# Literature Review

## CHAPTER 1

### 1. 1 Overview of HIV/AIDS

Human Immunodeficiency Virus (HIV) is a human pathogenic virus that cause AIDS (Acquired Immunodeficiency Syndrome). HIV/AIDS has been global pandemic for over the last three decades and is depicted as the modern day plague [Quinn, 1996 #666]. The United Nations Acquired Immune Deficiency Syndrome (UNAIDS) global report 2012 estimates that by the end of 2011 approximately 34 million people were living with HIV by the end of 2011 (WHO factsheet Number 360 (<http://www.who.int/mediacentre/factsheets/fs360/en/>)) and that over 95% of them are living in low and middle in come countries [Esparza, 2000 #782]. There is a significant variation in HIV prevalence among the countries around the globe (Figure 1.1). The UNAIDS 2012 report shows that although the global trend of new HIV infections and HIV-related deaths per year is declining, the current number of HIV infections is the highest since 1990 (Figure 1.2). The sub-Saharan region of Africa is the region most aggravated by the virus with 23.5 million people living with HIV [UNAIDS, #703]. UNAIDS estimates that approximately 1 in every 20 adults is HIV infected in this region [UNAIDS, #703]. This is 25 or more times the HIV prevalence in any other region of the world. Countries in Sub-Saharan Africa also have varying HIV prevalence with South Africa at the top followed by Nigeria [Esparza, #782]. The next severely affected regions, besides African continent, are Asia (China, Thailand, Indonesia), Caribbean and Eastern Europe, North America, western and central Europe [Esparza, #782].

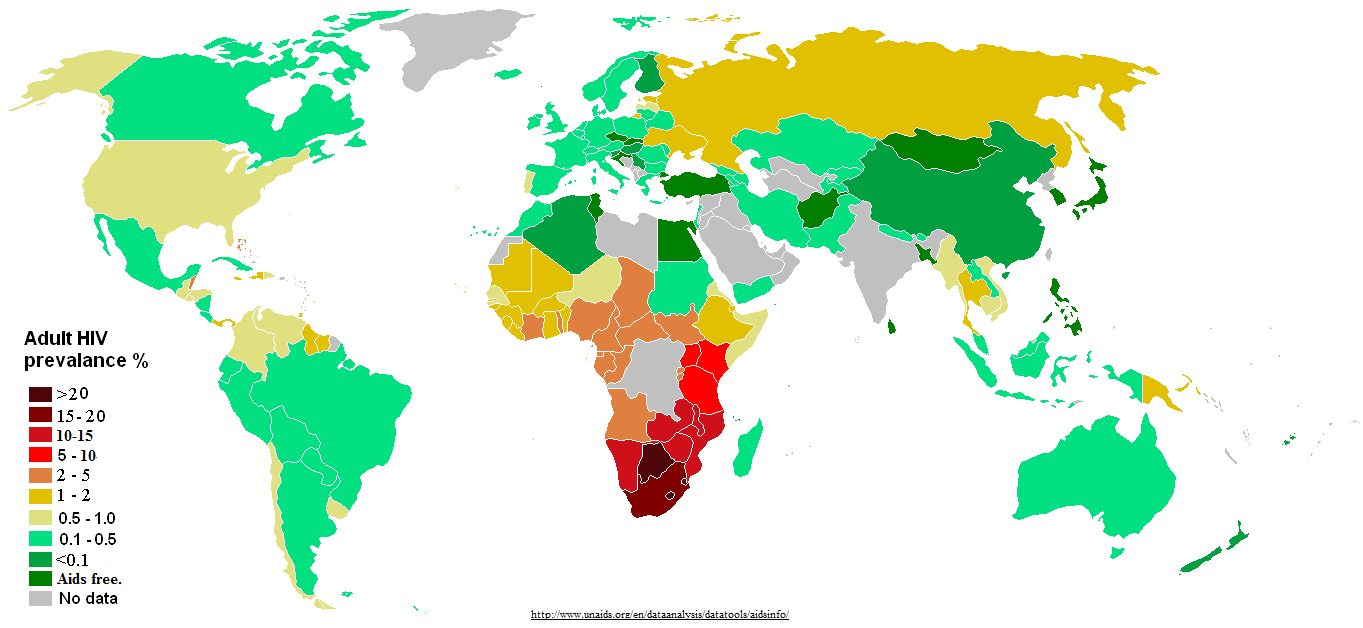


Figure 1.1: Distribution of HIV prevalence in adults around the world.

Source: http://www.unadis.org/en/dataanalysis/datatools/aidsinfo/

hiv global trend.pdf

Figure 1.2: Global trend of new HIV infections from 1990 to 2011. The number of people living with HIV globally is increasing (A) while the number of people newly infected with HIV (B) and the number of Adults and child deaths due to HIV are decreasing (C) globally in the time period. This scenario can be attributed to global scale up of drugs while infected people continue transmitting the virus to uninfected people [Zaidi, #709]. (Source: modified from UNAIDS 2012)

### 1.2 Discovery and characterization of HIV

As early as 1959, HIV infection cases had been documented but were unreported [Nahmias, 1986 #734;Corbitt, 1990 #736]. Curious doctors at that time collected patient blood samples and kept frozen, which were later shown to have HIV antibodies [Zhu, 1998 #735]. In June 1981, a case of acute immune depletion associated secondary infection was reported in some homosexuals in the USA [Friedman-Kien, 1981 #249;Friedman-Kien, 1981 #250]. Their infection was coupled with no lymphocyte proliferation [Gottlieb, 1981 #248]. Until 1983, the causative agent responsible for the severe immune depletion, named AIDS (Acquired Immunodeficiency Syndrome), was unknown [Francis, 1983 #95;Gallo, 1983 #93], when Luc Montagnier’s group at “Institut Pasteur” in Paris isolated the virus, which was initially named Human T-cell Leucamia Virus (HTLC) and later named as Human Immunodeficiency Virus (HIV) [Nahmias, 1986 #734]. Jay Levy’s group in San Francisco, USA also subsequently found the virus confirming the discovery in Paris [Levy, 1984 #92]. They found HIV was a lentivirus from Group VI retrovirus with two single strand RNA molecules [Baltimore, 1971 #204]; unique to any other previously isolated viruses and the virus can transmit from infected to healthy people [Wofsy, 1986 #90;Rogers, 1987 #262], mother to child through umbilical cord [Gallo, 1983 #93;Ziegler, 1985 #91].

Very soon, scientists around the world focused on this transmissible retrovirus. Complete sequencing of HIV genome in 1985 [Ratner, 1985 #89] led scientists to know more insights of HIV including its origin, genes/proteins and life cycle [Wain-Hobson, 1985 #88].

## 1.3 Origin and evolution of HIV

Exploration of the retrovirus led researchers to identify similarities between HIV and a retrovirus in African non-human primates that were then known as Simian Immunodeficiency Virus (SIV) [Gao, 1994 #656]. About 40 different primates, in Africa, are infected with SIV with some harboring multiple strain of SIVs [Apetrei, 2004 #737]. Phylogenetic analysis of SIV from African non-human primates and HIV in human provided remarkable understanding of viral transmission as zoonotic [Bailes, 2002 #659] and evolution of the virus in human after transmission [Gao, 1999 #738].

HIV is divided into two groups – HIV-1 and HIV-2 [Gao, 1999 #738]. Each group resulted from an independent cross species transmission from different African non-human primates to human [Sharp, 2010 #646]. HIV-2 is discovered in 1986. This group is transmitted from sooty mangabey monkeys (Cercocebus atys) [Hirsch, 1989 #657] and its prevalence is also high in the geographical location of these monkeys in West Africa [Santiago, 2005 #658]. Sooty mangabey monkeys are naturally infected by a strain of SIV [Hirsch, 1989 #657]. The phylogenetic analysis of HIV-2 strains shows that they closely group with the SIVsmm strain [Hirsch, 1989 #657] that are non-pathogentic to its host monkeys [Gao, 1992 #542]. SIVsmm evolved in its host to produce multiple strains and subsequent multiple zoonotic transmissions from sooty mangabey monkeys to human [Hahn, 2000 #655] gave rise to different subtypes of HIV-2. Although HIV-2 subtypes A to G are identified in human, it is assumed that more subtypes were introduced into human [Gürtler, 2004 #654] but are lost for low adaptation fitness [Damond, 2004 #653].

HIV-1 is the result of at least three cross species transmission events from chimpanzees (Pan troglodytes troglodytes (Ptt)) to human [Peeters, 1989 #652;Huet, 1990 #651]. Phylogenetic analysis of HIV-1 sequences has shown that three independent cross transmission of the virus in to the human population, each giving rises to three sub groups: group M (Major), group O (Outlier) and group N (Non M or Non O) [Keele, 2006 #497;Hahn, 2000 #655]. Recently a new HIV-1 strain, classified as group P, distinct from the previous three groups, has been discovered in a patient in Cameroon [Plantier, 2009 #762].

Group M is the most prevalent and accounts for 98% of all infections (reviewed in [Sharp, 2010 #646]). Its epicenter is thought to be Kinshasha of present day Democratic Republic of Congo [Sharp, #764;Sharp, 2010 #646]. Site stripping for clock detection method shows that group M and its closest simian relative branched out from their common ancestor in 17th century [Salemi, 2001 #774] whereas molecular clock analysis of group M shows that the origin of it’s most recent common ancestor dates back to late 1920s [Korber, 2000 #767]. By 1960, long before human discovered its presence, HIV-1 group M had already diversified substantially [Worobey, 2008 #770].

Group O and group N are rare and geographically confined to West African regions such as Cameroon and neighboring countries [Gao, 1999 #738]. It is still not understood about the non-pandemic characteristics of group O and N HIV-1 virus after the first zoonotic transmission [Ariën, 2005 #84]. It has been suggested that reduced replication capacity and transmission fitness are keys to their low prevalence[Ariën, 2005 #84]. Group O strain has at least 50% genetic identity with group M [Gürtler, 2004 #654;VANDEN HAESEVELDE, 1996 #661] and the molecular clock model of this group also showed that its origin dates back to 1920s [Lemey, 2004 #81]. The date of origin of the introduction of group N into human population has been estimated to be in 1960s [Simon, 1998 #487]. Phylogenetic analysis using genetic sequence under evolutionary pressure shows its close grouping with SIV from Chimpanzee [Gao, 1999 #738;Corbet, 2000 #645]. This indicates that group N might be a recombinant strain of SIV and HIV-1 group [Simon, 1998 #487].

Group P is transmitted from gorilla as it is closely related to its SIV [Plantier, 2009 #762]. A study of HIV infected people in Cameroon shows its low prevalence of 0.06% [Vallari, 2011 #80]. Although HIV group P is discovery only in Cameroon and confined there, it can still be pandemic as it can adapt in human [Vallari, 2011 #80].

### 1.4 HIV-1 Diversity

**1.4.1 HIV-1 subtypes**

HIV-1 group M is highly diversified and it is classified into nine subtypes: A, B, C, D, F, G, H, J and K (Figure 1.3) [Robertson, 2000 #768]. The subtype classification is based on the phylogenetic and sequence distance analyses using gene sequence data forming major clades [Robertson, 2000 #768]. “At least three epidemiologically unlinked sequences are required for defining a subtype” [Robertson, 2000 #505].



Figure 1.3: Phylogenetic tree showing HIV-1 group M diversification to subtypes A-D, F-H, J and K, inferred from nucleotide sequence alignments of gag, pol and env genes. Source: Robertson et al 2000 [Robertson, 2000 #505]

The classification of new subtype should also follow the same rule as “roughly equidistant from all previously characterized subtypes in all regions of the genome with a distinct pre-subtype branch similar to those of other subtypes” [Robertson, 2000 #768]. The predominating group M subtypes are A, B, C and D (reviewed in [McCutchan, 2006 #493]). The range of amino acid variation at gene level within a subtype and between subtypes differs from 15%– 20% and 25% - 35% respectively [Korber, 2001 #208]. Geographical locations of group M subtypes epidemic are show in Figure 1.4. The analysis from HIV samples collected from 70 countries in 2004 shows that “subtype C accounts for 50% of all infections worldwide” while subtypes A, B, G and D are found in decreasing order 12%, 10%, 6% and 3% respectively [Hemelaar, 2006 #213]. Subtypes F, H, J and K infections are rare and collectively account for only 0.94% infections [Hemelaar, 2006 #213].

Subtypes can be further classified to sub-subtype based on a distinct sister clade formation [Gao, 2001 #513] within a clade with the same rule of phylogenetic and distant analyses but not justifiable to call a subtype due to low genetic distance [Robertson, 2000 #768]. Only subtypes A and F exhibit distinct sister clades [Gao, 2001 #513]. Subtype A has sub-subtypes A1 and A2 (A3 and A4 are mentioned by Taylor et al 2008[Taylor, 2008 #214]); Subtype F has sub-subtypes F1 and F2 (Figure 1.3); sub-subtype F3 mentioned by Taylor et al 2008 [Taylor, 2008 #214].

The lower diversity observed in Group N [Ayouba, 2000 #825], O [Lemey, 2004 #826] and P [Vallari, 2011 #80] correlate to low prevalence and geographical confinement to Western African countries such as Cameroon. It is still unclear if group O can be sub divided into subtypes [Roques, 2002 #15]. Group N as well does not show distinct sub clade within itself (Figure 1.5).

**1.4.2 HIV-1 recombination**

Initially, HIV-1 group M subtypes E [Artenstein, 1995 #931;Tovanabutra, 2002 #933;Wasi, 1995 #932] and I (Figure 1.5) were also classified (reviewed in [McCutchan, 2006 #493]). With availability of complete HIV genome sequence and phylogenetic analysis from it, the subtypes E and I were reclassified as circular

### 

Figure 1.4: HIV diversity around the globe, its level of prevalence in the area and number of genome sequenced. Source: McCutchan 2006

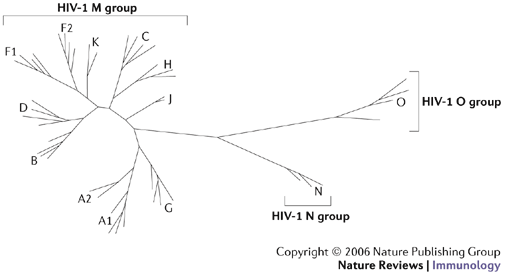


Figure 1.5: Neighboring joining phylogenetic tree showing HIV-1 group M, N and O. Group M shows distinct nine subtypes A-D, F-H, J, K while no specific subtype is observed in group N and O. Source: Letvin 2006 [Letvin, 2006 #12]

recombinant forms CRF01\_AE (recombinant form of subtype A and E) and CRF04\_cpx (recombinant form of more than two subtypes, designated by “cpx”) respectively (reviewed in [McCutchan, 2006 #493]). The same criterion of epidemiological unlinked isolates from three or more people applies for classification as a circular recombinant form (CRF) [Robertson, 2000 #505]. A recombinant isolate discovered in single patient is termed as Unique Recombinant Form (URF) (reviewed in [McCutchan, 2006 #493]). There are 55 CRFs listed in Los Alamos National Laboratory database for HIV sequences (<http://www.hiv.lanl.gov/content/sequence/HIV/CRFs/CRFs.html>) as on July 22, 2013. The recombination breakpoints to shuffle HIV genome [Archer, 2008 #11;McCutchan, 2002 #10;Zhang, 2010 #891;Salminen, 1995 #971] from different strains of the virus are listed in <http://www.hiv.lanl.gov/content/sequence/HIV/CRFs/breakpoints.html>, each appeared in a publication. CRFs account for at least 20% of all the HIV infections [Robertson, 2000 #505;Robertson, 1995 #7;Sharp, 1995 #518]. CRF02\_AG is the most prevalent circulating recombinant form infecting over 9 million people on the whole [McCutchan, 2000 #939] and geographically epidemic in the West African region (Figure 1.5). CRF01\_AE is the dominant circulating recombinant form in South-East Asia (Figure 1.5) [Taylor, 2008 #214].

**1.4.3 Intra-patient HIV diversity**

HIV infection initiates mostly with a single virion [Keele, 2008 #900;Fischer, #973]. Evidences of multiple HIV variants transmission are also recorded [Long, 2000 #910;Ping, 2000 #975]. Generally, HIV is genetically homogenous for a short post infection time [Delwart, 2002 #905;Haase, #904]. In the long-term post infection period, virus replicates rapidly to produce genetically heterogeneous population [Long, 2000 #910]. This heterogeneous viral population consisting of a swarm of highly similar but genetically non-identical HIV viruses is called the HIV quasispecies (reviewed in [McCutchan, 2006 #493]). It is observed that the diversity at a gene (*env*) in viral quasispecies can be approximately 30% [Fouchier, 1992 #978]. Factors that contribute to high genetic heterogeneity in viral quasispecies are high replication rate and turnover [Ho, 1995 #914], viral genome recombination [Gu, 1995 #980;Lole, 1999 #981;Fang, 2004 #898], higher mutation rate by erroneous reverse transcriptase [Roberts, 1988 #306;Bebenek, 1989 #982], and host immune selection [Price, 1997 #984;Borrow, 1997 #985]. On the whole, HIV replication (discussed in section 1.6) is the overall source of genetic heterogeneity in the viral population (reviewed in [Smyth, #919]).

Intra patientHIV genome recombination is a common event [Neher, #986;Fang, 2004 #898]. Two genomes from different viral strains from same subtype or different subtypes can be co-packed into single virion during replication [Stuhlmann, 1992 #922]. In the subsequent HIV replication, the ability of reverse transcriptase to switch between the two template genomes produces an intra subtype or inter-subtype recombined viruses at the end of the replication cycle [Ben-Artzi, 1996 #987;Kostrikis, 2002 #988]. Genetic recombination allows rapid and efficient shuffling of advantageous genes and removing deleterious mutations, thus, increasing the viral fitness in the host (reviewed in [Smyth, #919]). Successful transmission of the recombinant forms with high viral fitness to three or more people and circulates in human population establishes Circulating Recombinant Forms (CRFs) (reviewed in [Perrin, 2003 #930]), as discussed in section 1.4.2.

### 1.5 HIV genome and proteins – structures and functions

HIV has nine genes and produces 15 proteins [Frankel, 1998 #537]. The genes are broadly grouped as accessory (vif, vpr, vpu and nef), structural (*pol*, *gag* and *env*) and regulatory (*tat* and rev). The higher number of proteins than genes is a result of post - transcriptional proteolysis of the products of structural genes [Frankel, 1998 #537].

**1.5.1 Accessory genes:**

Vif promotes the viral infectivity to the host, but has no role in viral production [Jager, #989]. Vif is produced in the late stage of viral production [von Schwedler, 1993 #600;Sheehy, 2002 #603] to suppress the innate antiviral immunity of host [Simon, 1998 #602;Madani, 1998 #601]. It is observed that vif is expressed only when the virus infects immune cells that express cytidine deaminase APOBEC3G [Navarro, 2004 #228]. The reason is that Vif protein prevents APOBEC3 proteins from hyper mutating HIV reverse transcripts as a mechanism of defense [Conticello, 2003 #113;Mangeat, 2003 #597;Mariani, 2003 #591;Marin, 2003 #595;Mehle, 2004 #592;Schafer, 2004 #809;Sheehy, 2003 #593;Simon, 2005 #605;Stopak, 2003 #594;Wiegand, 2004 #498;Zhang, 2003 #604].

Vpr protein is packed in to nascent virions during budding out [Connor, 1995 #585]. Vpr is essential for viral core to enter and localize in the host cell nucleus after infecting the cell [Cohen, 1996 #797]. Vpr arrests the cell cycle during the transfer from G2 to M phase [Jowett, 1995 #586;Rogel, 1995 #587] by preventing the activation of the human p34cdc2/cyclin B complex [He, 1995 #588]. Vpr is also important for efficient viral replication in monocyte or macrophage cells, but T-cells [Connor, 1995 #585].

Vpu is a protein unique to HIV-1 [Cohen, 1988 #808] and is 16 kilo Dalton, 81 amino acids long [Strebel, 1988 #807]. Biological functions of vpu protein include degradation of CD4 in endoplasmic reticulum [Willey, 1992 #590], interference of host immune cell MHC class II antigen presentation on the cell surface allowing the virus for host immune escape [Hussain, 2008 #79;Nomaguchi, 2008 #78] and viral maturation and release from host cell membrane [Klimkait, 1990 #83].

Nef has no role in viral infectivity but plays a role during the biogenesis of viral particles [Laguette, 2009 #583] and virulence [Simmons, 2001 #990;Lenassi, #991]. Nef down regulates the production of major histocompatibility complex type I (MHC type I) in the host cell [Lewis, 2012 #486;Blagoveshchenskaya, 2002 #992;Greenberg, 1998 #993]. This impairs the function of cytotoxic T lymphocyte cells to clear the infected cells [Adnan, 2006 #576;Baur, 1994 #568;Collins, 1998 #578;Couillin, 1994 #573;Sawai, 1994 #571]. Nef also down regulates CD4 on host cell surface [Garcia, 1991 #565;Lama, 1999 #572] and modulates cellular activation to evade host immune system [Baur, 1994 #568;Sawai, 1994 #571].

**1.5.2 Structural genes and proteins**

The Gag gene produces a precursor polyprotein (pr55gag) of ~ 500 amino acids long and weighs 55 kilodalton [Briggs, 2004 #814]. The Gag precursor has all the building blocks to form a fully infectious virion, even in the absence of other viral products [Wang, 1993 #621]. A proteolytic cleavage of gag precursor yields the structural proteins – matrix, capsid, nucleocapsid and p6 (Figure 1.6) [Wiegers, 1998 #994]. The cleavage takes place in the nascent virus after budding out from host cell [Göttlinger, 1989 #77]. The matrix protein is at the N-terminal and p6 at the C-terminal of gag precursor [Borsetti, 1998 #995;Wiegers, 1998 #994].

All gag proteins play role at “post assembly and post processing stages in viral infectivity” [Wang, 1993 #621]. In the HIV replication cycle, matrix domain of gag plays role in targeting gag precursor to the plasma membrane of the host cell and the viral assembly at the site [Wang, 1993 #621;Gheysen, 1989 #810;Zhang, 1998 #70;Dawson, 1998 #69]. The highly basic region in matrix mediates electrostatic association with phospholipids in plasma membrane during assembly [Chukkapalli, #998;Zhou, 1996 #999]. The capsid is a curved and closed shell consisting 250 hexamers and 12 pentamers (reviewed in [Briggs, #1000]). Capsid packs viral proteins, nucleocapsid and viral genome during assembly to pass on to new HIV particles [Ganser-Pornillos, 2004 #820]. Nucleocapsid plays role in efficient viral assembly by making pr55–pr55 inter-protein contacts [Dawson, 1998 #69;Zhang, 1998 #70] and localization of viral proteins [Larsen, 2008 #817]. P6 protein plays role in detaching and releasing the newly formed HIV particles [Demirov, 2002 #811].

The *pol* polyprotein is produced by translational frame shift [Karacostas, 1993 #25] (Figure 1.5), such that, as much as 241 nucleotides at 5’ region of the gene overlaps with 3’ region of *gag* gene [Jacks, 1988 #231;Ratner, 1985 #89]. The proteolytic cleavage of *pol* precursor produces essential viral replication enzymes – protease (PR), reverse transcriptase (RT) and integrase (IN). The protease enzyme cleaves the *gag* and *pol* polyprotein to form the viral structure proteins and functional enzymes respectively [Darke, 1988 #1001;Nutt, 1988 #1002;ERICKSON-VIITANEN, 1989 #24]. The reverse transcriptase enzyme reverse transcribes the viral RNA to produce a cDNA molecule after infecting host cell [Jacobo-Molina, 1991 #63;Sarafianos, 2009 #426]. The RNase H domain in RT degrades the viral RNA molecule following cDNA production [Davies, 1991 #62]. The integrase enzyme removes two bases from 3’ DNA molecule and functions strand transfer during the process of integrating the proviral DNA into the host genome [Pruss, 1994 #60].

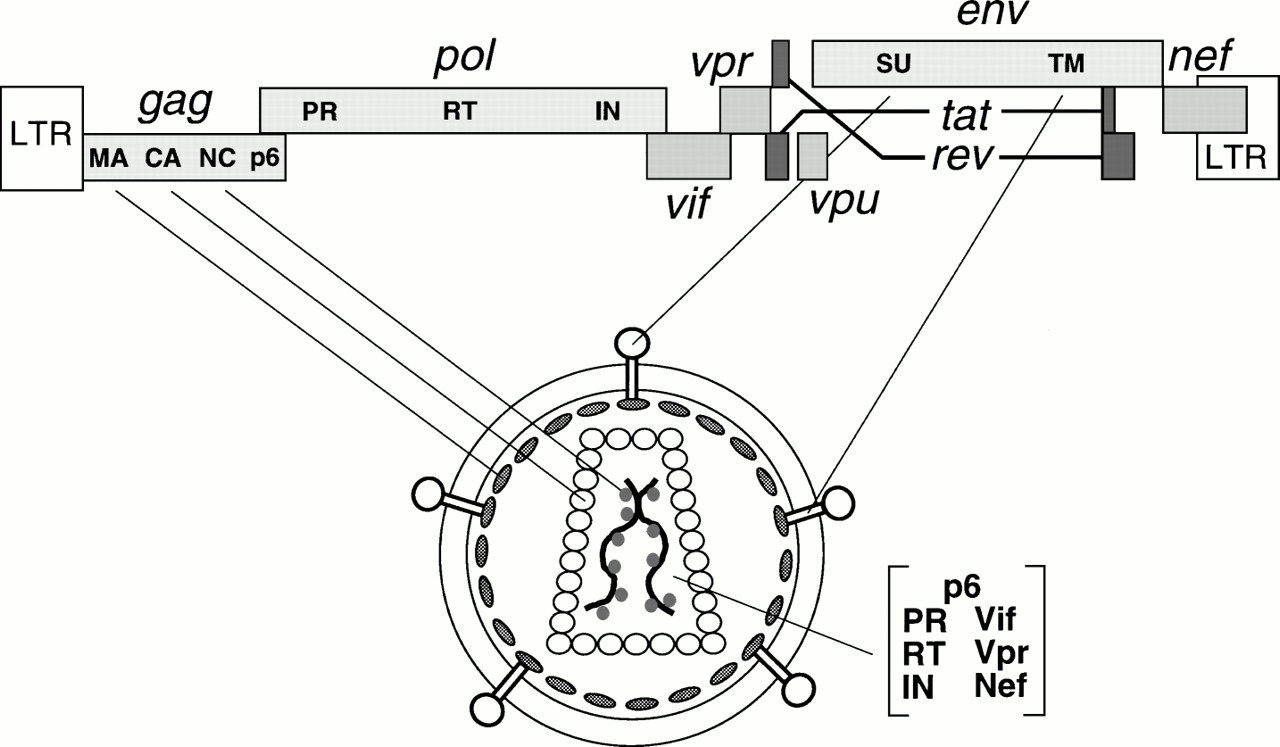


Figure 1.6: HIV genes and proteins positions in the viral genome and their viral parts. Source: Frankel and Young 1998 [Frankel, 1998 #537]

The *env* gene produces a precursor glycopolyprotein (gp160) that is processed at post-translational by human convertase enzymes - PC1 and furin to produce glycoprotein 120 (gp120, HIV-1 SU) and glycoprotein 41 (gp41, HIV-1 TM) [Decroly, 1994 #59]. Gp120 is a non-covalent complex of external protein and gp41 is a trans-membrane protein; both play vital role for initial steps in viral infection [Chan, 1997 #58]. Three gp120 molecules bound with three gp41 molecules to form envelop spikes [Pancera, #1003]. They are organized to form trimeric complexes on the surface of HIV and mediate HIV entry into the host cell [Liu, 2008 #56]. The exposed external complex gp120 binds to the CD4 receptor on the host immune cell [Rizzuto, 1998 #822]. This triggers a conformational shift of trimeric complex that enables a conserved gp120 region binding to a chemokine receptor, either CCR5 or CXCR4, to facilitate fusion of the viral and host membranes [Huang, 2007 #246;Rizzuto, 1998 #822;Wu, 1996 #823;Wu, 1997 #346]. The gp120-CD4 complex also triggers conformational change in gp41 trans-membrane protein from native non-fusion state to fusion state [Chan, 1997 #58;Kliger, 1997 #1004]. Gp41 plays role in the viral fusion and release of viral contents in to the host cell [Furuta, 1998 #46;Melikyan, 2008 #598],. The gp41 consists of heptad repeats - HR1 and HR2 that play role in fusion process [Tan, 1997 #53;Furuta, 1998 #46]. HR1 is a bundle of three helical motifs and HR2 is trimeric coiled coil structure [Dwyer, 2003 #828]. During fusion process, HR2 makes numerous contacts with HR1 to form stable six helical bundles [Melikyan, 2000 #48].

**1.5.3 Regulator Genes:**

*Tat* is a trans-activating factor localized in the nucleus for HIV gene expression [Rosen, 1990 #845;Roy, 1990 #1005]. The HIV proviral genome integrated in to the host genome is regulated by cellular as well as the viral transcription regulatory factors [Gaynor, 1992 #1006;Cullen, 1991 #1007]. Tat is the primary transcriptional regulatory factor [Marcello, 2001 #838]. An example of Tat action is the control of RNA polymerase II elongation during transcription, which otherwise disengages from the template DNA strand, terminating the transcription prematurely [Bourgeois, 2002 #1025;He, #1027;Chou, #1029]. Sodroski *et al*. (1985) first explained the function of Tat [Sodroski, 1985 #836].

Rev is a 19 kilo Dalton phosphoprotein [Malim, 1989 #843] trans-activating factor for HIV gene expression [Rosen, 1990 #845]. Like Tat, it is also mainly localized in the nucleus of host cell [Rosen, 1990 #844], but cycles rapidly between the nucleus and cytoplasm as it promotes nuclear export of the transcriptional products [Fischer, 1994 #848;Fischer, 1995 #846;Malim, 1989 #32;Henderson, 1997 #847]. Rev binds at the Rev Responsive Element (RRE), which is an RNA element encoded within the *env* region of the virus [Daly, 1989 #849;Malim, 1991 #1008].

### 1.6 HIV replication

There are 11 major events in HIV’s replication cycle (Figure 1.7). The initial step of viral entry in to a host immune cell includes HIV gp120 molecule binding to CD+ receptor followed by binding to a co receptor on the surface of the host cell and fusion of the viral and host cell membranes (see section 1.5.2 for more detail).

Following the fusion, the viral core enters in to the cytoplasm of host cell. The reverse transcriptase enzyme reverse transcribes the RNA molecule to cDNA (Figure 1.7 step 2) in the intact capsid of the viral core [McDonald, 2002 #555]. The reverse transcriptase enzyme is not perfect at copying mRNA molecule to cDNA and has no capability for error correction [Roberts, 1988 #306;Preston, 1988 #1150;Bebenek, 1989 #982;Bebenek, 1993 #1152]. The rate of errors generated is by reverse transcriptase is in the order of 10-5 per base per replication cycle [Mansky, 1995 #18]. This is a crucial step as it contributes to generation of variations in the viral quasispecies (reviewed in [Goodenow, 1989 #1155;Nowak, 1990 #1156]).

Subsequently the capsid is dissembled, termed as uncoating [McDonald, 2002 #555;Shah, 2013 #852], releasing the ribonucleoprotein complex in to the cytosol [Dismuke, 2006 #857]. The process can take an hour or less since time of post infection [Hulme, 2011 #853]. The capsid and nucleocapsid proteins dissociate from cDNA but the reverse transcription complex remains intact along with viral matrix, integrase, vpr and human protein high mobility group I (HMG I (Y)) forming pre-integration complex (PIC) [Bukrinsky, 1993 #858;Farnet, 1991 #860;Miller, 1997 #42]. The PIC protects cDNA from endonuclease degradation [Miller, 1997 #42]. In an ATP dependent process [Bukrinsky, 1992 #40], PIC is transported on host microtubules towards the nuclear membrane [McDonald, 2002 #555]. Integrase assists in nuclear import in association with nuclear import machinery like importin [Fassati, 2003 #870] and transportin-SR2 [Christ, 2008 #868]. It is now established that central polypurine tract-central termination sequence (cPPT-CTS) plays role in kinetics of nuclear import [Riviere, #862].

Post-nuclear entry, integrase processes the viral DNA for integration [LaFemina, 1992 #37] (Figure 1.7 step 5) into host genome as explained in section 1.5.2. Host transcriptional co-factor LEDGF/p75 and HIV integrase interact to tether to the host chromosome during the integration process [Emiliani, 2005 #34].

The integrated HIV provirus hijacks the host cell transcriptional machinery for viral genes to transcribe (Figure 1.7 step 6) [Davey, 2011 #33]. HIV protein Tat promotes the transcription of the viral DNA [Ott, 2011 #544;Razooky, 2011 #545]. The viral transcriptome encodes structural proteins, accessory proteins and viral enzymes necessary for a complete functional HIV (reviewed in [Karn, #871]).

The viral Rev protein facilitates exporting (Figure 1.7 step 7) of the unprocessed viral transcriptome to cytoplasm for translation [Malim, 1989 #32]. HIV has no translation system of its own; the host translational machinery is exploited for translation (Figure 1.7 step 8) of the viral transcriptome to its proteome [Cherry, 2005 #1031;Thompson, 2000 #1032].

Upon translation of all viral proteins, viral *gag* initiates virion assembly (Figure 1.7 step 9) at the cell membrane [Nermut, 1998 #1014;Saad, 2006 #942;Dong, 2005 #1016]. The complete assembled virion particles bud out and are released (Figure 1.7 step 10) from the plasma membrane by the host ESCRT machinery involving Tsg101 and ALIX regulatory proteins [Saksena, 2007 #877;Garrus, 2001 #875;Fujii, 2007 #30]. The maturation (Figure 1.7 step 11) of the nascent HIV virions begins concomitantly with budding out [Schubert, 2000 #879;Klimkait, 1990 #83]. *Gag* and *Pol* polyproteins are proteolytically cleaved by protease enzyme in the maturation step [Pettit, 2005 #1017;Darke, 1988 #65].

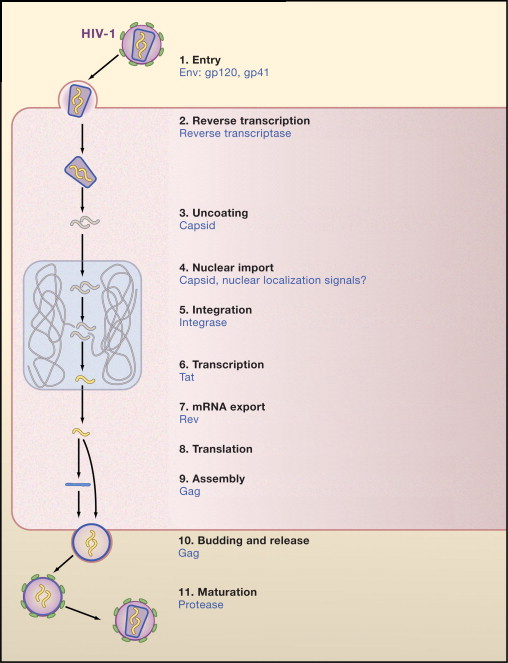


Figure 1.7: The HIV replication cycle showing major stages. Viral proteins that play role in each event are colored blue. (Source: modified from Ho and Bieniasz 2008) [Ho, 2008 #872]

Each HIV replication cycle releases new infectious virions in the order of 109 per day [Ho, 1995 #20]. The number of new infecting HIV determines the replication rate of the virus [Tersmette, 1989 #19]. A long post infection period shows higher turnover rate associated with CD4+ cell depletion and viral population expansion [Ho, 1995 #20].

### 1.7 Antiretroviral Drugs

**1.7.1 Reverse Transcriptase Inhibitors**

1. **Nucleoside Reverse Transcriptase Inhibitors (NRTIs)**

NRTIs are analogs of normal nucleotides but without 3’ hydroxyl group (reviewed in [Sarafianos, 2004 #293]). The drug is taken in unphosphorylated form, which cytokinases phosphorylates to form 5’ triophosphates (reviewed in [De Clercq, 2002 #414;Ilina, 2008 #413]). It leads to incomplete termination of cDNA synthesis [Sluis-Cremer, 2000 #5]. The list of NRTI drugs shown in Table 1.1.

1. **Non-nucleoside reverse transcriptase inhibitors (NNRTIs)**

NNRTIs specifically bind at an allosteric site 10 [Å](http://en.wikipedia.org/wiki/%C3%85) from polymerase active site of the HIV-1 reverse transcriptase [Himmel, 2006 #1018;Sarafianos, 2009 #426], close to substrate-binding site. The binding induces conformational change in the enzyme, which distorts the catalytic aspartate triad of active site and inhibits the function of the enzyme [Esnouf, 1995 #431;Balzarini, 2004 #442]. The list of NNRTI drugs is shown in Table 1.1

**Table 1.1: Antiretroviral drugs used in the treatment of HIV infection, their rechanisms of action and mechanisms of resistance.**

**Source: Adapted from François Clavel and Allan J. Hance 2004 [Clavel, 2004 #314] and [Ammaranond, 2012 #245;Colin, 2013 #1175]**

| **Drugs** | **Mechanism of Action** | **Mechanisms of Resistance** |
| --- | --- | --- |
| **Fusion and entry inhibitors** | | |
| Enfuvirtide (T-20) | 36 amino acid peptide derived from the HR2 domain of glycoprotein 41  Interferes with glycoprotein 41 dependent membrane fusion | Mutations affect HR1, a domain of glycoprotein 41 whose interaction with HR2 promotes membrane fusion |
| Maraviroc | Binds to CCR5 co receptor | Development of CXCR4 using HIV HIV-1 exploits CCR5 conformational heterogeneity |
| **Nucleoside reverse transcriptase inhibitors (NRTls)** | | |
| Zidovudine (AZT) | Analogues of normal nucleosides  Active as triphosphate derivatives  Incorporated into nascent viral DNA  Prematurely terminate HIV DNA synthesis | Thymidine analogue mutations promote ATP-mediated and pyrophosphate-mediated excision of the incorporated terminator |
| Didanosine (ddl) |
| Zalcitabine (ddC) |
| Stavudine (d4T) |
| Lamivudine (3TC) |
| Abacavir (ABC) |
| Tenofovir disoproxil (TVD) |
| Emtricitabine (FTC) |
| **Nucleotide reverse transcriptase inhibitors (NtRTls)** | | |
| Tenofovir | Same as nucleoside analogues | K65R impairs incorporation of tenofovir into DNA  Thymidine analogue mutations often associated with cross-resistance to tenofovir |
| **Non-nucleoside reverse transcriptase inhibitors (NNRTls)** | | |
| Nevirapine (NVP) | Bind a hydrophobic pocket of HIV type 1 reverse transcriptase  Block polymerization of viral DNA  Inactive against HIV type 2 | Mutations reduce affinity of the inhibitors for the enzyme  Single mutations generally sufficient to induce high level of resistance |
| Delavirdine (DLV) |
| Efavirenz (EFV) |
| Etravirine (ETR) |
| Protease Inhibitors | | |
| Saquinavir (SQV) | Structure derived from natural peptidic substrates of the HIV type 1 protease  Bind the active site of the protease | Mutations reduce affinity of the inhibitors for the enzyme  High-level resistance requires a accumulation of mutations |
| Ritonavir (RTV) |
| Indinavir (IDV) |
| Nelflnavir (NFV) |
| Amprenavir (APV) |
| Lopinavir + Ritonavir (LPV/r) |
| Fosamprenavir (FPV) |
| Tipranavir (TPV) |
| Darunavir (DRV) |
| Atazanavir (ATV) |
|  |  |  |
| **Integrase Inhibitors** | | |
| Raltegravir | Binds selectively to the enzyme complexes that results in inhibiting strand transfer of viral and host DNA | Mutations at conserved carboxylate residues (Asp64 and Asp116) |

**1.7.2 Protease Inhibitors (PI)**

Protease Inhibitors interfere with cleavage of *gag*-*pol* polypeptide [Seelmeier, 1988 #1020] as competitive peptidomimetic inhibitors. The hydroxyethylene core in the inhibitors prohibits the cleavage action of the HIV protease enzyme [Vacca, 1994 #401;Vacca, 1991 #1022]. However, as an adverse side effect, patients consuming the inhibitors have developed lipodystrophy and hyperlipidemia [Carr, 2000 #955;Carr, 1998 #953;Carr, 1998 #951;Carr, 1998 #952;Liang, 2001 #433;Miller, 2000 #954;Tsiodras, 2000 #956]. The list of protease inhibitors is show in Table 1.1.

**1.7.3 Integrase Inhibitors**

The feasibility and efficacy of integrase inhibitors have been tested in Rhesus Macaques [Hazuda, 2004 #266]. Most of the integrase inhibitors target strand transfer function of the enzyme [Bera, 2011 #559;Espeseth, 2000 #957;Hazuda, 2000 #374;McColl, 2010 #111;Pannecouque, 2002 #373]. X-ray structure of the integrase enzyme has revealed the active site model of the enzyme complexes with the DNA [Chen, 2008 #367]. The only integrase inhibitor showing promising anti retroviral effect through animal model is Raltegravir and is currently in clinical trials (reviewed in [Ammaranond, 2012 #245]).

**1.7.4 Cell entry inhibitors**

The cell entry inhibitors interfere viral binding or fusion of HIV to a host cell. Following are the two classes of cell entry inhibitors.

1. **CCR5 co receptor antagonist:**

Maraviroc is the only CCR5 antagonist in clinical use [De Clercq, 2005 #337;De Clercq, 2005 #416;Fätkenheuer, 2005 #335;Rosario, 2005 #336;Rosario, 2006 #333;Wheeler, 2007 #334]. It is the only anti retroviral drug that does not target any viral enzyme or protein molecule but, instead, binds to the host cell receptor CCR5 [Westby, 2005 #1157]. This binding prevents HIV gp120 binding to the co-receptor, thereby disabling the viral entry in to the cell [Fätkenheuer, 2005 #335]. However, it is important for HIV tropism test for the viral co-receptor use, before administrating this drug, as the drug is ineffective against CXCR4 co receptor using viruses [Raymond, 2010 #1176].

1. **Fusion Inhibitors**

Fusion inhibitor design is based on targeting the heptad regions HR1 or HR2 of gp41, which disables HIV to create fusion pore on host cell membrane (reviewed in [Baldwin, 2003 #959]). Enfuvirtide [Duffalo, 2003 #318;Poveda, 2005 #317] is a synthetic peptide, approved for clinical use in 2003 [Robertson, 2003 #354], which can bind to gp41 HR1 region [Wild, 1993 #321]. Emergence of Enfuvirtide resistant viral strains lead to its discontinued use for clinical use in 2004 [Briz, 2006 #362;Melby, 2007 #315]. Sifuvirtide is another HIV fusion inhibitor peptide under research [Wang, 2009 #316].

### 1.8 Brief history of antiretroviral treatment

The treatment of HIV infection started in 1986 with the only available drug – Zidovudine (AZT), an anti-cancer drug, which was once rejected for its toxic and unpleasant side effects [Richman, 1987 #1304]. Although Zidovudine was expensive, the drug was the only hope for HIV infected people until 1991 when a new drug Zalcitabine (ddC) was approved for use.

### 1.8 HIV Treatment

The World Health Organization (WHO) has produced year 2013 clinical guideline (<http://www.who.int/hiv/pub/guidelines/arv2013/art/en/index.html>) for HIV treatment. From the treatment point of view, the guideline has grouped HIV infected people as Adult, pregnant and breast feeding women, children less than 3 years old, children 3 or more years old and HIV people co-infected with other diseases. Regardless of the grouping, the guideline recommends treatment initiation to all confirmed HIV infected people with CD+ cell count less than or equal to 350 per mm3. However, in resource poor settings, where laboratory tests for CD+ count may not be feasible, WHO clinical stages should be used as a guide for treatment initiation [Weinberg, #1161].

According to the WHO consolidated anti retroviral guideline recommendation for HIV treatment, a combination of three reverse transcriptase inhibitors (2 NRTIs + 1 NNRTI) are used as first line therapy. Addition of protease inhibitor is recommended for children below 3 years old. On virologic failure to first line, second line drug regimen containing 2 NRTIs and a ritonavir boosted protease inhibitor is recommended. Following failure to second line therapy, a new drug class – integrase inhibitor is introduced in third line therapy along with a reverse transcriptase and a protease inhibitor.

Patients under ART therapy are monitored for effectiveness of the treatment at a defined interval. Laboratory test for viral load should be done for monitoring the treatment response. The viral load of greater than 1000 viral RNA copies/ml blood sample indicates virologic failure to the treatment and the patient is recommended to switch to new drug regimen (Figure 1.8). If viral load test is not feasible routinely, CD4+ count and clinical monitoring need to be used.

### 1.8 HIV Drug Resistance

Anti retroviral drug therapy has been successful at controlling HIV replication to low viral load within a host [Rosenberg, 2000 #960]. The therapy has been compromised and threatened for long term sustainability by the emergence of drug resistant HIV viral variants (reviewed in [Clavel, 2004 #314]). HIV mutates spontaneously in every replication cycle [Dumonceaux, 1998 #1168;di Marzo Veronese, 1993 #1169]; the error prone reverse transcriptase introduces mutations randomly at the rate of 10-5 per base per replication cycle [Mansky, 1996 #1170;Mansky, 1995 #18]. The accumulation of mutations and viral recombination contribute the most for the emergence of drug resistant variants (reviewed in [Menéndez-Arias, 2002 #312;Tebit, 2007 #469;Smyth, 2012 #919]). In the presence of anti retro viral drugs the HIV variants in the quasispecies get selected [Nijhuis, 1999 #2;Leslie, 2004 #192]. The selected variants proliferate to high number in subsequent replication cycles [Hosseinipour, 2013 #1167]. Transmission of resistant virus also confers drug resistance in drug naïve patients [Yerly, 1999 #4;Boden, 1998 #284;Wensing, 2005 #3]. Although drug resistant mutations escape drug effect, they confer decrease in viral fitness and replication rate at certain level [Lucas, 2005 #1;Mammano, 2000 #1171]. Other mutations counteract the effect of drug resistant mutations to boost up viral fitness and replication rate [Nijhuis, 1999 #2].

ART drugs are designed based on three dimension structures of the viral proteins to bind at a specific active region of the proteins and interfere with its normal function

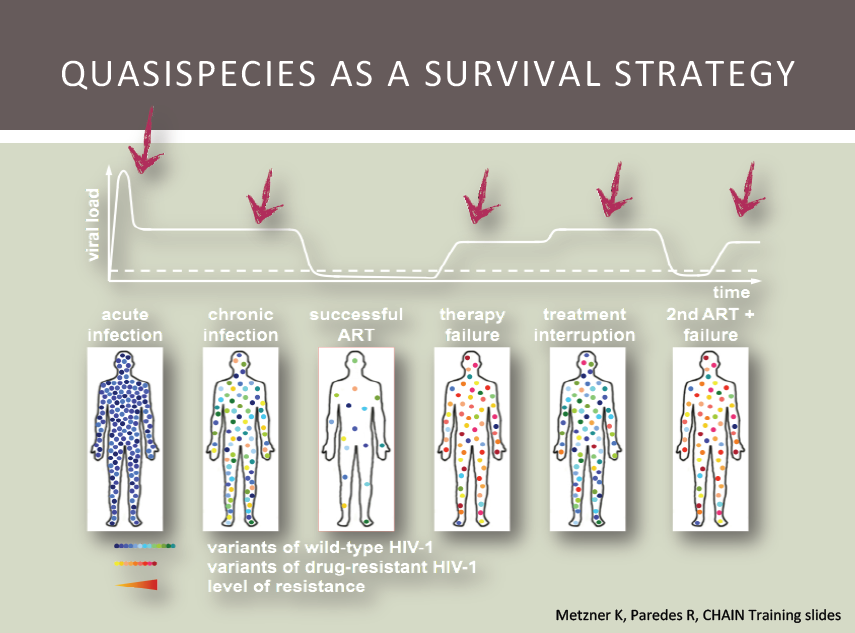


Figure 1.8: Dynamics of HIV viral load within an infected patient. Viral load peaks at acute infection followed by significant drop due to less CD4+ count. Introduction of highly active cocktail of anti retroviral drugs arrest viral replication cycle decreases the viral load to safe level. Spontaneous mutations give rise to resistant viruses that replicate with high turnover increasing the viral load and the therapy fails. Interruption of drugs contributes to higher viral load. Introduction of second line anti retroviral drugs suppress the viral load back to safe level again and the cycle of viral load repeats. Source: Roger Paredes personal communication

(reviewed in [Tantillo, 1994 #1033]). In general, mutations at the active site amino acid residues of HIV proteins change their three-dimensional shape, disabling the drug binding (Figure 1.9) [Miller, 2000 #965;Hsiou, 2001 #966;Tantillo, 1994 #1033;Fikkert, 2003 #265]. For more details on the mechanisms of HIV resistance, see Table 1.1.

The mutations attributing resistance to drugs have been well documented with associated drug and resistance level [Shafer, 2006 #310]. The viral quasispecies sequence information can be used to align against the reference sequence and infer the presence of any drug resistant mutation [Kuiken, 2003 #963].

### 1.9 HIV genotyping for Resistance Test

### Conventional Population Based HIV Drug Resistance Genotyping

The pattern of drug resistant mutations in HIV has strong relation with virologic failure [Lorenzi, 1999 #1045;Zolopa, 1999 #1046;van Leeuwen, 1995 #1047;Larder, 1989 #1049;Condra, 1995 #1050;Molla, 1996 #1051;Larder, 1989 #1052]. Given the database management of all the drug resistant mutations, the application of HIV genotype information on therapy optimization has shown virologic benefits in clinical settings [Baxter, 2000 #1053;Durant, 1999 #1055;Van Vaerenbergh, 2001 #229]. Conventional HIV genotyping involves Sanger dideoxy termination sequencing method [Ewing, 1998 #1114;Sanger, 1977 #1115;Metzker, 2005 #1116]. The method produces a chromatogram that represents an average genotype sequence of the viral population in the host [Struck, #1119].

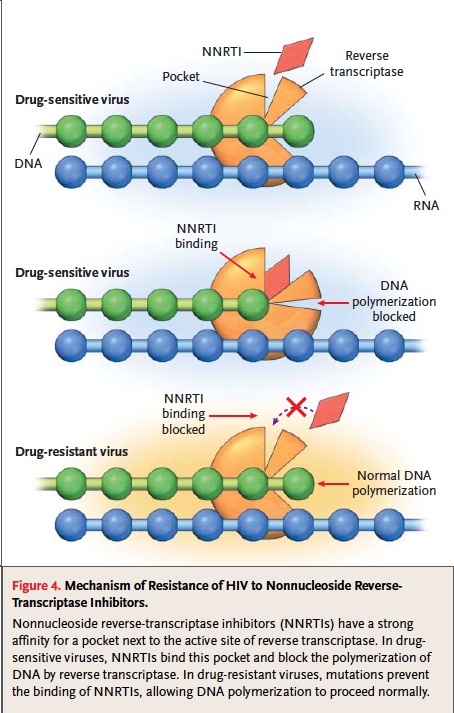


Figure 1.9: Development of HIV drug resistance against NNRTI. Accumulation of Drug resistant mutations changes three-dimensional confirmation of a viral protein, disabling drug binding and carrying out its normal function. Source: [Clavel, 2004 #314]

Conventional population based sequencing method has limited sensitivity, as it has been able to detect only the viral variants of frequency greater or equal to 20%; low prevalent variants go undetected [Ji, #1111;Johnson, #1112]. Thus, sequencing technology underestimates the total variants in the viral population [Palmer, 2005 #1118]. Undetected low frequency HIV variants are clinically significant for designing salvage therapy, as they are usually drug resistant [Simen, 2007 #313;Simen, 2009 #383;Rowley, #1122;Paredes, #1123]. Those minor variants rebound in the presence of drugs, leading to ultimate virologic failure [Rowley, #1122;Paredes, #1123]. This necessitates improved and highly sensitive sequencing technologies able to detect minor HIV variants in the viral quasispecies (reviewed in [Metzker, #1124]) .

### 1.10 Next Generation Sequencing technologies

The newer sequencing technologies developed as an alternative to automated Sanger sequencing is referred as Next Generation Sequencing (NGS) technologies. Commercially available NGS technologies in the market are Roche 454, Illumina, Applied Biosystems SOLiD technology and Ion torrent ). All the NGS systems has PCR step for DNA amplification but differ in other features [Shokralla, #1134]; each one outperforms the others at a particular application.

Roche 454 and Illumina implement ‘Sequencing by synthesis’ (SBS) technique for DNA sequencing [Margulies, 2005 #1135]. Both systems detect fluorescent light emitted from each nucleotide incorporation event during DNA synthesis [Margulies, 2005 #1127] (reviewed in [Shendure, 2008 #1132]). The intensity of fluorescent light emitted per base addition is too small for detection; this requires amplification of the template DNA to million copies. Simultaneous addition of same nucleotide in million copies produces higher intensity of detectable fluorescent light (reviewed in [Metzker, #1124;Shendure, 2008 #1132]). A defined order of free nucleotide molecules are flown in the reaction plate, allow nucleotide to incorporate, detect fluorescent light and wash off any unincorporated nucleotides for next cycle. Roche 454 and Illumina differ only at the sequencing step. In Roche 454, polymerase continues nucleotide addition reactions until the base flowing in the reaction plate is complementary to the template sequence. The intensity of fluorescent light emission is detected and the number of bases subsequently added, as a homopolymer run, is calculated based on the light intensity in a particular reaction cycle [Margulies, 2005 #1135]. Illumina, on the other hand, stops the reaction after single nucleotide addition [Bentley, 2008 #1136]. The light emission is detected that determines the base added in that particular cycle of base flow ([Bentley, 2008 #1136], reviewed in [Shendure, 2008 #1132]). These different sequencing techniques have high impact on sequence read length. Roche 454 produces low number of sequences but longer read length (Table 1.1). Comparatively, Illumina produces shorter reads in high quantity (Table 1.1).

Applied Biosystems SOLiD implements ‘Sequencing by ligation’ technique for DNA sequencing; thus, bypassing any DNA polymerase related sequencing error [Pandey, 2008 #1137]. The template DNA is amplified in similar way to previous NGS technologies. During sequencing, a universal primer and a library of pre-designed 1,2-probes of 8 nucleotides (or dibase probe) along with DNA ligase enzyme is added. The probes hybridize to the complementary template sequence and the fluorescence of the probe is read. The probe hybridization is repeated for seven cycles extending read length to 35 bases. In the next cycle, a new universal primer is hybridized at offset position of one base (n-1) to the previous primer position followed by ligation sequencing process. The primer resetting cycle is repeated five times providing dual measurement of each base and the final sequence is decoded from color code information using 4 by 4 color code (reviewed in [Mardis, 2008 #1126]).

Table 1:2 Comparative analysis of different NGS systems. Source: adapted from Shokralla et al 2012 [Shokralla, #1134], Niedringhaus et al 2011 [Niedringhaus, #1138] and Glenn 2011 [Glenn, 2011 #1140]

| Platform | Read length (bp) | Reads/run | Sequencing output/run | Run time | Advantages | Primary applications |
| --- | --- | --- | --- | --- | --- | --- |
| Roche 454 GS FLX | 400–500 | 1 × 106 | ≤500 Mb | 10 h | Longest read lengths among 2nd generation, high throughput compared to 1st generation sequencing | 1\*, 2, 3\*, 4, 7, 8\* |
| Roche 454 GS FLX+ | 600–800 | 1 × 106 | ≤700 Mb | 23 h |
| Roche 454 GS Junior | 400–450 | 1 × 105 | ∼35 Mb | 10 h |
| Illumina HiSeq 2000 | 100–200 | 6 × 109 | ≤540–600 Gb | 11 d | Very high throughput | 1\*, 2, 3\*, 4, 5, 6, 7, 8 |
| Illumina HiSeq 1000 | 100–200 | 3 × 109 | ≤270–300 Gb | 8.5 d |
| Illumina GAIIx | 50–75 | 6.4 × 108 | ≤95 Gb | 7.5–14.5 d |
| Illumina MiSeq | 100–150 | 7 × 106 | ≤1–2 Gb | 19–27 h |
| AB SOLiD 5500 system | 35–75 | 2.4 × 109 | ∼100 Gb | 4 d | Very high throughput; lowest reagent cost needed to reassemble a human genome among the widely accepted 2nd generation platforms, lower error rate | 3\*, 5, 6, 8 |
| AB SOLiD 5500 xl system | 35–75 | 6 × 109 | ∼250 Gb | 7–8 d |
| Ion Torrent -314 chip | 100–200 | 1 × 106 | ≥10 Mb | 3.5 h | Direct measurement of nucleotide incorporation events; DNA synthesis reaction operates under natural conditions (no need for modified DNA bases) | 1, 2, 3, 4, 8 |
| Ion Torrent -316 chip | 100–200 | 6 × 106 | ≥100 Mb | 4.7 h |
| Ion Torrent -318 chip | 100–200 | 11 × 106 | ≥1 Gb | 5.5 h |

Bold indicates applications that are most often used, economical or growing

1 = de novo BACs, plastids, microbial genomes. 2 = transcriptome characterization.

3 = targeted re-sequencing. 4 = de novo plant and animal genomes.

5 = re-sequencing and transcript counting. 6 = mutation detection.

7 = metagenomics. 8 = other

\*Pooling multiple samples with sequence tags (i.e. MIDs or indexes) is required for efficient use of this application

Ion torrent technology implements sequencing by synthesis method and electronic sensors connected to complementary metal-oxide-semiconductor integrated circuit with a microprocessor for signal processing [Jakobson, 2002 #148;Milgrew, 2004 #149]. The sequencing step is similar to Roche 454 homopolymer sequencing but the base detection is completely electronic. During DNA sequencing, a base incorporation releases a hydroxyl ion (H+) that shifts pH of the surrounding solution and this correlates directly to the number of nucleotides incorporated in that particular base flow cycle (reviewed in [Niedringhaus, #1138].

This change in pH is detected by sensor at the bottom of each well, converted to a voltage and digitalized by semi conductor CMOS integrated circuits [Pennisi, #1139]. Signal processing software is used to convert the data for measurement of incorporation in that flow using a physical model [Rothberg, 2011 #174].

### 1.11 HIV-1 Drug resistance Genotyping in the era of high throughput sequencing

The next generation sequencing technologies are highly demanded in the current genomic research in comparison to Sanger sequencing method for two reasons: low cost and ultra deep high throughput. The application of NGS has contributed to many research findings in organisms like HIV that exhibit high heterogeneity at genetic sequence. HIV mutations has strong link to viral fitness and survival in the presence of antiretroviral drugs and every mutation known to correlate drug resistance is maintained in a proper database. Therefore, it is highly essential to accurately detect HIV drug resistant mutations while assessing genetically diverse viral quasispecies.

Genotyping of HIV quasispecies are essential to identify the mutations responsible for drug resistance. HIV genotyping with NGS include amplicon-based approach of resequencing HIV genomic section that covers drug resistant mutations. Longer sequencing reads from NGS that cover the amplicon region is more advantageous to short reads for assessing drug resistant (reviewed in [Droege, 2008 #1189]). Roche 454 system with longest read length (Table 1.1) NGS system is, thus, increasingly used for genotyping HIV quasispecies.

Although Roche 454 produces large amount of sequence data with long reads, it is error prone and that increases the artificial diversity in the sample confounding the real variation. The PCR amplification step has embedded errors like DNA recombination [Kanagawa, 2003 #1179;Meyerhans, 1990 #1181;Yang, 1996 #1182], DNA synthesis error [Hughes, 2003 #1178;Mansky, 1995 #18] and DNA re-sampling error [Liu, 1996 #1183]. In addition to this, ultra deep sequencing underlying algorithm has inherent base miscall error [Quinlan, 2008 #1185]. The frequency of base call error is six times in homopolymeric region than non-homopolymeric region [Margulies, 2005 #177;Wang, 2007 #1187]. Enhancement in sequencing chemistry and base calling algorithm in Roche 454 Genome Sequencer FLX system improved accuracy of data to >99.5% including errors at homopolymer region (reviewed in [Droege, 2008 #1189]). At the coverage of 10x, the FLX system is able to identify all sequence variations that is known from Sanger sequencing [Bordoni, 2008 #1186]. An accuracy of 99.28% is reported, all errors at homopolymer region but given the coverage ≥ 30x, FLX system correctly identified all sequencing error [Bordoni, 2008 #1186].

Ultra deep high throughput sequencers associates quality score to every base. The score denotes probability that the corresponding base call is correct [Cock, 2010 #425]. It is essential to trim out the poor quality scored bases ensure high quality data in the downstream analysis.

Errors generated in early cycles of PCR get amplified in subsequent cycles and that they might be associated with high quality score during ultra deep sequencing. These erroneous bases can escape at quality trimming given their high quality score. A novel technique of tagging every template polynucleotide with a primer ID is introduced to track every ultra deep sequence read that originates from same source template sequence [Jabara, 2011 #1188]. Consensus sequence representing sequences with same Primer ID denotes the original source template and that corrects every known sequencing error [Jabara, 2011 #1188].

Ultra deep high throughput sequencers produce large volume of data that often is difficult to analyze at human scale. Analysis of data often requires multiple processing steps of specific task like grouping the data with similar information, quality control, mapping the sequence data to a reference sequence to infer any mutations. Reference mapped HIV quasispecies sequence data can be used to query a standard HIV drug resistance database e.g. Stanford database. Bioinformatics computational tools are used for each specific task but still they needs to be run manually. An automated computational pipeline where all the tasks can be run in step-wise fashion is a current necessity.

### 1.12 Thesis Outline

**Chapter 2**: The chapter introduces and describes, in detail, a novel algorithm for quality trimming called QTrim. The tool is designed currently to quality control Roche 454 ultra deep sequence data and optimized for both poor and high quality data. QTrim is compared to other widely used quality trimming tools and the comparative analysis result is also presented. The entire chapter is produced as a paper for publication entitled “**QTrim: A novel tool for the quality trimming of sequence reads generated using the Roche/454 sequencing platform” (Shrestha, RK and Travers, Simon; being reviewed)**.

**Chapter 3**: The chapter introduces Primer ID technique recently developed to correct ultra deep sequencing errors. The chapter also discusses a comparative analysis of sequence data with and without Primer ID approach and the effect on downstream analysis. The Primer ID approach sequence data is obtained from a study conducted to observe HIV virus response on the vaccine. The study was conducted in the University of Cape Town.

**Chapter 4**: The chapter introduces and describes the computational pipeline called Seq2Res that facilitates low cost HIV drug resistance test. The chapter describes a workflow of the pipeline, the requirements of the pipeline, HIV drug resistance output files and plots that summaries overall analysis. The test data is obtained from a study called CIPRA-SA, which is a trial of anti retroviral drug monitoring strategy in resource poor setting. The chapter is included in paper publication.

**Chapter 5**: This chapter discusses on comparison of clinical data and sequence data analyzed using Seq2Res pipeline. Excitingly, the sequence data analyzed using Seq2Res agrees to clinical data.

Need to reshuffle

Generally, protease inhibitors are added to second line ART therapy following failure of first line therapy [Pujades-Rodriguez, 2008 #1019].

Although antiretroviral therapy can suppress viral replication, they are expensive, can lead to multiple drug resistance and, thus, require high levels of adherence [Cohen, 2002 #376]. These necessitate alternative drugs that target the third enzyme, the integrase.

**Bibliography**