sanders-Buell et al, 1995
G60	CAGCCAAAATTACCCTATAGT GCAG	Forward	Sanders-Buell et al, 1995
G80	ATGAGAGAACCAAGGGGAAGT GA	Forward	Sanders-Buell et al, 1995
T7 primer	TAATACGACTCACTATAGGG	Forward	Eukaryotic TA Cloning © Kit (Bidirectional) <i>Version H</i>
BGH Reverse primer	TAGAAGGCACAGTCGAGG	Reverse	Eukaryotic TA Cloning © Kit (Bidirectional) <i>Version H</i>

Table 2.3Reference sequences of full-length HIV-1 subtype C *gag*.

Name	Acc. no.	Description	Year	Reference
92BR025	U52953	HIV-1, isolate from Brazil	1992	Gao et al, 1996
98BR004	AF286228	HIV-1 strain from Brazil	1998	Rodenburg et al, 2001
96BW06H51	AF290027	HIV-1 isolate 96BW06 from Botswana	1996	Ndung'u et al, 2000
BW.MJ4	AF321523	HIV-1 infectious molecular clone MJ4 from Botswana		Ndung'u et al, 2001
96.BW01B03	AF110959	HIV-1 isolate 96BW01, clone B03 from Botswana	1996	Novitsky et al, 1999
96BW01B21	AF110960	HIV-1 isolate 96BW01, clone B21 from Botswana	1996	Novitsky et al, 1999
96BW01B22	AF110961	HIV-1 isolate 96BW01, clone B22 from Botswana	1996	Novitsky et al, 1999
96BW0402	AF110962	HIV-1 isolate 96BW04, clone 02 from Botswana	1996	Novitsky et al, 1999
96BW0407	AF110963	HIV-1 isolate 96BW04, clone 07 from Botswana	1996	Novitsky et al, 1999
96BW0408	AF110964	HIV-1 isolate 96BW04, clone 08 from Botswana	1996	Novitsky et al, 1999
96BW0409	AF110965	HIV-1 isolate 96BW04, clone 09 from Botswana	1996	Novitsky et al, 1999
96BW0410	AF110966	HIV-1 isolate 96BW04, clone 10 from Botswana	1996	Novitsky et al, 1999
96BW0504	AF110968	HIV-1 isolate 96BW05, clone 04 from Botswana	1996	Novitsky et al, 1999
96BW1104	AF110969	HIV-1 isolate 96BW11, clone 04 from Botswana	1996	Novitsky et al, 1999
96BW1106	AF110970	HIV-1 isolate 96BW11, clone 06 from Botswana	1996	Novitsky et al, 1999
96BW11B01	AF110971	HIV-1 isolate 96BW11, clone B01 from Botswana	1996	Novitsky et al, 1999
96BW1210	AF110972	HIV-1 isolate 96BW12, clone 10 from Botswana	1996	Novitsky et al, 1999

96BW15B03	AF110973	HIV-1 isolate 96BW15, clone 03 from Botswana	1996	Novitsky et al, 1999
96BW15C02	AF110974	HIV-1 isolate 96BW15, clone C02 from Botswana	1996	Novitsky et al, 1999
96BW15C05	AF110975	HIV-1 isolate 96BW15, clone C05 from Botswana	1996	Novitsky et al, 1999
96BW16B01	AF110976	HIV-1 isolate 96BW16, clone B01 from Botswana	1996	Novitsky et al, 1999
96BW16D14	AF110977	HIV-1 isolate 96BW16, clone D14 from Botswana	1996	Novitsky et al, 1999
96BW1626	AF110978	HIV-1 isolate 96BW16, clone 26 from Botswana	1996	Novitsky et al, 1999
96BW17A09	AF110979	HIV-1 isolate 96BW17, clone A09 from Botswana	1996	Novitsky et al, 1999
96BW17B03	AF110980	HIV-1 isolate 96BW17, clone B03 from Botswana	1996	Novitsky et al, 1999
96BW17B05	AF110981	HIV-1 isolate 96BW17, clone B05 from Botswana	1996	Novitsky et al, 1999
96BW0502	AF110967	HIV-1 isolate 96BW05, clone 02 from Botswana	1996	Novitsky et al, 1999
96BW06K18	AF290030	HIV-1 isolate 96BW06, clone K18 from Botswana	1996	Ndung'u et al, 2000
96BW06J7	AF290029	HIV-1 isolate 96BW06, clone J7 from Botswana	1996	Ndung'u et al, 2000
96BW06J4	AF290028	HIV-1 isolate 96BW06, clone J4 from Botswana	1996	Ndung'u et al, 2000
86.ETH2220	U46016	HIV-1 isolate C2220 from Ethiopia	1986	Salminen et al, 1996
98IS002	AF286233	HIV-1 strain 98IS002 from Israel	1998	Rodenburg et al, 2001
IN.AF209990	AF209990	HIV-1 isolate C-Gag-221 from India		Gupta et al, 2001
93IN904	AF067157	HIV-1 isolate 301904 from India	1993	Lole et al, 1999
93IN905	AF067158	HIV-1 isolate 301905 from India	1993	Lole et al, 1999
93IN101	AB023804	HIV-1 subtype C genomic RNA	1993	Mochizuki et al, 1999
93IN999	AF067154	HIV-1 isolate 301999 from India	1993	Lole et al, 1999
94IN476	AF286223	HIV-1 strain 94IN476 from India	1994	Rodenburg et al, 2001
94IN11246	AF067159	HIV-1 isolate 11246 from India	1994	Lole et al, 1999
95IN21068	AF067155	HIV-1 isolate 21068 from India	1995	Lole et al, 1999

98IN012	AF286231	HIV-1 strain 98IN012 from India	1998	Rodenburg et al, 2001
98IN022	AF286232	HIV-1 strain 98IN022 from India	1998	Rodenburg et al, 2001
98TZ013	AF286234	HIV-1 strain 98TZ013 from Tanzania	1998	Rodenburg et al, 2001
98TZ017	AF286235	HIV-1 strain 98TZ017 from Tanzania	1998	Rodenburg et al, 2001
97ZA012	AF286227	HIV-1 strain 97ZA012 from South Africa	1997	Rodenburg et al, 2001
96ZM651	AF286224	HIV-1 strain 96ZM651 from Zambia	1996	Rodenburg et al, 2001
96ZM751	AF286225	HIV-1 strain 96ZM751 from Zambia	1996	Rodenburg et al, 2001

3.1.1 Amplification of gag

3.1.1.1 Optimisation of PCR conditions

3.1.1.2 PCR results of the different templates

3.1.1.2a Near full-length ~ 9 kb fragment as template

3.1.1.2b RNA as template

3.1.2 Cloning of gag

3.1.2.1 Cloning into pCR3.1

3.1.2.1a Screening for recombinants

3.1.2.1b Determination of the orientation of clones

3.1.2.2 Cloning into pODNA3 (p)

3.1.2.2a Restriction digest for downstream cloning

3.1.2.2b Screening for recombinants

3.1.3 Sequencing of gag

3.1.3.1 Sequencing Data

3.1.3.2 Sequencing Data Analyses

3.1.3.2a Nucleotide Sequence Analyses

3.1.3.2a(i) Restriction Digest Analyses

3.1.3.2a(ii) Subtyping

3.1.3.2a(iii) Phylogenetic Analyses

3.1.3.2b Predicted Amino Acid Sequence Analyses

3.1.4 DNA Quality Assurance

3.2 Expression Studies

3.2.1 *In vitro* Transcription/Translation

3.2.2 Cell culture and screening for myxomatoma contamination

CHAPTER THREE

3. Results

	Page
3.1 Characterisation studies	57
3.1.1 Amplification of <i>gag</i>	57
3.1.1.1 Optimisation of PCR conditions	57
3.1.1.2 PCR results of the different templates	58
3.1.1.2a Near full-length ~ 9 kb fragment as template	58
3.1.1.2b RNA as template	58
3.1.2 Cloning of <i>gag</i>	59
3.1.2.1 Cloning into pCR [®] 3.1	59
3.1.2.1a Screening for recombinants	59
3.1.2.1b Determination of the orientation of the inserts	59
3.1.2.2 Cloning into pcDNA3.1(-)	60
3.1.2.2a Restriction digest for directional cloning	60
3.1.2.2b Screening for recombinants	60
3.1.3 Sequencing of <i>gag</i>	61
3.1.3.1 Sequencing Data	61
3.1.3.2 Sequencing Data Analyses	61
3.1.3.2a Nucleotide Sequence Analyses	61
3.1.3.2a(i) Restriction Digest Analyses	61
3.1.3.2a(ii) Subtyping	62
3.1.3.2a(iii) Phylogenetic Analyses	62
3.1.3.2b Predicted Amino Acid Sequence Analyses	63
3.1.4 DNA Quality Assurance	63
3.2 Expression Studies	64
3.2.1 <i>In vitro</i> Transcription/ translation	64
3.2.2 Cell culture and screening for mycoplasma contamination	64

3.2.3	Evaluation of the transfection techniques	65
3.2.4	Transfection	65
	Tables	66
	Figures	72

3. Results

3.1. Characterisation Studies

3.1.1. Amplification of gag

3.1.1.1. Optimisation of PCR conditions

To optimise the conditions of the PCR reactions, a titration of DNA concentrations were used in reactions using the ~ 8 kb fragments of HIV-1 isolates as templates. The PCR products were visualised on a 1% agarose gel with two rows of 16 wells (Figure 3.1). All the PCR products in all wells at all concentrations yielded a strong band of approximately 1.5 kb in length. All the 2 mM MgCl₂ concentrations showed little or no visible background on the agarose gel. The PCR product of the isolate TV013, displayed a smaller non-specific band at a concentration ratio of 1:5 when compared to the ~ 1.5 kb band. The PCR products produced in 3 mM MgCl₂ (Mix 4) showed smearing possibly caused by non-specific primer binding. Two mM MgCl₂ produced PCR products with the least amount of background. The subsequent PCR reactions were done using a MgCl₂ concentration of 2 mM.

To confirm that the amplified PCR products were HIV-1 gag, a spot blot of the PCR products of the 2 mM MgCl₂ titration, were done and hybridised with an HIV-1 gag specific probe (Figure 3.2). All the PCR products were positive and the negative control was negative.

CHAPTER THREE

3. Results

3.1 Characterisation Studies

3.1.1 Amplification of *gag*

3.1.1.1 Optimisation of PCR conditions

To optimise the conditions of the PCR reactions, a titration of three different MgCl_2 concentrations were used in reactions using the ~ 9 kb fragments of six different HIV-1 isolates as templates. The PCR products were visualised on a 1% agarose gel with two rows of 16 wells (Figure 3.1). All six PCR products in all three MgCl_2 concentrations yielded a strong band of approximately 1.5 kb in length. All three MgCl_2 concentrations showed little or no visible background on the agarose gel. The PCR product of the isolate TV013, displayed a smaller non-specific band with a concentration ratio of 1:5 when compared to the ~ 1.5 kb band. The PCR products produced in 3 mM MgCl_2 (Mix 4) showed smearing possibly caused by non-specific primer binding. Two mM MgCl_2 produced PCR products with the least amount of background. The subsequent PCR reactions were done using a MgCl_2 concentration of 2 mM.

To confirm that the amplified PCR products were HIV-1 *gag*, a spot blot of the PCR products of the 2 mM MgCl_2 titration, were done and hybridised with an HIV-1 *gag* specific probe (Figure 3.2). All the PCR products were positive and the negative control was negative.

3.1.1.2 PCR results of the different templates

3.1.1.2a Near full-length ~ 9 kb fragment as template

PCR reactions, using the ~ 9 kb fragment as template were performed three times on separate occasions. The purpose of the first set of reactions was to optimise the PCR conditions (3.1.1.1). The second group of reactions was done to obtain PCR products for cloning into pCR[®]3.1 (Figure 3.3). Two of the products displayed smaller background bands, indicated by perforated arrows. The ratio of the non-specific band to the ~ 1.5 kb band for the PCR product of the isolate TV003 was 1:2, while the ratios of the two non-specific bands compared to the ~ 1.5 kb band for the product of the TV007 isolate, were 1:10 and 1:14 respectively. The third set of PCR reactions was performed to obtain PCR products for cloning into pcDNA3.1 (Figure 3.4). The PCR product of the isolate TV007, displayed a larger band, indicated by a perforated arrow, with a concentration ratio of 1:7 when compared to the ~ 1.5 kb band. All PCR products of the three sets were positive upon hybridisation with an HIV-1 *gag* specific probe (data not shown).

3.1.1.2b RNA as template

There were no suitable ~9kb PCR products for the isolates TV009 and TV014, and RNA was therefore used as templates. Faint bands, in the region of the ~ 1.5 kb positive band of the positive control, were observed for both isolates (Figure 3.5). Both PCR products were positive upon hybridisation with an HIV-1 *gag* specific probe. The PCR products were used in directional cloning into pcDNA3.1.

3.1.2 Cloning of gag

3.1.2.1 Cloning into pCR[®]3.1

3.1.2.1a Screening for recombinants

Clones were screened for inserts by evaluating the size of their plasmids. Plasmids with inserts were larger than the 5 kb native pCR[®]3.1 and therefore migrated slower on an agarose gel. Figure 3.6 is an example of such an evaluation. A plasmid containing a 1.5 kb *gag* insert would be ~6.5 kb in size. On a 1% agarose gel, the difference in migration between a 5 kb plasmid and a 6.5 kb plasmid is sufficient for selection. Figure 3.7 and Figure 3.8 are examples of the two confirmation processes, which were used to identify HIV-1 *gag* positive clones. Figure 3.7 illustrates an electrophoretic analysis of PCR confirmation, while Figure 3.8 illustrates confirmation by hybridisation. Table 3.1 displays the number of HIV-1 *gag* positive clones obtained for the different isolates.

3.1.2.1b Determination of the orientation of the inserts

The *gag* positive clones were digested with the restriction enzyme, *Xba*I to determine the orientation of the inserts. The enzyme, *Xba*I cut in the forward primer, GagF, and in the multiple cloning site, 43 bp downstream of the insert. Clones with inserts in the forward orientation would give a restriction pattern of a ~5 kb band and a ~1.5 kb band on a 1% agarose gel. Such patterns can be seen in Figures 3.9 (A and B) in lanes 3, 6, 15, 16, 21, 22, 23, 24, 26, 28, 45 and 51, which represent the following clones respectively: pTV012G34, pTV012G40, pTV019G20, pTV019G51, pTV013G2, pTV013G8, pTV013G11, pTV013G15, pTV001G8, pTV001G11, pTV009G97, pTV018G60. Clones with inserts in the

reverse orientation display a *Xba*I restriction pattern of a ~6.5 kb band without the smaller band on a 1% agarose gel. The number of HIV-1 positive clones with inserts in the forward orientation, in the vector pCR[®]3.1, is displayed in Table 3.1.

3.1.2.2 Cloning into pcDNA3.1(-)

3.1.2.2a Restriction digest for directional cloning

Purified PCR products of the isolates TV002, TV003, TV004, TV005, TV007, TV008, TV009, TV010, and TV014, as well as the vector, pcDNA3.1(-) were digested with the restriction enzymes *Xba*I and *Eco*RI. The digested PCR product of the isolate TV003 migrated faster through the agarose gel than the other products (Figure 3.10), indicating that the TV003 product was smaller than the rest. Figure 3.4 however showed that the undigested TV003 PCR product was the same size as the rest of the PCR products, suggesting that TV003 contained an *Eco*RI and/or *Xba*I cleavage site. To determine which of the restriction enzymes cleaved the product of TV003, restriction digestions of both enzymes, separate and together, were performed (Figure 3.11).

3.1.2.2b Screening for recombinants

Screening and confirmation of clones were done as described for pCR[®]3.1 (Figure 3.7 and Figure 3.8). The number of HIV-1 *gag* positive clones for the different isolates is given in Table 3.2.

3.1.1 Sequencing of *gag*

3.1.3.1 Sequencing Data

The full length *gag* sequences were assembled for the clones TV001G8, TV001G11, TV002G8, TV003G15, TV004G17, TV004G24, TV005G29, TV005G36, TV006G11, TV006G97, TV007G59, TV008G65, TV008G66, TV009G12, TV010G74, TV012G34, TV012G40, TV013G2, TV013G15, TV014G73, TV018G60, TV019G20, TV019G25. The nucleotide and amino acid sequences are listed in Appendix B1. All sequences, except for both clones of the isolate TV006, contained a single open reading frame (ORF) with no early stop codons. The clone TV006G11 contained a two nucleotide deletion in the beginning of the *gag* sequence, resulting in a frame shift. The clone TV006G97 contained an early stop codon, resulting in a truncated protein. Table 3.3 summarises the size of the different *gag* clones and their nucleotide compositions, which will be used to explain the instability of *gag* mRNA transcripts.

3.1.3.2 Sequencing Data Analyses

3.1.3.2a Nucleotide Sequence Analyses

3.1.3.2a(i) Restriction Digest Analyses

The sequences of the *gag* clones were analysed for recognition sites for the restriction enzymes, *AluI* and *Accl*. The restriction maps for the different *gag* clone sequences are given in Appendix B2.1. The unique cutting restriction enzymes as well as the non-cutting enzymes were identified for all the clone sequences. The list of unique and non-cutting enzymes for each sequence can be found in Appendix B2.2.

3.1.3.2a(ii) Subtyping

The results of the NCBI HIV-1 subtyping tool are given in Appendix B3, while representative plots are given in Figure 3.12. Fifteen of the isolates were identified as subtype C and one isolate, TV016 as subtype B. The similarity plot is colour coordinated with subtype C being a green bar and subtype B a red bar. Figures 3.12 (A & B) are representatives of subtype C sequences. The similarity plots for the isolate TV006 (Figures 3.12 C & D) are different from the rest of the subtype C similarity plots, which shows one window of 300 bp that is more similar to a subtype B. Bootscan graphs of the clones TV006G11 and TV006G97, constructed in SimPlot, are given in Figures 3.13 (A-F). Each bootscan graph shows the comparison of the TV006 sequence to two subtype C sequences (TV001G8 and TV004F17), and one subtype B sequence (K03455 or M17451 or TV016G95). A recombination breakpoint can be seen in Figures 3.13 A and C, at positions 600 bp to 1000 bp, where the subtype B sequence of K03455 in this region shows the highest percentage of permuted trees and both the subtype C reference sequences show a sharp drop in percentage permuted trees. This ~ 400 bp recombination area is not seen in the other bootscan graphs where the subtype B sequences are M17451 or TV016G95.

3.1.3.2a(iii) Phylogenetics

Phylogenetic analysis of the *gag* clone sequences compared to a set of HIV-1 reference sequences, is given in the form of a rooted phylogenetic tree (Figure 3.14). This tree confirms the NCBI subtyping, which shows that our 15 isolates cluster with the reference subtype C sequences and the isolate, TV016, clusters with the subtype B sequences. The rooted phylogenetic tree in Figure 3.15 displays the relationship between the different subtype C strains (Table 2.3) and our isolates. The distance matrices are given in Appendix B for the complete *gag*

gene (Appendix B4.1), the matrix (Appendix B4.2), the capsid (Appendix B4.3), the region between the nucleocapsid and p6 (Appendix B4.4), as well as for the three regions within the matrix containing published CTL epitopes (Appendix B4.5). The distance matrices for the CTL epitope areas, were calculated on the amino acid alignments of these areas. The rest of the matrices were calculated on the nucleotide alignments. The data obtained from the distance matrices are also summarised in Table 3.4, which shows the minimum and maximum variation found within the different areas. The variation was calculated as the difference between 100% and the percentage identical residues between two sequences.

3.1.3.2b Predicted Amino Acid Sequence Analyses

The amino acid alignment (Appendix B5) shows the variation of the amino acid sequences between the different subtype C strains, our isolates and two subtype B strains, of which one is the clone of the isolate TV016. Dots indicate identical sequence to the reference sequence, TV001G11, while mismatches are indicated by the amino acid at that particular position. Dashes indicate indels.

The hydrophilicity and hydrophobic graphs of the clones are displayed in Figures 3.16 (A & B) and Figures 3.16 (C & D) respectively. The hydrophobic helix of the major homology region found in the capsid domain is displayed in Figure 3.17. The amino acid sequence within the helix is as was found in the subtype C clones.

3.1.3.3 DNA Quality Assurance

A rooted phylogenetic tree (Figure 3.18 A & B) displays the relationship between the clone sequences and the sequences of the plasma and/or SNF. Figure 3.18A was calculated using all alignment positions, while Figure 3.18B was calculated

only using the synonymous substitutions. Bootstrap values above 50 % are indicated.

3.2 Expression studies

3.2.1 *In vitro* transcription/ translation

To determine whether the clones with inserts in the forward orientation produced proteins, an *in vitro* transcription/ translation was performed on them. The clones TV001G8, TV001G11, TV002G8, TV003G15, TV004G24, TV005G36, TV006G11, TV006G97, TV007G59, TV008G66, TV009G12, TV010G74, TV012G34, TV012G40, TV013G2, TV013G15, TV014G73, TV016G95, TV018G60, and TV019G20, were subjected to this translation system and the proteins were detected by Western blot. Figures 3.19 (A-E) represent the blots. The Western blots represented by Figures 3.19 (A) and (B) were done at the same time, while (C) and (D) were done a week later. A coloured band in the region of 55 kDa was considered as positive, and therefore the clones TV002G8, TV013G15, TV004G24, TV008G66, TV001G11, TV009G12 and TV014G73 were considered to be protein-producing clones. The *in vitro* transcription/ translation of the clone TV001G11 seen in Figure 3.20 (E) was performed with DNA that was older than four months. All the other translations were performed with DNA that was not older than two weeks.

3.2.2 Cell culture and screening for mycoplasma contamination

Both the Vero and CV-1 cell lines grew adherent. The Vero cells were fast growers and a monolayer of cells was achieved within two days after subculture. The CV-1 cells grew slower and monolayers were only achieved by day three or four after

subculture. All the cell lines used were negative for mycoplasma contamination and remained negative with routine testing thereafter.

3.2.3 Evaluation of the transfection techniques

The cell lines, Vero and CV-1, were transfected with the vector pCR[®]3.1/CAT, using the CaPO₄, Tfx[™] and electroporation methods. The transfection methods were evaluated on the concentration of the CAT protein produced. Table 3.5 summarises the results of the evaluation of the Tfx[™] method, while Table 3.6 summarises the results of the evaluation of the CaPO₄ and electroporation methods. The concentrations of the CAT production of the different transfection techniques were also plotted on graphs (Figure 3.20 A & B).

3.2.4 Transfection

Endotoxin free DNA was prepared of the clones TV001G11, TV002G8, TV004G24, TV008G66, TV009G12, TV013G15, and TV014G73 and were transfected into the Vero cell line, using the electroporation technique. A faint band above the 50 kDa size marker and corresponding height to the positive control was observed in the lanes representing the clones TV008G66 and TV013G8 (Figure 3.21 A). In a following Western blot, the positive control failed to be detected (data not shown). The positive *in vitro* transcription/ translation sample of the clone, TV009G12 was used as a positive control in subsequent Western blots, which was immunodetected with goat antibodies specific for HIV-1 Gag. The detection was done using ECL detection (Figure 3.21 B). Both of the clones, TV009G12 and TV014G73 were positive for production of a ~55 kDa protein, which correlated to a band in the positive control.

Table 3.1The number of HIV-1 *gag* positive pCR[®]3.1 clones

ISOLATE	No. minipreps	No. Positive clones	No. clones in forward orientation
TV001	20	4	2
TV002	30	3	0
TV004	17	5	0
TV005	12	1	0
TV006	44	6	1
TV007	37	3	0
TV008	74	5	0
TV012	40	5	2
TV013	9	4	4
TV018	5	2	1
TV019	60	6	2

Table 3.2

The number of HIV-1 gag positive pcDNA3.1 clones

ISOLATE	No. Minipreps	No. Positive clones
TV002	10	2
TV003	30	3
TV004	10	5
TV005	10	3
TV006	30	4
TV007	10	2
TV008	10	3
TV009	32	1
TV010	10	3
TV014	53	1

Table 3.3The nucleotide and amino acid composition of the HIV-1 subtype C *gag* clones

Clone	Size in bp ^c	A ^d	C ^d	G ^d	T ^d	No. aa ^e	MW (Da)
pTV001G8	1494	37	20	24	19	497	55 374
pTV001G11	1494	37	19	24	19	497	55 404
pcTV002G8	1464	37	19	24	20	487	54 503
pcTV003G15	1476	37	19	24	19	491	54 812
pcTV004G17	1479	37	20	25	19	492	54 889
pcTV004G24	1479	37	20	25	19	492	54 793
pcTV005G29	1479	37	19	24	19	492	55 109
pcTV005G36	1479	37	19	24	19	492	55 052
pcTV006G11	1468	37	19	25	19	488	54 593
pTV006G97	1460	37	19	24	19	485	54 315
pcTV007G59	1500	37	20	24	19	499	55 511
pTV008G65	1476	36	20	25	19	491	54 695
pcTV008G66	1476	37	20	24	19	491	54 899
pcTV009G12	1479	37	19	25	19	492	54 769
pcTV010G74	1482	37	20	24	19	493	55 088
pTV012G34	1491	36	20	24	19	496	55 405
pTV012G40	1503	37	20	24	19	500	55 951
pTV013G2	1527	37	20	24	19	508	56 952
pTV013G15	1527	37	20	24	19	508	56 910
pcTV014G73	1476	37	20	24	19	491	54 857
pTV018G60	1479	37	19	24	20	492	55 099
pTV019G20	1479	38	19	24	19	492	55 105
pTV019G25	1479	37	19	24	19	492	54 918

^c The size of the *gag* insert. ^d The percentage occurrence of the nucleotide in the *gag* sequence. ^e The number of amino acids in the predicted Gag protein.

Table 3.4The diversity of the HIV-1 *gag* gene

Region of sequence	Groups compared	Minimum variation ^(a)	Maximum variation ^(a)
Gag ^(b)	Subtype C <i>gag</i> clones	6	12
	Subtype C ^(g)	6	13
	Group M subtypes	8	19
	M, N, O	28	36
Matrix ^(b)	Subtype C ^(g)	6	15
	Group M subtypes	8	25
	M, N, O	20	42
Capsid ^(b)	Subtype C ^(g)	4	9
	Group M subtypes	5	15
	M, N, O	21	28
Nucleocapsid, Spacer peptides, p6 ^(b)	Subtype C ^(g)	3	14
	Group M subtypes	5	19
	M, N, O	27	35
CTL area 1 ^(d) in p17 ^(c)	Subtype C <i>gag</i> clones	4	26
CTL area 2 ^(e) in p17 ^(c)	Subtype C <i>gag</i> clones	0	18
CTL area 3 ^(f) in p17 ^(c)	Subtype C <i>gag</i> clones	0	45

^a Percentage diversity = 100% - percentage similarity (Distance matrix)

^b Distance matrix calculated on nucleotide diversity

^c Distance matrix calculated on amino acid diversity

^d Matrix area from the amino acid K18 to the amino acid E40

^e Matrix area from the amino acid G71 to the amino acid L101

^f Matrix area from the amino acid D121 to the amino acid Y132

^g Subtype C sequences in the Los Alamos HIV-1 *gag* reference subtypes set, plus the subtype C *gag* clones

Table 3.5

Evaluation of the Tfx™ transfection technique

Cell line	Tfx™ Reagent	Tfx™:DNA ratio	µg DNA	[CAT] in ng/ml
CV-1	Tfx™-50	2:1	0.25	0.012
			0.5	0.623
			0.75	1.028
			1.0	1.289
CV-1	Tfx™-50	4:1	0.25	0.027
			0.5	0.0
			0.75	1.005
			1.0	0.469
Vero	Tfx™-10	2:1	0.25	2.374
			0.5	2.461
			0.75	2.478
			1.0	2.094
Vero	Tfx™-10	4:1	0.25	2.184
			0.5	2.311
			0.75	2.019
			1.0	2.06

Table 3.6Evaluation of the CaPO₄ and Electroporation techniques

Transfection method	Culture growth area (cm ₂)	Cell line	[CAT] in ng/ml
CaPO₄	21	CV-1	2.51
	21	Vero	2.55
Electroporation	9.4	CV-1	1.13
	9.4	Vero	2.26

Figure 3.1 PCR results for the optimisation of the PCR reaction using the Expand High Fidelity PCR System at various MgCl₂ concentrations. Lanes marked 1 to 7 are the same for all three reactions. Lane 1: TV012; 2: TV012; 3: TV012; 4: TV025; 5: TV012; 6: TV012; 7: TV012. MgCl₂ concentration of 1.5 mM, lane 3 and 3.0 mM MgCl₂, lane 4 and 3.0 mM MgCl₂.

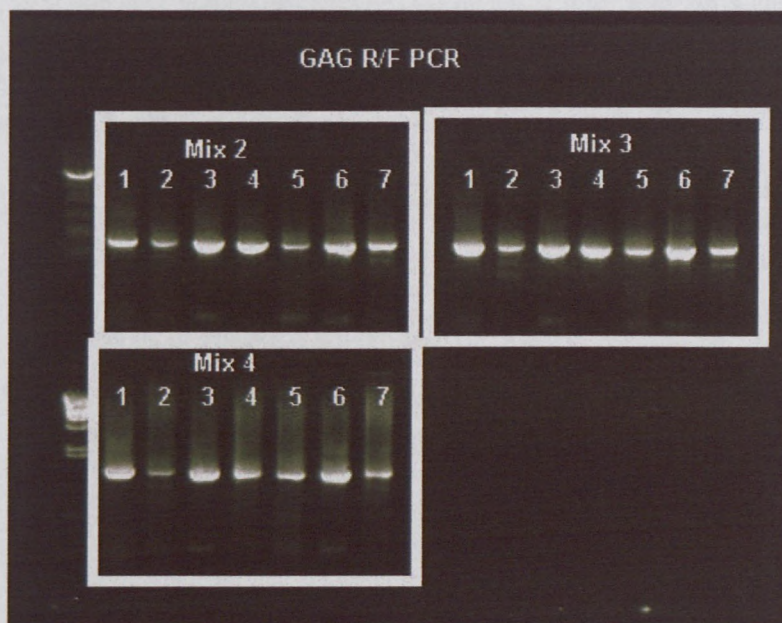


Figure 3.1 PCR results for the optimisation of the MgCl_2 concentration when using the Expand High Fidelity PCR System to amplify HIV-1 *gag*. The lanes marked 1 to 7 are the same for all three mixes. Lanes: 1, TV016; 2, TV012; 3, TV015; 4, TV008; 5, TV019; 6, TV016; 7, TV013. Mix 2 had a MgCl_2 concentration of 1.5 mM, mix 3 had 2.0 mM MgCl_2 and mix 4 had 3.0 mM MgCl_2 .



Figure 3.2 A spot blot of 6 *gag* PCR products for the optimisation of the PCR conditions. The numbers 1 to 4 represent the spot below the number. The number 5 represents the spot to right and number 6 the spot to the left of it. Spot 1, TV016; spot 2, TV012; spot 3, TV015; spot 4, TV008; spot 5, TV019; spot 6, TV013. The negative control is marked neg. A coloured spot is an indication of a positive reaction, while no coloration is an indication of a negative reaction.

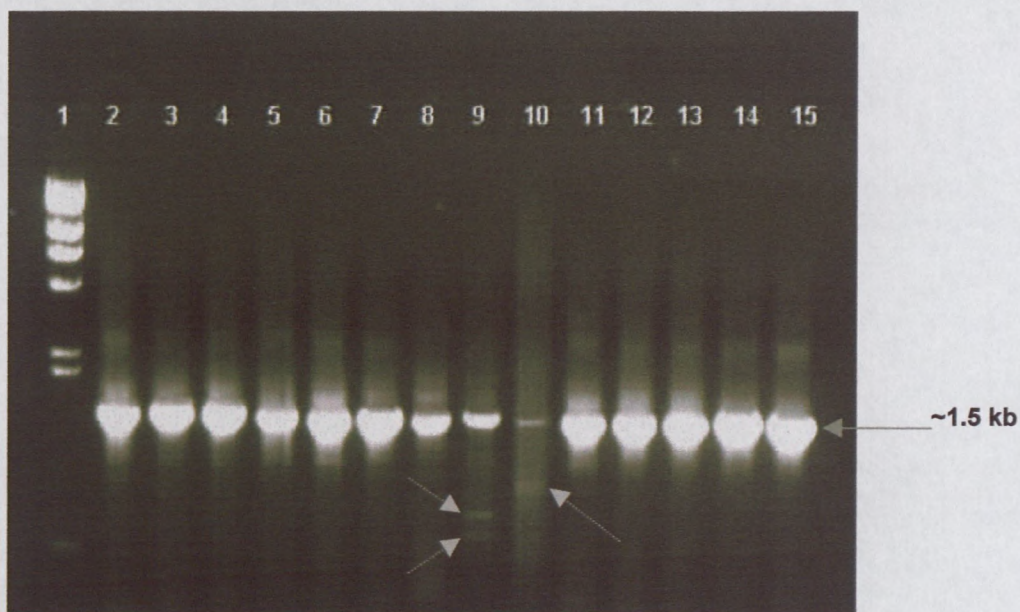


Figure 3.3 Agarose gel electrophoretic analyses of the PCR products to be used in TA cloning. Lanes: 1, Lambda/Hind III Molecular weight marker; 2, TV016; 3 and 4, TV013; 5, TV001; 6, TV006; 7, TV004; 8, TV002; 9, TV007; 10, TV003; 11, TV006; 12 and 13, TV005; 14 and 15, TV018. The perforated arrows indicate non-specific bands in lanes 9 and 10.

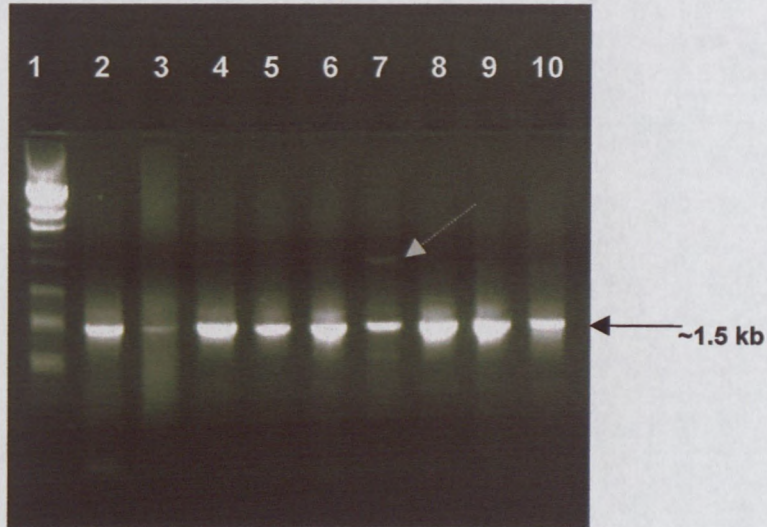


Figure 3.4 Agarose gel electrophoretic analyses of the amplified ~1.5 kb *gag* PCR product after purification for cloning into pcDNA3.1. Lanes: 1, Lambda/Hind III Molecular Weight Marker; 2, TV002; 3, TV003; 4, TV004; 5, TV005; 6, TV006; 7, TV007; 8, TV008; 9, TV010; 10, TV016. The perforated arrow indicates the non-specific band in lane 7.

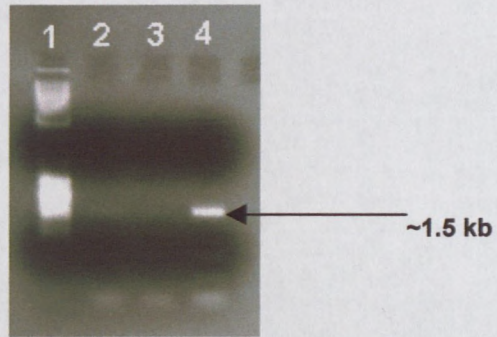


Figure 3.5 PCR results on RT-PCR products. Lanes: 1, Lambda/Hind III Molecular weight marker; 2, TV014; 3, TV009; 4, Positive control.

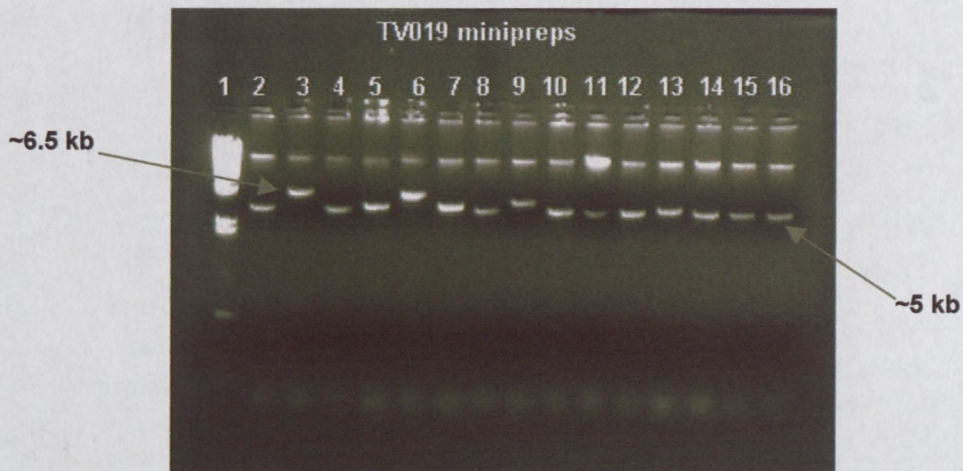


Figure 3.6 Agarose gel electrophoretic analyses of miniprep DNA for screening of clones. This is an example of how positive clones were distinguished from negative clones. Lane 1 displayed the Lambda/Hind III molecular weight marker. Lanes 2, 4, 5, 7, 8, 10, 11, 12, 13, 14, 15, and 16 are examples of negative clones, and lanes 3, 6, and 9 were considered to be positive clones. Negative clones, without inserts, were ~5 kb in size, while positive clones, with inserts, were ~6.5 kb in size.

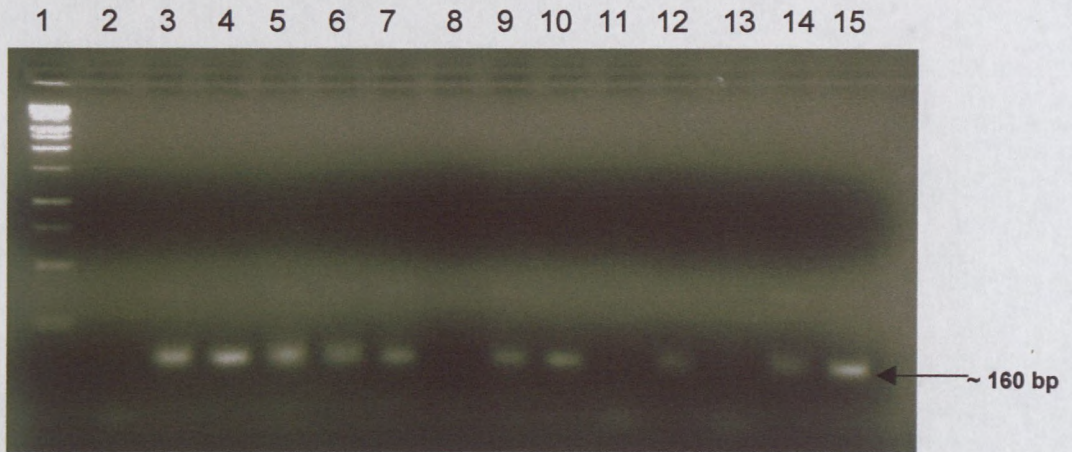


Figure 3.7 Agarose gel electrophoretic analyses of the PCR assay for confirmation of HIV-1 *gag* positive clones. This is an example of how the positive clones were distinguished from the negative clones. A positive clone would produce a 160 bp fragment after PCR. The smallest band of the 1 kb DNA ladder is 250 bp long. Lane 1 displayed the 1 kb DNA Ladder (Promega, USA) molecular weight marker. Lanes 3, 4, 5, 6, 7, 9, 10, 12, 14, and 15 were considered to be positive, while lanes 2, 8, 11, and 13 were all considered negative.

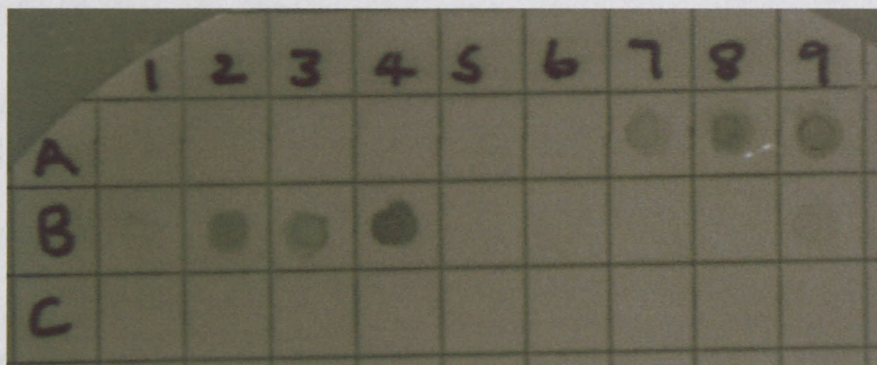


Figure 3.8 A spot blot assay for the confirmation of HIV-1 *gag* positive clones.

This is an example of how the positive clones were distinguished from the negative clones. A positive clone is identified as a coloured spot on the membrane, while a negative clone would have no coloration. A7, A8, A9, B2, B3, and B4 were considered to be positive, while B1 and B9 were considered as weak positives. A1 to A6, B5 to B8 and C1 to C9 were considered to be negative.

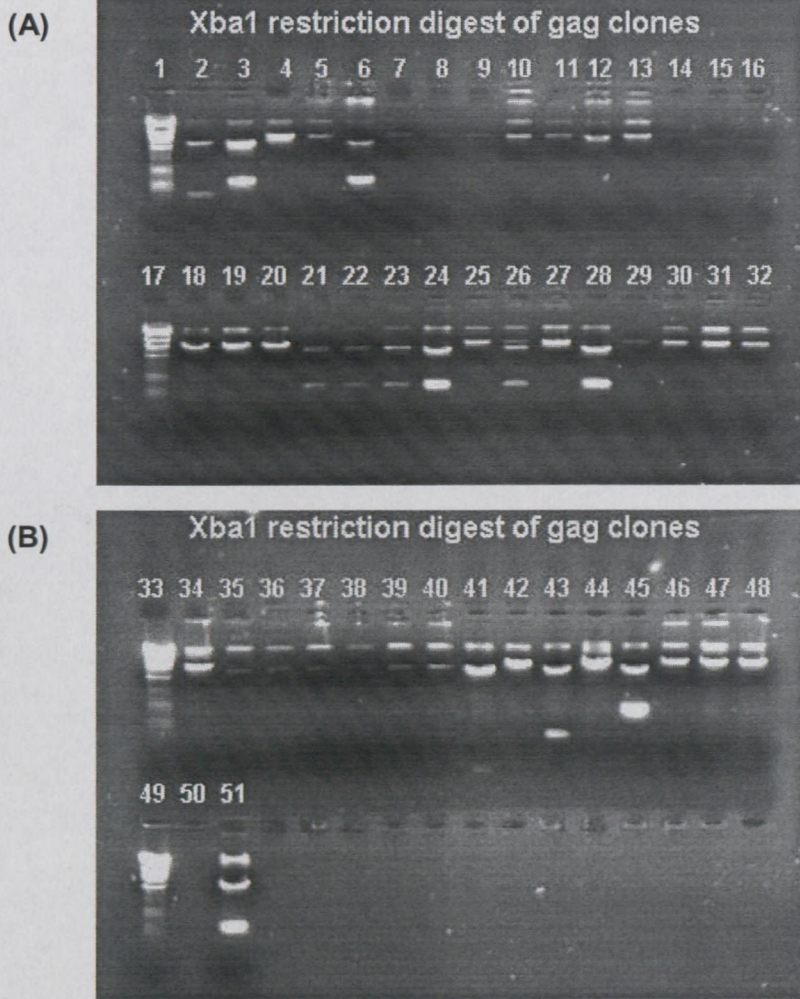


Figure 3.9 (A & B) Orientation determination of clones in the vector pCR[®]3.1, by restriction enzyme digestion with *Xba*1. The Lambda/Hind III molecular weight markers were loaded in the lanes 1, 17, 33, and 49. The pCR[®]3.1 clones were loaded as follows: lanes: 2, TV012G19; 3, TV012G34; 4, TV012G35; 5, TV012G38; 6, TV012G40; 7, TV008G59; 8, TV008G60; 9, TV008G62; 10, TV008G65; 11, TV008G70; 12, TV019G25; 13, TV019G10; 14, TV019G17; 15, TV019G20; 16, TV019G51; 18, TV016G61, 19, TV016G63; 20, TV016G67; 21, TV013G2; 22, TV013G8; 23, TV013G11; 24, TV013G15; 25, TV001G2; 26, TV001G8; 27, TV001G10; 28, TV001G11; 29, TV004G11; 30, TV004G12; 31, TV004G13; 32, TV004G16; 34, TV004G17; 35, TV002G1; 36, TV002G2; 37, TV002G16; 38, TV007G10; 39, TV007G11; 40, TV007G12; 41, TV006G71; 42, TV006G78; 43, TV006G88; 44, TV006G93; 45, TV006G97; 46, TV006G98; 47, TV005G29; 48, TV018G59; 50, blank; 51, TV018G60.

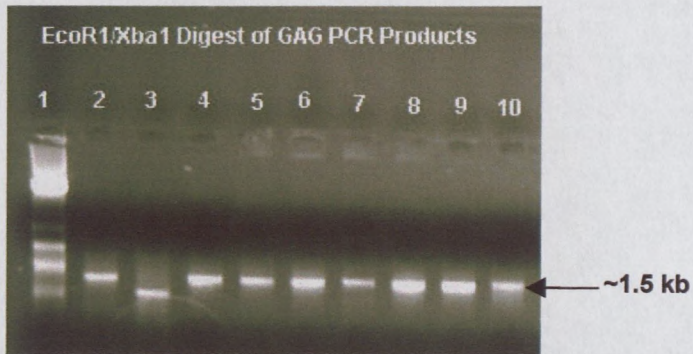


Figure 3.10 Agarose gel electrophoretic analyses of PCR products of *gag* after digestion with the restriction enzymes *Xba*I and *Eco*RI. Lanes: 1, Lambda/*Hind* III Molecular Weight Marker; 2, TV002; 3, TV003; 4, TV004; 5, TV005; 6, TV006; 7, TV007; 8, TV008; 9, TV010; 10, TV016. Although the lower end of the gel is not shown in this figure, no visible bands were observed beyond these borders when the photograph was taken.

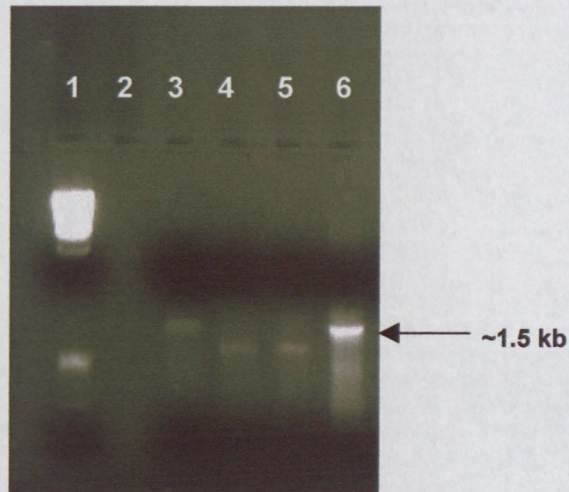


Figure 3.11 Restriction digest of the *gag* PCR product of the isolate TV003. Lanes: 1, Lambda/Hind III Molecular Weight Marker; 2, blank; 3, digestion with *EcoRI*; 4, digestion with *XbaI*; 5, digestion with both *EcoRI* and *XbaI*; 6, the undigested PCR product.

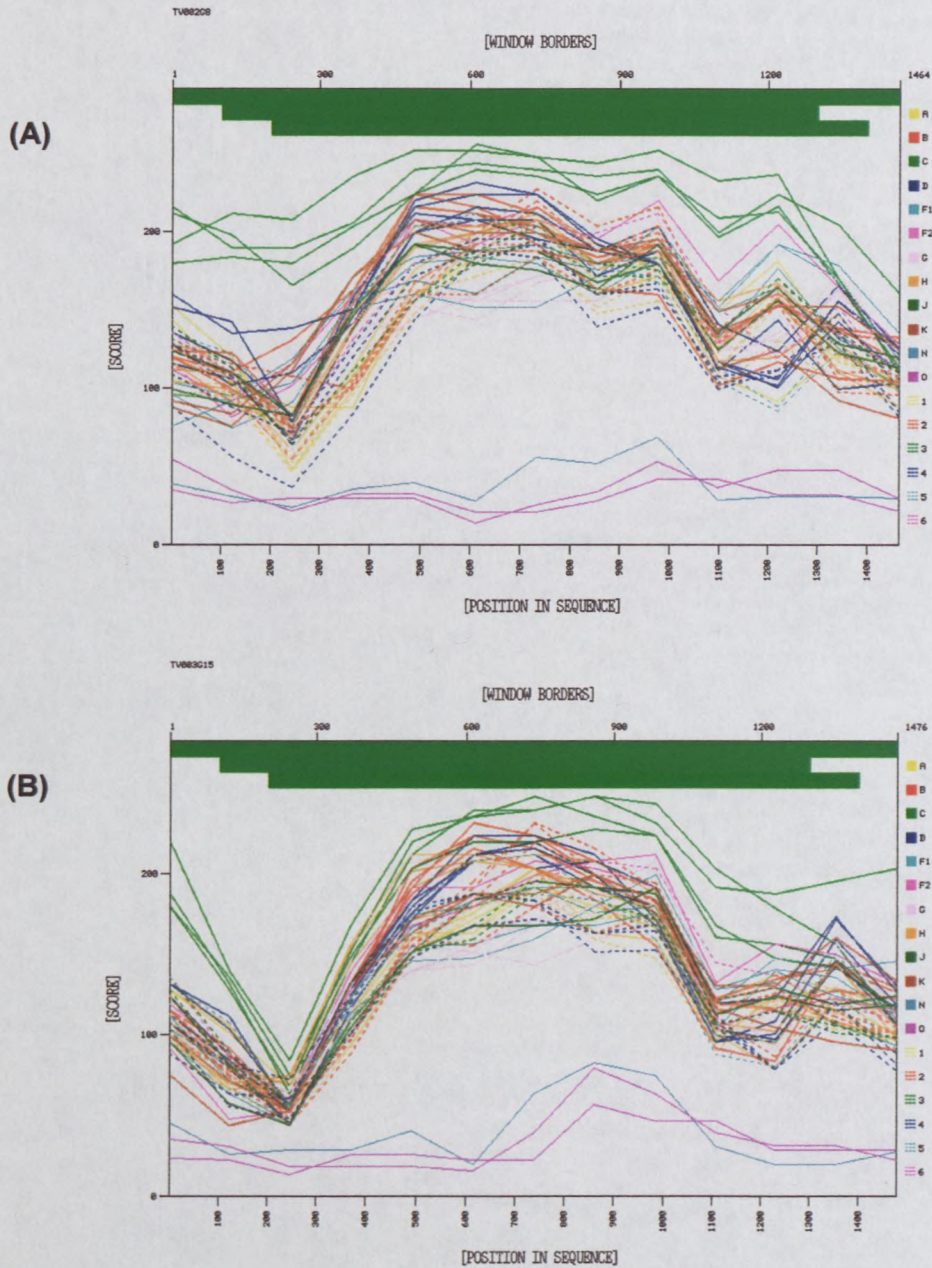


Figure 3.12 (A & B) The similarity plot of the NCBI HIV-1 subtyping database for the clones TV002G8 (A) and TV003G15 (B). The right hand legend indicates the different reference subtypes, which the query sequence was compared to. The solid bar at the top of the graph represents the sybtype most similar to the query sequence. The X-axis represents the nucleotide position, while the Y-axis represents the score of identity obtained when comparing the query sequence to the individual reference sequences. The similarity between the query and reference sequences increases with an increase in score.

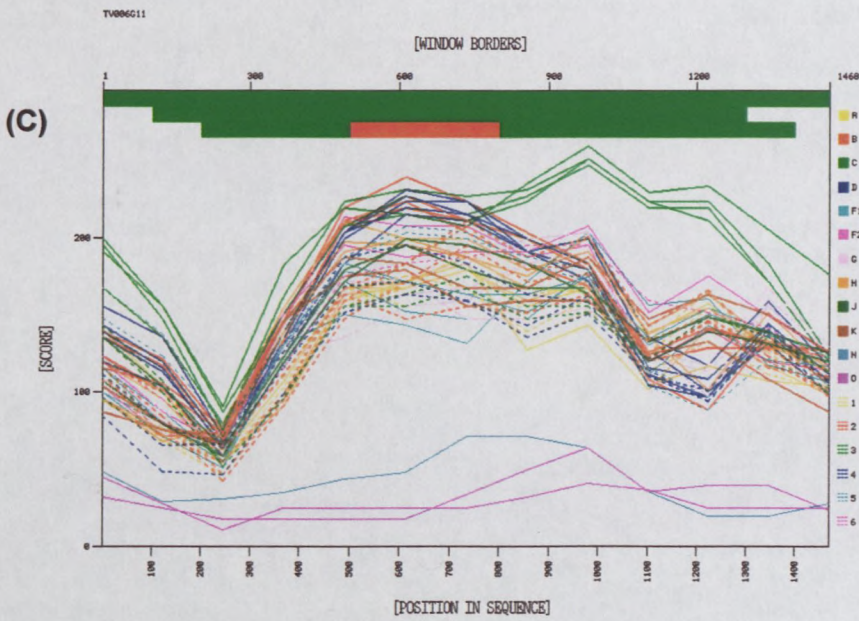


Figure 3.12 (C & D) The similarity plot of the NCBI HIV-1 subtyping database for the clones TV006G11 (C) and TV006G97 (D). The right hand legend indicates the different reference subtypes, which the query sequence was compared to. The solid bar at the top of the graph represents the sybtype most similar to the query sequence. The X-axis represents the sybtype most similar to the query sequence. The X-axis represents the nucleotide position, while the Y-axis represents the score of identity obtained when comparing the query sequence to the individual reference sequences. The similarity between the query and reference sequences increases with an increase in score.

(E)

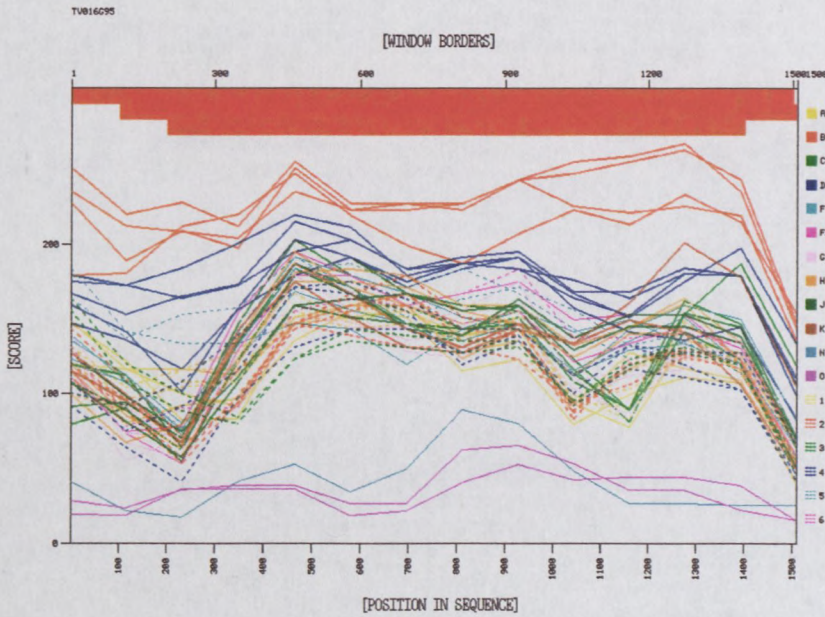


Figure 3.12 E The similarity plot of the NCBI HIV-1 subtyping database for the clone TV0016G95. The right hand legend indicates the different reference subtypes, which the query sequence was compared to. The solid bar at the top of the graph represents the sybtype most similar to the query sequence. The X-axis represents the nucleotide position, while the Y-axis represents the score of identity obtained when comparing the query sequence to the individual reference sequences. The similarity between the query and reference sequences increases with an increase in score.

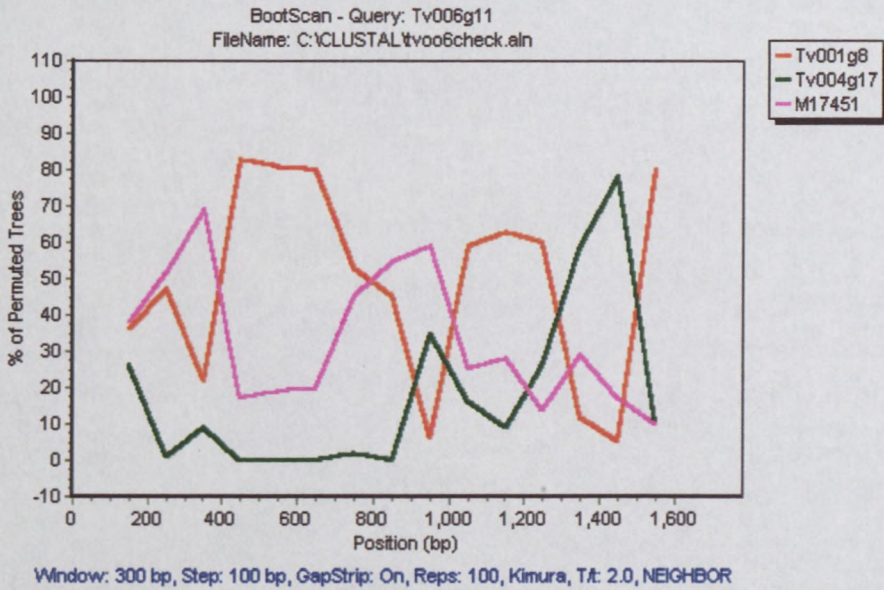
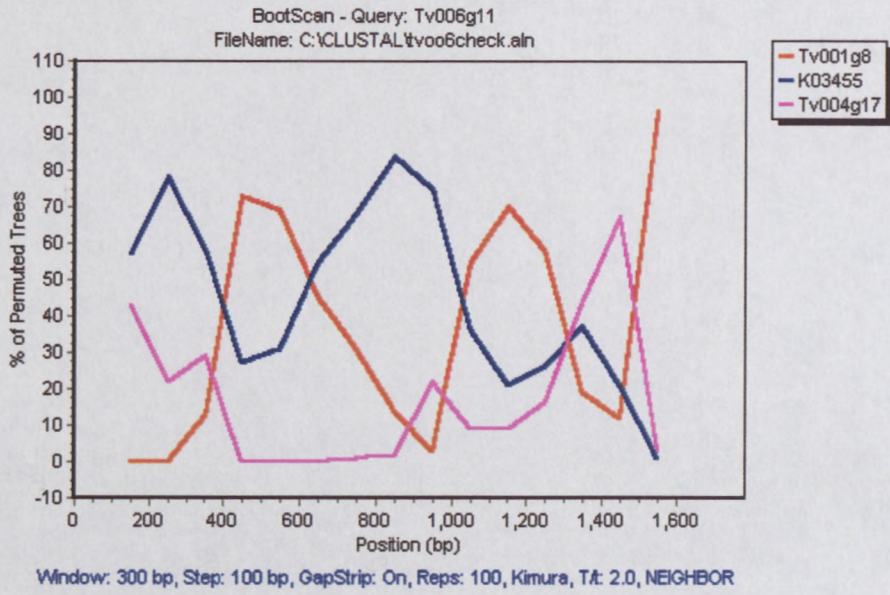
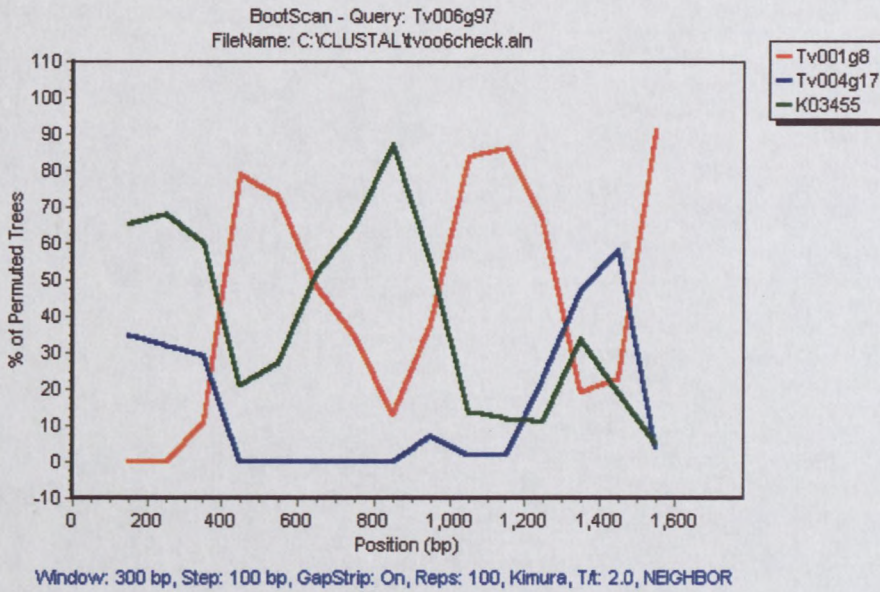


Figure 3.13 (A & B) The bootscan graph drawn in SimPlot for the comparison of the sequence of TV006G11 to the sequences of TV001G8, TV004G17 and K03455 (A) or M17451 (B). The right hand legend indicates the sequences that the sequence of TV006G11 was compared to. Similarity increases on the Y-axis, while the X-axis shows the nucleotide position.

(C)



(D)

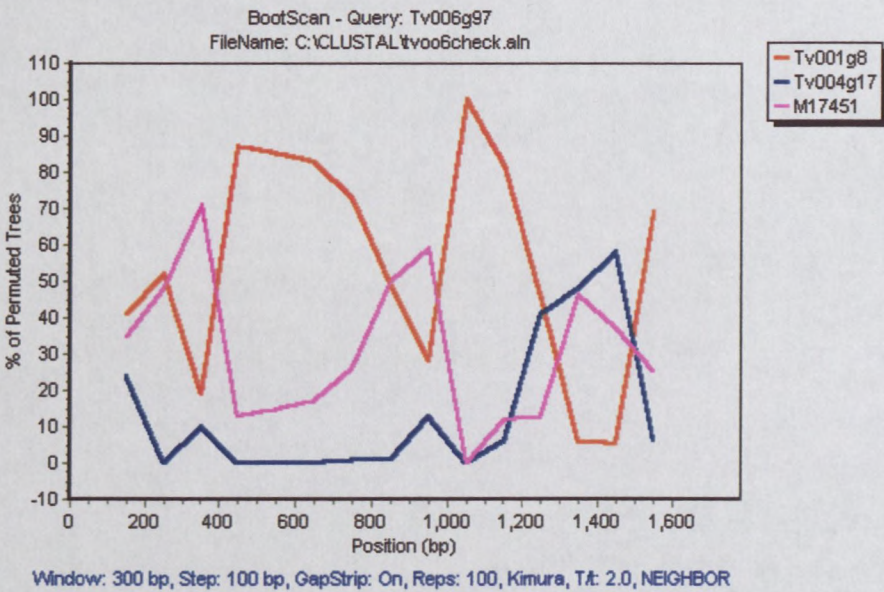


Figure 3.13 (C & D) The bootscan graph drawn in SimPlot for the comparison of the sequence of TV006G97 to the sequences of TV001G8, TV004G17 and K03455 (A) or M17451 (B). The right hand legend indicates the sequences that the sequence of TV006G97 was compared to. Similarity increases on the Y-axis, while the X-axis shows the nucleotide position.

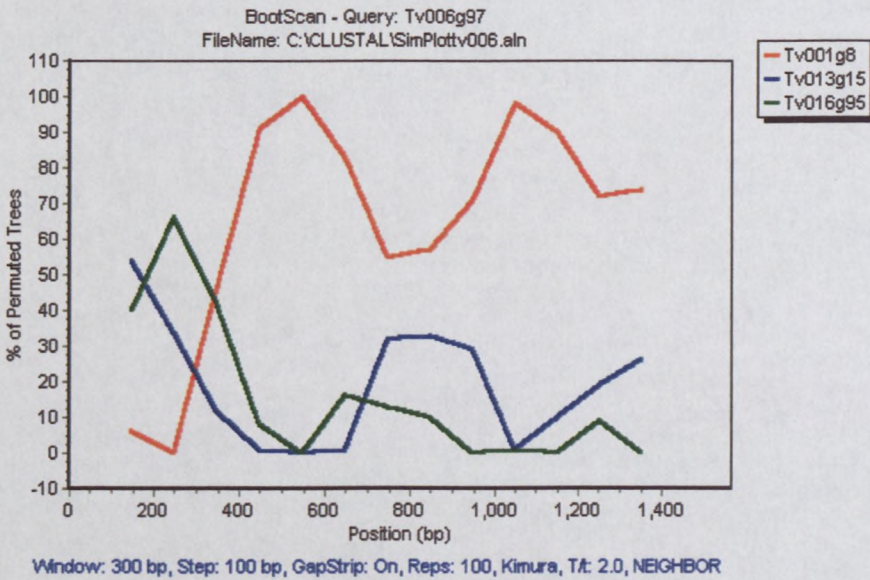
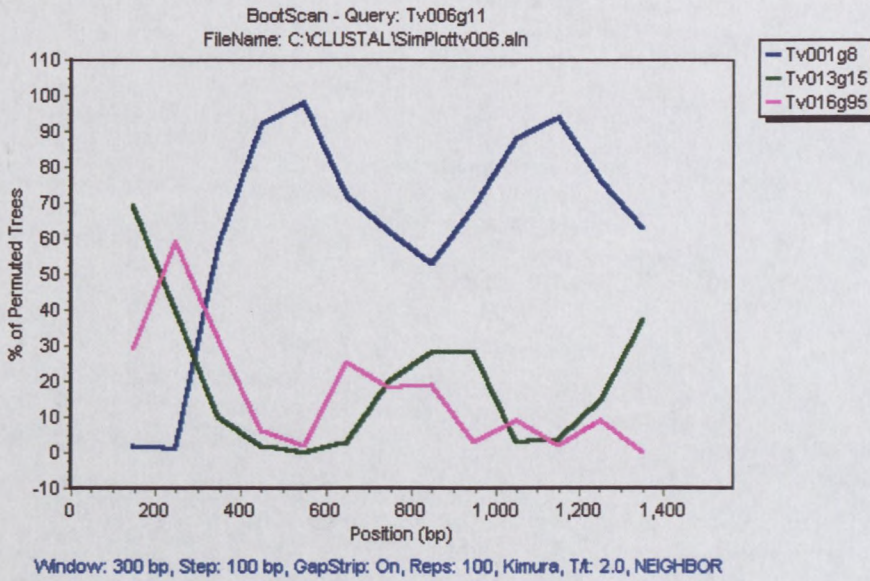


Figure 3.13 (E & F) The bootscan graph drawn in SimPlot for the comparison of the sequence of TV006G11 (E) and TV006G97 (F) to the sequences of TV001G8, TV013G15 and the clone of the subtype B isolate TV016. The right hand legend indicates the sequences that the sequence of the TV006 clones were compared to. Similarity increases on the Y-axis, while the X-axis shows the nucleotide position.

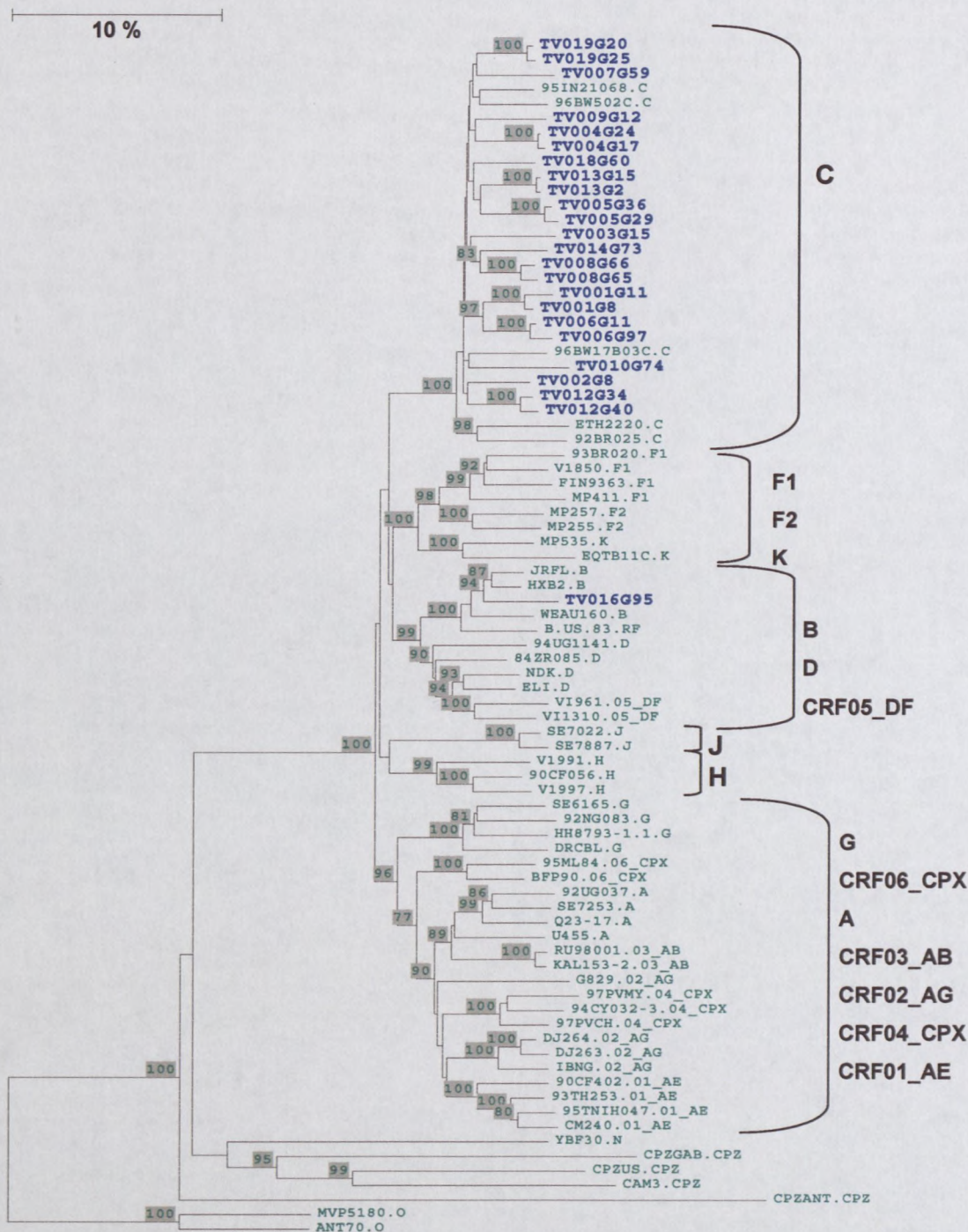


Figure 3.14 A neighbour-joining phylogenetic tree, constructed in TREECON, comparing the complete *gag* DNA sequences of the clones and the Los Alamos HIV-1 *gag* reference subtypes, with bootstrap values (shown as percentage) greater than 75 % indicated. The South African clones are indicated in bold, blue letters. The horizontal scale indicates the percentage variation, represented by the horizontal branch lengths between OTUs. Six clusters and the representing subtypes within group M are indicated by the brackets on the right hand side of the branches.

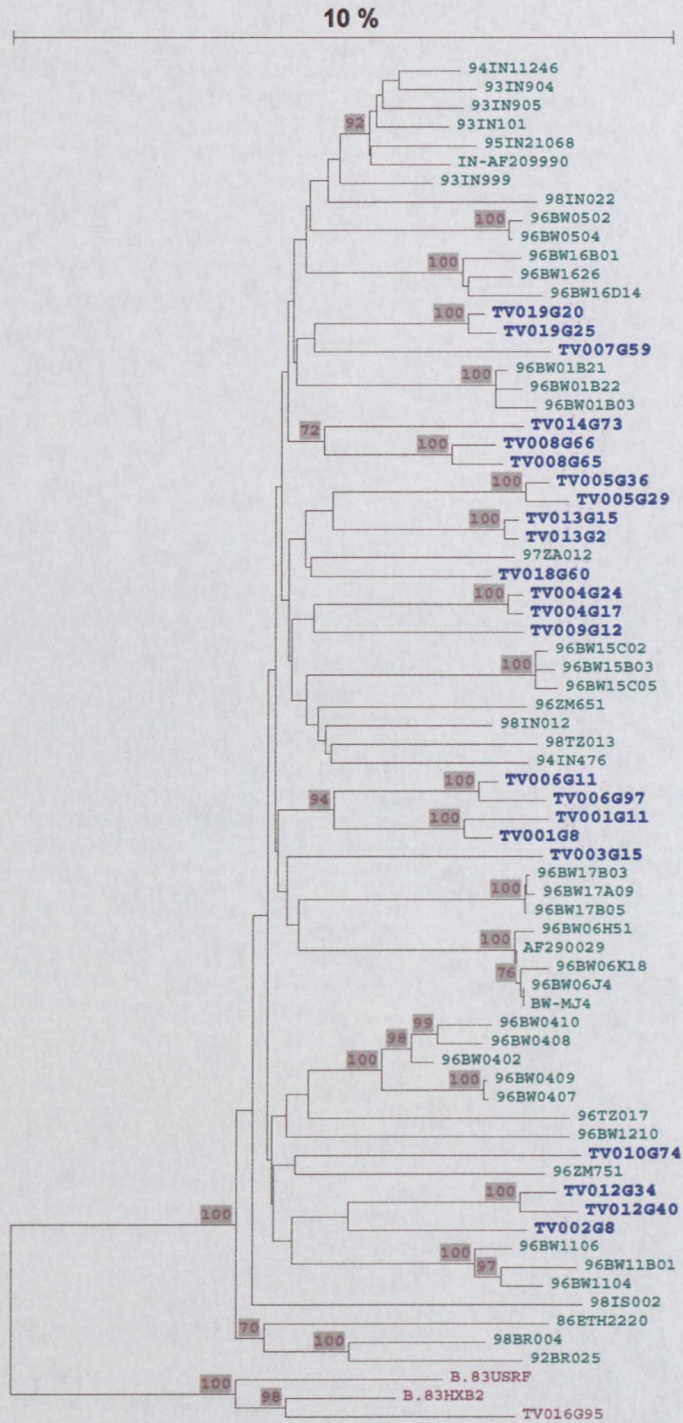


Figure 3.15 A neighbour-joining phylogenetic tree, constructed in TREECON, comparing the complete *gag* sequences of the South African clones and the 47 HIV-1 subtype C *gag* sequences in the Los Alamos databank (Table2.3). The tree is rooted with three subtype B sequences. The South African clones are indicated in bold, blue letters. The horizontal scale indicates the percentage variation, represented by the horizontal branch lengths between OTUs.

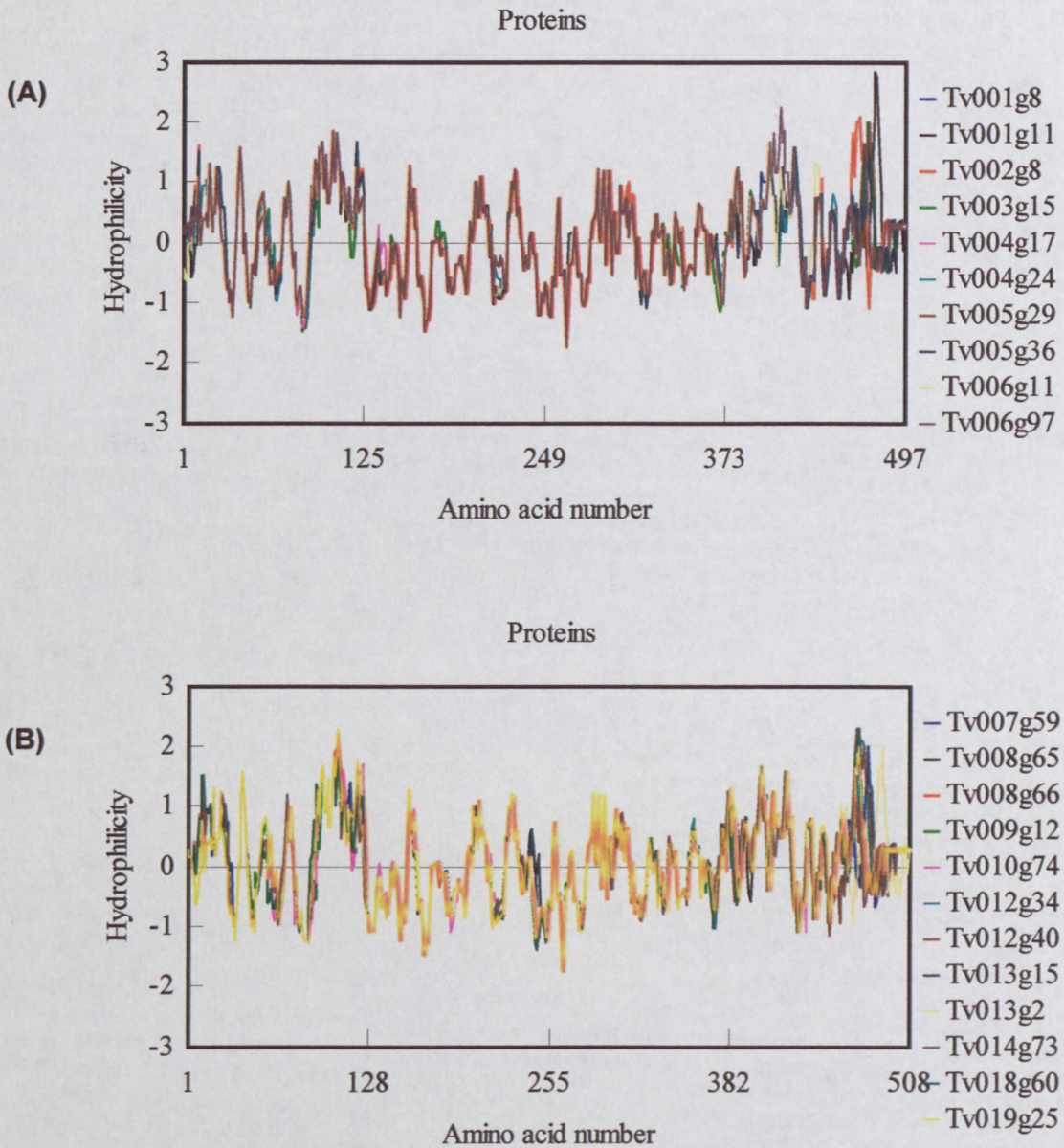


Figure 3.16 (A & B) The hydrophilicity profiles of the predicted amino acid sequences of the HIV-1 subtype C *gag* clones. In the right hand legend, Tvxxgxx represents TVxxGxx, as DNAMAN, which was used to construct the graphs, do not allow capital letters throughout a word.

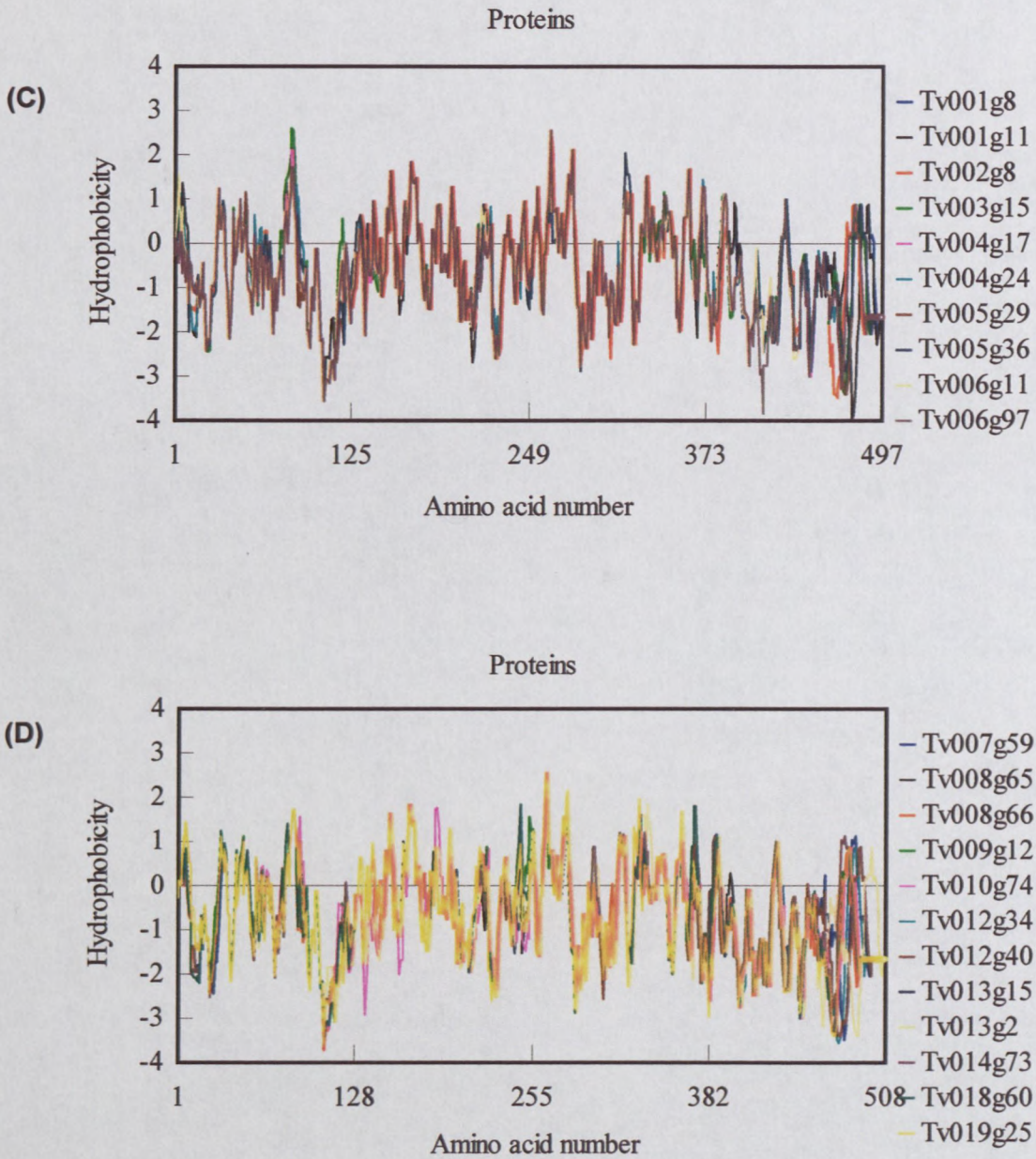


Figure 3.16 (C & D) The hydrophobicity profiles of the predicted amino acid sequences of the HIV-1 subtype C *gag* clones. In the right hand legend, Tvxxgxx represents TVxxGxx, as DNAMAN, which was used to construct the graphs, do not allow capital letters throughout a word.

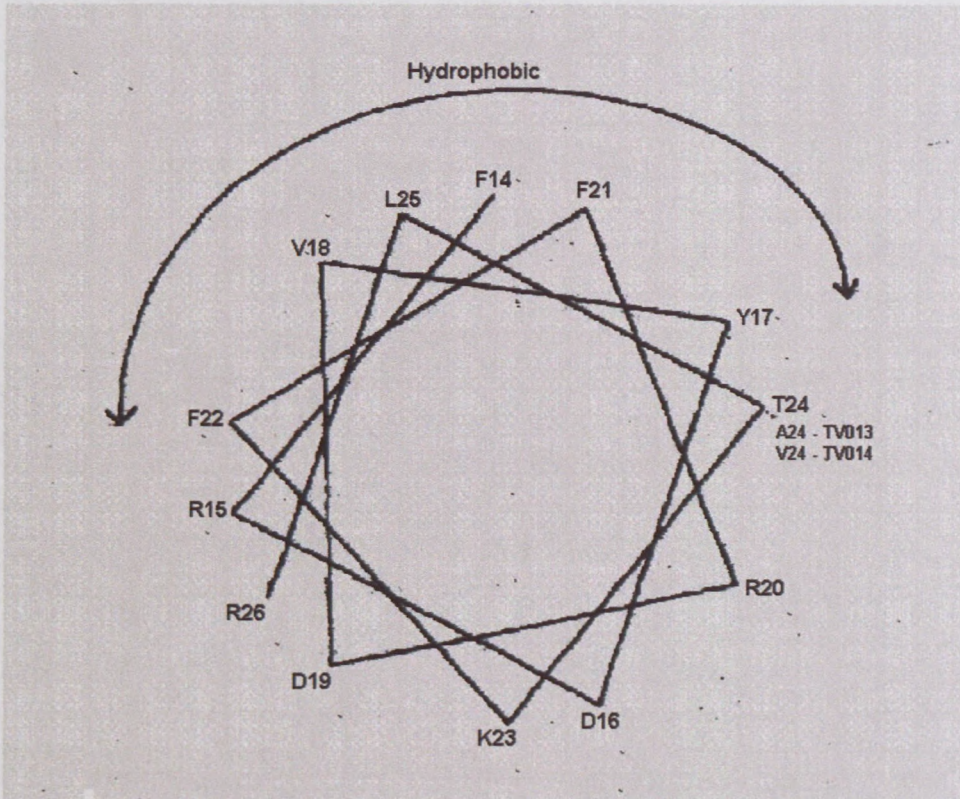


Figure 13.17 The amphipathic α -helix of the major homology region (MHR) in the capsid domain (Clish et al, 1996), as found in the predicted amino acid sequence of the HIV-1 subtype C *gag* clones. Each corner within the helix represents an amino acid, which is written in the single letter code, followed by the position of the amino acid in the MHR. A24-TV013 and V24-TV014, indicate the amino acid at that position for the isolates TV013 and TV014.

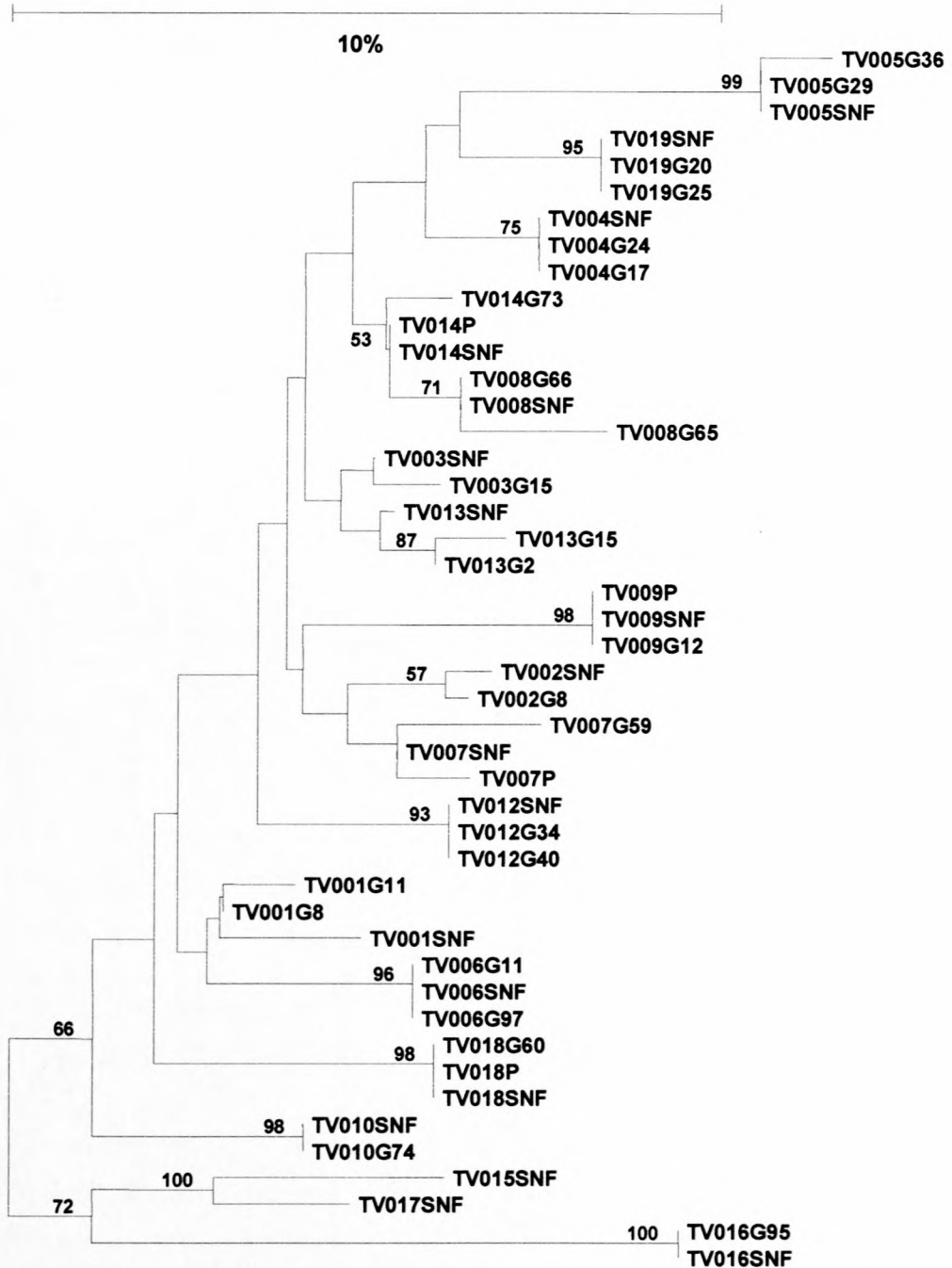


Figure 3.18A A neighbour-joining phylogenetic tree analysis, constructed in TREECON, of the quality control performed on the sequences of the clones, using all three codon positions in the alignment. The SNF at the back of sequence names indicates that the origin of the sequence is the earliest supernatant fluid available. The P indicates that the sequence is from the RNA isolated from the original plasma. Bootstrap values above 50% are indicated.

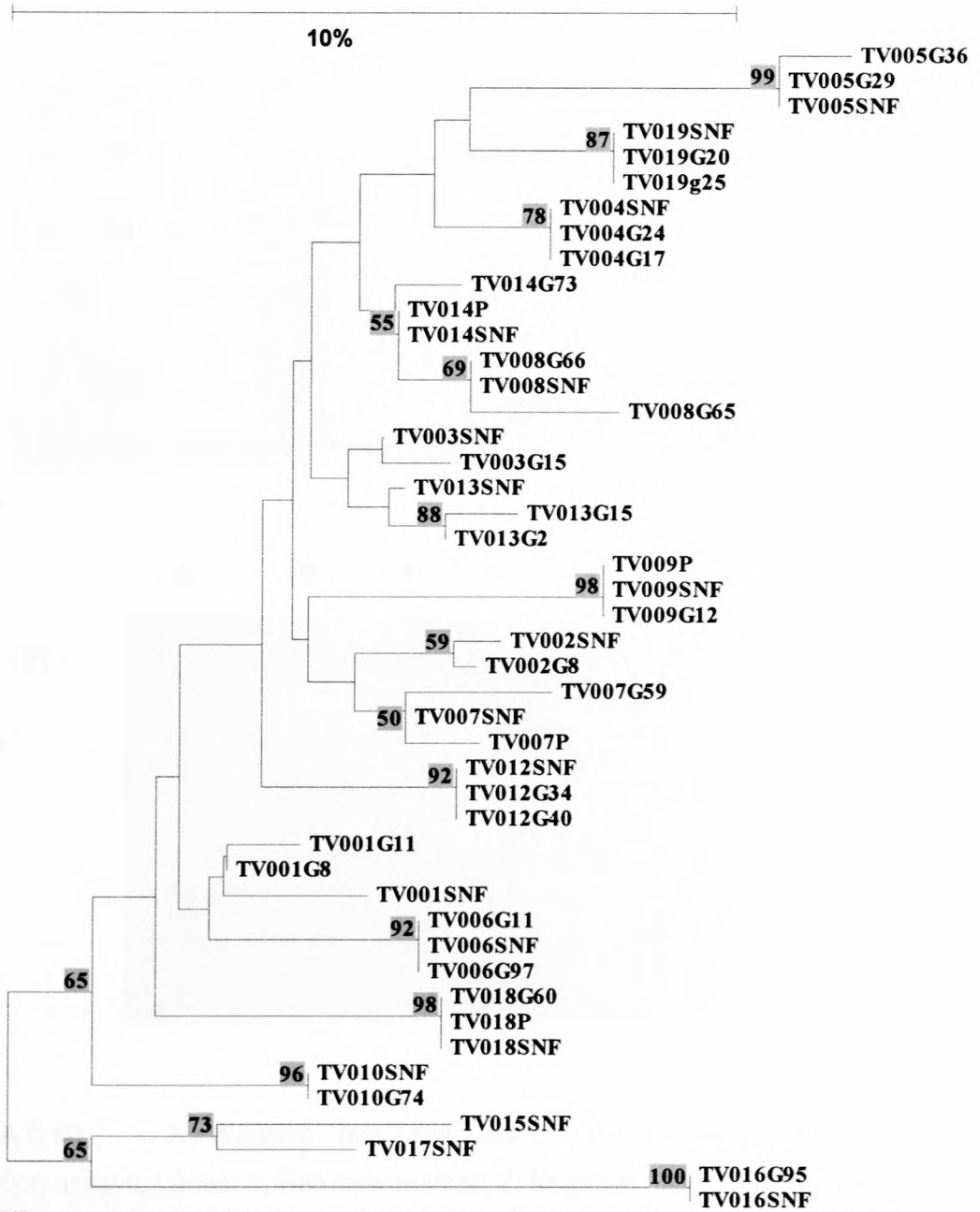


Figure 3.18B A neighbour-joining phylogenetic tree analysis, constructed in TREECON, of the quality control performed on the sequences of the clones, using only the synonymous substitutions in the alignment. The SNF at the back of sequence names indicates that the origin of the sequence is the earliest supernatant fluid available. The P indicates that the sequence is from the RNA isolated from the original plasma. Bootstrap values above 50% are indicated.

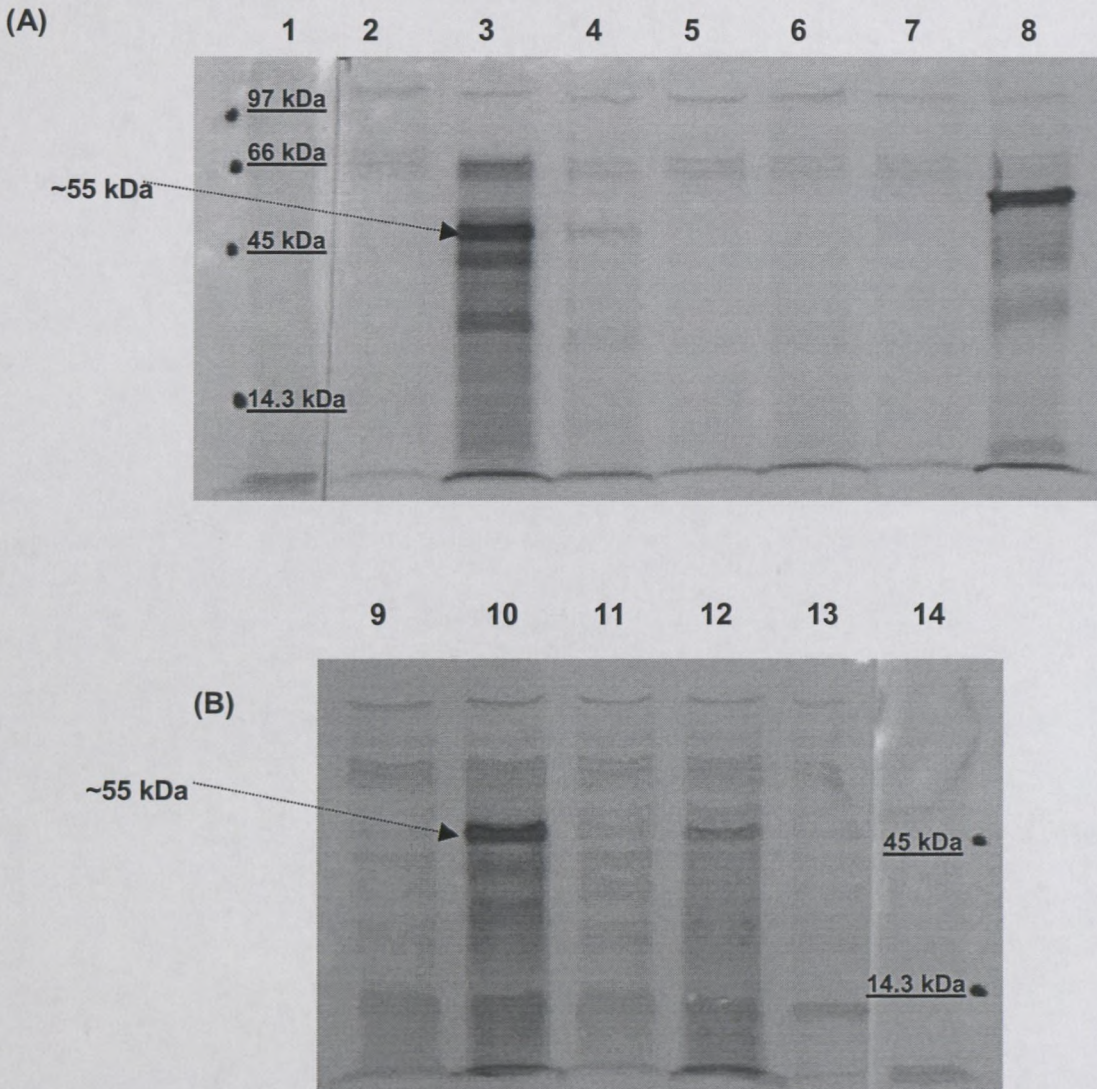


Figure 3.19 (A & B) A Western blot analyses of the *in vitro* transcription/translation assays. Lanes: 1, Rainbow marker; 2, Negative control; 3, TV002G8; 4, TV013G15; 5, TV007G59; 6, TV005G36; 7, TV019G20; 8, Luciferase control; 9, TV016G95; 10, TV004G24; 11, TV010G74; 12, TV008G66; 13, Negative control; 14, Rainbow marker. Dashed arrows indicate the position of a ~55 kDa protein.

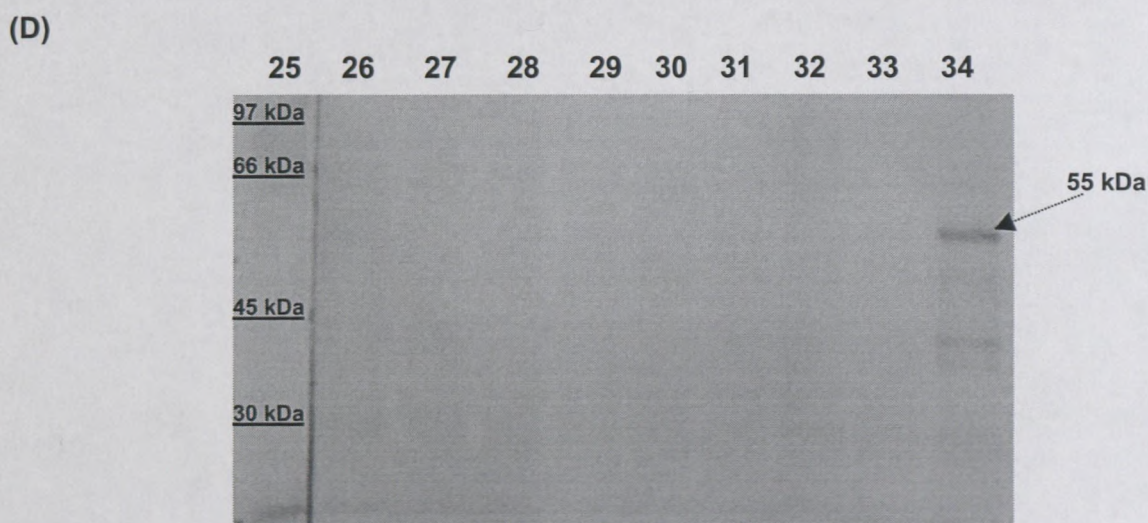
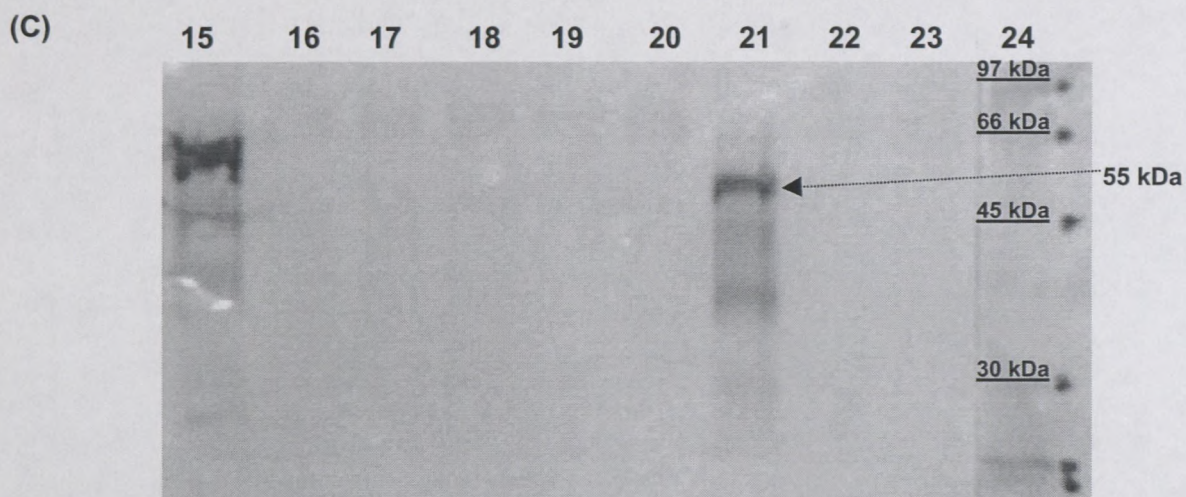


Figure 3.19 (C & D) A Western blot analyses of the *in vitro* transcription/translation assays. Lanes: 15, Luciferase control; 16, TV013G2; 17, TV006G97; 18, TV005G36; 19, TV003G15; 20, TV001G8; 21, TV001G11; 22, TV006G11; 23, Negative control; 24, Rainbow marker; 25, Rainbow marker; 26, TV012G34; 27, TV010G74; 28, TV012G40; 29, TV016G95; 30, TV018G60; 31, TV019G20; 32, TV007G59; 33, pCR[®]3.1; 34, TV001G11. Dashed arrows indicate the position of a ~55 kDa protein.

(E)

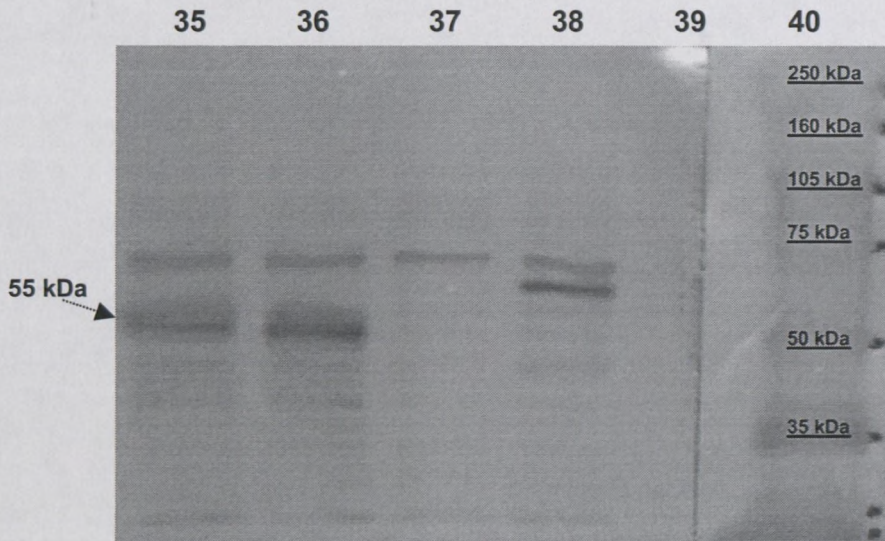


Figure 3.19 E A Western blot analyses of the *in vitro* transcription/ translation assays. Lanes: 35, TV009G12; 36, TV014G73; 37, TV001G11; 38, Luciferase control; 39, Blank; 40, Rainbow marker. Dashed arrows indicate the position of a ~55 kDa protein.

Expression of CAT - Tfxtransfection method

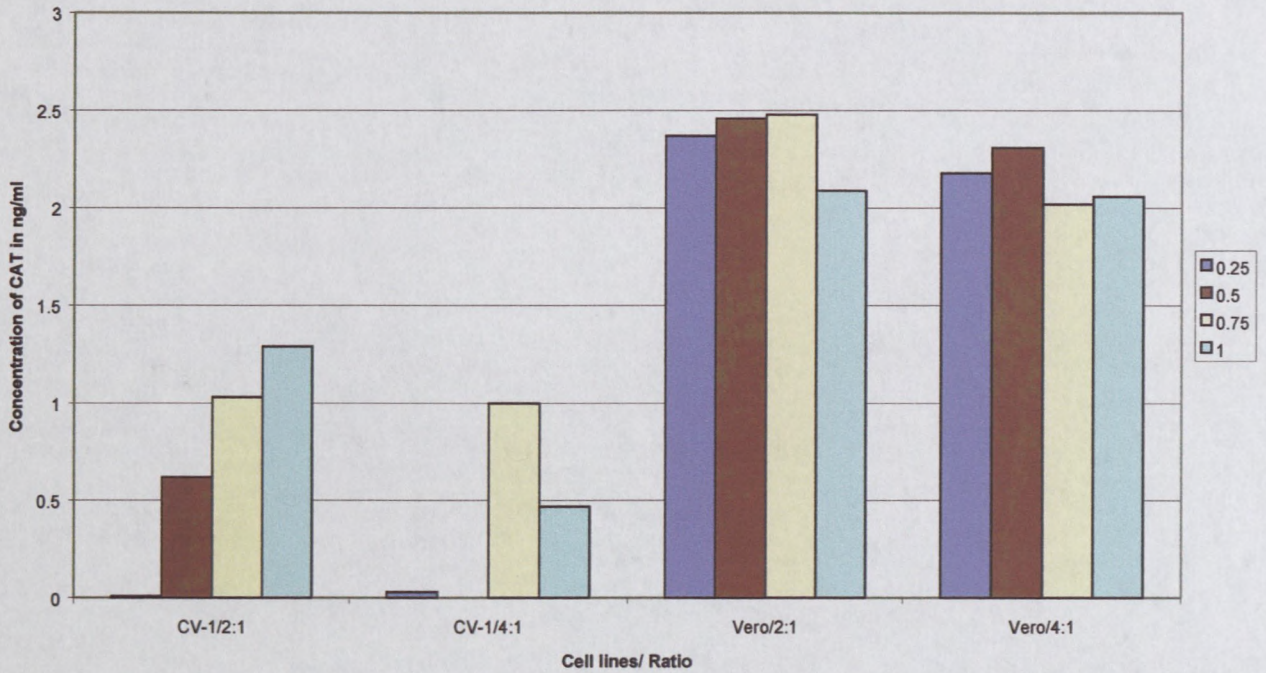


Figure 3.20A A graphical analyses of the production of the CAT protein in the Tfx transfection method and cell lines. The cell lines and ratio of transfection reagents to DNA used in the transfection are portrait on the X-axis, while the concentration of CAT protein is indicated on the Y-axis. The legend on the right indicates the concentration DNA used in the Tfx methods.

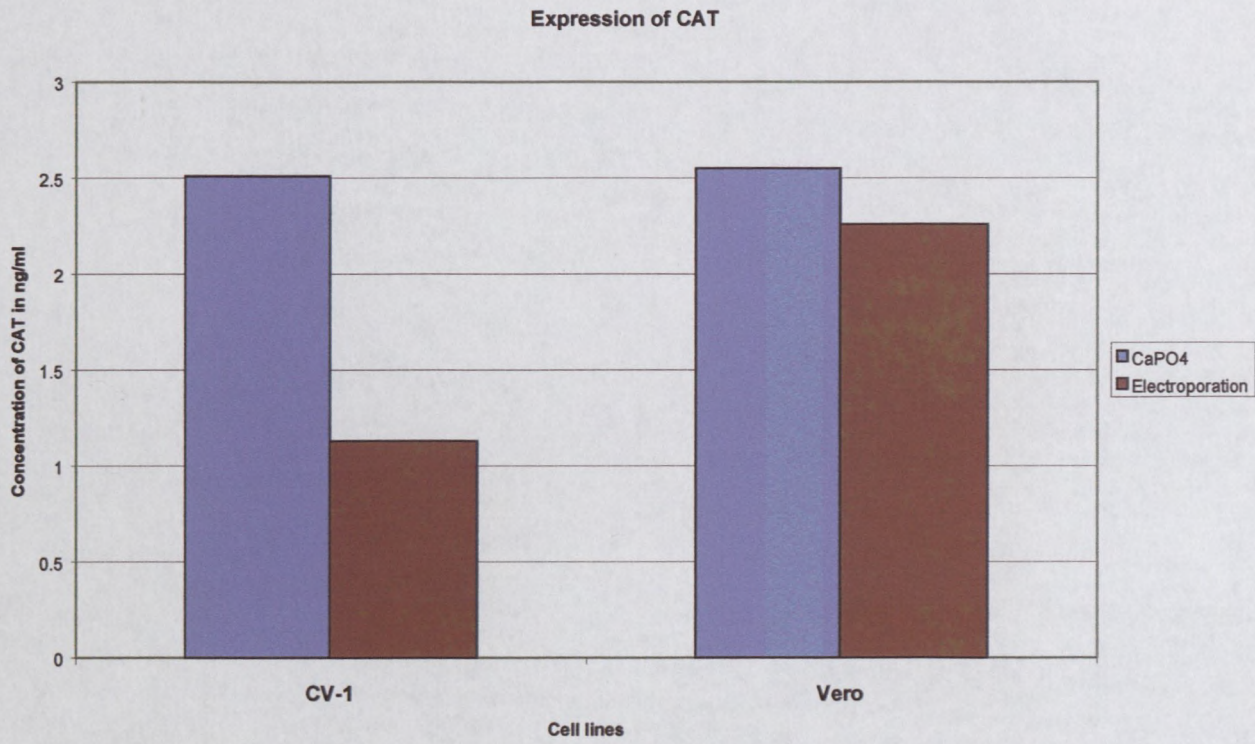


Figure 3.20B A graphical analyses of the production of the CAT protein in the CaPO₄ and electroporation transfection methods and cell lines. The cell lines are portrait on the X-axis, while the concentration of CAT protein is indicated on the Y-axis. The legend on the right indicates the transfection methods.

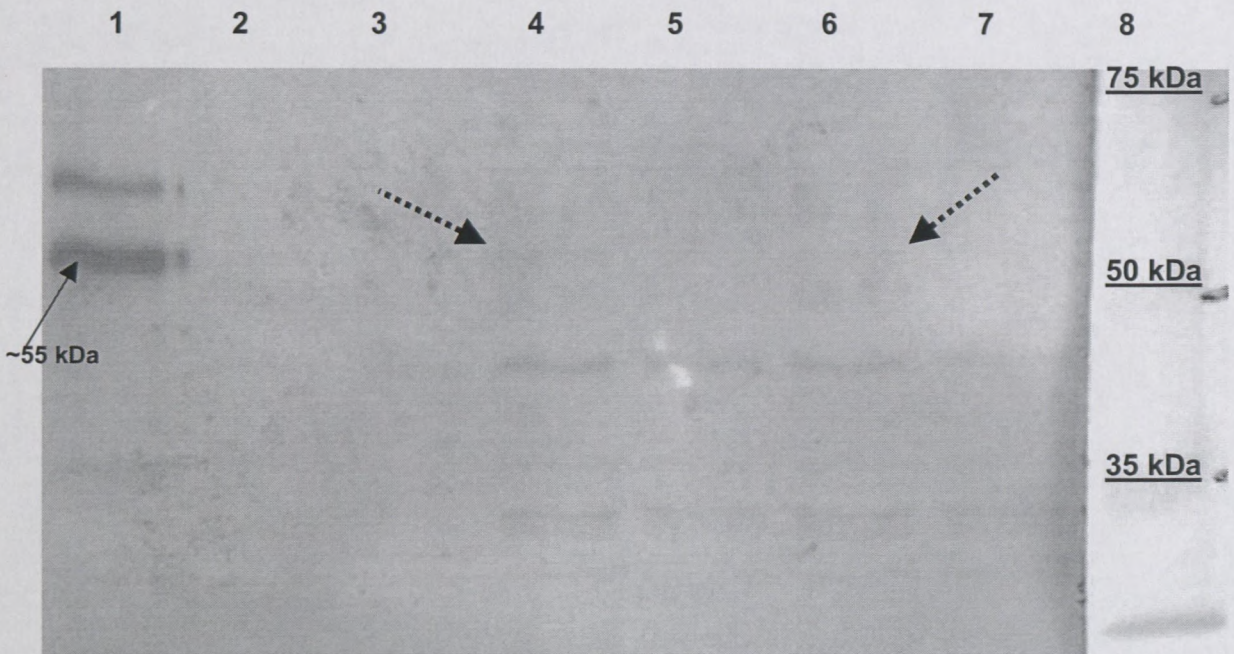


Figure 3.21 A A Western blot analyses of the harvested transfected vero cells. The cells were transfected with different HIV-1 subtype C gag clones. Blot A was detected by alkaline phosphatase. Lanes: 1, Pr55^{gag} positive control; 2, negative control; 3, TV001G11; 4, TV008G66; 5, TV004G24; 6, TV013G8; 7, TV002G8; 8, Rainbow marker. The dashed arrows indicate faint ~55 kDa bands in lanes 4 and 6.

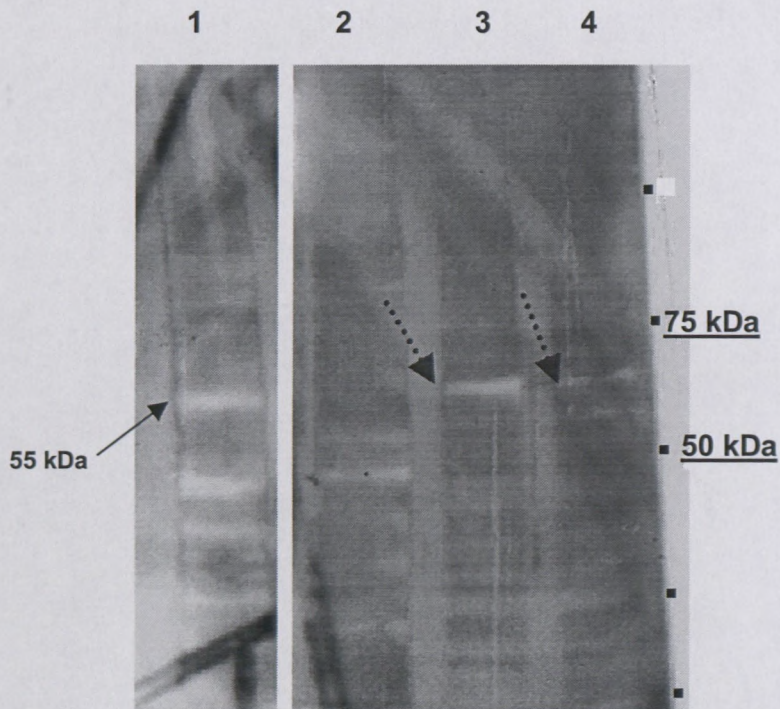


Figure 3.22 B A Western blot analyses of the harvested transfected vero cells. The cells were transfected with different HIV-1 subtype C gag clones. Blot B was detected by ECL. Lanes: 1, The positive *in vitro* transcription/translation reaction of TV009G12; 2, pCR[®]3.1; 3, TV009G12; 4, TV014G73. The Rainbow molecular weight size markers are indicated on the right hand side of the blot. The dashed arrows indicate ~55 kDa bands in lanes 3 and 4.

CHAPTER FOUR

4. Discussion and Conclusion

	Page
4.1 Discussion	104
4.1.1 Characterisation Studies	104
4.1.1.1 Amplification of <i>gag</i>	104
4.1.1.2 Cloning of <i>gag</i>	106
4.1.1.3 Sequencing of <i>gag</i>	107
4.1.1.3a Nucleotide Sequence Analyses	107
4.1.1.3a(i) Sequencing Data	107
4.1.1.3a(ii) Restriction Digest Analyses	108
4.1.1.3a(iii) Subtyping	110
4.1.1.3a(iv) Phylogenetic Analyses	111
4.1.1.3b Predicted Amino Acid Sequence Analyses	113
4.1.1.3b(i) Matrix	113
4.1.1.3b(ii) Capsid	115
4.1.1.3b(iii) Nucleocapsid	117
4.1.1.3b(iv) Spacer peptides and p6	117
4.1.1.3b(v) Protease cleavage sites	118
4.1.1.3b(vi) CTL epitopes	119
4.1.1.4 DNA Quality Assurance	120
4.1.2 Expression Studies	121
4.1.2.1 <i>In vitro</i> transcription/ translation	121
4.1.2.2 Evaluation of the transfection techniques	122
4.1.2.3 Gag protein production in cell culture	122
4.2 Conclusion	123

CHAPTER FOUR

4. Discussion and Conclusion

4.1 Discussion

4.1.1 Characterisation Studies

4.1.1.1 Amplification of *gag*

The Expand High Fidelity PCR System (Roche, Germany) was used for the amplification of *gag*. The system uses an enzyme mix containing thermostable *Taq* -and *Pwo* DNA polymerases. Due to the *Taq* DNA polymerase, the system generates PCR products containing either single 3' deoxyadenosine residues or blunt ends. The *Pwo* DNA polymerase, which has a 3'-5' exonuclease proofreading activity, causes the system to increase the fidelity of DNA synthesis three fold, thus introducing less PCR induced mutations.

The PCR cycle was designed with the length of the fragment to be amplified in mind. Longer starting denaturing and annealing times were required due to the large size of the starting template (~9kb). When the concentration of the smaller fragment (~1.5 kb) increased the time was reduced. The elongation time was kept long to reduce the risk of PCR mediated recombination, which can happen when the elongation process is terminated before completion (Pääbo et al, 1989, Overbaugh et al, 1990).

To ensure a high efficiency of cloning, it is important to use a PCR product that is homogenous in size. The less background bands visible on an agarose gel, after electrophoresis, the more specific the PCR reactions were. A single band of DNA, ~1.5 kb in length with no visible background bands, as seen in Figure 3.1, Mix 2, lanes one to six, was perceived to be homogenous enough for cloning. All the

PCR products of the isolate TV013 (Figure 3.1) displayed a smaller DNA band, which might have been produced by non-specific amplification. The ratio between the concentration of specific DNA to non-specific DNA (5:1) was not big enough to ignore the non-specific band. When cloning, the presence of non-specific DNA, especially shorter pieces, will result in clones with non-specific inserts, favouring the incorporation of the shorter piece of DNA (Sambrook et al, 1989). The PCR product of the isolate TV013 was discarded and the PCR repeated at a later stage, where the PCR product produced, was of high concentration and the non-specific band not visible by ethidium bromide staining (Figure 3.3).

It is also important to know the origin of the PCR product before cloning, as this will simplify the screening process thereafter. The origin of the PCR products of the isolates TV016, TV012, TV015, TV008, TV019, and TV013 were confirmed to be HIV-1 *gag*, by hybridisation with an HIV-1 *gag* specific probe (Figure 3.2). The probe is detected for by an alkaline phosphatase conjugated antibody, which causes a colour reaction in the presence of the substrate, 5-bromo-4-chloro-3-indolyl phosphate/ nitro blue tetrazolium (NBT/ BCIP). The presence of the conjugated antibody is identified as a coloured spot. Although non-specific background coloration is also possible, the colour development is not as strong as for the specific reaction. A positive reaction is therefore identified as a coloured spot, more intense in colour than the negative background stain.

The remainder of the subtype C isolates were amplified, using the optimised conditions (Figure 3.3). All the products, except for the product of TV003, were used for cloning. The PCR product of the isolate TV003 contained a smaller non-specific band of significant amount when compared to the ~1.5 kb band (2:1). The same reason for exclusion was used for TV003 as was used for the isolate TV013 (Figure 3.1). Although the PCR products of the isolates TV002 and TV007 contained several smaller non-specific bands, the products were used for cloning.

The PCR products of the isolates TV002, TV003, TV004, TV005, TV006, TV007, TV008, TV010 and TV016 (Figure 3.4), produced for cloning into pcDNA3.1, were all considered to be homogenous enough for cloning. Although the PCR product of the isolate TV007 displayed a larger non-specific DNA band (Figure 3.4), the product was used for cloning, as the ~1.5 kb band was smaller and would be favoured for incorporation during cloning. Although the isolate TV016 was an HIV-1 subtype B isolate, it was used to generate a clone for use as a subtype B reference. All the PCR products were confirmed to be of HIV-1 *gag* origin, by hybridisation with an HIV-1 *gag* specific probe.

The amplification of *gag* from RT-PCR products resulted in ~ 1.5 kb products that were visible as very weak bands after electrophoresis on an agarose gel (Figure 3.5). The products were however used to clone into the vector pcDNA3.1. The weak products were probably due to low concentration of starting template, as the RNA was only converted to DNA, without any amplification and the non-amplified DNA used as starting template in the PCR. The amplification of the other 13 subtype C *gag* genes used an amplified template and was therefore a nested PCR.

4.1.1.2 Cloning of *gag*

The reagents used in the PCR reaction, like free dNTPs, can interfere in the cloning reactions and must therefore be removed from the amplified DNA. The products were purified by using the Qiagen kit, which uses spin columns containing silica gel membranes that bind nucleic acids larger than 100 bp and efficiently remove unincorporated nucleotides, primers smaller than 40 bp and enzymes. Adjusting the pH on the silica membrane elutes the DNA. The sephadex columns used to purify the restriction digested PCR products worked on the same principle as the Qiagen cloumns.

The orientation of the insert was not important for the characterisation studies and *gag* was firstly cloned into the pCR[®]3.1 vector, using TA-cloning, where the insert could be inserted in the forward or reverse orientation. For the expression studies however, the insert needed to be in the forward orientation for expression of the gene. The isolates of which there was no pCR[®]3.1 clones with inserts in the forward orientation, were directionally cloned into the pcDNA3.1(-) vector. The vector pcDNA3.1(-) contained an *Xba*I recognition site upstream of the recognition site for the restriction enzyme, *Eco*RI, which correlates to the orientation of these two enzymes on the PCR products. Cleavage with *Xba*I and *Eco*RI results in incompatible ends, limiting the possibility of self-ligation of the vector onto itself and ensuring that the insert, cleaved with these two restriction enzymes, is cloned in only one possible orientation.

4.1.1.3 Sequencing of gag

4.1.1.3a Nucleotide Sequencing Analyses

4.1.1.3a(i) Sequencing Data

Sequencing analyses of clones from all 15 HIV-1 subtype C isolates showed that the cloned *gag* differed in size ranging from 1.464 kb to 1.527 kb, with an average size of 1.485 kb (Table 3.3). The difference in size is largely due to inserts found in the Gag p6 domain, at the C-terminal of the Gag polyprotein (Appendix B5). The nucleic acid composition of the *gag* clones correspond to the nucleic acid composition of lentiviruses with a composition of 36% A, 18% C, 24% G and 22% T (Foley, 2000). The *gag* clones had a slightly lower thymine percentage (19-20%). The highest occurring nucleotide in the cloned *gag* is adenosine (36-38%). It has been shown that elevated levels of AU content in human mRNAs resulted in instability and low expression levels (Hentze, 1991). The AT content in the *gag* clones, which corresponds to the AU content in their mRNA transcripts, was found

to range between 55-58% (Table 3.3), which accounts for the dependency of *gag* expression on the viral protein, Rev.

The two nucleotide deletion in the sequence of the clone TV006G11 seems to be an artefact of cloning, because the clone TV006G97 of the same isolate does not contain this deletion. The clone TV006G97 contained an early stop codon, resulting in a truncated protein. This early stop codon was the result of a single nucleotide change from G to A. These two clones, TV006G97 and TV006G11, were not cloned from the same PCR product and the deletion and mutation could have been produced at PCR level as well.

4.1.1.3a(ii) Restriction Digest Analyses

Subtyping of HIV-1 subtype C viruses by RFLP was previously described (Van Harmelen et al, 1999b), which looks at the restriction patterns of p17 after digestion with the restriction enzyme *AluI*. According to this study, subtype C strains have two *AluI* recognition sites, resulting in 3 fragments of which two can be seen on a 4% agarose gel and can be distinguished from subtypes A, B and D by the size of the upper band after digestion. They found that the size of the largest fragment after *AluI* digestion was ~240 bp for subtype C, while subtypes A, B and D produced a ~ 160 bp fragment. According to the restriction digest analyses of our subtype C clone sequences, only one isolate, TV018, as well as one clone of the isolate TV019 contained two *AluI* recognition sites, while 10 isolates plus the remaining clone of the isolate TV019 contained three recognition sites and three isolates contained four. Restriction digest of the p17 fragment with *AluI* of the subtype C isolate containing two recognition sites, as well as all of the subtype C isolates with three recognition sites, except for both clones of the isolate TV006, would produce a ~ 270 bp fragment as the largest fragment. The difference between ~ 240 bp and ~ 270 bp is probably negligible on a 4% agarose gel. The smallest fragment produced by the subtype C isolates containing three recognition

sites, is however 10 bp long, which would result in a restriction pattern close to what was observed in the RFLP study and would therefore have subtyped these clones as subtype C sequences. The clones of the isolate TV006, with three *AluI* recognition sites, would produce a slightly different restriction pattern with a largest fragment of ~ 240 bp and a smallest fragment of ~ 70 bp, but would still result in a similar pattern on a 4% agarose gel to the pattern observed in the RFLP study. Similar restriction digest patterns on a 4% agarose gel would also have been observed for the clones of the isolates TV001 and TV012, with four *AluI* recognition sites. Erroneous subtyping would however have been done on the clone of the isolate TV003, where the largest fragment produced after *AluI* digestion would have been ~ 170 bp in length, grouping this isolate with the subtypes A, B, and D. The 650 bp fragment, which includes the p17 region and a part of the p24, used to subtype the A, B and D subtypes in the RFLP study, would have identified the TV003 isolate as a subtype D sequence after digestion with the restriction enzymes *Accl* and *Swal*, which would have produced an uncut fragment after both digestion reactions. To identify a subtype D sequence, the final *XmnI* restriction digest of the 650 bp fragment would result in a two fragment pattern, not observed in the TV003 isolate, which would probably have resulted in this clone not being able to be subtyped. The subtype B isolate TV016 contained all the *AluI* and *Accl* recognition sites at the appropriate positions necessary for the identification of subtype B sequences.

Sequencing analyses confirmed the finding that the isolate TV003 had a cleavage site for the restriction enzyme *XbaI* and not for *EcoRI*. None of the other 14 isolates contained a recognition site for *XbaI*. The restriction enzymes, *EcoRI*, *HgaI*, *KpnI*, *NarI*, *NdeI*, *NotI*, *SacI*, *SacII*, *Sall*, *SmaI*, *XhoI* and *XmaI*, had no recognition sites in any of the 15 isolates, which is useful information when designing primers for directional cloning, or when sub cloning and making use of the multiple cloning site of the vector. Unique recognition sites for different restriction enzymes differed between the isolates as well as between clones from

the same isolate, giving less confidence to the use of restriction analyses for HIV identification.

4.1.1.3a(iii) Subtyping

All the sequences were submitted to the NCBI HIV-1 subtyping database and confirmed their HIV-1 subtype C status, and also the subtype B status of the TV016 clone (Appendix B3). The two clones of the isolate TV006, however showed a sequence stretch, between nucleotides 501 and 800, which was most similar to a subtype B strain (Figure 3.12 C & D), suggesting recombination in this area. On closer inspection of this region on the similarity plot, it became clear that the TV006 sequences in the region between 501 and 800, are most similar to only one subtype B sequence and was also quite similar to sequences of other subtypes as well.

Bootscreening is a procedure for the identification of HIV-1 recombinants and for the mapping of their recombination breakpoints (Salminen et al, 1995). The bootscreening tool in SimPlot delivered similar results for both clones of the isolate TV006 (Figure 3.13). Although the region 100 bp to 400 bp in the bootscreen graph showed a similarity to the subtype B strains, this was not considered as a recombination point, because there was no sharp drop in the similarity to one of the subtype C reference strains. In the bootscreen graphs, where the subtype B reference strain is K03455 (Figures 3.13 A & C), a definite recombination breakpoint can be seen in the region between 600 bp and 1000 bp, with both remaining reference sequences dropping sharply. This recombination breakpoint corresponds to the region of recombination seen in the similarity plots (Figures 3.12 C & D). When compared to the other subtype B reference strains, M17451 and TV016G95 (Figures 3.13 B, D, E, F), however, no recombination is observed in this region, suggesting that the TV006 sequences in that region were only similar to the subtype B strain, K01455 and not to the subtype B strains, M17451 or

TV016G65. The strain, K03455 is the lab strain, HXB2, which is not found in our laboratory and it is highly unlikely that there could have been recombination between the TV006 isolate and HXB2. The region between 600 bp and 1000 bp lies within the capsid domain, which is very conserved. It might be that the sequence of TV006 in this region is not typical subtype C, but similar to a range of subtypes, which is supported by the *AluI* and *Accl* restriction maps in this region of the TV006 clones that is comparable to the *AluI* and *Accl* restriction map for the subtype B isolate TV016. It was previously found that the *AluI* restriction pattern for the first ~ 700 bp of *gag* were similar for the subtypes A, B and D (Van Harmelen et al, 1999b). The sequence of TV006 was therefore not considered as a recombinant sequence, but a subtype C sequence with a small region of ~ 400 bp not typical subtype C, and would not lie as an outlier sequence on a phylogenetic tree comparing the subtype C sequences.

4.1.1.3a(iv) Phylogenetic Analyses

The phylogenetic tree, comparing the full-length *gag* DNA sequences of our subtype C isolates and the Los Alamos reference subtypes (Figure 3.14) also confirmed the NCBI HIV-1 subtyping. All of our isolates, except for the isolate TV016, clustered with the subtype C reference strains. As was expected, the isolate TV006 clustered with the rest of the subtype C strains, confirming that this sequence was indeed not a recombinant. In previous studies involving the South African subtype C isolates, it was established that the accessory genes (Scriba et al, 2001), the *env* gene (Engelbrecht et al, 2001) and the LTRs (Scriba et al, in press) of these isolates were all subtype C, strongly suggesting that the isolates are not recombinants. The isolate TV016 clustered with the subtype B strains.

Although the different subtypes all formed clusters, which were supported by high bootstrap values, there appears to be only six main subclusters within the M group for the *gag* sequences (Figure 3.14). Several subtypes formed clusters with other

subtypes, which were supported by high bootstrap values. The subtypes F and K formed a cluster, as well as subtypes B and D. The CRF05_DF formed a subcluster within the subtype D cluster, suggesting that the *gag* region of the CRF05_DF strains is strictly subtype D. The two sub-subtypes F1 and F2 formed subclusters within the subtype F cluster, indicating two definite subgroups of the F subtype. The subtypes G, A and CRFs 06_CPX, 03_AB, 02_AG, 04_CPX, and 01_AE are all part of one cluster with the subtype G forming a separate subcluster from the rest. The CRFs clustering with the subtype A strains are all recombinant forms containing subtype A sequences (Kuiken et al, 2000), which explains this clustering with subtype A. All however formed definite subclusters of their own, which were supported by high bootstrap values. The CRF 02_AG strain, G829, does however not cluster with the rest of the CRF 02_AG strains, suggesting that the recombination breakpoints in the G829 sequence might differ from that of the other CRF 02_AG sequences.

The clustering of different subtypes (Figure 3.14) indicate that these groups of subtypes have more recent common ancestors than the common ancestor for the Group M subtypes. The subtype C strains however form a single cluster, with the Brazilian and Ethiopian strains forming a separate subcluster within the subtype C cluster.

A tree was also constructed with our isolates and all the HIV-1 subtype C *gag* sequences available in the Los Alamos database and rooted with the subtype B strain, HXB2 and our isolate TV016 (Figure 3.15). Most of the high bootstrap values were on branches connecting clones. A significant clustering was however seen in the strains from India (93IN101, 93IN905, 93IN904, 94IN11246, AF209990 and 95IN21068), which are from different isolates, suggesting a single introduction which spread and resulted in this clustering. There are however two strains from India (98IN012 and 94IN476), which fall into a different subcluster, although not supported by high bootstrap values, indicating a different ancestor to the ancestor responsible for the majority of the Indian strains in the Los Alamos database.

Significant clustering was also seen between the clones from the isolates TV014 and TV008 and between the clones of the isolates TV001 and TV006 (Figure 3.15). Such clustering suggest a common ancestor more recent than the ancestor for all the subtype C strains. The only thing that the isolates TV014 and TV008 have in common is that they were infected outside the borders of South Africa. It is however impossible to make further assumptions with the available patient data. The South African subtype C isolates lie scattered between the other subtype C strains with no particular clustering with any of the countries represented by sequences on this tree. These data suggest multiple introductions of the HIV-1 subtype C virus into South Africa.

A distance matrix was calculated for the different alignments of the *gag* sequence. The nucleotide sequence of *gag* showed a variation of 6-12% between the clones and 6-13% within subtype C. The diversity between the clones is therefore comparable to the diversity seen between the the *gag* sequences of all the different strains of HIV-1 subtype C. An intersubtype variation of 8-19% was seen, while the inter-group variation was calculated between 28-36%, confirming that Gag is relatively conserved when compared to other viral proteins. The most variation was seen in the p17 matrix protein, which correlates to the finding that this protein is not essential for viral replication (Reil et al, 1998), as the most variable regions are usually dispensable regions.

4.1.1.3b Predicted Amino Acid Sequence Analyses

4.1.1.3b(i) Matrix

The HIV-1 matrix protein is involved in various processes in the viral life cycle. Freed and his colleagues did a series of analyses to establish the effects of single amino acid mutations on viral particle formation (Freed et al, 1994). The N-

terminal glycine residue (G1), involved in N-myristoylation (Göttlinger et al, 1989; Bryant & Ratner, 1990), was present in all the isolates except in the one clone of the isolate TV006 (TV006G11) (Appendix B5). All except two of the amino acid residues that Freed found to play a role in viral replication or particle production were conserved in all our isolates. The valine (V7) residue at position seven was present in only three of our isolates (TV002, TV005 and TV008). Freed found that a mutation at this position to a basic arginine residue resulted in a ten day delay of reverse transcriptase production (Freed et al, 1994). The remainder of the isolates contained an isoleucine (I7) residue at this position, with the exception of the clone TV006G97, which contained a threonine (T7). Both valine and isoleucine are aliphatic amino acids and are two of the most hydrophobic amino acids, with isoleucine the more hydrophobic of the two. The isoleucine at position seven was also found in all the subtype C strains, except for the strain of Botswana (96BW15C02), which had an arginine (R7) (Appendix B5). The Botswana strain, 96BW15C02, was a clone of the same patient as the strains, 96B15C05 and 96BW15B03 (Novitsky et al, 1999), which contained an isoleucine residue (I7), suggesting that the mutation might have been PCR induced. The Isoleucine (I7) was also found in both the HIV-1 subtype J reference strains, SE7022 and SE7887 (Kuiken et al, 2000). It can be postulated that if a mutation in position seven to a basic, hydrophilic residue like arginine, results in a delay in viral replication, then a mutation to a more hydrophobic residue could hasten viral replication. Isoleucine and valine are most probably too similar in their hydrophobic nature to result in any detectable difference, as was observed by Ono and his colleagues (Ono et al, 1997).

Two nuclear localisation signals (NLS) NLS-1 (²⁴GKKKYKLLKH) (Bukrinsky et al, 1993; von Schwedler et al, 1994) and NLS-2 (¹¹⁰KSKKKAQ) (Haffar et al, 2000) have been found within the matrix protein. The amino acid sequences of the published NLS regions are those of the HIV-1 subtype B strain, HXB2. Although only the isolate TV005 contained a lysine (K28) in NLS-1, the remaining isolates, except TV007 and TV009, contained either arginine or histidine residues (Appendix

B5), which are both basic amino acids. The isolate TV007 contained a threonine residue, which has a hydroxyl-containing side chain, while the isolate TV009 had the least hydrophobic, aliphatic amino acid, Glycine at the position 28. NLS-2, which contain a series of basic amino acids, was less conserved in all the isolates. The basic amino acid confers a hydrophilic property to this region and although the amino acid sequence was not conserved, the hydrophilic nature of this area was certainly maintained in all of the isolates (Figure 3.16).

The Leucine residues at positions 13 and 31, as well as the valine residue at position 35, all three critical for the incorporation of the HIV-1 transmembrane envelope glycoprotein into virion particles (Freed and Martin, 1995, 1996), were maintained in all our isolates (Appendix B5).

4.1.1.3b(ii) Capsid

The amino acid positions, mentioned below are as found within the capsid, starting from the first amino acid in the capsid domain as position one. The single mutation of a deoxyguanosine to a deoxyadenosine at position 551 found in clone TV006G97, resulting in a glycine to arginine mutation at position 89 of the capsid domain (G89R) (Appendix B5), could have been induced by either PCR or cloning. The clone TV006G11 does not contain this mutation. It was previously found that a mutation of the proline residue at position 90 (P90) or the glycine residue (G89) immediately preceding it in the capsid domain, disrupts the incorporation of cyclophilin A (CypA) into the virions (Braaten et al, 1996; Thali et al, 1994; Luban, 1996). All the clones of all the isolates contained a proline at position 90 in the capsid domain. Two additional sites, both containing the glycine-proline motif, for cyclophilin incorporation was found at positions 157 and 224 in the capsid domain (Endrich et al, 1999). The glycine-proline motifs at these two positions were conserved in all our isolates.

It was previously found that the phosphorylation of three serine residues at positions 109, 149 and 178 was essential for the viral uncoating process (Cartier et al, 1999). All our isolates contained serine residues at positions 109 and 149, but only the isolate TV002 contained a serine at position 178. The remainder of the isolates all contained a threonine (T178) at this position. Although six subtype B strains also contained threonine residues at position 178, the majority of 69 HIV-1 subtype B strains and all of the six subtype D strains in the Los Alamos database contained a serine (S178) residue at this position (Kuiken et al, 2000). The threonine residue was also the dominant amino acid at position 178 in the majority strains of the subtypes A, G, H, J and K. Both phosphoserines and phosphothreonines were previously found in the capsid protein, suggesting that both serine and threonine could be phosphorylated on their hydroxyl groups (Di Marzo Veronese et al, 1988).

The major homology region (MHR) in the capsid domain was well conserved in all of our subtype C isolates (Appendix B5), with the critical glutamine (Q), glutamic acid (E), and arginine (R) residues at positions 155, 159 and 167 respectively (Mammano et al, 1994), 100% conserved in all. At position 164, all the isolates contained the aromatic amino acid tyrosine, which correlated to the finding that different retroviruses contained an aromatic residue at this. Residues 161 to 173 of the capsid protein, which falls in the MHR, forms an amphipathic α -helix containing both an hydrophobic and hydrophilic side (Clish et al, 1996). Our isolates maintain this amphipathic helix-wheel model (Figure 3.17 and Appendix B5), with substitutions in the residues joining the hydrophilic- and hydrophobic sides (residues 22 and 24 in Figure 13.17), still maintaining the amphipathic character of the helix.

The stretch of amino acids 209 to 213 (ATLEE) and 218 and 219 (CQ) were found to be essential for the formation of HIV-1 particles (von Plobtzki et al, 1993). These regions were well conserved in our isolates, except for the residue at the position 210, which was either a threonine or serine (Appendix B5). Both amino

acids, threonine and serine contains an hydroxyl group in their side chains and are similar in structure, suggesting that particle formation would occur with a serine at position 210. The isolate TV003, however displayed a glycine residue at position 213, which is not similar to the preferred glutamate residue. Unfortunately there was only one clone sequenced for this isolate, but as this glutamate to glycine change was not observed in any of the other subtype C strains or of the other reference subtypes, this mutation can be seen as an artefact of cloning or PCR.

4.1.1.3b(iii) Nucleocapsid

The C-X₂-C-X₄-H-X₄-C motif of the zinc finger in the nucleocapsid (Berg, 1986) was maintained at both sites in all our isolates (Appendix B5). The cysteine to tyrosine change, seen in clone TV019G20 was seen to be a cloning artefact, as the other clone of this isolate did not contain this change. The consensus peptide sequence connecting the two zinc-binding domains was RAPRKKG. This sequence of the linker peptide was previously found to be important for virion formation and infectivity, possibly due to the spatial proximity it renders to the two zinc fingers (Morellet et al, 1994). Similar to the finding with the mutation in the clone TV019G20, the mutation in the linker peptide found in the clone TV012G34 was seen as a cloning artefact or PCR.

4.1.1.3b(iv) Spacer Peptides and p6

Amino acid alignments showed that four isolates (TV001, TV007, TV012 and TV013) contained a seven to ten amino acid repeat and one isolate (TV013) displayed a further four amino acid extension in the Gag p6 region (Appendix B5). These amino acid repeats were also observed previously (Barrie et al, 1996) and might also have a significant effect on protease function, as the domain upstream of the protease sequence influence the protease autoprocessing (Zybarth & Carter,

1995). The exact function of the sequence repeats or the influence it has is however not known.

4.1.1.3b(v) Protease cleavage sites

The matrix-capsid protease cleavage site (SQNY*PIVQN) was well conserved in all the isolates, except in the isolate TV005, where the C-terminal aromatic tyrosine was replaced with another aromatic amino acid, phenylalanine (Appendix B5). This correlates to the finding that the matrix-capsid cleavage site falls in the class I category where an aromatic amino acid and a proline residue flank the site of cleavage (Griffiths et al, 1992; Pettit et al, 1991). The proline residue at position one of the capsid domain, which is important in the refolding of the N-terminal into a β -hairpin/helix structure after proteolysis (von Schwedler et al, 1998) was conserved in all the isolates. This β -hairpin/helix structure is stabilised by the formation of a salt bridge between the proline residue and an aspartate residue at position 51, which was conserved in all the isolates. The second cleavage site between the capsid protein and the p2 spacer protein (KARVL*AEAMS) was also well conserved in the isolates, except for the one clone of TV012, TV012G34, which had a valine to alanine substitution at the C-terminal region of the capsid domain. The other TV012 clone did not have this change, suggesting that the alanine residue is not the natural occurring amino acid at this position, although both the amino acids valine and alanine are similar hydrophobic, aliphatic amino acids. The Gag p2-p7 nucleocapsid cleavage site was less conserved, especially in the p2 region. The Gag p2 region was also previously found to be variable in sequence and composition in different HIV-1 isolates (Louwagie et al, 1993.). The IM*MQ motif was however maintained in 11 of the 15 isolates. The hydrophobic/hydrophobic character of the cleavage site (Griffiths et al, 1992; Pettit et al, 1991) was also maintained in all the clones of the isolates TV012 (VM*MQ/ VM*IQ), TV008 (IL*MQ/ IF*MQ), TV006 (IL*MQ), and TV001 (IL*VQ). The p7-p1 and p1-p6 cleavage sites were further well conserved.

4.1.1.3b(vi) CTL Epitopes

Several studies have indicated the importance of HIV-specific and more specifically, Gag-specific CTL for controlling the virus in infected individuals. There are three major areas in the matrix domain in which published CTL epitopes fall. These three areas display a great percentage variation in their amino acid sequences (4-26%, 0-18%, 0-45%), with the most variation seen in the C-terminal area. Variation in CTL epitopes result in viral escape from the cell mediated immune response. Epitopes that are well conserved are therefore more desirable to include into a vaccine. The p24 capsid domain is quite conserved with the different HIV-1 groups varying between 21% and 28% from each other, while the subtype C strains varied between 4% and 9%. CTL epitopes have been described for almost the whole capsid domain, except for a 10 amino acid region at the C-terminal. The dominant CTL epitopes found in South African adults and children are also situated in the capsid (Goulder et al, 2000). The three p24 epitopes were **SALSEGATPQDLNMLNTVG**, **FRDYVDRFFKTLRAEQA**, and **SILDIKQGKEPFRDY**, of which the last two epitopes shared four amino acids. All three epitopes were well conserved in all our isolates, as well as in the rest of the subtype C strains. Interestingly 64 out of the 65 subtype C sequences in the predicted amino acid alignment (Appendix B5) contained a threonine in the first position of the first epitope (**TALSEGATPQDLNMLNTVG**), but the remainder of the amino acids in the epitope were identical to the consensus subtype B sequence. The Gag p2-p6 region contains three published CTL epitopes. The first region lies within the second CCHC motif of the zinc finger and is well conserved in all of our isolates. The second CTL epitope lies in the proline rich area where several repeat amino acid sequences occur and is therefore vary variable. The third area is again quite conserved. The recognition of CTL epitopes is dependent on the histocompatibility leukocyte antigen (HLA) haplotype of an individual cell and varies from person to person. The HLA recognises degraded peptides and displays the peptide on the

surface of the infected cell, where the peptide or epitope is recognised by the immune system. Each person would therefore recognise different epitopes and it would therefore be wise to include as many different epitopes especially of conserved regions in a vaccine. Inclusion of Gag in a vaccine would present 187 published epitopes, including the highly conserved 127 epitopes within the capsid domain, to the immune system (Korber et al, 2000).

4.1.1.4 DNA Quality Assurance

Careful lab work is not sufficient to prove that no contamination occurred and it is therefore important to include a sequence quality control to screen for possible contamination (<http://hiv-web.lanl.gov/seq-db.html>). A phylogenetic tree was constructed using the sequences of the *gag* clones and a 160 bp sequence within *gag* from the most original patient RNA available (Figure 3.18A). The *gag* clone sequences clustered with the corresponding patient RNA sequences. The branching order of 10 out of the 15 subtype C isolate clusters were supported by high bootstrap values. For conserved regions with little variation, the Los Alamos HIV sequence quality control protocol (http://hiv-web.lanl.gov/CONTAM/contam_main.html) suggests the construction of a Neighbour-joining tree, based on synonymous substitutions (Figure 3.18B). The phylogenetic tree based on synonymous substitutions delivered a tree with a similar topology as the tree based on all substitutions. The clustering of the isolates TV008 and TV014 was not seen as proof of contamination, as these two isolates were isolated and stored at different time points. The clustering of all of the sequences of these two isolates suggests that the contamination occurred at the time of virus isolation, which is very unlikely as these two samples were isolated four months apart from each other (data not shown). The variation between the full length *gag* sequences of the clones of the isolates TV008 and TV014 (6%) is significant enough to rule out contamination. Also, the fact that the phylogenetic tree was constructed with such

a small sequence length gives greater confidence to the areas with high bootstrap values. The branching orders with lower bootstrap values were accepted.

4.1.2 Expression Studies

4.1.2.1 *In vitro* transcription/ translation

Surprisingly only seven of all the clones could produce a ~55 kDa protein (Figure 3.19). One would expect that clones with such similar sequences would all be able to produce proteins. One would especially have expected the sister clones of TV013G15 and TV001G11, namely TV013G2 and TV001G8, to produce proteins as they are so similar in sequence. Two of the clones were in the vector pCR[®]3.1 and the remaining five were in the vector pcDNA3.1, suggesting that both vectors are capable of producing proteins, although pcDNA3.1 appears to be the better vector for protein production. The DNA quality might be considered a culprit in the low percentage of clones producing proteins. Although the protocol does not call for high quality DNA, it does warn to be mindful of RNase that can destroy the transcribed RNA. The normal miniprep DNA used in the *in vitro* transcription/ translation assay was prepared in a laboratory where the use of RNase is common practise and this might have contaminated certain tubes of DNA. There are also many bacterial proteins present in miniprep DNA, which might have some influence on either the transcription or translation processes. An indirect indication that DNA quality might play a role in outcome of *in vitro* transcription/ translation is the fact that the clone TV001G11 produced a protein with fresh DNA, but failed to do so again when repeated with older DNA. Some of the negative clones were repeated (Figure 3.19 D), but unfortunately with the same DNA and the outcome was the same.

4.1.2.2 Evaluation of the transfection techniques

Although the highest concentration of the CAT protein was produced in Vero cells transfected by using the CaPO₄ method, the electroporation of Vero cells was decided to be best method to continue with (Figure 3.20). The electroporated cells are plated on a smaller flask area than cells transfected by CaPO₄. The 2.26 ng/ml of CAT was therefore produced by fewer cells than the amount of cells that produced the 2.55 ng/ml of CAT (Table 3.5 and 3.6). The Vero cells produced the higher amount of protein in all of the transfection techniques tested (Figure 3.20). This correlates to the cell growth properties of the two cell lines where the Vero cells were the faster growing cells.

4.1.2.3 Gag protein production in cell culture

Four of seven protein producing clones showed production of a ~55 kDa protein in cell culture that was detectable by Western blot (Figure 3.21). The expression was very weak, which was however not a surprise as the native *gag* sequence contains various inhibitory sequences (Maldarelli et al, 1991; Olsen et al, 1992; Schneider et al, 1997), which decreases the stability of *gag* transcripts and translation efficiency, making the expression dependant of the viral protein Rev. Studies have shown that cellular proteins interact with these inhibitory sequences on the mRNA transcripts, forming a protein-mRNA complex, which is retained in the nucleus and degraded. As discussed earlier, the AT content of the cloned *gags* ranged between 55% and 58% (Table 3.3). For highly expressed genes a G or C is generally preferred over an A or T. Several studies have made *gag* constructs, involving the elimination of the inhibitory sequences and/ or codon optimising (Qiu et al, 1999; Kotsopoulou et al, 2000; Zur Megede et al, 2000), which increased the expression of Gag several fold. Although codon optimisation would have significantly improved expression of Gag, the process of codon optimising a 1.5 kb construct is quite expensive and we did not consider this for this study.

4.2 Conclusion

The full length *gag* gene of 23 *gag* clones of 15 HIV-1 subtype C isolates as well as one clone of one HIV-1 subtype B isolate was amplified, cloned and sequenced. The cloned *gag* genes varied between 1.464 kb to 1.527 kb in length and contained a nucleotide composition comparable to that of the different lentiviruses. The sequences of clones from 14 HIV-1 subtype C isolates contained intact open reading frames, while the clones of the isolate TV006 contained either a frame shift deletion at the beginning of the clone or an early stop codon resulting in a truncated protein. Phylogenetics confirmed that the 15 subtype C isolates were indeed HIV-1 subtype C strains. Subtyping by RFLP (Van Harmelen et al, 1999a) would have correctly subtyped 14 of the HIV-1 subtype C isolates as well as the HIV-1 subtype B strain. The clone of the isolate TV003 would have been erroneously subtyped using the RFLP method.

The sequences of the isolate TV006 showed a region, ~400 bp in length, of probable recombination, which could be described as a region atypical of subtype C, but similar to a range of subtypes. The greater part of the TV006 sequences was however typical HIV-1 subtype C sequences. When comparing the predicted amino acid sequences of the two clones of the isolate TV006 to previously published mutational analyses of different regions of Gag, the two sequences gave in more than one cases inconsistent results. This might be due to fact that the clones were generated from two different PCR products. One could also postulate that the patient, from whom the isolate TV006 was isolated, contained quasispecies, resulting in a virus culture containing diverse strains within the culture. When a clone is established, an homogenous population is isolated and when compared to other clones obtained from the same culture, the diversity between the clones can be determined. The two clones of the isolate TV006 varied 98% between each other, which is not significantly different. The *gag* gene is however well conserved and it would be interesting to look at less conserved regions of the HIV-1 genome to establish the homogeneity of the TV006 culture.

The South African HIV-1 subtype C strains do not form a cluster with any of the other subtype C strains from various countries and therefore suggest multiple introduction of the subtype C strain into South Africa. The strains from Brazil and Ethiopia fall in a different subtype C subcluster as the South African strains, suggesting that these strains are more distantly related than the strains from Botswana, India, Israel, Tanzania and Zambia.

The native *gag* gene does not elicit good expression of a protein, due to the high AT content and the presence of inhibitory sequences within the gene. Sequence modification whereby the GC content is increased without altering the amino acid sequence, have been shown to increase protein expression several times (Zur Megede et al, 2000; Qiu et al, 1999; Kotsopoulou et al, 2000). It has also been shown that good CTL responses can be induced in humans by immunisation with a plasmid based DNA vaccine (Calarota et al, 1998, 1999; MacGregor et al, 1998). The sequence of any of the subtype C strains can be used to modify for use as a Southern African DNA vaccine, as the diversity between the South African strains are comparable to the diversity between all the different subtype C strains. It would also be wise to include the full length modified *gag* gene in the DNA vaccine, as this would present 187 published CTL epitopes to the immune system. In such a racially heterogeneous country as South Africa, it would be important to include as many CTL epitopes in the vaccine, as people with different HLA haplotypes, recognise different epitopes, and limiting the number of epitopes would limit the range and efficiency of the vaccine.

The second aim of this study was to establish the transfection technique and Western blot protocol in our laboratory. Since the completion of this study, other cloned HIV-1 genes have successfully been transfected into mammalian cells using the electroporation technique and the proteins produced were screened for by Western blot.

CHAPTER FIVE

5. References

- Aldovini A, Feinberg MB. Transfection of molecularly cloned HIV genomes. In: Aldovini A, Walker BD, eds. *Techniques in HIV research*. Stockton Press. New York. 1990; 147-175.
- Bachand F, Yao X-J, Hrimech M, Rougeau N, Cohens EA. Incorporation of Vpr into Human Immunodeficiency virus type 1 requires a direct interaction with the p6 domain of the p55 Gag precursor. *J. Biol. Chem.* 1999; 274; 9083-9091.
- Barnett SW, Lu S, Srivastava I, et al. The ability of an oligomeric Human Immunodeficiency virus type 1 (HIV-1) envelope antigen to elicit neutralizing antibodies against primary HIV-1 isolates is improved following partial deletion of the second hypervariable region. *J. Virol.* 2001; 75; 5526-5540.
- Barré-Sinoussi F, Chermann JC, Rey F, et al. Isolation of a T-Lymphotropic Retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science.* 1983; 220; 868-871.
- Barrie KA, Perez EE, Lamers SL, et al. 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.* 1996; 219; 407-416.

Becker M.L., De Jager, G. and Becker, W.B. Analysis of partial *gag* and *env* gene sequences of HIV type 1 strains from southern Africa. *AIDS Res. Hum. Retroviruses*. 1995: 11; 1265-1267.

Berg JM. Potential metal-binding domains in nucleic acid binding proteins. *Science*. 1986: 232; 485-487.

Borrow P, Lewicki H, Hahn BH, Shaw GM, Oldstone MBA. Virus-specific CD8+ cytotoxic T-lymphocyte activity associated with control of viremia in primary Human Immunodeficiency virus type 1 infection. *J. Virol*. 1994: 68; 6103-6110.

Braaten D, Aberham C, Franke EK, Yin L, Phares W, Luban J. Cyclosporin A-resistant Human Immunodeficiency virus type 1 mutants demonstrate that Gag encodes the functional target of cyclophilin A. *J. Virol*. 1996: 70; 5170-5176.

Bredell H, Williamson C, Sonnenberg P, Martin DJ, Morris L. Genetic characterisation of HIV type 1 from migrant workers in three South African gold mines. *AIDS Res. Hum. Retroviruses*. 1998: 14; 677-684.

Broder S, Gallo RC. A pathogenic retrovirus (HTLV-III) linked to AIDS. *N. Engl. J. Med*. 1984: 311; 1292-1297.

Bryant M, Ratner L. Myristoylation-dependant replication and assembly of human immunodeficiency virus 1. *Proc. Natl. Acad. Sci. USA*. 1990: 87; 523-527.

Bukrinsky MI, Sharova N, McDonald TL, et al. Association of integrase, matrix, and reverse transcriptase antigens of human immunodeficiency virus type 1 with viral nucleic acids following acute infection. *Proc. Natl. Acad. Sci. USA*. 1993: 90; 6125-6129.

Buseyne F, Chaix M-L, Rouzioux C, Blanche S, Rivière Y. Patient-specific cytotoxic T-lymphocyte cross-recognition of naturally occurring variants of a Human Immunodeficiency virus type 1 (HIV-1) p24gag epitope by HIV-1 infected children. *J. Virol.* 2001; 75; 4941-4946.

Calarota S, Bratt G, Nordlund S, et al. Cellular cytotoxic response induced by DNA vaccination in HIV-1 infected patients. *Lancet.* 1998; 351; 1320-1325.

Calarota SA, Leandersson A-C, Bratt G, et al. Immune responses in asymptomatic HIV-1 infected patients after HIV-DNA immunization followed by highly active antiretroviral treatment. *J. Immunol.* 1999; 163; 2330-2338.

Cao J, Sullivan N, Desjardin E, et al. Replication and neutralization of Human Immunodeficiency virus type 1 lacking the V1 and V2 variable loops of the gp120 envelope glycoprotein. *J. Virol.* 1997; 71; 9808-9812.

Cartier C, Sivard P, Tranchat C, et al. Identification of three major phosphorylation sites within HIV-1 capsid: Role of phosphorylation during the early steps of infection. *J. Biol. Chem.* 1999; 274; 19434-19440.

Chan DC, Kim PS. HIV entry and its inhibition. *Cell.* 1998; 93; 681-684.

Claassen M. The growth of HIV-1 C subtype isolates in large volumes at high concentrations for further studies. BScMedScHons. University of Stellenbosch. 1999.

Clavel F, Guetard D, Brun-Vezinet F, et al. Isolation of a new human retrovirus from West African patients with AIDS. *Science.* 1986; 233; 343-346.

Clerici M, Shearer. Correlates of protection in HIV infection and the progression of HIV infection to AIDS. *Immunol. Lett.* 1996: 51; 69-73.

Clish CB, Peyton DH, Barklis E. Spectroscopic study of an HIV-1 capsid protein major homology region peptide analog. *FEBS Lett.* 1996: 378; 43-47.

Coffin J, Haase A, Levy JA, et al. Human Immunodeficiency viruses. *Science.* 1986a: 232; 697.

Coffin J, Haase A, Levy JA, et al. What to call the AIDS virus? *Nature.* 1986b: 321; 10.

Coffin JM. Retroviridae: The Viruses and Their Replication. In: Fields BN, Knipe DM, Howley PM, eds. *Fields Virology*. Third Edition. Lippincott-Raven Publishers. Philadelphia, USA. 1996; 1767-1847.

Cohen SN, Chang AC, Boyer HW, Helling RB. Construction of biologically functional bacterial plasmids in vitro. *Proc Natl Acad Sci USA* 1973: 11; 3240-4.

Connor R, Sheridan K, Ceraldini D, Choe S, Landau N. Changes in coreceptor use correlates with disease progression in HIV-1-infected individuals. *J. Exp. Med.* 1997: 185; 621-628.

De Cock KM, Binkin NJ, Zuber PLF, Tappero JW, Castro KG. Research issues involving HIV-associated tuberculosis in resource-poor countries. *JAMA.* 1996: 276; 1502-1507.

De Leys R, Vanderborght B, van den Haesevelde M, et al. Isolation and partial characterization of an unusual Human Immunodeficiency Retrovirus from two persons of West-Central African origin. *J. Virol.* 1990: 64; 1207-1216.

Department of Health / Directorate Health Systems Research: Seventh National HIV survey of women attending antenatal Clinics of the public health services in South Africa. October/November 2000. Pretoria: Directorate Health systems Research, Department of Health; 2001.

Dettenhofer M, Yu X-F. Proline residues in Human Immunodeficiency virus type 1 p6^{Gag} exerts a cell type-dependent effect on viral replication and virion incorporation of Pol proteins. *J. Virol.* 1999; 73; 4696-4704.

Di Marzo Veronese F, Copeland TD, Oroszlan S, Gallo RC, Sarngadharan MG. Biochemical and Immunological analysis of Human Immunodeficiency virus gag gene products p17 and p24. *J. Virol.* 1988; 62; 795-801.

Durali D, Morvan F, Letourneur D, et al. Cross-reactions between the cytotoxic T-lymphocyte responses of human immunodeficiency virus-infected African and European patients. *J. Virol.* 1998; 72; 3547-3553.

Endrich MM, Gehrig P, Gehring H. Maturation-induced conformational changes of the HIV-1 capsid protein and identification of two high affinity sites for cyclophilins in the C-terminal domain. *J. Biol. Chem.* 1999; 274; 5326-5332.

Engelbrecht S, Van Rensburg EJ. Detection of Southern African human immunodeficiency virus type 1 subtypes by polymerase chain reaction: evaluation of different primer pairs and conditions. *J. Virol. Meth.* 1995; 55; 391-400.

Engelbrecht S, Smith TL, Kasper P, et al. HIV type 1 V3 domain serotyping in Gauteng, Mpumalanga, KwaZulu-Natal, and Western Cape Provinces of South Africa. *AIDS Res. Hum. Retroviruses.* 1999; 15; 325-328.

Engelbrecht S, De Villiers T, Sampson CC, Zur Megede J, Barnett SW, Janse van Rensburg E. Genetic analysis of the complete *gag* and *env* genes of HIV-1 subtype C primary isolates from South Africa. *AIDS Res. Hum. Retroviruses*. 2001: in press.

Esparza J, Bhamarapavati N. Accelerating the development and future availability of HIV-1 vaccines: Why, when, where and how? *Lancet*. 2000: 355; 2061-2066.

Farnet CM, Bushman FD. HIV cDNA integration: molecular biology and inhibitor development. *AIDS (Suppl. A)*. 1996: 10; 3-11.

Felsenstein J. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution*. 1985: 39; 783-791.

Fitch WM. Toward defining the course of evolution: minimum change for a specific tree topology. *Syst. Zool.* 1971: 20; 406-416.

Fitzon T, Leschonsky B, Bieler K, et al. Proline residues in the HIV-1 NH₂-terminal capsid domain structure determinants for proper core assembly and subsequent steps of early replication. *Virology*. 2000: 268; 294-307.

Foley BT. An overview of the molecular phylogeny of lentiviruses. In: Kuiken C, Foley B, Hahn B, et al, editors. *HIV Sequence Compendium 2000*. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory. 2000.

Fraley R, Subramani S, Berg P, Papahadjopoulos D. Introduction of liposome-encapsulated SV40 DNA into cells. *J. Biol. Chem.* 1980: 255; 10431-10435.

Freed EO, Orenstein JM, Buckler-White AJ, Martin MA. Single amino acid changes in the Human Immunodeficiency Virus type 1 matrix protein block virus particle production. *J. Virol.* 1994; 68; 5311-5320.

Freed EO, Martin MA. Virion incorporation of envelope glycoproteins with long but short cytoplasmic tails is blocked by specific, single amino acid substitutions in the Human Immunodeficiency virus type 1 matrix. *J. Virol.* 1995; 69; 1984-1989.

Freed EO, Martin MA. Domains of the Human Immunodeficiency virus type 1 matrix and gp41 cytoplasmic tail required for envelope incorporation into virions. *J. Virol.* 1996; 70; 341-351.

Freed EO. HIV-1 Gag proteins: Diverse functions in the virus life cycle. *Virology.* 1998; 251; 1-15.

Freshney RI. *Culture of Animal Cells. A Manual of Basic Technique.* Second Edition. Alan R. Liss, Inc. New York. 1987.

Friedman SM, Felman YM, Rothenberg R, et al. Follow-Up on Kaposi's Sarcoma and Pneumocystis Pneumonia. *MMWR.* 1981; 30; 409-410.

Friedman-Kien A, Laubenstein L, Marmor M, et al. Kaposi's Sarcoma and Pneumocystis Pneumonia among homosexual men – New York City and California. *MMWR.* 1981; 30; 306-308.

Gallay P, Swingler S, Song J, Bushman F, Trono D. HIV nuclear import is governed by the phosphotyrosine-mediated binding of matrix to the core domain of integrase. *Cell.* 1995; 83; 569-576.

Gamble TR, Vajdos FF, Yoo S, et al. Crystal structure of human cyclophilin A bound to the amino-terminal domain of the HIV-1 capsid. *Cell*. 1996; 87; 1285-1294.

Gamble TR, Yoo S, Vajdos FF, et al. Structure of the carboxyl-terminal dimerisation domain of the HIV-1 capsid protein. *Science*. 1997; 278; 849-853.

Gao F, Yue L, White AT, et al. Human infection by genetically-diverse SIVsm-related HIV-2 in West Africa. *Nature*. 1992; 358; 495-499.

Gao F, Morrison SG, Robertson DL, et al. Molecular cloning and analysis of functional envelope genes from Human Immunodeficiency virus type 1 sequence subtypes A through G. *J. Virol*. 1996; 70; 1651-1667.

Gao F, Robertson DL, Carruthers CD, et al. A comprehensive panel of near-full-length clones and reference sequences for non-subtype B isolates of human immunodeficiency virus type 1. *J. Virol*. 1998; 72; 5680-5698.

Gao F, Bailes E, Robertson DL, et al. Origin of HIV-1 in *Pan troglodytes troglodytes*. *Nature*. 1999; 397; 436-441.

Gartner S, Popovic M. Virus Isolation and Production. In: Aldovini A, Walker BD, eds. *Techniques in HIV research*. Stockton Press. New York. 1990; 53-70.

Gitti RK, Lee BM, Walker J, Summers MF, Yoo S, Sundquist WI. Structure of the amino-terminal core domain of the HIV-1 capsid protein. *Science*. 1996; 273; 231-235.

Gottlieb MS, Schroff R, Schanker HM, et al. Pneumocystis carinii pneumonia and mucosal candidiasis in previously healthy homosexual men. Evidence of a new acquired cellular immunodeficiency. N. Engl. J. Med. 1981a: 305; 1425-1431.

Gottlieb MS, Schanker HM, Fan PT, Saxon A, Weisman JD. Pneumocystis Pneumonia – Los Angeles. MMWR. 1981b: 30; 250-252.

Göttlinger HG, Sodroski JG, Haseltine WA. Role of capsid precursor processing and myristoylation in morphogenesis and infectivity of Human Immunodeficiency virus type 1. Proc. Natl. Acad. Sci. USA. 1989: 86; 5781-5785.

Goulder PJR, Brander C, Annamalai K, et al. Differential narrow focusing of immunodominant Human Immunodeficiency virus Gag-specific cytotoxic T-lymphocyte responses in infected African and Caucasoid adults and children. J Virol. 2000: 74; 5679-90.

Graham FL, Van der Eb AJ. A new technique for the assay of infectivity of Human Adenovirus 5 DNA. Virology. 1973: 52; 456-467.

Grant AD, Djomant G, De Cock KM. Natural history and spectrum of disease in adults with HIV/AIDS in africa. AIDS. 1997: 11 (suppl B); S43-S54.

Griffiths JT, Phylip LH, Konvalinka PJ, et al. Different requirements for productive interaction between the active site of HIV-1 proteinase and substrates containing – hydrophobic*hydrophobic- or –aromatic*pro- cleavage sites. Biochemistry. 1992: 31; 5193-5200.

Gupta S, Arora K, Gupta A, Chaudhary VK. Gag-derived proteins of HIV-1 isolates from Indian patients: cloning, expression, and purification of p17 of B- and C-subtypes. Protein Expr. Purif. 2001: 21; 378-385.

Gürtler LG, Hauser PH, Eberle J, et al. A new subtype of Human Immunodeficiency virus type 1 (MVP-5180) from Cameroon. *J. Virol.* 1994; 68; 1581-1585.

Haffar OK, Popov S, Dubrovsky L, et al. Two nuclear localisation signals in the HIV-1 matrix protein regulate nuclear import of the HIV-1 pre-integration complex. *L. Mol. Biol.* 2000; 299; 359-368.

Hahn BH, Shaw GM, Taylor ME, et al. Genetic variation in HTLV-III/LAV over time in patients with AIDS or at risk for AIDS. *Science.* 1986; 232:1548-1553.

Hahn BH, Shaw GM, De Cock KM, Sharp PM. AIDS as a Zoonosis: Scientific and public health implications. *Science.* 2000; 287; 607-614.

Hanahan D. Techniques for transformation of *E. coli*. In: Glover DM, ed. DNA cloning. Volume 1. A practical approach. IRL Press Limited. England. 1985; 109-135.

Harasawa R, Mizusawa H, Nozawa K, Nakagawa T, Asada K, Kato I. Detection and tentative identification of dominant mycoplasma species in cell cultures by restriction analysis of the 16S-23S rRNA intergenic spacer regions. *Res. Microbiol.* 1993; 144; 489-493.

Hentze M. Determinants and regulation of cytoplasmic mRNA stability in eukaryotic cells. *Biochim. Biophys. Acta.* 1991; 1090; 281-292.

Hirsch V, Olmsted R, Murpey-Corb M, Pырcell, Johnson P. An African primate lentivirus (SIVsm) closely related to HIV-2. *Nature.* 1989; 339; 389-392.

Hu WS, Temin HM. Retroviral recombination and reverse transcription. *Science.* 1990; 250; 1227-1233.

Hughes WT. Pneumocystis carinii pneumonia. N. Engl. J. Med. 1977; 297; 1381-1383.

Hunt G, Tiemessen CT. Occurrence of additional NF-kappaB-binding motifs in the long terminal repeat region of South African HIV-1 genetic subtype C isolates. AIDS Res. Hum. Retroviruses. 2000; 16; 305-306.

Jenkins Y, McEntee M, Weis K, Greene WC. Characterization of HIV-1 Vpr nuclear import: analysis of signals and pathways. J. Cell Biol. 1998; 143; 875-885.

Jensen FC. Infection of human and simian tissue cultures with Rous Sarcoma virus. Proc. Natl. Acad. Sci. USA. 1964; 52; 53-59.

Johansson B, Sherefa K, Sonnerborg A. Multiple enhancer motifs in HIV type 1 strains from Ethiopia. AIDS Res. Hum. Retroviruses. 1995; 11; 761-764.

Kaiser K, Murray NE. The use of phage lambda replacement vectors in the construction of representative genomic DNA libraries. In: Glover DM, ed. DNA cloning. Volume 1. A practical approach. IRL Press Limited. England. 1985; 1-47.

Karacostas V, Nagashima K, Gonda MA, Moss B. Human Immunodeficiency virus-like particles produced by vaccinia virus expression vector. Proc. Natl. Acad. Sci. USA. 1989; 86; 8964-8967.

Kaul R, Rowland-Jones SL, Kimani J, et al. Late seroconversion in HIV-resistant Nairobi prostitutes despite pre-existing HIV-specific CD8+ responses. J. Clin. Invest. 2001; 107; 341-349.

Kemp DJ, Smith DB, Foote SJ, Samaras N, Peterson MG. Colorimetric detection of specific DNA segments amplified by polymerase chain reactions. *Proc. Natl. Acad. Sci. USA.* 1989; 86; 2423-2427.

Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 1980; 16; 111-120.

Klein MR, Van Baalen CA, Holwerda AM, et al. Kinetics of Gag-specific cytotoxic T lymphocyte responses during the clinical course of HIV-1 infection: A longitudinal analysis of rapid progressors and long-term asymptomatics. *J. Exp. Med.* 1995; 181; 1365-1372.

Kondo E, Göttlinger HG. A conserved LXXLF sequence is the major determinant in p6gag required for the incorporation of Human Immunodeficiency virus type 1 Vpr. *J. Virol.* 1996; 70; 159-164.

Korber B, Brander C, Haynes B, et al, eds. *HIV Molecular Immunology 2000*. Los Alamos National Laboratory: Theoretical Biology and Biophysics, Los Alamos, New Mexico. 2000.

Kotsopoulou E, Narry KV, Kingsman AJ, Kingsman SM, Mitrophanous KA. A rev-independent human immunodeficiency virus type 1 (HIV-1)-based vector that exploits a codon-optimized HIV-1 gag-pol gene. *J. Virol.* 2000; 74; 4839-4852.

Koup RA, Safrit JT, Cao Y, et al. Temporal association of cellular immune responses with the initial control of viremia in primary Human Immunodeficiency virus type 1 syndrome. *J. Virol.* 1994; 68; 4650-4655.

Kräusslich H-G, Fäcke M, Heuser A-M, Konvalinka J, Zentgraf H. The spacer peptide between Human Immunodeficiency virus capsid and nucleocapsid proteins is essential for ordered assembly and viral infectivity. *J. Virol.* 1995; 69; 3407-3419.

Kuiken C, Foley B, Hahn B, et al editors. HIV Sequence Compendium 2000. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory. 2000.

Laemmli, UK. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature.* 1970: 227; 680-685.

Langlade-Demoyen P, Ngo-Giang-Huong N, Ferchal F, Oksenhendler E. Human immunodeficiency virus (HIV) nef-specific cytotoxic T lymphocytes in noninfected heterosexual contact of HIV-infected patients. *J. Clin. Invest.* 1994; 93; 1293-1297.

Li W-H, Graur D, eds. *Fundamentals of Molecular Evolution.* Sinauer Associates, Inc., USA. 1991.

Lole KS, Bollinger RC, Paranjape RS, et al. Full-length Human Immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. *J. Virol.* 1999; 73; 152-160.

Louwagie J, McCutchan FE, Peeters M, et al. Phylogenetic analysis of *gag* genes from 70 international HIV-1 isolates provides evidence for multiple genotypes. *AIDS.* 1993; 7; 769-780.

Louwagie J, Janssens W, Mascola J, et al. Genetic diversity of the envelope glycoprotein from human immunodeficiency virus type 1 isolates of African origin. *J. Virol.* 1995; 69; 263-71.

Loyter A, Scangos GA, Ruddle FH. Mechanisms of DNA uptake by mammalian cells: fate of exogenously added DNA monitored by the use of fluorescent dyes. *Proc. Natl. Acad. Sci. USA.* 1982; 79; 422-426.

Luban J. Absconding with the Chaperone: Essential Cyclophilin-Gag interaction in HIV-1 virions. *Cell.* 1996; 87; 1157-1159.

MacGregor RR, Boyer JD, Ugen KE, et al. First human trial of a DNA-based vaccine for treatment of human immunodeficiency virus type 1 infection: safety and host response. *J. Infect. Dis.* 1998; 178; 92-100.

Maddon PJ, Dalgleish AG, McDougal JS, Clapham PR, Weiss RA, Axel R. The T4 gene encodes the AIDS virus receptor and is expressed in the immune system and brain. *Cell.* 1986; 47; 333-348.

Maldarelli F, Martin MA, Strebel K. Identification of posttranscriptionally active inhibitory sequences in human immunodeficiency virus type 1 RNA: Novel level of gene regulation. *J. Virol.* 1991; 65; 5732-5743.

Mammano F, Öhagen A, Höglund S, Göttlinger HG. Role of the Major Homology Region of Human Immunodeficiency virus type 1 in virion morphogenesis. *J. Virol.* 1994; 68; 4927-4936.

Massiah MA, Worthylake D, Christensen AM, et al. Comparison of the NMR and X-ray structures of the HIV-1 matrix protein: Evidence for conformational changes during viral assembly. *Protein Science.* 1996; 5; 2391-2398.

Masur H, Michelis MA, Greene JB, et al. An outbreak of community-acquired *pneumocystis carinii* pneumonia. Initial manifestation of cellular immune dysfunction. *N. Engl. J. Med.* 1981; 305; 1431-1438.

McAdam S, Kaleebu P, Krausa P, et al. Cross-clade recognition of p55 by cytotoxic T lymphocytes in HIV-1 infection. *AIDS*. 1998; 12; 571-579.

Mochizuki N, Otsuka N, Matsuo K, et al. An infectious DNA clone of HIV type 1 subtype C. *AIDS Res. Hum. Retroviruses*. 1999; 15; 1321-1324.

Momany C, Kovari LC, Prongay AJ, et al. Crystal structure of dimeric HIV-1 capsid protein. *Nat. Struct. Biol.* 1996; 3; 763-770.

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

Morellet N, de Rocquigny H, Mély Y, et al. Conformational behaviour of the active and inactive forms of the nucleocapsid NCp7 of HIV-1 studied by ¹H NMR. *J. Mol. Biol.* 1994; 235; 287-301.

Morikawa Y, Hockley DJ, Nermut MV, Jones IM. Roles of matrix, p2, and N-terminal myristoylation in human immunodeficiency virus type 1 Gag assembly. *J. Virol.* 2000; 74; 16-23.

Morris L, Van der Ryst E, Gray C, Williamson C. Should South Africa be preparing for HIV-1 vaccine efficacy trials? *S. Afr. Med. J.* 1997; 87; 285-290.

Morris L, Cilliers T, Bredell H, Phoswa M, Martin D. CCR5 is the major coreceptor used by HIV-1 subtype C isolates from patients with active tuberculosis. *AIDS Res. Hum. Retroviruses*. 2001; 17; 697-701.

Morris L, Williamson C. Host and viral factors that impact on HIV-1 transmission and disease progression in South Africa. *S. Afr. Med. J.* 2001; 91; 212-215.

Mullis KB, Faloona FA. Specific synthesis of DNA *in vitro* via a polymerase-catalyzed chain reaction. *Methods Enzymol* 1987: 155; 335-50.

Musey L, Hughes J, Schacker T, Shea T, Corey L, McElrath MJ. Cytotoxic-T-cell responses, viral load, and disease progression in early human immunodeficiency virus type 1 infection. *N. Engl. J. Med.* 1997: 337; 1267-1274.

Ndung'u T, Renjifo B, Novitsky VA, Mclane MF, Gaolekwe S, Essex M. Molecular cloning and biological characterisation of full-length HIV-1 subtype C from Botswana. *Virology.* 2000: 278; 390-399.

Ndung'u T, Renjifo B, Essex M. Construction and analysis of an infectious Human Immunodeficiency virus type 1 subtype C molecular clone. *J. Virol.* 2001: 75; 4964-4972.

Nicholas K, Nicholas H. GeneDoc, version 2.5.00, a program for the shading and editing of multiple sequence alignments. 1997: <http://www.psc.edu/biomed/genedoc>

Novitsky VA, Montano MA, McLane MF, et al. Molecular cloning and phylogenetic analysis of Human Immunodeficiency virus type 1 subtype C: a set of 23 full-length clones from Botswana. *J. Virol.* 1999: 73; 4427-4432.

Okamoto T, Benter T, Josephs SF, Sadaie MR, Wong-Staal F. Transcriptional activation from the long-terminal repeat of Human Immunodeficiency virus *in vitro*. *Virology.* 1990: 177; 606-614.

Olsen HS, Cochrane AW, Rosen C. Interaction of cellular factors with intragenic *Cis*-acting repressive sequences within the HIV genome. *Virology.* 1992: 191; 709-715.

Ono A, Huang M, Freed EO. Characterisation of Human Immunodeficiency Virus Type 1 matrix revertants: Effects on virus assembly, Gag processing, and Env incorporation into virions. *J. Virol.* 1997; 71; 4409-4418.

Ono A, Orenstein JM, Freed EO. Role of the Gag matrix domain in targeting Human Immunodeficiency virus type 1 assembly. *J. Virol.* 2000; 74; 2855-2866.

Ott DE, Chertova EN, Busch LK, Coren LV, Gagliardi TD, Johnson DG. Mutational analysis of the hydrophobic tail of the Human Immunodeficiency virus type 1 p6^{Gag} protein produces a mutant that fails to package its Envelope protein. *J. Virol.* 1999; 73; 19-28.

Ou C-Y, Kwok S, Mitchell SW, et al. DNA amplification for direct detection of HIV-1 in DNA of peripheral blood mononuclear cells. *Science.* 1988; 239; 295-297.

Overbaugh J, Dewhurst S, Mullins JI. Proviral DNA Cloning. In: Aldovini A, Walker BD, eds. *Techniques in HIV Research.* Stock Press. New York. 1990; 131-139.

Pääbo S, Higuchi RG, Wilson AC. Ancient DNA and the polymerase chain reaction. The emerging field of molecular archaeology. *J. Biol. Chem.* 1989; 264; 9709-9712.

Page RDM, Holmes EC. *Molecular Evolution. A phylogenetic approach.* Blackwell Science Ltd. Oxford. 1998.

Peeters M, Vincent R, Perret J-L, et al. Evidence for differences in MT2 cell tropism according to genetic subtypes of HIV-1: syncytium-inducing variants seem rare among subtype C HIV-1 viruses. *J. Acquir. Imm. Def. Synd.* 1999; 20; 115-121.

Peeters M, Sharp PM. Genetic diversity of HIV-1: the moving target. *AIDS*. 2000; 14 (suppl 3); S129-S140.

Pettit SC, Simsic CJ, Loeb DD, et al. Analysis of retroviral protease cleavage sites reveals two types of cleavage sites and the structural requirements of the P1 amino acid. *J. Biol. Chem.* 1991; 266; 14539-14547.

Pettit SC, Moody MD, Wehbie RS, et al. The p2 domain of Human Immunodeficiency virus type 1 Gag regulates sequential proteolytic processing and is required to produce fully infectious virions. *J. Virol.* 1994; 68; 8017-8027.

Qiu J-T, Song R, Dettenhofer M, et al. Evaluation of novel human immunodeficiency virus type 1 Gag DNA vaccines for protein expression mammalian cells and induction of immune responses. *J. Virol.* 1999; 73; 9145-9152.

Ratner L, Gallo RC, Wong-Staal F. HTLV-III, LAV, ARV are variants of same AIDS virus. *Nature*. 1985; 313; 636-637.

Ray SC. SimPlot for Windows (version 1.6). Baltimore, Md. Distributed by author. 1999. (<http://www.welch.jhu.edu/~sray/download>).

Reil H, Bukovsky AA, Gelderblom HR, Göttlinger H. Efficient HIV-1 replication can occur in the absence of the viral matrix protein. *EMBO*. 1998; 17; 2699-2708.

Robertson DL, Sharp PM, McMurphy FE, Hahn BH. Recombination in HIV-1. *Nature*. 1995; 374; 124-126.

Rodenburg CM, Li Y, Trask SA, et al. Near full-length clones and reference sequences for subtype C isolates of HIV type 1 from three different continents. *Aids. Res. Hum. Retroviruses*. 2001; 17; 161-168.

Rovinski B, Klein M. Genetically engineered Human Immunodeficiency virus type 1 (HIV-1) vaccines. In: Kurstak, ed. *Modern Vaccinology*. Plenum Medical Book Company, N.Y., USA. 1994; 181-212.

Rowland-Jones SL, Nixon DF, Aldhous MC, et al. HIV-specific cytotoxic T-cell activity in an HIV-exposed but uninfected infant. *Lancet*. 1993; 341; 860-861.

Rowland-Jones SL, Dong T, Dorrell L, et al. Broadly cross-reactive HIV-1 specific cytotoxic T-lymphocytes in highly-exposed persistently seronegative donors. *Immunol Lett*. 1999; 66; 9-14.

Saitou N, Nei M. The neighbour-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol*. 1987; 4; 406-425.

Sakai K, Dewhurst S, Ma X, Volsky DJ. Differences in cytopathogenicity and host cell range among infectious molecular clones in human immunodeficiency virus type 1 simultaneously isolated from an individual. *J. Virol*. 1988; 62:4078-4085.

Salminen MO, Carr JK, Burke DS, McCutchan FE. Identification of breakpoints in intergenotypic recombinants of HIV-1 type 1 by Bootscanning. *AIDS Res. Hum. Retroviruses*. 1995; 11; 1423-1425.

Salminen MO, Johansson B, Sönnernborg A, et al. Full-length sequence of an Ethiopian Human Immunodeficiency virus type 1 (HIV-1) isolate of genetic subtype C. *AIDS Res. Hum. Retroviruses*. 1996; 12; 1329-1339.

Sambrook J, Fritsch EF, Maniatis T. Molecular Cloning: A Laboratory Manual second edition. New York: Cold Spring Harbor Laboratory Press, 1989.

Sanders RW, Schiffner L, Master A, et al. Variable-loop-deleted variants of the Human Immunodeficiency virus type 1 envelope glycoprotein can be stabilized by an intermolecular disulfide bond between the gp120 and gp41 subunits. *J. Virol.* 2000; 74; 5091-5100.

Sanders-Buell, Salminen MO, McCutchan FE. Sequencing primers for HIV-1. In: Myers G, Korber B, Hahn BH, Jeang K-T, Mellors JW, McCutchan FE, Henderson LE and Pavlakis GN, editors. Human Retroviruses and AIDS 1995: A Compilation and Analysis of Nucleic Acid and Amino Acid Sequences. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM. 1995.

Scarlata S, Ehrlich LS, Carter CA. Membrane-induced alterations in HIV-1 Gag and matrix protein-protein interactions. *J. Mol. Biol.* 1998; 277; 161-169.

Schneider R, Cambell M, Nasioulas G, Felber BK, Pavlakis GN. Inactivation of the human immunodeficiency virus type 1 inhibitory elements allows Rev-independent expression of Gag and Gag/protease and particle formation. *J. Virol.* 1997; 71; 4892-4903.

Scriba TJ, Treurnicht FK, Zeier M, Engelbrecht S, Janse van Rensburg E. Characterization and phylogenetic analysis of South African HIV-1 subtype C accessory genes. *AIDS Res. Hum. Retroviruses.* 2001; 17; 775-781.

Scriba TJ, De Villiers T, Treurnicht FK, et al. Characterisation of the South African HIV-1 subtype C complete 5' LTR, *nef* and regulatory genes. *AIDS Res. Hum. Retroviruses.* 2001: In press.

Shafer RW, Edlin BR. Tuberculosis in patients infected with human immunodeficiency virus: perspective on the past decade. Clin. Infect. Dis. 1996; 22; 683-704.

Shigekawa K, Dower WJ. Electroporation of eukaryotes and prokaryotes: a general approach to the introduction of macromolecules into cells. Biotechniques. 1988; 6; 742-751.

Siegal FP, Lopez C, Hammer GS, et al. Severe acquired immunodeficiency in male homosexuals, manifested by chronic perianal ulcerative herpes simplex lesions. N. Engl. J. Med. 1981; 305; 1439-1444.

Simon F, Mauclore P, Roques P, et al. Identification of a new human immunodeficiency virus type 1 distinct from group M and group O. Nat Med 1998; 9; 1032-1037.

Stamatatos L, Cheng-Mayer. An envelope modification that renders a primary, neutralization resistant clade B Human Immunodeficiency virus type 1 isolate highly susceptible to neutralization by sera from other clades. J. Virol. 1998; 72; 7840-7845.

Swerdlow H, Gesteland R. Capillary gel electrophoresis for rapid, high resolution DNA sequencing. Nucleic Acids Res. 1990;18; 1415-1419.

Tang J, Hu M, Lee S, Roblin R. A polymerase chain reaction based method for detecting *Mycoplasma/ Acholeplasma* contaminants in cell culture. J. Microbiol. Methods. 2000; 39; 121-126.

Tateno Y, Nei M, Tajima F. Accuracy of estimated phylogenetic trees from molecular data. I. Distantly related species. J. Mol. Evol. 1982; 18; 387-404.

Thali M, Bukovsky A, Kondo E, et al. Functional association of cyclophilin A with HIV-1 virions. *Nature*. 1994; 372; 363-365.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research*. 1997; 25; 4876-4882.

Towbin H, Staehelin T, Gordon J. Electrophoretic transfer of proteins from polyacrylamids gels to nitrocellulose sheets: procedure and some applications. *Proc. Natl. Acad. Sci. USA*. 1979; 76; 4350-4354.

Treurnicht FK, Smith T-L, Engelbrecht S, et al. Genotypic and phenotypic analysis of the *env* gene of South African HIV-1 subtype B and C isolates. *J. Med. Virol.* 2001: in press.

Tritel M, Resh MD. Kinetic analysis of Human Immunodeficiency virus type 1 assembly reveals the presence of sequential intermediates. *J. Virol.* 2000; 74; 5845-5855.

Tscherning C, Alaeus A, Fredriksson R, et al. Differences in chemokine coreceptor usage between genetic subtypes of HIV-1. *Virology*. 1998; 241; 181-188.

UNAIDS/WHO working group on global HIV/AIDS and STD surveillance: Report on the global HIV/AIDS epidemic – December 2000. UNAIDS, Geneva, 2000. (http://www.unaids.org/epidemic_update/report/index.html)

Van de Peer Y, Wachter R. TREECON for Windows: a software package for the construction and drawing of evolutionary trees for the Microsoft Windows environment. *Comput. Appl. Biosci.* 1994; 10; 569-570.

Van der Loeff M, Aaby P. Towards a better understanding of the epidemiology of HIV-2. *AIDS*. 1999; 13 (suppl A); S69-S84.

Van Harmelen J, Wood R, Lambrick M, Rybicki EP, Williamson A-L, Williamson C. An association between HIV-1 subtypes and mode of transmission in Cape Town, South Africa. *AIDS*. 1997; 11; 81-87.

Van Harmelen JH, Van der Ryst E, Loubser AS, et al. A predominantly HIV type 1 subtype C-restricted epidemic in South African urban populations. *AIDS Res. Hum. Retroviruses*. 1999a; 15; 395-398.

Van Harmelen J, Van der Ryst E, Wood R, Lyons SF, Williamson C. Restriction fragment length polymorphism analysis for rapid gag subtype determination of human immunodeficiency virus type 1 in South Africa. *J. Virol. Meth.* 1999b; 78; 51-59.

Vermund SH, Schultz AM, Johnston M. Evaluation of new vaccines: challenges for HIV/AIDS. In: Cutts FT, Smith PG, eds. *Vaccination and world health*. John Wiley & Sons. Inc. England. 1994: 89-122.

Von Pöbltzki A, Wagner R, Niedrig M et al. Identification of a region in the Pr55^{gag}-polyprotein essential for HIV-1 particle formation. *Virology*. 1993; 193; 981-985.

Von Schwedler U, Kornbluth RS, Trono D. The nuclear localization signal of the matrix protein of human immunodeficiency virus type 1 allows the establishment of infection in macrophages and quiescent lymphocytes. *Proc. Natl. Acad. Sci. USA*. 1994; 91; 6992-6996.

Von Schwedler UK, Stemmler TL, Klishko VY, et al. Proteolytic refolding of the HIV-1 capsid protein amino-terminus facilitates viral core assembly. *EMBO J.* 1998; 17; 1555-1568.

Weaver, RF, Philip WH. Gene Cloning and Manipulation. In: Kane K, ed. *Genetics*, Second Edition. Dubuque, IA: Wm. C. Brown Publishers, 1992: 437-485.

Wieggers K, Rutter G, Kottler H, Tessmer U, Hohenberg H, Kräusslich H-G. Sequential steps in Human Immunodeficiency virus particle maturation revealed by alterations of individual Gag polyprotein cleavage sites. *J. Virol.* 1998; 72; 2846-2854.

Williamson C, Engelbrecht S, Lambrick M, et al. HIV-1 subtypes in different risk groups in South Africa. *Lancet.* 1995; 346: 782.

Wills JW, Craven RC. Form, function, and use of retroviral Gag proteins. *AIDS.* 1991; 5; 639-654.

Wong TK, Neumann E. Electric field mediated gene transfer. *Biochem. Biophys. Res. Commun.* 1982; 107; 584-587.

Yuan X, Yu X, Lee T-H, Essex M. Mutations in the N-terminal region of Human Immunodeficiency virus type 1 matrix protein block intracellular transport of the Gag precursor. *J. Virol.* 1993; 67; 6387-6394.

Zhou MY, Clark SE, Gomez-Sanchez CE. Universal cloning method by TA strategy. *Biotechniques.* 1995; 19; 34-35.

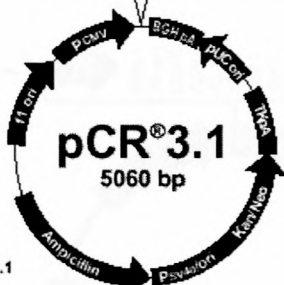
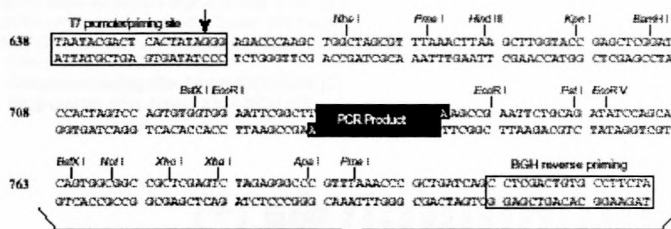
Zhou W, Parent LJ, Wills JW, Resh MD. Identification of a membrane-binding domain within the amino-terminal region of human immunodeficiency virus type 1

Gag protein which interacts with acidic phospholipids. J. Virol. 1994; 68; 2556-2569.

Zur Megede J, Chen M-C, Doe B, et al. Increased expression and immunogenicity of sequence-modified human immunodeficiency virus type 1 *gag* gene. J. Virol. 2000; 74; 2628-2635.

Zybarth G, Carter C. Domains upstream of the protease in Human Immunodeficiency virus type 1 Gag-Pol influence PR autoprocessing. J. Virol. 1995; 69; 3878-3884.

A1. Map and summary of pCR[®]3.1



Comments for pCR[®]3.1
5060 nucleotides

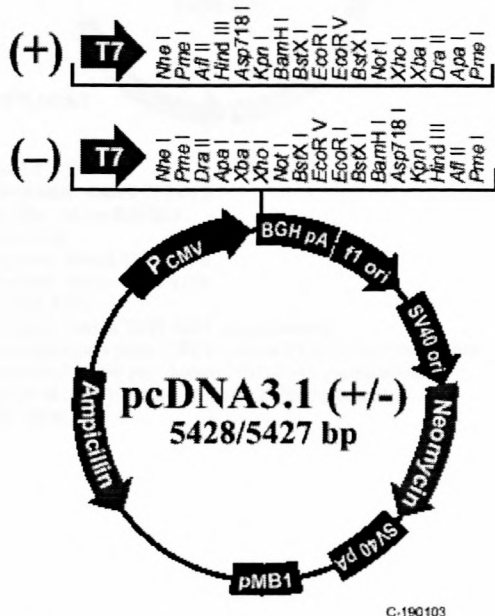
- CMV promoter: bases 1-596
- Putative transcriptional start: bases 620-625
- T7 promoter/priming site: bases 638-657
- Multiple cloning site: bases 670-801
- TA Cloning[®] site: 737-738
- BGH reverse priming site: bases 813-831
- BGH polyadenylation site: bases 812-1026
- pUC origin: bases 1116-1789
- SV40 promoter and origin: bases 3194-3532 (complement)
- Neomycin/kanamycin resistance gene (ORF): bases 2371-3159 (complement)
- Thymidine kinase polyadenylation site: bases 1926-2196 (complement)
- Ampicillin resistance gene (ORF): bases 3611-4471 (complement)
- f1 origin: bases 4602-5058

(Eukaryotic TA Cloning Kit Catalog, Invitrogen)

A2. Map and summary of pcDNA3.1

Comments for pcDNA3.1 (+)
5428 nucleotides

CMV promoter: bases 232-819
 T7 promoter/priming site: bases 863-882
 Multiple cloning site: bases 895-1010
 pcDNA3.1/BGH reverse priming site: bases 1022-1039
 BGH polyadenylation sequence: bases 1028-1252
 f1 origin: bases 1298-1726
 SV40 promoter and origin: bases 1731-2074
 Neomycin resistance gene (ORF): bases 2136-2930
 SV40 polyadenylation signal: bases 3104-3234
 pMB1 origin (pUC-derived): bases 3617-4287
 Ampicillin resistance gene (*bla*): bases 4432-5428 (C)
 ORF: bases 4432-5292 (C)
 Ribosome binding site: bases 5300-5304 (C)
bla promoter (P3): bases 5327-5333 (C)

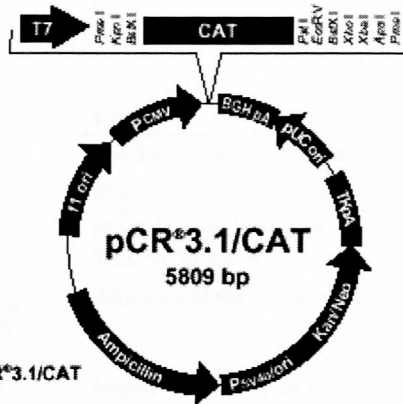


C-190103

pcDNA3.1(+) and pcDNA3.1(-) have been completely sequenced. If you suspect an error in the sequences, please contact Invitrogen's Technical Services Department.

(pcDNA3.1 catalog, Invitrogen)

A3. Map and summary of pCR[®]3.1/CAT



Comments for pCR[®]3.1/CAT
5809 nucleotides

- CMV promoter: bases 1-596
- Putative transcriptional start: bases 620-625
- T7 promoter/priming site: bases 638-657
- CAT ORF: bases 738-1486
- BGH reverse priming site: bases 1562-1579
- BGH polyadenylation site: bases 1561-1775
- pUC origin: bases 1906-2489
- SV40 promoter and origin: bases 3943-4281 (complement)
- Neomycin/kanamycin resistance gene (ORF): bases 3120-3908 (complement)
- Thymidine kinase polyadenylation site: bases 2675-2945 (complement)
- Ampicillin resistance gene (ORF): bases 4360-5220 (complement)
- f1 origin: bases 5351-5807

(Eukaryotic TA Cloning Kit Catalog, Invitrogen)

A4. List of genotypes and corresponding phenotypes of bacterial strains



Table 13. Genetic Markers in Frequently Used *E. coli* Strains.

Symbol	Description	Effect
<i>ara-14</i>	Mutation in arabinose metabolism	Blocks arabinose catabolism.
<i>dam</i>	Adenine methylase mutation	Blocks methylation of adenine residues in the sequence 5'...G ^m ATC...3'.
<i>dcm</i>	Cytosine methylase mutation	Blocks methylation of cytosine in the sequence 5'...C ^m CAGG...3' or 5'...C ^m CTGG...3'.
<i>deoR</i>	Regulatory gene mutation allowing constitutive expression of genes for deoxyribose synthesis	Allows uptake of large plasmids.
<i>endA1</i>	Endonuclease mutation	Improves quality of plasmid DNA isolations.
<i>galK</i>	Galactokinase mutation	Blocks catabolism of galactose.
<i>gyrA96</i>	DNA gyrase mutation	Confers resistance to nalidixic acid.
<i>hflA150</i>	Mutation leading to a high frequency of lysogeny	Leads to greatly enhanced frequency of lysogeny in λ phages containing a normal repressor (<i>cI</i>) gene (1).
<i>hsdR</i> (r_K^- , m_K^+)	Restriction minus, modification positive	Allows cloning without cleavage of transformed DNA by endogenous restriction endonucleases. DNA prepared from this strain can be used to transform r_K^+ <i>E. coli</i> strains.
<i>hsdS20</i> (r_B^- , m_B^-)	Restriction minus, modification minus	Allows cloning without cleavage of transformed DNA by endogenous restriction endonucleases. DNA prepared from this strain is unmethylated by the <i>hsdS20</i> methylases.
<i>lacI⁺</i>	Overproduction of the <i>lac</i> repressor protein	Leads to high levels of the <i>lac</i> repressor protein, inhibiting transcription from the <i>lac</i> promoter.
<i>lacY</i>	Galactoside permease mutation	Blocks lactose utilization.
<i>lacY1</i>	β -D-galactosidase mutation	Blocks lactose utilization.
<i>lacZ</i> Δ M15	Partial deletion of β -D-galactosidase gene	Allows complementation of β -galactosidase activity by α -complementation sequence in pGEM ⁺ -Z Vectors. Allows blue/white selection for recombinant colonies when plated on X-Gal.

A4. List of genotypes and corresponding phenotypes of bacterial strains

Symbol	Description	Effect
<i>leu B</i>	β -isopropyl malate dehydrogenase mutation	Requires leucine for growth on minimal media.
$\Delta(lon)$	Deletion of <i>lon</i> protease	Reduces proteolysis of expressed fusion proteins.
<i>mcr A</i>	Mutation in restriction system	Blocks restriction of DNA methylated at the sequence 5'..G ^m CGC..3'.
<i>mcr B</i>	Mutation in restriction system	Blocks restriction of DNA methylated at the sequence 5'..AG ^m CT..3'.
<i>met B</i>	Cystathionine γ -synthase mutation	Requires methionine for growth on minimal media.
<i>mntl-1</i>	Mutation in mannitol metabolism	Blocks catabolism of mannitol.
<i>mut S</i>	Mismatch repair minus strain	Prevents repair of the newly synthesized, unmethylated strand.
P2	P2 bacteriophage lysogen present in host	λ phages containing the <i>red</i> and <i>gam</i> genes of λ are growth inhibited by P2 lysogens (2).
<i>pro AB</i>	Mutations in proline metabolism	Requires proline for growth in minimal media.
<i>rec A1, rec A13</i>	Mutation in recombination	Prevents recombination of introduced DNA with host DNA, ensuring stability of inserts. Inserts are more stable in <i>rec A1</i> than <i>rec A13</i> hosts.
<i>rel A</i>	Relaxed phenotype; mutation eliminating stringent factor	Allows RNA synthesis in the absence of protein synthesis.
<i>rec B, rec C</i>	Exonuclease V mutations	Reduces general recombination and affects repair of radiation damage.
<i>ops L</i>	Mutation in subunit S12 of 30S ribosome	Confers resistance to streptomycin.
<i>rec B</i>	Exonuclease I mutation	Allows general recombination in <i>rec BC</i> mutant strains.
<i>sup B, sup C, sup G, sup L, sup M, sup N, sup O</i>	Suppressor mutations	Suppress ochre (UAA) and amber (UAG) mutations.
<i>sup D, sup E, sup F</i>	Suppressor mutations	Suppress amber (UAG) mutations.
<i>thi-1</i>	Mutation in thiamine metabolism	Thiamine required for growth in minimal media.
<i>tn5</i>	Transposon	Encodes resistance to kanamycin.
<i>tn10</i>	Transposon	Encodes resistance to tetracycline.
<i>trt A</i>	Mutation in outer membrane protein	Confers resistance to bacteriophage T1.
<i>trf D36</i>	Transfer factor mutation	Prevents transfer of F' episome.
<i>xyl-5</i>	Mutation in xylose metabolism	Blocks catabolism of xylose.

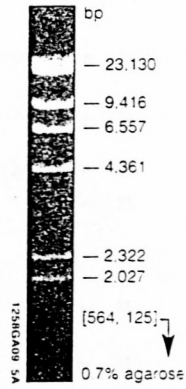
References

- Hoyt, A. (1982) *Cell* **31**, 565.
 Kaiser, K. and Murray, N. (1985) In: *DNA Cloning, Vol. 1*, Glover, D., ed., IRL Press Ltd., Oxford, UK.

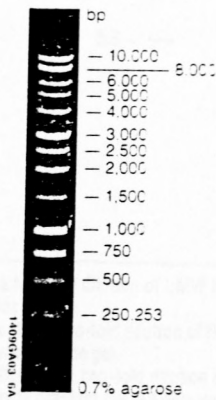
A5. DNA Molecular Weight Markers



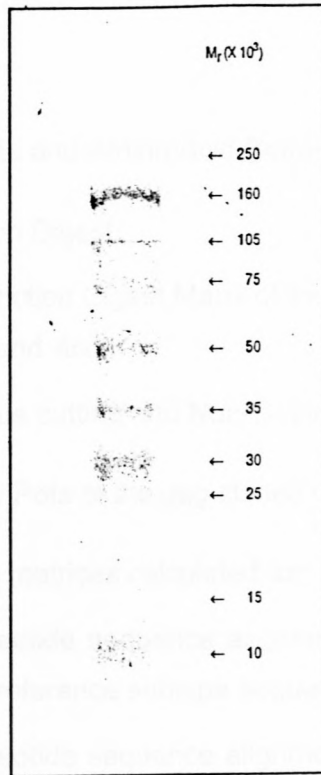
Lambda DNA/*Hind* III Markers



1kb DNA Ladder



Rainbow Molecular Weight Markers



The ladder of proteins in the Full-Range Rainbow markers on a 12% SDS-PAGE gel.

Protein and Peptide Molecular Weight Markers

LMW	HMW-SDS	HMW-Native
97.0 - --	220 - ---	
66.0 - --	170 - ---	
45.0 - --	116 - ---	669 - ---
30.0 - --	76 - ---	440 - ---
20.1 - --	53 - ---	232 - ---
14.4 - --		140 - ---
		66 - ---

LMW: A 3 μ l aliquot of a two-fold dilution of LMW Marker Kit was separated on a 15% T, 2.7% C polyacrylamide gel.
HMW-SDS: A 10 μ l aliquot of a two-fold dilution of HMW-SDS Marker Kit was separated on 7.5% T, 2.7% C polyacrylamide gel.
HMW-Native: A 10 μ l aliquot of a two-fold dilution of HMW Marker Kit was separated on 5-12.5% polyacrylamide gradient gel. All gels were stained with PhastGel Blue R. Numbers represent molecular weights (M_r) in thousands.

Appendix B

- B1. Nucleotide and Amino Acid Sequences of the HIV-1 *gag* clones
- B2. Restriction Digest
 - B2.1 Restriction Digest Maps of the *gag* clones for the Restriction enzymes, *A**l**u**I* and *A**c**c**I*
 - B2.2 Unique cutting and Non-Cutting restriction enzymes
- B3. Similarity Pots of the *gag* clones obtained from the NCBI subtyping tool
- B4. Distance matrices calculated for:
 - B4.1 Nucleotide sequence alignment of the whole *gag* for the *gag* clones and reference subtype sequences
 - B4.2 Nucleotide sequence alignment of the p17 matrix domain for the *gag* clones and reference subtype sequences
 - B4.3 Nucleotide sequence alignment of the p24 capsid domain for the *gag* clones and reference subtype sequences
 - B4.4 Nucleotide sequence alignment of the region from the nucleocapsid to p1 spacer peptide for the *gag* clones and reference subtype sequences
 - B4.5 Amino acid sequence alignment of the three areas in p17 containing published CTL epitopes
- B5. Predicted amino acid sequence alignment of the complete *gag*

B1. Nucleotide and Amino Acid Sequences of gag clones

Translation of TV001G8 (1-1494)

Universal code

Total amino acid number: 497, MW=55374

Max ORF: 1-1491, 497 AA, MW=55374

```
1      ATGGGTGCGAGAGCGTCAATATTAAGCGGCGGAAAATTAGATAAATGGGAAAGAATTAGG
1      M G A R R A S I L S G G K L D K W E R I R
61     TTAAGGCCAGGGGAAAGAAACATTATATGTTAAAACATCTAGTATGGGCAAGCAGGGAG
21     L R P G G G K K H Y M L K H L V W A S R E
121    CTGGAAGATTGCACTTAAACCTGGCCTGTTAGAAACATCAGAAGGCTGTAACAAATA
41     L E R F A L N P G L L E T S E G C K Q I
181    ATAAAAAGCTACAACCGCTCTTCAGACAGGAACAGAGGAACTTAGATCATTATTCAAC
61     I K Q L Q P A L Q T G T E E L R S L F N
241    ACAGTAGCAACTCTCTATTGTGTACATAAAGGGATAAAGGTACGAGACACCAAGGAAGCC
81     T V A T L Y C V H K G I K V R D T K E A
301    TTAGACAAGATAGAGGAAGAACAACAATGTGCAAGCAAGCACAGCAGGCAAAAGCG
101    L D K I E E E Q N K C Q Q K A Q Q A K A
361    GCTGACGAAAAGGTCAGTCAAAATATCCTATAGTACAGAATGCCAAGGGCAAAATGGTA
121    A D E K V S Q N Y P I V Q N A Q G Q M V
421    CACCAAGCTATATCACCTAGAACATTGAATGCATGGGTAAAAGTAATAGAGGAGAAGGCT
141    H Q A I S P R T L N A W V K V I E E K A
481    TTCAACCCAGAGGTAATACCCATGTTTACAGCATTATCAGAAGGAGCCACCCACAAAGAT
161    F N P E V I P M F T A L S E G A T P Q D
541    TTAACACCATGTTAAATACAGTGGGGGACATCAAGCAGCCATGCAAAATGTTAAAAGAT
181    L N T M L N T V G G H Q A A M Q M L K D
601    ACCATCAATGAGGAGGCTGCAGAATGGGATAGGACACATCCAGTGCATGCAGGGCCTGTT
201    T I N E E A A E W D R T H P V H A G P V
661    GCACCAGGCCAGATGAGAGAACCAAGGGGAAGTACATAGCAGGAACACTAGTACCCCTT
221    A P G Q M R E P R G S D I A G T T S T L
721    CAGGAACAAATAGCATGGATGACAAGTAATCCACCTATCCAGTAGGAGACATCTATAAA
241    Q E Q I A W M T S N P P I P V G D I Y K
781    AGATGGATAATCTGGGGTTAAATAAAATAGTAAGAATGTATAGCCCTGTCAGCATTTTG
261    R W I I L G L N K I V R M Y S P V S I L
841    GACATAAAACAGGGCCAAAAGAACCCTTTAGAGATTATGTAGATCGGTTCTTTAAAAC
281    D I K Q G P K E P F R D Y V D R F F K T
901    TTAAGAGCTGAACAAGCTACACAAGATGTAAAAAATGGATGACAGACACCTTGTGGTC
301    L R A E Q A T Q D V K N W M T D T L L V
961    CAAAATGCCAACCAGATTGTAAGACCATTTAAGAGCATTAGGACCAGGGGCTTCATTA
321    Q N A N P D C K T I L R A L G P G A S L
1021   GAAGAAATGATGACAGCATGTCAGGGAGTGGGAGGACCTAGCCATAAAGCAAGGGTGTG
341   E E M M T A C Q G V G G P S H K A R V L
1081   GCTGAGGCAATGAGCCAAACAAACAGTAACATACTAGTGCAGAGAAGCAATTTTAAAGGC
361   A E A M S Q T N S N I L V Q R S N F K G
1141   CCTAACAGAATTGTTAAATGTTTCAACTGTGGCAAAGTAGGGCACATAGCCAGAAAGTGC
381   P N R I V K C F N C G K V G H I A R K C
1201   AGGGCCCTAGGAAAAGGGCTGTTGGAATGTGGACAGGAAGGCCACCAATGAAAGAC
401   R A P R K K G C W K C G Q E G H Q M K D
1261   TGTACTGAGAGGCAGGCTAATTTTTAGGGAAAATCTGGCCTTCCCACAAGGGGAGGCCA
421   C T E R Q A N F L G K I W P S H K G R P
1321   GGGAAATTCCTCCGAACAGACCAGAGCCAACAGCCCCACCAGCAGGCCAACAGCCCCA
441   G N F L Q N R P E P T A P P A E P T A P
```


B1. Nucleotide and Amino Acid Sequences of gag clones

1381 CCAGCAGAGAGCTTCAGGTTTCGAGGAGACAACCCCGTGCCGAGGAAGGAGAAAGACAGG
461 P A E S F R F E E T T P V P R K E K D R

1441 GAACCTTAACTCCCTCAAATCACTCTTGGCAGCGACCCCTCGTCACAATA
481 E P L T S L K S L F G S D P S S Q *

Translation of TV001G11 (1-1494)

Universal code
Total amino acid number: 497, MW=55404
Max ORF: 1-1491, 497 AA, MW=55404

1 ATGGGTGCGAGAGCGTCAATATTAAGCGGCGGAAAATTAGATAAATGGGAAAAGATTAGG
1 M G A R A S I L S G G K L D K W E R I R

61 TTAAGGCCAGGGGAAAGAAACATTATATGTTAAAACATTTAGTATGGGCAAGCAGAGAG
21 L R P G G K K H Y M L K H L V W A S R E

121 CTGGAAGATTGTCACTTAAACCTGGCCTGTAGAGACAGCAGAAGGCTGTAACAATA
41 L E R F A L N P G L L E T A E G C K Q I

181 ATAAAACAGCTACAACCGCTCTTCAGACAGGAACAGAGGAACCTTAGATCATTATTCAAC
61 I K Q L Q P A L Q T G T E E L R S L F N

241 ACAGTAGCAACTCTCTATTGTGTACATAAAGGAATAGAGGTACGAGACACCAAGGAAGCC
81 T V A T L Y C V H K G I E V R D T K E A

301 TTAGACAAGATAGAGGAAGAACAACAATGTCAACAAAAGGCACAACAGGCAAAAGCG
101 L D K I E E E Q N K C Q Q K A Q Q A K A

361 GCTGATGAAAAGTTCAGTCAAAATTATCCTATAGTACAGAATGCCCAAGGGCAAATGGTA
121 A D E K V S Q N Y P I V Q N A Q G Q M V

421 CACCAAGCTATATCACCTAGAACATTGAATGCATGGGTAAAAGTAATAGAGGAGAAGGCT
141 H Q A I S P R T L N A W V K V I E E K A

481 TTCAACCCAGAGGTGATACCCATGTTTACAGCATTATCAGAAGGAGCCACCCACAAGAT
161 F N P E V I P M F T A L S E G A T P Q D

541 TTAACACAATGTTAAATACAGTGGGGGACATCAAGCAGCCATGCAAAATGTTAAAAGAT
181 L N T M L N T V G G H Q A A M Q M L K D

601 ACCATCAATGAGGAGGCTGCAGAATGGGATAGGACACATCCAGTGCATGCAGGCCTGTT
201 T I N E E A A E W D R T H P V H A G P V

661 GCACCAGCCAGATGAGAGAACAAGGGGAAGTGACATAGCAGGAACACTAGTACCCTT
221 A P G Q M R E P R G S D I A G T T S T L

721 CAGGAACAATAGCATGGATGACAAGTAATCCACCTATTCCAGTAGGGGACATCTATAAA
241 Q E Q I A W M T S N P P I P V G D I Y K

781 AGATGGATAAATCTGGGGTTAAATAAAATAGTAAGAATGTATAGCCCTGTTAGCATTG
261 R W I I L G L N K I V R M Y S P V S I L

841 GACATAAAAACAAGGGCCAAAAGAACCCTTTAGAGATTATGTAGATCGGTTCTTTAAACT
281 D I K Q G P K E P F R D Y V D R F F K T

901 TTAAGAGCTGAACAAGCTACACAAGATGTAATAAATGGATGACAGACACCTTGTGGTC
301 L R A E Q A T Q D V K N W M T D T L L V

961 CAAAATGCGAACCCAGATTGTAAGACCATTTTAAGAGCATTAGGACCAGGGCTTCATTA
321 Q N A N P D C K T I L R A L G P G A S L

1021 GAAGAAATGATGACAGCATGTGAGGAGTGGGAGGACCTAGCCATAAAGCAAGGTGTTG
341 E E M M T A C Q G V G G P S H K A R V L

1081 GCTGAGGCAATGAGCCAAAACAACAGTAACATACTAGTGCAGAGAAGCAATTTTAAAGGC
361 A E A M S Q T N S N I L V Q R S N F K G

1141 TCTAACAGAATTGTTAAATGTTTCAACTGTGGCAAGGTGGGGCACATAGTCAGAAAATGC
381 S N R I V K C F N C G K V G H I V R N C

B1. Nucleotide and Amino Acid Sequences of gag clones

1201 AGGGCCCCTAGGAAAAAGGGCTGTTGAAATGTGGACAGGAAGGGCACCAAATGAAAGAC
401 R A P R K K G C W K C G Q E G H Q M K D

1261 TGTACTGAGAGACAGGCTAATTTTTAGGAAAATCTGGCCTTCCCACAAGGGGAGGCCA
421 C T E R Q A N F L G K I W P S H K G R P

1321 GGGAAATTCCTCCAGAACAGACCAGGCCAACAGCCCCACCAGCAGAACCAACAGCCCCA
441 G N F L Q N R P E P T A P P A E P T A P

1381 CCAGCAGAGAGCTTCAGGTTTCGAGGAGACAACCCCGTCCGGAAGAGGGAGAAGAGAGG
461 P A E S F R F E E T T P V P K R E K E R

1441 GAACCTTTAACTCCCTCAAATCACTCTTTGGCAACGACCCCTCGTCAATAA
481 E P L T S L K S L F G N D P S S Q *

Translation of TV002G8 (1-1464)

Universal code

Total amino acid number: 487, MW=54503

Max ORF: 1-1461, 487 AA, MW=54503

1 ATGGGTGCGAGAGCGTCAGTATTGAAAGGGAAAAAATTAGATACATGGGAAAGAATTAGG
1 M G A R A S V L K G K K L D T W E R I R

61 TTAAGGCCAGGGGAAAGAAACACTATATGCTAAAACCTAGTATGGGCAAGCAGGGAG
21 L R P G G K K H Y M L K H L V W A S R E

121 CTGAAAGATTGCACTTAAACCTGGCCTTTTAGAAAACAGCAGAAGGCTGTAACAATA
41 L E R F A L N P G L L E T A E G C K Q I

181 ATGCAACAGCTACAATCAGCTCTTCAGACAGGAACAGAGGAAGTATAGATCATTATATAAC
61 M Q Q L Q S A L Q T G T E E L R S L Y N

241 ACAGTAGCAACTCTCTATTGTGTACATAAAGAGATAGATGTACGAGACACCAAGGAAGCC
81 T V A T L Y C V H K E I D V R D T K E A

301 TTAGACAAGATAGAGGAAGAACAATAAGAGTCAGCAAAAAACAGCAAGCAGAAGCG
101 L D K I E E E Q N K S Q Q K T Q Q A E A

361 GCTGACAAAGGAAAGGTCACTCAAATATCCAAATAGTGCAGAATCTCCAAGGGCAAATG
121 A D K G K V S Q N Y P I V Q N L Q G Q M

421 GTACACCAGGCCATATCACCGAGAAGTTAAATGCATGGGTAAGTAATAGAAGAGAAG
141 V H Q A I S P R T L N A W V K V I E E K

481 GCTTTCAGCCCAGAGGTAATACCCATGTTTACAGCATTATCAGAAGGAGCTACCCACAA
161 A F S P E V I P M F T A L S E G A T P Q

541 GATTTAAACACCATGTTAAATACAGTGGGGGACACCAAGCAGCCATGCAAAATGTTAAA
181 D L N T M L N T V G G H Q A A M Q M L K

601 GATACCATCAATGAGGAGGCTGCAGAATGGGATAGGTTACATCCAGTGCATGCAGGCCT
201 D T I N E E A A E W D R L H P V H A G P

661 ATTGCAACCAGGCCAATGAGAGAACCAAGGGGAAGTGACATAGCAGGAACTACTAGTACC
221 I A P G Q M R E P R G S D I A G T T S T

721 CTTCAGAACAATAGCATGGATGACAAGTAACCCACCTATCCGGTGGGAGACATCTAT
241 L Q E Q I A W M T S N P P I P V G D I Y

781 AAAAGATGGATAATCTGGGGTTAAATAAAATAGTAAGAAATGTATAGCCCTGTGAGCATT
261 K R W I I L G L N K I V R M Y S P V S I

841 TTGGACATAAAAAGGGCCAAAAGAACCCTTTAGAGACTATGTAGACCGATTCTTTAAA
281 L D I K Q G P K E P F R D Y V D R F F K

901 ACTTTAAGGGCTGAACAATCTTACAAGAGGTAAAAAATTGGATGACAGACACCTTGTG
301 T L R A E Q S S Q E V K N W M T D T L L

961 GTCCAAAATGCAAAACCCAGATTGTAAGACCATTTTAAGAGCATTAGGACCAGGGGCTACA
321 V Q N A N P D C K T I L R A L G P G A T

1021 TTAGAAGAAATGATGACAGCATGTCAGGGAGTGGGAGGACCTGGCCACAAAGCAAGAGTT
341 L E E M M T A C Q G V G G P G H K A R V

B1. Nucleotide and Amino Acid Sequences of gag clones

1081 TTGGCTGAGGCAATGAGCCAAGCAAATACAAACATAATGATGCAGAAAAGCAATTTTAAA
361 L A E A M S Q A N T N I M M Q K S N F K

1141 GGCCTAAAAGAAGCTGTTAAATGTTTCAATTGTGGCAAGGAAGGCATATAGCCAGAAAT
381 G P K R T V K C F N C G K E G H I A R N

1201 TGCAGGGCCCTAGGAAAAAGGCTGTTGAAATGTGGAAGGAAGGACACCAAATGAAA
401 C R A P R K K G C W K C G K E G H Q M K

1261 GACTGTACTGAAAGGCGGCTAATTTTTTAGGAAAATTTGGCCTTCCTACAAGGGGAGG
421 D C T E R Q A N F L G K I W P S Y K G R

1321 TCGGGGAATTCCTTCAGAGCAGACCAGGCCATCAGCTCCACCAGCAGAGAGCTTCAGG
441 S G N F L Q S R P E P S A P P A E S F R

1381 TTCAGGAGCGGGAGCCGAAAGACAGGAACCCCTTAACTTCCTCAAATCAGCTCTT
461 F E E R E P K D K E P P L T S L K S L F

1441 GGCAGCGACCCCTCGTCACAATAA
481 G S D P S S Q *

Translation of TV003G15(1-1476)

Universal code

Total amino acid number: 491, MW=54812

Max ORF: 1-1473, 491 AA, MW=54812

1 ATGGTGCAGAGCGTCAATATTAAGAGGGGAAAATTAGATAAATGGGAAAAATTAGG
1 M G A R A S I L R G G K L D K W E K I R

61 TTAAGCCAGGGGAAAAGAAACGCTATATGATAAAACACCTAGTATGGGCAAGCAGAGAG
21 L R P G G K K R Y M I K H L V W A S R E

121 CTGAAAAATTCGCACTTAAACCTGGCCTTTTAGAGACATCAGAAGGATGTAACAGATA
41 L E K F A L N P G L L E T S E G C K Q I

181 ATGAAACAGCTACAACCGCTCTTCAGACAGGAACAGAGGAAGTCTAGATCATTATTCAAC
61 M K Q L Q P A L Q T G T E E L R S L F N

241 ACCATAGCAGTTCTCTATTGTGTACATGAAAAGATAGAGGTACAGACACCAAGGAAGCC
81 T I A V L Y C V H E K I E V Q D T K E A

301 TTAGACAAGATAGAGGAAGAAACAAAACAAAAGTCAGCAAAAAACACAGCAGGCAGCAGCA
101 L D K I E E E Q N K S Q Q K T Q Q A A A

361 GCTGACGGAAAAGTCAGTCAAAATATCCTATAGTGCAGAAATGCCCAAGGGCAAAATGGTG
121 A D G K V S Q N Y P I V Q N A Q G Q M V

421 CACCAGGCATATCACCTAGGACTTTGAATGCATGGGTAAAAGTAATAGAGGAGAAGGCT
141 H Q S I S P R T L N A W V K V I E E K A

481 TTTAGCCCAGAGGTAATACCCATGTTTACAGCATTATCAGAAGGAGCCACCTCACAAGAC
161 F S P E V I P M F T A L S E G A T S Q D

541 TTAACACCATGCTAAATACAGTGGGGGACATCAAGCAGCCATGCAAAATGTTAAAAGAT
181 L N T M L N T V G G H Q A A M Q M L K D

601 ACCATCAATGAGGAGGCTGCAGAATGGGATAGAAATACATCCAGTACATGCGGGGCCTATT
201 T I N E E A A E W D R I H P V H A G P I

661 GCACCAGCCAAATGAGAGAACCAAGGGGAAGTGCATAGCAGGAAGTACTAGTACCCTT
221 A P G Q M R E P R G S D I A G T T S T L

721 CAGGAACAATAGCATGGATGACAAGTAATCCACCTATCCCAGTGGGAGACATCTATAAA
241 Q E Q I A W M T S N P P I P V G D I Y K

781 AGATGGATAATTTTGGGGTTAAATAAAATAGTAAGAATGTATAGCCCTGTCAGCATTG
261 R W I I L G L N K I V R M Y S P V S I L

841 GACATAAAACAAGGGCCAAAGGAACCTTTAGAGACTATGTAGACAGGTTCTTTAAAAC
281 D I K Q G P K E P F R D Y V D R F F K T

901 TTAAGAGCTGAACAAGCTACACAAGATGTAATAAAATGGATGACAGAAACCTTGTGGTC
301 L R A E Q A T Q D V K N W M T E T L L V

B1. Nucleotide and Amino Acid Sequences of gag clones

961 CAAAATGCAAACCCAGATTGTAAGACCATTTTAAGAGGGTTAGGAACAGGGGCTACATTA
321 Q N A N P D C K T I L R G L G T G A T L
1021 GAGGGAATGATGACAGCATGTCAGGGAGTGGGAGGACCTGGCCATAAAGCAAGAGTGTTA
341 E G M M T A C Q G V G G P G H K A R V L
1081 GCTGAAGCAATGAGCCAAGCAACATATAACATAATGATGCAGAGAAGCAATTTTAAAGGC
361 A E A M S Q A T Y N I M M Q R S N F K G
1141 TCTAGAAAAATGTTAAATGTTTCAACTGTGGCAGGAAAGGGCACATAGCCAGAAATTGC
381 S R K I V K C F N C G R K G H I A R N C
1201 AGGGCCCTAGAAAAAGGGCTGTTGAAATGTGAAAGGAAGGACACCAATGAGAGAA
401 R A P R K K G C W K C G K E G H Q M R E
1261 TGTAAGGAAAGCAGGCTAATTTTATAGGAAATTTGGCCTTCCACAAGGGGAGGCCA
421 C T E K Q A N F L G K I W P S H K G R P
1321 GGAATTTCTTCAGAGCAGACCAGGCCAACAGCCCCACCAGCAGAGAGCTTCAGGTTCC
441 G N F L Q S R P E P T A P P A E S F R F
1381 GAGGAGACACCCCGCGATGAAGCAGGAACCGAAAGCAGGGAACCCCTTAACTCCCTC
461 E E T P P A M K Q E P K D R E P L T S L
1441 AAATCACTCTTGGCAGCGACCCCTCGTCACAATAA
481 K S L F G S D P S S Q *

Translation of TV004G17 (1-1479)

Universal code
Total amino acid number: 492, MW=54889
Max ORF: 1-1476, 492 AA, MW=54889

1 ATGGTGCAGAGCGTCAATATTAAGAGGGGAAAATTAGATAAATGGAAAAAATTAGG
1 M G A R A S I L R G G K L D K W E K I R
61 TTAAGGCCAGGGGAAAAGAAACATTATATGATAAAACACCTAGTATGGCAAGCAGGGAG
21 L R P G G K K H Y M I K H L V W A S R E
121 CTGAAAGATTGCACTTAAACCTGGCCTTTAGAGACAGCAGAGGGCTGTAACAAATA
41 L E R F A L N P G L L E T A E G C K Q I
181 ATAAAACAGCTACATCCAGCTCTTCAGACAGGAACAGAGGAACCTAGATCATTATACAAC
61 I K Q L H P A L Q T G T E E L R S L Y N
241 ACCGTGGTAACTCTTTATTGCGTACATGCAGAGATAGAGGTACGAGACACCAAGGAAGCC
81 T V V T L Y C V H A E I E V R D T K E A
301 TTAGACAAGATAGAGGAAGAAACAAAACAAAAGTCAGCAAAAAACACAGCAGGCAAAAGCG
101 L D K I E E E Q N K S Q Q K T Q Q A K A
361 GCTGACGGAAGTCAAGTCAAAATATCCTATAGTACAGAATCTCCAGGGCGAATGGTA
121 A D G K V S Q N Y P I V Q N L Q G R M V
421 CACCAAGCCATATCACCTAGAACCTTGAATGCATGGGTAAAAGTAATAGAGGAAAAGGCT
141 H Q A I S P R T L N A W V K V I E E K A
481 TTAGCCCAGAGGTAATACCCATGTTTACAGCATTATCAGAAGGAGCCACCCCAAGAC
161 F S P E V I P M F T A L S E G A T P Q D
541 TTAACACCACTGTTAAATACAGTGGGGGACATCAAGCAGCCATGCAATGTTAAAGAT
181 L N T M L N T V G G H Q A A M Q M L K D
601 ACCATCAACGAGGAGGCTGCAGAATGGGATAGATTACATCCAGCACAGGCAGGGCCTGTT
201 T I N E E A A E W D R L H P A Q A G P V
661 GCACCAGGCCAAATAGAGAACCAAGGGGAAGTGCATAGCAGGAACCTACTAGTACCCTT
221 A P G Q I R E P R G S D I A G T T S T L
721 CAGGAACAAATAACATGGATGACAAGTAACCCACCTGTTCCAGTGGGAGAAATCTATAAA
241 Q E Q I T W M T S N P P V P V G E I Y K
781 AGATGGATAATTCTGGGGTTAAATAAAATAGTAAGGATGTATAGCCCTGTCAGCATTTG
261 R W I I L G L N K I V R M Y S P V S I L

B1. Nucleotide and Amino Acid Sequences of gag clones

841 GACATAAAACAAGGGCCAAAGGAACCCCTTAGAGACTATGTAGACCGGTTCTTTAAACT
281 D I K Q G P K E P F R D Y V D R F F K T

901 TTAAGAGCTGAACAGGCTACACAAGAAGTAAAGGCTGGATGACAGACACCTTATTGGTC
301 L R A E Q A T Q E V K G W M T D T L L V

961 CAAAATGCGAACCCAGATTGTAAGACCATTTTAAGAGCATTAGGACCAGGGGCTACACTA
321 Q N A N P D C K T I L R A L G P G A T L

1021 GAAGAAATGATGACAGCATGTCAGGGAGTGGGAGGACCTAGCCACAAGGCAAGAGTGTG
341 E E M M T A C Q G V G G P S H K A R V L

1081 GCTGAGGCAATGAGCCAAACAAACAGTGAAGCATAATGATGCAGAAAAGCAATTTTAAA
361 A E A M S Q T N S A S I M M Q K S N F K

1141 GGAGCCAAAAGAAATGTTAAATGCTTCAACTGTGGCAAGGAGGGGCACATAGCCAGAAAT
381 G A K R I V K C F N C G K E G H I A R N

1201 TGCAGGGCCCTTAGGAAAAAGGCTGTTGAAATGTGGACAGGAAGGACCAATGAAA
401 C R A P R K K G C W K C G Q E G H Q M K

1261 GACTGTACTGAGAGGCAGGCTAATTTTTAGGGAAAATTTGGCCTTCCCAAAAGGAAGG
421 D C T E R Q A N F L G K I W P S H K G R

1321 CCAGGGAATTCCTTCAGAACAGACCAGCCAAACAGCACCACCAGCAGAGAGCTTCAGG
441 P G N F L Q N R P E P T A P P A E S F R

1381 TTCGAGGAGACAACACCCTCCGAAGCAGGAGCCGAAGGACAGGGAACTTTAACTCC
461 F E E T T P T P K Q E P K D R E P L T S

1441 CTCAAATCACTCTTTGGCAGCGACCCCTCGTCACAATA
481 L K S L F G S D P S S Q *

Translation of TV004G24 (1-1479)

Universal code

Total amino acid number: 492, MW=54793

Max ORF: 1-1476, 492 AA, MW=54793

1 ATGGTGCGAGAGCGTCAATATTAAGAGGGGAAAATTAGATAAATGGGAAAAAATTAGG
1 M G A R A S I L R G G K L D K W E K I R

61 TTAAGGCCAGGGGAAAGAACATTATATGATAAAACCTAGTATGGGCAAGCAGGGAG
21 L R P G G K K H Y M I K H L V W A S R E

121 CTGAAAAGATTGCACCTTAACCTGGCCTTTAGAGACAGCAGGGCTGTAACAATA
41 L E R F A L N P G L L E T A E G C K Q I

181 ATAAAACAGCTACATCCAGCTCTCAGACAGGAACAGGGAAGCTAGATCATTATATAAC
61 I K Q L H P A L Q T G T E E L R S L Y N

241 ACCGTGGCAACTCTTTATTGCGTACATGCAGAGATAGAGGTACGAGACCCAAGGAAGCC
81 T V A T L Y C V H A E I E V R D T K E A

301 TTAGACAAGATAGAGGAAGAACAAAACAAAAGTCAGCAAAAACACAGCAGGCAAAAGCG
101 L D K I E E E Q N K S Q Q K T Q Q A K A

361 GCTGACGGAAAAGTCAGTCAAATATCCTATAGTACAGAATCTCCAAGGGCAAATGGTA
121 A D G K V S Q N Y P I V Q N L Q G Q M V

421 CACCAGCCATATCACCTAGAACCTTGAATGCATGGGTAAGTAATAGAGGAAAAGGCT
141 H Q A I S P R T L N A W V K V I E E K A

481 TTTAGCCCAGAGGTAATACCCATGTTTACAGCATTATCAGAAGGAGCCACCCCAAGAC
161 F S P E V I P M F T A L S E G A T P Q D

541 TTAACACCATGTTAAATACAGTGGGGGACATCAAGCAGCCATGCAAAATGTTAAAGAT
181 L N T M L N T V G G H Q A A M Q M L K D

601 ACCATCAACGAGGAGGCTGCAGAATGGGATAGATTACATCCAGCACAGGCGGGCCTGTT
201 T I N E E A A E W D R L H P A Q A G P V

661 GCACCAGGCCAAAATAAGAGAACCAAGGGGAAGTGACATAGCAGGAACTACTAGTACCCCT
221 A P G Q I R E P R G S D I A G T T S T L

B1. Nucleotide and Amino Acid Sequences of gag clones

721 CAGGAACAAATAACATGGATGACAAGTAACCCACCTGTTCCAGTGGGAGAAATCTATAAA
241 Q E Q I T W M T S N P P V P V G E I Y K

781 AGATGGATAATTCTGGGGTTAAATAAAATAGTAAGGATGTATAGCCCTGTCAGCATTTTG
261 R W I I L G L N K I V R M Y S P V S I L

841 GACATAAAACAAGGGCCAAAGGAACCCCTTAGAGACTATGTAGACCGGTTCTTTAAAACT
281 D I K Q G P K E P F R D Y V D R F F K T

901 TTAAGACTGAACAGGCTACACAAGAAGTAAAGGCTGGATGACAGACACCTTATGGTC
301 L R A E Q A T Q E V K G W M T D T L L V

961 CAAATGCGAACCCAGATTGTAAGACCATTTTAAGAGCATTAGACCAGGGCTACACTA
321 Q N A N P D C K T I L R A L G P G A T L

1021 GAAGAAATGATGACAGCATGTCAGGAGTGGGAGGACCTAGCCACAAGGCAAGAGTGTG
341 E E M M T A C Q G V G G P S H K A R V L

1081 GCTGAGCAATGAGCCAAACAACAGTGAAGCATAATGATGAGAAAGCAATTTTAAA
361 A E A M S Q T N S A S I M M Q K S N F K

1141 GGAGCCAAAAGAATTGTTAAATGCTTCAACTGTGGCAAGGAGGGGCACATAGCCAGAAAT
381 G A K R I V K C F N C G K E G H I A R N

1201 TGCAGGGCCCTAGGAAAAAAGGCTGTTGAAAATGTGGACAGGAAGGACCAAAATGAAA
401 C R A P R K K G C W K C G Q E G H Q M K

1261 GACTGTACTGAGAGACAGGCTAATTTTTAGGGAAAATTTGGCCTTCCCACAAAGGAAG
421 D C T E R Q A N F L G K I W P S H K G R

1321 CCAGGGAATTTCCCTCAGAACAGACCAGAGTCAACAGCACCACCAGCAGAGGCTTCAGG
441 P G N F L Q N R P E S T A P P A E S F R

1381 TTCAGGAGACAACCCACTCCGAAGCAGGAGCCGAAGGACAGGGAACCTTTAGCTTCC
461 F E E T T P T P K Q E P K D R E P L A S

1441 CTCAAATCACTCTTTGGCAGCGACCCCTCGTCACAATAA
481 L K S L F G S D P S S Q *

Translation of TV005G29(1-1479)

Universal code

Total amino acid number: 492, MW=55109

Max ORF: 1-1476, 492 AA, MW=55109

1 ATGGGTGCGAGAGCGTCAGTATTGAGAGGGGAAAAATTAGATGCATGGGAAAAAATTAGG
1 M G A R A S V L R G E K L D A W E K I R

61 TTAAGGCCAGGGGGAAAAAGTATATGTTAAAACACATAGTATGGGCAAGCAGGGAG
21 L R P G G K K K Y M L K H I V W A S R E

121 CTGAAAAGATTTGCACTTAACCCCTGGTCTTTTAGAAAACATTAGAAGGCTGTAACAAATA
41 L E R F A L N P G L L E T L E G C K Q I

181 ATGCAACAGCTACAACCAGCTCTCAGACAGGAACAGAGGAACCTTAAGTCATTATACAAC
61 M Q Q L Q P A L Q T G T E E L K S L Y N

241 ACAGTAGCAACTCTCTATTGTGCACAAAAAGGATAGATGTACGAGACACCAAGGAAGCC
81 T V A T L Y C A H K R I D V R D T K E A

301 TTAGACAAGATAGAGGAAGAACAACAAAAGTCAAGCAAAAAACACAGCAGACAAAAACG
101 L D K I E E E Q N K S Q Q K T Q Q T K T

361 GCTGACGAAAAGGTCAGTCAAATTTCCCTATAGTGCAGAATCTTCAAGGGCAAATGGTA
121 A D E K V S Q N F P I V Q N L Q G Q M V

421 CATCAAGCCATATCACCTAGAACCTTGAATGCATGGGTAAAGGTAATAGAGGAAAAGGCT
141 H Q A I S P R T L N A W V K V I E E K A

481 TTTAGCCCAGAGGTAATACCTATGTTTACAGCATTATCAGAAGGAGCCACCCACAAGAT
161 F S P E V I P M F T A L S E G A T P Q D

541 TTAACACCATGTTAAATACGGTGGGGGACATCAAGCAGCCATGCAAAATGTTAAAAGAT
181 L N T M L N T V G G H Q A A M Q M L K D

B1. Nucleotide and Amino Acid Sequences of gag clones

601 ACCATCAATGAAGAGGCTGCAGAATGGGATAGATTACATCCAGTACATGCGGGCCTATT
 201 T I N E E A A E W D R L H P V H A G P I

661 GCACCAGGCCAAATGAGAGAACCAAGGGGAAGTGACATAGCAGGAACTACTAGTACCCTT
 221 A P G Q M R E P R G S D I A G T T S T L

721 CAGGAACAAATAGCATGGATGACAAACAACCCACCTGTTCCAGTGGGAGACATCTATAAA
 241 Q E Q I A W M T N N P P V P V G D I Y K

781 AGATGGATAATTTTGGGGCTAAATAAATAGTGAAGATGTATAGCCCTGTCAGCATTTTG
 261 R W I I L G L N K I V R M Y S P V S I L

841 GATATAAGACAAGGCCAAAGGAACCTTTTAGAGACTATGTAGACCGGTTCTTTAAAAT
 281 D I R Q G P K E P F R D Y V D R F F K T

901 TTAAGAGCTGAACAAGCTACACAAGAGGTAAAAAATGGATGACAGACACCTTGTGTATC
 301 L R A E Q A T Q E V K N W M T D T L L I

961 CAAAATGCGAACCCAGATGTAAGACCATTTTAAGAGCATTAGGACCAGGGCTAGTTTA
 321 Q N A N P D C K T I L R A L G P G A S L

1021 GAAGAAATGATGACAGCATGTCAAGGAGTGGGAGGACCTGGCCACAAAGCAAGAGTGTG
 341 E E M M T A C Q G V G G P G H K A R V L

1081 GCTGAGGCAATGAGCCAGGCAAAACAATGCACACATAATGATGCAAGAAGCAATTTTAAA
 361 A E A M S Q A N N A H I M M Q R S N F K

1141 GGCTCAAAAAGAATTTGTTAAATGCTTCAACTGCGCAAGGAAGGGCACATAGCAAAAAAT
 381 G S K R I V K C F N C G K E G H I A K N

1201 TGCAGAGCCCCCAGGAAAAAAGGCTGTTGGAAATGTGAAAGGAAGGACCAAAATGAAA
 401 C R A P R K K G C W K C G K E G H Q M K

1261 GATTGTACTGAGAGACAGGCTAATTTTTTAGGGAAAATTTGGCCTTCCCACAAGGGGAG
 421 D C T E R Q A N F L G K I W P S H K G R

1321 CCAGGGAATTTCTTTCAGAGCAGGCCAGGCCAACCGGCCACCAGCAGAGAGCTTCAGG
 441 P G N F L Q S R P E P T A P P A E S F R

1381 TTCGAGGAGACAACATCAGCTCCACAGCAGGAGCCGAAAGACAGGGAACCCCTTAACTTCC
 461 F E E T T S A P Q Q E P K D R E P L T S

1441 CTCAAATCACTCTTTGGCAGCGACCCCTCGTCACAATAA
 481 L K S L F G S D P S S Q *

Translation of TV005G36(1-1479)

Universal code

Total amino acid number: 492, MW=55052

Max ORF: 1-1476, 492 AA, MW=55052

1 ATGGGTGCGAGAGCGTCAGTATTGAGAGGGGAAAAATTAGATGCATGGGAAAAAATTAGG
 1 M G A R A S V L R G E K L D A W E K I R

61 TTAAGGCCAGGGGAAAGAAAAAGTATATGTTAAAAACACATAGTATGGGCAAGCAGGGAG
 21 L R P G G K K K Y M L K H I V W A S R E

121 CTGGAAAGATTTGCACTTAAACCTGGTCTTTTAGAAACATCAGAAGGCTGTAACAATA
 41 L E R F A L N P G L L E T S E G C K Q I

181 ATGCAACAGCTACAACAGCTCTTCAGACAGGAACAGAGGAACTTAAGTCATTATACAAC
 61 M Q Q L Q P A L Q T G T E E L K S L Y N

241 ACAGTAGCAACTCTCTATTGTGTACACAAGGGATAGATGTACGAGACACCAAGGAAGCC
 81 T V A T L Y C V H K G I D V R D T K E A

301 TTAGACAAGATAGAGGAAGAACAACAAGTCAAGCAAAAAACACAGCAGGCAAAAAAG
 101 L D K I E E E Q N K S Q Q K T Q Q A K T

361 GCTGACGAAAAAGTCAGTCAAAATTTCCCTATAGTGCAAGATCTTCAGGGCAAAATGTA
 121 A D E K V S Q N F P I V Q N L Q G Q M V

421 CATCAAGCCATATCACCTAGAACCTTGAATGCATGGGTAAAAGTAATAGAGGAAAAGGCT
 141 H Q A I S P R T L N A W V K V I E E K A

B1. Nucleotide and Amino Acid Sequences of gag clones

481 TTTAGCCAGAGGTAATACCTATGTTTACAGCATTATCAGAAGGAGCCACCCACAAGAT
161 F S P E V I P M F T A L S E G A T P Q D

541 TTAACACCATGTTAAATACGGTGGGGGACATCAAGCAGCCATGCAAAATGTTAAAGAT
181 L N T M L N T V G G H Q A A M Q M L K D

601 ACCATCAATGAAGAGGCTGCAGAATGGGATAGATTACATCCAGTACATGCGGGCCTATT
201 T I N E E A A E W D R L H P V H A G P I

661 GCACCGCCAAATGAGAGAACAAGGGGAAGTGACATAGCAGGAACACTAGTACCCCTT
221 A P G Q M R E P R G S D I A G T T S T L

721 CAGGAACAAATAGCATGGATGACAAACACCCACCTGTCCAGTGGGAGACATCTATAAA
241 Q E Q I A W M T N N P P V P V G D I Y K

781 AGATGGATAATTTTGGGCTAAATAAATAGTGAGAATGTATAGCCCTGTGAGCATTTTG
261 R W I I L R L N K I V R M Y S P V S I L

841 GATATAAGACAAGGGCCAAAGGAACCTTTAGAGACTATGTAGACCGGTTCTTTAAAACT
281 D I R Q G P K E P F R D Y V D R F F K T

901 TTAAGAGCTGAACAAGCTACACAAGAGGTAATAAATGGATGACAGACATCTTGTGGTC
301 L R A E Q A T Q E V K N W M T D I L L V

961 CAAAATGCGAACCCAGATTGTAAGACCATTTTAAGAGCATTAGGACCAGGGGCTAGTTTA
321 Q N A N P D C K T I L R A L G P G A S L

1021 GAAGAAATGATGACAGCATGTCAAGGAGTGGGAGGACCTGGCCCAAAGCAAGAGTGTG
341 E E M M T A C Q G V G G P G H K A R V L

1081 GCTGAGGCAATGAGCCAGGCAAAACATGCACACATAATGATGCAGAGAAGCAATTTAAA
361 A E A M S Q A N N A H I M M Q R S N F K

1141 GGCTCAAAAAGAATTTGTTAAATGCTTCAACTGTGGCAAGGAAGGGCACATAGCCAAAAAT
381 G S K R I V K C F N C G K E G H I A K N

1201 TGCAGAGCCCCAGGAAAAAGGCTGTTGAAAATGTGGAAAGGAAGGACACCAAATGAAA
401 C R A P R K K G C W K C G K E G H Q M K

1261 GATTGTACTGAGAGACAGGCTAATTTTTTAGGAAAATTTGGCCTTCCCACAAGGGGAGG
421 D C T E R Q A N F L G K I W P S H K G R

1321 CCAGGAAATTTCTTCAGAGCAGGCCAGGCCAACGGCCCCACCAGCAGAGAGCTTCAGG
441 P G N F L Q S R P E P T A P P A E S F R

1381 TTCGAGGAGACAACATCAGCTCCAAAGCAGGAGCCGACAGACAGGAACCCCTTAACCTCC
461 F E E T T S A P K Q E P T D R E P L T S

1441 CTCAAATCACTCTTTGGCAGCGACCCCTCGTCACAATAA
481 L K S L F G S D P S S Q *

Translation of TV006G11(1-1468)

Universal code

Total amino acid number: 488, MW=54593

Max ORF: 2-1465, 488 AA, MW=54593

1 ATGGGTGCGAGCGTCAATATTTAAAGGGGGAAAAATTAGATGCATGGGAAAGAATTAGGTT
1 W V R A S I L K G G K L D A W E R I R L

61 AAGGCCAGGGGAAAGAAACACTATATGATAAAACATTTAGTATGGGCAAGCAGGGAGCT
21 R P G G K K H Y M I K H L V W A S R E L

121 GGAAAGATTTGCACTTAACCTGGCCTGTAGAGACATCAGAAGGATGTAACAAATAAT
41 E R F A L N P G L L E T S E G C K Q I M

181 GAACCGCTACAACCATCTCTTCAGACAGGAACAGAAGAACTTAGATCATTATACAACAC
61 N Q L Q P S L Q T G T E E L R S L Y N T

241 AGTAGCAACTCTTATTGTGTACATGAAAAGATAGAGGTACGAGACCAAGGAGCCTT
81 V A T L Y C V H E K I E V R D T K E A L

301 AGACAAGATAGAGGAAGAACAAAACAAAAGCCAGCAAAAAACACAGGCAAAAGCGGC
101 D K I E E E Q N K S Q Q K T Q Q A K A A

B1. Nucleotide and Amino Acid Sequences of gag clones

361 TGGCGAAAAGGTCAGTCAAATATCTATAGTGCAGAATGCCAAGGGCAAATGGTACA
 121 G E K V S Q N Y P I V Q N A Q G Q M V H

421 CCAAGCTATATACCTAGAAGCTTAAATGCATGGGTAAAAGTAATAGAGGAGAAGGCTTT
 141 Q A I S P R T L N A W V K V I E E K A F

481 CAGCCCAGAGGTAATACCCATGTTTACAGCATTATCAGAAGGAGCCACCCACAGATTT
 161 S P E V I P M F T A L S E G A T P Q D L

541 AAACACCATGTTAAATACAGTGGGAGGACATCAAGCAGCCATGCAAATGTTAAAAGATAC
 181 N T M L N T V G G H Q A A M Q M L K D T

601 CATCAATGAGGAAGCTGCAGAATGGGATAGGGTACATCCAGTGCATGCAGGGCCTGTGTC
 201 I N E E A A E W D R V H P V H A G P V A

661 ACCAGGACAGATGAGAGAACCAGGGGAAGTACATAGCAGGAAGTACTAGTACCCCTGCA
 221 P G Q M R E P R G S D I A G T T S T L Q

721 GGAACAAATAGCATGGATGACAGTAATCCACCTATTCCAGTAGGAGAAATTTATAAAAG
 241 E Q I A W M T S N P P I P V G E I Y K R

781 ATGGATAATCTGGGGTTAAATAAAATAGTAAGAATGTATAGCCCTGTCAGCATCTTGGGA
 261 W I I L G L N K I V R M Y S P V S I L D

841 CATAAACAAAGGCCAAAGGAACCCCTTTAGGGACTATGTAGACCGGTTCTTTAAACTTT
 281 I K Q G P K E P F R D Y V D R F F K T L

901 AAGAGCCGAACAGGCTACACAAGATGTAATAAAATGGATGACAGACCCCTGTGGTCCA
 301 R A E Q A T Q D V K N W M T D T L L V Q

961 AAATGCCAAGCCAGATTGTAAGACCAATTTAAGAGCATTAGGACCAGGGGCTTCATTAGA
 321 N A N P D C K T I L R A L G P G A S L E

1021 AGAAATGATGACAGCATGTCCAGGAGTGGGAGGACCTAGCCACAAGCAAGAGTGTGGC
 341 E M M T A C Q G V G G P S H K A R V L A

1081 TGAGGCAATGAGCCAAGCAAACAATATAAACATACTGATGCAGAGAAGCAATTTAAGGG
 361 E A M S Q A N N I N I L M Q R S N F K G

1141 CTCTAAGAGAATTGTTAAATGCTTCAACTGTGGCAAGGAAGGGCACATAGCCAGAAATG
 381 S K R I V K C F N C G K E G H I A R N C

1201 CAGGGCCCCTAGGAAAAGGGCTGTTGGAATGTGGAAGGAAGGACACCAATAAAGA
 401 R A P R K K G C W K C G K E G H Q I K D

1261 CTGTACTGAGAGGAGGCTAATTTTTAGGGAAAATTTGGCCTTCCCAGGGGAGGCC
 421 C T E R Q A N F L G K I W P S R K G R P

1321 AGGGAATTTCCCTCAGAACAGGCCAGAGCCAACAGCCCCACCAGCAGAAAGCTTCAGTT
 441 G N F L Q N R P E P T A P P A E S F R F

1381 CGAGGAGACAACCCCTGCGCCGAAGCAGGACAAGGAACCCCTTAACCTCCCTCAAATCACT
 461 E E T T P A P K Q D K E P L T S L K S L

1441 CTTTGGCAGCGACCCCTCGTCAATAA
 481 F G S D P S S Q *

Translation of TV006G97(1-1460)

Universal code

Total amino acid number: 485, MW=54316

Max ORF: 1-1227, 409 AA, MW=45623

1 ATGGTGCAGAGCGTCAACATTAAGGGGAAAATAGATGCATGGGAAAGAATTAGG
 1 M G A R A S T L K G G K L D A W E R I R

61 TTAAGGCCAGGGGAAAGAAACACTATATGATAAAACATTTAGTATGGGCAAGCAGGGAG
 21 L R P G G K K H Y M I K H L V W A S R E

121 CTGGAAGATTTGCACTTAACCCCTGGCCTGTAGAGACATCAGAAGGATGTAACAAATA
 41 L E R F A L N P G L L E T S E G C K Q I

181 ATGAACAGCTACAACCTCTCTTCAGACAGGAACAGAAGAAGTCTAGATCATTATACAAC
 61 M N Q L Q P S L Q T G T E E L R S L Y N

B1. Nucleotide and Amino Acid Sequences of gag clones

241 ACAGTAGCAACTCTCTATTGTGTACATGAAAAGATAGAGGTACGAGACACCAAGGAAGCC
81 T V A T L Y C V H E K I E V R D T K E A

301 TTAGACAAGATAGAGGAAGAACAAAACAAAAGCCAGCAAAAAACACAACAGGCAAGGCG
101 L D K I E E E Q N K S Q Q K T Q Q A K A

361 GCTGGCGAAAAGGTCAGTCAAAATTATCCTATAGTGCAGAATGCCAAGGGCAAATGGTA
121 A G E K V S Q N Y P I V Q N A Q G Q M V

421 CACCAAGCTATATCGCTAGAACGTTAAATGCATGGGTAAAAGTAATAGAGGAGAAGGCT
141 H Q A I S P R T L N A W V K V I E E K A

481 TTCAGCCAGAGGTAATACCCATGTTTACAGCATTATCAGAAGGAGCCACCCACAAGAT
161 F S P E V I P M F T A L S E G A T P Q D

541 TTAACACCATGTTAAATACAGTGGGAGGACATCAAGCAGCTATGCAAAATGTTAAAGAT
181 L N T M L N T V G G H Q A A M Q M L K D

601 ACCATCAATGAGGAAGCTGCAGAATGGGATAGGGTACATCCAGTGCATGCAAGGCCTGTT
201 T I N E E A A E W D R V H P V H A R P V

661 GCACCAGGACAGATGAGAGAACCAAGGGGAAGTGACATAGCAGGAACACTAGTACCCTG
221 A P G Q M R E P R G S D I A G T T S T L

721 CAGGAACAAATAGCATGGATGACAAGTAATCCACCTATCCAGTAGGAGAAATTTATAAA
241 Q E Q I A W M T S N P P I P V G E I Y K

781 AGATGGATAATTCTGGGGTTAAATAAAAATAGTAAGAATGTATAGCCCTGTGAGCATCTG
261 R W I I L G L N K I V R M Y S P V S I L

841 GACATAAAACAAGGGCCAAAGGAACCCCTTAGGGACTATGTAGACCGGTTCTTTAAAACT
281 D I K Q G P K E P F R D Y V D R F F K T

901 TTAAGAGCTGAACAAGCTACACAAGATGTAAAAAATGGATGACAGACACCTTGTGGTC
301 L R A E Q A T Q D V K N W M T D T L L V

961 CAAAATGCCAACCCAGATTGTAAGACATTTAAGAGCATTAGGGCCAGGGGCTTCATTA
321 Q N A N P D C K T I L R A L G P G A S L

1021 GAAGAAATGATAACAGCATGTCAGGGAGTGGGAGGACCTAGCCCAAAGCAAGAGTGTG
341 E E M I T A C Q G V G G P S H K A R V L

1081 GCTGAGCAATGAGCCAAGCAAACAATATAAACATACTGATGCAGAGAAGCAATTTAAG
361 A E A M S Q A N N I N I L M Q R S N F K

1141 GGCTCTAAGAGAATTGTTAAATGCTTCAACTGTGGCAAGGAAGGCACATAGCCAAAAAT
381 G S K R I V K C F N C G K E G H I A K N

1201 TGCAGACCCCTAGGAAAAAGGCTGTTGAAAATGTAGAAAAGAAAGACCAAAATGAAA
401 C R A P R K K G C * K C R K E R H Q M K

1261 GACTGTACTGAAAGGCAGGCTAATTTTTTAGGAAAATTTGGCCTTCCCAAGGGGAGG
421 D C T E R Q A N F L G K I W P S H K G R

1321 CCAGGGAATTTCCCTCAGAACAGGCCAGGCCAACAGCCCCACCAGCAGAAAAGCTTCAGG
441 P G N F L Q N R P E P T A P P A E S F R

1381 TTCGAGAAGACAACCCCTGCGCCGAAGCAGGACAAGGAACCCCTTAACTTCCCTCAAATCA
461 F E K T T P A P K Q D K E P L T S L K S

1441 CTCTTTGGCAGCGACCCCTC
481 L F G S D P

Translation of TV007G59 (1-1500)

Universal code

Total amino acid number: 499, MW=55511

Max ORF: 1-1497, 499 AA, MW=55511

1 ATGGGTGCGAGAGCGTCAATATTAAGAGGGGAAAATAGATAAATGGGAAGAAATTAGG
1 M G A R A S I L R G G K L D K W E E I R

61 TTAAGGCCAGGGGAAAGAAAACCTATAGGCTAAAACATCTAGTATGGGCAAGCAGGGAG
21 L R P G G K K T Y R L K H L V W A S R E

B1. Nucleotide and Amino Acid Sequences of gag clones

121 CTGGAAAGATTGCACTTAACCTGGCCTTTTAGAGACAGCAGAAGGCTGTAAACAAATA
41 L E R F A L N P G L L E T A E G C K Q I

181 ATAAGACAGCTACACCCAGCTCTTCAGACAGGAACCGAGGAAGCTTAGATCATTATACAAC
61 I R Q L H P A L Q T G T E E L R S L Y N

241 ACAGTAGCAACTCTCTATTGTGTACATGCAACATAGAGGTAAAAGACACCAAGGAAGCC
81 T V A T L Y C V H A N I E V K D T K E A

301 TTAGACAAGATAGAGGAAGAACAAAACAAAAGTCAGCAAAAATCAGAGCAGGCAAAAAGTA
101 L D K I E E E Q N K S Q Q K S E Q A K V

361 GGTAACGAAAAGATCAGTCAAAAATTATCCTATAGTGCAGAATCTCCAAGGGCAAATGGTA
121 G N E K I S Q N Y P I V Q N L Q G Q M V

421 CACCAGGCCTTATCACCTAGAACTTTGAATGCATGGGTAAAAGTAATAGAGGAGAAGGCT
141 H Q A L S P R T L N A W V K V I E E K A

481 TTCAGCCCAGAGGTAATACCCATGTTTACAGCATTATCAGAAGGAGCCACCCCAAGAT
161 F S P E V I P M F T A L S E G A T P Q D

541 TTAAACACCATGTTAAACACAGTGGGGGGCATCAAGCAGCCATGCAAAATGTTAAAGAC
181 L N T M L N T V G G H Q A A M Q M L K D

601 ACCATCAATGAAGAGGCTGCAGAATGGGATCGATTACACCCAGTACATGCAGGGCCTATT
201 T I N E E A A E W D R L H P V H A G P I

661 GCACCAGGCCAAATGAGAGAACCAAGGGGAAGTGACATAGCAGGAACACTACTAGCACCCCT
221 A P G Q M R E P R G S D I A G T T S T L

721 CAGGAACAAATAGCATGGATGACAAGTAACCCACCTATTCGGTGGGAGATATCTATAAA
241 Q E Q I A W M T S N P P I P V G D I Y K

781 AGATGGATAATCTGGGGTTAAATAAAATAGTAAGAATGTATAGCCCTGTCAGCATTTTG
261 R W I I L G L N K I V R M Y S P V S I L

841 GACATTAACAAGGGCCAAAGGAACCCCTTGTAGACTATGTAGACCGGTTCTTTAAACT
281 D I K Q G P K E P F R D Y V D R F F K T

901 TTAAGAGCTGAACAAGCTACACAAGATGTAATAAATGGATGACAGACACCTTGTGGTC
301 L R A E Q A T Q D V K N W M T D T L L V

961 CAAAATGCGAACCCAGATTGTAAGATCATTTTAAAGAGATTAGACCAGGGGCTACATTA
321 Q N A N P D C K I I L R G L G P G A T L

1021 GAAGAAATGATGACAGCATGTCAGGGAGTGGGAGGACCTAGCCCAAAGCAAGAGTGTG
341 E E M M T A C Q G V G G P S H K A R V L

1081 GCTGAGCAATGAGCCAAGCAAACAGTGGAAACATAATGATGCAGAAAAGCAATTTTAGA
361 A E A M S Q A N S G N I M M Q K S N F R

1141 GGCTCTAAAAGAATTATTAATGTTTAACTGTGGCAAGGAAGGGCACATAGCCAAAAAT
381 G S K R I I K C F N C G K E G H I A K N

1201 TGTAAGCCCTTAGGAAAAGAGGCTGTGGAAATGTGGAAAGGAAGGACCAAAATGAAA
401 C K A P R K R G C W K C G K E G H Q M K

1261 GACTGTACTGAAAGACAGGCTAATTTTTAGGGAAAATTTGGCCTTCTGCAAGGGGAGG
421 D C T E R Q A N F L G K I W P S C K G R

1321 CCAGGGAATTTCTTCAGAACAGGCCAGCCAAACAGCCCCACCAGCAGAGCCAAACAGCC
441 P G N F L Q N R P E P T A P P A E P T A

1381 CCACCAGCAGAGAGCCTCAGGATCGAGGAAACAACCCCGCTCCGAAGCCGGAGCCGAGG
461 P P A E S L R I E E T T P A P K P E P R

1441 GACAGGAAACCCCTTAATCTCCCTCAAATCACCCCTTGGCAGCGACCCCTCGTCACAATAA
481 D R E P L I S L K S P F G S D P S S Q *

Translation of TV008G65 (1-1476)

Universal code

Total amino acid number: 491, MW=54695

Max ORF: 1-1473, 491 AA, MW=54695

B1. Nucleotide and Amino Acid Sequences of gag clones

1 ATGGGTGCGAGAGCGTCAGTATTAAAGAGGCGAAAAATTAGATACATGGGAAAAATTAGG
1 M G A R A S V L R G E K L D T W E K I R

61 TTAAGGCCAGGGGAAAGAAACGCTATATGCTAAAACACATAGTATGGGCAAGCAGGGAG
21 L R P G G K K R Y M L K H I V W A S R E

121 CTGGAAGATTGCACTTAAACCCTGGCCTTTTAGAGACATCAGAAGGCTGTAACAATA
41 L E R F A L N P G L L E T S E G C K Q I

181 ATACAACAGCTACAACCAGCTCTTCAGACAGGAACAGAGGAAGCTAAATCGTTATCAAC
61 I Q Q L Q P A L Q T G T E E L K S L F N

241 ACAGTAGCAACTCTCTATTGTGTACATAAAAAGATAGAGGTTTCAGACACCAAGGAAGCC
81 T V A T L Y C V H K K I E V R D T K E A

301 TTAGACAAGATAGAGGAAGAACAAAACAAAGTCAGCAAAAACACAGCAGGCAGGAAGCG
101 L D K I E E E Q N K S Q Q K T Q Q A E A

361 GCTGACAAAAGGTCAGTCAAAATTATCCTATAGTACAGAAGCTCCAAGGGCAATGGTA
121 A D K K V S Q N Y P I V Q N L Q G Q M V

421 CACCAAGCCCTATCACCTAGAACTTTGAATGCATGGGTAAAAGTAATAGAGGAGAAGGCT
141 H Q A L S P R T L N A W V K V I E E K A

481 TTTGGCCAGAGGTAATACCCATGTTTACAGCATTATCAGAAGGAGCCACCCAGCAGAT
161 F G P E V I P M F T A L S E G A T P A D

541 TTAACACCATGTTAAATACAGTGGGGGACATCAGGCAGCCATGCAGATGTTAAAGAT
181 L N T M L N T V G G H Q A A M Q M L K D

601 ACCATCAATGAGGAGGCTGCAGAATGGGACAGATTACCCAGTACATGCAGGGCCTACT
201 T I N E E A A E W D R L H P V H A G P T

661 GCACCAGGCCAAATGAGAGAACCTAGGGGAAGTGACATAGCAGGAAGTACTAGTACCCTT
221 A P C Q M R E P R G S D I A G T T S T L

721 CAGGAACAAATAGCTCGGATGACAAGTAACCCACCTGTCCAGTGGGAGACATCTATAAA
241 Q E Q I A R M T S N P P V P V G D I Y K

781 AGATGGATAAATCTAGGGTTAAATAAAATAGTAAGAATGTATAGCCCTGTCCAGCATTTG
261 R W I I L G L N K I V R M Y S P V S I L

841 GACATAAAACAGGGGCCAAAAGAACCCTTTAGAGACTATGTAGACCGGTTCTTTAAACT
281 D I K Q G P K E P F R D Y V D R F F K T

901 TTAAGAGCTGAACAGCTACACAAGAGGTAAGGTTGGATGACAGACACCTTGTGGTC
301 L R A E Q A T Q E V K G W M T D T L L V

961 CAAAATGCGAACCCAGATTGTAAGACCATTTTAAAGCATTAGACCAGGGGCTACATTA
321 Q N A N P D C K T I L R A L G P G A T L

1021 GAAGAAATGATGACAGCATGTCAGGGAGTGGGAGGACCTGGCCACAAGCCAGAGTGTG
341 E E M M T A C Q G V G G P G H K A R V L

1081 GCTGAGGCAATGAGCCAAGCAAACAGTAACATACTTATGCAGAGAAGCAATTTAAAGGC
361 A E A M S Q A N S N I L M Q R S N F K G

1141 TCTAAAAGAATTTGTTAAATGTTCAACTGTGGCAAGGAAGGCACATAGCCGAAATGCG
381 S K R I V K C F N C G K E G H I A G N C

1201 AGGGCCCTAGAAAAAGGGCTGTTGGAAATGTGGAAAAGGACACCAATGAAAGAA
401 R A P R K K G C W K C G K E G H Q M K E

1261 TGTAAGTGAAGGAGGCTAATTTTTAGGGAAAATTTGGCCTTCCCAAGGGGAGGCA
421 C T E R Q A N F L G K I W P S H K G R P

1321 GGGAAATTCCTCCAGAGCAGACCAGGCCAACAGCCCCACCAGCAGAGAGCTTCAGGTT
441 G N F L Q S R P E P T A P P A E S F R F

1381 GAGGAGACAACCCCGCTCCGAAGCAGGAGTCGAAAGACAGGGAGCCCTTAACTTCCCTC
461 E E T T P A P K Q E S K D R E P L T S L

1441 AGATCACTCTTTGGCAACGACCCCTCGTCACAATAA
481 R S L F G N D P S S Q *

B1. Nucleotide and Amino Acid Sequences of gag clones

Translation of TV008G66 (1-1476)

Universal code

Total amino acid number: 491, MW=54899

Max ORF: 1-1473, 491 AA, MW=54899

```
1      ATGGGTGCGAGAGCGTCA GTATTAAGAGGCGAAAAATTGGATACATGGGAAAAGATTAGG
1      M G A R A S V L R G E K L D T W E K I R

61     TTAAGGCCAGGGGAAAGAACGCTATATGCTAAAACACATAGTATGGGCAAGCAGGGAG
21     L R P G G K K R Y M L K H I V W A S R E

121    CTGAAAAGATTGCACCTAACCCCTGGCCTTTTAGAGACATCAGAAGGCTGTAACAAATA
41     L E R F A L N P G L L E T S E G C K Q I

181    ATACAACAGCTACAACCAGCTCTTCAGACAGGAACAGAGGAAGCTAAATCATTATTCAAC
61     I Q Q L Q P A L Q T G T E E L K S L F N

241    ACAGTAGCAACTCTCTATTGTGTACACAGAAAGATAGAGGTACGAGACACCAAGAAGCC
81     T V A T L Y C V H R K I E V R D T K E A

301    TTAGACAAGATAGAGGAAGAACGAAACAAAAGTCAGCAAAAAACACAGCAGGCAGAAAGCG
101    L D K I E E E R N K S Q Q K T Q Q A E A

361    GCTGACAAAAAGGTCAGTCAAATATCCTATAGTACAGAATCTCCAAGGGCAAATGGTA
121    A D K K V S Q N Y P I V Q N L Q G Q M V

421    CACCAGGCCATACCTAGAACTTTGAATGCATGGGTAAGTAATAGAGGAGAAGGCT
141    H Q A L S P R T L N A W V K V I E E K A

481    TTTAGCCAGAGGTAATACCCATGTTTACAGCATTATCAGAAGGAGCCACCCAGCAGAT
161    F S P E V I P M F T A L S E G A T P A D

541    TTAACACCATGTTAAATACAGTGGGGGACATCAAGCAGCCATGCAGATGTTAAAGAT
181    L N T M L N T V G G H Q A A M Q M L K D

601    ACCATCAATGAGGAGGCTGCAGAATGGGACAGATTACACCCAGTACATGCAGGGCCTGCT
201    T I N E E A A E W D R L H P V H A G P A

661    GCACCAGGCCAAATGAGAGAACCTAGGGGAAGTGACATAGCAGGAAGTACTAGTACCCCTT
221    A P G Q M R E P R G S D I A G T T S T L

721    CAGGAACAAATAGCATGGATGACAAGTAACCCACCTGTCCCAGTGGGAGACATCTATAAA
241    Q E Q I A W M T S N P P V P V G D I Y K

781    AGATGGATAAATCTAGGGTTAAATAAAAATAGTAAGAATGTATAGCCCTGTCAGCATTTTG
261    R W I I L G L N K I V R M Y S P V S I L

841    GACATAAAACAGGGGCCAAAAGAACCCCTTTAGAGACTATGTAGACCGGTTCTTTAAACT
281    D I K Q G P K E P F R D Y V D R F F K T

901    TTAAGAGCTGAACAAGCTACACAAGAGGTAAAAGGTTGGATGACAGACCTTGTGGTC
301    L R A E Q A T Q E V K G W M T D T L L V

961    CAAAATGCGAACCCAGATTTGTAAGACCAATTTAAGAGCATTAGGACCAGGGCTACATTA
321    Q N A N P D C K T I L R A L G P G A T L

1021   GAAGAAATGATGACAGCATGTGAGGAGTGGGAGGACCTGGCCCAAAGCCAGAGTATTG
341   E E M M T A C Q G V G G P G H K A R V L

1081   GCTGAGGCAATGAGCCAAGCAAACAGTAACATATTTATGCAGAGAAGCAATTTAAAGGC
361   A E A M S Q A N S N I F M Q R S N F K G

1141   TCTAAAAGAATGTGTAATGTTTCAACTGTGGCAAGGAAGGCACATAGCCAAAAATTGC
381   S K R I V K C F N C G K E G H I A K N C

1201   AGGGCCCTAGAAAAAGGGCTGTTGAAATGTGAAAAGAAGGACCAAAATGAAAGAC
401   R A P R K K G C W K C G K E G H Q M K D

1261   TGTACTGAAAGGCGAGCTAATTTTTAGGGAAAAATTGGCCTTCCCAAGGGGAGGCCA
421   C T E R Q A N F L G K I W P S H K G R P

1321   GGGAAATTCCTCCAGAGCAGACCAGGCCAACAGCCCCACCAGCAGAGAAGTTCAGGTTCC
441   G N F L Q S R P E P T A P P A E N F R F
```

B1. Nucleotide and Amino Acid Sequences of gag clones

1381 GAGGAGACAACCCCGCTCCGAAGCAGGAGTCGAAAGACAGGGAGCCCTTAACCTCCCTC
461 E E T T P A P K Q E S K D R E P L T S L
1441 AGATCACTCTTTGGCAACGACCCCTCGTCACAATAA
481 R S L F G N D P S S Q *

Translation of TV009G12 (1-1479)

Universal code
Total amino acid number: 492, MW=54769
Max ORF: 1-1476, 492 AA, MW=54769

1 ATGGGTGCGAGAGCGTCAATATTAAGAGGGGAAAATTAGATAAATGGGAAAAAATTAGG
1 M G A R A S I L R G G K L D K W E K I R
61 TTAAGGCCAGGGGAAAAGGATATATGATAAAACACTTAGTATGGGCAAGCAGGGAG
21 L R P G G K K G Y M I K H L V W A S R E
121 CTGGAAGATTGCACTTAACTCTGGCCTTTTAGAAACAGCAGACGGCTGTAAACAAATA
41 L E R F A L N S G L L E T A D G C K Q I
181 CTACAACAGCTACAACCAGCTCTCCAGACAGGGACAGAGGAACCTTAGATCATTATATAAC
61 L Q Q L Q P A L Q T G T E E L R S L Y N
241 ACAGTGGCAACTCTCTATTGTGTACATAGCAATATTGGGGTACGAGACACCAAGGAAGCC
81 T V A T L Y C V H S N I G V R D T K E A
301 TTAGACAAGATAGAGGAAGAACAAAACAAAAGTCAGCAAAAAACACAGCAGGCAAAAGCG
101 L D K I E E E Q N K S Q Q K T Q Q A K A
361 GCTGAAGGAAAGGTCAGTCAAAATTATCCTATAGTGCAGAATCTTCAAGGGCAGATGGTA
121 A E G K V S Q N Y P I V Q N L Q G Q M V
421 CATCAGGCCATATCACCTAGAACCTTGAATGCATGGGTAAAAGTAATAGAGGAGAAGGCT
141 H Q A I S P R T L N A W V K V I E E K A
481 TTCAGCCAGAGGTAATACCCATGTTTACAGCATTATCAGAAGGAGCCACCCACAAGAT
161 F S P E V I P M F T A L S E G A T P Q D
541 CTAACACTATGTTAAATACAGTGGGGGACATCAAGCAGCCATGCAAAATGTTAAAAGAT
181 L N T M L N T V G G H Q A A M Q M L K D
601 ACCATCAATGAGGAGGCTGCGGAATGGGATAGGTTACATCCAGTACAGGACAGGCGCTATT
201 T I N E E A A E W D R L H P V Q A G P I
661 GCACCAGCCAAATAAGAGAACCAAGGGAAGTGACATAGCAGGAACTACTAGTACCCTT
221 A P G Q I R E P R G S D I A G T T S T L
721 CAGGAACAAATAGCATGGATGACAGGTAACCCAGTTATTCAGTGGGAGACATCTATAAA
241 Q E Q I A W M T G N P V I P V G D I Y K
781 AGATGGATAATTATGGGGTTAAATAAAATAGTAAGAATGTATAGCCCTGTGAGCATTG
261 R W I I M G L N K I V R M Y S P V S I L
841 GACATAAGACAAGGACCAAAGAACCCCTTTAGAGACTATGTAGACCGGTTCTTTAAAAC
281 D I R Q G P K E P F R D Y V D R F F K T
901 TTAAGAGCTGAACAAGCTACCCAAGAAAGTAAAAAATTGGATGACAGACACCTTGTGGTC
301 L R A E Q A T Q E V K N W M T D T L L V
961 CAAAATGCGAACCCAGATTGTAAGACCATTTTAAGAGCATTAGGACCAGGGGCTACATTA
321 Q N A N P D C K T I L R A L G P G A T L
1021 GAAGAAATGATGACAGCATGTCAGGGAGTGGGAGGACCTGGCCACAAAGCAAGATTTTG
341 E E M M T A C Q G V G G P G H K A R V L
1081 GCTGAGGCAATGAGCCAAGTAAACAATGCAAAACATAATGATGCAAAAGAAATTTTAAA
361 A E A M S Q V N N A N I M M Q R S N F K
1141 GGCTCCAAAAGAAATGTTAAATGTTTCAACTGTGGCAAGAAGGGCACATAGCCAGAAAT
381 G S K R I V K C F N C G K E G H I A R N
1201 TGCAGGGCCCTAGGAAAAAAGGCTGTTGAAATGTGGAAAGGAAGGACACCAATGAAA
401 C R A P R K K G C W K C G K E G H Q M K

B1. Nucleotide and Amino Acid Sequences of gag clones

1261 GACTGTACTGAGAGACAGGCTAATTTTTAGGGAAAATTGGCCTTCCCACAAGGGGAGG
421 D C T E R Q A N F L G K I W P S H K G R
1321 CCAGGGAATTTCTCCAGAGCAGACCCGAGCCATCAGCCCCGCCAGCAGAGACTTCAGG
441 P G N F L Q S R P E P S A P P A E S F R
1381 TTCGAGGAGACAACCCCACTCCGAAGCAGGAGTCGAAAAGACAGGGAAACCTTAACCTCC
461 F E E T T P T P K Q E S K D R E T L T S
1441 CTCAAATCACTCTTTGGCAGCGACCCCTCGTCACAATAA
481 L K S L F G S D P S S Q *

Translation of TV010G74(1-1482)

Universal code
Total amino acid number: 493, MW=55088
Max ORF: 1-1479, 493 AA, MW=55088

1 ATGGGTGCGAGAGCGTCAATATTAAGAGGCGGAAAATTAGATAAAATGGGAAAAAATTAGA
1 M G A R A S I L R G G K L D K W E K I R
61 TTAAGGCCAGGGGAAAAGAAACACTATATGTTAAAAACACATAGTATGGGCAAGCAGGGAG
21 L R P G G K K H Y M L K H I V W A S R E
121 CTGAAAAGATTGCACTTAACCCCTGGCCTTTTAGAGACATCAGAAGGCTGTAACAATA
41 L E R F A L N P G L L E T S E G C K Q I
181 ATACAACAGCTACACACAGCTCTTAAGACAGGAACAGAGGAACCTTACATCATTATACAAC
61 I Q Q L H T A L K T G T E E L T S L Y N
241 ACAGTAGCAACTCTCTACTGTGTACATGCAGGATAGAGGTACGAGACACCAAGGAGGCC
81 T V A T L Y C V H A G I E V R D T K E A
301 TTAGACAAGATAGAGGAGGAGCAAAAACAAAAGTCAGAAAAAATGCAGCAAGCAGAAGTG
101 L D K I E E E Q N K S Q K K M Q Q A E V
361 GCTGACAAAAAGAGGTCAAGTCAAAAATTATCTATAGTACAGAATCACCAAGGGCAAATG
121 A D K K K V S Q N Y P I V Q N H Q G Q M
421 GTACACCAGAACATATCACCAAGAACTTTAAATGCATGGGTAAAAGTAATAGAGGAGAAG
141 V H Q N I S P R T L N A W V K V I E E K
481 GGTTCACCCAGAGGTAATACCCATGTTTACAGCATTATCAGAGGGAGCCACCCCTTCT
161 G F N P E V I P M F T A L S E G A T P S
541 GATCTGAACCCATGTTTAAATATAGTGGGGGACATCAAGCAGCCATGCAAAATGTTAAAA
181 D L N T M L N I V G G H Q A A M Q M L K
601 GATACCATCAATGAGGAGGCTGCAGAATGGGATAGATTACCCAGCACAGGCAGGCCT
201 D T I N E E A A E W D R L H P A Q A G P
661 GTTGCAACAGGCCAAATCAGAGATCCAAGGGGAAGTGACATAGCAGGAACCTACTAGTACC
221 V A P G Q I R D P R G S D I A G T T S T
721 CTTAGGAAACAGTAACATGGATGACAAAATAACCCACCTATCCAGTAGGAGACATCTAT
241 L Q E Q V T W M T N N P P I P V G D I Y
781 AAAAGATGGATAATCTGGGATTAATAAAAATAGTAAGAATGTATAGCCCTGTCAGCATT
261 K R W I I L G L N K I V R M Y S P V S I
841 TTGACATTAGACAAGGACCAAAGGAGCCTTTTAGAGACTATGTAGATCGGTTCTTTAAA
281 L D I R Q G P K E P F R D Y V D R F F K
901 ACTTAAAGAGCTGAACAAGCTACACAAGATGTAATAAAATGGATGACAGACACCTTGTG
301 T L R A E Q A T Q D V K N W M T D T L L
961 GTCAAAATGCAACCCAGATTGTAAGACCATTTAAGAGCATTAGGACCAGGGGCTACA
321 V Q N A N P D C K T I L R A L G P G A T
1021 TTAGAAGAAATGATGACAGCATGTCAAGGAGTGGGAGGACCTAGCCACAAAGCAAGAGTC
341 L E E M M T A C Q G V G G P S H K A R V
1081 TTGGCTGAGCAATGAGCCAAGCAGGCAATACAAACATAATGATGCAGAAAAGCAATTC
361 L A E A M S Q A G N T N I M M Q K S N F

B1. Nucleotide and Amino Acid Sequences of gag clones

1141 AAAGGCCCTAGAAGAACTATTAATGCTTCAACTGTGGCAAGGAAGGACACCTAGCCAGA
 381 K G P R R T I K C F N C G K E G H L A R

1201 AATTGCAGGGCCCCCTAGGAAAAAGGCTGTTGGAAATGTGGAAAAGGAAGGACACCAAATG
 401 N C R A P R K K G C W K C G K E G H Q M

1261 AAAGACTGTACTGAGAGGCAGGCTAATTTTTAGGGAAAAATTTGGCCTTCCCACTCGGGG
 421 K D C T E R Q A N F L G K I W P S H S G

1321 AGGCCAGGGAACCTCCTTCAGAACAGACCAGGCCAACAGCCCCACCAGCAGAGAGCTTC
 441 R P G N F L Q N R P E P T A P P A E S F

1381 AGGTTTCGAGGAGACAACCCCGCTCAGAAGCAGGAGCCGCAAGACAGGGAACCCCTAACT
 461 R F E E T T P A Q K Q E P Q D R E P L T

1441 TCCCTCAAATCACTCTTTGGCGGCGACCCCTCGTCAACAATA
 481 S L K S L F G G D P S S Q *

Translation of TV012G34(1-1491)

Universal code
 Total amino acid number: 496, MW=55405
 Max ORF: 1-1488, 496 AA, MW=55405

1 ATGGGTGCGAGAGCGTCAATATTAAGAGGGGGAAAATTAGATAAATGGGAAAAAATTAGG
 1 M G A R A S I L R G G K L D K W E K I R

61 TTAAGGCCAGGGGGAAAAAACACTATATGCTAAAAACACCTAGTATGGGCAAGCAGAGAG
 21 L R P G G K K H Y M L K H L V W A S R E

121 CTGAAAGATTTCAGTTAACCCCTGGCCTTTTAGAGACATCAGACGGATGTAGACAAATA
 41 L E R F A V N P G L L E T S D G C R Q I

181 ATAAAACAGCTACAACCAGCTCTTCAGACAGGAACAGAGGAAATTAGATCATTATTTAAC
 61 I K Q L Q P A L Q T G T E E I R S L F N

241 ACAGTAGCAACTCTCTATTGTGTACATGAAGGGATAGATGTACGAGACCAAGGAAGCC
 81 T V A T L Y C V H E G I D V R D T K E A

301 TTAGACAAGTTGGAGGAGGAACAAAACAAATGTCAGCAAAAAACAGCAGGCAGAAGCG
 101 L D K L E E E Q N K C Q Q K T Q Q A E A

361 GCTGACAAAAAGGTCAGTCAAAATTATCCTATAGTGCAGAACCTCCAAGGGCAAATGGTA
 121 A D K K V S Q N Y P I V Q N L Q G Q M V

421 CACCAGGCCATATCACCTAGAACCTTGAATGCATGGGTAAGTAATAGAGGAGAAGGCT
 141 H Q A I S P R T L N A W V K V I E E K A

481 TTTAGCCCAGAGGTAATACCCATGTTTACAGCATTATCAGAAGGAGCCACCCACAAGAT
 161 F S P E V I P M F T A L S E G A T P Q D

541 TTAACACCATGTTAAATACAGTGGGGGACATCAAGCAGCCATGCAAATGTTAAAGAT
 181 L N T M L N T V G G H Q A A M Q M L K D

601 ACCATCAATGAGGAGGCTGCGAATGGGATAGGTTACATCCAGTACATGACAGGCCTGTT
 201 T I N E E A A E W D R L H P V H A G P V

661 GCACCAGCCAGATGAGAGAACCAAGGGGAAGTGACATAGCAGAACTACTAGTACCCTT
 221 A P G Q M R E P R G S D I A E T T S T L

721 CAAGAACAATAGCATGGATGACAAGTAACCCACCTATCCAGTAGGAGACATCTATAAA
 241 Q E Q I A W M T S N P P I P V G D I Y K

781 AGGTGGATAATTCTGGGGTTAAATAAAAATAGTAAGAAATGTACAGCCCTGTACAGATTTG
 261 R W I I L G L N K I V R M Y S P V S I L

841 GACATAAAACAAGGACCAAGGAACCCCTTTAGAGACTATGTAGACCGGTTCTTCAAAACT
 281 D I K Q G P K E P F R D Y V D R F F K T

901 TTAAGAGCTGAACAATCTACACAAGAGGTAATAAAATGGATGACAGACACCTTGTAGTC
 301 L R A E Q S T Q E V K N W M T D T L L V

961 CAAAATGCCAACCCAGATTGTAAGACCATTTTAAGAGCATTAGGACCAGGGGCTTCATTA
 321 Q N A N P D C K T I L R A L G P G A S L

B1. Nucleotide and Amino Acid Sequences of gag clones

1021 GAAGAAATGATGACAGCATGTCAGGGAGTGGGAGGACCTAGCCACAAAGCAAGAGCTTTG
341 E E M M T A C Q G V G G P S H K A R A L

1081 GCTGAGGCAATGAGCCAAGCAACAATGCAAGTGAATGATGCAGAAAAGCAATTTTAAA
361 A E A M S Q A N N A S V M M Q K S N F K

1141 GGCCCTAGAAGTACTGTTAAATGTTTCAACTGTGGCAAGGAAGGCACATAGCCAGGAAT
381 G P R S T V K C F N C G K E G H I A R N

1201 TGCAGGGCCCTAGGAAAAAGGACTGTTGGAATGTGGAAGGAAGGACACCAATGAAA
401 C R A P R K K D C W K C G K E G H Q M K

1261 GACTGTACTGAGAGACAGGCTAATTTTATAGGAAAATTTGGCCTTCCCACAGGGGAGG
421 D C T E R Q A N F L G K I W P S H K G R

1321 CCAGGGAATTTCTTCAGAGCAGGCCAGAGCCAAGCAGCCCACTAGAGCCAAGCC
441 P G N F L Q S R P E P T A P P L E P T A

1381 CCACCAGCAGAGAGCTTCAAGTTCGAGGAGACTCCGAAGCGGGAGCCGAAAGACAGGGAA
461 P P A E S F K F E E T P K R E P K D R E

1441 CCCTTAACTCCCTCAAATCACTCTTTGGCAGCGACCCCTCGTACAATAA
481 P L T S L K S L F G S D P S S Q *

Translation of TV012G40(1-1503)

Universal code

Total amino acid number: 500, MW=55961

Max ORF: 1-1500, 500 AA, MW=55961

1 ATGGGTGCGAGAGCGTCAATATTAAGAGGGGGAAAATTAGACAAATGGGAAAAAATTAGG
1 M G A R A S I L R G G K L D K W E K I R

61 TTAAGCCAGGGGGAAAAACGCTATATGCTAAAACACCTAGTATGGGCAAGCAGAGAG
21 L R P G G K K R Y M L K H L V W A S R E

121 CTGGCAGATTTCAGTTAACCTGGCCTTTTAGAGACATCAGACGGATGTAGACAAATA
41 L D R F A V N P G L L E T S D G C R Q I

181 ATAAAACAGCTACAACAGCTCTTCAGACAGGAACAGAGGAAATTAGATCATTATTTAAC
61 I K Q L Q P A L Q T G T E E I R S L F N

241 ACAGTAGCAACTCTCTATTGTGTACATAAAGGGATAGATGTACGAGACACCAAGGAAGCC
81 T V A T L Y C V H K G I D V R D T K E A

301 TTAGACAAGATAGAGGAGGAACAAAACAAATGCCAGCAAAAAACACAGCAGGCGGAAGCG
101 L D K I E E E Q N K C Q Q K T Q Q A E A

361 GCTGACAAAAAGGTGAGTCAAAATATCCTATAGTGCAGAACCTCCAAGGGCAATGGTA
121 A D K K V S Q N Y P I V Q N L Q G Q M V

421 CACCAGGCCATATCACCTAGAACCTTGAATGCATGGGTAAAAGTAATAGAGGAGAAGGCT
141 H Q A I S P R T L N A W V K V I E E K A

481 TTTAGCCCAGAGGTAATACCCATGTTTACAGCATTATCAGAAGGAGCCACCCACAAGAT
161 F S P E V I P M F T A L S E G A T P Q D

541 TTAACACCATGTTAAATACAGTGGGGGACATCAAGCAGCCATGCAATGTTAAAAGAT
181 L N T M L N T V G G H Q A A M Q M L K D

601 ACCATCAATGAGGAGGCTGCCGAATGGGATAGGTTACATCCAGTACATGAGGGCCTGTT
201 T I N E E A A E W D R L H P V H A G P V

661 GCACCAGGCCAGATGAGAGAACCAAGGGGAAGTGACATAGCAGAACTACTAGTACCCTT
221 A P G Q M R E P R G S D I A E T T S T L

721 CAAGAACAAATAGCATGGATGACAAGTAACCCACCTATCCAGTAGGAGACATCTATAAA
241 Q E Q I A W M T S N P P I P V G D I Y K

781 AGGTGGATAATTTCTGGGGTTAAATAAAATAGTAAGAATGTACAGCCCTGTACAGATTTTG
261 R W I I L G L N K I V R M Y S P V S I L

841 GACATAAAAACAGGACCAAAAAGAACCTTTTAGAGACTATGTAGACCGGTTCTTCAAACCT
281 D I K Q G P K E P F R D Y V D R F F K T

B1. Nucleotide and Amino Acid Sequences of gag clones

901 TTAAGAGCTGAACAATCTACACAAGAGGTAAAAAATTGGATGACAGACACCTTGTAGTC
301 L R A E Q S T Q E V K N W M T D T L L V

961 CAAAATGCGAACCAGATTGTAAGACCATTTAAGAGCATTAGGACCAGGGCTTCATTA
321 Q N A N P D C K T I L R A L G P G A S L

1021 GAAGAAATGATGACAGCATGTCAGGGAGTGGGAGGACCTACCCACAAAGCAAGATTTTG
341 E E M M T A C Q G V G G P T H K A R V L

1081 GCTGAGGCAATGAGCCAAGCAAACAATACAAGTGAATGATACAGAAAAGCAATTTAAA
361 A E A M S Q A N N T S V M I Q K S N F K

1141 GGCCCTAGAAGAGCTGTTAAATGTTTCAACTGTGGCAAGGAAGGGCACATAGCCAGGAAT
381 G P R R A V K C F N C G K E G H I A R N

1201 TGCAGGGCCCTAGGAAAAAGGGCTGTTGAAATGTGAAAGGAAGGACACCAAATGAAA
401 C R A P R K K G C W K C G K E G H Q M K

1261 GACTGTACTGAGAGACAGGCTAATTTTTTAGGGAAAAATTTGGCCTCCCAAGGAAGG
421 D C T E R Q A N F L G K I W P S H K G R

1321 CCAGGGAATTTCTTCAGAGCAGACCAGAGCCAACAGCCCCACCCTAGAACCAACAGCC
441 P G N F L Q S R P E P T A P P L E P T A

1381 CCACCAGCAGAGAGCTTCAAGTTCGAGGAGACTCCGAAGCAGGAGCCGAAAGACAGGGAA
461 P P A E S F K F E E T P K Q E P K D R E

1441 CCCTACAGGGAACCCCTTAACTTCCCTCAAATCACTCTTTGGCAGCGACCCCTCGTCAAA
481 P Y R E P L T S L K S L F G S D P S S Q

1501 TAA
501 *

Translation of TV013G2 (1-1527)

Universal code

Total amino acid number: 508, MW=56952

Max ORF: 1-1524, 508 AA, MW=56952

1 ATGGGTGCGAGAGCGTCAATATTAAGAGGGACGAAATTAGATGCATGGGAAAAAATTAGG
1 M G A R A S I L R G T K L D A W E K I R

61 TTAAGGCCAGGGGAAAGAAACATTATATGTTAAAAACACCTAGTATGGGCAAGCAGGGAG
21 L R P G G K K H Y M L K H L V W A S R E

121 CTGGAAAGATTGCGACTTAAACCTGGCCTTTTGAACATCGGAAGGCTGTAACAATA
41 L E R F A L N P G L L E T S E G C K Q I

181 ATGAAACAGCTACACCCAGCTCTTACAGCAGGAACAGAGGAACCTTAAATCATTATACAAC
61 M K Q L H P A L Q T G T E E L K S L Y N

241 ACAGTAGCAACTCTCTATTGTGTACATGAAAGCATAAAGGTACGAGACCCAAGGAAGCC
81 T V A T L Y C V H E S I K V R D T K E A

301 TTAGACAAGATAGAGGAAGAACAAAACAAATTTAAAGTCAGCAAAAAACACAGCAGGCA
101 L D K I E E E Q N K I K S Q Q K T Q Q A

361 AAAGCGGCTGACGAAAAAGTCAGTCAAATATCCTATAGTGCAGAAATCTTCAAGGGCAA
121 K A A D E K V S Q N Y P I V Q N L Q G Q

421 ATGGTACATCAGAACCTATCACCTAGAACCTTGAATGCATGGGTAAAAGTAATAGAGGAG
141 M V H Q N L S P R T L N A W V K V I E E

481 AAGGCTTTTAGCCAGAGGTAATACCCATGTTTACAGCATTATCAGAAGGAGCCACCCCA
161 K A F S P E V I P M F T A L S E G A T P

541 CAAGATTTAAACACCATGTTAAATACGGTGGGGGACATCAAGCAGCCATGCAAAATGTTA
181 Q D L N T M L N T V G G H Q A A M Q M L

601 AAAGATCCCATCAATGAAGAGGCTGCAGAATGGGATAGATTACACCCAGTCCATGCGGGG
201 K D P I N E E A A E W D R L H P V H A G

661 CCTATGGCACCAGGCCAATTGAGAGAACCAAGGGGAAGTGACATAGCAGGAACACTACTAGT
221 P M A P G Q L R E P R G S D I A G T T S

B1. Nucleotide and Amino Acid Sequences of gag clones

721 ACCCTTCAGGAACAAATAGCATGGATGACAAGTAATCCACCTATCCCAGTGGGAGACATC
 241 T L Q E Q I A W M T S N P P I P V G D I

781 TATAAAAGATGGATAATCTGGGGTTAAATAAAAATAGTGAGAATGTATAGCCCTATCAGC
 261 Y K R W I I L G L N K I V R M Y S P I S

841 ATTTTGGACATAAGACAAGGGCCAAAGGAACCCCTTGTAGACTATGTAGACCGGTTCTTT
 281 I L D I R Q G P K E P F R D Y V D R F F

901 AAAGCCTTAAGAGCTGAACAAGCTACACAAGATGTAAAAAATGGATGACAGAAACCTTG
 301 K A L R A E Q A T Q D V K N W M T E T L

961 CTGGTCCAAAATGCGAACCCAGATTGTAAGACCATTTTAAAAGCATTAGGAATAGGGGCT
 321 L V Q N A N P D C K T I L K A L G I G A

1021 ACATTGGAAGAAATGATGACAGCATGTCAGGGAGTGGGGGACCTAGTCACAAAGCAAGA
 341 T L E E M M T A C Q G V G G P S H K A R

1081 GTGTTAGCTGAGGCAATGAGCCAGCAAACAATAACAACATAATGATGCAGAGAAGCAAT
 361 V L A E A M S Q A N N T N I M M Q R S N

1141 TTTAAAAGCTCAAAAAGAAATGTTAAATGTTTCAACTGTGGCAAGGAAGGGCATATAGCC
 381 F K S S K R I V K C F N C G K E G H I A

1201 AGAAATGTCAGGGCCCTAGGAAAAGGGCTGTTGAAAATGTGAAAAGGAAGGACCCAA
 401 R N C R A P R K K G C W K C G K E G H Q

1261 ATGAAAGATTGTACTGAGAGGCAGGCAAATTTTTAGGGAAAATTTGGCCTCCCAAG
 421 M K D C T E R Q A N F L G K I W P S H K

1321 GGGAGCCAGGGAATTTCTTCAGAACAGACCAGAGCCAACAGCCCCACCAGCAGAGAGT
 441 G R P G N F L Q N R P E P T A P P A E S

1381 TTCAGGAACAGACCAGGCCAACGGCTCCACCAGCAGAGAGCTTCAGGTTTCAGGAGACA
 461 F R N R P E P T A P P A E S F R F E E T

1441 ACCCCCACTCCGAAGCAGGAGCCGAAAAGACAGGGATCCCTTAACTTCCCTCAAATCACTC
 481 T P T P K Q E P K D R D P L T S L K S L

1501 TTTGGCAGCGACCCCTCGTCACAATAA
 501 F G S D P S S Q *

Translation of TV013G15(1-1527)

Universal code

Total amino acid number: 508, MW=56910

Max ORF: 1-1524, 508 AA, MW=56910

1 ATGGGTGCGAGAGCGTCAATATTAAGAGGGACGAAATTAGATGCATGGGAAAAAATTAGG
 1 M G A R A S I L R G T K L D A W E K I R

61 TTAAGGCCAGGGGAAAGAAACATTATATGTTAAAAACACCTAGTATGGGCAAGCAGGGAG
 21 L R P G G K K H Y M L K H L V W A S R E

121 CTGGAAGATTGCACTTAACCTGGCCTTTTAGAAACATCAGAAGGCTGTAAACAATA
 41 L E R F A L N P G L L E T S E G C K Q I

181 ATGAAACAGCTACACCCAGCTCTTCAGACAGGAACAGAGGAACCTTAAATCATTATACAAC
 61 M K Q L H P A L Q T G T E E L K S L Y N

241 ACAGTAGCAACTCTCTATTGTGTACATGAAAACATAAAGGTACGAGACCCAAGGAAGCC
 81 T V A T L Y C V H E N I K V R D T K E A

301 TTAGACAAGATAGAGGAAGAACAACAATAAATTTAAAGTCAGCAAAAAACACAGCAGGCA
 101 L D K I E E E Q N K I K S Q Q K T Q Q A

361 AAAGCGGCTGACGAAAAGTCAGTCAAATATCCTATAGTGCAGAATCTTCAAGGGCAA
 121 K A A D E K V S Q N Y P I V Q N L Q G Q

421 ATGGTACATCAGAACCTATCACCTAGAACCTTGAATGCATGGGTAAAAGTAATAGAGGAG
 141 M V H Q N L S P R T L N A W V K V I E E

481 AAGGCTTTTAGCCCAGAGGTAATACCCATGTTTACAGCATTATCAGAAGGAGCCACCCCA
 161 K A F S P E V I P M F T A L S E G A T P

B1. Nucleotide and Amino Acid Sequences of gag clones

541 CAAGATTTAAGCACCATGTGTTAAATACGGTGGGGGGACATCAAGCAGCCATGCAAATGTTA
 181 Q D L S T M L N T V G G H Q A A M Q M L

601 AAAGATACCATCAATGAAGAGGCTGCAGAATGGGATAGATTACCCCAGTCCATGCGGGG
 201 K D T I N E E A A E W D R L H P V H A G

661 CCTATGGCACCAGGCCAATTGAGAGAACCAAGGGGAAGTGACATAGCAGGAACACTAGT
 221 P M A P G Q L R E P R G S D I A G T T S

721 ACCCTTCGGGAACAAATAGCATGGATGACAAGTAATCCACCTATCCCAGTGGGAGACATC
 241 T L R E Q I A W M T S N P P I P V G D I

781 TATAAAAGATGGATAATTCTGGGGTTAAATAAAATAGTGAGAATGTATAGCCCTGTCAGC
 261 Y K R W I I L G L N K I V R M Y S P V S

841 ATTTTGGACATAAGACAAGGGCCAAAGGAACCCCTTAGAGACTATGTAGACCGGTTCTTT
 281 I L D I R Q G P K E P F R D Y V D R F F

901 AAAGCCTTAAGAGCTGAACAAGCTACACAAGATGTAAAAAATTGGATGACAGAAACCTTG
 301 K A L R A E Q A T Q D V K N W M T E T L

961 CTGGTCCAAAATGCGAACCCAGATTGTAAGACCATTTTAAAAGCATTAGGAATAGGGGCT
 321 L V Q N A N P D C K T I L K A L G I G A

1021 ACATTGGAAGAAATGATGACAGCATGTCAGGGAGTGGGGGGACCTAGTCACAAAGCAAGA
 341 T L E E M M T A C Q G V G G P S H K A R

1081 GTGTTAGCTGAGGCAATGAGCCAAGCAAACAATAACAACATAATGATGACAGAGAAGCAAT
 361 V L A E A M S Q A N N T N I M M Q R S N

1141 TTTAAAAGCTCAAAAAGAATTGTTAAATGTTCCAAGTGGCAAGGAAGGGCATATAGCC
 381 F K S S K R I V K C S N C G K E G H I A

1201 AGAAATGCGAGGGCCCTAGGAAAAAGGGCTGTTGGAATGTGGAAGGAAGGACACCAA
 401 R N C R A P R K K G C W K C G K E G H Q

1261 ATGAAAGATTGTAAGTACTGAGAGGCGAGCAAAATTTTTAGGAAAATTTGGCCTTCCACAAG
 421 M K D C T E R Q A N F L G K I W P S H K

1321 GGGAGGCCAGGGAATTTCCCTTCAGAACAGACCAGAGCCAAACAGCCCCACCAGCAGAGAGT
 441 G R P G N F L Q N R P E P T A P P A E S

1381 TTCAGGAACAGACCAGAGCCAACGGCTCCACCAGCAGAGAGCTTCAGGTTCCAGGAGACA
 461 F R N R P E P T A P P A E S F R F E E T

1441 ACCCCCACTCCGAAGCAGGAGCCGAAAGACAGGGATCCCTTAAGTCCCTCAAACTACTC
 481 T P T P K Q E P K D R D P L T S L K S L

1501 TTGGCAGCGACCCCTCGTCACAATA
 501 F G S D P S S Q *

Translation of TV014G73 (1-1476)

Universal code
 Total amino acid number: 491, MW=54857
 Max ORF: 1-1473, 491 AA, MW=54857

1 ATGGGTGCGAGAGCGTCAATATTAAGAGGGGAAAAATTAGATAAATGGGAGAAAATTAGG
 1 M G A R A S I L R G E K L D K W E K I R

61 CTAAGGCCAGGGGAAGGAAACACTATATGCTAAAACATCTAGTATGGGCAAGCAGAGAG
 21 L R P G G R K H Y M L K H L V W A S R E

121 CTGGAAGATTGCACTTAACCCCTGGCCTTTAGAGACATCAAGGCTGTAACAATA
 41 L E R F A L N P G L L E T S Q G C K Q I

181 ATAAAACAGCTACCCCAGCTCTTAAGACAGGAACAGAGGAACCTTAGGTCATTATACAAC
 61 I K Q L H P A L K T G T E E L R S L Y N

241 ACAGTAGCAACTCTCTATGTGTACATGAAAACATAGAGGTACGAGACCAAGGAGGCC
 81 T V A T L Y C V H E N I E V R D T K E A

301 TTAGACAAGATAGAGGAAGAACAACAAGTCAAGCAAAAACACAGCAGGCAAAAGCG
 101 L D K I E E E Q N K S Q Q K T Q Q A K A

B1. Nucleotide and Amino Acid Sequences of gag clones

361 GCTGACGAAAGGAGTCACTCAAATATCCCATAGTGCAGAATCTCCAAGGGCAAATGGTA
 121 A D E G V S Q N Y P I V Q N L Q G Q M V

421 CACCAGGCCATATCACCTAGAACTTTGAATGCATGGGTGAAAGTAATAGAGGAGAAGGCT
 141 H Q A I S P R T L N A W V K V I E E K A

481 TTTAGCCCAGAAGTAATACCCATGTTTACAGCATTATCAGAAGGAGCCACCCACAAGAT
 161 F S P E V I P M F T A L S E G A T P Q D

541 TTAAACACCATGTTAAATACAGTAGGGGACATCAAGCAGCCATGCAGATGTTAAAAGAT
 181 L N T M L N T V G G H Q A A M Q M L K D

601 ACCATCAATGAGGAGGCTGCAGAATGGGATAGATTACATCCAGTCCATGCAGGGCCTGCT
 201 T I N E E A A E W D R L H P V H A G P A

661 GCACCAGGCCAAATGAGGGAACCTAGAGGAAGTGACATAGCAGGACTACTAGTACCCTT
 221 A P G Q M R E P R G S D I A G T T S T L

721 CAGGAACAAATAGCATGGATGACAGGTAACCCACCTGTCCCAGTGGGAGACATCTATAAA
 241 Q E Q I A W M T G N P P V P V G D I Y K

781 AGATGGATAATCTGGGGTTAAATAAAATAGTAAGAATGTATAGCCCTGTCAGCATTTTG
 261 R W I I L G L N K I V R M Y S P V S I L

841 GACATAAAACAAGGGCCAAAGGAACCCCTTTAGAGACTATGTAGATCGGTTCTTTAAAGTT
 281 D I K Q G P K E P F R D Y V D R F F K V

901 TTAAGAGCTGAACAAGCTACACAAGATGTAATAAAATGGATGACAGACACCTTGTGATC
 301 L R A E Q A T Q D V K N W M T D T L L I

961 CAAAATGCCAACCAGATTGTAAGACCATCTTAAAGGCATTGGGACCAGCGGCTTCATTA
 321 Q N A N P D C K T I L K A L G P A A S L

1021 GAAGAAATGATGACAGCATGTCAGGAGTGGGAGGACCTGGCCACAAGCAAGAGTGTG
 341 E E M M T A C Q G V G G P G H K A R V L

1081 GCTGAGGCAATGAGCCAAGCAACAGTAACATAATGATGCAGAGAAGCAATTTTAAAGGA
 361 A E A M S Q A N S N I M M Q R S N F K G

1141 TCTAAAAGAATGTTAAATGTTCAACTGTGGCAAGGAAGGGCACATAGCCAGAATGTC
 381 S K R I V K C F N C G K E G H I A R N C

1201 AGGGCCCCTAGAAAAGGGCTGTGGAAATGTGGACAAGAAGGACACCAATGAAAGAC
 401 R A P R K K G C W K C G Q E G H Q M K D

1261 TGTACTGAAAGGCAGGCTAATTTTTAGGAAAATTTGGCCTTCCACAGGGGAGGCCA
 421 C T E R Q A N F L G K I W P S H K G R P

1321 GGGAATTTCTCCAGAGCAGGCCAGGCCAACAGCCCCACCAGCAGAGGCTTCAGGTTT
 441 G N F L Q S R P E P T A P P A E S F R F

1381 GAGGAACAACCCCGCTCCGAAACAGGAGTCGAAGGACAGGGAACCCCTAATTTCCCTC
 461 E E T T P A P K Q E S K D R E P L I S L

1441 AAATCACTCTTTGGCAGCGACCCCTCGTCACAATAA
 481 K S L F G S D P S S Q *

Translation of TV018G60(1-1479)

Universal code

Total amino acid number: 492, MW=55099

Max ORF: 1-1476, 492 AA, MW=55099

1 ATGGGTGCGAGAGCGTCAATATTTAAAGGCGAAAAATAGATAGATGGGAAAGAATTAGG
 1 M G A R A S I L K G E K L D R W E R I R

61 TTAAGGCCAGGGGAAAGAAACATTATATGTTAAAACACATAGTATGGGCAAGCAGGGAG
 21 L R P G G K K H Y M L K H I V W A S R E

121 TTGGAAAAATTTGCACTTAACCTGGCCTTTTAGAAAACAGCAGAAGGCTGTAATCAAATA
 41 L E K F A L N P G L L E T A E G C N Q I

181 ATGAACCAGCTACAACCAGCTCTTCAGACAGGAACAGAGGAACCTTAAATCATTATTCAAC
 61 M N Q L Q P A L Q T G T E E L K S L F N

B1. Nucleotide and Amino Acid Sequences of gag clones

241 ACAGTAGCAACTCTCTATTGTGTACATAAAAAAGATAGATGTACGAGACACCAAGGAAGCC
81 T V A T L Y C V H K K I D V R D T K E A

301 TTAGATAAGATAGAGGAAGAACAAAAACAAAGTCAGCAAAAAACACAGCAGGCAAAAGCG
101 L D K I E E E Q N K S Q Q K T Q Q A K A

361 GCTGACGAAAAGGTGAGTCAAAATTATCCTATAGTACAAAATCTCCAAGGGCAATGGTA
121 A D E K V S Q N Y P I V Q N L Q G Q M V

421 CATCAAGCCATATCACCTAGAACCTTGAATGCATGGGTAAAAGTAATAGAGGAGAAGGCC
141 H Q A I S P R T L N A W V K V I E E K A

481 TTAGCCCAGAGGTAATACCCATGTTTACAGCATTATCAGAAGGAGCCACCCACAAAGAT
161 F S P E V I P M F T A L S E G A T P Q D

541 TTAACACCATGTTAAATACGGTGGGGGACATCAAGCAGCCATGCAAAATGTTAAAGAT
181 L N T M L N T V G G H Q A A M Q M L K D

601 ACCATCAATGAGGAGGCTGCAGAATGGGATAGATTACATCCAGTACATGCGGGCCTGTT
201 T I N E E A A E W D R L H P V H A G P V

661 GCACCAGGCCAAATGAGAGAACCAAGGGGAAGTGACATAGCAGAACTACTAGTACCCTT
221 A P G Q M R E P R G S D I A G T T S T L

721 CAGGAACAAATAGCATGGATTACAGCTAACCCACCTATCCAGTAGGAGAAATCTATAAA
241 Q E Q I A W I T A N P P I P V G E I Y K

781 AGATGGATAAATCTGGGGTTAAATAAAATAGTGAAGATGTATAGCCCTGTCAGCATTTTG
261 R W I I L G L N K I V R M Y S P V S I L

841 GACATAAGACAAGGACCAAGGAACCCCTTTAGAGACTATGTAGATCGGTTCTTTAAACT
281 D I R Q G P K E P F R D Y V D R F F K T

901 TTAAGAGCTGAACAAGCTACACAAGATGTAAAAAATGGATGACAGACACCTTGTGGTC
301 L R A E Q A T Q D V K N W M T D T L L V

961 CAAAATGCGAACCAGATTGTAAGACCATTTAAGAGCATTAGGACCAGGGGCTACATTA
321 Q N A N P D C K T I L R A L G P G A T L

1021 GAAGAAATGATGACAGCATGTCAGGGAGTGGGAGGACCTAGCCACAAAGCAAGAGTTTTG
341 E E M M T A C Q G V G G P S H K A R V L

1081 GCTGAGGCAATGAGCCAAGCAAAACAATGCAGTCATAATGATGCAGAAAAGCAATTTTAAA
361 A E A M S Q A N N A V I M M Q K S N F K

1141 GGTCTAGAAAAATATTAGATGTTTCAACTGTGGTAAGGAAGGCACATAGCCGAAAC
381 G P R K I I R C F N C G K E G H I A R N

1201 TGCAGGGCCCCTAGGAAAAAGGCTGTTGGAAATGTGGAAGGAGGGACACCAATGAAA
401 C R A P R K K G C W K C G K E G H Q M K

1261 GACTGTACTGAAAGGCAGGCTAATTTTTAGGAAAAATTTGGCCTTCCACAGGGGAGG
421 D C T E R Q A N F L G K I W P S H K G R

1321 CCAGGGAATTTCTTCAGAACAGACCAGGCCAACGCCCCACCAGCAGAGAGCTTCAAG
441 P G N F L Q N R P E P T A P P A E S F K

1381 TTCGAGGAGACAACCCCACTCCGAGGCAGGAGTCGAAAAGACAGGGAACCCCTAACTTCC
461 F E E T T P T P R Q E S K D R E P L T S

1441 CTCAAATCACTCTTTGGCAGCGACCCCTCGTCACAATAA
481 L K S L F G S D P S S Q *

Translation of TV019G20(1-1479)

Universal code

Total amino acid number: 492, MW=55105

Max ORF: 1-1476, 492 AA, MW=55105

1 ATGGGTGCGAGAGCGTCAATATTAAGAGGCGGAAAATTAGATACATGGGAAAAAATTAGG
1 M G A R A S I L R G G K L D T W E K I R

61 TTAAGGCCAGGGGAAAGAAACACTATATGCTAAAACATCTAGTATGGGCAAGCAGGGAG
21 L R P G G K K H Y M L K H L V W A S R E

B1. Nucleotide and Amino Acid Sequences of gag clones

121 CTGGAAGATTTGCACTTAACCCTGGCCTTTTAGAGACATCAGAAGGCTGTAACAAATA
41 L E R F A L N P G L L E T S E G C K Q I

181 ATAAGACAGCTACAACCAGCTCTTCAGACAGGAACAGAGGAACCTAAATCATTATATAAC
61 I R Q L Q P A L Q T G T E E L K S L Y N

241 ACAGTAGCAACTCTCTATTGTGTACATGCAAAGATAGAGGTACGAGACACCAAGGAAGCC
81 T V A T L Y C V H A K I E V R D T K E A

301 TTAGACAGGATAGAGGAAGAACAGAAAAATGTCAGCAAAAAACACAGCAGGCAAAAGAG
101 L D R I E E E Q K K C Q Q K T Q Q A K E

361 GCTGACGGGAAGATCAGTCAAATATCCTATAGTGCAGAACTTCTCAAGGGCAAATGGTA
121 A D G K I S Q N Y P I V Q N L Q G Q M V

421 CACCAGGCCATATCACCTAGAACTTTGAATGCATGGGTAAAAGTAATAGAGGAGAAGGCT
141 H Q A I S P R T L N A W V K V I E E K A

481 TTAGCCCAGAAGTAATACCCATGTTTACAGCATTATCAGAAGGAGCCACCCCAAGAT
161 F S P E V I P M F T A L S E G A T P Q D

541 TTAACACCATGCTAAATACAGTGGGGGACATCAAGCAGCCATGCAAAATGTTAAAAGAT
181 L N T M L N T V G G H Q A A M Q M L K D

601 ACCATCAATGAGGAGGCTGCAGAATGGGACAGAATACATCCAGTACATGCAGGGCCTATT
201 T I N E E A A E W D R I H P V H A G P I

661 GCACCAGGCCAAATGAGAGAACCAGGGGAAGTACATAGCAGGAACCTACTAGTACCCCT
221 A P G Q M R E P R G S D I A G T T S T L

721 CAGGAACAAATAGCATGGATGACAAGTAACCCACCTGTTCCAGTGGGAGAAATCTATAAA
241 Q E Q I A W M T S N P P V P V G E I Y K

781 AGATGGATAATCTGGGCCTAAATAAATAGTAAGAATGTATAGCCCTGTCAGCATTGTTG
261 R W I I L G L N K I V R M Y S P V S I L

841 GACATAAAACAAGGCCAAAGGAACCCCTTTAGAGATTATGTAGATCGGTTCTTTAAACT
281 D I K Q G P K E P F R D Y V D R F F K T

901 TTAAGAGCCGAACAAGCTACACAAGATGTAAAAAATGGATGACAGACACCTTGTGGTC
301 L R A E A A T Q D V K N W M T D T L L V

961 CAAAATGCGAACCCAGATTGTAAGATCATTTTAAGAGGATTAGGACCAGGGCTACATTA
321 Q N A N P D C K I I L R G L G P G A T L

1021 GAAGAAATGATGACAGCATGTGAGGAGTGGGAGGACCTGGCCACAAGCAAGAGTGTG
341 E E M M T A C Q G V G G P G H K A R V L

1081 GCTGAGGCAATGAGCCAAGCAAACAGTACAAATATAATGATGACAGAGGCAATTTTAAA
361 A E A M S Q A N S T N I M M Q R G N F K

1141 GGCCCTAAAAGAAACATTAATGTTTTAACTGTGGCAAGGAAGGCACCTAGCCAGAAAT
381 G P K R N I K C F N C G K E G H L A R N

1201 TACAGGGCCCTAGGAAAAAGGTTGTTGAAATGTGAAAAGAAGGACACCAATGAAA
401 Y R A P R K K G C W K C G K E G H Q M K

1261 GACTGTACAGAGAGACAGGCTAATTTTTAGGAAAATTTGGCCTTCCCACAAGGAAGG
421 D C T E R Q A N F L G K I W P S H K G R

1321 CCAGGGAACCTCCTTCAGAACAGAACAGGCCAACAGCCCAAGGACAGAGGCTTCAGG
441 P G N F L Q N R T E P T A P P A E S F R

1381 TTCGAGGAGACAAACCTGCTCCGAAGCAGGAGCCGAAAGACAGGGAACCCCTAACTTCC
461 F E E T N P A P K Q E P K D R E P L T S

1441 CTCAAATCACTCTTTGGCAGCGACCCCTCGTCACAATAA
481 L K S L F G S D P S S Q *

B1. Nucleotide and Amino Acid Sequences of gag clones

Translation of TV019G25(1-1479)

Universal code

Total amino acid number: 492, MW=54918

Max ORF: 1-1476, 492 AA, MW=54918

```
1      ATGGGTGCGAGAGCGTCAATATTAGGAGGCGGAAAAATTAGATACATGGGAAAAAATTAGG
1      M G A R A S I L G G G K L D T W E K I R

61     TTAAAGCCAGGGGAAAGAAACACTATATGCTAAAACATCTAGTATGGGCAAGCAGGGAG
21     L R P G G K K H Y M L K H L V W A S R E

121    CTGGAAGATTGTCACCTAACCCCTGGCCTTTTAGAGACATCAGAAGGCTGTAACAATA
41     L E R F A L N P G L L E T S E G C K Q I

181    ATAAGACAACCTACAACAGCTCTTCAGACAGGAACAGAGGAACCTAAATCATTATACAAC
61     I R Q L Q P A L Q T G T E E L K S L Y N

241    ACAGTAGCAACTCTCTATGTGTACATGCAAAGATAGAGGTACGAGACACCAAGGAAGCC
81     T V A T L Y C V H A K I E V R D T K E A

301    TTAGATAAGATAGAGGAAGAACAGAAAAATGTCAGCAAAAAACACAGCAGGCAAAAAGAG
101    L D K I E E E Q K K C Q Q K T Q Q A K E

361    GCTGACGGGAAGATCAGTCAAAATTATCCTATAGTGCAGAATCTTCAAGGGCAAATGGTA
121    A D G K I S Q N Y P I V Q N L Q G Q M V

421    CACCAGGCCATATCACCTAGAACTTTGAATGCATGGGTAAAAGTAATAGAGGAGAAGGCT
141    H Q A I S P R T L N A W V K V I E E K A

481    TTTAGCCCAGAAGTAATACCCATGTTTACAGCATTATCAGAAGGAGCCACCCCAAGAT
161    F S P E V I P M F T A L S E G A T P Q D

541    TTAACACCATGCTAAATACAGTGGGGGACATCAAGCAGCCATGCAAAATGTTAAAAGAT
181    L N T M L N T V G G H Q A A M Q M L K D

601    ACCATCAATGAGGAGGCTGCAGAATGGGACAGAATACATCCAGTACATGCAGGGCCTATT
201    T I N E E A A E W D R I H P V H A G P I

661    GCACCAGGCCAAATGAGAGAACCAAGGGGAAGTGACATAGCAGGAACACTAGTACCCTT
221    A P G Q M R E P R G S D I A G T T S T L

721    CAGGAACAAATAGCATGGATGACAAGTAACCCACCTGTTCCAGTGGGAGAAATCTATAAA
241    Q E Q I A W M T S N P P V P V G E I Y K

781    AGATGGATAAATCTGGGCCTAAATAAATAGTAAGAATGTATAGCCCTGTCAGCATTG
261    R W I I L G L N K I V R M Y S P V S I L

841    GACATAAAACAAGGACCAAGCAACCCCTTAGAGATTATGTAGACCGGTTCTTTAAACT
281    D I K Q G P K E P F R D Y V D R F F K T

901    TTAAGAGCCGAACAAGCTACACAAGATGTAATAAATGGAATGACAGACACCTTGTGGTC
301    L R A E Q A T Q D V K N W M T D T L L V

961    CAAAATGCGAACCCAGATTGTAAGATCATTTTAAGAGGATTAGGACCAGGGGCTACATTA
321    Q N A N P D C K I I L R G L G P G A T L

1021   GAAGAAATGATGACAGCATGTCAGGGAGTGGGAGGACCTGGCCACAAAGCAAGAGTGTG
341   E E M M T A C Q G V G G P G H K A R V L

1081   GCTGAGGCAATGAGCCAAGCAACAGTACAAATATAATGATGACAGAGGCAATTTTAAA
361   A E A M S Q A N S T N I M M Q R G N F K

1141   GGCCCTAAAAGAAACATTAATGTTTTAACTGTGGCAAGGAAGGCACCTAGCCAGAAAT
381   G P K R N I K C F N C G K E G H L A R N

1201   TGCAGGGCCCTAGGAAAAAGGTTGTTGGAAATGTGGAAAAGGACACCAAAATGAAA
401   C R A P R K K G C W K C G K E G H Q M K

1261   GACTGTACAGAGAGACAGGCTAATTTTTAGGGAAAAATTTGGCCTTCCCACAAGGAAGA
421   D C T E R Q A N F L G K I W P S H K G R

1321   CCAGGAAACTTCCCTCAGAACCGAACAGAGCCAACAGCCCCACCAGCAGAGAGCTTCAGG
441   P G N F L Q N R T E P T A P P A E S F R

1381   TTCGAGGAGACAAACCTGCTCCGAAGCAGGAGCCGAAAGACAGGGAACCTTAACTTCC
```


B1. Nucleotide and Amino Acid Sequences of gag clones

461 F E E T N P A P K Q E P K D R E P L T S
 1441 CTCAAATCACTCTTTGGCAGCGACCCCTCGTCACAATAA
 481 L K S L F G S D P S S Q *

Translation of TV016G95 (1-1508)

Universal code

Total amino acid number: 502, MW=55966

Max ORF: 1-1506, 502 AA, MW=55966

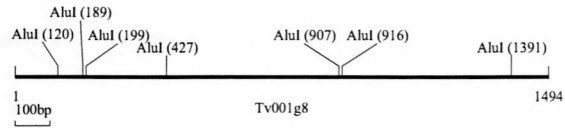
1 ATGGGTGCGAGAGCGTTCGGTATTAAGCGGGGAGAATTAGATAGATGGGAAAAAATTCGG
 1 M G A R A S V L S G G E L D R W E K I R
 61 TTGAGGCCAGGGGAAAGAAAAAGTATCAATTAACAATATAGTATGGGCAAGCAGGGAG
 21 L R P G G K K K Y Q L K H I V W A S R E
 121 CTAGAACGCTTCGCAGTTAACCCCTGGCCTGTTAGAAACATCAGAAGGTTGTAGACAAATA
 41 L E R F A V N P G L L E T S E G C R Q I
 181 CTGGGACAGCTACAGCCAGCCCTTCAGACAGGATCAGAAGGACTTAGATCATTATATAAT
 61 L G Q L Q P A L Q T G S E G L R S L Y N
 241 ACAGTAGCAACCCCTCTATTGTGTGCATCAAAGGATAGAGGTAAAAGACACCAAGGAAGCT
 81 T V A T L Y C V H Q R I E V K D T K E A
 301 TTAGAGAAAAAGAGGAAGAGCAAAAAGTAAGAAAAAGGCACAGCAAGCAGCCGCT
 101 L E K I E E E Q N K S K K K A Q Q A A A
 361 GACACAGGAAACAGTAGCAGCAGCCAGGTCAGCCAAAATTACCCCTATAGTCAGAACATG
 121 D T G N S S S S Q V S Q N Y P I V Q N M
 421 CAGGGGCAAAATGGTACATCAGGCCATATCACCTAGAACTTTAAATGCATGGGTAAAAGTA
 141 Q G Q M V H Q A I S P R T L N A W V K V
 481 GTGGAAGAGAAGGCTTTAGTCCAGAAGTAATACCCATGTTTGCAGCATTATCAGAAGGA
 161 V E E K A F S P E V I P M F A A L S E G
 541 GCCACCCACAAGATTTGAACACCATGCTAAATACAGTGGGGGACATCAGGCAGCCATG
 181 A T P Q D L N T M L N T V G G H Q A A M
 601 CAAATGTTAAAAGAGACCATCAATGAGGAAGCTGCAGAATGGGATAGAATGCATCCAGTG
 201 Q M L K E T I N E E A A E W D R M H P V
 661 CATGCAGGCTATTGCACCAGGCCAGATGAGAGGCCAAGGGGAGCGACATAGCAGGG
 221 H A G P I A P G Q M R E P R G S D I A G
 721 ACTACTAGTACCCTTCAGGAACAGATAAACTGGATGACAGCTAATCCACCTACCCAGTA
 241 T T S T L Q E Q I N W M T A N P P T P V
 781 GGAGAAATCTATAAAAGATGGATAATCCTGGGACTAAATAAAATAGTAAGGATGTATAGT
 261 G E I Y K R W I I L G L N K I V R M Y S
 841 CCTACTAGCATTCTGGACATAAGACAAGGACCAAAAAGAACCCCTTTAGAGATTATGTAGAC
 281 P T S I L D I R Q G P K E P F R D Y V D
 901 CGGTTCTATAAACTCTAAGAGCCGAGCAAGCTTCACAGGAGGTAAAAAATGGATGACA
 301 R F Y K T L R A E Q A S Q E V K N W M T
 961 GAAACCTTGTTAGTCCAGAATGCAAACCCAGATTGTAAGACTATTTAAAAGCGTTGGGA
 321 E T L L V Q N A N P D C K T I L K A L G
 1021 CCAGCAGCTACACTAGAAGAAATGATGACAGCATGTCAGGGAGTAGGAGGCCCGGCCAT
 341 P A A T L E E M M T A C Q G V G G P G H
 1081 AAAGCAAGAGTGTAGCTGAAGCAATGAGCCAAGTAAACAAATTCAGCTACCATATGATG
 361 K A R V L A E A M S Q V T N S A T I M M
 1141 CAGAAAGGCAATTTTAGGAACCAAGAAAAATGTTAAGTGCTCAATTGTGGCAAAGAA
 381 Q K G N F R N Q R K I V K C F N C G K E
 1201 GGGCACATAGCCAGAAATTGCAGGGCCCTAGGAAAAAGGCTGTTGGAAATGTGGAAAG
 401 G H I A R N C R A P R K K G C W K C G K
 1261 GAAGGACACCAATGAAAGATTGACTGAGAGACAAGCTAATTTTTAGGGAAAACTGG

B1. Nucleotide and Amino Acid Sequences of gag clones

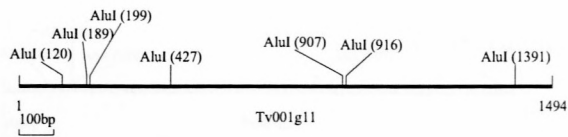
421 E G H Q M K D C T E R Q A N F L G K I W
1321 CCTTCCCACAAGGGAAGGCCAGGGAATTTTCTTCAGAGCAGACCAGAGCCAACAGCCCCA
441 P S H K G R P G N F L Q S R P E P T A P
1381 CCAGAAGAGAGCTTCAGTTTGGGGAGGAGACAACAACCCCTCTCAGAAGCAGGAGACG
461 P E E S F R F G E E T T T P S Q K Q E T
1441 GTAGACAAGGACCTGTATCCTTTAGTTTCCCTCAAATCACTCTTTGGCAACGACCCCTCG
481 V D K D L Y P L V S L K S L F G N D P S
1501 TACAATAA
501 Y N

B2.1 Restriction Digest Maps for *AluI* and *AccI* on the *gag* clone sequences

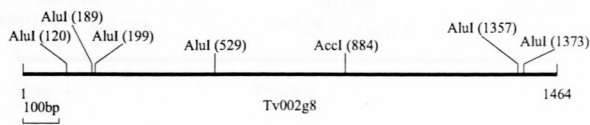
TV001G8



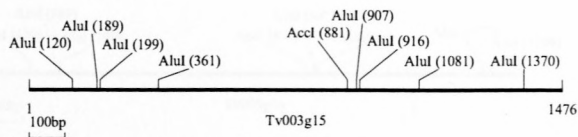
TV001G11



TV002G8

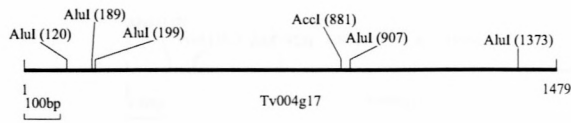


TV003G15

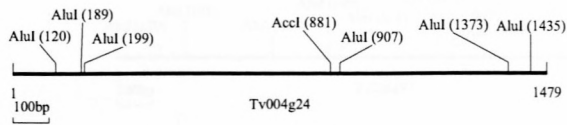


B2.1 Restriction Digest Maps for *AluI* and *AccI* on the *gag* clone sequences

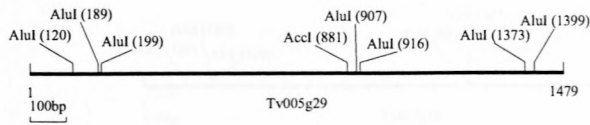
TV004G17



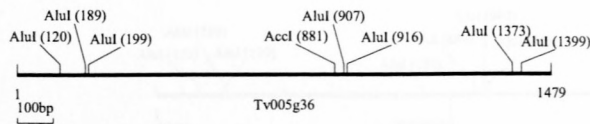
TV004G24



TV005G29



TV005G36

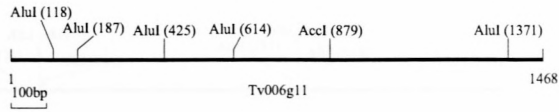


TV005G50

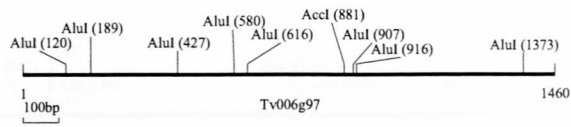


B2.1 Restriction Digest Maps for *AluI* and *AccI* on the *gag* clone sequences

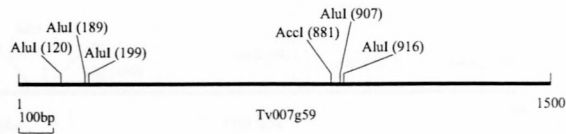
TV006G11



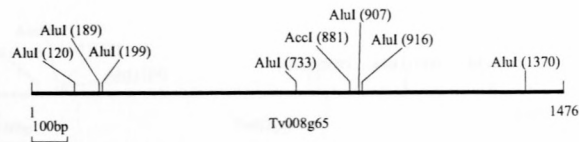
TV006G97



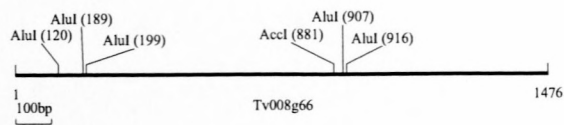
TV007G59



TV008G65

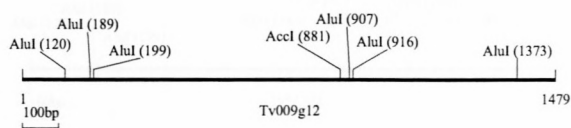


TV008G66

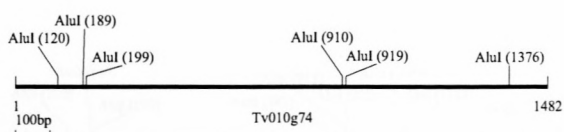


B2.1 Restriction Digest Maps for *AluI* and *AccI* on the *gag* clone sequences

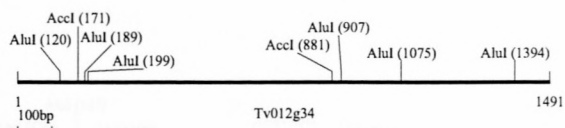
TV009G12



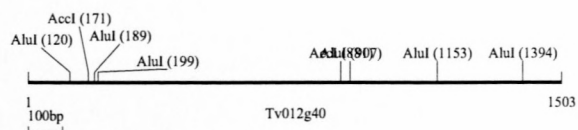
TV010G74



TV012G34

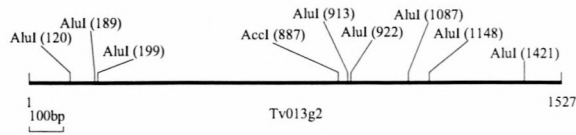


TV012G40

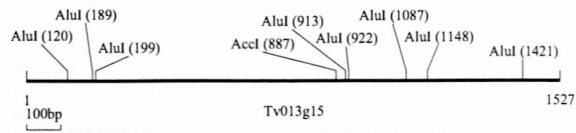


B2.1 Restriction Digest Maps for *AluI* and *AccI* on the *gag* clone sequences

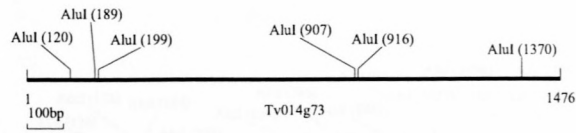
TV013G2



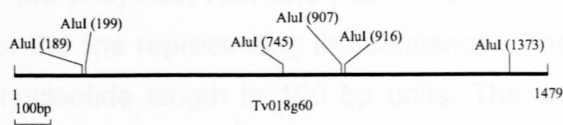
TV013G15



TV014G73

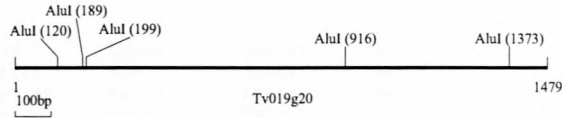


TV018G60

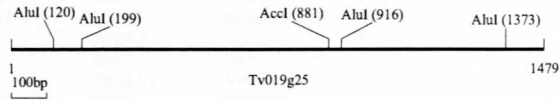


B2.1 Restriction Digest Maps for *AluI* and *AccI* on the *gag* clone sequences

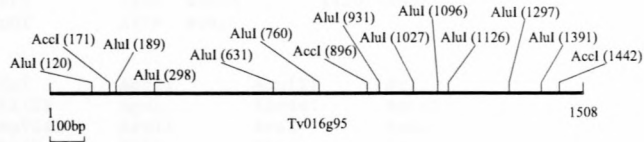
TV019G20



TV019G25



TV016G95



Appendix B2.1 A restriction enzyme recognition site map of the *gag* clone sequences for the enzymes, *AluI* and *AccI*. The names of the maps are indicated under the line representing the sequence. The horizontal scale indicates the nucleotide length in 100 bp units. The lines on top of the sequence line indicate the recognition sites, and the position of the recognition site is shown in brackets.

B2.2 Unique cutting and non-cutting Restriction Digest Enzymes for the gag clones

Restriction analysis on TV001G8

Enzymes with >1 sites are not shown

16 sites found

List by Site Order

20	SspI	425	HphI	649	SphI	1206	ApaI
206	SapI	453	NsiI	655	AlwNI	1207	AvrII
261	BspI407I	478	XmnI	1054	PpuMI	1244	BanI
370	DrdI	621	PstI	1206	BanII	1400	TaqI

Non Cut Enzymes

AatII	Acc65I	AccI	AccII	AccIII	AclI
AcyI	AflIII	AflIII	AgeI	Alw44I	ApaBI
ApaLI	AscI	Asp718I	AsuII	AvaI	BalI
BamHI	BbeI	BbvII	BclI	BglI	BglII
Bpu1102I	BsaHI	BsaOI	Bsc91I	BsiI	BspHI
BspMI	BspMII	BssHII	BstD102I	BstEII	BstXI
Bsu36I	Cfr10I	CfrI	Clal	Csp45I	CspI
CvnI	DraIII	EagI	Ecl136II	Eco31I	Eco47III
Eco52I	Eco56I	Eco72I	EcoHI	EcoICRI	EcoNI
EcoRI	EcoRV	EheI	EspI	FnuDII	FseI
HaeII	HgaI	HgiAI	HhaI	HindII	HindIII
HinfI	HinPII	HpaI	HpaII	I-PpoI	KpnI
MaeII	MfeI	Mlu113I	MluI	MscI	MspAlI
MspI	MstI	MstII	NaeI	NarI	NcoI
NdeI	NheI	NotI	NruI	NspBII	PacI
PflMI	PinAI	PleI	PmaCI	PmeI	PvuI
PvuII	SacI	SacII	SalI	SauI	ScaI
SciI	SfaNI	SfiI	SgrAI	SmaI	SnaBI
SplI	SpoI	SrfI	SstI	SstII	StuI
SunI	SwaI	ThaI	Tth111I	VspI	XbaI
XcmI	XhoI	XhoII	XmaI	XmaIII	XorII

Restriction analysis on TV001G11

Enzymes with >1 sites are not shown

15 sites found

List by Site Order

20	SspI	453	NsiI	655	AlwNI	1207	AvrII
206	SapI	478	XmnI	1054	PpuMI	1244	BanI
261	BspI407I	621	PstI	1206	BanII	1400	TaqI
334	HindII	649	SphI	1206	ApaI		

Non Cut Enzymes

AatII	Acc65I	AccI	AccII	AccIII	AclI
AcyI	AflIII	AflIII	AgeI	Alw44I	ApaBI
ApaLI	AscI	Asp718I	AsuII	AvaI	BalI
BamHI	BbeI	BbvII	BclI	BglI	BglII
Bpu1102I	BsaHI	BsaOI	Bsc91I	BsiI	BspHI
BspMI	BspMII	BssHII	BstD102I	BstEII	BstXI
Bsu36I	Cfr10I	CfrI	Clal	Csp45I	CspI
CvnI	DraIII	DrdI	EagI	Ecl136II	Eco31I
Eco47III	Eco52I	Eco56I	Eco72I	EcoHI	EcoICRI
EcoNI	EcoRI	EcoRV	EheI	EspI	FnuDII
FseI	HaeII	HgaI	HgiAI	HhaI	HindIII
HinfI	HinPII	HpaI	HpaII	I-PpoI	KpnI
MaeII	MfeI	Mlu113I	MluI	MscI	MspAlI
MspI	MstI	MstII	NaeI	NarI	NcoI
NdeI	NheI	NotI	NruI	NspBII	PacI
PflMI	PinAI	PleI	PmaCI	PmeI	PvuI
PvuII	SacI	SacII	SalI	SauI	ScaI
SciI	SfaNI	SfiI	SgrAI	SmaI	SnaBI
SplI	SpoI	SrfI	SstI	SstII	StuI
SunI	SwaI	ThaI	Tth111I	VspI	XbaI
XcmI	XhoI	XhoII	XmaI	XmaIII	XorII

Restriction analysis on TV002G8

Enzymes with >1 sites are not shown

29 sites found

List by Site Order

206	SapI	481	XmnI	884	AccI	1209	SduI
-----	------	-----	------	-----	------	------	------

B2.2 Unique cutting and non-cutting Restriction Digest Enzymes for the gag clones

226	MboI	551	RleAI	1057	PpuMI	1209	Bsp1286I
226	Sau3AI	624	PstI	1062	CfrI	1209	BanII
228	DpnI	652	SphI	1064	BalI	1209	ApaI
261	Bsp1407I	712	SpeI	1064	MscI	1210	AvrII
338	PleI	763	HpaII	1109	SfaNI	1382	TaqI
428	HphI	763	MspI	1167	MfeI	1389	BstD102I
456	NsiI						

Non Cut Enzymes

AatII	Acc65I	AccII	AccIII	AclI	AcyI
AflII	AflIII	AgeI	Alw44I	AlwNI	ApaBI
ApaLI	AscI	Asp718I	AsuII	AvaI	BamHI
BanI	BbeI	BbvII	BclI	BglI	BglII
Bpu1102I	BsaHI	BsaOI	Bsc91I	BsiI	BsmI
BspHI	BspMI	BspMII	BssHII	BstEII	BstXI
Bsu36I	Cfr10I	Clal	Csp45I	CspI	CvnI
DraIII	DrdI	EagI	Ecl136II	Eco31I	Eco47III
Eco52I	Eco56I	Eco72I	EcoHI	EcoICRI	EcoNI
EcoRI	EcoRV	EheI	EspI	FnuDII	FseI
HaeII	HgaI	HgiAI	HhaI	HindII	HindIII
HinPII	HpaI	I-PpoI	KpnI	MaeII	Mlu113I
MluI	MspAI	MstI	MstII	NaeI	NarI
NcoI	NdeI	NheI	NotI	NruI	NspBII
PacI	PflMI	PinAI	PmaCI	PmeI	PvuI
PvuII	SacI	SacII	SalI	SauI	ScaI
SciI	SfiI	SgrAI	SmaI	SnaBI	SplI
SpoI	SrfI	SspI	SstI	SstII	StuI
SunI	SwaI	ThaI	Tth111I	VspI	XbaI
XcmI	XhoI	XhoII	XmaI	XmaIII	XorII

Restriction analysis on TV003G15

Enzymes with >1 sites are not shown

34 sites found

List by Site Order

20	SspI	361	NspBII	621	PstI	1141	XbaI
206	SapI	370	DrdI	709	SpeI	1206	BanII
226	MboI	418	ApaLI	881	AccI	1206	ApaI
226	Sau3AI	418	ApaBI	1054	PpuMI	1321	RleAI
228	DpnI	418	Alw44I	1059	CfrI	1379	TaqI
261	Bsp1407I	422	HgiAI	1061	BalI	1396	ThaI
361	PvuII	425	HphI	1061	MscI	1396	FnuDII
361	AlwNI	436	AvrII	1106	SfaNI	1396	AccII
361	MspAI	453	NsiI				

Non Cut Enzymes

AatII	Acc65I	AccIII	AclI	AcyI	AflII
AflIII	AgeI	AscI	Asp718I	AsuII	AvaI
BamHI	BanI	BbeI	BbvII	BclI	BglI
BglII	Bpu1102I	BsaHI	BsaOI	Bsc91I	BsiI
BspHI	BspMI	BspMII	BssHII	BstD102I	BstEII
BstXI	Bsu36I	Cfr10I	Clal	Csp45I	CspI
CvnI	DraIII	EagI	Ecl136II	Eco31I	Eco47III
Eco52I	Eco56I	Eco72I	EcoHI	EcoICRI	EcoNI
EcoRI	EcoRV	EheI	EspI	FseI	HaeII
HgaI	HhaI	HindII	HindIII	HinfI	HinPII
HpaI	HpaII	I-PpoI	KpnI	MaeII	MfeI
Mlu113I	MluI	MspI	MstI	MstII	NaeI
NarI	NcoI	NdeI	NheI	NotI	NruI
PacI	PflMI	PinAI	PleI	PmaCI	PmeI
PvuI	SacI	SacII	SalI	SauI	ScaI
SciI	SfiI	SgrAI	SmaI	SnaBI	SphI
SplI	SpoI	SrfI	SstI	SstII	StuI
SunI	SwaI	Tth111I	VspI	XcmI	XhoI
XhoII	XmaI	XmaIII	XmNI	XorII	

Restriction analysis on TV004G17

Enzymes with >1 sites are not shown

26 sites found

List by Site Order

20	SspI	425	HphI	884	Cfr10I	1109	SfaNI
206	SapI	453	NsiI	884	PinAI	1209	ApaI
226	MboI	453	BsmI	884	AgeI	1209	BanII
226	Sau3AI	621	PstI	885	HpaII	1210	AvrII

B2.2 Unique cutting and non-cutting Restriction Digest Enzymes for the gag clones

228	DpnI	655	AlwNI	885	MspI	1324	RleAI
370	DrdI	709	SpeI	1054	PpuMI	1382	TaqI
399	HinfI	881	AccI				

Non Cut Enzymes

AatII	Acc65I	AccII	AccIII	AclI	AcyI
AflII	AflIII	Alw44I	ApaBI	ApaLI	AscI
Asp718I	AsuII	AvaI	BalI	BamHI	BanI
BbeI	BbvII	BclI	BglI	BglII	Bpull102I
BsaHI	BsaOI	Bsc91I	BsiI	Bsp1407I	BspHI
BspMI	BspMII	BssHII	BstD102I	BstEII	BstXI
Bsu36I	CfrI	ClalI	Csp45I	CspI	CvnI
DraIII	EagI	Ecl136II	Eco31I	Eco47III	Eco52I
Eco56I	Eco72I	EcoHI	EcoICRI	EcoNI	EcoRI
EcoRV	EheI	EspI	FnuDII	FseI	HaeII
HgaI	HgiAI	HhaI	HindIII	HindIII	HinPII
HpaI	I-PpoI	KpnI	MaeII	MfeI	Mlu113I
MluI	MscI	MspAlI	MstI	MstII	NaeI
NarI	NcoI	NdeI	NheI	NotI	NruI
NspBII	PacI	PflMI	PleI	PmaCI	PmeI
PvuI	PvuII	SacI	SacII	SalI	SauI
ScaI	SciI	SfiI	SgrAI	SmaI	SnaBI
SphI	SplI	SpoI	SrfI	SstI	SstII
StuI	SunI	SwaI	ThaI	Tth111I	VspI
XbaI	XcmI	XhoI	XhoII	XmaI	XmaIII
XmnI	XorII				

Restriction analysis on TV004G24

Enzymes with >1 sites are not shown

27 sites found

List by Site Order

20	SspI	453	BsmI	884	PinAI	1209	BanII
206	SapI	453	NsiI	884	Cfr10I	1210	AvrII
226	MboI	621	PstI	885	HpaII	1324	RleAI
226	Sau3AI	655	AlwNI	885	MspI	1352	HindII
228	DpnI	709	SpeI	1054	PpuMI	1356	PleI
370	DrdI	881	AccI	1109	SfaNI	1382	TaqI
425	HphI	884	AgeI	1209	ApaI		

Non Cut Enzymes

AatII	Acc65I	AccII	AccIII	AclI	AcyI
AflII	AflIII	Alw44I	ApaBI	ApaLI	AscI
Asp718I	AsuII	AvaI	BalI	BamHI	BanI
BbeI	BbvII	BclI	BglI	BglII	Bpull102I
BsaHI	BsaOI	Bsc91I	BsiI	Bsp1407I	BspHI
BspMI	BspMII	BssHII	BstD102I	BstEII	BstXI
Bsu36I	CfrI	ClalI	Csp45I	CspI	CvnI
DraIII	EagI	Ecl136II	Eco31I	Eco47III	Eco52I
Eco56I	Eco72I	EcoHI	EcoICRI	EcoNI	EcoRI
EcoRV	EheI	EspI	FnuDII	FseI	HaeII
HgaI	HgiAI	HhaI	HindIII	HinPII	HpaI
I-PpoI	KpnI	MaeII	MfeI	Mlu113I	MluI
MscI	MspAlI	MstI	MstII	NaeI	NarI
NcoI	NdeI	NheI	NotI	NruI	NspBII
PacI	PflMI	PmaCI	PmeI	PvuI	PvuII
SacI	SacII	SalI	SauI	ScaI	SciI
SfiI	SgrAI	SmaI	SnaBI	SphI	SplI
SpoI	SrfI	SstI	SstII	StuI	SunI
SwaI	ThaI	Tth111I	VspI	XbaI	XcmI
XhoI	XhoII	XmaI	XmaIII	XmnI	XorII

Restriction analysis on TV005G29

Enzymes with >1 sites are not shown

27 sites found

List by Site Order

206	SapI	425	HphI	884	AgeI	1059	CfrI
223	AflII	453	BsmI	885	MspI	1061	MscI
260	Alw44I	621	PstI	885	HpaII	1061	BalI
260	ApaLI	709	SpeI	956	MboI	1106	ApaBI
264	HgiAI	881	AccI	956	Sau3AI	1209	BanII
370	DrdI	884	Cfr10I	958	DpnI	1382	TaqI
399	HinfI	884	PinAI	1054	PpuMI		

B2.2 Unique cutting and non-cutting Restriction Digest Enzymes for the gag clones

Non Cut Enzymes

AatII	Acc65I	AccII	AccIII	AclI	AcyI
AflIII	AlwNI	ApaI	AscI	Asp718I	AsuII
AvaI	AvrII	BamHI	BanI	BbeI	BbvII
BclI	BglI	BglII	Bpu1102I	BsaHI	BsaOI
Bsc91I	BsiI	Bsp1407I	BspHI	BspMI	BspMII
BssHII	BstD102I	BstEII	BstXI	Bsu36I	Clal
Csp45I	CspI	CvnI	DraIII	EagI	Ecl136II
Eco31I	Eco47III	Eco52I	Eco56I	Eco72I	EcoHI
EcoICRI	EcoNI	EcoRI	EcoRV	EheI	EspI
FnuDII	FseI	HaeII	HgaI	HhaI	HindII
HindIII	HinPII	HpaI	I-PpoI	KpnI	MaeII
MfeI	Mlu113I	MluI	MspAlI	MstI	MstII
NaeI	NarI	NcoI	NdeI	NheI	NotI
NruI	NspBII	PacI	PflMI	PleI	PmaCI
PmeI	PvuI	PvuII	SacI	SacII	SalI
SauI	ScaI	SciI	SfiI	SgrAI	SmaI
SnaBI	SphI	SplI	SpoI	SrfI	SspI
SstI	SstII	StuI	SunI	SwaI	ThaI
Tth111I	VspI	XbaI	XcmI	XhoI	XhoII
XmaI	XmaIII	XmnI	XorII		

Restriction analysis on TV005G36

Enzymes with >1 sites are not shown

22 sites found

List by Site Order

206	SapI	453	BsmI	884	AgeI	1061	BalI
223	AflII	621	PstI	885	HpaII	1061	MscI
261	Bsp1407I	709	SpeI	885	MspI	1106	ApaBI
370	DrdI	881	AccI	1054	PpuMI	1209	BanII
399	HinfI	884	Cfr10I	1059	CfrI	1382	TaqI
425	HphI	884	PinAI				

Non Cut Enzymes

AatII	Acc65I	AccII	AccIII	AclI	AcyI
AflIII	Alw44I	AlwNI	ApaI	ApaLI	AscI
Asp718I	AsuII	AvaI	AvrII	BamHI	BanI
BbeI	BbvII	BclI	BglI	BglII	Bpu1102I
BsaHI	BsaOI	Bsc91I	BsiI	BspHI	BspMI
BspMII	BssHII	BstD102I	BstEII	BstXI	Bsu36I
Clal	Csp45I	CspI	CvnI	DpnI	DraIII
EagI	Ecl136II	Eco31I	Eco47III	Eco52I	Eco56I
Eco72I	EcoHI	EcoICRI	EcoNI	EcoRI	EcoRV
EheI	EspI	FnuDII	FseI	HaeII	HgaI
HgiAI	HhaI	HindII	HindIII	HinPII	HpaI
I-PpoI	KpnI	MaeII	MboI	MfeI	Mlu113I
MluI	MspAlI	MstI	MstII	NaeI	NarI
NcoI	NdeI	NheI	NotI	NruI	NspBII
PacI	PflMI	PleI	PmaCI	PmeI	PvuI
PvuII	SacI	SacII	SalI	Sau3AI	SauI
ScaI	SciI	SfiI	SgrAI	SmaI	SnaBI
SphI	SplI	SpoI	SrfI	SspI	SstI
SstII	StuI	SunI	SwaI	ThaI	Tth111I
VspI	XbaI	XcmI	XhoI	XhoII	XmaI
XmaIII	XmnI	XorII			

Restriction analysis on TV006G11

Enzymes with >1 sites are not shown

28 sites found

List by Site Order

18	SspI	440	AclI	879	AccI	1052	PpuMI
224	MboI	440	MaeII	882	PinAI	1207	ApaI
224	Sau3AI	476	XmnI	882	AgeI	1208	AvrII
226	DpnI	546	RleAI	882	Cfr10I	1369	HindIII
259	Bsp1407I	647	SphI	883	MspI	1380	TaqI
403	BsmI	653	AlwNI	883	HpaII	1397	HinPII
423	HphI	707	SpeI	911	BglI	1399	HhaI

Non Cut Enzymes

AatII	Acc65I	AccII	AccIII	AcyI	AflII
AflIII	Alw44I	ApaBI	ApaLI	AscI	Asp718I
AsuII	AvaI	BalI	BamHI	BanI	BbeI
BbvII	BclI	BglII	Bpu1102I	BsaHI	BsaOI

B2.2 Unique cutting and non-cutting Restriction Digest Enzymes for the gag clones

Bsc91I	BsiI	BspHI	BspMI	BspMII	BssHII
BstD102I	BstEII	BstXI	Bsu36I	CfrI	Clal
Csp45I	CspI	CvnI	DraIII	DrdI	EagI
Ecl136II	Eco31I	Eco47III	Eco52I	Eco56I	Eco72I
EcoHI	EcoICRI	EcoNI	EcoRI	EcoRV	EheI
EspI	FnuDII	FseI	HaeII	HgaI	HgiAI
HindII	HinfI	HpaI	I-PpoI	KpnI	MfeI
Mlu113I	MluI	MscI	MspAlI	MstI	MstII
NaeI	NarI	NcoI	NdeI	NheI	NotI
NruI	NspBII	PacI	PflMI	PleI	PmaCI
PmeI	PvuI	PvuII	SacI	SacII	SalI
SapI	SauI	ScaI	SciI	SfiI	SgrAI
SmaI	SnaBI	SplI	SpoI	SrfI	SstI
SstII	StuI	SunI	SwaI	ThaI	Tth111I
VspI	XbaI	XcmI	XhoI	XhoII	XmaI
XmaIII	XorII				

Restriction analysis on TV006G97

Enzymes with >1 sites are not shown

29 sites found

List by Site Order

17	HindII	478	XmnI	884	Cfr10I	1210	AvrII
226	MboI	649	SphI	884	PinAI	1371	HindIII
226	Sau3AI	654	StuI	885	HpaII	1382	TaqI
228	DpnI	691	MaeIII	885	MspI	1393	BbvII
261	Bsp1407I	709	SpeI	1054	DraII	1393	Bsc91I
405	BsmI	881	AccI	1054	PpuMI	1399	HinPII
442	MaeII	884	AgeI	1057	PssI	1401	HhaI
442	AclI						

Non Cut Enzymes

AatII	Acc65I	AccII	AccIII	AcyI	AflII
AflIII	Alw44I	AlwNI	ApaBI	ApaI	ApaLI
AscI	Asp718I	AsuII	AvaI	BalI	BamHI
BanI	BbeI	BclI	BglI	BglII	Bpu1102I
BsaHI	BsaOI	BsiI	BspHI	BspMI	BspMII
BssHII	BstD102I	BstEII	BstXI	Bsu36I	CfrI
Clal	Csp45I	CspI	CvnI	DraIII	DrdI
EagI	Eam1105I	Ecl136II	Eco31I	Eco47III	Eco52I
Eco56I	Eco72I	EcoHI	EcoICRI	EcoNI	EcoRI
EcoRV	EheI	EspI	FnuDII	FseI	HaeII
HgaI	HgiAI	HinfI	HpaI	HphI	I-PpoI
KpnI	MfeI	Mlu113I	MluI	MscI	MspAlI
MstI	MstII	NaeI	NarI	NcoI	NdeI
NheI	NotI	NruI	NspBII	PacI	PflMI
PleI	PmaCI	PmeI	PvuI	PvuII	SacI
SacII	SalI	SapI	SauI	ScaI	SciI
SfiI	SgrAI	SmaI	SnaBI	SplI	SpoI
SrfI	SspI	SstI	SstII	SunI	SwaI
ThaI	Tth111I	VspI	XbaI	XcmI	XhoI
XhoII	XmaI	XmaIII	XorII		

Restriction analysis on TV007G59

Enzymes with >1 sites are not shown

26 sites found

List by Site Order

20	SspI	478	XmnI	884	PinAI	1210	AvrII
206	SapI	548	RleAI	884	AgeI	1396	Bsu36I
261	Bsp1407I	621	PstI	1054	PpuMI	1396	CvnI
399	HinfI	630	Clal	1102	XcmI	1396	SauI
427	StuI	771	EcoRV	1187	Bsp1286I	1396	MstII
453	BsmI	881	AccI	1187	SduI	1420	BstD102I
453	NsiI	884	Cfr10I				

Non Cut Enzymes

AatII	Acc65I	AccII	AccIII	AclI	AcyI
AflII	AflIII	Alw44I	AlwNI	ApaBI	ApaI
ApaLI	AscI	Asp718I	AsuII	AvaI	BalI
BamHI	BanI	BanII	BbeI	BbvII	BclI
BglI	BglII	Bpu1102I	BsaHI	BsaOI	Bsc91I
BsiI	BspHI	BspMI	BspMII	BssHII	BstEII
BstXI	CfrI	Csp45I	CspI	DraIII	DrdI
EagI	Ecl136II	Eco31I	Eco47III	Eco52I	Eco56I

B2.2 Unique cutting and non-cutting Restriction Digest Enzymes for the gag clones

Eco72I	EcoHI	EcoICRI	EcoNI	EcoRI	EheI
EspI	FnuDII	FseI	HaeII	HgaI	HgiAI
HhaI	HindII	HindIII	HinPII	HpaI	I-PpoI
KpnI	MaeII	MfeI	MluI13I	MluI	MscI
MspAII	MstI	NaeI	NarI	NcoI	NdeI
NheI	NotI	NruI	NspBII	PacI	PflMI
PleI	PmaCI	PmeI	PvuI	PvuII	SacI
SacII	SalI	ScaI	SciI	SfiI	SgrAI
SmaI	SnaBI	SpeI	SphI	SplI	SpoI
SrfI	SstI	SstII	SunI	SwaI	ThaI
Tth111I	VspI	XbaI	XhoI	XhoII	XmaI
XmaIII	XorII				

Restriction analysis on TV008G65

Enzymes with >1 sites are not shown

26 sites found

List by Site Order

206	SapI	682	AvrII	1059	CfrI	1396	BstD102I
261	Bsp1407I	709	SpeI	1061	MscI	1408	HinfI
370	DrdI	881	AccI	1061	BalI	1416	PleI
425	HphI	884	Cfr10I	1077	BstXI	1441	Sau3AI
453	NsiI	884	AgeI	1206	ApaI	1441	MboI
453	BsmI	884	PinAI	1321	RleAI	1443	DpnI
621	PstI	1054	PpuMI				

Non Cut Enzymes

AatII	Acc65I	AccII	AccIII	AclI	AcyI
AflII	AflIII	Alw44I	AlwNI	ApaBI	ApaLI
AscI	Asp718I	AsuII	AvaI	BamHI	BanI
BbeI	BbvII	BclI	BglI	BglII	Bpu1102I
BsaHI	BsaOI	Bsc91I	BsiI	BspHI	BspMI
BspMII	BssHII	BstEII	Bsu36I	Clal	Csp45I
CspI	CvnI	DraIII	EagI	Ecl136II	Eco31I
Eco47III	Eco52I	Eco56I	Eco72I	EcoHI	EcoICRI
EcoNI	EcoRI	EcoRV	EheI	EspI	FnuDII
FseI	HaeII	HgaI	HgiAI	HhaI	HindII
HindIII	HinPII	HpaI	I-PpoI	KpnI	MaeII
MfeI	Mlu113I	MluI	MspAII	MstI	MstII
NaeI	NarI	NcoI	NdeI	NheI	NotI
NruI	NspBII	PacI	PflMI	PmaCI	PmeI
PvuI	PvuII	SacI	SacII	SalI	SauI
ScaI	SciI	SfaNI	SfiI	SgrAI	SmaI
SnaBI	SphI	SplI	SpoI	SrfI	SspI
SstI	SstII	StuI	SunI	SwaI	ThaI
Tth111I	VspI	XbaI	XcmI	XhoI	XhoII
XmaI	XmaIII	XmnI	XorII		

Restriction analysis on TV008G66

Enzymes with >1 sites are not shown

29 sites found

List by Site Order

206	SapI	656	ApaBI	885	HpaII	1206	ApaI
261	Bsp1407I	682	AvrII	885	MspI	1321	RleAI
370	DrdI	709	SpeI	1054	PpuMI	1396	BstD102I
425	HphI	881	AccI	1059	CfrI	1416	PleI
453	NsiI	884	AgeI	1061	BalI	1441	MboI
453	BsmI	884	Cfr10I	1061	MscI	1441	Sau3AI
621	PstI	884	PinAI	1077	BstXI	1443	DpnI
655	AlwNI						

Non Cut Enzymes

AatII	Acc65I	AccII	AccIII	AclI	AcyI
AflII	AflIII	Alw44I	ApaLI	AscI	Asp718I
AsuII	AvaI	BamHI	BanI	BbeI	BbvII
BclI	BglI	BglII	Bpu1102I	BsaHI	BsaOI
Bsc91I	BsiI	BspHI	BspMI	BspMII	BssHII
BstEII	Bsu36I	Clal	Csp45I	CspI	CvnI
DraIII	EagI	Ecl136II	Eco31I	Eco47III	Eco52I
Eco56I	Eco72I	EcoHI	EcoICRI	EcoNI	EcoRI
EcoRV	EheI	EspI	FnuDII	FseI	HaeII
HgaI	HgiAI	HhaI	HindII	HindIII	HinPII
HpaI	I-PpoI	KpnI	MaeII	MfeI	Mlu113I
MluI	MspAII	MstI	MstII	NaeI	NarI

B2.2 Unique cutting and non-cutting Restriction Digest Enzymes for the gag clones

NcoI	NdeI	NheI	NotI	NruI	NspBII
PacI	PflMI	PmaCI	PmeI	PvuI	PvuII
SacI	SacII	SalI	SauI	ScaI	SciI
SfaNI	SfiI	SgrAI	SmaI	SnaBI	SphI
SplI	SpoI	SrfI	SspI	SstI	SstII
StuI	SunI	SwaI	ThaI	Tth111I	VspI
XbaI	XcmI	XhoI	XhoII	XmaI	XmaIII
XmnI	XorII				

Restriction analysis on TV009G12

Enzymes with >1 sites are not shown

26 sites found

List by Site Order

261	BspI407I	709	SpeI	885	MspI	1109	SfaNI
425	HphI	745	BstEII	1040	NspI	1209	BanII
453	NsiI	881	AccI	1054	PpuMI	1209	ApaI
453	BsmI	884	Cfr10I	1059	CfrI	1210	AvrII
478	XmnI	884	AgeI	1061	MscI	1345	AvaI
537	XhoII	884	PinAI	1061	BalI	1419	PleI
537	BglII	885	HpaII				

Non Cut Enzymes

AatII	Acc65I	AccII	AccIII	AclI	AcyI
AflII	AflIII	Alw44I	AlwNI	ApaBI	ApaLI
AscI	Asp718I	AsuII	BamHI	BanI	BbeI
BbvII	BclI	BglI	Bpu1102I	BsaHI	BsaOI
Bsc91I	BsiI	BspHI	BspMI	BspMII	BssHII
BstD102I	BstXI	Bsu36I	ClaI	Csp45I	CspI
CvnI	DraIII	DrdI	EagI	Ecl136II	Eco31I
Eco47III	Eco52I	Eco56I	Eco72I	EcoHI	EcoICRI
EcoNI	EcoRI	EcoRV	EheI	EspI	FnuDII
FseI	HaeII	HgaI	HgiAI	HhaI	HindII
HindIII	HinPII	HpaI	I-PpoI	KpnI	MaeII
MfeI	Mlu113I	MluI	MspAlI	MstI	MstII
NaeI	NarI	NcoI	NdeI	NheI	NotI
NruI	NspBII	PacI	PflMI	PmaCI	PmeI
PstI	PvuI	PvuII	SacI	SacII	SalI
SapI	SauI	ScaI	SciI	SfiI	SgrAI
SmaI	SnaBI	SphI	SplI	SpoI	SrfI
SstI	SstII	StuI	SunI	SwaI	ThaI
Tth111I	VspI	XbaI	XcmI	XhoI	XmaI
XmaIII	XorII				

Restriction analysis on TV010G74

Enzymes with >1 sites are not shown

22 sites found

List by Site Order

20	SspI	624	PstI	1103	BglI	1212	ApaI
202	AflII	658	AlwNI	1112	SfaNI	1213	AvrII
261	BspI407I	681	XhoII	1212	SduI	1314	AvaI
298	StuI	712	SpeI	1212	Bspl286I	1385	TaqI
456	NsiI	1057	PpuMI	1212	BanII	1402	BstD102I
481	XmnI	1084	PleI				

Non Cut Enzymes

AatII	Acc65I	AccI	AccII	AccIII	AclI
AcyI	AflIII	AgeI	Alw44I	ApaBI	ApaLI
AscI	Asp718I	AsuII	BalI	BamHI	BanI
BbeI	BbvII	BclI	BglII	Bpu1102I	BsaHI
BsaOI	Bsc91I	BsiI	BsmI	BspHI	BspMI
BspMII	BssHII	BstEII	BstXI	Bsu36I	Cfr10I
CfrI	ClaI	Csp45I	CspI	CvnI	DraIII
DrdI	EagI	Ecl136II	Eco31I	Eco47III	Eco52I
Eco56I	Eco72I	EcoHI	EcoICRI	EcoNI	EcoRI
EcoRV	EheI	EspI	FnuDII	FseI	HaeII
HgaI	HgiAI	HhaI	HindII	HindIII	HinPII
HpaI	HpaII	I-PpoI	KpnI	MaeII	MfeI
Mlu113I	MluI	MscI	MspAlI	MspI	MstI
MstII	NaeI	NarI	NcoI	NdeI	NheI
NotI	NruI	NspBII	PacI	PflMI	PinAI
PmaCI	PmeI	PvuI	PvuII	RleAI	SacI
SacII	SalI	SapI	SauI	ScaI	SciI
SfiI	SgrAI	SmaI	SnaBI	SphI	SplI

B2.2 Unique cutting and non-cutting Restriction Digest Enzymes for the gag clones

SpoI	SrfI	SstI	SstII	SunI	SwaI
ThaI	Tth111I	VspI	XbaI	XcmI	XhoI
XmaI	XmaIII	XorII			

Restriction analysis on TV012G34

Enzymes with >1 sites are not shown

28 sites found

List by Site Order

20	SspI	370	DrdI	884	AgeI	1152	ScaI
138	HpaI	425	HphI	884	Cfr10I	1209	ApaI
138	HindII	453	BsmI	885	HpaII	1209	BanII
206	SapI	453	NsiI	885	MspI	1210	AvrII
226	MboI	655	AlwNI	1054	PpuMI	1403	TaqI
226	Sau3AI	709	SpeI	1106	ApaBI	1404	PleI
228	DpnI	884	PinAI	1109	SfaNI	1410	HinfI

Non Cut Enzymes

AatII	Acc65I	AccII	AccIII	AclI	AcyI
AflII	AflIII	Alw44I	ApaLI	AscI	Asp718I
AsuII	AvaI	BalI	BamHI	BanI	BbeI
BbvII	BclI	BglI	BglII	Bpu1102I	BsaHI
BsaOI	Bsc91I	BsiI	BspHI	BspMI	BspMII
BssHII	BstD102I	BstEII	BstXI	Bsu36I	CfrI
Clal	Csp45I	CspI	CvnI	DraIII	EagI
Ecl136II	Eco31I	Eco47III	Eco52I	Eco56I	Eco72I
EcoHI	EcoICRI	EcoNI	EcoRI	EcoRV	EheI
EspI	FnuDII	FseI	HaeII	HgaI	HgiAI
HhaI	HindIII	HinPII	I-PpoI	KpnI	MaeII
MfeI	Mlu113I	MluI	MscI	MspAII	MstI
MstII	NaeI	NarI	NcoI	NdeI	NheI
NotI	NruI	NspBII	PacI	PflMI	PmaCI
PmeI	PstI	PvuI	PvuII	SacI	SacII
SalI	SauI	SciI	SfiI	SgrAI	SmaI
SnaBI	SphI	SplI	SpoI	SrfI	SstI
SstII	StuI	SunI	SwaI	ThaI	Tth111I
VspI	XbaI	XcmI	XhoI	XhoII	XmaI
XmaIII	XmnI	XorII			

Restriction analysis on TV012G40

Enzymes with >1 sites are not shown

24 sites found

List by Site Order

20	SspI	370	DrdI	884	Cfr10I	1209	ApaI
138	HpaI	425	HphI	884	PinAI	1209	BanII
138	HindII	453	BsmI	884	AgeI	1210	AvrII
226	Sau3AI	453	NsiI	885	MspI	1403	TaqI
226	MboI	655	AlwNI	885	HpaII	1404	PleI
228	DpnI	709	SpeI	1054	PpuMI	1410	HinfI

Non Cut Enzymes

AatII	Acc65I	AccII	AccIII	AclI	AcyI
AflII	AflIII	Alw44I	ApaBI	ApaLI	AscI
Asp718I	AsuII	AvaI	BalI	BamHI	BanI
BbeI	BbvII	BclI	BglI	BglII	Bpu1102I
BsaHI	BsaOI	Bsc91I	BsiI	BspHI	BspMI
BspMII	BssHII	BstD102I	BstEII	BstXI	Bsu36I
CfrI	Clal	Csp45I	CspI	CvnI	DraIII
EagI	Ecl136II	Eco31I	Eco47III	Eco52I	Eco56I
Eco72I	EcoHI	EcoICRI	EcoNI	EcoRI	EcoRV
EheI	EspI	FnuDII	FseI	HaeII	HgaI
HgiAI	HhaI	HindIII	HinPII	I-PpoI	KpnI
MaeII	MfeI	Mlu113I	MluI	MscI	MspAII
MstI	MstII	NaeI	NarI	NcoI	NdeI
NheI	NotI	NruI	NspBII	PacI	PflMI
PmaCI	PmeI	PstI	PvuI	PvuII	SacI
SacII	SalI	SauI	ScaI	SciI	SfaNI
SfiI	SgrAI	SmaI	SnaBI	SphI	SplI
SpoI	SrfI	SstI	SstII	StuI	SunI
SwaI	ThaI	Tth111I	VspI	XbaI	XcmI
XhoI	XhoII	XmaI	XmaIII	XmnI	XorII

B2.2 Unique cutting and non-cutting Restriction Digest Enzymes for the gag clones

Restriction analysis on TV013G2

Enzymes with >1 sites are not shown

28 sites found

List by Site Order

20	SspI	627	PstI	890	PinAI	1215	Bsp1286I
206	SapI	666	BanI	891	HpaII	1215	SduI
261	Bsp1407I	676	MfeI	891	MspI	1215	ApaI
376	DrdI	715	SpeI	906	AflIII	1215	BanII
405	HinfI	887	AccI	1046	NspI	1216	AvrII
431	HphI	890	AgeI	1060	PpuMI	1430	TaqI
459	BsmI	890	Cfr10I	1064	Tth111I	1473	BamHI

Non Cut Enzymes

AatII	Acc65I	AccII	AccIII	AcI	AcyI
AflIII	Alw44I	AlwNI	ApaBI	ApaLI	AscI
Asp718I	AsuII	AvaI	BalI	BbeI	BbvII
BclI	BglI	BglII	Bpu1102I	BsaHI	BsaOI
Bsc91I	BsiI	BspHI	BspMI	BspMII	BssHII
BstD102I	BstEII	BstXI	Bsu36I	CfrI	Clai
Csp45I	CspI	CvnI	DraIII	EagI	Ecl136II
Eco31I	Eco47III	Eco52I	Eco56I	Eco72I	EcoHI
EcoICRI	EcoNI	EcoRI	EcoRV	EheI	EspI
FnuDII	FseI	HaeII	HgaI	HgiAI	HhaI
HindII	HindIII	HinPII	HpaI	I-PpoI	KpnI
MaeII	Mlu113I	MluI	MscI	MspAlI	MstI
MstII	NaeI	NarI	NcoI	NdeI	NheI
NotI	NruI	NspBII	PacI	PflMI	PleI
PmaCI	PmeI	PvuI	PvuII	SacI	SacII
SalI	SauI	ScaI	SciI	SfiI	SgrAI
SmaI	SnaBI	SphI	SplI	SpoI	SrfI
SstI	SstII	StuI	SunI	SwaI	ThaI
VspI	XbaI	XcmI	XhoI	XmaI	XmaIII
XmnI	XorII				

Restriction analysis on TV013G15

Enzymes with >1 sites are not shown

32 sites found

List by Site Order

20	SspI	666	BanI	891	MspI	1215	SduI
206	SapI	676	MfeI	906	AflIII	1216	AvrII
261	Bsp1407I	715	SpeI	1046	NspI	1430	TaqI
376	DrdI	887	AccI	1060	PpuMI	1473	XhoII
405	HinfI	890	PinAI	1064	Tth111I	1473	BamHI
431	HphI	890	AgeI	1215	ApaI	1473	Sau3AI
459	BsmI	890	Cfr10I	1215	Bsp1286I	1473	MboI
627	PstI	891	HpaII	1215	BanII	1475	DpnI

Non Cut Enzymes

AatII	Acc65I	AccII	AccIII	AcI	AcyI
AflIII	Alw44I	AlwNI	ApaBI	ApaLI	AscI
Asp718I	AsuII	AvaI	BalI	BbeI	BbvII
BclI	BglI	BglII	Bpu1102I	BsaHI	BsaOI
Bsc91I	BsiI	BspHI	BspMI	BspMII	BssHII
BstD102I	BstEII	BstXI	Bsu36I	CfrI	Clai
Csp45I	CspI	CvnI	DraIII	EagI	Ecl136II
Eco31I	Eco47III	Eco52I	Eco56I	Eco72I	EcoHI
EcoICRI	EcoNI	EcoRI	EcoRV	EheI	EspI
FnuDII	FseI	HaeII	HgaI	HgiAI	HhaI
HindII	HindIII	HinPII	HpaI	I-PpoI	KpnI
MaeII	Mlu113I	MluI	MscI	MspAlI	MstI
MstII	NaeI	NarI	NcoI	NdeI	NheI
NotI	NruI	NspBII	PacI	PflMI	PleI
PmaCI	PmeI	PvuI	PvuII	SacI	SacII
SalI	SauI	ScaI	SciI	SfiI	SgrAI
SmaI	SnaBI	SphI	SplI	SpoI	SrfI
SstI	SstII	StuI	SunI	SwaI	ThaI
VspI	XbaI	XcmI	XhoI	XmaI	XmaIII
XmnI	XorII				

B2.2 Unique cutting and non-cutting Restriction Digest Enzymes for the gag clones

Restriction analysis on TV014G73

Enzymes with >1 sites are not shown

24 sites found

List by Site Order

20	SspI	453	BsmI	1009	MspAII	1061	MscI
202	AflIII	621	PstI	1009	NspBII	1106	SfaNI
261	Bsp1407I	655	AlwNI	1040	NspI	1138	XhoII
298	StuI	656	ApaBI	1054	PpuMI	1206	ApaI
370	DrdI	709	SpeI	1059	CfrI	1206	BanII
453	NsiI	745	BstEII	1061	BalI	1396	BstD102I

Non Cut Enzymes

AatII	Acc65I	AccI	AccII	AccIII	AclI
AcyI	AflIII	AgeI	Alw44I	ApaLI	AscI
Asp718I	AsuII	AvaI	AvrII	BamHI	BanI
BbeI	BbvII	BclI	BglI	BglII	Bpu1102I
BsaHI	BsaOI	Bsc91I	BsiI	BspHI	BspMI
BspMII	BssHII	BstXI	Bsu36I	Cfr10I	Clal
Csp45I	CspI	CvnI	DraIII	EagI	Ecl136II
Eco31I	Eco47III	Eco52I	Eco56I	Eco72I	EcoHI
EcoICRI	EcoNI	EcoRI	EcoRV	EheI	EspI
FnuDII	FseI	HaeII	HgaI	HgiAI	HhaI
HindII	HindIII	HinPII	HpaI	HpaII	I-PpoI
KpnI	MaeII	MfeI	Mlu113I	MluI	MspI
MstI	MstII	NaeI	NarI	NcoI	NdeI
NheI	NotI	NruI	PacI	PflMI	PinAI
PmaCI	PmeI	PvuI	PvuII	SacI	SacII
SalI	SapI	SauI	ScaI	SciI	SfiI
SgrAI	SmaI	SnaBI	SphI	SplI	SpoI
SrfI	SstI	SstII	SunI	SwaI	ThaI
Tth111I	VspI	XbaI	XcmI	XhoI	XmaI
XmaIII	XmnI	XorII			

Restriction analysis on TV018G60

Enzymes with >1 sites are not shown

20 sites found

List by Site Order

20	SspI	453	NsiI	882	Sau3AI	1209	ApaI
206	SapI	453	BsmI	884	DpnI	1209	BanII
261	Bsp1407I	478	StuI	1106	ApaBI	1210	AvrII
370	DrdI	709	SpeI	1109	SfaNI	1411	HinfI
425	HphI	882	MboI	1199	AlwNI	1419	PleI

Non Cut Enzymes

AatII	Acc65I	AccI	AccII	AccIII	AclI
AcyI	AflIII	AflIII	AgeI	Alw44I	ApaLI
AscI	Asp718I	AsuII	AvaI	BalI	BamHI
BanI	BbeI	BbvII	BclI	BglI	BglII
Bpu1102I	BsaHI	BsaOI	Bsc91I	BsiI	BspHI
BspMI	BspMII	BssHII	BstD102I	BstEII	BstXI
Bsu36I	Cfr10I	CfrI	Clal	Csp45I	CspI
CvnI	DraIII	EagI	Ecl136II	Eco31I	Eco47III
Eco52I	Eco56I	Eco72I	EcoHI	EcoICRI	EcoNI
EcoRI	EcoRV	EheI	EspI	FnuDII	FseI
HaeII	HgaI	HgiAI	HhaI	HindII	HindIII
HinPII	HpaI	HpaII	I-PpoI	KpnI	MaeII
MfeI	Mlu113I	MluI	MscI	MspAII	MspI
MstI	MstII	NaeI	NarI	NcoI	NdeI
NheI	NotI	NruI	NspBII	PacI	PflMI
PinAI	PmaCI	PmeI	PvuI	PvuII	SacI
SacII	SalI	SauI	ScaI	SciI	SfiI
SgrAI	SmaI	SnaBI	SphI	SplI	SpoI
SrfI	SstI	SstII	SunI	SwaI	ThaI
Tth111I	VspI	XbaI	XcmI	XhoI	XhoII
XmaI	XmaIII	XmnI	XorII		

Restriction analysis on TV019G20

Enzymes with >1 sites are not shown

19 sites found

List by Site Order

20	SspI	453	BsmI	1059	CfrI	1209	ApaI
206	SapI	453	NsiI	1061	BalI	1209	BanII
303	EcoNI	621	PstI	1061	MscI	1210	AvrII

B2.2 Unique cutting and non-cutting Restriction Digest Enzymes for the gag clones

399	HinfI	709	SpeI	1109	SfaNI	1382	TaqI
425	HphI	1054	PpuMI	1184	BanI		

Non Cut Enzymes

AatII	Acc65I	AccI	AccII	AccIII	AclI
AcyI	AflIII	AflIII	AgeI	Alw44I	AlwNI
ApaBI	ApaLI	AscI	Asp718I	AsuII	AvaI
BamHI	BbeI	BbvII	BclI	BglI	BglII
Bpu1102I	BsaHI	BsaOI	Bsc91I	BsiI	BspHI
BspMI	BspMII	BssHII	BstD102I	BstEII	BstXI
Bsu36I	Cfr10I	Clal	Csp45I	CspI	CvnI
DraIII	DrdI	EagI	Ecl136II	Eco31I	Eco47III
Eco52I	Eco56I	Eco72I	EcoHI	EcoICRI	EcoRI
EcoRV	EheI	EspI	FnuDII	FseI	HaeII
HgaI	HgiAI	HhaI	HindII	HindIII	HinPII
HpaI	HpaII	I-PpoI	KpnI	MaeII	MfeI
Mlu113I	MluI	MspAlI	MspI	MstI	MstII
NaeI	NarI	NcoI	NdeI	NheI	NotI
NruI	NspBII	PacI	PflMI	PinAI	PleI
PmaCI	PmeI	PvuI	PvuII	SacI	SacII
SauI	SauI	ScaI	SciI	SfiI	SgrAI
SmaI	SnaBI	SphI	SplI	SpoI	SrfI
SstI	SstII	StuI	SunI	SwaI	ThaI
Tth111I	VspI	XbaI	XcmI	XhoI	XhoII
XmaI	XmaIII	XmnI	XorII		

Restriction analysis on TV019G25

Enzymes with >1 sites are not shown

26 sites found

List by Site Order

20	SspI	709	SpeI	1054	PpuMI	1209	ApaI
206	SapI	881	AccI	1059	CfrI	1209	BanII
399	HinfI	884	Cfr10I	1061	BalI	1210	AvrII
425	HphI	884	AgeI	1061	MscI	1323	Bsc91I
453	NsiI	884	PinAI	1109	SfaNI	1323	BbvII
453	BsmI	885	HpaII	1184	BanI	1382	TaqI
621	PstI	885	MspI				

Non Cut Enzymes

AatII	Acc65I	AccII	AccIII	AclI	AcyI
AflIII	AflIII	Alw44I	AlwNI	ApaBI	ApaLI
AscI	Asp718I	AsuII	AvaI	BamHI	BbeI
BclI	BglI	BglII	Bpu1102I	BsaHI	BsaOI
BsiI	BspHI	BspMI	BspMII	BssHII	BstD102I
BstEII	BstXI	Bsu36I	Clal	Csp45I	CspI
CvnI	DraIII	DrdI	EagI	Ecl136II	Eco31I
Eco47III	Eco52I	Eco56I	Eco72I	EcoHI	EcoICRI
EcoNI	EcoRI	EcoRV	EheI	EspI	FnuDII
FseI	HaeII	HgaI	HgiAI	HhaI	HindII
HindIII	HinPII	HpaI	I-PpoI	KpnI	MaeII
MfeI	Mlu113I	MluI	MspAlI	MstI	MstII
NaeI	NarI	NcoI	NdeI	NheI	NotI
NruI	NspBII	PacI	PflMI	PleI	PmaCI
PmeI	PvuI	PvuII	SacI	SacII	SalI
SauI	ScaI	ScaI	SciI	SfiI	SgrAI
SnaBI	SphI	SplI	SpoI	SrfI	SstI
SstII	StuI	SunI	SwaI	ThaI	Tth111I
VspI	XbaI	XcmI	XhoI	XhoII	XmaI
XmaIII	XmnI	XorII			

Restriction analysis on TV016G95

Enzymes with >1 sites are not shown

23 sites found

List by Site Order

53	XmnI	358	AlwNI	899	PinAI	1113	MaeIII
138	HpaI	440	HphI	899	Cfr10I	1185	MfeI
138	HindII	607	Eco31I	899	AgeI	1227	ApaI
311	SapI	636	PstI	1052	Eam1105I	1227	BanII
358	NspBII	664	SphI	1071	EcoHI	1228	AvrII
358	MspAlI	724	SpeI	1074	CfrI		

Non Cut Enzymes

AatII	Acc65I	AccII	AccIII	AclI	AcyI
-------	--------	-------	--------	------	------

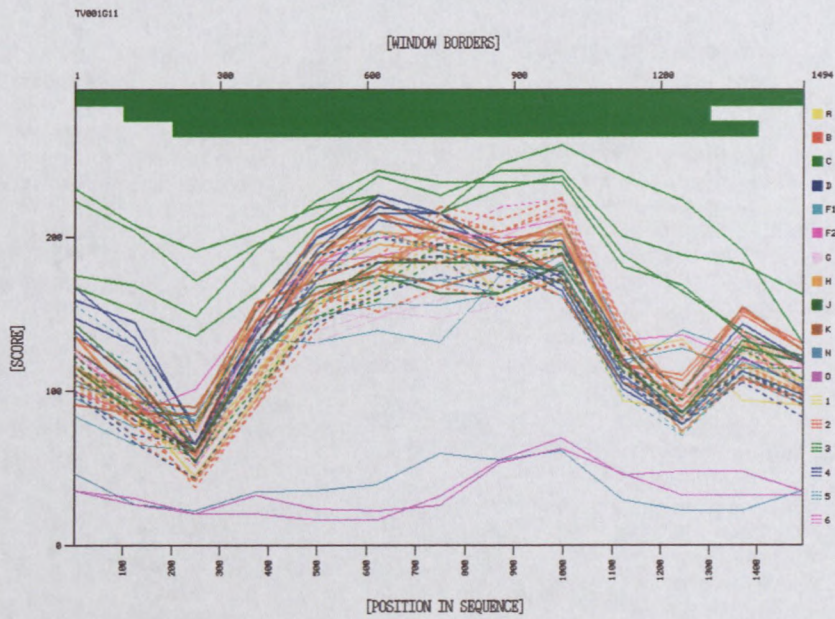
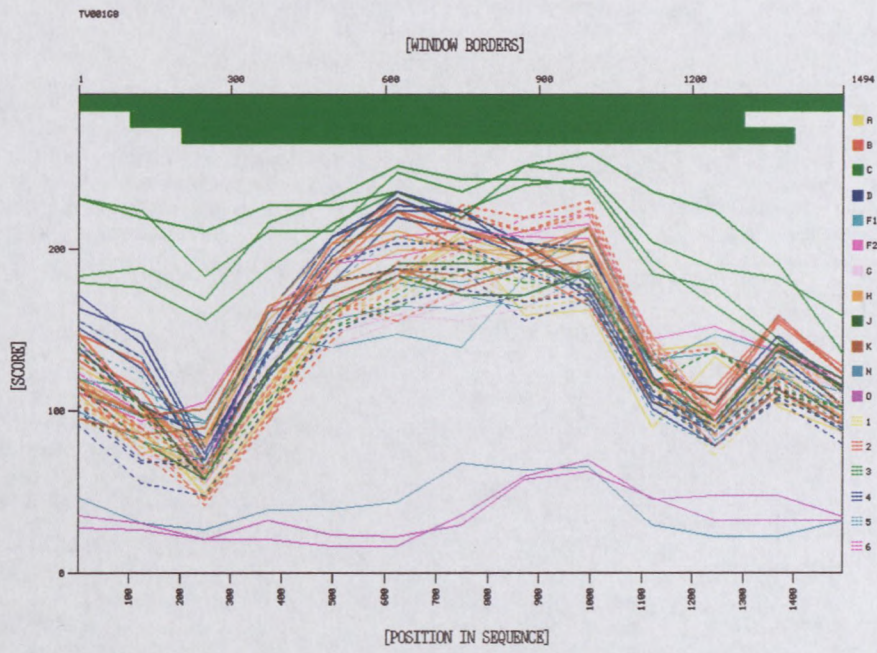
B2.2 Unique cutting and non-cutting Restriction Digest Enzymes for the *gag* clones

AflII	AflIII	Alw44I	ApaLI	AscI	Asp718I
AsuII	AvaI	BalI	BamHI	BanI	BbeI
BbvII	BclI	BglI	BglII	Bpu1102I	BsaHI
BsaOI	Bsc91I	BsiI	Bsp1407I	BspHI	BspMI
BspMII	BssHII	BstD102I	BstEII	BstXI	Bsu36I
ClaI	Csp45I	CspI	CvnI	DraIII	DrdI
EagI	Ecl136II	Eco47III	Eco52I	Eco56I	Eco72I
EcoICRI	EcoNI	EcoRI	EcoRV	EheI	EspI
FnuDII	FseI	HaeII	HgaI	HgiAI	HhaI
HinI	HinPII	I-PpoI	KpnI	MaeII	Mlu113I
MluI	MscI	MstI	MstII	NaeI	NarI
NcoI	NdeI	NheI	NotI	NruI	PacI
PflMI	PleI	PmaCI	PmeI	PvuI	PvuII
SacI	SacII	SalI	SauI	ScaI	SciI
SfiI	SgrAI	SmaI	SnaBI	SplI	SpoI
SrfI	SspI	SstI	SstII	StuI	SunI
SwaI	TaqI	ThaI	Tth111I	VspI	XbaI
XcmI	XhoI	XhoII	XmaI	XmaIII	XorII

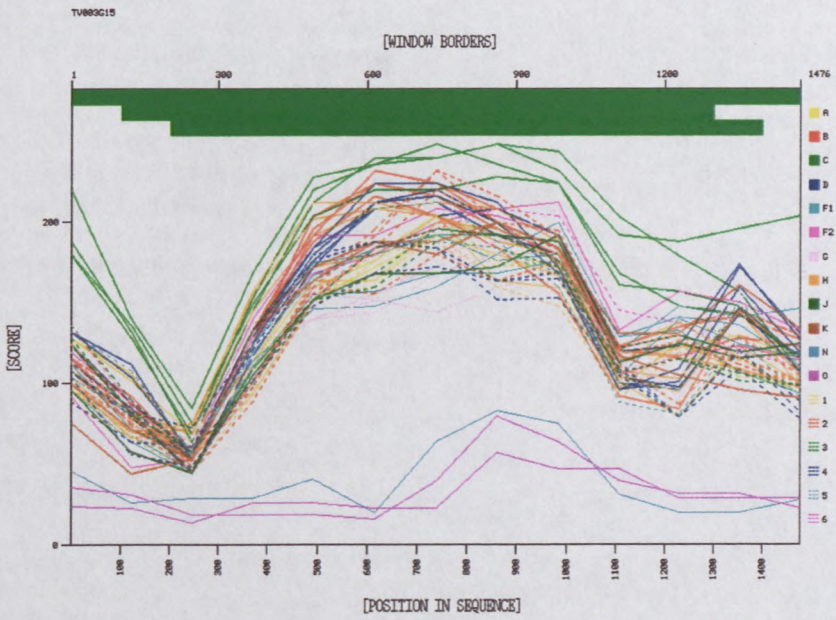
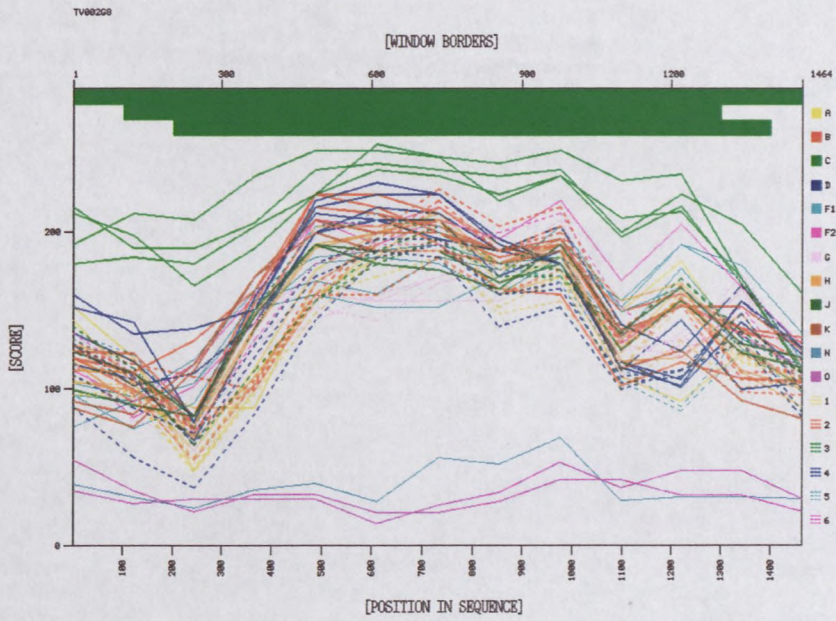
B3. Similarity Plots of the *gag* clones obtained from the NCBI subtyping tool

- B3. The similarity plots of the NCBI HIV-1 subtyping database for the different *gag* clones. The right hand legend indicates the different reference subtypes, which the query sequence was compared to. The solid bar at the top of the graph represents the sybtype most similar to the query sequence. The X-axis represents the nucleotide position, while the Y-axis represents the score of identity obtained when comparing the query sequence to the individual reference sequences. The similarity between the query and reference sequences increases with an increase in score.

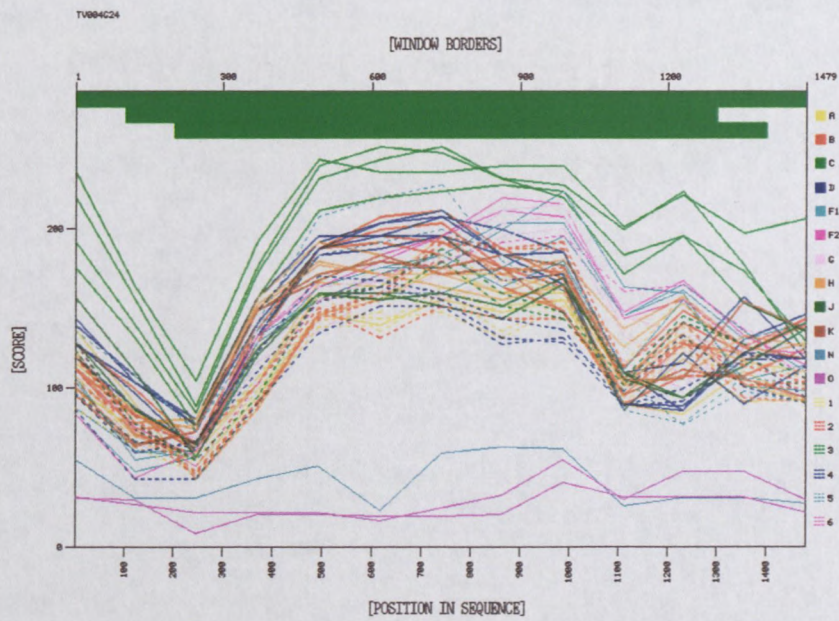
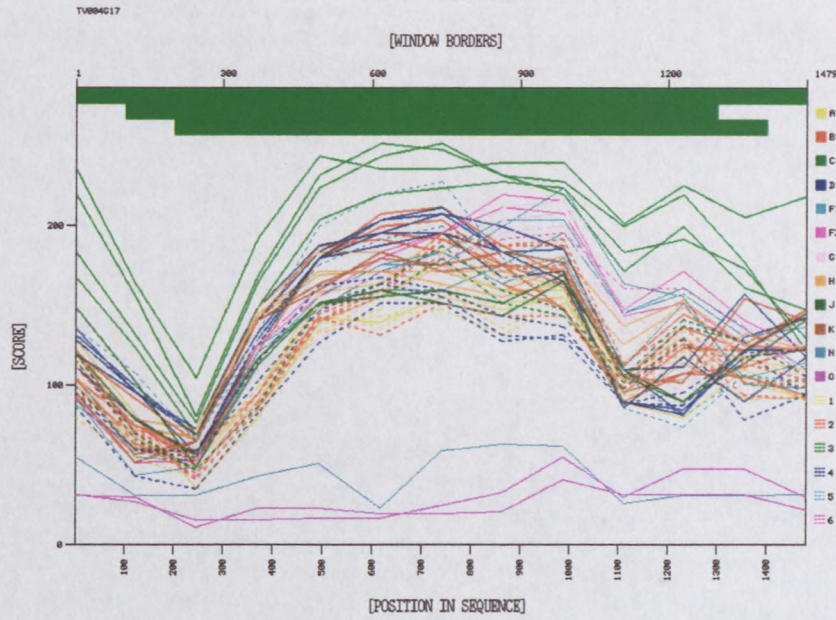
B3. Similarity Plots of the *gag* clones obtained from the NCBI subtyping tool



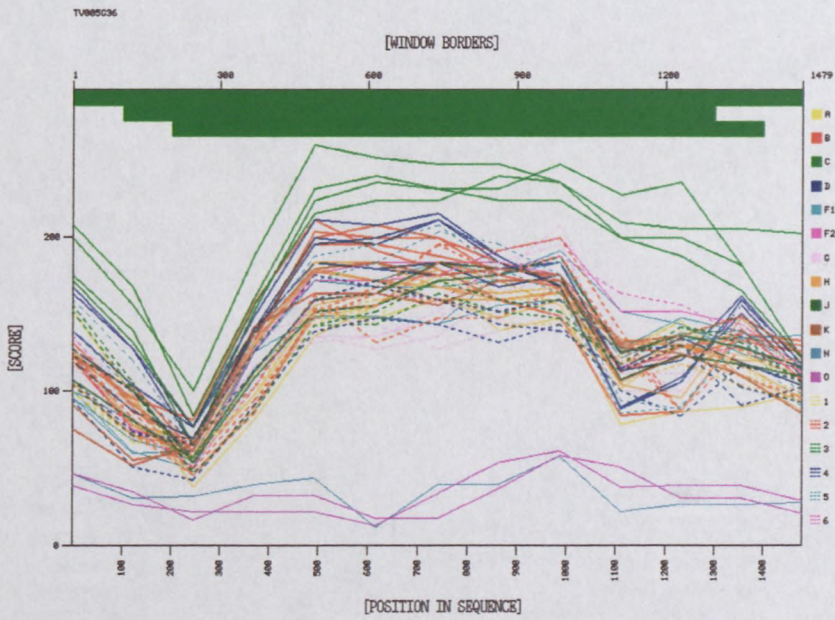
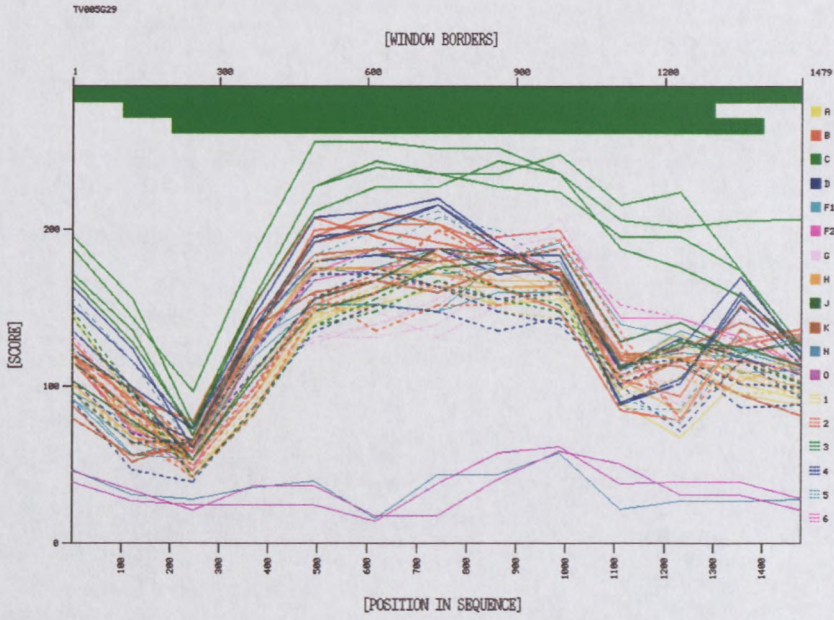
B3. Similarity Plots of the gag clones obtained from the NCBI subtyping tool



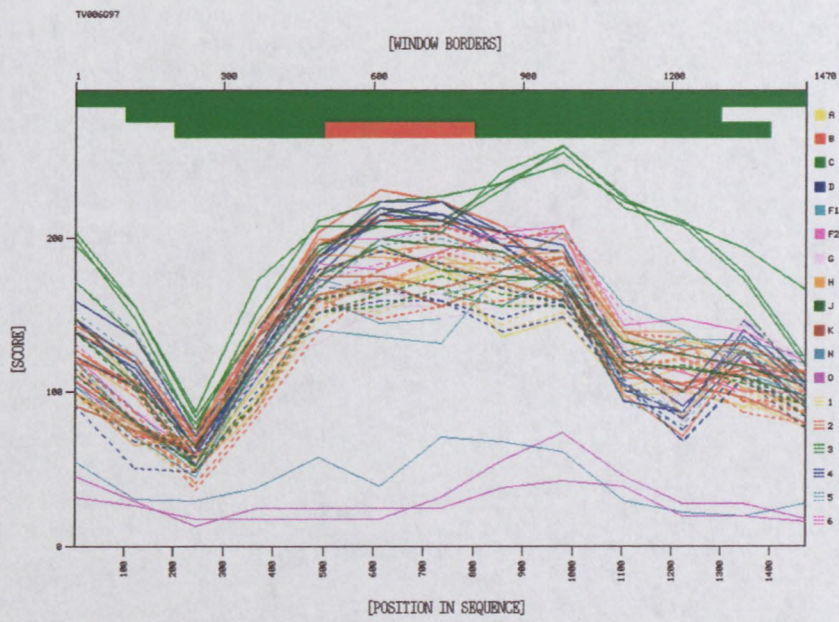
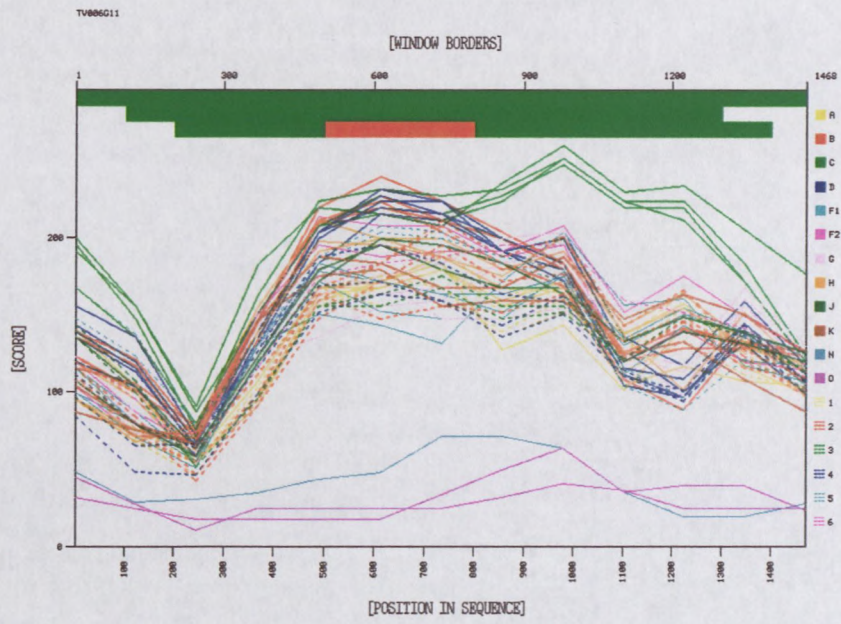
B3. Similarity Plots of the gag clones obtained from the NCBI subtyping tool



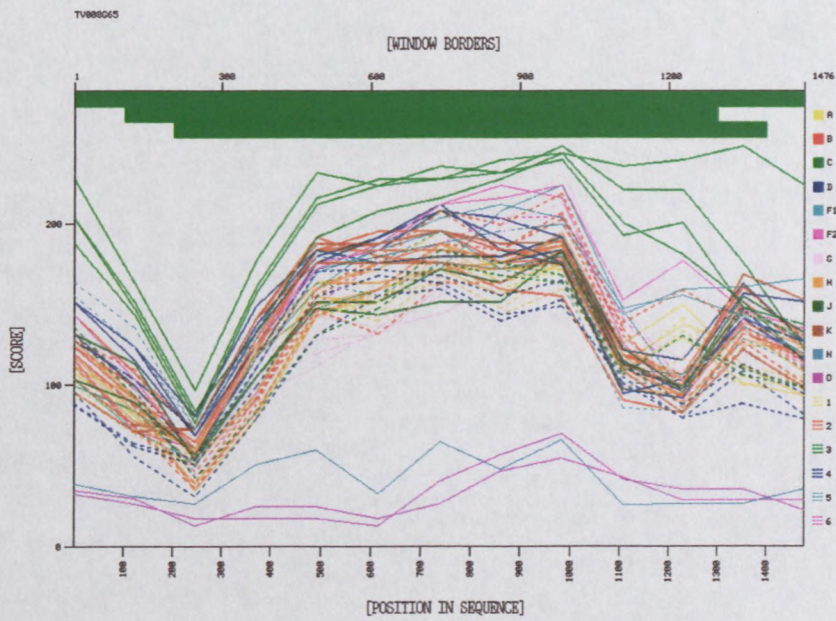
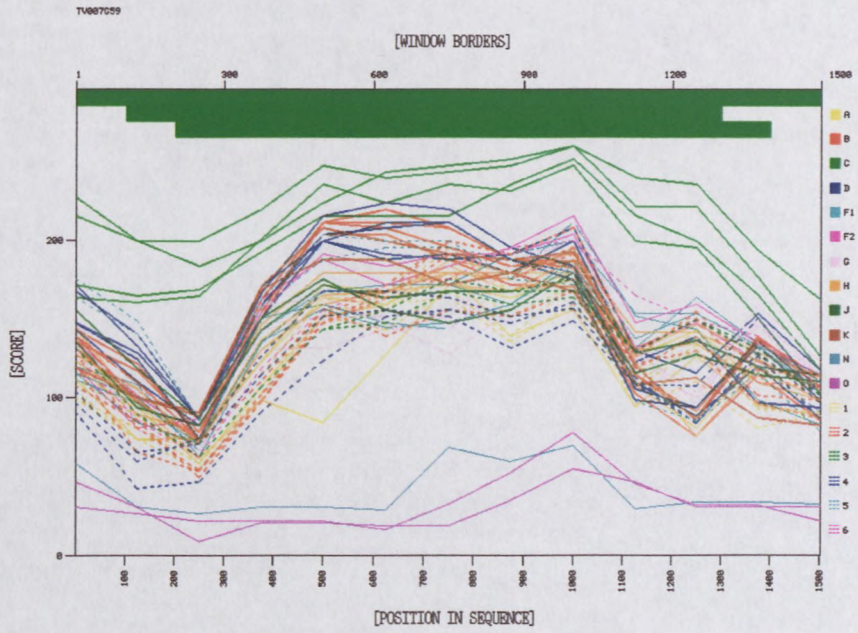
B3. Similarity Plots of the gag clones obtained from the NCBI subtyping tool



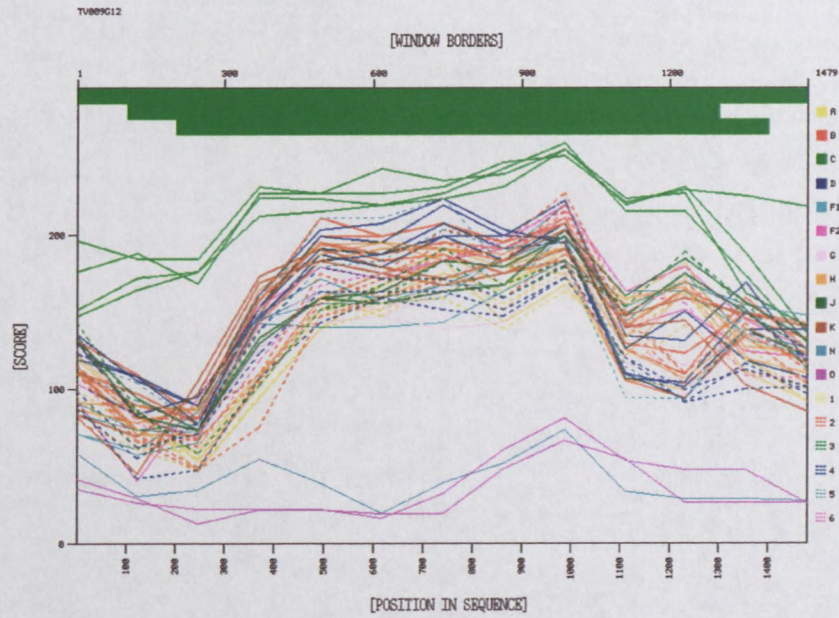
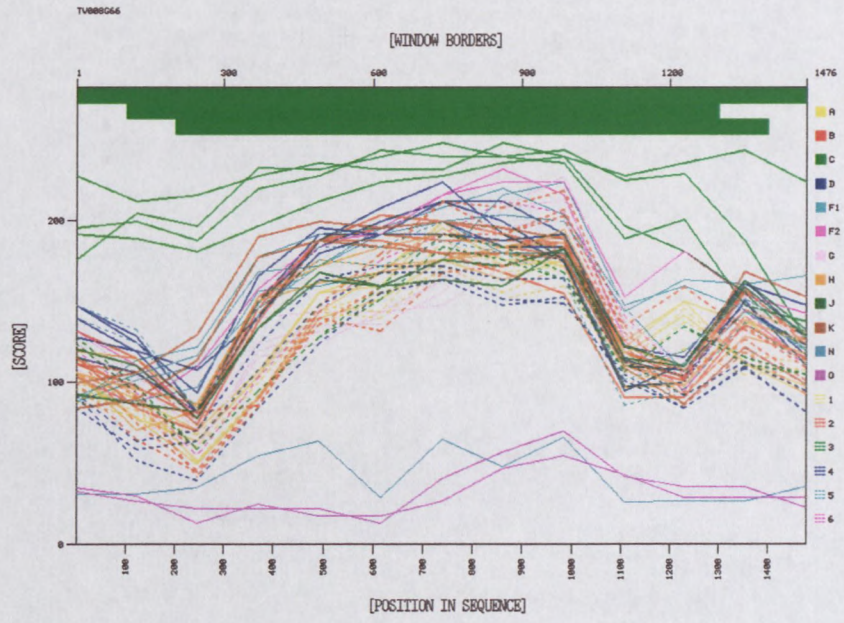
B3. Similarity Plots of the gag clones obtained from the NCBI subtyping tool



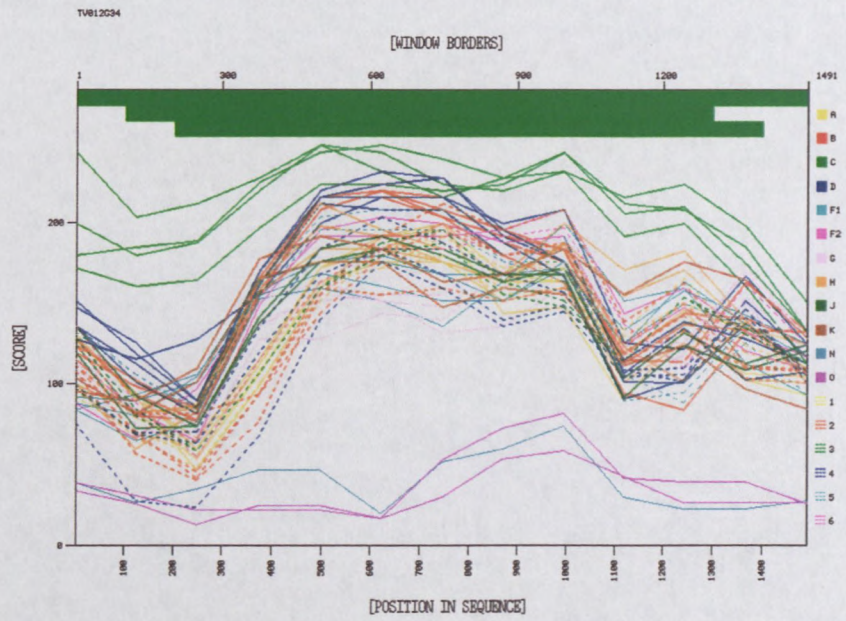
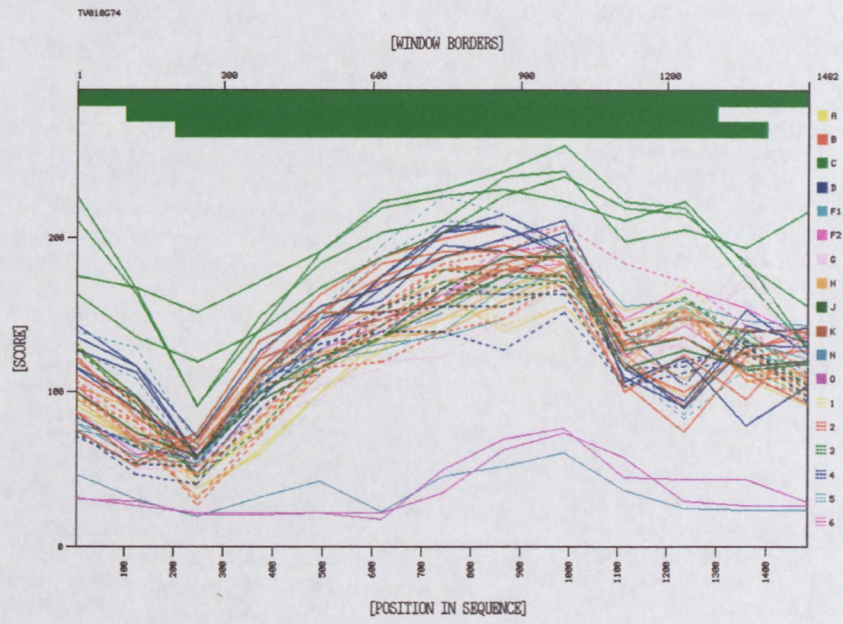
B3. Similarity Plots of the gag clones obtained from the NCBI subtyping tool



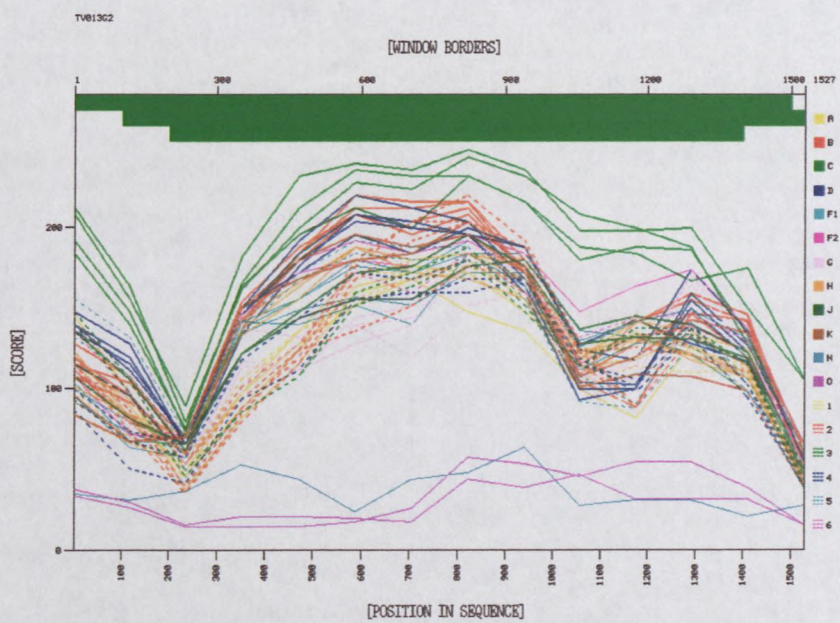
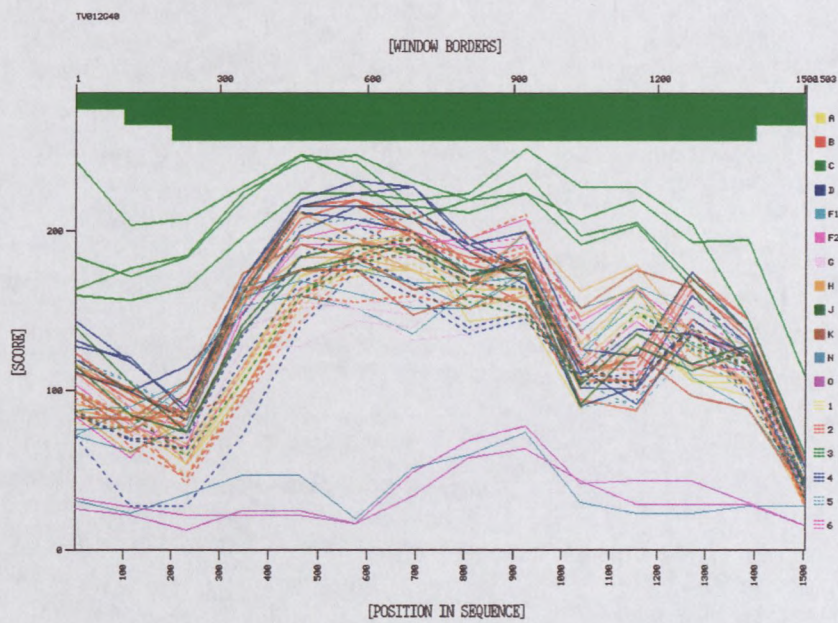
B3. Similarity Plots of the gag clones obtained from the NCBI subtyping tool



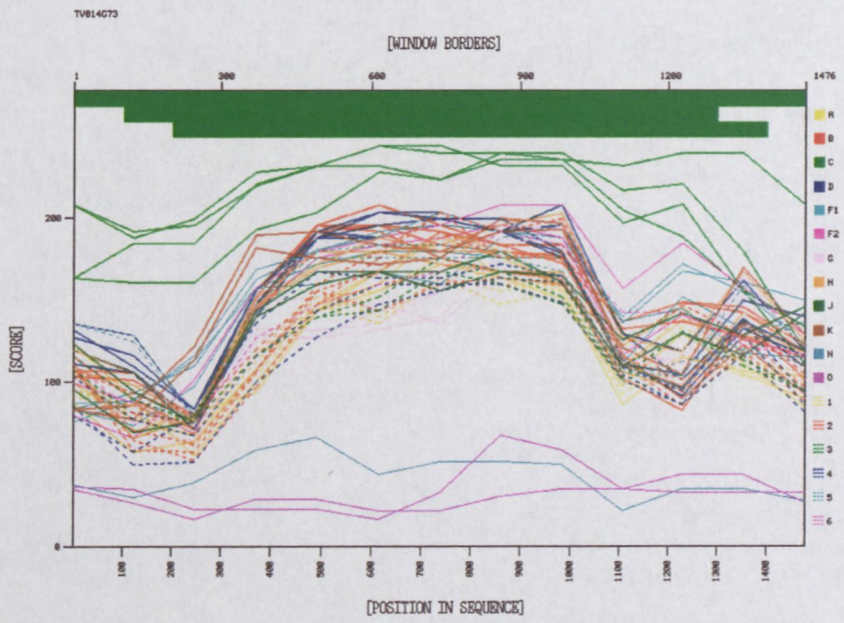
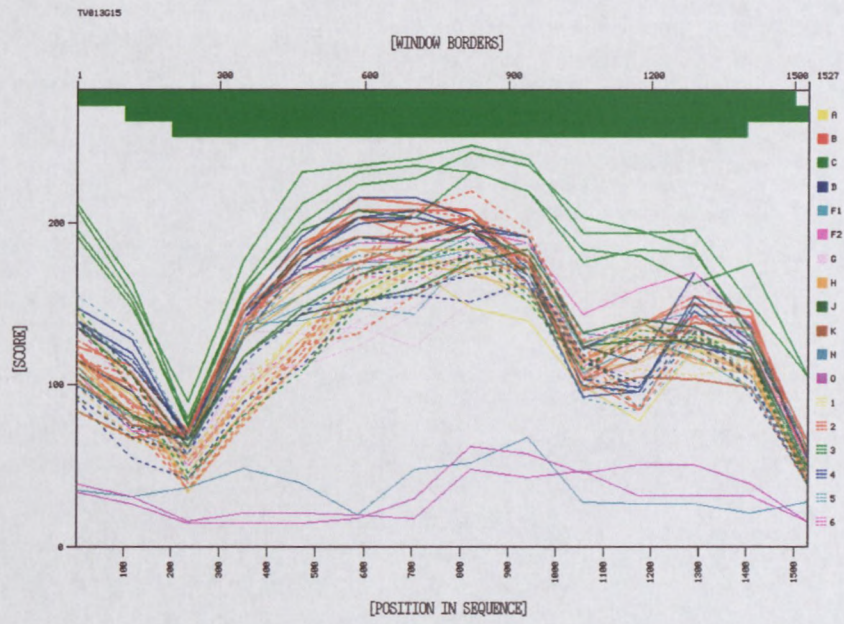
B3. Similarity Plots of the gag clones obtained from the NCBI subtyping tool



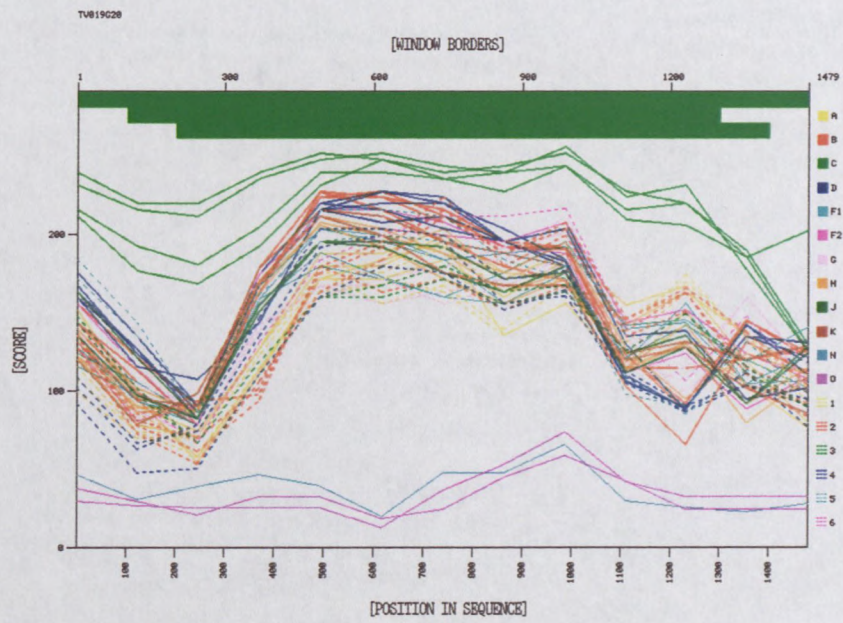
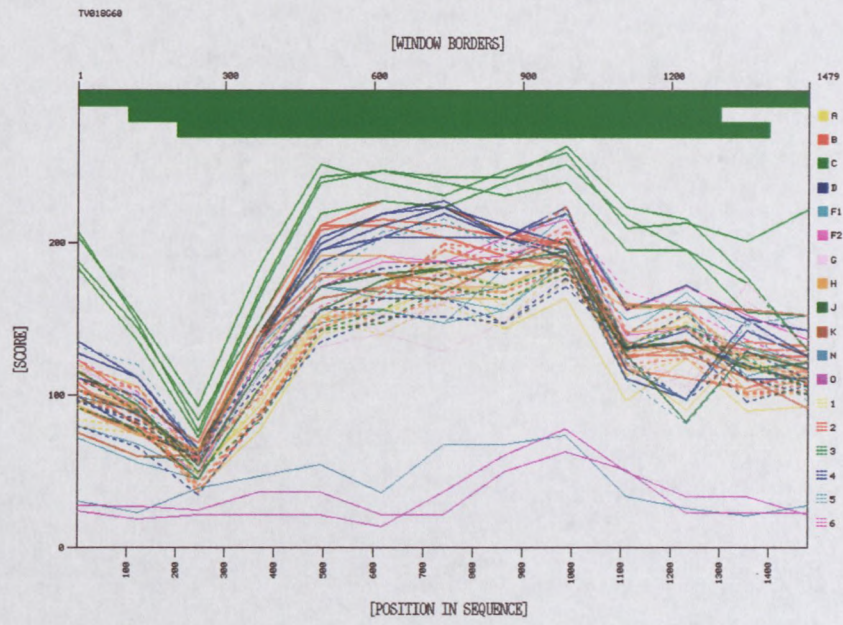
B3. Similarity Plots of the gag clones obtained from the NCBI subtyping tool



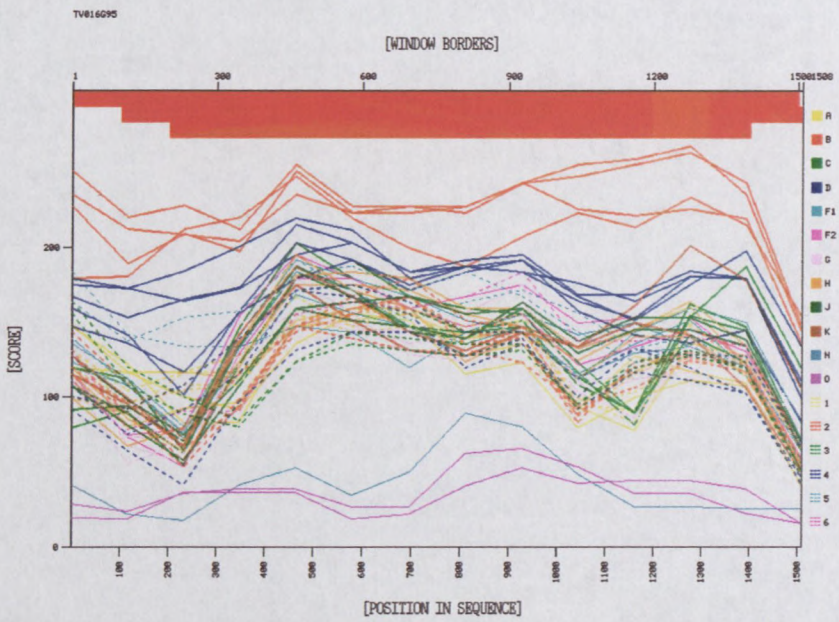
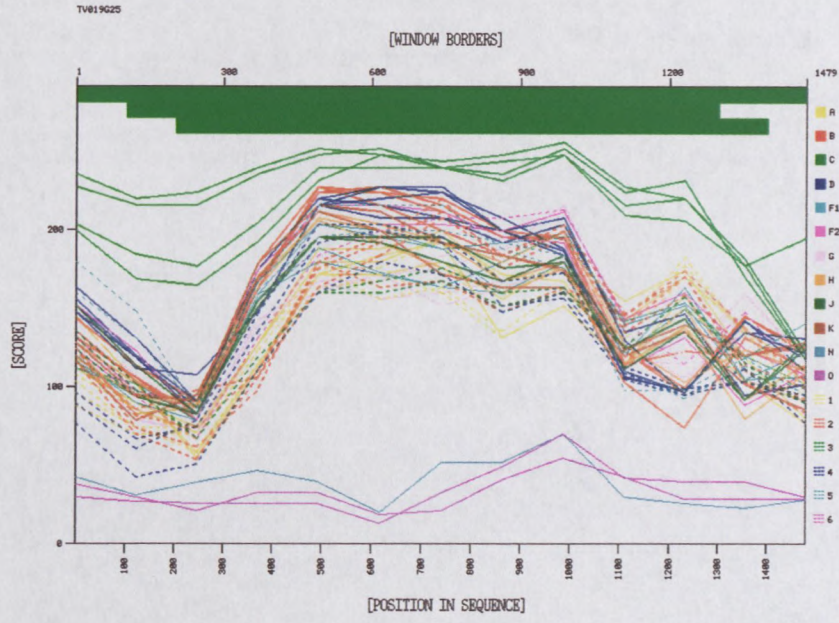
B3. Similarity Plots of the gag clones obtained from the NCBI subtyping tool



B3. Similarity Plots of the gag clones obtained from the NCBI subtyping tool



B3. Similarity Plots of the gag clones obtained from the NCBI subtyping tool



B4. Distance Matrices

The distance matrices are generated by GeneDoc and is a function describing the relationships between pairs of sequences. Three numbers are generated:

1. The number of residues that match exactly (identical residues) between the two sequences.
2. The number of residues that match similar residues (conservative substitutions) between two sequences. This number differs from the first number when amino acids are compared, but are identical to the first when nucleotides are compared.
3. The number of residues lined up with a gap character.

Each number is expressed as a count and as percentage on different sides of the matrix diagonal. The diagonal shows how many locations have at least one residue for the single sequence, i.e., the sequence length.

B4.1 Distance Matrix for the complete gag

Gag

	05_DF_V091		D_MU01141		F1_80R020		F1_MP411		F2_MP257		K_MP535		H_80CF056		J_SE7847		N_YB30		CPZ_CPZ15		CPZ_CPZ24B	
	05_DF_V1310	D_B4ZRG05	F1_V1850	F1_FN9M3	F2_MP255	KEQT811C	H_V1907	H_V1901	J_SE7022	CPZ_CAM3	CPZ_CPZ24B	CPZ_CPZ24B	CPZ_CPZ24B	CPZ_CPZ24B	CPZ_CPZ24B	CPZ_CPZ24B	CPZ_CPZ24B	CPZ_CPZ24B	CPZ_CPZ24B	CPZ_CPZ24B	CPZ_CPZ24B	CPZ_CPZ24B
	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22
1	076	076	080	076	080	080	080	080	080	080	080	080	080	080	080	080	080	080	080	080	080	080
2	076	076	080	076	080	080	080	080	080	080	080	080	080	080	080	080	080	080	080	080	080	080
3	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084
4	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084
5	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082
6	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084
7	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084
8	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084
9	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082
10	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082
11	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082
12	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084
13	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082
14	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082
15	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082
16	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082
17	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082
18	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082
19	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082
20	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082
21	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084
22	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082
23	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082
24	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084
25	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082
26	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082
27	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082
28	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082
29	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084
30	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082
31	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082
32	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082
33	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082
34	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082
35	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082
36	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082
37	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082
38	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082	082
39	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084
40	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084
41	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084	084

B4.1 Distance Matrix for the complete gag

Gag	TV00G24	C20R25	TV01G40	TV00G28	C20R50C2	TV01G25	TV00G59	TV00G06	TV00G07	TV01G4	TV00G15	B4R82	TV01G05	B4R22	B4R25	B4R33	DELJ	D10K																							
1	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60											
2	1382	1383	1378	1351	1384	1366	1368	1352	1357	1382	1381	1479	90%	82%	80%	83%	87%	81%	82%	91%	80%	82%	82%	80%	85%	84%	86%	87%	84%	86%	84%	86%	84%	86%	84%						
3	1384	1387	1376	1351	1384	1366	1368	1352	1357	1382	1381	0	90%	82%	80%	83%	87%	81%	82%	91%	80%	82%	82%	80%	85%	84%	86%	87%	84%	86%	84%	86%	84%	86%	84%						
4	8	8	8	28	44	58	48	27	30	23	0	0	1%	0%	0%	0%	1%	1%	2%	2%	0%	1%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%						
5	1384	1387	1376	1351	1384	1366	1368	1352	1357	1382	1381	1386	1383	1469	0	479	92%	92%	92%	92%	91%	92%	91%	90%	90%	89%	89%	84%	84%	84%	85%	85%	86%	86%	86%	86%					
6	1377	1378	1371	1354	1341	1348	1355	1347	1349	1395	1383	1381	1	303	1500	0	91%	91%	91%	91%	90%	90%	89%	89%	89%	89%	84%	84%	84%	85%	85%	86%	86%	86%	86%	86%					
7	25	25	25	45	61	75	65	46	44	15	24	21	21	0	1	476	86%	84%	91%	91%	91%	91%	91%	82%	81%	82%	81%	85%	84%	85%	85%	85%	85%	85%	85%	85%					
8	1378	1375	1376	1347	1349	1367	1364	1361	1352	1376	1364	1368	1	388	1371	0	90%	84%	91%	91%	91%	91%	91%	82%	81%	82%	81%	85%	84%	85%	85%	85%	85%	85%	85%	85%					
9	1379	1378	1379	1350	1352	1369	1368	1363	1357	1379	1364	1367	1	367	1374	1	455	1478	84%	91%	91%	91%	91%	91%	82%	81%	82%	81%	85%	84%	85%	85%	85%	85%	85%	85%	85%				
10	1378	1377	1369	1341	1338	1355	1360	1348	1342	1378	1374	1381	1	383	1378	1	398	1368	1	478	91%	91%	91%	91%	90%	82%	81%	82%	81%	85%	84%	85%	85%	85%	85%	85%	85%	85%			
11	1378	1377	1369	1341	1338	1355	1360	1348	1342	1378	1374	1381	1	383	1378	1	398	1368	1	478	91%	91%	91%	91%	90%	82%	81%	82%	81%	85%	84%	85%	85%	85%	85%	85%	85%	85%			
12	1359	1356	1354	1345	1335	1349	1350	1345	1338	1352	1351	1342	1	362	1355	1	358	1360	1	359	1470	84%	92%	92%	91%	91%	85%	84%	83%	84%	84%	84%	84%	84%	84%	84%	84%	84%			
13	1359	1356	1354	1345	1335	1349	1350	1345	1338	1352	1351	1342	1	362	1355	1	358	1360	1	359	1470	84%	92%	92%	91%	91%	85%	84%	83%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%		
14	1370	1367	1363	1349	1342	1358	1359	1352	1345	1355	1358	1371	1	371	1362	1	363	1363	1	364	1449	1	468	87%	82%	82%	81%	80%	81%	81%	81%	81%	81%	81%	81%	81%	81%	81%	81%	81%	
15	1360	1377	1372	1350	1341	1368	1367	1363	1362	1374	1374	1378	1	378	1383	1	379	1379	1	373	1386	1	395	1464	98%	91%	80%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	
16	1360	1377	1372	1350	1341	1368	1367	1363	1362	1374	1374	1378	1	378	1383	1	379	1379	1	373	1386	1	395	1464	98%	91%	80%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%
17	1360	1377	1372	1350	1341	1368	1367	1363	1362	1374	1374	1378	1	378	1383	1	379	1379	1	373	1386	1	395	1464	98%	91%	80%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%
18	1360	1377	1372	1350	1341	1368	1367	1363	1362	1374	1374	1378	1	378	1383	1	379	1379	1	373	1386	1	395	1464	98%	91%	80%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%
19	1360	1377	1372	1350	1341	1368	1367	1363	1362	1374	1374	1378	1	378	1383	1	379	1379	1	373	1386	1	395	1464	98%	91%	80%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%
20	1360	1377	1372	1350	1341	1368	1367	1363	1362	1374	1374	1378	1	378	1383	1	379	1379	1	373	1386	1	395	1464	98%	91%	80%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%
21	1360	1377	1372	1350	1341	1368	1367	1363	1362	1374	1374	1378	1	378	1383	1	379	1379	1	373	1386	1	395	1464	98%	91%	80%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%
22	1360	1377	1372	1350	1341	1368	1367	1363	1362	1374	1374	1378	1	378	1383	1	379	1379	1	373	1386	1	395	1464	98%	91%	80%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%
23	1360	1377	1372	1350	1341	1368	1367	1363	1362	1374	1374	1378	1	378	1383	1	379	1379	1	373	1386	1	395	1464	98%	91%	80%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%
24	1360	1377	1372	1350	1341	1368	1367	1363	1362	1374	1374	1378	1	378	1383	1	379	1379	1	373	1386	1	395	1464	98%	91%	80%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%
25	1360	1377	1372	1350	1341	1368	1367	1363	1362	1374	1374	1378	1	378	1383	1	379	1379	1	373	1386	1	395	1464	98%	91%	80%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%
26	1360	1377	1372	1350	1341	1368	1367	1363	1362	1374	1374	1378	1	378	1383	1	379	1379	1	373	1386	1	395	1464	98%	91%	80%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%
27	1360	1377	1372	1350	1341	1368	1367	1363	1362	1374	1374	1378	1	378	1383	1	379	1379	1	373	1386	1	395	1464	98%	91%	80%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%
28	1360	1377	1372	1350	1341	1368	1367	1363	1362	1374	1374	1378	1	378	1383	1	379	1379	1	373	1386	1	395	1464	98%	91%	80%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%
29	1360	1377	1372	1350	1341	1368	1367	1363	1362	1374	1374	1378	1	378	1383	1	379	1379	1	373	1386	1	395	1464	98%	91%	80%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%
30	1360	1377	1372	1350	1341	1368	1367	1363	1362	1374	1374	1378	1	378	1383	1	379	1379	1	373	1386	1	395	1464	98%	91%	80%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%
31	1360	1377	1372	1350	1341	1368	1367	1363	1362	1374	1374	1378	1	378	1383	1	379	1379	1	373	1386	1	395	1464	98%	91%	80%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%
32	1360	1377	1372	1350	1341	1368	1367	1363	1362	1374	1374	1378	1	378	1383	1	379	1379	1	373	1386	1	395	1464	98%	91%	80%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%
33	1360	1377	1372	1350	1341	1368	1367	1363	1362	1374	1374	1378	1	378	1383	1	379	1379	1	373	1386	1	395	1464	98%	91%	80%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%
34	1360	1377	1372	1350	1341	1368	1367	1363	1362	1374	1374	1378	1	378	1383	1	379	1379	1	373	1386	1	395	1464	98%	91%	80%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%
35	1360	1377	1372	1350	1341	1368	1367	1363	1362	1374	1374	1378	1	378	1383	1	379	1379	1	373	1386	1	395	1464	98%	91%	80%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%
36	1360	1377	1372	1350	1341	1368	1367	1363	1362	1374	1374	1378	1	378	1383	1	379	1379	1	373	1386	1	395	1464	98%	91%	80%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%
37	1360	1377	1372	1350	1341	1368	1367	1363	1362	1374	1374	1378	1	378	1383	1	379	1379	1	373	1386	1	395	1464	98%	91%	80%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%
38	1360	1377	1372	1350	1341	1368	1367	1363	1362	1374	1374	1378	1	378	1383	1	379	1379	1	373	1386	1	395	1464	98%	91%	80%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%
39	1360	1377	1372	1350	1341	1368	1367	1363	1362	1374	1374	1378	1	378	1383	1	37																								

B4.1 Distance Matrix for the complete gag

Gag	05_DF_V981	D_84U1G141	F1_R5BR20	F1_MP411	F2_MP257	K_MP535	H_9CQF056	J_SET867	N_Y8F30	CP2_CPF25	CP2_CPF28								
05_DF_V1310	D_84CR065	F1_V1850	F1_FIN063	F2_MP255	KEQT811C	H_V1967	H_V1981	J_SET022	CP2_CAM3	CP2_CPF2AB									
81	82	83	84	85	86	87	88	89	90	91	92								
85%	85%	86%	86%	85%	85%	84%	85%	85%	85%	84%	85%	85%	86%	87%	88%	89%	90%	91%	92%
85%	85%	86%	86%	85%	85%	84%	85%	85%	85%	84%	85%	85%	86%	87%	88%	89%	90%	91%	92%
2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%
43	85%	85%	86%	86%	85%	85%	84%	85%	85%	84%	85%	85%	86%	87%	88%	89%	90%	91%	92%
85%	85%	86%	86%	85%	85%	84%	85%	85%	85%	84%	85%	85%	86%	87%	88%	89%	90%	91%	92%
85%	85%	86%	86%	85%	85%	84%	85%	85%	85%	84%	85%	85%	86%	87%	88%	89%	90%	91%	92%
2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%
44	84%	83%	84%	84%	83%	83%	83%	83%	84%	83%	83%	83%	83%	83%	83%	83%	83%	83%	83%
84%	83%	84%	84%	83%	83%	83%	83%	84%	83%	83%	83%	83%	83%	83%	83%	83%	83%	83%	83%
3%	3%	3%	3%	3%	3%	3%	3%	3%	3%	3%	3%	3%	3%	3%	3%	3%	3%	3%	3%
45	85%	84%	86%	85%	85%	84%	84%	85%	85%	83%	85%	85%	84%	84%	84%	84%	84%	84%	84%
85%	84%	86%	85%	85%	84%	84%	84%	85%	85%	83%	85%	85%	84%	84%	84%	84%	84%	84%	84%
2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%
46	85%	85%	86%	85%	86%	85%	84%	85%	86%	83%	85%	85%	84%	84%	85%	85%	85%	85%	85%
85%	85%	86%	85%	86%	85%	84%	85%	86%	83%	85%	85%	84%	84%	85%	85%	85%	85%	85%	85%
2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%
47	84%	84%	85%	85%	84%	84%	84%	85%	85%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%
84%	84%	85%	85%	84%	84%	84%	84%	85%	85%	84%	84%	84%	84%	84%	84%	84%	84%	84%	84%
2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%
48	85%	84%	86%	85%	84%	83%	83%	83%	85%	83%	84%	85%	84%	84%	84%	84%	84%	84%	84%
85%	84%	86%	85%	84%	83%	83%	83%	85%	83%	84%	85%	84%	84%	84%	84%	84%	84%	84%	84%
2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%
49	85%	84%	86%	85%	84%	84%	84%	85%	85%	83%	85%	85%	84%	84%	85%	85%	85%	85%	85%
85%	84%	86%	85%	84%	84%	84%	84%	85%	85%	83%	85%	85%	84%	84%	85%	85%	85%	85%	85%
2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%
50	84%	83%	85%	85%	84%	83%	83%	83%	84%	83%	84%	85%	84%	83%	84%	84%	84%	84%	84%
84%	83%	85%	85%	84%	83%	83%	83%	84%	83%	84%	85%	84%	83%	84%	84%	84%	84%	84%	84%
3%	3%	3%	3%	3%	3%	3%	3%	3%	3%	3%	3%	3%	3%	3%	3%	3%	3%	3%	3%
51	83%	83%	85%	84%	83%	82%	82%	82%	83%	84%	83%	84%	84%	83%	83%	83%	83%	83%	83%
83%	83%	85%	84%	83%	82%	82%	82%	83%	84%	83%	84%	84%	83%	83%	83%	83%	83%	83%	83%
3%	3%	3%	3%	3%	3%	3%	3%	3%	3%	3%	3%	3%	3%	3%	3%	3%	3%	3%	3%
52	84%	84%	86%	85%	85%	83%	84%	83%	85%	83%	85%	85%	84%	84%	84%	84%	84%	84%	84%
84%	84%	86%	85%	85%	83%	84%	83%	85%	83%	85%	85%	84%	84%	84%	84%	84%	84%	84%	84%
2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%
53	85%	84%	85%	85%	84%	84%	83%	85%	85%	84%	85%	85%	83%	84%	84%	84%	84%	84%	84%
85%	84%	85%	85%	84%	84%	83%	85%	85%	84%	85%	85%	83%	84%	84%	84%	84%	84%	84%	84%
3%	3%	3%	3%	3%	3%	3%	3%	3%	3%	3%	3%	3%	3%	3%	3%	3%	3%	3%	3%
54	80%	80%	81%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%
80%	80%	81%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%
0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
55	85%	83%	86%	86%	85%	85%	86%	85%	86%	87%	85%	87%	86%	86%	86%	86%	86%	86%	86%
85%	83%	86%	86%	85%	85%	86%	85%	86%	87%	85%	87%	86%	86%	86%	86%	86%	86%	86%	86%
0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
56	87%	87%	89%	87%	85%	85%	84%	85%	85%	84%	86%	85%	84%	84%	85%	85%	85%	85%	85%
87%	87%	89%	87%	85%	85%	84%	85%	85%	84%	86%	85%	84%	84%	85%	85%	85%	85%	85%	85%
1%	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%
57	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%
80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%
0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
58	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%
80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%
0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
59	83%	82%	83%	82%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%
83%	82%	83%	82%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%
0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
60	82%	82%	82%	82%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%
82%	82%	82%	82%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%
0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
61	1500	80%	91%	80%	80%	80%	85%	85%	87%	86%	85%	87%	85%	85%	85%	85%	84%	84%	85%
0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
62	1415	1500	91%	80%	85%	86%	85%	86%	84%	85%	85%	85%	85%	85%	85%	85%	84%	84%	85%
0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
63	1387	1380	1508	82%	80%	80%	85%	86%	87%	87%	87%	87%	87%	87%	87%	87%	87%	87%	87%
82%	82%	82%	82%	80%	80%	80%	85%	86%	87%	87%	87%	87%	87%	87%	87%	87%	87%	87%	87%
1%	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%
64	1372	1368	1384	1509	80%	85%	84%	85%	86%	87%	85%	86%	86%	86%	86%	86%	86%	86%	86%
12	15	13	0	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
65	1301	1289	1310	1303	1485	93%	93%	92%	91%	90%	88%	89%	85%	85%	84%	85%	84%	84%	85%
24	21	31	0	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
66	1310	1298	1302	1291	1398	1488	92%	91%	90%	90%	87%	88%	85%	85%	84%	84%	84%	84%	85%
13	12	18	28	27	3	0	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
67	1293	1292	1301	1285	1399	1380	1491	91%	90%	89%	86%	86%	84%	85%	84%	84%	84%	84%	85%
30	27	37	36	6	9	0	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
68	1300	1295	1306	1295	1378	1370	1494	90%	89%	89%	86%	86%	84%	85%	84%	84%	84%	84%	85%
27	24	34	33	9	15	0	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
69	1314	1300	1319	1310	1355	1348	1345	1488	91%	90%	89%	86%	86%	84%	85%	84%	84%	84%	85%
19	16	28	27	3	6	0	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
70	1312	1308	1328	1321	1353	1357	1353	1343	1399	1500	86%	86%	86%	86%	86%	86%	86%	86%	86%
9	6	16	15	12	21	18	12	0	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
71	1296	1281	1298	1290	1310	1298	1297	1324	1323	1488	82%	84%	84%	84%	83%	84%	84%	84%	85%
21	18	28	27	3	0	8	0	12	0	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
72	1321	1305	1321	1313	1338	1333	1335	1315	1344	1348	1372	1488	86%	86%	85%	84%	85%	85%	86%
21	18	28	27	3	0	8	0	12	0	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
73	1288	1283	1323	1302	1284	1288	1283	1282	1285	1300	1277	1299	1500	85%	81%	80%	80%	81%	82%
13	16	14	13	31	28	37	32	28	18	28	28	0	0%	0%	0%	0%	0%	0%	0%
74																			

B4.2 Distance Matrix for the p17 matrix domain

p17

	TV00424	TV00104	C_#BHW17B3C	TV000538	TV01840	TV000596	TV012040	C_#S421088	TV018020	C_#BHW502C	TV013015	TV008087	TV003015	C_#T4220																					
	TV004017	TV006012	TV010011	TV005029	TV002058	TV008085	TV007074	TV012034	TV018025	TV007058	TV013022	TV014073	TV006011	C_#B2B025	G_#2H0083																				
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29						
1	387	80%	87%	81%	82%	82%	80%	81%	81%	82%	82%	81%	81%	81%	81%	82%	84%	83%	82%	82%	82%	82%	82%	82%	82%	82%	82%	80%	80%	80%	80%	78%			
2	0	80%	83%	83%	82%	82%	80%	81%	81%	82%	82%	81%	81%	81%	81%	82%	84%	83%	82%	82%	82%	82%	82%	82%	82%	82%	82%	80%	80%	80%	80%	78%			
3	0	0	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%		
4	385	387	83%	83%	82%	82%	80%	81%	81%	82%	82%	81%	81%	81%	81%	82%	84%	83%	82%	82%	82%	82%	82%	82%	82%	82%	82%	80%	80%	80%	80%	78%			
5	381	383	0	81%	81%	80%	80%	80%	80%	81%	82%	81%	81%	81%	81%	82%	84%	83%	82%	82%	82%	82%	82%	82%	82%	82%	82%	80%	80%	80%	80%	78%			
6	383	383	387	81%	81%	80%	80%	80%	80%	81%	82%	81%	81%	81%	81%	82%	84%	83%	82%	82%	82%	82%	82%	82%	82%	82%	82%	80%	80%	80%	80%	78%			
7	383	383	355	387	87%	82%	81%	82%	82%	84%	84%	81%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	80%	80%	80%	80%	78%			
8	383	383	355	0	87%	82%	81%	82%	82%	84%	84%	81%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	80%	80%	80%	80%	78%			
9	358	358	353	377	387	82%	80%	80%	80%	83%	82%	81%	80%	81%	82%	82%	81%	82%	80%	82%	80%	81%	82%	82%	82%	82%	80%	80%	80%	80%	78%				
10	358	358	353	377	0	82%	80%	80%	80%	83%	82%	81%	80%	81%	82%	82%	81%	82%	80%	82%	80%	81%	82%	82%	82%	82%	80%	80%	80%	80%	78%				
11	357	358	348	357	358	378	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	
12	357	358	348	357	358	378	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%
13	350	350	348	358	348	345	387	82%	82%	83%	82%	81%	80%	81%	82%	82%	81%	82%	80%	82%	80%	81%	82%	82%	82%	82%	82%	80%	80%	80%	80%	80%	80%	80%	80%
14	350	350	348	358	348	345	0	88%	82%	83%	82%	81%	80%	81%	82%	82%	81%	82%	80%	82%	80%	81%	82%	82%	82%	82%	82%	80%	80%	80%	80%	80%	80%	80%	80%
15	355	355	350	358	352	350	382	387	82%	83%	83%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	81%	82%	82%	82%	82%	82%	80%	80%	80%	80%	80%	80%	80%	80%
16	355	355	350	358	352	350	382	0	82%	83%	83%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	81%	82%	82%	82%	82%	82%	80%	80%	80%	80%	80%	80%	80%	80%
17	356	356	353	358	353	348	358	382	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%
18	356	356	353	358	353	348	358	382	0	82%	82%	82%	80%	81%	82%	82%	81%	82%	80%	82%	80%	81%	82%	82%	82%	82%	82%	80%	80%	80%	80%	80%	80%	80%	80%
19	358	358	353	368	361	350	380	381	84%	83%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%
20	358	358	353	368	361	350	380	381	0	84%	83%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%
21	358	358	353	368	361	350	380	381	0	84%	83%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%
22	359	359	357	364	359	351	382	383	386	387	87%	82%	83%	83%	82%	83%	84%	81%	81%	81%	82%	82%	82%	82%	82%	82%	82%	82%	80%	80%	80%	80%	80%	80%	80%
23	359	359	357	364	359	351	382	383	386	387	0	87%	82%	83%	83%	82%	83%	84%	81%	81%	81%	82%	82%	82%	82%	82%	82%	82%	82%	80%	80%	80%	80%	80%	80%
24	358	358	355	360	355	347	380	381	382	382	379	87%	81%	81%	81%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	80%	80%	80%	80%	80%	80%	80%
25	358	358	355	360	355	347	380	381	382	382	379	0	81%	81%	81%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	80%	80%	80%	80%	80%	80%	80%
26	358	358	355	360	355	347	380	381	382	382	379	0	81%	81%	81%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	80%	80%	80%	80%	80%	80%	80%
27	358	358	355	360	355	347	380	381	382	382	379	0	81%	81%	81%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	80%	80%	80%	80%	80%	80%	80%
28	358	358	355	360	355	347	380	381	382	382	379	0	81%	81%	81%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	80%	80%	80%	80%	80%	80%	80%
29	358	358	355	360	355	347	380	381	382	382	379	0	81%	81%	81%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	80%	80%	80%	80%	80%	80%	80%
30	358	358	355	360	355	347	380	381	382	382	379	0	81%	81%	81%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	80%	80%	80%	80%	80%	80%	80%
31	358	358	355	360	355	347	380	381	382	382	379	0	81%	81%	81%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	80%	80%	80%	80%	80%	80%	80%
32	358	358	355	360	355	347	380	381	382	382	379	0	81%	81%	81%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	80%	80%	80%	80%	80%	80%	80%
33	358	358	355	360	355	347	380	381	382	382	379	0	81%	81%	81%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	80%	80%	80%	80%	80%	80%	80%
34	358	358	355	360	355	347	380	381	382	382	379	0	81%	81%	81%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	80%	80%	80%	80%	80%	80%	80%
35	358	358	355	360	355	347	380	381	382	382	379	0	81%	81%	81%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	80%	80%	80%	80%	80%	80%	80%
36	358	358	355	360	355	347	380	381	382	382	379	0	81%	81%	81%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	80%	80%	80%	80%	80%	80%	80%
37	358	358	355	360	355	347	380	381	382	382	379	0	81%	81%	81%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	80%	80%	80%	80%	80%	80%	80%
38	358	358	355	360	355	347	380	381	382	382	379	0	81%	81%	81%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	80%	80%	80%	80%	80%	80%	80%
39	358	358	355	360	355	347	380	381	382	382	379	0	81%	81%	81%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	82%	80%	80%	80%	80%	80%	80%	80%
40	358	358	355	360	355	347	380	381	382	382	379	0	81%	81																					

B4.2 Distance Matrix for the p17 matrix domain

p17	01_AB_KAL153-2	06_CPX8FP90	01_AE_CM240	01_AE_B9TH253	02_AG_DJ283	02_AG_DJ284	02_AG_DJ285	04_CPX87PVMY	A_U455	02_AG_D629	CPZ_CP2JUS	CPZ_CAM3	CPZ_CPZANT	O_MVP180						
	02_AB_RUR001	06_CPX85ML84	01_AE_B9TH047	01_AE_B9CF402	02_AG_DJ284	02_AG_DJ285	04_CPX84CY033-3	04_CPX87PVMY	02_AG_D629	CPZ_CP2JUS	CPZ_CAM3	N_Y8F30	O_ANTO							
	80	81	82	83	84	85	86	87	88	89	90	91	92	93						
1	81%	80%	80%	80%	79%	79%	81%	80%	79%	79%	78%	79%	77%	82%	81%	83%	80%	53%	81%	85%
2	81%	81%	80%	80%	80%	79%	81%	81%	79%	79%	78%	79%	77%	82%	81%	83%	80%	53%	81%	85%
3	81%	82%	80%	80%	79%	78%	80%	80%	78%	78%	77%	78%	80%	82%	80%	83%	80%	53%	81%	85%
4	82%	82%	81%	82%	80%	79%	80%	81%	79%	78%	79%	77%	79%	80%	80%	80%	80%	55%	80%	86%
5	81%	81%	80%	81%	80%	80%	81%	81%	78%	78%	78%	80%	77%	82%	80%	80%	80%	54%	81%	85%
6	80%	80%	79%	79%	78%	77%	80%	79%	77%	78%	78%	79%	81%	80%	80%	80%	80%	54%	81%	85%
7	83%	83%	80%	81%	80%	79%	81%	80%	79%	79%	79%	77%	79%	81%	78%	83%	80%	53%	80%	86%
8	83%	83%	81%	81%	81%	80%	82%	80%	80%	80%	79%	80%	81%	78%	83%	80%	80%	54%	80%	86%
9	83%	82%	81%	82%	80%	80%	81%	81%	79%	82%	79%	80%	82%	79%	82%	80%	80%	54%	80%	87%
10	81%	81%	80%	80%	80%	79%	80%	81%	78%	79%	78%	80%	78%	80%	81%	78%	81%	53%	82%	86%
11	82%	82%	81%	81%	80%	80%	81%	80%	79%	78%	79%	80%	79%	82%	80%	80%	80%	54%	81%	87%
12	81%	81%	80%	80%	80%	79%	80%	79%	78%	77%	78%	77%	79%	77%	82%	80%	80%	54%	81%	87%
13	80%	80%	79%	79%	78%	77%	79%	78%	77%	78%	78%	77%	75%	77%	75%	81%	50%	82%	87%	84%
14	80%	80%	80%	79%	78%	78%	79%	80%	78%	77%	78%	79%	77%	81%	80%	80%	80%	53%	81%	85%
15	81%	80%	80%	80%	79%	78%	79%	80%	78%	78%	78%	79%	78%	82%	80%	80%	80%	53%	82%	85%
16	80%	80%	80%	80%	78%	77%	78%	79%	78%	78%	77%	78%	79%	77%	83%	80%	80%	53%	82%	85%
17	81%	81%	79%	80%	79%	78%	80%	79%	79%	77%	78%	78%	80%	78%	82%	80%	80%	53%	82%	86%
18	81%	81%	80%	80%	79%	78%	80%	79%	79%	78%	78%	79%	80%	78%	83%	80%	80%	54%	79%	87%
19	81%	80%	80%	80%	79%	79%	80%	79%	79%	78%	78%	77%	78%	80%	77%	82%	80%	54%	80%	86%
20	81%	81%	79%	80%	79%	78%	80%	79%	78%	77%	78%	79%	80%	77%	82%	80%	80%	53%	81%	84%
21	81%	81%	80%	80%	79%	78%	80%	79%	78%	77%	78%	78%	80%	77%	81%	80%	80%	53%	80%	85%
22	81%	81%	80%	80%	79%	78%	80%	79%	78%	77%	78%	78%	80%	77%	81%	80%	80%	53%	81%	85%
23	80%	80%	79%	79%	80%	80%	81%	79%	78%	77%	78%	80%	78%	81%	81%	80%	80%	52%	82%	85%
24	83%	82%	81%	80%	80%	81%	81%	79%	80%	78%	80%	78%	79%	80%	80%	80%	80%	53%	82%	86%
25	83%	82%	80%	80%	80%	81%	81%	79%	79%	78%	80%	78%	79%	80%	80%	80%	80%	53%	82%	86%
26	82%	82%	81%	81%	80%	80%	81%	82%	80%	79%	80%	79%	82%	78%	82%	80%	80%	54%	83%	86%
27	81%	81%	79%	80%	78%	77%	80%	80%	77%	78%	77%	79%	79%	77%	81%	80%	80%	53%	80%	87%
28	82%	81%	79%	79%	78%	78%	80%	80%	78%	78%	78%	79%	80%	77%	81%	80%	80%	53%	80%	87%
29	80%	79%	79%	79%	79%	80%	81%	78%	78%	78%	79%	78%	79%	77%	82%	80%	80%	54%	82%	81%
30	83%	82%	81%	82%	81%	80%	83%	80%	79%	79%	80%	80%	81%	80%	79%	79%	82%	54%	80%	82%
31	82%	81%	80%	81%	81%	80%	82%	81%	78%	79%	80%	81%	81%	82%	79%	78%	83%	54%	80%	81%
32	80%	79%	79%	80%	78%	78%	78%	78%	78%	77%	77%	78%	78%	77%	80%	80%	80%	54%	80%	81%
33	85%	84%	83%	84%	84%	83%	84%	84%	82%	82%	82%	84%	82%	82%	80%	84%	82%	60%	80%	84%
34	84%	83%	81%	82%	82%	81%	82%	82%	81%	81%	83%	80%	82%	83%	79%	83%	82%	60%	80%	85%
35	83%	82%	81%	82%	82%	81%	83%	81%	81%	82%	81%	82%	81%	82%	81%	80%	80%	54%	80%	82%
36	84%	83%	82%	82%	80%	79%	81%	81%	80%	80%	80%	80%	79%	80%	80%	78%	80%	54%	80%	83%
37	84%	84%	81%	82%	81%	79%	82%	81%	81%	80%	80%	80%	79%	80%	80%	80%	80%	54%	80%	83%
38	87%	86%	84%	84%	83%	83%	84%	82%	82%	83%	83%	81%	83%	84%	81%	81%	80%	60%	80%	84%
39	85%	84%	82%	83%	82%	81%	82%	81%	81%	82%	80%	82%	84%	81%	81%	80%	80%	60%	80%	84%
40	85%	85%	82%	83%	82%	81%	83%	82%	81%	81%	81%	82%	82%	82%	79%	80%	80%	60%	80%	84%
41	85%	84%	82%	82%	83%	83%	84%	82%	81%	81%	81%	82%	82%	81%	82%	80%	80%	60%	80%	82%

B4.2 Distance Matrix for the p17 matrix domain

p17

	TV004024	TV00104	C_9B6W17B03C	TV006038	TV018040	TV006086	TV012040	C_9B5N1008	TV019020	C_9B6W602C	TV013015	TV006087	TV003015	C_9B2H025	C_9B2H063													
TV004017	TV008012	TV010011	TV006028	TV00208	TV006085	TV010074	TV012034	TV019025	TV007058	TV01302	TV014073	TV006011	C_9B2H025	C_9B2H063														
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
42	528	328	330	331	330	321	326	327	327	328	328	328	328	328	328	328	328	328	328	328	328	328	328	328	328	328	328	328
43	325	327	328	328	328	323	324	335	335	335	335	335	335	335	335	335	335	335	335	335	335	335	335	335	335	335	335	335
44	328	330	329	328	328	327	325	330	333	335	330	334	330	332	328	332	329	328	330	335	330	329	329	327	327	327	327	327
45	319	321	318	325	324	318	319	320	324	317	326	324	319	317	323	323	324	328	325	322	320	318	325	325	320	317	317	311
46	329	329	324	334	335	327	330	331	327	328	336	331	330	321	325	329	325	335	335	329	333	327	334	328	324	328	318	318
47	325	327	325	328	328	320	325	326	328	321	334	329	325	320	324	324	328	330	330	323	325	327	326	327	327	326	325	314
48	329	329	327	327	328	320	328	327	325	324	329	325	324	323	327	325	329	332	331	327	333	333	327	328	326	325	322	309
49	308	308	305	310	305	304	308	310	307	304	310	306	299	301	303	305	315	317	313	308	310	312	307	309	308	311	304	309
50	314	316	311	315	310	307	310	312	314	309	315	311	305	306	308	309	317	318	315	313	315	317	311	315	315	320	308	311
51	312	314	314	318	313	308	312	314	308	309	315	311	305	310	310	313	315	318	315	318	315	315	317	311	312	312	320	317
52	309	309	305	317	313	304	309	311	311	313	313	309	304	307	310	306	313	313	315	311	311	313	307	314	314	315	309	312
53	303	301	300	307	305	300	308	308	310	303	310	306	298	300	300	303	312	308	307	305	304	306	299	308	304	305	301	291
54	321	321	320	323	322	317	321	322	319	322	327	323	312	314	318	320	329	323	320	322	319	324	322	323	314	319	302	302
55	313	315	311	316	315	312	307	307	308	307	311	307	301	308	310	314	315	316	319	318	310	310	308	308	307	303	306	296
56	310	312	305	309	309	308	307	309	308	305	313	311	301	304	308	308	316	317	314	313	311	313	308	312	312	315	308	313
57	313	315	312	320	318	308	315	317	322	313	316	312	310	315	316	311	311	312	314	314	316	316	310	320	320	313	316	319
58	324	326	323	329	325	322	328	330	331	324	328	322	322	323	324	321	325	324	325	328	329	329	322	329	329	325	323	326
59	315	317	315	318	318	311	319	322	322	318	318	315	306	319	318	315	315	314	314	317	319	319	311	320	320	318	317	318
60	321	323	324	328	323	320	330	332	330	323	327	323	319	320	321	319	322	323	321	322	329	329	319	330	330	327	324	328
61	320	322	325	325	322	319	329	331	329	322	326	324	320	320	319	321	322	320	321	322	320	321	327	317	328	328	322	323
62	317	319	318	324	320	314	320	322	322	320	321	317	314	317	318	320	318	317	318	318	322	322	314	321	320	323	318	317
63	318	320	319	325	321	315	323	324	326	320	321	319	317	318	317	317	317	318	319	318	325	325	316	320	320	323	319	318
64	315	317	314	317	318	311	318	321	321	317	320	318	310	315	316	310	314	315	316	313	319	319	318	318	318	319	312	317
65	315	315	311	315	317	308	315	318	319	314	317	315	308	311	312	308	310	311	313	311	315	315	317	317	317	318	309	313
66	321	323	320	320	321	317	323	326	322	320	322	320	316	315	316	312	318	319	319	317	324	324	318	324	324	323	318	319
67	320	322	318	322	321	314	317	320	323	322	320	318	312	318	319	315	316	317	318	315	321	321	321	322	322	325	318	319
68	314	318	312	315	312	308	315	317	321	312	317	315	307	310	310	310	313	315	315	312	315	315	314	315	313	317	306	313
69	314	318	310	314	311	309	315	317	322	314	315	311	305	309	310	312	314	315	314	317	317	317	317	315	317	315	317	314
70	310	312	306	312	309	306	313	313	317	310	312	308	303	306	306	306	308	308	310	308	312	312	309	310	310	313	306	312
71	313	318	312	316	317	311	316	318	328	317	314	312	308	312	313	312	313	314	313	316	317	317	316	319	319	320	317	311
72	304	306	306	306	306	304	307	309	312	308	312	308	301	305	305	305	303	304	305	306	306	305	306	305	309	315	303	305
73	310	312	310	316	315	310	315	317	321	318	315	315	308	312	314	312	312	313	312	318	315	315	310	316	318	315	314	314
74	308	310	313	313	312	307	316	318	323	317	315	311	303	306	310	309	314	313	315	314	316	318	313	315	315	319	310	314
75	305	307	310	305	305	301	309	311	315	309	314	307	299	305	308	304	308	309	305	306	306	310	306	310	310	308	307	305
76	257	259	260	258	257	254	259	261	256	253	256	256	252	252	256	260	257	258	258	256	254	254	252	256	256	252	251	257
77	257	259	264	256	253	248	253	258	253	253	254	252	249	255	253	256	256	254	252	256	256	254	252	256	256	254	243	245
78	258	261	265	265	271	260	260	262	262	260	262	260	263	264	262	262	262	262	262	262	262	262	262	262	262	265	261	258
79	280	282	287	281	278	273	274	278	279	275	278	275	274	278	278	280	281	282	281	274	278	278	278	275	284	282	285	272
80	240	241	235	248	241	235	240	244	245	238	244	243	234	237	239	238	238	238	238	244	244	236	238	238	245	238	240	235
81	243	242	254	247	245	243	245	246	250	248	245	240	243	243	244	244	244	245	242	243	250	245	244	244	244	248	237	240
82	255	256	269	262	264	256	267	268	266	261	265	264	255	255	257	258	262	263	260	253	262	262	257	261	261	259	250	246
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29

B4.2 Distance Matrix for the p17 matrix domain

p17

	01_AB_KAL153.2	02_CPX_BFP90	01_AE_C4040	01_AE_BST053	02_AG_D2063	02_AG_BHG0	01_CPX_B7PVNY	A_U455	CPZ_C2P28	CPZ_C4M3	CPZ_C2P21	O_MVPS160
	01_AB_R80001	01_CPX_B9M44	01_AE_BST0407	01_AE_BSC042	02_AG_D2064	01_CPX_B4CV032.3	01_CPX_B7PVCH	02_AG_G429	CPZ_C2P28B	N_YBF30	O_ANT70	
80	81	82	83	84	85	86	87	88	89	90	91	92
82	85	84%	82%	83%	82%	81%	82%	80%	81%	80%	80%	82%
42	85%	84%	82%	83%	82%	81%	82%	80%	81%	80%	80%	82%
	0%	1%	0%	1%	0%	0%	2%	2%	0%	2%	0%	3%
43	87%	87%	84%	85%	86%	85%	84%	83%	84%	82%	84%	86%
	87%	87%	84%	85%	86%	85%	84%	83%	84%	82%	84%	86%
	0%	1%	0%	1%	0%	0%	2%	2%	0%	2%	0%	3%
44	85%	85%	84%	84%	84%	83%	85%	82%	82%	83%	80%	84%
	85%	85%	84%	84%	84%	83%	85%	82%	82%	83%	80%	84%
	0%	0%	1%	0%	0%	0%	2%	2%	0%	2%	0%	3%
45	83%	82%	80%	81%	82%	81%	82%	80%	81%	79%	80%	82%
	83%	82%	80%	81%	82%	81%	82%	80%	81%	79%	80%	82%
	0%	2%	3%	2%	2%	2%	5%	5%	2%	1%	2%	3%
46	85%	84%	82%	83%	82%	84%	83%	82%	82%	81%	81%	80%
	85%	84%	82%	83%	82%	84%	83%	82%	82%	81%	81%	80%
	1%	1%	2%	1%	1%	1%	3%	3%	1%	2%	1%	2%
47	84%	84%	82%	82%	83%	82%	83%	81%	81%	79%	81%	80%
	84%	84%	82%	82%	83%	82%	83%	81%	81%	79%	81%	80%
	1%	1%	2%	1%	1%	1%	3%	3%	1%	2%	1%	2%
48	84%	83%	83%	84%	82%	81%	83%	83%	81%	81%	79%	82%
	84%	83%	83%	84%	82%	81%	83%	83%	81%	81%	79%	82%
	0%	0%	1%	0%	0%	0%	2%	2%	0%	2%	0%	3%
49	81%	80%	79%	79%	79%	79%	81%	79%	82%	81%	81%	79%
	81%	80%	79%	79%	79%	79%	81%	79%	82%	81%	81%	79%
	3%	3%	4%	3%	3%	3%	1%	1%	1%	3%	3%	2%
50	81%	80%	78%	79%	79%	81%	79%	81%	80%	80%	78%	79%
	81%	80%	78%	79%	79%	81%	79%	81%	80%	80%	78%	79%
	4%	4%	5%	4%	4%	4%	2%	2%	2%	4%	4%	3%
51	81%	80%	81%	80%	80%	79%	82%	80%	81%	80%	78%	77%
	81%	80%	81%	80%	80%	79%	82%	80%	81%	80%	78%	77%
	3%	3%	4%	3%	3%	3%	1%	1%	1%	3%	3%	2%
52	80%	80%	78%	78%	79%	78%	81%	79%	80%	80%	78%	78%
	80%	80%	78%	78%	79%	78%	81%	79%	80%	80%	78%	78%
	3%	3%	4%	3%	3%	3%	1%	1%	1%	3%	3%	2%
53	80%	79%	79%	79%	79%	79%	80%	80%	77%	78%	77%	81%
	80%	79%	79%	79%	79%	79%	80%	80%	77%	78%	77%	81%
	3%	3%	4%	3%	3%	3%	1%	1%	3%	3%	2%	2%
54	81%	80%	79%	80%	80%	80%	81%	80%	78%	78%	77%	79%
	81%	80%	79%	80%	80%	80%	81%	80%	78%	78%	77%	79%
	0%	0%	1%	0%	0%	0%	2%	2%	0%	2%	0%	3%
55	78%	77%	78%	77%	79%	78%	79%	80%	79%	77%	75%	78%
	78%	77%	78%	77%	79%	78%	79%	80%	79%	77%	75%	78%
	3%	3%	4%	3%	3%	3%	1%	1%	3%	3%	2%	2%
56	81%	80%	80%	79%	81%	81%	83%	82%	82%	81%	80%	77%
	81%	80%	80%	79%	81%	81%	83%	82%	82%	81%	80%	77%
	3%	3%	4%	3%	3%	3%	1%	1%	3%	3%	2%	2%
57	81%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%
	81%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%
	1%	1%	1%	1%	1%	1%	3%	3%	1%	2%	1%	4%
58	82%	81%	80%	80%	80%	80%	81%	80%	81%	80%	81%	80%
	82%	81%	80%	80%	80%	80%	81%	80%	81%	80%	81%	80%
	0%	0%	0%	0%	0%	0%	2%	2%	0%	1%	0%	3%
59	82%	81%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%
	82%	81%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%
	0%	0%	0%	0%	0%	0%	2%	2%	0%	1%	0%	3%
60	80%	80%	81%	81%	80%	80%	80%	80%	80%	80%	80%	80%
	80%	80%	81%	81%	80%	80%	80%	80%	80%	80%	80%	80%
	0%	0%	0%	0%	0%	0%	2%	2%	0%	1%	0%	3%
61	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%
	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%
	0%	0%	0%	0%	0%	0%	2%	2%	0%	1%	0%	3%
62	81%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%
	81%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%
	3%	3%	4%	3%	3%	3%	1%	1%	3%	3%	2%	2%
63	82%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%
	82%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%	80%
	0%	0%	0%	0%	0%	0%	2%	2%	0%	1%	0%	3%
64	85%	85%	85%	84%	84%	85%	84%	84%	84%	84%	84%	84%
	85%	85%	85%	84%	84%	85%	84%	84%	84%	84%	84%	84%
	0%	0%	0%	0%	0%	0%	2%	2%	0%	1%	0%	3%
65	85%	85%	85%	84%	84%	85%	84%	84%	84%	84%	84%	84%
	85%	85%	85%	84%	84%	85%	84%	84%	84%	84%	84%	84%
	0%	0%	0%	0%	0%	0%	2%	2%	0%	1%	0%	3%
66	85%	85%	85%	84%	84%	85%	84%	84%	84%	84%	84%	84%
	85%	85%	85%	84%	84%	85%	84%	84%	84%	84%	84%	84%
	0%	0%	0%	0%	0%	0%	2%	2%	0%	1%	0%	3%
67	85%	85%	85%	84%	84%	85%	84%	84%	84%	84%	84%	84%
	85%	85%	85%	84%	84%	85%	84%	84%	84%	84%	84%	84%
	0%	0%	0%	0%	0%	0%	2%	2%	0%	1%	0%	3%
68	85%	85%	85%	84%	84%	85%	84%	84%	84%	84%	84%	84%
	85%	85%	85%	84%	84%	85%	84%	84%	84%	84%	84%	84%
	0%	0%	0%	0%	0%	0%	2%	2%	0%	1%	0%	3%
69	85%	85%	85%	84%	84%	85%	84%	84%	84%	84%	84%	84%
	85%	85%	85%	84%	84%	85%	84%	84%	84%	84%	84%	84%
	0%	0%	0%	0%	0%	0%	2%	2%	0%	1%	0%	3%
70	85%	85%	85%	84%	84%	85%	84%	84%	84%	84%	84%	84%
	85%	85%	85%	84%	84%	85%	84%	84%	84%	84%	84%	84%
	0%	0%	0%	0%	0%	0%	2%	2%	0%	1%	0%	3%
71	85%	85%	85%	84%	84%	85%	84%	84%	84%	84%	84%	84%
	85%	85%	85%	84%	84%	85%	84%	84%	84%	84%	84%	84%
	0%	0%	0%	0%	0%	0%	2%	2%	0%	1%	0%	3%
72	85%	85%	85%	84%	84%	85%	84%	84%	84%	84%	84%	84%
	85%	85%	85%	84%	84%	85%	84%	84%	84%	84%	84%	84%
	0%	0%	0%	0%	0%	0%	2%	2%	0%	1%	0%	3%
73	85%	85%	85%	84%	84%	85%	84%	84%	84%	84%	84%	84%
	85%	85%	85%	84%	84%	85%	84%	84%	84%	84%	84%	84%
	0%	0%	0%	0%	0%	0%	2%	2%	0%	1%	0%	3%
74	85%	85%	85%	84%	84%	85%	84%	84%	84%	84%	84%	84%
	85%	85%	85%	84%	84%	85%	84%	84%	84%	84%	84%	84%
	0%	0%	0%	0%	0%	0%	2%	2%	0%	1%	0%	3%
75	85%	85%	85%	84%	84%	85%	84%	84%	84%	84%	84%	84%
	85%	85%	85%	84%	84%	85%	84%	84%	84%	84%	84%	84%
	0%	0%	0%	0%	0%	0%	2%	2%	0%	1%	0%	3%
76	85%	85%	85%	84%	84%	85%	84%	84%	84%	84%	84%	84%
	85%	85%	85%	84%	84%	85%	84%	84%	84%	84%	84%	84%
	0%	0%	0%	0%	0%	0%	2%	2%	0%	1%	0%	3%
77	85%	85%	85%	84%	84%	85%	84%	84%	84%	84%	84%	84%
	85%	85%	85%	84%	84%	85%	84%	84%	84%	84%	84%	84%
	0%	0%	0%	0%	0%	0%	2%	2%	0%	1%	0%	3%
78	85%	85%	85%	84%	84%	85%	84%	84%	84%	84%	84%	84%
	85%	85%	85%	84%	84%	85%	84%	84%	84%	84%	84%	84%
	0%	0%	0%	0%	0%	0%	2%	2%	0%	1%	0%	3%
79	85%	85%	85%	84%	84%	85%	84%	84%	84%	84%	84%	84%
	85%	85%	85%	84%	84%	85%	84%	84%	84%	84%	84%	84%
	0%	0%	0%	0%	0%	0%	2%	2%	0%	1%	0%	3%
80	85%	85%	85%	84%	84%	85%	84%	84%	84%	84%	84%	84%
	85%	85%	85%	84%	84%	85%	84%	84%	84%	84%	84%	84%
	0%	0%	0%	0%	0%	0%	2%	2%	0%	1%	0%	3%
81	85%	85%	85%	84%	84%	85%	84%	84%	84%	84%	84%	84%
	85%	85%	85%	84%	84%	85%	84%	84%	84%	84%	84%	84%
	0%	0%	0%	0%	0%	0%	2%	2%	0%	1%	0%	3%
82	85%	85%	85%	84%	84%	85%	84%	84%	84%	84%	84%	84%
	85%	85%	85%	84%	84%	85%	84%	84%	84%	84%	84%	84%
	0%	0%	0%	0%	0%	0%	2%	2%	0%	1%	0%	3%
83	85%	85%	85%	84%	84%	85%	84%	84%	84%	84%	84%	84%
	85%	85%	85%	84%	84%	85%	84%	84%	84%	84%	84%	84%
	0%	0%	0%	0%	0%	0%	2%	2%	0%	1%	0%	3%
84	85%	85%	85%	84%	84%	85%	84%	84%	84%	84%	84%	84%
	85%	85%	85%	84%	84%	85%	84%	84%	84%	84%	84%	84%
	0%	0%	0%	0%	0%	0%	2%	2%	0%	1%	0%	3%

B4.3 Distance Matrix for the p24 capsid domain

824

	03_AB_RU8001		01_AE_CM240		01_AE_95TH047		02_AG_DJ283		02_AG_IBNG		04_CPK_87PVMY		02_AG_GA29		06_CPK_95MLM4		G_S68165		G_DRCLB		CPZ_CPZANT	
	03_AB_KA1153-2		U_A455		01_AE_95TH053		01_AE_95RCF02		02_AG_DJ264		04_CPK_84CY032-3		04_CPK_87PVMY		06_CPK_BFP90		G_82H5063		G_H48783-1.1		N_V8F30	
	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82
1	75%	75%	74%	75%	75%	75%	74%	75%	75%	74%	76%	76%	75%	75%	77%	75%	76%	75%	76%	75%	76%	71%
2	74%	75%	74%	74%	74%	74%	73%	74%	74%	73%	73%	74%	74%	75%	75%	75%	77%	75%	76%	75%	76%	69%
3	75%	75%	76%	76%	76%	76%	75%	75%	75%	75%	73%	76%	76%	77%	75%	74%	73%	74%	75%	76%	71%	71%
4	76%	77%	76%	77%	76%	76%	76%	77%	76%	77%	76%	76%	78%	78%	77%	77%	76%	76%	76%	76%	77%	70%
5	75%	76%	77%	76%	75%	76%	75%	75%	76%	76%	75%	76%	75%	77%	77%	77%	77%	76%	77%	77%	73%	73%
6	87%	87%	87%	87%	87%	87%	86%	89%	89%	86%	87%	86%	88%	87%	89%	88%	88%	88%	88%	87%	78%	70%
7	87%	86%	87%	87%	87%	87%	86%	89%	89%	86%	87%	86%	88%	87%	89%	88%	88%	88%	88%	87%	78%	71%
8	87%	86%	86%	87%	87%	87%	86%	89%	89%	86%	87%	86%	88%	87%	89%	88%	88%	88%	88%	87%	78%	72%
9	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	79%	71%
10	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	79%	71%
11	87%	86%	86%	86%	87%	87%	86%	89%	89%	86%	87%	86%	88%	87%	89%	88%	88%	88%	88%	87%	78%	71%
12	86%	87%	87%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	79%	71%
13	86%	87%	87%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	79%	71%
14	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	78%	71%
15	87%	87%	87%	87%	87%	87%	86%	89%	89%	86%	87%	86%	88%	87%	89%	88%	88%	88%	87%	88%	87%	78%
16	87%	87%	87%	87%	87%	87%	86%	89%	89%	86%	87%	86%	88%	87%	89%	88%	88%	88%	87%	88%	87%	78%
17	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	79%	71%
18	87%	87%	86%	86%	86%	87%	87%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	78%	71%
19	87%	87%	87%	87%	87%	87%	86%	89%	89%	86%	87%	86%	88%	87%	89%	88%	88%	88%	87%	88%	87%	72%
20	87%	87%	86%	87%	86%	87%	86%	87%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	77%	71%
21	89%	89%	88%	88%	87%	88%	88%	89%	90%	89%	88%	87%	89%	89%	88%	88%	88%	88%	88%	88%	79%	71%
22	87%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	79%	71%
23	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	78%	71%
24	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	78%	71%
25	89%	89%	88%	88%	88%	88%	88%	89%	90%	89%	88%	87%	89%	89%	88%	88%	88%	88%	88%	88%	78%	71%
26	88%	88%	88%	88%	87%	88%	88%	89%	90%	89%	88%	87%	89%	89%	88%	88%	88%	88%	88%	88%	79%	71%
27	87%	86%	86%	87%	87%	87%	86%	89%	89%	86%	87%	86%	88%	87%	89%	88%	88%	88%	88%	88%	79%	71%
28	87%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	79%	72%
29	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	79%	72%
30	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	78%	72%
31	87%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	78%	71%
32	87%	86%	86%	87%	87%	87%	86%	89%	89%	86%	87%	86%	88%	87%	89%	88%	88%	88%	88%	88%	78%	72%
33	87%	86%	86%	87%	87%	87%	86%	89%	89%	86%	87%	86%	88%	87%	89%	88%	88%	88%	88%	88%	79%	72%
34	89%	89%	88%	88%	88%	88%	88%	89%	90%	89%	88%	87%	89%	89%	88%	88%	88%	88%	88%	88%	78%	72%
35	87%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	78%	72%
36	85%	85%	86%	87%	86%	87%	87%	86%	87%	86%	87%	86%	86%	86%	86%	86%	86%	86%	86%	86%	87%	72%
37	87%	87%	87%	86%	86%	86%	86%	87%	86%	86%	87%	86%	86%	86%	86%	86%	86%	86%	86%	86%	87%	72%
38	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	79%	72%
39	86%	86%	86%	86%	87%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	86%	78%	72%
40	89%	89%	88%	88%	87%	88%	88%	89%	90%	89%	88%	87%	89%	89%	88%	88%	88%	88%	88%	88%	78%	72%
41	86%	86%	87%	87%	87%	87%	87%	86%	86%	86%	87%	86%	86%	86%	86%	86%	86%	86%	86%	86%	89%	71%

B4.3 Distance Matrix for the p24 capsid domain

p24

03_AB_RUM001 01_AE_CMG40 01_AE_95TH047 02_AG_DJ063 02_AG_018WG 04_CPX_97PVVWY 02_AG_G039 06_CPX_95ML84 G_SEE185 G_DRCBL CPZ_CPZANT

03_AB_R153-2	A_U455	01_AE_95TH053	01_AE_90CF402	02_AG_DJ064	04_CPX_94C0332-3	04_CPX_97PVVCH	06_CPX_BFP90	G_92NG083	G_H48785-1-1	N_Y8F30											
61	62	63	64	65	66	70	71	72	73	74	75	76	77	78	79	80	81	82			
42	86%	86%	87%	87%	87%	86%	86%	86%	87%	87%	86%	86%	86%	86%	86%	86%	86%	77%	72%		
	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%		
43	89%	89%	89%	89%	88%	89%	89%	89%	89%	89%	89%	89%	89%	89%	89%	89%	89%	79%	72%		
	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%		
44	86%	86%	86%	86%	86%	86%	86%	86%	86%	87%	86%	86%	86%	86%	86%	86%	86%	79%	72%		
	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%		
45	86%	86%	86%	86%	87%	86%	86%	86%	87%	87%	86%	87%	86%	87%	87%	87%	87%	78%	73%		
	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%		
46	86%	87%	86%	86%	86%	86%	87%	87%	86%	85%	86%	86%	86%	86%	86%	86%	86%	78%	73%		
	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%		
47	87%	87%	87%	87%	87%	87%	86%	86%	86%	87%	86%	86%	86%	86%	87%	86%	86%	78%	73%		
	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%		
48	88%	88%	88%	88%	88%	88%	88%	88%	88%	88%	88%	88%	88%	88%	88%	88%	88%	78%	73%		
	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%		
49	88%	89%	89%	89%	88%	88%	89%	89%	89%	89%	88%	88%	89%	89%	89%	89%	89%	79%	73%		
	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%		
50	89%	89%	89%	89%	88%	88%	89%	89%	89%	89%	88%	89%	89%	89%	89%	89%	89%	79%	73%		
	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%		
51	88%	88%	88%	87%	87%	88%	88%	88%	88%	87%	88%	88%	88%	88%	88%	88%	88%	78%	72%		
	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%		
52	87%	86%	87%	87%	87%	87%	86%	86%	86%	87%	87%	86%	86%	86%	87%	86%	86%	78%	72%		
	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%		
53	88%	88%	89%	88%	88%	88%	88%	88%	88%	87%	87%	88%	88%	88%	88%	88%	88%	78%	72%		
	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%		
54	88%	89%	89%	88%	88%	88%	89%	89%	89%	88%	88%	89%	89%	89%	89%	89%	89%	79%	73%		
	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%		
55	88%	88%	89%	89%	89%	89%	89%	89%	89%	88%	88%	89%	89%	89%	89%	89%	89%	78%	73%		
	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%		
56	89%	89%	89%	89%	88%	88%	89%	89%	89%	89%	88%	89%	89%	89%	89%	89%	89%	78%	72%		
	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%		
57	88%	88%	89%	88%	87%	87%	88%	88%	88%	87%	88%	88%	88%	88%	88%	88%	88%	78%	72%		
	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%		
58	93%	93%	93%	91%	91%	91%	92%	92%	92%	91%	91%	92%	92%	91%	90%	89%	88%	77%	71%		
	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%		
59	92%	92%	92%	91%	90%	90%	91%	92%	92%	91%	91%	90%	92%	91%	92%	91%	89%	88%	87%	72%	
	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%		
60	93%	94%	94%	91%	91%	91%	92%	93%	93%	93%	91%	91%	92%	91%	93%	92%	89%	89%	88%	77%	71%
	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	
61	90%	90%	93%	91%	91%	91%	92%	93%	93%	90%	90%	91%	90%	92%	92%	88%	88%	87%	77%	72%	
	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	
62	90%	90%	93%	92%	91%	92%	91%	93%	93%	91%	91%	92%	91%	92%	92%	89%	89%	87%	78%	72%	
	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	
63	94%	94%	94%	92%	92%	92%	93%	93%	93%	91%	91%	92%	91%	92%	92%	89%	89%	88%	77%	72%	
	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	
64	93%	93%	94%	92%	92%	92%	93%	94%	94%	92%	92%	93%	91%	91%	90%	88%	87%	87%	78%	72%	
	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	
65	93%	93%	93%	92%	92%	92%	93%	93%	93%	92%	92%	93%	90%	90%	87%	86%	86%	86%	78%	72%	
	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	
66	93%	93%	94%	92%	92%	92%	93%	93%	93%	90%	91%	90%	88%	87%	87%	87%	87%	78%	72%		
	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	
67	93%	93%	94%	92%	92%	92%	93%	93%	93%	91%	91%	91%	91%	91%	87%	87%	88%	87%	78%	71%	
	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	
68	94%	94%	94%	93%	93%	93%	94%	94%	94%	92%	92%	92%	92%	92%	89%	88%	88%	88%	77%	72%	
	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	
69	94%	94%	94%	93%	93%	93%	94%	94%	94%	92%	92%	92%	92%	92%	89%	88%	88%	88%	77%	72%	
	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	
70	94%	94%	94%	93%	93%	93%	94%	94%	94%	92%	92%	92%	92%	92%	89%	88%	88%	88%	77%	72%	
	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	
71	93%	93%	93%	92%	92%	92%	93%	93%	93%	91%	92%	91%	89%	89%	86%	86%	86%	86%	78%	71%	
	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	
72	93%	93%	93%	92%	92%	92%	93%	93%	93%	91%	90%	88%	88%	88%	86%	86%	86%	87%	76%	70%	
	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	
73	93%	93%	94%	92%	92%	92%	93%	93%	93%	91%	91%	91%	91%	91%	87%	87%	88%	87%	78%	71%	
	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	
74	93%	93%	93%	92%	92%	92%	93%	93%	93%	91%	90%	88%	88%	88%	86%	86%	86%	86%	78%	71%	
	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	
75	94%	94%	93%	92%	92%	92%	93%	93%	93%	91%	90%	88%	88%	88%	86%	86%	86%	86%	79%	72%	
	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	
76	94%	94%	93%	92%	92%	92%	93%	93%	93%	91%	90%	88%	88%	88%	86%	86%	86%	86%	79%	72%	
	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	
77	91%	91%	92%	91%	90%	91%	90%	92%	92%	90%	89%	87%	87%	87%	84%	84%	84%	84%	78%	73%	
	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	
78	91%	91%	92%	91%	90%	91%	90%	92%	92%	90%	89%	87%	87%	87%	84%	84%	84%	84%	78%	72%	
	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	
79	91%	91%	92%	91%	90%	91%	90%	92%	92%	90%	89%	87%	87%	87%	84%	84%	84%	84%	78%	72%	
	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	
80	90%	90%	91%	90%	90%	90%	91%	91%	91%	89%	88%	86%	86%	86%	84%	84%	84%	84%	77%	72%	
	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	
81	93%	93%	93%	92%	92%	92%	93%	93%	93%	91%	90%	88%	88%	88%	86%	86%	86%	86%	79%	72%	
	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	
82	90%	90%	91%	90%	90%	90%	91%	91%	91%	89%	88%	86%	86%	86%	84%	84%	84%	84%	78%	72%	
	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	
81	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82

B4.4 Distance Matrix for the region from the nucleocapsid to the p1 spacer peptide

p2-01

	B.WE180		J.SET847		J.SET022		D.NCK		D.R4UG1141		O5_DF_V1310		O4_CPX84Y0323		O4_CPX87P7CH		G.H48785-1		G.SEE185		O6_CPX85M34		F1.FIN053		F1.M411		F2.MP257		H.OCF056		K.EGT811C	
TV018095	B.US.RF	J.SET022	D.NCK	D.R4UG1141	O5_DF_V1310	O4_CPX84Y0323	O4_CPX87P7CH	G.H48785-1	G.SEE185	O6_CPX85M34	F1.FIN053	F1.M411	F2.MP257	H.OCF056	K.EGT811C																	
31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60			
1	78%	78%	78%	77%	78%	78%	78%	77%	78%	77%	78%	78%	77%	78%	78%	77%	78%	77%	78%	77%	78%	77%	79%	77%	79%	79%	79%	78%	78%	78%		
2	76%	75%	72%	76%	76%	75%	76%	74%	75%	74%	75%	76%	75%	77%	77%	75%	76%	73%	77%	76%	77%	76%	77%	74%	78%	76%	76%	75%	78%	74%		
3	80%	79%	78%	79%	79%	79%	80%	79%	80%	79%	78%	77%	78%	79%	80%	80%	79%	80%	81%	79%	82%	80%	81%	79%	82%	80%	80%	80%	80%	78%		
4	80%	79%	78%	79%	80%	79%	81%	79%	80%	79%	78%	77%	78%	79%	80%	80%	78%	80%	77%	83%	80%	82%	80%	82%	80%	80%	80%	80%	80%	78%		
5	80%	79%	77%	80%	80%	79%	80%	78%	79%	79%	78%	77%	79%	79%	80%	81%	80%	79%	79%	77%	82%	81%	81%	80%	82%	80%	80%	80%	80%	78%		
6	79%	78%	75%	75%	75%	75%	74%	75%	74%	74%	74%	74%	74%	74%	74%	75%	76%	75%	75%	75%	77%	77%	77%	74%	78%	77%	76%	75%	76%	75%		
7	76%	76%	74%	74%	74%	75%	74%	75%	73%	73%	73%	74%	74%	75%	76%	76%	74%	76%	73%	77%	76%	76%	73%	77%	76%	76%	75%	76%	74%	74%		
8	78%	78%	75%	78%	78%	75%	78%	74%	75%	74%	74%	74%	74%	74%	74%	77%	78%	78%	77%	77%	75%	78%	78%	77%	77%	78%	78%	77%	76%	78%	73%	
9	81%	80%	78%	79%	79%	78%	80%	78%	79%	77%	77%	79%	79%	80%	79%	81%	80%	79%	80%	77%	81%	80%	80%	79%	81%	81%	81%	80%	81%	78%	81%	
10	81%	80%	79%	78%	79%	78%	80%	77%	79%	77%	77%	79%	79%	80%	79%	81%	80%	79%	80%	77%	81%	80%	80%	79%	81%	81%	81%	80%	80%	77%	81%	
11	85%	84%	82%	82%	82%	82%	82%	80%	81%	80%	79%	79%	80%	79%	79%	80%	80%	79%	79%	78%	82%	81%	81%	80%	82%	82%	82%	82%	80%	81%	81%	
12	83%	83%	81%	81%	81%	82%	82%	80%	82%	80%	79%	82%	81%	81%	83%	83%	82%	83%	81%	84%	82%	84%	81%	84%	83%	83%	83%	83%	83%	81%	81%	
13	82%	82%	80%	82%	81%	82%	83%	81%	81%	80%	79%	81%	80%	80%	82%	82%	81%	82%	80%	83%	82%	83%	82%	83%	81%	83%	82%	82%	82%	80%	80%	
14	82%	82%	80%	82%	81%	82%	83%	81%	81%	80%	79%	81%	80%	80%	82%	82%	81%	82%	80%	83%	82%	83%	82%	83%	81%	83%	82%	82%	82%	80%	80%	
15	77%	77%	75%	75%	75%	77%	77%	76%	76%	74%	75%	75%	76%	77%	74%	76%	75%	75%	77%	75%	78%	77%	77%	77%	77%	77%	76%	77%	76%	74%	74%	
16	78%	77%	76%	78%	78%	77%	78%	76%	77%	77%	78%	75%	76%	76%	76%	76%	75%	75%	76%	74%	78%	78%	77%	77%	77%	77%	77%	77%	77%	74%	74%	
17	78%	77%	76%	77%	77%	77%	78%	75%	76%	75%	75%	76%	76%	77%	77%	76%	76%	77%	76%	75%	78%	78%	77%	75%	78%	78%	77%	76%	77%	74%	74%	
18	79%	79%	77%	80%	81%	79%	80%	78%	80%	77%	76%	77%	77%	78%	80%	79%	78%	79%	77%	82%	81%	80%	78%	82%	80%	81%	79%	80%	76%	76%	76%	
19	83%	82%	81%	81%	81%	82%	83%	80%	82%	82%	82%	82%	82%	81%	84%	83%	81%	83%	80%	85%	83%	83%	82%	84%	83%	83%	83%	82%	83%	81%	81%	
20	77%	76%	74%	74%	75%	76%	74%	75%	74%	73%	74%	73%	74%	75%	75%	76%	76%	75%	76%	73%	77%	76%	76%	74%	76%	76%	76%	75%	75%	73%	73%	
21	76%	76%	74%	74%	74%	74%	74%	74%	74%	72%	74%	74%	74%	74%	74%	75%	74%	75%	73%	77%	76%	76%	74%	76%	76%	76%	75%	75%	73%	73%	73%	
22	83%	81%	80%	82%	82%	80%	81%	79%	81%	81%	79%	79%	80%	80%	79%	81%	80%	80%	80%	77%	84%	82%	83%	81%	83%	82%	83%	82%	81%	79%	79%	
23	81%	80%	79%	79%	79%	80%	80%	79%	80%	78%	78%	79%	80%	80%	79%	80%	80%	79%	80%	77%	81%	80%	82%	79%	81%	80%	81%	80%	79%	77%	77%	
24	82%	80%	79%	79%	79%	80%	79%	80%	79%	80%	79%	80%	81%	81%	80%	81%	81%	80%	80%	78%	82%	81%	83%	80%	81%	81%	82%	81%	80%	78%	78%	
25	80%	78%	77%	79%	79%	78%	79%	79%	78%	78%	78%	78%	78%	78%	78%	80%	80%	79%	79%	77%	81%	80%	82%	80%	81%	79%	81%	80%	80%	78%	78%	
26	81%	80%	78%	80%	81%	80%	80%	79%	80%	79%	79%	80%	80%	80%	81%	81%	81%	80%	79%	82%	81%	82%	80%	82%	81%	82%	81%	82%	81%	78%	78%	
27	82%	81%	80%	81%	81%	81%	82%	80%	82%	80%	79%	79%	80%	80%	79%	82%	82%	80%	81%	80%	84%	82%	81%	80%	83%	82%	82%	82%	81%	79%	79%	
28	81%	81%	79%	81%	81%	81%	82%	80%	80%	80%	80%	82%	82%	81%	84%	84%	82%	84%	80%	84%	82%	84%	83%	84%	83%	84%	82%	83%	82%	84%	81%	
29	86%	82%	81%	85%	85%	85%	86%	85%	85%	83%	83%	79%	78%	81%	79%	80%	79%	79%	79%	78%	81%	81%	81%	81%	81%	81%	81%	81%	81%	81%	81%	
30	86%	84%	81%	86%	86%	86%	86%	85%	85%	84%	84%	83%	80%	79%	81%	79%	79%	79%	79%	79%	81%	82%	81%	82%	81%	82%	84%	82%	82%	83%	83%	
31	413	93%	82%	85%	85%	86%	89%	84%	87%	84%	83%	80%	79%	82%	81%	81%	80%	80%	80%	82%	81%	82%	80%	82%	82%	84%	83%	82%	83%	83%		
32	386	414	86%	84%	84%	87%	86%	84%	84%	83%	83%	80%	79%	81%	80%	80%	80%	80%	80%	81%	81%	80%	79%	80%	81%	83%	81%	83%	82%	82%		
33	389	378	417	84%	84%	86%	86%	85%	87%	84%	83%	78%	78%	80%	78%	79%	79%	77%	80%	79%	81%	81%	81%	81%	81%	81%	81%	81%	81%	81%	81%	
34	353	349	353	405	96%	85%	85%	83%	84%	82%	81%	79%	78%	80%	80%	81%	80%	79%	78%	78%	82%	82%	82%	81%	82%	82%	85%	84%	83%	81%	81%	
35	353	349	351	400	405	86%	85%	83%	83%	82%	81%	79%	78%	80%	80%	82%	80%	79%	79%	78%	82%	82%	82%	81%	82%	82%	85%	84%	83%	81%	81%	
36	369	366	371	354	352	414	83%	80%	80%	80%	81%	79%	81%	80%	81%	80%	81%	82%	81%	82%	81%	82%	81%	82%	81%	84%	83%	83%	82%	82%		
37	373	369	374	354	354	387	414	81%	81%	80%	80%	81%	81%	82%	84%	83%	81%	83%	83%	84%	82%	82%	82%	82%	83%	85%	83%	84%	83%	83%		
38	359	357	364	349	349	375	383	420	88%	86%	84%	77%	77%	79%	80%	80%	80%	80%	78%	81%	78%	81%	79%	80%	80%	81%	80%	82%	81%	82%	82%	
39	369	369	371	354	351	380	385	378	410	87%	86%	80%	81%	82%	81%	82%	81%	82%	81%	82%	81%	82%	81%	82%	81%	84%	83%	83%	82%	82%		
40	356	349	359	344	342	373	372	365	369	417	86%	79%	79%	79%	81%	81%	79%	81%	81%	83%	81%	83%	81%	80%	83%	82%	82%	81%	82%	81%	82%	
41	350	347	352	338	338	365	373	359	364	379	414	78%	80%	80%	80%	82%	81%	81%	81%	81%	83%	81%	83%	81%	82%	82%	81%	82%	82%	80%	80%	

B4.4 Distance Matrix for the region from the nucleocapsid to the p1 spacer peptide

pr-1

	A SET53	A Q23-17	A U455	D1_AE_C0420	D1_AE_R0TH047	D1_AE_KA1153-1	D2_AG_D1203	D2_AG_BHG	CPZ_CAM3	N YBF30	O MVPS180	CPZ_CPZ048										
K MP55	A R2UG037	A U455	D1_AE_R0TH053	D1_AE_K0CF402	D1_AE_RL06001	D2_AG_D1204	D2_AG_G829	CPZ_CPZUS	O ANT10	CPZ_CPZANT												
	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82
1	80%	74%	75%	76%	77%	72%	73%	76%	76%	77%	7%	75%	75%	75%	76%	80%	81%	81%	52%	51%	52%	55%
2	78%	73%	72%	74%	75%	71%	72%	75%	74%	74%	7%	72%	73%	73%	74%	57%	59%	59%	51%	49%	50%	57%
3	82%	77%	76%	78%	79%	74%	74%	78%	78%	77%	7%	78%	78%	78%	78%	80%	81%	82%	54%	53%	54%	59%
4	82%	77%	77%	78%	80%	74%	75%	78%	78%	77%	7%	78%	78%	78%	78%	80%	82%	82%	55%	54%	54%	59%
5	81%	77%	77%	78%	80%	74%	75%	79%	79%	78%	7%	78%	77%	77%	78%	82%	82%	81%	55%	54%	55%	59%
6	79%	73%	73%	74%	75%	71%	71%	74%	75%	74%	7%	72%	73%	72%	74%	56%	60%	60%	52%	52%	51%	57%
7	78%	72%	72%	73%	74%	70%	71%	74%	74%	73%	7%	72%	72%	72%	73%	56%	60%	60%	52%	51%	53%	57%
8	78%	74%	74%	76%	77%	71%	72%	75%	76%	76%	7%	75%	76%	76%	74%	58%	61%	60%	53%	52%	53%	58%
9	81%	76%	76%	77%	78%	75%	76%	79%	79%	78%	7%	77%	77%	77%	77%	81%	84%	82%	54%	53%	54%	60%
10	81%	76%	76%	78%	79%	75%	76%	79%	79%	78%	7%	77%	77%	77%	77%	81%	84%	82%	54%	54%	54%	60%
11	83%	78%	78%	80%	81%	76%	76%	81%	80%	80%	8%	80%	80%	80%	84%	85%	84%	57%	56%	56%	62%	
12	83%	79%	78%	81%	82%	77%	78%	81%	81%	81%	8%	80%	80%	80%	84%	85%	84%	57%	56%	56%	62%	
13	83%	80%	79%	81%	82%	78%	79%	82%	81%	80%	8%	80%	79%	79%	81%	80%	87%	85%	55%	55%	56%	62%
14	83%	80%	79%	81%	82%	78%	79%	82%	81%	80%	8%	80%	79%	79%	81%	80%	87%	85%	55%	54%	56%	62%
15	77%	74%	74%	76%	77%	72%	73%	76%	76%	76%	7%	74%	74%	74%	50%	61%	58%	51%	50%	51%	56%	
16	79%	75%	74%	75%	77%	72%	73%	76%	76%	76%	7%	74%	74%	74%	50%	61%	58%	51%	50%	51%	56%	
17	79%	74%	74%	76%	77%	72%	73%	76%	76%	76%	7%	74%	74%	74%	50%	61%	58%	51%	50%	51%	56%	
18	80%	76%	76%	78%	79%	74%	75%	78%	78%	78%	7%	77%	77%	77%	82%	84%	82%	55%	54%	54%	59%	
19	85%	79%	77%	80%	81%	76%	77%	80%	80%	80%	7%	79%	79%	78%	78%	83%	85%	83%	55%	55%	56%	61%
20	76%	72%	72%	74%	74%	70%	71%	74%	73%	74%	7%	72%	72%	72%	73%	56%	59%	58%	52%	51%	51%	56%
21	76%	72%	72%	74%	74%	70%	71%	74%	73%	73%	7%	72%	72%	72%	73%	57%	59%	58%	52%	51%	51%	56%
22	83%	78%	78%	80%	80%	75%	77%	80%	80%	81%	8%	79%	79%	79%	80%	82%	84%	84%	56%	54%	54%	61%
23	81%	76%	76%	78%	79%	74%	75%	78%	78%	77%	7%	77%	77%	77%	77%	82%	82%	82%	55%	55%	54%	61%
24	82%	77%	77%	79%	79%	75%	76%	80%	79%	79%	7%	78%	78%	78%	78%	81%	82%	82%	55%	54%	54%	62%
25	81%	76%	75%	77%	78%	74%	75%	78%	77%	77%	7%	77%	77%	77%	77%	82%	81%	81%	54%	54%	53%	60%
26	82%	77%	77%	79%	79%	75%	76%	79%	79%	79%	7%	78%	78%	78%	78%	82%	84%	81%	54%	54%	54%	60%
27	83%	77%	77%	79%	80%	75%	76%	79%	79%	79%	7%	78%	78%	78%	78%	81%	82%	82%	55%	53%	53%	60%
28	83%	79%	79%	81%	82%	77%	77%	81%	81%	80%	8%	79%	79%	79%	79%	82%	85%	82%	55%	55%	55%	62%
29	86%	80%	79%	80%	82%	78%	77%	82%	81%	81%	8%	80%	82%	82%	81%	83%	86%	85%	57%	57%	56%	60%
30	87%	80%	79%	80%	81%	78%	77%	82%	81%	81%	8%	80%	82%	82%	81%	83%	86%	86%	56%	56%	55%	60%
31	88%	80%	80%	80%	82%	79%	78%	83%	82%	81%	8%	82%	82%	82%	81%	83%	86%	85%	56%	57%	57%	62%
32	85%	79%	80%	81%	82%	78%	77%	82%	81%	81%	8%	82%	82%	82%	81%	84%	87%	86%	57%	56%	56%	62%
33	80%	78%	79%	80%	81%	78%	78%	81%	81%	81%	8%	81%	81%	80%	80%	85%	87%	87%	56%	57%	56%	62%
34	84%	78%	81%	80%	82%	77%	78%	81%	81%	82%	8%	80%	81%	81%	82%	83%	86%	84%	57%	57%	54%	61%
35	84%	80%	81%	80%	82%	77%	78%	81%	81%	82%	8%	80%	81%	80%	82%	83%	86%	85%	57%	56%	53%	61%
36	87%	80%	79%	80%	82%	78%	78%	83%	81%	82%	8%	80%	81%	81%	80%	82%	85%	85%	57%	56%	57%	61%
37	87%	81%	81%	81%	84%	80%	79%	84%	82%	82%	8%	81%	82%	81%	81%	82%	85%	86%	56%	57%	56%	61%
38	84%	80%	80%	79%	81%	77%	77%	82%	81%	80%	8%	79%	79%	79%	79%	81%	83%	86%	56%	56%	56%	62%
39	85%	78%	78%	80%	82%	78%	78%	82%	81%	82%	8%	80%	79%	80%	81%	84%	85%	85%	55%	55%	55%	60%
40	85%	80%	79%	79%	81%	77%	78%	80%	80%	79%	7%	79%	80%	80%	80%	81%	83%	84%	55%	56%	56%	60%
41	83%	78%	78%	79%	80%	76%	76%	79%	80%	79%	7%	79%	80%	79%	80%	82%	84%	84%	55%	55%	54%	58%

B4.4 Distance Matrix for the region from the nucleocapsid to the p1 spacer peptide

p#01

	C 95N21088																														B_HFL							
	C 96W022	TV00065	TV00066	TV00067	TV00068	TV00069	TV00070	TV00071	TV00072	TV00073	TV00074	TV00075	TV00076	TV00077	TV00078	TV00079	TV00080	TV00081	TV00082	TV00083	TV00084	TV00085	TV00086	TV00087	TV00088	TV00089	TV00090	TV00091	TV00092	TV00093		TV00094	TV00095	TV00096	TV00097	TV00098	TV00099	TV00100
42	338	321	323	323	326	322	320	334	333	332	334	341	331	331	337	337	333	334	321	341	341	340	332	331	336	324	332	330	344	337	339	356	321	323	326	322	320	
43	343	325	328	328	331	325	323	338	332	331	337	341	337	337	341	336	337	324	343	342	341	335	333	338	325	333	332	334	344	333	338	324	348	326	334	342	336	
44	339	328	330	330	331	328	326	342	338	337	333	340	336	336	345	337	336	326	341	344	343	335	335	336	326	334	332	343	344	346	353	344	346	348	349	348	349	
45	342	325	335	334	339	330	330	336	337	338	338	344	338	338	338	336	336	330	343	347	346	337	337	339	331	337	335	346	337	339	339	339	339	339	339	339	339	
46	345	329	337	337	340	331	331	343	338	340	338	344	343	342	338	340	335	348	347	346	338	336	338	334	338	340	332	340	332	337	341	341	341	341	341	341	341	
47	349	332	340	338	341	335	333	340	340	337	342	350	345	345	341	338	338	334	351	351	350	338	336	340	336	340	343	354	338	340	340	340	340	340	340	340	340	
48	342	322	328	328	332	327	326	335	331	331	337	341	339	338	338	333	334	327	339	344	343	333	331	333	330	336	332	343	333	335	335	335	335	335	335	335	335	
49	344	328	335	335	334	330	331	337	337	337	335	347	341	341	337	337	333	341	346	347	346	334	336	331	335	338	351	336	338	338	338	338	338	338	338	338	338	
50	344	321	329	329	330	327	328	335	332	333	345	342	343	335	334	328	342	343	343	342	343	328	342	331	330	332	328	339	343	341	344	344	344	344	344	344	344	344
51	340	330	345	345	342	335	333	338	338	339	341	348	345	345	350	348	342	338	351	352	351	347	338	342	338	338	345	349	345	348	348	348	348	348	348	348	348	
52	338	325	336	336	337	334	331	339	334	336	340	341	338	338	348	341	337	336	343	348	345	347	332	333	337	332	334	337	342	345	345	345	345	345	345	345	345	345
53	345	327	339	341	338	333	331	335	335	336	340	347	345	345	342	341	338	332	346	348	347	346	338	343	338	340	337	349	342	345	345	345	345	345	345	345	345	345
54	342	322	334	334	338	327	327	323	341	334	335	338	341	340	340	338	338	337	328	345	344	343	341	334	336	335	335	348	336	339	339	339	339	339	339	339	339	339
55	345	332	342	344	343	339	333	338	340	339	347	348	342	343	341	338	339	349	355	354	343	335	336	337	341	343	351	342	345	345	345	345	345	345	345	345	345	345
56	343	324	333	333	336	335	331	337	338	337	341	342	339	339	341	340	337	334	344	348	347	346	333	336	327	333	336	343	345	348	348	348	348	348	348	348	348	348
57	347	326	335	335	336	334	334	337	340	341	342	344	339	339	347	344	342	336	345	348	347	346	338	342	333	341	339	348	350	353	353	353	353	353	353	353	353	353
58	348	325	336	336	337	330	328	334	339	340	345	345	342	347	343	338	331	342	347	348	343	342	349	345	339	339	355	341	337	347	347	347	347	347	347	347	347	347
59	343	327	336	336	335	334	331	333	338	337	335	342	340	347	343	338	331	344	344	345	340	330	332	330	336	335	350	346	348	348	348	348	348	348	348	348	348	348
60	331	316	324	324	328	328	321	320	327	328	338	336	332	332	333	328	325	318	337	335	334	330	324	328	323	327	328	338	345	347	347	347	347	347	347	347	347	347
61	349	332	343	343	340	343	339	344	341	340	347	344	344	349	348	348	349	346	333	353	349	348	347	340	342	336	342	346	350	352	352	352	352	352	352	352	352	352
62	325	314	322	322	326	318	316	325	320	321	329	328	332	332	334	331	328	317	328	330	329	326	318	322	316	320	322	322	322	322	322	322	322	322	322	322	322	322
63	329	313	321	324	324	322	320	328	322	323	331	328	330	330	334	327	323	319	322	333	332	328	321	325	316	322	322	322	322	322	322	322	322	322	322	322	322	322
64	331	319	327	327	328	323	321	322	325	326	326	337	337	339	331	328	328	333	340	338	334	328	328	321	327	330	340	337	338	338	338	338	338	338	338	338	338	338
65	337	321	330	333	333	326	324	334	328	329	337	339	340	344	336	333	328	334	339	338	335	328	330	326	328	328	330	343	342	341	341	341	341	341	341	341	341	341
66	325	314	321	323	322	320	318	321	324	325	330	332	335	335	336	328	323	320	327	331	330	326	322	324	317	323	325	332	339	340	340	340	340	340	340	340	340	340
67	315	308	310	312	311	310	308	311	315	316	317	322	327	327	325	325	321	314	318	321	320	319	314	315	308	314	320	324	325	324	325	325	325	325	325	325	325	325
68	334	320	327	329	330	325	323	328	330	329	340	338	342	341	335	328	328	331	336	337	333	329	333	323	329	330	337	344	344	344	344	344	344	344	344	344	344	344
69	330	319	329	329	330	326	324	330	328	327	335	336	337	337	340	334	329	326	334	336	335	333	324	328	320	328	339	341	342	342	342	342	342	342	342	342	342	342
70	335	319	323	325	328	323	321	334	330	329	336	338	332	332	343	330	326	324	331	337	336	337	323	327	322	330	339	341	342	342	342	342	342	342	342	342	342	342
71	332	319	323	325	326	322	322	334	329	328	335	335	331	331	340	328	323	330	332	333	330	337	323	327	320	328	338	333	338	339	339	339	339	339	339	339	339	339
72	327	310	321	320	321	315	313	328	318	319	334	328	328	332	326	321	315	328	333	332	329	330	325	316	324	323	330	344	343	343	343	343	343	343	343	343	343	343
73	329	314	321	321	322	318	316	330	322	323	331	330	330	332	328	323	317	329	332	331	330	325	317	325	324	323	334	341	341	341	341	341	341	341	341	341	341	341
74	329	314	321	321	322	318	314	331	321	322	335	328	329	329	333	328	325	319	325	331	330	326	322	325	316	322	324	329	344	345	345	345	345	345	345	345	345	345
75	311	318	328	328	327	326	322	326	324	324	334	332	335	334	337	332	320	325	334	337	332	320	325	334	323	323	330	342	341	341	341	341	341	341	341	341	341	341
76	276	262	269	270	276	271	269	270	274	273	283	279	290	288	279	282	276	271	277	279	278	274	273	271	274	268	276	282	282	282	282	282	282	282	282	282	282	282
77	280	263	269	272	274	274	273	277	282	281	283	284	289	289	285	285	277	275	280	282	281	278	270	271	271	277	285	288	290	290	290	290	290	290	290	290	290	290
78	286	270	280	280	278	281	280	275	278	287	284	289	288	281	288	278	277	283	284	283	286	278	278	271	275	278	281	282	281	281	281	281	281	281	281	281	281	281
79	243	235	245	247	247	243	245	248	243	244	248	247	248	245																								

B4.4 Distance Matrix for the region for the nucleocapsid to the p1 spacer peptide

p1

TV018G65	B_U83 RF	J5E7022	D_ELI	D_4GUG1141	D_5F_V1310	D_4_CPX84C203-2	D_4_CPX87PV1C	G_H4973-1.1	G_SEE165	D_6_CPX89M4	F1 FN6383	F1 MP411	F2 MP257	H_90C7056	K_EQT811C															
31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	
340	339	334	334	334	338	339	332	344	335	333	414	644	644	644	676	676	676	676	676	676	676	676	676	676	676	676	676	676	676	
338	339	334	332	332	340	344	333	340	340	341	392	414	644	644	676	676	676	676	676	676	676	676	676	676	676	676	676	676	676	
18	18	23	23	23	18	22	23	15	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
348	345	342	341	341	346	346	340	351	338	341	393	280	414	644	676	676	676	676	676	676	676	676	676	676	676	676	676	676	676	
346	345	342	341	341	346	346	340	351	338	341	393	280	414	644	676	676	676	676	676	676	676	676	676	676	676	676	676	676	676	
18	18	23	23	23	18	22	23	15	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
346	341	338	338	338	338	350	342	349	337	341	358	363	360	417	696	696	696	696	696	696	696	696	696	696	696	696	696	696	696	
346	341	338	338	338	338	350	342	349	337	341	358	363	360	0	896	896	896	896	896	896	896	896	896	896	896	896	896	896	896	
18	17	22	22	22	17	23	24	14	11	15	15	15	0	0	3%	4%	4%	4%	4%	4%	4%	4%	4%	4%	4%	4%	4%	4%	4%	
343	341	338	343	344	342	354	343	355	344	347	363	367	364	360	414	676	676	676	676	676	676	676	676	676	676	676	676	676	676	
343	341	338	343	344	342	354	343	355	344	347	363	367	364	360	0	676	676	676	676	676	676	676	676	676	676	676	676	676	676	
15	14	19	18	18	12	12	12	15	14	4	4	15	0	0	1%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	
342	342	341	341	340	347	353	345	359	348	348	369	378	368	363	366	417	676	676	676	676	676	676	676	676	676	676	676	676	676	
342	342	341	341	340	347	353	345	359	348	348	369	378	368	363	366	0	676	676	676	676	676	676	676	676	676	676	676	676	676	
22	21	26	26	26	17	23	24	18	19	3	3	3	18	7	0	0	2%	0%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	
339	339	333	335	336	338	346	335	348	338	343	362	365	359	373	380	365	408	676	676	676	676	676	676	676	676	676	676	676	676	
339	339	333	335	336	338	346	335	348	338	343	362	365	359	373	380	365	0	676	676	676	676	676	676	676	676	676	676	676	676	
25	24	29	29	29	22	28	29	21	22	6	6	8	21	10	9	0	1%	4%	4%	4%	4%	4%	4%	4%	4%	4%	4%	4%	4%	
341	340	342	332	333	343	354	347	356	344	345	368	369	365	375	380	387	381	414	696	696	696	696	696	696	696	696	696	696	696	
341	340	342	332	333	343	354	347	356	344	345	368	369	365	375	380	387	381	0	696	696	696	696	696	696	696	696	696	696	696	
18	18	23	23	23	18	22	23	15	16	0	0	0	0	0	0	0	0	0	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	
348	344	345	336	335	351	356	341	360	346	350	371	372	373	379	378	387	378	420	676	676	676	676	676	676	676	676	676	676	676	
348	344	345	336	335	351	356	341	360	346	350	371	372	373	379	378	387	378	0	676	676	676	676	676	676	676	676	676	676	676	
25	24	29	29	29	20	20	26	27	17	20	12	12	12	15	12	9	18	12	0	4%	4%	4%	4%	4%	4%	4%	4%	4%	4%	4%
350	344	348	345	348	348	356	349	362	353	349	342	345	346	345	356	354	348	353	350	408	644	624	614	604	604	604	604	604	604	
350	344	348	345	348	348	356	349	362	353	349	342	345	346	345	356	354	348	353	350	0	644	624	614	604	604	604	604	604	604	
23	22	27	21	21	22	22	28	29	21	18	12	12	15	18	12	18	15	12	18	0	0	0	0	0	0	0	0	0	0	
348	348	349	344	344	345	347	341	353	345	341	338	341	342	337	348	347	344	344	384	408	676	676	676	676	676	676	676	676	676	
348	348	349	344	344	345	347	341	353	345	341	338	341	342	337	348	347	344	344	0	676	676	676	676	676	676	676	676	676	676	
23	22	27	21	21	22	22	28	29	21	18	12	12	15	18	15	18	15	18	0	0	0	0	0	0	0	0	0	0	0	
349	339	348	345	344	347	348	344	352	350	349	343	341	338	342	351	351	341	347	347	380	372	411	604	604	604	604	604	604	604	
349	339	348	345	344	347	348	344	352	350	349	343	341	338	342	351	351	341	347	347	0	604	604	604	604	604	604	604	604	604	
20	19	24	18	18	19	19	25	26	18	15	9	9	9	12	11	12	15	3	7	0	1%	0%	0%	0%	0%	0%	0%	0%	0%	
341	339	342	342	343	345	350	344	352	352	349	350	351	345	350	363	356	351	351	352	380	370	377	417	676	676	676	676	676	676	
341	339	342	342	343	345	350	344	352	352	349	350	351	345	350	363	356	351	351	352	0	676	676	676	676	676	676	676	676	676	
18	17	22	18	17	17	23	24	14	11	15	15	15	15	18	21	15	15	9	8	0	1%	1%	4%	4%	4%	4%	4%	4%	4%	
349	339	346	349	350	347	354	349	364	351	348	348	350	348	354	365	364	363	367	365	366	370	411	676	676	676	676	676	676	676	
349	339	346	349	350	347	354	349	364	351	348	348	350	348	354	365	364	363	367	365	366	370	0	676	676	676	676	676	676	676	
20	19	24	18	18	19	19	25	26	18	15	9	9	9	12	11	12	15	3	7	0	0	0	0	0	0	0	0	0	0	
348	344	348	345	348	348	356	349	362	353	349	342	345	346	345	356	354	348	353	350	408	644	624	614	604	604	604	604	604	604	
348	344	348	345	348	348	356	349	362	353	349	342	345	346	345	356	354	348	353	350	0	644	624	614	604	604	604	604	604	604	
23	22	27	21	21	22	22	28	29	21	18	12	12	15	18	12	18	15	12	18	0	0	0	0	0	0	0	0	0	0	
348	348	349	344	344	345	347	341	353	345	341	338	341	342	337	348	347	344	344	384	408	676	676	676	676	676	676	676	676	676	
348	348	349	344	344	345	347	341	353	345	341	338	341	342	337	348	347	344	344	0	676	676	676	676	676	676	676	676	676	676	
23	22	27	21	21	22	22	28	29	21	18	12	12	15	18	15	18	15	18	0	0	0	0	0	0	0	0	0	0	0	
349	339	348	345	344	347	348	344	352	350	349	343	341	338	342	351	351	341	347	347	380	372	411	604	604	604	604	604	604	604	
349	339	348	345	344	347	348	344	352	350	349	343	341	338	342	351	351	341	347	347	0	604	604	604	604	604	604	604	604	604	
20	19	24	18	18	19	19	25	26	18	15	9	9	9	12	11	12	15	3	7	0	1%	0%	0%	0%	0%	0%	0%	0%	0%	
341	339	342	342	343	345	350	344	352	352	349	350	351	345	350	363	356	351	351	352	380	370	377	417	676	676	676	676	676	676	
341	339	342	342	343	345	350	344	352	352	349	350	351	345	350	363	356	351	351	352	0	676	676	676	676	676	676	676	676	676	
18	17	22	18	17	17	23	24	14	11	15	15	15	15	18	21	15	15	9	8	0	1%	1%	4%	4%	4%	4%	4%	4%	4%	
349	339	346	349	350	347	354	349	364	351	348	348	350	348	354	365	364	36													

B4.4 Distance Matrix for the region from the nucleocapsid to the p1 spacer peptide

#7-#1	A SE2753	A Q23-17	A U455	01_AE CM240	01_AE B9TH047	01_AB KAL153-2	02_AG D.0263	02_AG 18HG	CPZ CM3	N YF530	O MYP5180	CPZ CP2548
K.MP335	A.92UG037	A.U455	01_AE B9TH053	01_AE B9CF042	01_AB_RU86001	02_AG_D.0264	02_AG_G459	CPZ CP25	OANT70	CPZ CP2ANT		
81	82	83	84	85	86	87	88	89	90	91	92	93
42	82%	79%	79%	79%	77%	76%	81%	80%	83%	81%	80%	81%
	82%	80%	79%	80%	77%	78%	82%	80%	82%	81%	80%	81%
	3%	3%	4%	4%	4%	10%	4%	4%	4%	4%	4%	4%
43	82%	80%	79%	80%	77%	78%	82%	80%	82%	81%	80%	81%
	82%	80%	80%	80%	77%	78%	82%	80%	82%	81%	80%	81%
	3%	3%	4%	4%	4%	10%	4%	4%	4%	4%	4%	4%
44	83%	80%	80%	80%	81%	79%	77%	83%	80%	83%	82%	80%
	83%	80%	80%	80%	81%	79%	77%	83%	80%	83%	82%	80%
	3%	3%	4%	4%	4%	7%	10%	4%	4%	4%	4%	4%
45	84%	79%	79%	80%	80%	78%	77%	82%	81%	81%	81%	80%
	84%	79%	79%	80%	80%	78%	77%	82%	81%	81%	81%	80%
	1%	2%	4%	3%	4%	8%	3%	3%	3%	3%	3%	3%
46	85%	80%	79%	81%	81%	79%	78%	82%	81%	82%	81%	80%
	85%	80%	79%	81%	81%	79%	78%	82%	81%	82%	81%	80%
	2%	2%	4%	3%	3%	8%	3%	3%	3%	3%	3%	3%
47	83%	79%	79%	80%	80%	77%	75%	81%	80%	81%	80%	80%
	83%	79%	79%	80%	80%	77%	75%	81%	80%	81%	80%	80%
	4%	4%	5%	4%	5%	10%	4%	4%	4%	4%	4%	4%
48	82%	78%	78%	79%	80%	77%	76%	80%	78%	80%	79%	79%
	82%	78%	78%	79%	80%	77%	76%	80%	78%	80%	79%	79%
	5%	5%	6%	5%	5%	11%	5%	5%	5%	5%	5%	5%
49	83%	80%	79%	81%	81%	79%	77%	82%	81%	82%	81%	80%
	83%	80%	79%	81%	81%	79%	77%	82%	81%	82%	81%	80%
	3%	3%	4%	4%	4%	7%	10%	4%	4%	4%	4%	4%
50	83%	78%	78%	80%	79%	77%	76%	81%	80%	82%	81%	79%
	83%	78%	78%	80%	79%	77%	76%	81%	80%	82%	81%	79%
	4%	4%	5%	4%	4%	7%	10%	4%	4%	4%	4%	4%
51	85%	82%	82%	82%	83%	79%	78%	83%	85%	83%	83%	82%
	85%	82%	82%	82%	83%	79%	78%	83%	85%	83%	83%	82%
	4%	4%	5%	4%	7%	10%	4%	4%	4%	4%	4%	4%
52	83%	82%	80%	81%	82%	78%	78%	82%	83%	82%	81%	81%
	83%	82%	80%	81%	82%	78%	78%	82%	83%	82%	81%	81%
	4%	4%	5%	4%	4%	7%	10%	4%	4%	4%	4%	4%
53	85%	80%	80%	80%	80%	78%	77%	82%	82%	80%	81%	79%
	85%	80%	80%	80%	80%	78%	77%	82%	82%	80%	81%	79%
	3%	3%	3%	3%	3%	8%	3%	3%	3%	3%	3%	3%
54	84%	79%	78%	78%	79%	77%	77%	81%	81%	81%	81%	79%
	84%	79%	78%	78%	79%	77%	77%	81%	81%	81%	81%	79%
	2%	2%	3%	3%	3%	8%	3%	3%	3%	3%	3%	3%
55	85%	81%	81%	82%	82%	80%	80%	84%	84%	83%	83%	81%
	85%	81%	81%	82%	82%	80%	80%	84%	84%	83%	83%	81%
	3%	3%	4%	4%	3%	8%	3%	3%	3%	3%	3%	3%
56	86%	81%	81%	81%	81%	79%	79%	83%	83%	82%	82%	80%
	86%	81%	81%	81%	81%	79%	79%	83%	83%	82%	82%	80%
	3%	3%	3%	3%	3%	8%	3%	3%	3%	3%	3%	3%
57	86%	81%	82%	81%	83%	79%	78%	83%	84%	84%	84%	82%
	86%	81%	82%	81%	83%	79%	78%	83%	84%	84%	84%	82%
	4%	4%	5%	4%	7%	10%	4%	4%	4%	4%	4%	4%
58	86%	80%	81%	81%	82%	78%	77%	82%	83%	82%	81%	82%
	86%	80%	81%	81%	82%	78%	77%	82%	83%	82%	81%	82%
	4%	4%	5%	5%	5%	8%	10%	5%	5%	5%	5%	5%
59	85%	79%	79%	81%	82%	78%	77%	82%	82%	81%	81%	80%
	85%	79%	79%	81%	82%	78%	77%	82%	82%	81%	81%	80%
	4%	4%	5%	4%	4%	7%	10%	4%	4%	4%	4%	4%
60	80%	81%	80%	81%	83%	79%	78%	85%	82%	84%	83%	82%
	80%	81%	80%	81%	83%	79%	78%	85%	82%	84%	83%	82%
	0%	1%	2%	2%	2%	5%	2%	2%	2%	2%	2%	2%
61	411	83%	82%	82%	84%	82%	81%	86%	84%	83%	83%	84%
	0	83%	82%	82%	84%	82%	81%	86%	84%	83%	83%	84%
	0	1%	2%	2%	2%	5%	2%	2%	2%	2%	2%	2%
62	347	411	91%	91%	90%	84%	83%	87%	89%	87%	87%	80%
	347	0	91%	91%	90%	84%	83%	87%	89%	87%	87%	80%
	8	0	1%	0%	1%	4%	6%	0%	0%	0%	0%	0%
63	345	379	411	91%	90%	84%	83%	87%	89%	87%	87%	80%
	345	379	0	91%	90%	84%	83%	87%	89%	87%	87%	80%
	12	8	0	1%	4%	6%	0%	0%	0%	0%	0%	0%
64	342	375	375	408	90%	85%	85%	90%	89%	88%	87%	80%
	342	375	375	0	90%	85%	85%	90%	89%	88%	87%	80%
	9	3	3	0	3%	5%	0%	0%	0%	0%	0%	0%
65	350	373	370	369	405	86%	86%	89%	89%	88%	87%	80%
	350	373	370	369	0	86%	86%	89%	89%	88%	87%	80%
	12	8	8	0	4%	8%	0%	0%	0%	0%	0%	0%
66	354	380	358	361	365	423	86%	91%	89%	85%	85%	82%
	354	380	358	361	365	0	86%	91%	89%	85%	85%	82%
	24	18	18	15	18	0	8%	3%	3%	3%	3%	3%
67	337	343	344	349	347	367	364	89%	86%	85%	81%	82%
	337	343	344	349	347	367	364	0	89%	86%	85%	81%
	33	27	27	24	27	38	0	5%	5%	5%	5%	5%
68	358	361	361	368	364	368	367	408	91%	90%	90%	87%
	358	361	361	368	364	368	367	408	0	91%	90%	87%
	8	3	3	3	3	3	3	0	0%	0%	0%	0%
69	357	367	370	367	368	374	353	375	408	89%	89%	88%
	357	367	370	367	368	374	353	375	408	0	89%	89%
	9	3	3	3	3	15	24	0	0%	0%	0%	0%
70	348	361	367	360	368	362	348	369	366	408	89%	87%
	348	361	367	360	368	362	348	369	366	408	0	89%
	9	3	3	3	3	15	24	0	0%	0%	0%	0%
71	347	360	365	359	365	361	347	368	365	401	408	86%
	347	360	365	359	365	361	347	368	365	401	408	0
	9	3	3	3	3	15	24	0	0%	0%	0%	0%
72	344	357	358	362	358	351	334	357	361	355	354	408
	344	357	358	362	358	351	334	357	361	355	354	408
	9	3	3	3	3	15	24	0	0%	0%	0%	0%
73	348	362	361	363	363	355	338	361	363	361	360	368
	348	362	361	363	363	355	338	361	363	361	360	368
	9	3	3	3	3	15	24	0	0%	0%	0%	0%
74	347	357	355	358	358	349	338	358	358	358	385	387
	347	357	355	358	358	349	338	358	358	358	385	387
	9	3	3	3	3	15	24	0	0%	0%	0%	0%
75	347	352	357	353	356	356	337	358	357	362	361	368
	347	352	357	353	356	356	337	358	357	362	361	368
	9	3	3	3	3	15	24	0	0%	0%	0%	0%
76	282	288	281	286	277	286	282	280	286	282	281	278
	282	288	281	286	277	286	282	280	286	282	281	278
	32	32	38	35	38	44	53	35	35	35	35	35
77	285	280	275	282	278	277	270	283	282	279	277	280
	285	280	275	282	278	277	270	283	282	279	277	280
	29	29	35	32	33	47	50	32	32	32	32	32
78	297	290	287	285	289	284	274	292	296	285	287	289
	297	290	287	285	289	2						

B4.5 Distance Matrix for the three areas of the p17 matrix containing published CTL epitopes

p17 CTL Area 3		TV001G8	TV006G97	TV013G2	TV004G17	TV010G74	TV019G25	TV012G34	TV002G8	TV006G29	TV008G65	TV009G12	TV016G65													
		TV001G11	TV006G11	TV013G15	TV014G73	TV004G24	TV019G20	TV007G59	TV012G40	TV018G90	TV005G36	TV008G66	TV003G15													
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	
1	8	100%	87%	87%	100%	100%	87%	87%	87%	77%	75%	75%	87%	87%	77%	100%	87%	87%	87%	87%	75%	87%	55%	87%	86%	
	0	100%	87%	87%	100%	100%	87%	87%	87%	88%	87%	87%	100%	100%	100%	100%	88%	100%	100%	100%	100%	87%	87%	86%	11%	
	0	0%	0%	0%	0%	0%	0%	0%	0%	11%	0%	0%	0%	0%	0%	11%	0%	0%	0%	0%	0%	0%	0%	0%	11%	
2	8	8	87%	87%	100%	100%	87%	87%	87%	77%	75%	75%	87%	87%	77%	100%	87%	87%	87%	87%	75%	87%	55%	87%	86%	
	8	0	87%	87%	100%	100%	87%	87%	87%	86%	87%	87%	100%	100%	100%	88%	100%	100%	100%	100%	100%	87%	87%	86%	11%	
	0	0	0%	0%	0%	0%	0%	0%	0%	11%	0%	0%	0%	0%	0%	11%	0%	0%	0%	0%	0%	0%	0%	0%	11%	
3	7	7	8	100%	87%	87%	75%	75%	75%	86%	82%	82%	75%	75%	75%	86%	87%	75%	75%	75%	75%	75%	75%	55%	86%	
	7	7	0	100%	87%	87%	75%	75%	75%	77%	75%	75%	87%	87%	77%	87%	87%	87%	87%	87%	75%	75%	75%	55%	86%	
	0	0	0	0%	0%	0%	0%	0%	0%	11%	0%	0%	0%	0%	0%	11%	0%	0%	0%	0%	0%	0%	0%	0%	11%	
4	7	7	8	8	87%	87%	75%	75%	75%	86%	82%	82%	75%	75%	75%	86%	87%	75%	75%	75%	75%	75%	75%	55%	86%	
	7	7	8	8	0	87%	87%	75%	75%	77%	75%	75%	87%	87%	77%	87%	87%	87%	87%	87%	75%	75%	75%	55%	86%	
	0	0	0	0	0%	0%	0%	0%	0%	11%	0%	0%	0%	0%	0%	11%	0%	0%	0%	0%	0%	0%	0%	0%	11%	
5	8	8	7	7	8	100%	87%	87%	87%	77%	75%	75%	87%	87%	77%	100%	87%	87%	87%	87%	75%	87%	55%	87%	86%	
	8	8	7	7	0	100%	87%	87%	87%	87%	87%	87%	100%	100%	88%	100%	100%	100%	100%	100%	100%	100%	87%	87%	11%	
	0	0	0	0	0	0%	0%	0%	0%	11%	0%	0%	0%	0%	0%	11%	0%	0%	0%	0%	0%	0%	0%	0%	11%	
6	8	8	7	7	8	8	87%	87%	87%	77%	75%	75%	87%	87%	77%	100%	87%	87%	87%	87%	75%	87%	55%	87%	86%	
	8	8	7	7	8	0	87%	87%	87%	87%	87%	87%	100%	100%	88%	100%	100%	100%	100%	100%	100%	100%	87%	87%	11%	
	0	0	0	0	0	0	0%	0%	0%	11%	0%	0%	0%	0%	0%	11%	0%	0%	0%	0%	0%	0%	0%	0%	11%	
7	7	7	8	8	7	7	8	75%	75%	86%	82%	82%	75%	75%	75%	86%	87%	75%	75%	75%	75%	75%	62%	75%	55%	
	7	7	8	8	7	7	0	75%	75%	77%	75%	75%	87%	87%	77%	87%	87%	87%	87%	87%	75%	75%	62%	75%	55%	
	0	0	0	0	0	0	0	0%	0%	11%	0%	0%	0%	0%	0%	11%	0%	0%	0%	0%	0%	0%	0%	0%	11%	
8	7	7	8	8	7	7	8	100%	77%	87%	87%	82%	87%	87%	77%	87%	87%	75%	75%	87%	87%	87%	100%	55%	86%	
	7	7	8	8	7	7	8	0	100%	77%	100%	100%	87%	87%	77%	87%	87%	87%	87%	87%	75%	75%	100%	55%	86%	
	0	0	0	0	0	0	0	0	0%	11%	0%	0%	0%	0%	0%	11%	0%	0%	0%	0%	0%	0%	0%	0%	11%	
9	7	7	8	8	7	7	8	8	77%	87%	87%	82%	87%	87%	77%	87%	87%	75%	75%	87%	87%	87%	100%	55%	86%	
	7	7	8	8	7	7	8	0	77%	100%	100%	100%	87%	87%	77%	87%	87%	87%	87%	87%	75%	75%	100%	55%	86%	
	0	0	0	0	0	0	0	0	11%	0%	0%	0%	0%	0%	0%	11%	0%	0%	0%	0%	0%	0%	0%	0%	11%	
10	8	8	7	7	8	7	7	8	86%	86%	55%	86%	86%	86%	86%	77%	86%	86%	86%	86%	86%	86%	86%	77%	55%	
	8	8	7	7	8	7	7	0	77%	77%	88%	88%	88%	88%	88%	88%	88%	88%	88%	88%	88%	88%	88%	77%	55%	
	1	1	1	1	1	1	1	1	11%	11%	11%	11%	11%	11%	11%	11%	11%	11%	11%	11%	11%	11%	11%	11%	0%	
11	8	8	5	5	8	8	5	7	7	6	8	100%	75%	75%	75%	86%	75%	82%	82%	75%	75%	75%	87%	44%	86%	
	7	7	8	8	7	7	8	8	7	8	8	0	100%	87%	87%	77%	87%	87%	87%	87%	87%	87%	100%	44%	86%	
	0	0	0	0	0	0	0	0	0	1	0	0%	0%	0%	0%	11%	0%	0%	0%	0%	0%	0%	0%	0%	11%	
12	8	8	5	5	8	8	5	7	7	8	8	75%	75%	75%	86%	75%	82%	82%	75%	75%	75%	75%	87%	44%	86%	
	7	7	8	8	7	7	8	8	7	8	8	0	87%	87%	77%	87%	87%	87%	87%	87%	87%	87%	100%	44%	86%	
	0	0	0	0	0	0	0	0	0	1	0	0%	0%	0%	0%	11%	0%	0%	0%	0%	0%	0%	0%	0%	11%	
13	8	8	8	8	7	7	8	8	7	7	7	8	100%	82%	82%	55%	75%	82%	82%	82%	82%	82%	82%	82%	44%	77%
	8	8	7	7	8	8	7	7	7	8	7	0	100%	100%	88%	100%	100%	100%	100%	100%	100%	100%	100%	75%	87%	
	0	0	0	0	0	0	0	0	0	1	0	0	0%	0%	0%	11%	0%	0%	0%	0%	0%	0%	0%	0%	11%	
14	7	7	8	8	7	7	8	7	7	8	8	8	100%	88%	87%	75%	75%	100%	100%	100%	100%	100%	75%	87%	55%	
	8	8	7	7	8	8	7	7	7	8	7	8	0	100%	88%	100%	100%	100%	100%	100%	100%	100%	87%	87%	86%	
	0	0	0	0	0	0	0	0	0	1	0	0	0	0%	0%	11%	0%	0%	0%	0%	0%	0%	0%	0%	11%	
15	7	7	8	8	7	7	8	7	7	8	8	8	88%	87%	75%	75%	100%	100%	100%	100%	100%	75%	87%	55%	86%	
	8	8	7	7	8	8	7	7	7	8	7	8	0	88%	100%	100%	100%	100%	100%	100%	100%	100%	87%	87%	86%	
	0	0	0	0	0	0	0	0	0	1	0	0	0	0%	0%	11%	0%	0%	0%	0%	0%	0%	0%	0%	11%	
16	7	7	8	8	7	7	8	7	7	8	8	8	8	8	8	0	77%	86%	86%	86%	86%	86%	86%	77%	55%	
	8	8	7	7	8	8	7	7	7	8	7	8	8	8	8	0	88%	88%	88%	88%	88%	88%	88%	77%	86%	
	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	11%	11%	11%	11%	11%	11%	11%	11%	0%	
17	8	8	7	7	8	8	7	7	7	7	7	8	8	8	8	7	7	8	87%	87%	87%	87%	75%	87%	55%	
	8	8	7	7	8	8	7	7	7	8	7	8	8	8	8	8	0	100%	100%	100%	100%	100%	87%	87%	86%	
	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0%	0%	0%	0%	0%	0%	0%	11%	
18	7	7	8	8	7	7	8	8	8	8	8	8	8	8	8	8	7	8	100%	75%	75%	82%	75%	44%	86%	
	8	8	7	7	8	8	7	7	7	8	7	8	8	8	8	8	0	100%	100%	100%	100%	100%	87%	87%	86%	
	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0%	0%	0%	0%	0%	0%	0%	11%	
19	7	7	8	8	7	7	8	8	8	8	8	8	8	8	8	8	7	8	8	75%	75%	82%	75%	44%	86%	
	8	8	7	7	8	8	7	7	7	8	7	8	8	8	8	8	0	100%	100%	100%	100%	100%	87%	87%	86%	
	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0%	0%	0%	0%	0%	0%	0%	11%	
20	7	7	8	8	7	7	8	7	7	8	8	8	8	8	8	8	7	8	8	8	8	8	100%	75%	87%	55%
	8	8	7	7	8																					

B5. Predicted amino acid sequence alignment of the complete Gag of the subtype C strains

B5. The predicted amino acid sequence alignment of the complete Gag of the HIV-1 subtype C *gag* clones, HIV-1 subtype C strains in the Los Alamos database as well as a subtype B strain and the subtype B *gag* clone of the isolate TV016. The amino acid sequence of the clone, TV001G8 is the reference sequence. Dots indicate similarity to the reference sequence, while dashes indicate a deletion at that position. A mismatch is indicated by the amino acid at that particular position of the sequence. The positions of the individual sequences are given on the right hand side. A general position indication is given in multiples of 20 at the top of the sequences. The different domains of Gag is indicated by a two way arrow on top of the alignment. The protease cleavage sites are indicated by a green line down the sequences.

B5. Predicted amino acid sequence alignment of the complete Gag of the subtype C strains

	← Matrix	↔	Capsid	→				
	20	*	140	*	160	*		
Tv001g8	: AQQAKAADE	----	KVSNYPIVQNA	QGQMVHQAI	S	PRTLNAWVKVIEEKAFNPEVI	PM : 168	
Tv001g11	:	-----	-----	-----	-----	-----	: 168	
Tv002g8	: T	KG-----	-----	L	-----	S	: 169	
Tv003g15	: T	A	G-----	-----	S	S	: 168	
Tv004g17	: T	G-----	-----	L	R	S	: 168	
Tv004g24	: T	G-----	-----	L	-----	S	: 168	
Tv005g29	: T	T	-----	F	L	S	: 168	
Tv005g36	: T	T	-----	F	L	S	: 168	
Tv006g11	: T	G-----	-----	-----	-----	S	: 167	
Tv006g97	: T	G-----	-----	-----	-----	S	: 168	
Tv007g59	: SE . . .	VGN . . .	-----	I	L	L	S : 168	
Tv008g65	: T	E	K-----	-----	L	L	G : 168	
Tv008g66	: T	E	K-----	-----	L	L	S : 168	
Tv009g12	: T	EG-----	-----	-----	L	S	: 168	
Tv010g74	: M	EV . . .	KK-----	-----	H	N	G : 169	
Tv012g34	: T	E	K-----	-----	L	-----	S : 168	
Tv012g40	: T	E	K-----	-----	L	-----	S : 168	
Tv013g2	: T	-----	-----	-----	L	NL	S : 170	
Tv013g15	: T	-----	-----	-----	L	NL	S : 170	
Tv014g73	: T	-----	G-----	-----	L	-----	S : 168	
Tv018g60	: T	-----	-----	-----	L	-----	S : 168	
Tv019g20	: T	E	G-----	-----	I	L	S : 168	
Tv019g25	: T	E	G-----	-----	I	L	S : 168	
92BR025	: T	E	KG-----	-----	L	P	A V S : 169	
98BR004	: T	E	KG-----	-----	R	L	PM A V S : 169	
96BW1210	: E	E	AKG-----	-----	L	-----	S I : 169	
BW.MJ4	: T	E	AG-----	-----	L	P	G S : 168	
96BW06J7	: T	E	AG-----	-----	L	P	G S : 168	
96BW06J4	: T	E	AG-----	-----	L	P	G S : 168	
96BW06K18	: T	E	AG-----	-----	L	P	G S : 168	
96BW06H51	: T	E	AG-----	-----	L	P	Q G I : 168	
96BW15B03	: T	E	AG-----	-----	I	L	S : 168	
96BW15C02	: T	E	AG-----	-----	I	L	S : 168	
96BW15C05	: T	E	AG-----	-----	I	L	S : 168	
96BW0504	: T	E	G-----	-----	L	-----	S : 167	
96BW0502	: T	E	G-----	-----	L	-----	S : 168	
96BW16D14	: T	ETT-----	G-----	-----	LH	-----	S : 168	
96BW1626	: T	E	TG-----	-----	L	S	S : 168	
96BW16B01	: T	E	TG-----	-----	L	-----	S : 168	
96.BW01B03	: T	TD . . .	G-----	-----	I	A	L PL S : 168	
96BW01B22	: T	TD . . .	G-----	-----	I	L	PL S : 168	
96BW01B21	: T	TD . . .	G-----	-----	I	P	PL K S : 168	
96BW0402	: I	E	KG-----	-----	L	-----	S : 169	
96BW0408	: I	E	KG-----	-----	L	-----	S : 169	
96BW0407	: I	E	KG-----	-----	L	-----	S : 169	
96BW0409	: I	E	KG-----	-----	L	-----	S : 169	
96BW1106	: T	E	K-----	-----	L	-----	S : 168	
96BW11B01	: T	VE . . .	K-----	-----	G	L	S : 168	
96BW1104	: T	E	K-----	-----	R	L	S : 168	
86.ETH2220	: T	G	RG-----	-----	D	M	P A V S : 169	
98IN022	: M	T	E	G-----	-----	L	V S I : 168	
93IN905	: T	E	G-----	-----	-----	L	S : 168	
93IN101	: T	E	G-----	-----	-----	L	S : 168	
IN.AF20999	: T	NE . . .	G-----	-----	-----	L	S : 168	
93IN904	: T	E	-----	-----	-----	L	L S : 168	
94IN11246	: T	E	-----	-----	-----	L	S : 165	
95IN21068	: T	ED . . .	G-----	-----	-----	L	S : 168	
93IN999	: T	E	G-----	-----	-----	L	P S : 168	
94IN476	: I	E	G-----	-----	-----	L	PL S : 168	
98IN012	: T	E	KK-----	-----	V	L	L S : 169	
98IS002	: T	-----	-----	-----	-----	P	Y SL S : 168	
98TZ013	: T	-----	V-----	-----	I	LR	-----	S : 168
98TZ017	: T	E	GG-----	-----	-----	L	-----	V N S : 172
97ZA012	: T	-----	KE-----	-----	-----	L	L	S : 169
96ZM651	: T	-----	G-----	-----	-----	L	KL	S : 170
B.FR.83.HX	:	A	DTGHSN	-----	-----	-----	-----	V S : 171
Tv016g95	:	A	DTGNSSSSQ	-----	-----	-----	-----	M V S : 173

B5. Predicted amino acid sequence alignment of the complete Gag of the subtype C strains

		← Capsid →										
		180	*	200	*	220	*					
Tv001g8	:	FTALSEGATPQDLN	TMLNTV	GGHQAAMQ	LKDTINEE	AAEWDR	THP	VHAGPVP	APGQMRE	:	227	
Tv001g11	:									:	227
Tv002g8	:									:	228
Tv003g15	:S.....									:	227
Tv004g17	:L..AQ.....I..									:	227
Tv004g24	:L..AQ.....I..									:	227
Tv005g29	:L.....I.....									:	227
Tv005g36	:L.....I.....									:	227
Tv006g11	:V.....									:	226
Tv006g97	:V.....R.....									:	227
Tv007g59	:L.....I.....									:	227
Tv008g65	:A.....L.....T.....									:	227
Tv008g66	:A.....L.....A.....									:	227
Tv009g12	:L..Q...I...I..									:	227
Tv010g74	:S.....I.....L..AQ.....I..D									:	228
Tv012g34	:L.....									:	227
Tv012g40	:L.....									:	227
Tv013g2	:P.....L.....M...L..									:	229
Tv013g15	:S.....L.....M...L..									:	229
Tv014g73	:L.....A.....									:	227
Tv018g60	:L.....									:	227
Tv019g20	:I.....I.....									:	227
Tv019g25	:I.....I.....									:	227
92BR025	:L.....									:	228
98BR004	:L..Q...I...I..									:	228
96BW1210	:G..L.....									:	228
BW.MJ4	:L..Q.....D									:	227
96BW06J7	:L..Q.....D									:	227
96BW06J4	:L..Q.....D									:	227
96BW06K18	:Q.....P.....D..Q..L..Q.....C..D									:	227
96BW06H51	:L..Q.....D									:	227
96BW15B03	:L.....I.....									:	227
96BW15C02	:L.....I.....									:	227
96BW15C05	:L.....I.....									:	227
96BW0504	:L..Q.....D									:	226
96BW0502	:L..Q.....D									:	227
96BW16D14	:V.....I.....									:	227
96BW1626	:V.....I.....									:	227
96BW16B01	:I.....V.....I.....									:	227
96.BW01B03	:L.....									:	227
96BW01B22	:L.G									:	227
96BW01B21	:L.....									:	227
96BW0402	:L.....I.....									:	228
96BW0408	:L.....I.....									:	228
96BW0407	:V.....I.....									:	228
96BW0409	:V.....I.....									:	228
96BW1106	:L.....									:	227
96BW11B01	:T.....L.....									:	227
96BW1104	:T.....L.....									:	227
86.ETH2220	:L.....D									:	228
98IN022	:R.....L..Q...IP.....									:	227
93IN905	:L.....I.....									:	227
93IN101	:L..I...I.....									:	227
IN.AF20999	:L.....I.....									:	227
93IN904	:L..I...I.....									:	227
94IN11246	:I.....I.....									:	224
95IN21068	:L..P...I...L..									:	227
93IN999	:L.....I...I..									:	227
94IN476	:S.....L.....NP.....									:	227
98IN012	:L.....									:	228
98IS002	:S..A.....V.....I.....									:	227
98T2013	:L.....I.....									:	227
98T2017	:DE.....S..A.....D..L.....I.....									:	231
97ZA012	:L.....A.....									:	228
96ZM651	:L.....I.....									:	229
B.FR.83.HX	:S.....E.....V.....I.....									:	230
Tv016g95	:A.....E.....M.....I.....									:	232

B5. Predicted amino acid sequence alignment of the complete Gag of the subtype C strains

	← Capsid →													
	240	*	260	*	280									
Tv001g8	:	PRGSDIAGTTSTLQEQIAWMTSNPPI	PVGD	IYKR	WII LGLN	KIVR	MYSP	VSI	LDIK	QGP	:	286		
Tv001g11	:	:	286		
Tv002g8	:	:	287		
Tv003g15	:	:	286		
Tv004g17	:	T	V	E	:	286		
Tv004g24	:	T	V	E	:	286		
Tv005g29	:	N	V	R	286		
Tv005g36	:	N	V	R	286		
Tv006g11	:	E	:	285		
Tv006g97	:	E	:	286		
Tv007g59	:	:	286		
Tv008g65	:	R	V	:	286		
Tv008g66	:	V	:	286		
Tv009g12	:	G	V	M	R	286		
Tv010g74	:	VT	N	R	287		
Tv012g34	:	E	:	286		
Tv012g40	:	E	:	286		
Tv013g2	:	I	R	288		
Tv013g15	:	R	R	288		
Tv014g73	:	G	V	:	286		
Tv018g60	:	I	A	E	R	286		
Tv019g20	:	V	E	:	286		
Tv019g25	:	V	E	:	286		
92BR025	:	T	N	V	:	287		
98BR004	:	N	V	E	T	287		
96BW1210	:	N	N	:	287		
BW.MJ4	:	G	H	:	286		
96BW06J7	:	G	H	:	286		
96BW06J4	:	G	H	:	286		
96BW06K18	:	H	G	H	:	286		
96BW06H51	:	G	H	:	286		
96BW15B03	:	R	286		
96BW15C02	:	R	286		
96BW15C05	:	T	R	286		
96BW0504	:	A	V	R	285		
96BW0502	:	A	V	R	286		
96BW16D14	:	V	E	R	286		
96BW1626	:	V	E	R	286		
96BW16B01	:	V	E	R	286		
96.BW01B03	:	N	A	V	:	286		
96BW01B22	:	N	A	V	:	286		
96BW01B21	:	N	A	V	:	286		
96BW0402	:	:	287		
96BW0408	:	:	287		
96BW0407	:	:	287		
96BW0409	:	:	287		
96BW1106	:	:	286		
96BW11B01	:	:	286		
96BW1104	:	:	286		
86.ETH2220	:	G	V	:	287		
98IN022	:	V	E	M	T	R	286
93IN905	:	S	G	V	E	:	286	
93IN101	:	S	G	V	R	:	286	
IN.AF20999	:	S	G	V	:	286		
93IN904	:	S	G	V	:	286		
94IN11246	:	V	E	T	:	283	
95IN21068	:	N	V	R	:	286	
93IN999	:	G	V	:	286		
94IN476	:	G	R	:	286	
98IN012	:	N	R	:	287	
98IS002	:	N	V	E	:	286	
98T2013	:	G	V	R	:	286	
98T2017	:	V	E	R	:	290	
97ZA012	:	M	:	287	
96ZM651	:	:	288		
B.FR.83.HX	:	G	N	E	T	R	289	
Tv016g95	:	N	A	T	E	T	R	291

B5. Predicted amino acid sequence alignment of the complete Gag of the subtype C strains

	← Capsid →					
	300	*	320	*	340	*
Tv001g8	:	KEPFRDYVDRFFKTLRAEQATQDVKNWMTDTLLVQNANPDCKTILRALGPGASLEEMMT	:			345
Tv001g11	:				345
Tv002g8	:	SS.E.....			T.....
Tv003g15	:	E.....		G..T..T..G...
Tv004g17	:	E..G.....		T.....
Tv004g24	:	E..G.....		T.....
Tv005g29	:	E.....		I.....
Tv005g36	:	E.....		I.....
Tv006g11	:
Tv006g97	:I.....
Tv007g59	:		I..G....T.....
Tv008g65	:	E..G.....		T.....
Tv008g66	:	E..G.....		T.....
Tv009g12	:	E.....		T.....
Tv010g74	:		T.....
Tv012g34	:	S..E.....	
Tv012g40	:	S..E.....	
Tv013g2	:	A.....		E.....K..I..T.....
Tv013g15	:	A.....		E.....K..I..T.....
Tv014g73	:	V.....		I.....K..A.....
Tv018g60	:		T.....
Tv019g20	:		I..G....T.....
Tv019g25	:		I..G....T.....
92BR025	:
98BR004	:
96BW1210	:
BW.MJ4	:	..S.....	C.....		T.....
96BW06J7	:	..S.....	C.....		T.....
96BW06J4	:	..S.....	C.....		T.....
96BW06K18	:	..S.....	C.....		T.....
96BW06H51	:	..S.....	C.....		T.....
96BW15B03	:		P.....T.....
96BW15C02	:		T.....
96BW15C05	:		T.....
96BW0504	:		E.....T.....
96BW0502	:		E.....T.....
96BW16D14	:D.....
96BW1626	:S.....
96BW16B01	:
96.BW01B03	:		A.....I.....
96BW01B22	:I.....
96BW01B21	:I.....
96BW0402	:	S..E.....	
96BW0408	:	S..E.....	
96BW0407	:	S..E.....	
96BW0409	:	S..E.....	
96BW1106	:	..S.....	SS.E.....		R..KT.....
96BW11B01	:	..S.....	SS.E.R.....		R..K.....
96BW1104	:	..S.....	SS.E.....		R..K.....
86.ETH2220	:
98IN022	:		E.....
93IN905	:A.....
93IN101	:
IN.AF20999	:
93IN904	:
94IN11246	:
95IN21068	:		E.....
93IN999	:		R.....
94IN476	:		E..G.....V.....
98IN012	:	S.....		A.....T.....
98IS002	:	E.....		K.....T.....
98TZ013	:A.T.....
98TZ017	:G....T.....
97ZA012	:	..S.....		K.....
96ZM651	:		E.....K.....T.....
B.FR.83.HX	:	Y.....		S.E.....E.....K....A.T.....
Tv016g95	:	Y.....		S.E.....E.....K....A.T.....

B5. Predicted amino acid sequence alignment of the complete Gag of the subtype C strains

	← Capsid	p2	Nucleocapsid	→			
	360	*	380	*	400	*	
Tv001g8	: ACQGVGGP	SHKARV	LAEAMSQ	TN--	SNII	VQRSNFKGPNRIVKCFNC	GKVGHIARKCRA : 402
Tv001g11	:			--		S.....	V.N... : 402
Tv002g8	:	G.....	A--T..	MM.K...	K.T.....	E.....	N... : 403
Tv003g15	:	G.....	ATY--	MM.....	SRK.....	RK.....	N... : 402
Tv004g17	:		S-AS..	MM.K...	AK.....	E.....	N... : 403
Tv004g24	:		S-AS..	MM.K...	AK.....	E.....	N... : 403
Tv005g29	:	G.....	A.N-AH	MM.....	SK.....	E.....	KN... : 403
Tv005g36	:	G.....	A.N-AH	MM.....	SK.....	E.....	KN... : 403
Tv006g11	:		A.N-I..	M.....	SK.....	E.....	N... : 402
Tv006g97	:		A.N-I..	M.....	SK.....	E.....	KN... : 403
Tv007g59	:		A.S-G..	MM.K...	R.SK..	I.....	E... KN.K. : 403
Tv008g65	:	G.....	A.--..	M.....	SK.....	E.....	GN... : 402
Tv008g66	:	G.....	A.--..	EM.....	SK.....	E.....	KN... : 402
Tv009g12	:	G.....	V.N-A..	MM.....	SK.....	E.....	N... : 403
Tv010g74	:		AGN-T..	MM.K...	R.TI...	E...L..	N... : 404
Tv012g34	:	A.....	A.N-ASV	MM.K...	RST...	E...L..	N... : 403
Tv012g40	:	T.....	A.N-TSVM	I.K...	R.A...	E...L..	N... : 403
Tv013g2	:		A.N-T..	MM.....	SSK.....	E.....	N... : 405
Tv013g15	:		A.N-T..	MM.....	SSK.....	S.....	E...N... : 405
Tv014g73	:	G.....	A.--..	MM.....	SK.....	E.....	N... : 402
Tv018g60	:		A.N-AV..	MM.K...	RK.IR..	E...L..	N... : 403
Tv019g20	:	G.....	A.S-T..	MM..G...	K.NI...	E...L..	NY... : 403
Tv019g25	:	G.....	A.S-T..	MM..G...	K.NI...	E...L..	N... : 403
92BR025	:	G.....	KV.N-T..	MM...C...	K.TI...	E...L..	N... : 404
98BR004	:	G.....	V.N-T..	MM..G...	K.I...	E...L..	KN... : 404
96BW1210	:		HAGN-AG	MM..G...	RK.P...	E...L..	N... : 404
BW.MJ4	:		A.S-TS..	M..G...	K.I...	E...L..	KN... : 403
96BW06J7	:		A.S-TS..	M..G...	K.I...	E...L..	KN... : 403
96BW06J4	:		A.S-TS..	M..G...	K.I...	E...L..	KN... : 403
96BW06K18	:		A.S-TS..	M..G...	K.I...	E...L..	KN... : 403
96BW06H51	:		A.S-TS..	M..G...	K.I...	E...L..	KN... : 403
96BW15B03	:	G.....	ATS-A..	M.....	K.I...	E...L..	N... : 403
96BW15C02	:	G.....	ATS-A..	M.....	K.I...	E...L..	N... : 403
96BW15C05	:	G.....	ATS-A..	M.....	K.I...	E...L..	N... : 403
96BW0504	:	G.....	A.S-V..	MM.K...	R.N...	E...L..	KN... : 402
96BW0502	:	G.....	A.S-V..	MM.K...	R.N...	E...L..	KN... : 403
96BW16D14	:		A.N-T.V..	I..G...	R.S...	E...L..	N... : 403
96BW16Z6	:	G.....	A.N-T..	MI.....	R.S...	E...L..	N... : 403
96BW16B01	:		A.N-T..	MI.....	R.S...	E...L..	N... : 403
96.BW01B03	:		A.S-M..	MM.....	N.K...	E...L..	N... : 403
96BW01B22	:		G.A.S-M..	MM.....	N.K...	E...L..	N... : 403
96BW01B21	:		A.S-M..	MM.....	N.K...	E...L..	N... : 403
96BW0402	:		--T.VMM	R.....	E...L..	N... : 403
96BW0408	:		--T.VMM	R.....	E...L..	N... : 403
96BW0407	:	G.....	A.--T.VMM	K.....	R.....	E...L..	N... : 403
96BW0409	:		A.--T.SVMM	K.....	R.....	E...L..	N... : 403
96BW1106	:	I.....	A.N-P..	MM.KN...	R.....	E...L..	N.K. : 403
96BW11B01	:	I.....	A.N-P..	MM.KN...	R.....	E...L..	N.K. : 403
96BW1104	:	I.....	A.N-P..	MM.KN...	T.....	E...L..	N.K. : 403
86.ETH2220	:	A.....	V.N-TT..	MM.K...	K.AI...	E...L..	N... : 404
98IN022	:	G.....	A.S-T..	MM..G...	K.....	E...L..	KN... : 402
93IN905	:	G.....	A.N-T..	M.....	SK.....	E...L..	KN... : 402
93IN101	:	G.....	A.S-T..	M.....	SK.....	E...L..	KN... : 402
IN.AF20999	:		S-T..	M.....	K.....	E...L..	KN... : 402
93IN904	:		S-A..	M.K...	SK..I...	E...L..	N... : 402
94IN11246	:	K.....	S-A..	M.....	SK..I...	E...L..	N... : 399
95IN21068	:		S-A..	M.....	SK.....	E...L..	N... : 402
93IN999	:	G.....	A.S--..	M.....	SK.T...	E...L..	N... : 402
94IN476	:		SHS--..	MM..G...	K.....	E...L..	N... : 402
98IN012	:		GS-T..	MM.....	SK.T...	E...L..	N... : 403
98IS002	:		A.N-T..	MM.K...	R.TI...	E...L..	N... : 403
98T2013	:		TS-T..	M.....	K.T...	E...L..	N... : 403
98T2017	:	G...I..	A.N-A..	MM.....	TRKT...	E...L..	N... : 407
97ZA012	:		A.N-T..	MM.K...	YK.....	E...L..	KN... : 404
96ZM651	:		S-V..	M.K...	NK.M...	E...L..	N... : 405
B.FR.83.HX	:	G.....	VTNSAT	MM..G...	RNQRK...	E...L..	T..N... : 407
Tv016g95	:	G.....	VTNSAT	MM.KG...	RNQRK...	E...L..	N... : 409

B5. Predicted amino acid sequence alignment of the complete Gag of the subtype C strains

	← Nucleocapsid →	p1	p6 →	
	420	440	460	
Tv001g8	: PRKKGCGWKCQEGHQMKDCTERQANFLGKIWPSHKG-RPGNE	LQNRPEPTAPPAE----	: 456	
Tv001g11	:	: 456	
Tv002g8	:K.....Y..S..	: 457	
Tv003g15	:K.....RE..K.....	: 456	
Tv004g17	:	: 457	
Tv004g24	:	: 457	
Tv005g29	:K.....	: 457	
Tv005g36	:K.....	: 457	
Tv006g11	:K...I.....R..	: 456	
Tv006g97	:RK.R.....	: 456	
Tv007g59	: ...R.....K.....C..	: 457	
Tv008g65	:K.....E.....	: 456	
Tv008g66	:K.....	: 456	
Tv009g12	:K.....	: 457	
Tv010g74	:K.....S..	: 458	
Tv012g34	: ...D....K.....	: 457	
Tv012g40	:K.....	: 457	
Tv013g2	:K.....	: 463	
Tv013g15	:K.....	: 463	
Tv014g73	:	: 456	
Tv018g60	:K.....	: 457	
Tv019g20	:K.....	: 457	
Tv019g25	:K.....	: 457	
92BR025	:K...V.....R..L..	: 458	
98BR004	:K.....	: 458	
96BW1210	:K.....S.G.....	: 458	
BW.MJ4	:K.....G..	: 458	
96BW06J7	:K.....G..	: 458	
96BW06J4	:K.....G..	: 458	
96BW06K18	:K.....G..	: 458	
96BW06H51	:K.....G..	: 458	
96BW15B03	:K.....T..	: 457	
96BW15C02	:K.....T..	: 457	
96BW15C05	:K.....T..	: 457	
96BW0504	:K.....S..A..TV-	: 455	
96BW0502	:K.....S..A..TV-	: 456	
96BW16D14	:K.....E.....	: 457	
96BW1626	:K.....D.....	: 457	
96BW16B01	:K.....S..	: 457	
96.BW01B03	:K.....L..S..	: 457	
96BW01B22	:K.....L..S..	: 457	
96BW01B21	:K.....L..S..	: 457	
96BW0402	:K.....S..	: 457	
96BW0408	:K.....S..	: 457	
96BW0407	:K.....S..	: 457	
96BW0409	:K.....S..	: 457	
96BW1106	:K.....S..	: 457	
96BW11B01	:K.G.....S..	: 457	
96BW1104	:K.....R.....S..	: 457	
86.ETH2220	:K.....RL..N..	: 461	
98IN022	:K.....R.....	: 452	
93IN905	:K.....S..	: 456	
93IN101	:K.....S..	: 456	
IN.AF20999	:K.....S..	: 456	
93IN904	:K.....S..	: 456	
94IN11246	:K.....S..	: 453	
95IN21068	:K.....S..	: 456	
93IN999	:K.....R----	: 456	
94IN476	: ...R.....	: 456	
98IN012	:K.....A.....	: 457	
98IS002	: ...RSR...K.....E.....S..	: 457	
98TZ013	:R.....G..	: 457	
98TZ017	:K.....S..L-	: 460	
97ZA012	:K.....R.....S..	: 458	
96ZM651	:K.....	: 459	
B.FR.83.HX	:K.....Y..	: 461	
Tv016g95	:K.....S..E----	: 463	

B5. Predicted amino acid sequence alignment of the complete Gag of the subtype C strains

	← p6 →	
	480 * 500 * 520	
Tv001g8	: ---PTAPPAESFRFEETT--PVPRKE-KDRE---PLTSLKSLFGSDPSSQ-	: 497
Tv001g11	: -----KR.-.E.-----N.-----	: 497
Tv002g8	: -----RE.----.K.---P.-----	: 487
Tv003g15	: -----P---AMKQ.P.-----	: 491
Tv004g17	: -----T.KQ.P.-----	: 492
Tv004g24	: -----T.KQ.P.-----A.-----	: 492
Tv005g29	: -----SA.QQ.P.-----	: 492
Tv005g36	: -----SA.KQ.PT.-----	: 492
Tv006g11	: -----A.KQ---K.-----	: 488
Tv006g97	: -----K.---A.KQ---K.-----	: 488
Tv007g59	: ---L.I....A.KP.PR....I...P.-----	: 499
Tv008g65	: -----A.KQ.S....R...N.-----	: 491
Tv008g66	: -----N....A.KQ.S....R...N.-----	: 491
Tv009g12	: -----T.KQ.S....T.-----	: 492
Tv010g74	: -----AQKQ.PQ.-----G.-----	: 493
Tv012g34	: ---K....KR.P.-----	: 496
Tv012g40	: ---K....KQ.P...PYRE.-----	: 500
Tv013g2	: RPE.....T.KQ.P...D-----	: 508
Tv013g15	: RPE.....T.KQ.P...D-----	: 508
Tv014g73	: -----A.KQ.S....I.-----	: 491
Tv018g60	: -----K....T..Q.S....	: 492
Tv019g20	: -----N--A.KQ.P.-----	: 492
Tv019g25	: -----N--A.KQ.P.-----	: 492
92BR025	: -----G.E.TT.SRKQ.TI.K.L---L.T-	: 496
98BR004	: -----S.KQ.Q....L.-	: 493
96BW1210	: -----AQKQ.P...P---A...N..L.-	: 494
BW.MJ4	: -----ALKQ.P.K.-----P...L.-	: 493
96BW06J7	: -----ALKQ.P.K.-----P...L.-	: 493
96BW06J4	: -----ALKQ.P.K.-----P...L.-	: 493
96BW06K18	: -----ALKQ.P.K.-----P...L.-	: 493
96BW06H51	: -----ALKQ.P.K.-----L.-	: 493
96BW15B03	: -----K....A.KQ.P....I.....L.-	: 492
96BW15C02	: -----K....A.KQ.P....I.....L.-	: 492
96BW15C05	: -----K....A.KQ.P....I.....L.-	: 492
96BW0504	: -----A.KQ.P...PYRE...A.R...G.L.-	: 501
96BW0502	: -----A.KQ.P...PYRE...A.R...G.L.-	: 502
96BW16D14	: -----G.G....A.KQ.P....R...N..L.-	: 492
96BW1626	: -----G....A.KQ.P....R...N..L.-	: 492
96BW16B01	: -----G.G....A.KQ.P....H...R...N..L.-	: 492
96.BW01B03	: -----A.KQ.P....R...L.-	: 492
96BW01B22	: -----A.KQ.P....R...L.-	: 492
96BW01B21	: -----A.KQ.P....R...L.-	: 492
96BW0402	: -----KQ.P.-----L.-	: 492
96BW0408	: -----QKQ.P.----T.....L.-	: 492
96BW0407	: -----GQKQ.S....T.....N..L.-	: 492
96BW0409	: -----GQKQ.S....T.....N..L.-	: 492
96BW1106	: -----L---A.KQ.T....I.....L.-	: 490
96BW11B01	: -----L---A.KQ.M....L.-	: 490
96BW1104	: -----A.KQ.T....I.....L.-	: 490
86.ETH2220	: --E....P....A---S.KQ.L....A.....N.HLL-	: 504
98IN022	: -----G.G....A.KQ.S....L.-	: 494
93IN905	: -----A.KQ.P....R...L.-	: 491
93IN101	: -----A.KQ.P....LL.-	: 491
IN.AF20999	: -----ALQQGP....R...L.-	: 491
93IN904	: -----P---A.KQ.P....R...L.-	: 491
94IN11246	: -----P---A.KQ.P.E....R...L.-	: 488
95IN21068	: -----A.KQ.P....R...L.-	: 491
93IN999	: -PE.....ALKQ.P.-----L.-	: 500
94IN476	: -----K....A.KQ.S....L.-	: 491
98IN012	: -----A.KQ.L....L.-	: 492
98IS002	: -----A---S.KP.Q..K.PYKE..SA.....N..L.-	: 496
98T2013	: -----G....A.KQ.P....N..LF.-	: 499
98T2017	: -----A---LARKQ.L....I.....L.-	: 502
97ZA012	: -----K....L..QKQ.S....T.....L.-	: 495
96ZM651	: -----A.KQ.S....A.....L.-	: 494
B.FR.83.HX	: -----SGVE.TT.PQKQ.PI.K.LY---R...N.-----	: 500
Tv016g95	: -----G.E.TT.SQKQ.TV.KDLY---V.....N...YN-	: 502



sampson_characterisation_2002

