DOCTORAL THESIS 2013

Evolution of the Human Immunodeficiency Virus type I

Protease and Integrase:

Effects on Viral Replication Capacity and Robustness.

Elena Capel Malo

IrsiCaixa
Institut de Recerca de la Sida

UAB
Universitat Autònoma de Barcelona

Universitat Autònoma de Barcelona,

Faculty of Biosciences,

Department of Genetics and Microbiology.

Evolution of the Human Immunodeficiency Virus type I Protease and Integrase: Effects on Viral Replication Capacity and Robustness.

Elena Capel Malo

AIDS Research Institute IrsiCaixa,

Hospital Germans Trias i Pujol

2013

Thesis to obtain the PhD degree in Microbiology of the Universitat Autònoma de Barcelona

Director: Dr. Miguel Ángel Martínez de la Sierra

Tutor: Dr. Antonio Villaverde Corrales

Summary

The rapid spread of human immunodeficiency virus type 1 (HIV-1) has been accompanied by continuous extensive viral genetic diversification. Little is known about how virus diversification is influencing the viral replication capacity (RC) over time. The aim of this study was to investigate the impact of HIV-1 diversification on HIV-1 RC. HIV-1 protease (PR) and integrase (IN), two essential viral enzymes that are responsible for the maturation of the budding virion and the persistence of the viral DNA in the host cell, respectively, were analysed. In addition, the mutational robustness of the HIV-1 PR was also investigated by comparing the tolerance to single random amino acid mutations of an artificial in vitro-generated HIV-1 PR with the wild-type PR. First, the genetic variability of hundred and thirty nine HIV-1 PR, forty HIV-1 gag and forty-seven HIV-1 IN sequences from early HIV-1 naïve isolates was compared with fifty HIV-1 PR, forty-one gag and forty-seven IN sequences from late naïve isolates obtained 15 years apart. Then, thirty-three HIV-1 PR sequences were amplified from three groups of virus: two naïve sample groups isolated 15 years apart plus a third group of PR inhibitor-(PI) resistant samples. The amplified PRs were recombined with an HXB2 infectious clone and RC was determined in MT-4 cells. RC was also measured in these three groups after random mutagenesis in vitro using error-prone PCR. No significant RC differences were observed between recombinant viruses from either early or late naïve isolates (P=0.5729), even though the PRs from the late isolates had significantly lower sequence conservation scores compared with a subtype B ancestral sequence (P<0.0001). Randomly mutated recombinant viruses from the three groups exhibited significantly lower RC values than the corresponding wild-type viruses (P<0.0001). There was no significant difference regarding viral infectivity reduction between viruses carrying randomly mutated naïve PRs from early or recent sample isolates (P=0.8035). Interestingly, a significantly greater loss of RC was observed in the PI-resistant PR group (P=0.0400). These results demonstrate that PR sequence diversification has not affected HIV-1 RC or PR robustness and indicate that PRs carrying PI resistance substitutions are less robust than naïve PRs. The third study of this thesis analysed the effect of the IN genetic diversification in the virus RC. Forty-seven HIV-1 IN sequences were amplified from two groups of virus from naïve patients isolated 15 years apart. The amplified IN were recombined with a pNL4-3 infectious clone and RC was determined in CEM-GFP cells, a tat driven GFP reporter system. Significant RC differences were observed between recombinant viruses from either early or late

naïve isolates (P=0.0286), even though the IN from the late isolates had significantly lower sequence conservation scores compared with a subtype B ancestral sequence (P<0.0001). These results suggest that the IN genetic diversification over time has influenced in some cases the ex vivo viral RC Interestingly, some IN polymorphisms S17N, I72V, S119P, and D256E were found to be linked to a reduced viral RC and their additive effect might contribute to impair the IN function. Finally, the artificial evolution of the HIV-1 PR was evaluated. These results demonstrated that wild-type natural HIV-1 PR was as vulnerable to the addition of single random amino acid mutations as an artificial in vitro-generated HIV-1 PR.

Resum

La ràpida propagació del virus de la immunodeficiència humana tipus 1 (VIH-1) ha estat acompanyada d'una continua i extensa diversificació genètica vírica. Se sap poc sobre com la diversificació viral està influenciant la capacitat replicativa (CR) viral al llarg del temps. L'objectiu d'aquesta tesi va ser investigar l'impacte de la diversificació del VIH-1 sobre la seva CR. Es van analitzar dos enzims vírics essencials, la proteasa (PR) i la integrasa (IN) del VIH-1, responsables de la maduració del virió emergent i de la persistència de l'ADN víric a la cèl·lula hoste respectivament. A més, es va investigar la robustesa mutacional de la PR del VIH-1, comparant la tolerància a mutacions aleatòries úniques d'una PR del VIH-1 generada artificialment in vitro amb la PR salvatge. Primerament, es va comparar la variabilitat genètica de cent trenta-nou PR, quaranta gag i quaranta-set IN del VIH-1 procedents d'aïllats clínics naïve primerencs amb cinquanta PR, quaranta-una gag i quaranta-set IN del VIH-1 procedents d'aïllats naïve tardans obtinguts 15 anys després. A continuació, trenta-tres seqüències de la PR del VIH-1 van ser amplificades a partir de tres grups de virus: dos grups de mostres naïve aïllades amb 15 anys de diferència i un tercer grup de mostres resistents a inhibidors de PR (IP). Les PR amplificades van ser recombinades amb un clon infecciós HXB2 i la seva CR va ser determinada en cèl·lules MT4. La CR també va ser mesurada en aquests tres grups després d'haver aplicat una mutagènesi aleatòria in vitro mitjançant una PCR propensa a errors. No es van observar diferències significatives en les CR dels diferents virus recombinants tant dels aïllats naïve primerencs com tardans (P=0.5729), tot i que les PR dels aïllats tardans tenien una conservació de seqüència significativament inferior a la dels aïllats primerencs (P<0.0001). Els virus recombinants mutats aleatòriament dels tres grups mostraren una CR significativament inferior als virus salvatges corresponents (P<0.0001). No hi havia una diferència significativa en la reducció de la infectivitat viral entre els virus portadors de les PR naïve mutades aleatòriament i les mostres aïllades primer o després (P=0.8035). Notablement, es va observar una pèrdua significativament més gran de la CR en el grup de les PR resistents a IP (P=0.0400). Aquests resultats demostren que la diversificació de la seqüència de la PR no ha afectat la CR del VIH-1 o la robustesa de la PR i indiquen que les PR portadores de substitucions de resistència a IP són menys robustes que les PR naïve. Posteriorment, es van amplificar quaranta-set seqüències de la IN del VIH-1 a partir de dos grups de virus provinents de pacients naïve aïllats amb 15 anys de diferència. Les IN amplificades van ser recombinades amb un clon infecciós pNL4-3 i la CR va ser

determinada amb cèl·lules CEM-GFP, un sistema reporter de GFP determinat per tat. Es van observar diferències significatives en la CR tant dels virus recombinants dels aïllats primerencs com dels aïllats tardans (P=0.0286), tot i que les IN dels aïllats tardans tenien una conservació de seqüència significativament inferior respecte a una seqüència ancestral de subtipus B (P<0.0001). Aquests resultats suggereixen que la diversificació genètica de la IN al llarg del temps ha influenciat en alguns casos la CR viral ex vivo. Remarcablement, es va trobar que alguns polimorfismes S17N, I72V, S119P, i D256E estaven relacionats amb una reducció de la CR viral i el seu efecte additiu podria perjudicar la funció de la IN. Finalment, es va avaluar l'evolució artificial de la PR del VIH-1. Es va demostrar que la PR salvatge del VIH-1 és tant vulnerable a l'addició de mutacions úniques aleatòries com una PR artificial del VIH-1 generada in vitro.

Resumen

La rápida propagación del virus de la inmunodeficiencia humana tipo 1 (VIH-1) ha estado acompañada de una continua y extensa diversificación genética viral. Se sabe poco sobre cómo está influenciando la diversificación viral la capacidad replicativa (CR) viral a lo largo del tiempo. El objeto de este trabajo fue investigar el impacto de la diversificación del VIH-1 sobre su CR. Analizamos dos enzimas víricos esenciales, la proteasa (PR) i la integrasa (IN) del VIH-1, responsables de la maduración del virión emergente y de la persistencia del ADN vírico en la célula hospedadora respectivamente. Además, se investigó la robustez mutacional de la PR del VIH-1, comparando la tolerancia a mutaciones aleatorias únicas de una PR del VIH-1 artificial generada in vitro y la PR salvaje. Primero, se comparó la variabilidad genética de ciento treinta y nueve PR, cuarenta gag i cuarenta y siete IN del VIH-1 de aislados clínicos naïve tempranos con cincuenta PR, cuarenta y un gag y cuarenta y siete IN del VIH-1 de aislados naïve tardíos obtenidos 15 años después. A continuación, treinta tres secuencias de la PR del VIH-1 fueron amplificadas a partir de tres grupos de virus: dos grupos de muestras naïve aisladas con 15 años de diferencia y un tercer grupo de muestras resistentes a inhibidores de PR (IP). Las PR amplificadas fueron recombinadas con un clon infeccioso HXB2 y su CR fue determinada en celulas MT4. La CR también fue medida en estos tres grups después de haver aplicado una mutagénesis aleatoria in vitro mediante una PCR propensa a errores. No se observaron diferencias significativas en las CR de los diferentes virus recombinantes tanto de los aislados naïve tempranos como tardíos (P=0.5729), a pesar de que las PR de los aislados tardíos tenían una conservación de secuencia significativamente inferior a la de los aislados tempranos (P<0.0001). Los virus recombinantes mutados aleatoriamente de los tres grupos mostraron una CR significativamente inferior en los virus salvajes correspondientes (P<0.0001). No hubo una diferencia significativa respecto a la reducción de la infectividad viral entre los virus portadores de las PR naïve mutadas aleatoriamente de las muestras aisladas primero o después (P=0.8035). Notablemente, se observó una pérdida significativamente más grande de la CR en el grupo de las PR resistentes a IP (P=0.0400). Estos resultados demuestran que la diversificación de la secuencia de la PR no ha afectado a la CR del VIH-1 o a la robustesa de la PR e indican que las PR portadoras de substituciones de resistencia a IP son menos robustas que las PR naïve. Cuarenta y siete secuencias de la IN del VIH-1 fueron amplificadas a partir de dos grupos de virus provenientes de pacientes naïve aislados con 15 años de

diferencia. Las IN amplificadas fueron recombinadas con un clon infeccioso pNL4-3 y la CR fue determinada con células CEM-GFP, un sistema reporter de GFP determinado por tat. Se observó diferencias significativas en la CR tanto de los virus recombinantes de los aislados tempranos como de los aislados tardíos (P=0.0286), a pesar de que las IN de los aislados tardíos tenían una conservación de secuencia significativamente inferior respecto a una secuencia ancestral de subtipo B (P<0.0001). Estos resultados sugieren que la diversificación genética de la IN a lo largo del tiempo ha influido en algunos casos la CR viral ex vivo. Notablemente, se encontró que algunos polimorfismos S17N, I72V, S119P, y D256E estaban relacionados con una reducción de la CR viral i su efecto aditivo podría perjudicar la función de la IN. Finalmente, se evaluó la evolución artificial de la PR del VIH-1. Se demostró que la PR salvaje del VIH-1 es tan vulnerable a la adición de mutaciones únicas aleatorias como una PR del VIH-1 artificial generada in vitro.

Introduction

The discovery of the Human Immunodeficiency Virus (HIV)

In the summer of 1981, clinicians in California and New York observed among young individuals, previously healthy, homosexual men an unusual clustering of rare diseases, notably Kaposi sarcoma and opportunistic infections with *Pneumocystis carinii* pneumonia (now renamed *Pneumocystis jiroveci* pneumonia), as well as cases of unexplained, persistent lymphoadenopathy (Centers for Disease Control (CDC), 1981; Friedman-Kien, 1981). It soon became evident that these individuals had a common immunological deficit in cell-mediated immunity, resulting predominantly from a significant decrease of circulating CD4+ T cells (Gottlieb *et al.*, 1981; Masur *et al.*, 1981). These cases where the first to describe a new disease, that would later be called the Acquired Immunodeficiency Syndrome (AIDS). Early suggestions that AIDS resulted from behavior specific to gay men were largely dismissed when the syndrome was observed in distinctly different groups in the United States.

After several false leads, many investigators concluded that the clustering of AIDS cases and their occurrence in diverse risk groups could be explained only if AIDS were caused by an infectious microorganism transmitted by intimate contact, for example through sexual (Francis *et al.*, 1983) activity or blood. As with many emerging infectious diseases, the initial and most powerful tool to illuminate the etiology of the disease was classic epidemiology. Initial observations regarding the immunopathogenesis of AIDS, together with a growing understanding of human and animal retroviruses, suggested that the disease might have a retroviral etiology (Gallo, 2002; Montagnier, 2002). Two retroviruses, human T-lymphotrophic virus (HTLV-I and HTLV-II), which had been recently recognized at that time, were the only viruses known to preferentially infect CD4+ T cells. The transmission pattern of HTLV was similar to that seen among individuals with AIDS; in addition, HTLV-I and related retroviruses were known to cause varying degrees of immune deficiency in humans and animals (Varmus, 1988). Thus, the search for a new retrovirus was undertaken seriously.

In 1983, experimental data indicating an association between a retrovirus and AIDS were published by a research team in France led by Luc Montagnier (Barré-Sinoussi *et al.*, 1983). This retrovirus isolated from a biopsy of a patient with lymphoadenopathy was later called Lymphoadenopathy-Associated Virus (LAV). In 1984, researchers at the US National Institutes of Health, led by Robert C. Gallo, published seminal papers that established, with virological and epidemiological evidence, that the virus isolated from patients with AIDS or pre-AIDS, called HTLV-III by the US group, was the cause of AIDS (Gallo *et al.*, 1984; Popovic *et al.*, 1984). At the same time, the virus was also isolated independently by Jay Levy and co-workers in California from both individuals affected with AIDS and asymptomatic individuals from groups at high risk for AIDS (Levy *et al.*, 1984) and named it AIDS-associated Retorviruses (ARVs). Within a short time, the three prototype viruses (LAV, HTLV-III, and ARV) were recognized as members of the same group of retroviruses, and their properties identified them as lentiviruses (Ratner *et al.*, 1985). Finally, in 1986 the International Committee on Taxonomy of Viruses recommended giving the AIDS virus its separate name, Human Immunodeficiency Virus or HIV in its abbreviated form (Coffin *et al.*, 1986). Actually, the HTLV-III isolate was confirmed to be the same virus as the LAV isolate (Gallo, 2002; Montagnier, 2002; Wain-Hobson *et al.*, 1991), because of a contamination of the tissue samples exchanged between the French and US groups.

HIV was identified and shown to be the cause of AIDS less than 2 and half years after the disease was first identified. It took only another 2 years for blood tests to become commercially available, reducing almost to zero the transmission of AIDS through blood transfusion in developed countries. In 1987, the first anti-HIV drug, azidothymidine (AZT), a nucleoside analogue of the reverse transcriptase, which blocks HIV RT activity, was introduced. With the arrival of the HIV protease inhibitors and triple drug therapy in 1995, many patients are alive today who would otherwise have died.

AIDS epidemic

The infection caused by the now known HIV became epidemic in late 1983 and early 1984. The epidemic was first reported in the West countries, but several retrospective studies showed the origin of the infection in Central Africa (Clumck *et al.*, 1985; Curran *et al.*, 1985; Montagnier, 1985; Piot *et al.*, 1984). At the end of 2011, 34 million people around the world were living with HIV, according to the last United Nations-AIDS (UNAIDS) report on the global AIDS epidemic (‘Global report: UNAIDS report on the global AIDS epidemic 2012’, 2012). Unfortunately, the burden of the epidemic continues to vary considerably between countries and regions. Sub-Saharan Africa remains most severely affected, with nearly 1 in every 20 adults (4.9%) living with HIV and accounting for 69% of the people living with HIV worldwide (Fig 1).

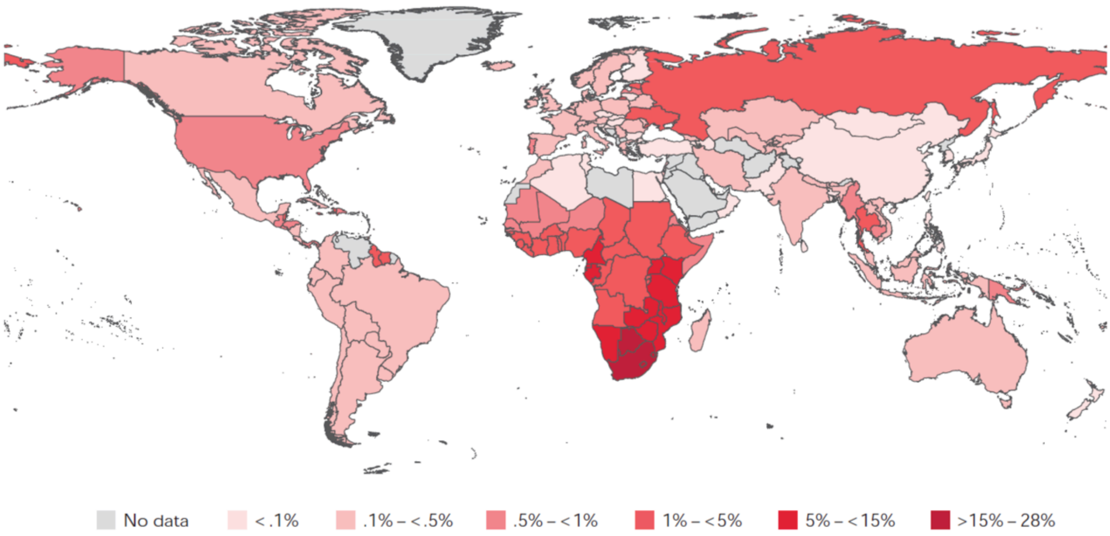


Figure 1. Global prevalence of HIV infection in 2009. Percentage of HIV adult prevalence is shown in a range of grey and red colours (see legend on the bottom). Taken from the Joint United Program on HIV/AIDS (UNAIDS) and the World Health Organization (WHO) 2010. AIDS epidemic update: November 2010 (www.unaids.org).

The number of people dying from AIDS-related causes began to decline in the mid-2000s because of scaled-up antiretroviral therapy and the steady decline in HIV incidence since the peak in 1997. In 2011, this decline continued, with evidence that the drop in the number of people dying from AIDS-related causes is accelerating in several countries. In 2011, 1.7 million people died from AIDS-related causes worldwide. This represents a 24% decline in AIDS-related mortality compared with 2005 (when 2.3 million deaths occurred).

Biology of the virus

Classification

HIV is a retrovirus that belongs to the *Lentivirus* genus. The mature HIV virion carries two copies of a single-stranded, positive-sense, and enveloped RNA. HIV comprises two distinct viruses, HIV-1 and HIV-2, which differ in origin and gene sequence. However, both viruses cause AIDS with a similar spectrum of symptoms, though Central Nervous System disease may be more frequent in HIV-2 disease (Lucas *et al.*, 1993). HIV-1 is more virulent, more infective (Gilbert *et al.*, 2003), and is the cause of the majority of HIV infections globally. It appears that HIV-2 infection takes longer to progress to AIDS (Whittle *et al.*, 1994). The lower infectivity of HIV-2 compared to HIV-1 implies that fewer of those exposed to HIV-2 will be infected per exposure.

Because of its relatively poor capacity for transmission, HIV-2 is largely confined to West Africa (Reeves & Doms, 2002). HIV-2 was first isolated in 1986 (Clavel *et al.*, 1986).

HIV-1 is closely related to Simian Immunodeficiency Virus of chimpanzees (SIVcpz). It is classed phylogenetically into three groups -M (for main), N (for non-M, non-O) and O (for outlier) (Robertson *et al.*, 2000)- which differ from each other in genetic sequence as much as each does from SIVcpz, indicating that each group represents a separate chimpanzee-to-human transfer (Fig 2). HIV-2, in contrast, resembles Simian Immunodeficiency Virus of the sooty mangabey monkey (SIVsm), with at least six separate transfers of this virus to humans. Whereas HIV-1 groups N and O remain localized in Gabon and Cameroon, HIV-2 is present mainly in West Africa (with some spread to Europe and India) and HIV-1 group M has given rise to the worldwide pandemic, diverging into various clades or subtypes, known as A–K and into circulating recombinant forms (CRF). Globally, subtype C is now the most successful of the HIV-1 M lineages and today accounts for approximately 50% of infections, whereas subtypes A and B each account for over 10% of worldwide HIV infections. Subtype C mostly occurs in Southern Africa and Asia, whereas subtype A is mainly distributed in eastern Europe and central Asia, and subtype B dominates in North America, the Caribbean, Latin America, western and central Europe and Australia. Subtypes D and G, CRF01_AE, and CRF02_AG account for only between 2% and 6% each. Subtypes F, H, J, K, other CRFs, and all other unclassified recombinant forms individually make only a minor contribution to the global HIV population (<1% each) but together account for the remaining 15% of worldwide HIV infections (Hemelaar *et al.*, 2011). It is not yet clear what has made HIV-1 M fitter for pandemic spread (Weiss, 2003).

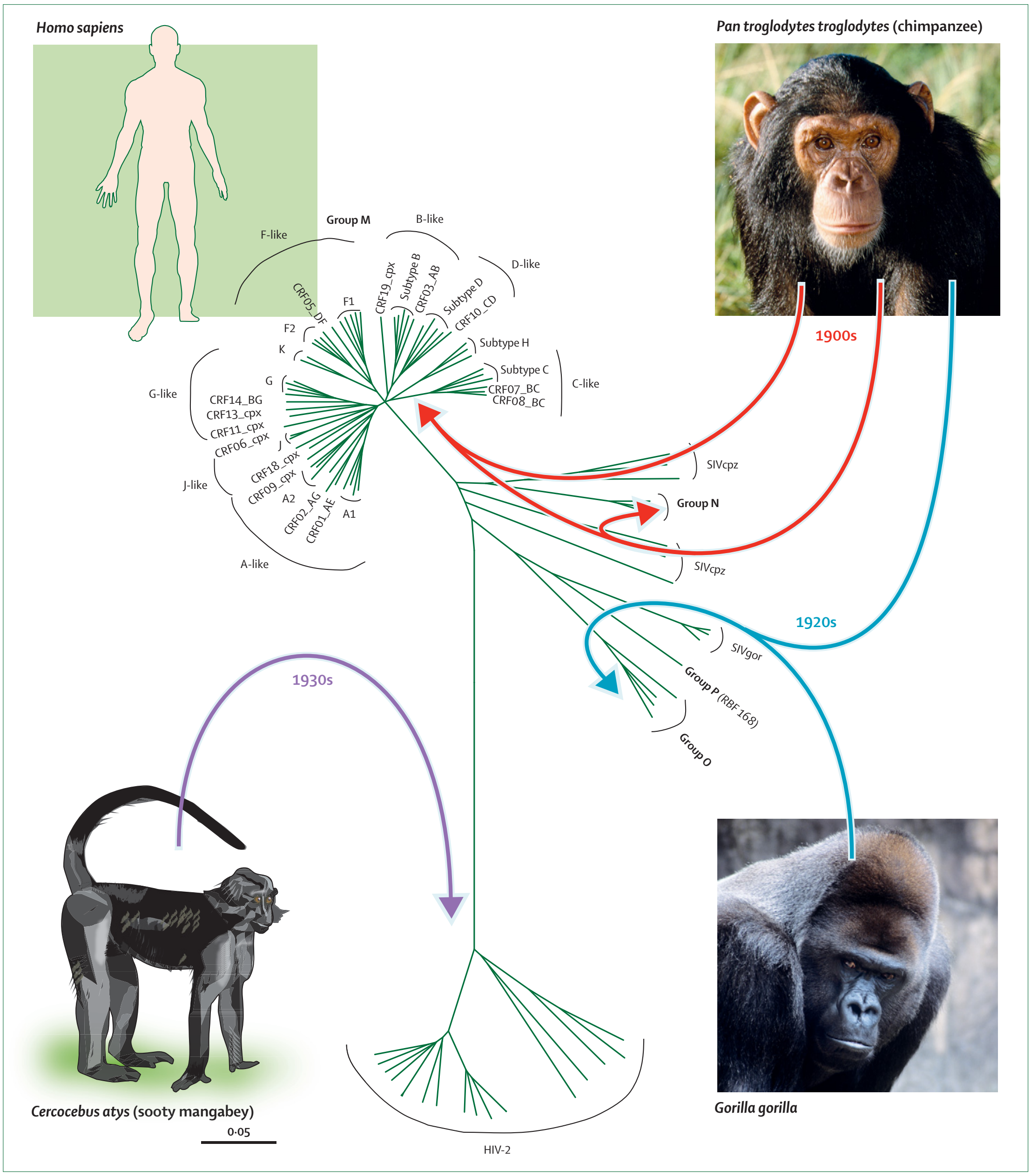


Figure 2: Relations between and genetic diversity in HIV-1 groups M, N, O, and P, HIV-2, and SIVs, and patterns of cross-species transmission. CRF = circulating recombinant form; cpz = chimpanzee; gor = gorilla; cpx = complex; SIV = simian immunodeficiency virus. Obtained from (Tebit & Arts, 2010).

Origins of the HIV and AIDS epidemic

Since HIV-1 was first discovered, the reasons for its sudden emergence, epidemic spread, and unique pathogenicity have been a subject of intense study. The discovery of an antigenically distinct virus, termed HIV-2, that was found to cause AIDS in patients in western Africa (Clavel *et al.*, 1986) gave the first clue of the possible emergence of AIDS from cross-species infections with lentiviruses from different primate species (Sharp *et al.*, 1994). HIV-2 was only distantly related to HIV-

1, but was closely related to a simian virus that caused immunodeficiency in captive macaques (Chakrabarti *et al.*, 1987; Guyader *et al.*, 1987). Moreover, close simian relatives of HIV-1 and HIV-2 were found in chimpanzees (Huet *et al.*, 1990) and sooty mangabeys (Hirsch *et al.*, 1989), respectively. Soon thereafter, additional viruses, collectively termed simian immunodeficiency viruses (SIVs) with a suffix to denote their species of origin, were found in various different primates from sub-Saharan Africa, including African green monkeys, sooty mangabeys, mandrills, chimpanzees, and others. Surprisingly, these viruses appeared to be largely nonpathogenic in their natural hosts, despite clustering together with the human and simian AIDS viruses in a single phylogenetic lineage within the radiation of lentiviruses (Fig 2). These relationships provided the first evidence that AIDS had emerged in both humans and macaques as a consequence of cross-species infections with lentiviruses from different primate species (Sharp *et al.*, 1994). Indeed, subsequent studies confirmed that SIVmac was not a natural pathogen of macaques (which are Asian primates), but had been generated inadvertently in US primate centers by inoculating various species of macaques with blood and/or tissues from naturally infected sooty mangabeys (Apetrei *et al.*, 2005; 2006). Similarly, it became clear that HIV-1 and HIV-2 were the result of zoonotic transfers of viruses infecting primates in Africa (Hahn *et al.*, 2000).

Although there is compelling evidence that both HIV types emerged from two dissimilar SIVs in separate geographical regions of Africa, each of the two HIVs has its own simian progenitor and specific genetic precursor, and all of the primates that carry these SIVs have been in close contact with humans for thousands of years without the emergence of epidemic HIV. Some modern event must have aided in the transition of SIV to HIV. The research held by Marx and co-workers (Marx *et al.*, 2001) revealed that serial passage of partially adapted SIV between humans could produce the series of cumulative mutations sufficient for the emergence of epidemic HIV strains. Among all these many primate-to-human lentivirus transfers occurred in recent history, the only one for which we have a reasonably accurate starting time is the pandemic strain, HIV-1 group M. The first known positive human sample dates from 1959 in Kinshasa, Zaire, but from detailed phylogenetic studies of existent strains, a date for the species jump can be estimated as 1931 ± 12 years (Korber, 2000).

Structure and genome

Like all retroviruses HIV contains a genome composed of two copies of single stranded positive-sense RNA of approximately 9.8 kilobases (kb) enclosed in a cone-shaped core surrounded by a membrane envelope. The HIV-1 genome encodes nine open reading frames but fifteen proteins are made in all (Fig 3)(Frankel & Young, 1998). Both ends of the provirus are flanked by a repeated sequence known as the long terminal repeats (LTRs), which are required for the proviral DNA to integrate to the host-cell DNA and they have binding sites for the transcription factors necessary to express the viral genes. Some genes are translated into large polyproteins, Gag, Pol and Env, which are then cleaved by a virus-encoded protease as well as cellular proteases into smaller proteins. The four Gag proteins, MA (matrix or p17), CA (capsid or p24), NC (nucleocapsid or p7), and p6, and the two Env proteins, SU (surface or gp120) and TM (transmembrane or gp41), are structural components that make up the core of the virion and outer membrane envelope. MA forms the inner shell in the particle just below the viral membrane, CA forms the conical core enclosing the viral genomic RNA, and NC interacts with viral RNA inside the capsid. The envelope glycoproteins are responsible of the recognition of the target-cell (gp120) and promote the fusion of viral and cellular membranes (gp41) that result in the release of the viral contents into the host cell. They are made from the precursor gp160, which is a singly spliced message from the full-length viral mRNA and cellular enzymes mediate the proteolytic cleavage of gp160. The three Pol proteins, PR (protease), RT (reverse transcriptase), and IN (integrase), provide essential enzymatic functions and are also encapsulated within the particle. HIV-1 encodes six additional proteins, often called accessory proteins, three of which (Vif, Vpr, and Nef) are found in the viral particle. Two other accessory proteins, Tat and Rev, provide essential gene regulatory functions, and the last protein, Vpu, indirectly assists in assembly of the virion.

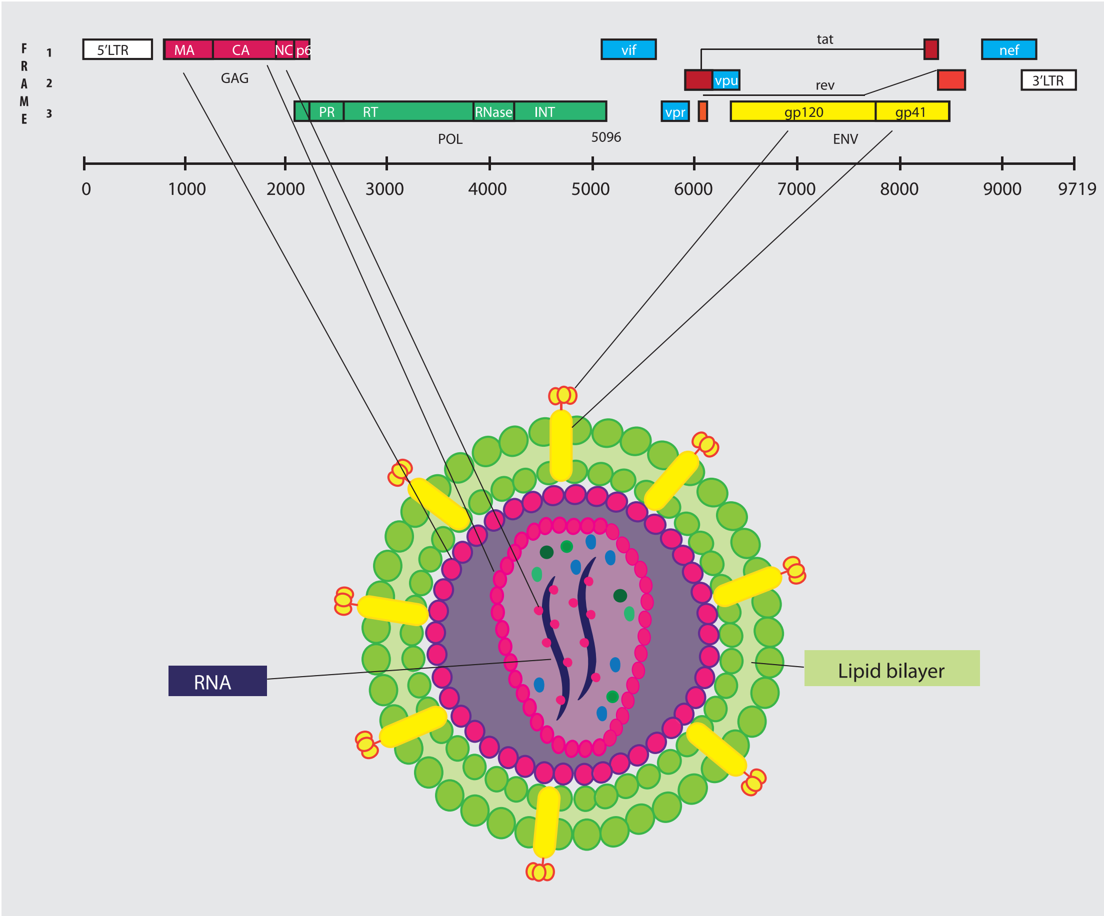


Figure 3: Organisation of the HIV-1 genome and mature virion. A schematic diagram of the HIV-1 HXB2 genome is shown at the top of the figure. A schematic diagram of the mature HIV-1 virion is shown below. Modified from (Frankel & Young, 1998).

Tropism

HIV is a retrovirus that infects preferentially CD4+ T lymphocytes and causes their destruction, leading to an immunological failure of the infected person. Although it can infect a variety of immune cells such as macrophages and microglial cells. Entry of HIV-1 into a host cell is a multi-step process, with the viral envelope gp120 and gp41 acting sequentially to mediate the viral attachment, CD4 cell receptor binding, chemokine co-receptor binding (CCR5 or CXCR4), and fusion of the viral and host membranes (Kwong *et al.*, 1998). Virus tropism is defined by the use of chemokine co-receptor: T cell line-tropic (TCL-tropic), generally syncytium-inducing that use CXCR4 (called X4 virus); macrophage-tropic (M-tropic), or non-syncytium-inducing that use CCR5 (called R5 virus); or dual-tropic, that replicate efficiently in both target cell types (called dual X4R5). However, the use of co-receptor alone does not explain viral tropism, as not all R5 viruses are able to use CCR5 on macrophages for a productive infection (Coakley *et al.*, 2005) and HIV-1 can also infect a subtype of myeloid dendritic cells (Knight *et al.*, 1990), which probably constitute a reservoir that maintains infection when CD4^{+} T cell numbers have declined to extremely low levels.

Genetic findings have yielded major insights into the in vivo roles of individual co-receptors and their ligands; of particular importance is the discovery of an inactivating mutation in the CCR5 gene, the CCR5-Δ32 mutation, a null allele resulting from a 32 base pair deletion in the open reading frame of CCR5, which in homozygous form confers strong resistance to HIV-1 infection with R5 virus, as the mutation stops HIV-1 from binding to this co-receptor, reducing its ability to infect target cells (Dean *et al.*, 1996; Liu *et al.*, 1996; Samson *et al.*, 1996). Actually, this has been confirmed recently by the report of the only patient that has been cured of its HIV infection (Hütter *et al.*, 2009). The well-known Berlin patient that was HIV-1 positive for more than 10 years was diagnosed with acute myelogenous leukemia (AML) in Germany. His doctor, Dr. Hütter decided to perform a bone marrow transplantation that matched the patient HLA along with the Δ32 mutation. The patient has been followed and to date no signs of viral relapse have been reported and he continues to be seronegative, although he was seropositive prior to bone marrow transplantation.

Replication cycle

HIV-1 life cycle follows different steps (Fig 4). First, the virus surface envelope protein gp120 recognizes and binds to the target cell CD4 receptor, which promotes a conformational change that elicits the virus to bind the chemokine co-receptor, CCR5 or CXCR4 (Kwong *et al.*, 1998). This last contact allows the membrane fusion by the transmembrane envelope protein gp41 and the internalisation of the virus into the host cell. After fusion, the nucleocapsid loses its structure and its content is released into the cytoplasm, where the dimeric single-stranded RNA is copied by the virion reverse transcriptase into a complementary double-stranded DNA (cDNA). This molecule enters then the cell nucleus, where the virion integrase covalently joins the viral DNA to cellular DNA, creating the integrated provirus. Integration predominantly takes place in transcriptionally active regions of the genome (Schröder *et al.*, 2002). The provirus can be in a latent state (inactive), or undergo active viral production, depending on the activation state of the cellular polymerases, being transcribed by the host cell machinery to give rise to new HIV-1 genomes and mRNAs. Viral mRNAs are translated into regulatory and “accessory” proteins, such as Nef, Tat, Rev, Vpu and polyprotein precursors of structural genes, such as *gag*, *gag-pol*, and *env*. Finally, these viral proteins synthesized in the cytoplasm and the new viral RNA are assembled into new virions in lipid rafts on cellular membranes (Nguyen & Hildreth, 2000) and bud from the host cell. The further maturation of virions occurs after the formation of active protease dimers, which cleave Gag and Pol polyprotein precursors into their functional subunits. The virus assumes its mature shape with a clearly defined inner cone-shaped core and outer dodecahedral envelope.

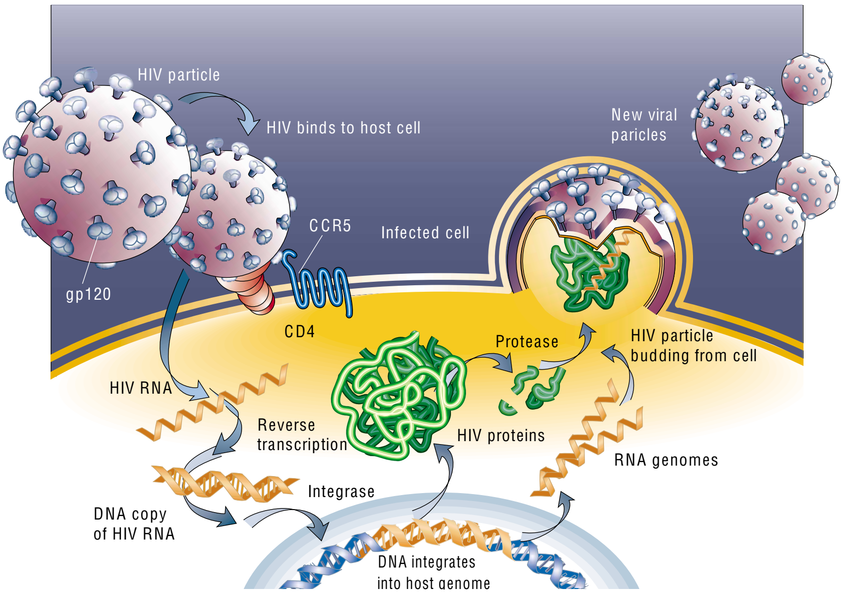


Figure 4: The HIV replication cycle. Reproduced from (Weiss, 2003).

Clinical course of HIV infection

The natural course of an HIV infection is divided into three clinical stages (Fauci *et al.*, 1996) (Fig 5):

‣ **Acute phase:** Soon after HIV-1 enters the body, it is widely disseminated, predominantly to lymphoid tissues (Pantaleo *et al.*, 1993), where HIV-1 virus infects a large number of CD4 cells and replicates rapidly. This is explained by a sharp rise of HIV RNA in blood (viral load,VL) and a consequent depletion of CD4+ T cells. The primary symptoms correspond to flu-like symptoms.

‣ **Asymptomatic phase:** Two to four weeks after exposure to the virus, the immune system fights back with killer T cells (CD8+ T lymphocyte cells) and B-lymphocyte-cell-produced antibodies. During this time, HIV-1 levels in the blood drop dramatically (Tindall & Cooper, 1991) and CD4 cell counts rebound. However, the resulting immune response to suppress the virus is only partially successful and some virus escapes, hiding and lying dormant in infected cells for months or years in a state of chronic, persistent viral replication ensues.

‣ **AIDS phase:** After a period of time, the immune system deteriorates to the point where the body is unable to fight off other infections and AIDS develops. The HIV-1 viral load in the blood dramatically increases while the number of CD4 cells drop to dangerously low levels (fewer than 200 CD4 cells per mm^3 of blood). When this happens, symptomatic diseases and opportunistic infections arise, that can be the cause of death.

Since the use of highly active antiretroviral treatment (HAART) the asymptomatic phase has been extended. The CD4 cell count and the plasma VL (pVL) determine the course of HIV-1 infection. Thus, monitoring of CD4 counts and pVL is important at all stages of infection, as it is used to help clinicians determine when to start preventive chemotherapy for opportunistic infections and when to start antiretroviral therapy (ART). But progression to AIDS among HIV-1 infected individuals is highly heterogeneous due to host and viral factors (Casado *et al.*, 2010; Dalmau *et al.*, 2009).

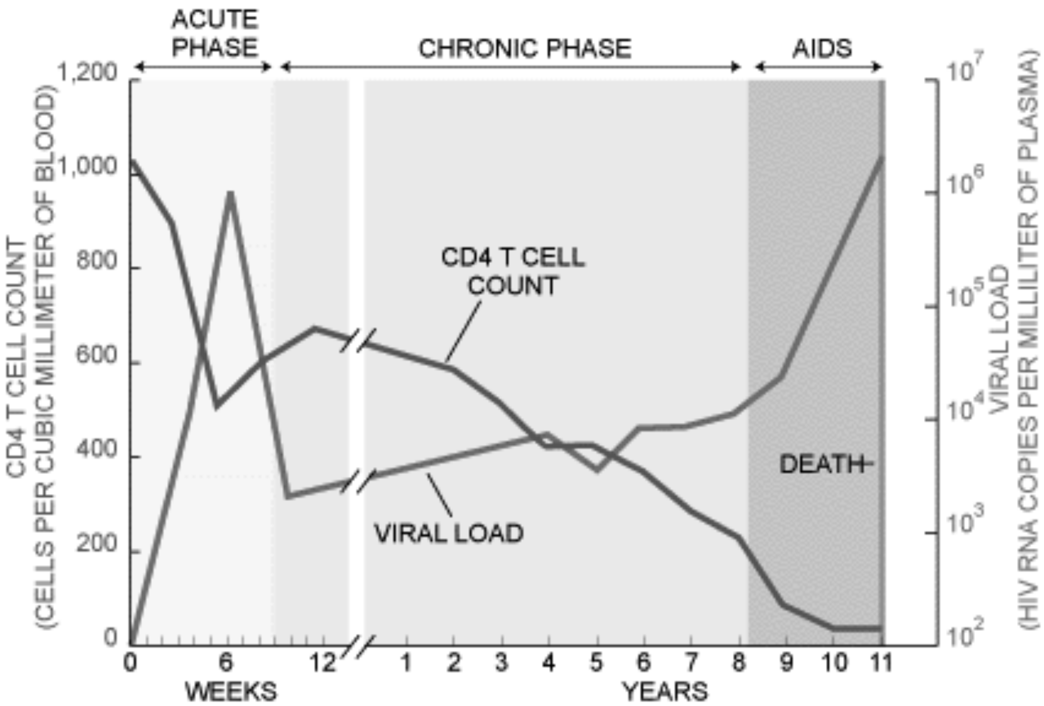


Figure 5: HIV clinical course of infection. Image source: http://www.scientificamerican.com

Genetic variability

As an RNA virus, HIV virus population does not consist of a single genotype, rather, it is an ensemble of related sequences, termed quasispecies (Domingo & Holland, 1997; Eigen, 1993; Holland *et al.*, 1982; 1992). Thirty-five years ago, Manfred Eigen and Peter Schuster described quasispecies as “a given distribution of macromolecular species with closely interrelated sequences, dominated by one or several (degenerate) master copies” (Eigen & Schuster, 1977). In other words, self-replicating entities, such as RNA viruses, exist as a cloud of related genotypes. Quasispecies arise from rapid genomic evolution from an initially limited number of infectious particles powered by the high mutation rate of RNA viral replication (Domingo *et al.*, 1978). Although a high mutation rate is dangerous for a virus because it results in nonviable individuals, it has been hypothesized that high mutation rates create a ‘cloud’ of potentially beneficial mutations at the population level, which afford the viral quasispecies a greater probability to evolve and adapt to new environments and challenges during infection (Coffin, 1995; Eigen, 1993).

Genetic diversity remains one of the major obstacles to eradication of HIV. The viral quasispecies can respond rapidly to selective pressures, such as that imposed by the immune system and antiretroviral therapy, and frustrates vaccine design efforts. Although influenza virus has been the paradigm of a variable virus, yet the HIV population present in a single individual, six years after infection, can be as great as the global variation for an influenza outbreak (Fig 6) (Korber *et al.*, 2001). Two unique features of retroviral replication are responsible for the variation generated during infection. First, mutations are frequently introduced into the viral genome by the error prone viral reverse transcriptase which generates a high rate of incorrect nucleotide substitutions (10^{-4} to 10^{-5} mutations per nucleotide and per replication cycle) (Mansky & Temin, 1995) and through the actions of host cellular factors, such as the APOBEC family of nucleic acid editing enzymes. Second, the HIV reverse transcriptase can utilize both copies of the co-packaged viral genome in a process termed retroviral recombination (Jung *et al.*, 2002). When the co-packaged viral genomes are genetically different, retroviral recombination can lead to the shuffling of mutations between viral genomes in the quasispecies (Smyth *et al.*, 2012). In addition, the rapid viral turnover of CD4 T cells contributes to accelerate viral replication (Ho *et al.*, 1995; Wei *et al.*, 1995) and the generation of genetic variability.

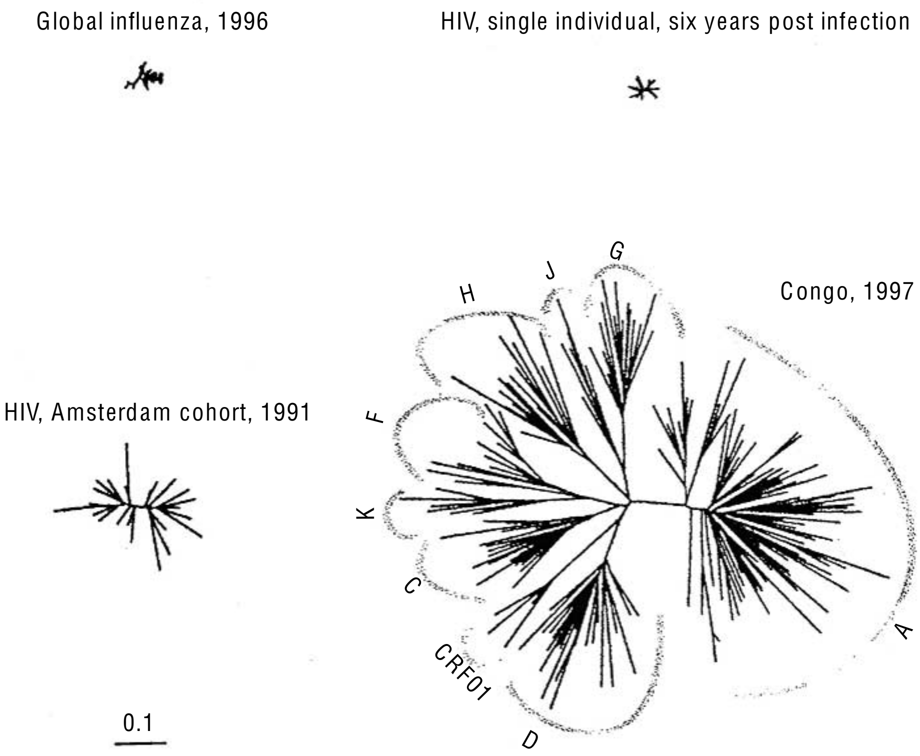


Figure 6: The scale of HIV variation. Sequence divergence of envelope glycoproteins of HIV (gp120 V2-C5) compared with that of influenza A H3 (HA1). The length of the spokes indicates the degree of divergence, with the scale shown. HIV variation in a single person six years after infection (nine genomes analysed) is similar to that of worldwide influenza A (96 genomes) in a single year. The greatest amount of variation is in the Democratic Republic of Congo, where HIV first developed and has diversified into subtypes A–K (except for subtype B, which is prevalent in the West, and E, which is prevalent in Thailand). CRF01, circulating recombinant form. [(Weiss, 2003) Adapted from(Korber *et al.*, 2001)]

Antiretroviral treatment and drug resistance development

Antiretroviral treatment targets the different phases of the HIV-1 replication cycle. In 1987, the US Food and Drug Administration (FDA) approved the first antiretroviral drug, a nucleoside analogue of the reverse transcriptase, known as azidothymidine (AZT) (Wright, 1986; Young, 1988). The use of AZT managed to reduce viral replication and contributed to reduce AIDS morbidity at that time. But soon after, the emergence of HIV-1 strains with reduced drug susceptibility (Larder *et al.*, 1989), also known as resistance variants, urged the development of new antiretroviral drugs (Myers, 1990). The treatment of HIV-1 infection was revolutionized in the mid-1990s by the development of inhibitors of the reverse transcriptase and protease, two of three essential enzymes of HIV-1, and the introduction of drug regimens that combined these agents to enhance the overall efficacy and durability of therapy. The introduction of combination therapy, also known as HAART, for the treatment of HIV-1 infection had a huge impact in reducing the morbidity and mortality associated with HIV-1 infection and AIDS (Collier *et al.*, 1996; D'Aquila *et al.*, 1996; Staszewski *et al.*, 1996). The key of this success was the suppression of viral replication and the reduction of the plasma HIV-1 viral load to undetectable levels of the most sensitive clinical assays (<50 RNA copies/ mL) resulting in a significant reconstitution of the immune system (Autran, 1997; Komanduri *et al.*, 1998; Lederman *et al.*, 1998) as measured by an increase in circulating CD4+ T-lymphocytes. Importantly, combination therapy using three antiretroviral agents directed against at least two distinct molecular targets is the underlying basis for preventing the evolution drug resistance.

Recently a large number of inhibitors targeting different steps of HIV-1 cycle have been developed. To date, an arsenal of 24 FDA-approved drugs are available for treatment of HIV-1 infections (Arts & Hazuda, 2012). These drugs are distributed into six distinct classes based on their molecular mechanism and resistance profiles: (1) nucleoside-analog reverse transcriptase inhibitors (NRTIs), (2) non–nucleoside reverse transcriptase inhibitors (NNRTIs), (3) integrase inhibitors (INIs or INSTIs), (4) protease inhibitors (PIs), (5) fusion inhibitors, and (6) coreceptor antagonists.

‣ **Reverse transcriptase inhibitors:** NRTIs are nucleoside-analogs that compete with cellular nucleotides and are responsible for the termination of the growing viral DNA chain, whereas NNRTIs form a hydrophobic pocket in the RT that reduces its polymerase activity.

Table 1: Reverse transciptase inhibitors approved by the US FDA and under development. Modified from www.fda.gov and www.aidsmeds.com.

Brand Name	Generic Name	Manufacturer Name	Approval Date
		Multi-class Combination Products	
Atripla	efavirenz, emtricitabine and tenofovir disoproxil fumarate	Bristol-Myers Squibb and Gilead Sciences	12-July-06
Complera	emtricitabine, rilpivirine, and tenofovir disoproxil fumarate	Gilead Sciences	10-August-11
Stribild	elvitegravir, cobicistat, emtricitabine, tenofovir disoproxil fumarate	Gilead Sciences	27-August-12
		Nucleoside Reverse Transcriptase Inhibitors (NRTIs)	
Combivir	lamivudine and zidovudine	GlaxoSmithKline	27-sep-97
Emtriva	emtricitabine, FTC	Gilead Sciences	2-jul-03
Epivir	lamivudine, 3TC	GlaxoSmithKline	17-nov-95
Epzicom	abacavir and lamivudine	GlaxoSmithKline	02-Aug-04
Hivid	zalcitabine, dideoxycytidine, ddC (no longer marketed)	Hoffmann-La Roche	19-jun-92
Retrovir	zidovudine, azidothymidine, AZT, ZDV	GlaxoSmithKline	19-mar-87
Trizivir	abacavir, zidovudine, and lamivudine	GlaxoSmithKline	14-nov-00
Truvada	tenofovir disoproxil fumarate and emtricitabine	Gilead Sciences	02-Aug-04
Videx EC	enteric coated didanosine, ddI EC	Bristol Myers-Squibb	31-oct-00
Videx	didanosine, dideoxyinosine, ddI	Bristol Myers-Squibb	9-oct-91
Viread	tenofovir disoproxil fumarate, TDF	Gilead Sciences	26-oct-01
Zerit	stavudine, d4T	Bristol Myers-Squibb	24-jun-94
Ziagen	abacavir sulfate, ABC	GlaxoSmithKline	17-Dec-98
Amdoxovir	AMDX, DAPD	RFS Pharma	In development
Tenofovir alafenamide fumarate (TAF)	GS 7340	Gilead Sciences	In development
		Non-nucleoside Reverse Transcriptase Inhibitors (NNRTIs)	
Edurant	rilpivirine, RVP	Tibotec Therapeutics	20-may-11
Intelence	etravirine, ETR	Tibotec Therapeutics	18-Jan-08
Rescriptor	delavirdine, DLV	Pfizer	4-Apr-97
Sustiva	efavirenz, EFV	Bristol Myers-Squibb	17-sep-98
Viramune (Immediate Release)	nevirapine, NVP	Boehringer Ingelheim	21-jun-96
Viramune XR (Extended Release)	nevirapine, NVP	Boehringer Ingelheim	25-mar-11
UK-453061	Lersivirine	ViiV Healthcare	In development

- **Integrase inhibitors:** All integrase inhibitors target the strand transfer reaction catalyzed by the viral integrase. They are thus referred to as either INIs or more specifically, integrase strand transfer inhibitors (INSTIs) (Espeseth *et al.*, 2000; Hazuda *et al.*, 2004a, b; McColl & Chen, 2010). Basically, INSTIs have two mechanisms of action: (1) bind only to the specific complex between integrase and the viral DNA and (2) interact with the two essential magnesium metal ion cofactors in the integrase active site and also the DNA (Arts & Hazuda, 2012).

Table 2: Integrase inhibitors approved by the US FDA and under development. Modified from www.fda.gov and www.aidsmeds.com.

HIV integrase strand transfer inhibitors (INSTIs)			
Brand Name	**Generic Name**	**Manufacturer Name**	**Approval Date**
Isentress	raltegravir	Merck & Co., Inc.	12-oct-07
GSK-572	Dolutegravir	ViiV Healthcare and Japan-based Shionogi & Co.	In development
Stribild	Elvitegravir	Gilead Sciences	In development

▸ **Protease inhibitors:** PIs compete with the protease natural substrate, preventing the correct function of the enzyme. Thus preventing the maturation of new budding virions that leads to the formation of non-infectious particles.

Table 3: Protease inhibitors approved by the US FDA and under development. Modified from www.fda.gov and www.aidsmeds.com.

Protease Inhibitors (PIs)			
Brand Name	**Generic Name**	**Manufacturer Name**	**Approval Date**
Agenerase	amprenavir, APV (no longer marketed)	GlaxoSmithKline	15-Apr-99
Aptivus	tipranavir, TPV	Boehringer Ingelheim	22-jun-05
Crixivan	indinavir, IDV,	Merck	13-mar-96
Fortovase	saquinavir (no longer marketed)	Hoffmann-La Roche	7-nov-97
Invirase	saquinavir mesylate, SQV	Hoffmann-La Roche	6-Dec-95
Kaletra	lopinavir and ritonavir, LPV/RTV	Abbott Laboratories	15-sep-00
Lexiva	Telzir, fosamprenavir, FPV	GlaxoSmithKline	20-oct-03
Norvir	ritonavir, RTV	Abbott Laboratories	1-mar-96
Prezista	darunavir, DRV	Tibotec, Inc.	23-jun-06
Reyataz	atazanavir sulfate, ATV	Bristol-Myers Squibb	20-jun-03
Viracept	nelfinavir mesylate, NFV	Agouron Pharmaceuticals	14-mar-97

▸ **Entry inhibitors:** Two types of entry inhibitors have been developed depending on the step of the viral cell cycle they are implicated. Fusion inhibitors prevent the interaction of the two subunits of the envelope transmembrane protein, gp41, thus impeding the fusion of the viral envelope with the cellular membrane. Whereas CCR5 coreceptor antagonists act as allosteric inhibitors by altering the conformation of the second extracellular loop of the receptor and prevents interaction with the V3 stem loop of the envelope surface protein, gp120.

Table 4: Entry inhibitors approved by the US FDA and under development. Modified from www.fda.gov and www.aidsmeds.com.

Brand Name	Generic Name	Manufacturer Name	Approval Date
		Fusion Inhibitors	
Fuzeon	enfuvirtide, T-20	Hoffmann-La Roche & Trimeris	13-mar-03
		CCR5 coreceptor antagonist	
Selzentry	maraviroc	Pfizer	06-aug-07
TBR-652, TAK-652	Cenicriviroc	Tobira Therapeutics	In development
TMB-355	Ibalizumab	Taimed Biologics	In development
PRO 140	-	Progenics Pharmaceuticals, Inc.	In development

Although the success of HAART in reducing AIDS morbidity, drug resistance continues to be documented in patients failing therapy as well as in therapy-naïve patients infected with transmitted drug-resistant viruses. Moreover, all the anti-HIV compounds present long-term toxicity and adverse effects (Carr, 2003), lowering the adherence to treatment, what leads to a suboptimal concentration of the compounds and subsequent viral failure with development of resistances. The drug-resistance emergence together with the latency of HIV and the presence of viral reservoirs (Stevenson, 2003), where the drugs cannot achieve the optimal concentrations, make the current treatments unable to eradicate the virus from infected individuals. It has been demonstrated that viral replication persists in infected patients under suppressive HAART and drives immune activation (Buzón *et al.*, 2010a; Sigal *et al.*, 2011). The International AIDS Society has convened a group of international experts to develop a strategy for research towards an HIV cure (Deeks, 2012). The main goal nowadays is to eradicate viral latency.

Virus fitness

A replicative fitness cost is associated with nearly all RTI, PI and INI resistance mutations when the respective drug is absent. HIV-1 resistant variants are typically present in the infecting virus population (‘‘swarm’’ or ‘‘quasispecies’’) prior to treatment but are maintained at low frequency due to their low fitness. Fitness is the parameter that defines the replicative capacity of the virus in a given environment (Quiñones-Mateu *et al.*, 2008). During viral replication within a defined microenvironment, different genomes encode virus that replicate at high rates, continually mutate, but generally remain under the same selective pressures (Domingo *et al.*, 1999). Positive (Darwinian) selection implies that one or more members of the quasispecies are better suited to a given environment, whereas negative selection eliminates unfit variants (Domingo & Holland, 1997; Domingo *et* *al.*, 1996; 1999). In the case of HIV, each individual member of the quasispecies has an intrinsic growth rate, known as replicative fitness.

Selective pressure in the form of drug therapy often leads to dramatic shifts in the quasispecies distribution, as a virus that was poorly fit in the absence of drug can rapidly emerge as the most fit in the presence of drug (Coffin, 1995; Domingo & Holland, 1997; Domingo *et al.*, 1997). During this in vivo selection, several drug-resistant variants may emerge and compete for dominance. These resistant isolates will pass through the drug-induced bottleneck and initiate a new quasispecies distribution that will again be governed by replication efficiency, now in the presence of drugs (Coffin, 1995; Loveday & Hill, 1995). For example, an HIV-1 clone harboring an M184V mutation in the RT coding region is likely present at very low frequency in the intrapatient HIV-1 population due to its low fitness. Upon administration of 3TC (or lamivudine), this 3TC-resistant, M184V HIV-1 variant is immediately selected in the population and can be considered the most ‘‘fit’’ clone in this environment. Several studies have suggested the presence of a lower viral load in a 3TC treated patient harboring M184V HIV-1 variant as compared to an untreated patient harboring ‘‘wild-type’’ HIV-1 (Deval *et al.*, 2004; White *et al.*, 2002). The M184V HIV-1 is basically insensitive to 3TC so this reduction in virus load is attributed to a reduced replicative capacity of the virus.

Virulence is typically defined as the rate in host mortality as a consequence of infection (Bull, 1994), which can be further refined to reproduction rate and pathogenic potential of the parasite (Bremermann & Pickering, 1983). In contrast, a parasite’s fitness is dependent on its survival and adaptability in a given environment. Hence, there is often confusion between the principles of virulence and fitness when applied to the interaction and survival of both parasite and host. Viruses are obligate parasites that require a living cell for reproduction and survival. Thus, higher fitness within a host is dependent on mechanisms that enhance spread between susceptible host cells such as improved replication efficiency and increased transmission efficiency. In HIV research, a topic that is often overlooked is the impact of fitness on HIV transmission, disease progression, evolution, and prevalence in the human population. The rate of new infections continues to increase, people progress to AIDS and die, and yet little is still known about the phenotypic differences between the heterogeneous etiological agent. Recent studies suggest that the nature of the virus itself, and not solely manifestations of host factors and the immune response is contributing to HIV-1 disease progression

(Quinones-Mateu *et al.*, 2000). Therefore, in this thesis, we aim to better know how HIV-1 is evolving among naïve infected patients over time. We will focus on two key enzymes for the virus, the protease and the integrase.

HIV-1 protease

The HIV-1 protease is an aspartic protease of 11kDa consisting of 2 identical 99-amino acid monomers (Miller *et al.*, 1989; Navia *et al.*, 1989; Wlodawer *et al.*, 1989). The dimer is stabilized by a four-stranded antiparallel β sheet formed by amino (N)- and carboxyl (C)-terminal β-strands. The enzyme active site is formed at the interface of the two subunits and contains a catalytic triad (Asp25-Thr26-Gly27) responsible for the cleavage reactions of PR. Each monomer contains a "flap" comprising two antiparallel β-strands connected by a β-turn (residues 49 to 52) and situated on top of the catalytic site (Fig 7). The conformation of the flap differs significantly in the PR and PR-inhibitor complexes, with some backbone C^{α} atoms being displaced by up to 7 Å (angstrom) (Miller *et al.*, 1989).

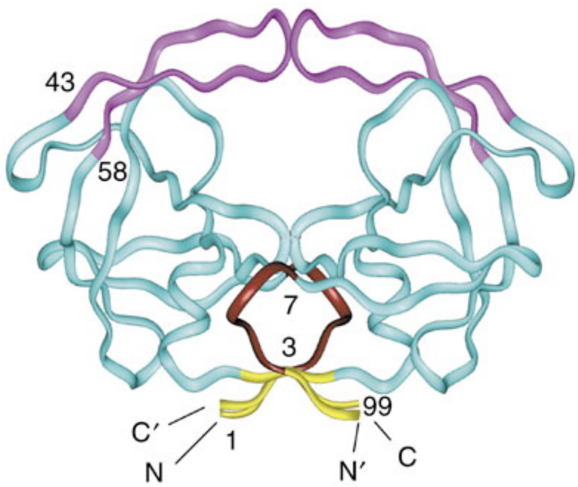


Figure 7: Ribbon drawing depicting the backbone crystal structure of the free HIV-1 protease. Segments of the structure that contain residues that are flexible in solution on the millisecond‐microsecond timescale are color-coded. In the terminal domain residues 4‐7 in the autoproteolysis-sensitive loop are in red and residues 1‐3 and 96‐99, which form the interfacial four-stranded β sheet, are in yellow. In the flap domain residues 43‐58, which cover the substrate-binding site, are in purple. The drawing was generated using the program Insight II (Molecular Simulations Inc., San Diego) and the heavy-atom coordinates (PDB accession code 3PHV). Obtained from (Ishima *et al.*, 1999).

In summary, HIV-1 protease is formed by five functional conserved domains: (1,2) the amino and carboxyl terminal residues (residues 1-9, 94-99 respectively) which are involved in the protease dimer stabilization, (3) the catalytic site (residues 21-

32), (4) the flap situated on top of the catalytic site (residues 44-56), and (5) the substrate binding site (residues 78-88) (Fig 8).

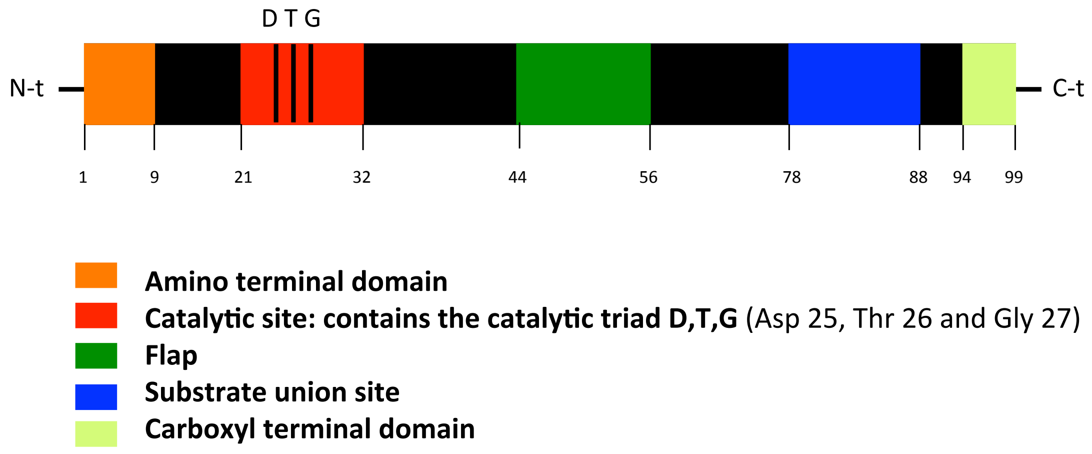


Figure 8: HIV-1 protease conserved regions and residues. Conserved regions of the amino acid sequence of HIV-1 protease (99 amino acids) of clade B consensus (shown as a reference) are boxed.

The viral encoded protease regulates proteolytic processing of virion components during particle assembly. The active protease is capable of recognizing and cleaving a diverse array of amino acid sequences; at least 11 cleavage sites within the Gag and Gag-Pol polyproteins have been reported and a single site in Nef (de Oliveira *et al.*, 2003). The primary transcript of HIV-1 is a full-length viral mRNA, which is translated into the gag (Pr55gag) and gag-pol (Pr160gag-pol) polyproteins (Fig 9). The processing of Pr55Gag and Pr160Gag-Pol into functional, mature subunits is a complex and essential step in HIV-1 maturation (Kohl *et al.*, 1988). Both precursors are translated from the same mRNA by a ribosomal frame-shifting mechanism that allows the Pr160gag-pol precursor to be synthesized as a carboxyl-terminal extension of Pr55gag (Jacks *et al.*, 1988). It has been suggested that the polyprotein precursors are cleaved by the viral protease encoded in Pr160gag-pol into their final products before the particles have budded from the cell surface (Kaplan *et al.*, 1994). Pr55gag is cleaved into p17 (MA), p24 (CA), p2, p9 (NC), p1, and p6. Cleavage of Pr160gag-pol yields MA, CA, and NC, as well as p6* (protease upstream region), p10 (PR), p66/51 (RT), and p32 (IN) proteins (Gelderblom, 1991; Wills & Craven, 1991). Each reaction occurs at a unique cleavage site that differs in amino acid composition (Billich *et al.*, 1988). Some cleavage sites contain phosphorylated Ser/Thr or Tyr residues that alter the sites' susceptibilities to cleavage (Tözsér *et al.*, 1999).

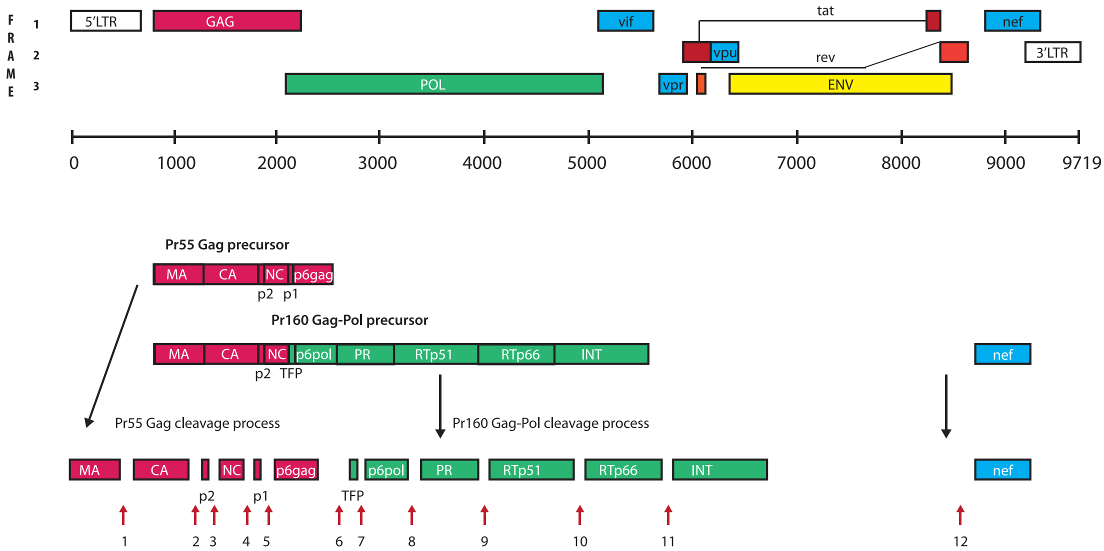


Figure 9: HIV-1 protease cleavage sites. The HIV-1 genome is shown above. The main structural proteins are formed by cleavage of the Pr55gag polyprotein into matrix (MA; p17), capsid (CA; p24), nucleocapsid (NC; p7), p6gag, and two spacer peptides, p2 and p1. The viral enzymes are formed by cleavage of Pr160gag-pol, a fusion protein derived by ribosomal frame shifting (13). Although Pr160gag-pol also contains p17, p24, and p2, its C-terminal cleavage products are NC, a transframe protein (TFP), p6pol, protease (PR), reverse transcriptase (RTp51), RNase H (RTp66), and integrase (IN). Red arrows indicate the 12 proteolytic reactions required to generate a mature infectious virion. Adapted from (de Oliveira *et al.*, 2003).

A detailed scheme of the HIV virion assembly, budding and maturation process orchestrated by the viral protease is shown in figure 10.

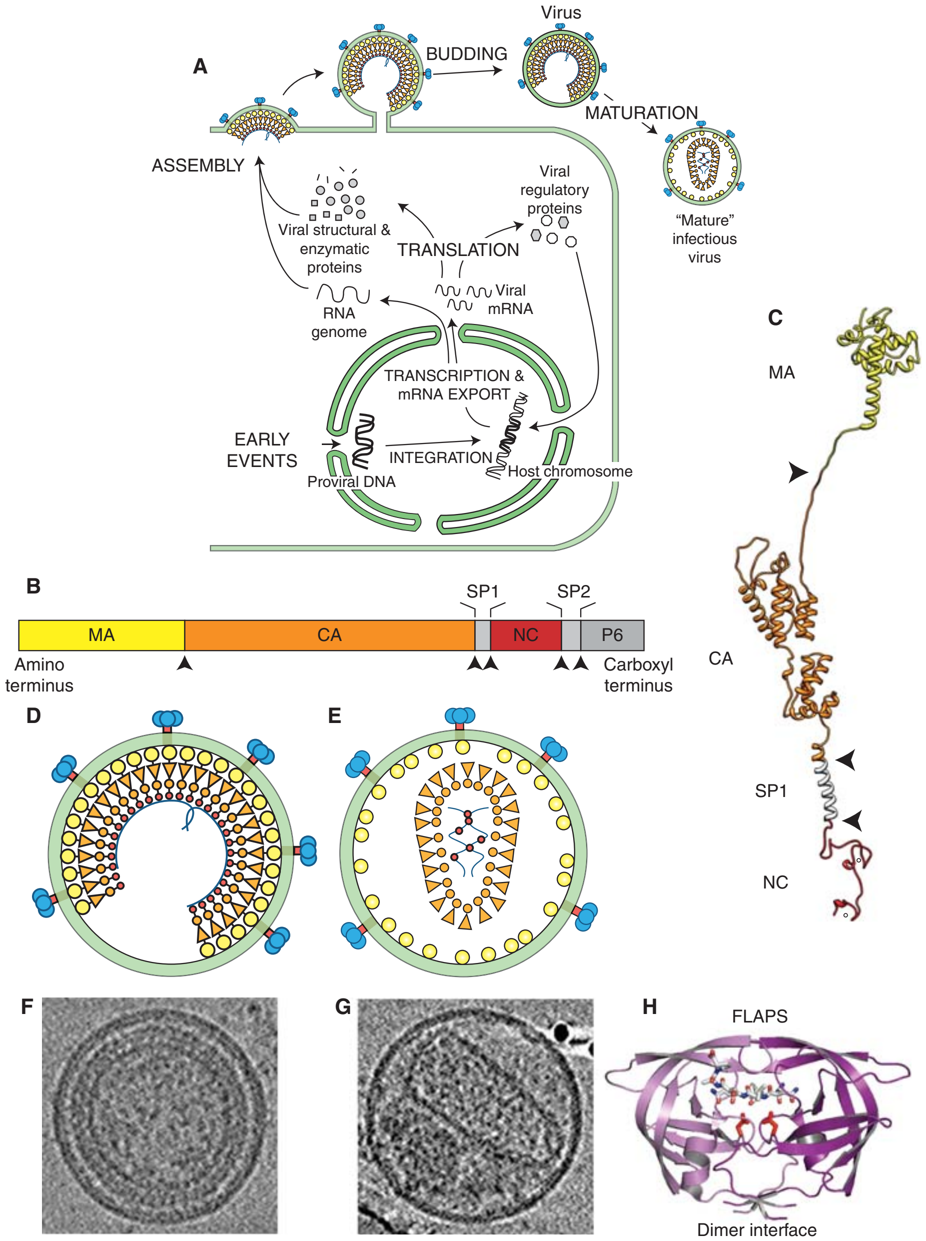


Figure 10: HIV-1 assembly, budding, and maturation. (A) Schematic illustration showing the different stages of HIV-1 assembly, budding, and maturation. (B) Domain structure of the HIV-1 Gag protein; arrows denote the five sites that are cleaved by the viral PR during maturation. (C) Structural model of the HIV-1 Gag protein, created by combining structures of the isolated MA-CANTD (2GOL), CACTD (1BAJ), and NC (1MFS) proteins, with a helical model for SP1. (D) Schematic model showing the organization of the immature HIV-1 virion. (E) Schematic model showing the organization of the mature HIV-1 virion. (F) Central section from a cryo-EM tomographic reconstruction of an immature HIV-1 virion. (G) Central section from a tomographic reconstruction of a mature HIV-1 virion. (H) Structure of HIV-1 protease (PR, 3D3T). The two subunits in the dimer are shown in different shades of purple, the "flap" and dimerization interfaces are labeled, positions of the active site Asp25 residues are shown in red, and a bound peptide corresponding to the SP2-p6 cleavage site is shown as a stick model, with oxygen atoms in red and nitrogen atoms in blue. Obtained from (Sundquist & Kräusslich, 2012).

HIV-1 protease variability

Evolutionary change in proteins is based on genetic variation introduced by mutations. Much of the research in this area has dealt with the history of individual genes, which can poorly reflect the history of the organism, particularly when rates of recombination are high. Besides, evolutionary rates are most easily studied in rapidly evolving organisms, either those with high mutation rates or with short generation times, both of which facilitate the rapid generation of genetic diversity (Duffy *et al.*, 2008). The protease gene has great plasticity, with polymorphisms observed in 49 of the 99 codons, and more than 20 substitutions known to be associated with resistance to PIs (Robert W Shafer, 2001). Numerous studies have described HIV-1 protease variability and polymorphisms found in naïve or PI-treated infected individuals (Ceccherini-Silberstein *et al.*, 2004) (Fig 11). Those reports have also shown that the HIV-1 protease is an enzyme that can accept a great number of amino acid changes without losing its enzyme activity (Loeb *et al.*, 1989).

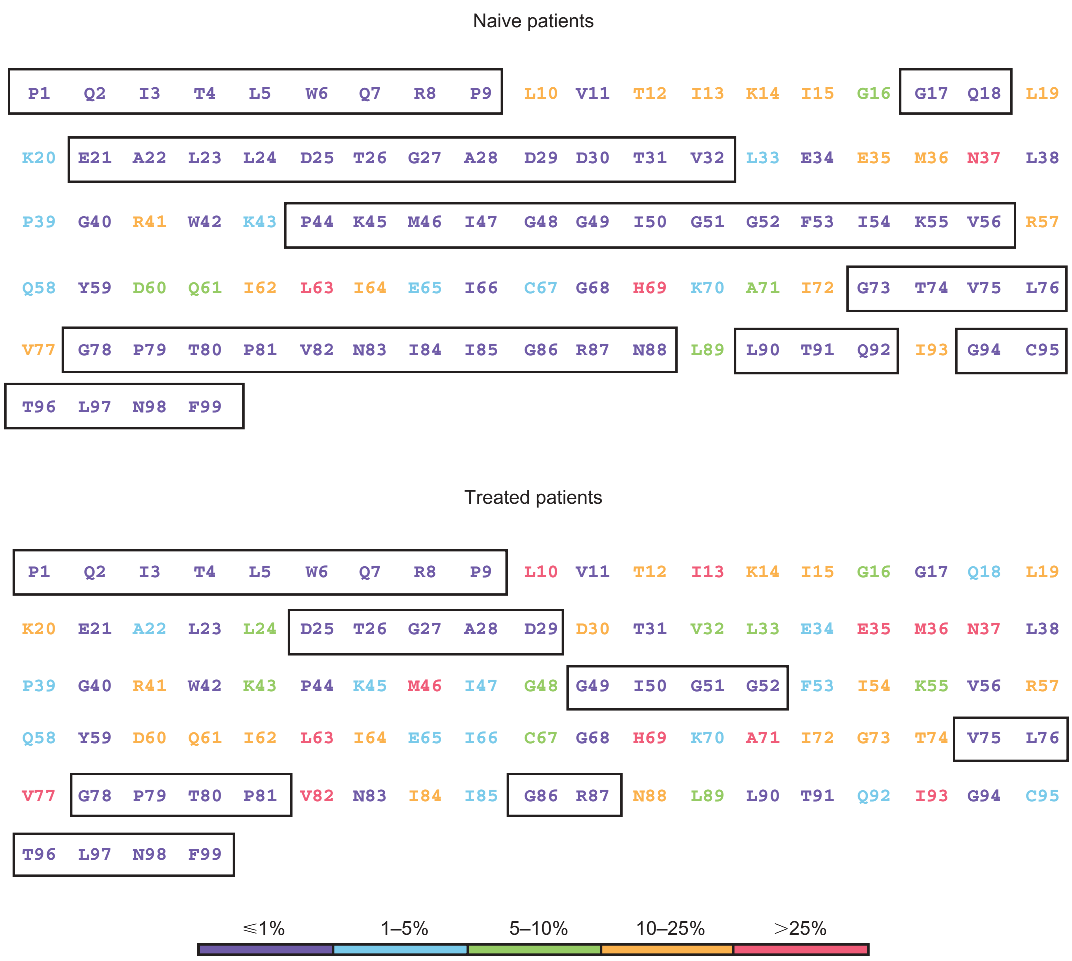


Figure 11: Conserved regions of HIV-1 protease in drug-naïve and drug-treated HIV-1 infected patients (Ceccherini-Silberstein *et al.*, 2004). The amino acid sequence of HIV-1 PR (99 amino acids) of clade B consensus (shown as a reference) is coloured according to the frequency rate of mutations observed in plasma samples from 457 drug-naïve and 639 PI-

treated patients. Conserved domains or stretches of amino acids are boxed. The bar indicates the frequency rate of mutations (%) relative to the colours used in the figure.

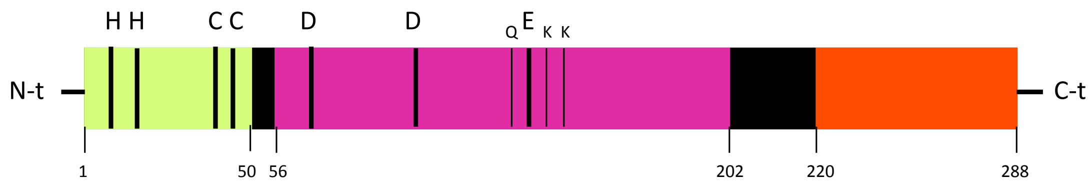
Robustness defined as a protein's tolerance to substitutions is fundamental to understand natural or artificial protein evolution. Recently, it has been shown that mutational robustness can enhance protein's capacity to evolve (Aharoni *et al.*, 2004; Bloom *et al.*, 2006). It has been claimed that selection might favor proteins that are robust to genetic change (Taverna & Goldstein, 2002), which may lead to an increase in selectively neutral variants. Proteins are to a certain extent tolerant to single mutations. Although former studies have provided important information about protein tolerance to amino acid changes, they may not reflect the protein tolerance to single substitutions because when several mutations are present simultaneously in the same genotype, they interact to determine the overall fitness, a phenomenon also called epistasis (Phillips *et al.*, 2000).

Recently, it has been shown the low tolerability to single random mutations for the wild-type HIV-1 protease (Parera *et al.*, 2006), in contrast with its in vivo ability to generate an adaptive variation. Furthermore, analysis of epistatic interactions among pairs of deleterious mutations in the viral protease showed high frequencies of lethality and negative epistasis, thus indicating that the HIV-1 protease is highly sensitive to the effects of deleterious mutations (Parera *et al.*, 2009). Therefore, proteins may not be as robust to mutational change as is usually expected. Moreover, rapid evolution of HIV-1 has contributed to the diversification of the viral population (Hemelaar *et al.*, 2011; Korber, 2000). As a consequence, site-specific sequence conservation values have decreased over time at the population level. This raises the question of to what extent is viral diversification influencing HIV-1 fitness and virulence over the course of the epidemic.

Two studies of this thesis analyse the HIV-1 protease evolution and the relationship with its ex vivo viral replication capacity and in vitro catalytic efficiency, respectively. One study investigates the natural evolution of the HIV-1 protease by deciphering the impact of HIV-1 protease diversification on ex vivo HIV-1 replication capacity in early and recent naïve samples isolated 15 years apart, as well as in samples from HIV-1 infected patients carrying PI resistance mutations. Whereas the last study explores its artificial evolution by comparing the capacity of an in vitro mutated HIV-1 protease and of the wild type protease to tolerate one single amino acid substitution. Moreover, the sequence diversification over time of different HIV-1 genes: gag, protease and integrase is analysed, using the sequences of the HIV-1 naïve patients infected 15 years apart mentioned above.

HIV-1 integrase

The HIV-1 integrase protein plays an essential role in the viral life cycle by catalyzing the chromosomal integration of the newly synthesized double-stranded viral DNA into the host genomic DNA (Coffin *et al.*, 1997; Rice *et al.*, 1996). Thus enabling HIV-1 to establish a permanent genetic reservoir that can both initiate new virus production and replicate through cellular mitosis. Therefore it has been a critical target for new antiviral therapies as explained before (Nguyen *et al.*, 2011; Powderly, 2010). HIV-1 integrase is a 32 kDa protein of 288 amino acids, comprising three functional domains: the amino (N)-terminal domain (amino acids 1-50), the catalytic core domain (amino acids 56-202), and the carboxyl (C)-terminal domain (amino acids 220-288) (Fig 12) (Engelman & Craigie, 1992; Jaskolski *et al.*, 2009). Interestingly, although the intact HIV-1 integrase appears to function as a tetramer (Jenkins *et al.*, 1996; Zheng *et al.*, 1996), each of the isolated domains forms stable dimers. Peptides that inhibit IN activity may function by binding to exposed hydrophobic residues via coiled-coil interactions and inhibiting higher-order oligomerization (Sourgen *et al.*, 1996). A model of the tetramer generated from the independently determined structures of the IN domains is shown in figure 13 (Dyda *et al.*, 1994).



N-terminal domain, HHCC region: Hys 12 and 16; Cys 40 and 43.
Catalytic domain: D,D,E motif (Asp 64 and 116, Glu 152)
C-terminal domain: nonspecific DNA binding domain (aa 220 to 270)

Figure 12: HIV-1 integrase conserved regions and residues. Conserved regions of the amino acid sequence of HIV-1 integrase (288 amino acids) of clade B consensus (shown as a reference) are boxed. Residues reported to be relevant for the Integrase catalytic activity are indicated (Q148, K156, K159). Abbreviation "aa" stands for amino acid.

The **N-terminal domain** is composed of four α helices that contain a highly conserved zinc-binding Hys12-Hys16-Cys40-Cys43 motif (Cai *et al.*, 1997; Dyda *et al.*, 1994; Polard & Chandler, 1995; Rice *et al.*, 1996) involved in the stabilization folding and proper oligomerization of the integrase subunits (Burke *et al.*, 1992; Lee *et al.*, 1997; Zheng *et al.*, 1996) (Fig 13, A).

The **catalytic core domain**, which plays a critical role in integrase enzymatic activity, contains the catalytic Asp64-Asp116-Glu152 motif that is conserved in all retroviral integrases, as well as in retro-transposons from plants, animals and fungi and in some bacterial transposases (Avidan & Hizi, 2008; Kulkosky *et al.*, 1995; Polard & Chandler, 1995; Rice *et al.*, 1996). The structure of the core domain crystallizes as a dimer with each monomer consisting of a central five-strand β-sheet and six α-helices (Dyda *et al.*, 1994) (Fig 13, A). It also contains other functional domains and residues such as the nuclear localization signal, a critical sequence mediating the nuclear import of the integrase in the context of the pre-integration complex (Bouyac-Bertoia *et al.*, 2001); the Lys186-Arg187-Lys188 multimerization motif at the dimer-dimer interface (Berthoux *et al.*, 2007; Wang, 2001); and several important residues involved in binding to the human lens epithelium-derived growth factor (LEDGF/p75), which is an essential cellular cofactor for HIV integration, linking the integrase to chromatin (Busschots *et al.*, 2007; Cherepanov *et al.*, 2005; Hombrouck *et al.*, 2007; Maertens, 2003; Rahman *et al.*, 2007).

The **C-terminal domain** has strong but nonspecific DNA-binding activity and is involved in the binding with viral and cellular DNA with the minimal nonspecific DNA binding region (MDBD 220-270 aa) (Engelman *et al.*, 1994; Lutzke *et al.*, 1994; Lutzke & Plasterk, 1998; Vink *et al.*, 1993). This domain, required for the integration reaction, is involved also in protein oligomerization and interactions with the reverse transcriptase (Lutzke & Plasterk, 1998). The structure of the C-terminal DNA-binding domain of IN was determined independently by two groups using NMR methods (Eijkelenboom *et al.*, 1995; Lodi *et al.*, 1995). The two structures appear to be essentially identical. Each of the monomeric subunits of the symmetrical dimer is composed of five antiparallel b-strands arranged in a b-barrel and folded in a way that is topologically similar to the Src-homology-3 (SH3) domain. An extensive hydrophobic dimer interface is formed by the “face-to-face” packing of three β-sheet strands from each monomer (Fig 13, A).

Integration is a complex process that that takes place in a stepwise manner: (1) assembly of the integrase enzyme at the end of the HIV long terminal repeat (LTR), forming the pre-integration complex; (2) cleavage of the terminal GT dinucleotide from the 3' end of each LTR (known as 3' processing); (3) translocation of the pre-integration complex to the nucleus through the nucleopore; (4) covalent linkage of the viral DNA into the cellular chromosome (known as strand transfer); and (5) gap repair by cellular enzymes (Brown, 1990). Although the integrase catalyses both 3' processing and strand transfer, current INSTIs only inhibit its strand transfer activity (Dyda *et al.*, 1994; Grant & Zolopa, 2008; Hazuda *et al.*, 2000).

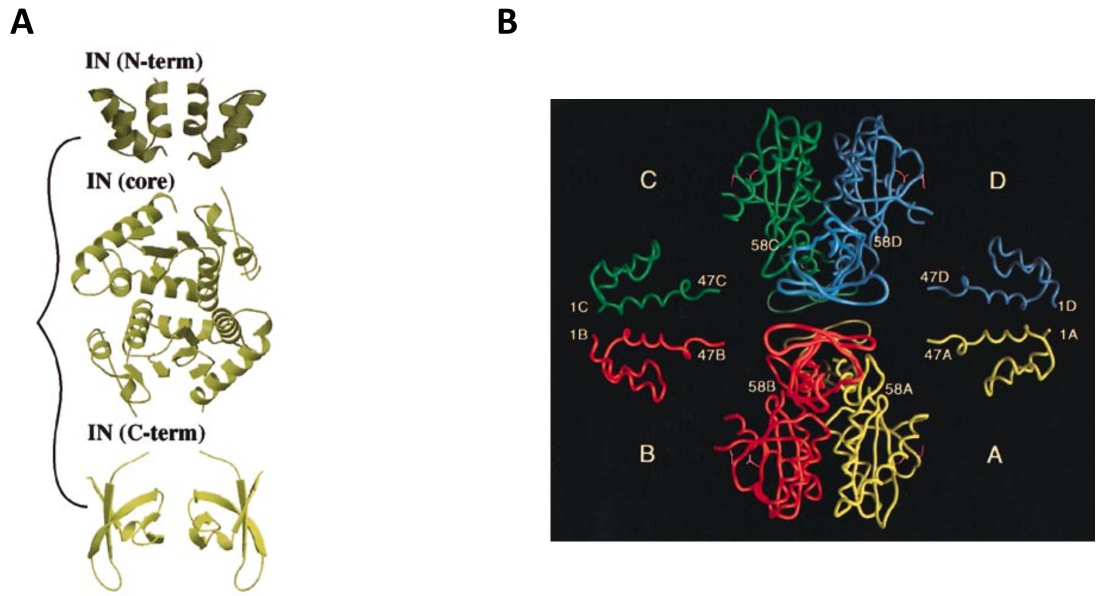


Figure 13: Structure of the HIV-1 integrase. (A) Ribbon drawing depicting the backbone crystal structure of the free HIV-1 integrase. (B) Model of the HIV-1 IN tetramer generated using independently solved structures of the N-terminal, core, and C-terminal domains. The four molecules of the tetramer (A, B, C and D) are displayed in different colors. The active site residues (Asp 64 and Asp 116) of each catalytic core are shown in magenta. (Note that the third catalytic residue, Glu 152, is not visible in the electron density map). Obtained from (Cai *et al.*, 1997; Turner & Summers, 1999).

HIV-1 integrase variability

As for the HIV-1 protease, various studies have described HIV-1 integrase variability and polymorphisms found in naïve or PI-treated infected individuals (Ceccherini-Silberstein *et al.*, 2009) that are shown in figure 14. Viral fitness defects resulting from mutations in the Pol gene are well documented following antiretroviral therapy (Quiñones-Mateu *et al.*, 2008), and Raltegravir resistance mutations N115H and

Q148H in integrase have been associated recently with reduced viral RC (Hu & Kuritzkes, 2010).

Mutations in the IN gene can alter the integration process and the resistance to INIs, but also affect the Gag–Pol precursor protein and alter assembly, maturation and other subsequent viral events (Bukovsky & Göttlinger, 1996; Engelman *et al.*, 1995; Mandal *et al.*, 2008; Quillent *et al.*, 1996; Wu *et al.*, 1999; Zhu *et al.*, 2004). For instance, it has been reported that mutations at conserved residues of the IN could cause a significant block of viral reverse transcription (Wu *et al.*, 1999; Zhu *et al.*, 2004). Therefore, IN mutations are pleiotropic and may alter virus replication through various mechanisms at different stages in the virus life cycle. At least in part, this probably explains the diverse phenotypes that have been reported for IN mutant viruses. These have included viruses with defects in assembly, virion morphology, reverse transcription, nuclear import, integration of the provirus and excessive RNA splicing phenotype, Gag protein processing and virus particle production (Bukovsky & Göttlinger, 1996; Engelman *et al.*, 1995; Mandal *et al.*, 2008; Quillent *et al.*, 1996; Wu *et al.*, 1999).

The third study presented in this thesis investigates the impact of the HIV-1 integrase diversification on the ex vivo HIV-1 replication capacity in early and late samples isolated 15 years apart. The selected HIV-1 infected patients are naïve for antiretroviral therapy with integrase inhibitors. Therefore, the HIV-1 natural evolution can be explored since no selective pressure such as antiretroviral treatment has been applied.

	F1	L2	D3	G4	I5	D6	K7	A8	Q9	E10	E11	H12	E13	K14	Y15
	H16	S17	N18	W19	R20	A21	M22	A23	S24	D25	F26	N27	L28	P29	P30
	V31	V32	A33	K34	E35	I36	V37	A38	S39	C40	D41	K42	C43	Q44	L45
	K46	G47	E48	A49	M50	H51	G52	Q53	V54	D55	C56	S57	P58	G59	I60
I	W61	Q62	L63	D64	C65	T66	H67	L68	E69	G70	K71	I72	I73	L74	V75
II	A76	V77	H78	V79	A80	S81	G82	Y83	I84	E85	A86	E87	V88	I89	P90
	A91	E92	T93	G94	Q95	E96	T97	A98	Y99	F100	L101	L102	K103	L104	A105
	G106	R107	W108	P109	V110	K111	T112	I113	H114	T115	D116	N117	G118	S119	N120
III	F121	T122	S123	T124	T125	V126	K127	A128	A129	C130	W131	W132	A133	G134	I135
IV	K136	Q137	E138	F139	G140	I141	P142	Y143	N144	P145	Q146	S147	Q148	G149	V150
	V151	E152	S153	M154	N155	K156	E157	L158	K159	K160	I161	I162	G163	Q164	V165
	R166	D167	Q168	A169	E170	H171	L172	K173	T174	A175	V176	Q177	M178	A179	V180
	F181	I182	H183	N184	F185	K186	R187	K188	G189	G190	I191	G192	G193	Y194	S195
	A196	G197	E198	R199	I200	V201	D202	I203	I204	A205	T206	D207	I208	Q209	T210
	K211	E212	L213	Q214	K215	Q216	I217	T218	K219	I220	Q221	N222	F223	R224	V225
	Y226	Y227	R228	D229	S230	R231	D232	P233	L234	W235	K236	G237	P238	A239	K240
V	L241	L242	W243	K244	G245	E246	G247	A248	V249	V250	I251	Q252	D253	N254	S255
VI	D256	I257	K258	V259	V260	P261	R262	R263	K264	A265	K266	I267	I268	R269	D270
	Y271	G272	K273	Q274	M275	A276	G277	D278	D279	C280	V281	A282	S283	R284	Q285
	D286	E287	D288												

≤ 1%   1.1-5%   5.1-10%   10.1-25%   > 25%

Figure 14: Conserved regions and residues of HIV-1 integrase in HIV-1-infected patients naïve to integrase inhibitors (Ceccherini-Silberstein *et al.*, 2009)**.** The amino acid sequence of HIV-1 integrase (288 amino acids) of clade B consensus (shown as a reference) is coloured according to the frequency rate of mutations observed in plasma samples from 448 integrase inhibitor-naïve patients. Residues associated with integrase inhibitor resistance are underlined. Conserved regions of amino acids are boxed. The bar indicates the frequency rate of mutations (%) relative to the colours used in the figure.