CHAPTER 1

# Literature Review

## Overview of HIV/AIDS

Human Immunodeficiency Virus (HIV) is a human pathogenic virus that causes Acquired Immunodeficiency Syndrome (AIDS). 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 (WHO factsheet Number 360 (<http://www.who.int/mediacentre/> factsheets/fs360/en/)) and that over 95% of them are living in low and middle income 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 [UNAIDS, #703]. Countries in Sub-Saharan Africa also have varying HIV prevalence with South Africa at the top followed by Nigeria [Esparza, 2000 #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, 2000 #782].

## 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 Acquired Immunodeficiency Syndrome (AIDS), 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 were researching 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].

## 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, were 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 was discovered in 1986. This group was transmitted from sooty mangabey monkeys (Cercocebus atys) [Hirsch, 1989 #657] and its prevalence was also high in the geographical location of these monkeys in West Africa [Santiago, 2005 #658]. Sooty mangabey monkeys were naturally infected by a strain of SIV [Hirsch, 1989 #657]. The phylogenetic analysis of HIV-2 strains showed that they closely group with the SIVsmm strain [Hirsch, 1989 #657] that were 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 were identified in human, it was assumed that more subtypes were introduced into human [Gürtler, 2004 #654] but were 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 branch 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].

## HIV-1 Diversity

### 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].

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 [Taylor, 2008 #214]); Subtype F has sub-subtypes F1 and F2 (Figure 1.3); sub-subtype F3 mentioned by Taylor et al [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).

### 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 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 that is 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 Los Alamos National Laboratory website (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].

### 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, for example *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 is the overall source of genetic heterogeneity in the viral population (reviewed in [Smyth, #919]).

Intra patient HIV 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]).

## 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].

### 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].

### 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].

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].

### 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].

## 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 molecules 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 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. 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].

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].

## Antiretroviral Drugs

### Reverse Transcriptase Inhibitors

#### Nucleoside Reverse Transcriptase Inhibitors (NRTIs)

NRTIs are analogs of 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]). This then, leads to the incomplete termination of HIV-1 cDNA synthesis [Sluis-Cremer, 2000 #5]. The action of the drugs is shown in **Figure 1.8** and the list of approved NRTI drugs shown in **Table 1.1**.

#### Non-nucleoside reverse transcriptase inhibitors (NNRTIs)

NNRTIs specifically bind at an allosteric site 10 Å from the polymerase active site of the HIV-1 reverse transcriptase [Himmel, 2006 #1018;Sarafianos, 2009 #426], close to the substrate-binding site. The binding induces conformational changes in the enzyme, which distorts the catalytic aspartate triad of its active site and inhibits the function of the enzyme (**Figure 1.9**) [Esnouf, 1995 #431;Balzarini, 2004 #442]. The list of approved NNRTI drugs is shown in **Table 1.1,**

### Protease Inhibitors (PI)

Protease Inhibitors interfere with the cleavage of the *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 that have used these 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].

### Integrase Inhibitors

The feasibility and efficacy of integrase inhibitors have been tested in Rhesus Macaques [Hazuda, 2004 #266]. Most of the integrase inhibitors target the strand transfer function of the enzyme [Bera, 2011 #559;Espeseth, 2000 #957;Hazuda, 2000 #374;McColl, 2010 #111;Pannecouque, 2002 #373]. An 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 that has shown a promising antiretroviral effect is Raltegravir, which was tested on animal models and is currently undergoing clinical trials (reviewed in [Ammaranond, 2012 #245]).

### Cell entry inhibitors

The cell entry inhibitors interfere with the viral binding or fusion of HIV to a host cell. The two classes of cell entry inhibitors are listed below:

#### 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 also the only antiretroviral 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 to carry out an 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].

#### Fusion Inhibitors

Fusion inhibitor design is based on targeting the heptad regions HR1 or HR2 of gp41, which prevents HIV from creating a 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 the gp41 HR1 region [Wild, 1993 #321]. However the emergence of Enfuvirtide resistant viral strains lead to its discontinuation for clinical use in 2004 [Briz, 2006 #362]. Sifuvirtide is another HIV fusion inhibitor peptide under research [Wang, 2009 #316].

## HIV Treatment

### Brief history of antiretroviral treatment

The treatment of HIV infection has been a great challenge and still remains as an unsolved problem [Sandstrom, 1987 #1324]. In 1985, an assay for diagnosis of HIV antibody was developed for the confirmation of HIV infection [Ward, 1986 #1325]. Clinical treatment for those with confirmed HIV infection started with the only available NRTI drug – azidothymidine (AZT), (later called Zidovudine (ZDV)). The drug is characterized for its toxic and unpleasant side effects [Richman, 1987 #1304;Koch, 1992 #1326]. Nonetheless, the drug was the only hope for HIV infected people at the chronic stages of infection in mid 1980’s and was approved for use but the survival benefits lasted less than a year [Fischl, 1990 #1327;Fischl, 1993 #1329;Volberding, 1990 #1332;Lundgren, 1994 #1333;Volberding, 1995 #1334]. Other NRTI drugs including didanosine (ddI) in 1991, Zalcitabine (ddC) in 1992, stavudine (d4T) in 1994 and lamivudine (3TC) in 1995 - were approved for use (**Figure 1.10**) but were toxic as well. The administration of the drugs was altered to reduce the toxicity of each drug but the approach remained ineffective [Skowron, 1993 #1335]. Then, a combination therapy containing two NRTI drugs [Saravolatz, 1996 #1338], for example zidovudine with didanosine or zalcitabine showed some improvement, characterized by increased CD+ and better survival but with less durability and poor tolerability [Hammer, 1996 #1336]. Triple NRTI combination therapy containing 3TC, ZDV and d4T was better tolerated but could not control HIV reproduction [Kuritzkes, 1999 #1337]. A good result obtained from using those NRTI drugs was the substantial reduction in HIV transmission from mother to child at birth [McIntyre, 2009 #1342;McGowan, 2000 #1343;Connor, 1994 #1341].

A notable advancement in antiretroviral treatment was observed after the development of NNRTI drugs and PI drugs that interacted directly with the viral proteins reverse transcriptase and protease to inhibit their action. Clinical trials were conducted with triple combination therapy contained 2 NRTIs and a NNRTI drug or 2 NRTIs and a PI drug [Staszewski, 1999 #1375;Montaner, 1998 #1377;Montaner, 1998 #1344]. Besides antiretroviral activity, combination therapy was also studied for toxicity and tolerability [Montaner, 1998 #1377;Staszewski, 1999 #1350;Staszewski, 1999 #1375]. The triple combination therapy of Nevirapine/efavirenz (NNRTI drug) with two NRTI drugs showed a good viral suppressing result [Staszewski, 1999 #1350;Staszewski, 1999 #1375] and was superior to monotherapy and dual therapy [Robbins, 2003 #1340].

A drug cocktail with 2 NRTIs and a protease inhibitor showed highly effective result [Cameron, 1999 #1346;Merry, 1997 #1347] with viral suppression time longer than the study period [Hammer, 1997 #304;Gulick, 2000 #1348]. The concept of highly active antiretroviral therapy was conceived after the cocktail of three drugs from different classes showed effective results [Hammer, 1997 #304;Gulick, 1997 #305;Gulick, 1998 #1400]. The success of triple drug therapy was reported in Vancouver AIDS conference in 1996. In a short time, recommendations for antiretroviral therapy were published to manage HIV infections [Carpenter, 1997 #1353]. More drugs from NRTI, NNRTI and protease inhibitors were developed with lower toxicity and higher potency than the earlier drugs. After years of researching different drug combination, the first drug regimen for ‘standard-of-care’ is available consisting of two NRTI drugs and a third drug from any other drug class [Vella, 2012 #1355].

### Treatment guideline

The World Health Organization (WHO) has produced the 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 individuals as adult, pregnant and breast feeding women, children less than 3 years old, children 3 or more years old and HIV infected individuals co-infected with other diseases. Regardless of the grouping, the guideline recommends treatment initiation to all confirmed HIV infected people with CD4+ cell count less than or equal to 350 cells/mm3 in resource poor countries and less or equal to 500 cells/mm3 in resource rich countries. However, where laboratory tests for CD4+ count may not be feasible, the WHO clinical stages should be used as a guide for treatment initiation [Weinberg, #1161].

The WHO antiretroviral guideline recommends a combination of 2 NRTIs and 1 NNRTI as first line therapy. The addition of a protease inhibitor is recommended for children below 3 years old. On virologic failure to first line treatment, a 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 – the integrase inhibitor is introduced in third line therapy along with a reverse transcriptase and a protease inhibitor.

Patients undergoing ART therapy are monitored for effectiveness of the treatment at a defined interval. Laboratory tests for viral load should be done for monitoring the treatment response. A 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 1.11). If a viral load test is not feasible routinely, CD4+ count and clinical monitoring need to be used [Weinberg, 2010 #1161].

## HIV Drug Resistance

Currently there are 20 approved antiretroviral drugs that include 8 PIs, 7 NRTIs, 4 NNRTIs and 1 integrase inhibitor. Antiretroviral treatment using a drug or a combination of different class drugs, selects the viral variants with mutations that emerge out to become a major variant in the viral population [Simen, 2009 #383][Shirasaka, 1995 #1371] that could lead to virologic failure (Figure 1.11).

A study by Richman et al [Richman, 1994 #1351] showed that all 38 HIV infected individuals treated with nevirapine developed reduced susceptibility towards the drug by 12 weeks of treatment with known drug resistant mutations. 14 patients when treated with a combination of nevirapine and azidovudine showed that 57% developed K103N mutation in RT in 32 weeks [Richman, 1994 #1351].

Gupta et al [Gupta, 2008 #1732] showed that when 7970 patients were treated with a combination of two NRTIs and a NNRTI or a PI, 4.9% of patients receiving NNRTI and 5.3% of patients receiving PI had virologic failure after 48 weeks. Genotyping of the samples from the virologic failure patients showed that M184V mutation was found in 35.5% of patients receiving NNRTI in the drug combination and 21.0% in patients receiving PI in the drug combination [Gupta, 2008 #1732].

The error prone nature of the reverse transcriptase [Dumonceaux, 1998 #1168;di Marzo Veronese, 1993 #1169] and high turnover per replication per viral cycle [Mansky, 1996 #1170;Mansky, 1995 #18] are two major driving forces that result in multiple mutations conferring resistance to the drug [Larder, 1989 #1049;Larder, 1991 #1094;Kellam, 1994 #1365;Tisdale, 1993 #1366].

Studies showed that HIV replication was suppressed for longer periods than earlier monotherapy after the introduction of NNRTI drugs or protease drugs, or both inclusive triple combination therapies [Montaner, 1998 #1377;Staszewski, 1999 #1375;Collier, 1996 #1357]. However, 700 of 1402 patients treated with NRTI, NNRTIs and a protease inhibitor included triple combined therapy had virologic failure at 12 months and 94 died of AIDs related sickness after 14 month [Grabar, 2000 #1733].

The drug resistant mutations change the viral protein structure that disables the drug binding to the enzyme (**Figure 1.12**). This limits the therapy options and drug failure with new combinations in short time [Hammer, 2008 #1413]. Even the combination therapy of five drugs including 2 NRTIs, one NNRTI and 2 PIs has resulted in poor virologic response in just 24 weeks in a study by Piketty et al [Piketty, 1999 #1388]. These studies also show that the drug resistant viral variants can vary from high level to undetectable level and that suggests the necessity of drug resistance testing before initiating antiretroviral therapy [Hanna, 2001 #1390].

Phenotypic and genotypic assays are available for drug resistance testing. A phenotypic assay includes viral stock generation from peripheral blood mononuclear cells (PBMCs), titration of stock to get viral infectivity, infection of cell culture with known concentrations of antiretroviral drugs and calculation of inhibitory concentration (IC) 50 and 90, based on a measure of infection. The limitations of the method include: its labor intensive, minimum of six weeks time requirement, in vitro viral selection pressure during the assay period and use of PBMCs only (not virus in plasma) for drug susceptibility test [Hanna, 2001 #1390].

The limitations led to development of HIV resistance assays based on recombination of the virus from plasma samples [Hertogs, 1998 #1364;Kellam, 1994 #1391;Shi, 1997 #1392;Martinez-Picado, 1999 #1393;Petropoulos, 2000 #1394]. The recombinant assays are based on extraction of the plasma viral genome, amplification of PR and RT regions, insertion of the sequence into a HIV vector to produce recombinant virus that are used for infection of cell culture on which drug susceptibility test is done at IC50 and IC90. Antivirogram assay (Virco, Mechelen, Belgium) [Hertogs, 1998 #1364] and PhenoSense assay (ViroLogic, South San Francisco, California) [Petropoulos, 2000 #1394] are two automated recombinant assays; both require up to 10 days to complete the resistance test. These assays sample the predominant variant in the viral population while minor variants may go undetected that could lead to drug failure [Simen, 2009 #1397]. The assays involve complexities and are expensive (reviewed in [Hirsch MS, 2000 #451]).

Genotypic assays are based on mutations inferred from gene sequences. Specific mutations in HIV-1 provide resistance to related antiretroviral drugs (**Figure 1.13** and **Figure 1.14**). Drug resistant mutations from HIV test sequences can be inferred comparing the sequences with the HIV *pol* reference sequence and the mutations can be compared with a database **(**e.g. the Stanford HIV database [Rhee, 2003 #1224]) of known drug resistant mutations using genotypic interpretation algorithm [Rhee, 2009 #1526]. The interpretation algorithm provides resistance scores for the combination of drug resistant mutations that indicates the level of resistance to the associated drugs. The known drug susceptibility information on the combination of drug resistant mutations, can be used to infer the drug susceptibility of the HIV genotypic sequence data classed as susceptible, resistant and intermediate resistant [Mayer, 2001 #1419;Larder, 1999 #1420].

Sanger based technology [Sanger, 1977 #183] has been the standard for sequencing HIV-1 genes for drug resistance genotyping. Oligonucleotide hybridization based genotypic assays, as in GeneChip (Affymetrix) [Kozal, 1996 #1398] and LiPA (InnoGenetics) [Stuyver, 1997 #1399], were in used but limited to preselected drug resistant mutation codons.

## HIV drug resistance genotyping

### Conventional Population Based HIV Drug Resistance Genotyping

The presence/absence of certain drug resistant mutations (**Figure 1.13** and **Figure 1.14**) 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] and characterization of these drug resistance mutations (DRMs) can be used to optimize the antiretroviral therapy [Baxter, 2000 #1053;Durant, 1999 #1055;Van Vaerenbergh, 2001 #229]. Conventional HIV genotyping involves Sanger dideoxy termination based population sequencing that produces consensus sequence of the viral population in a sample [Ewing, 1998 #1114;Sanger, 1977 #1115;Metzker, 2005 #1116].

The Sanger technology outputs a chromatogram that shows a peak for all the bases at a particular position of a gene sequence for the viral population [Struck, #1119]. The sequence of the mixed population is determined based on the peaks generated for the bases called. At the position of nucleotide mixture, besides the highest peak, the lower peaks with greater or equal to 20% height of the highest peak are also marked and the ambiguous base representing the marked bases are added to the sequence [Woods, 2012 #1108]. Thus, conventional population based sequencing method has limited sensitivity; the low frequency variants below 20% prevalence are not detected reliably [Ji, #1111;Johnson, #1112;Palmer, 2005 #1118]; and it underestimates the total number of variants in the viral population [Palmer, 2005 #1118]. Undetected low frequency HIV variants have been shown to be clinically significant [Rowley, 2010 #1122;Simen, 2007 #313;Simen, 2009 #1547;Paredes, 2010 #1123]. These minor variants rebound in the presence of drugs, leading to ultimate virologic failure [Rowley, #1122;Paredes, #1123]. Besides this sensitivity limitation, the conventional population based genotyping method is highly expensive [Dias-Neto, 2009 #1704][Liu, 2012 #1705] and this limits the application of the technology for resistance genotyping in high HIV prevalent resource limited settings. This necessitates improved, highly sensitive sequencing and cost-effective technologies able to detect minor HIV variants in the viral quasispecies [Dudley, 2012 #1209][Wang, 2007 #1103].

## Next Generation sequencing technologies

The sequencing technologies developed with much higher throughput than automated Sanger sequencing are known as next generation High Throughput Sequencing (HTS) technologies. Commercially available next generation HTS technologies in the market are Roche/454, Illumina, Applied Biosystems SOLiD technology, Ion torrent) and the recent third generation single molecule real time HTS technology – PacBio.. HTS systems differ from each other in terms of total raw sequence reads output, sequencing error rate, read length, sequencing time (Table 1.2), sequencing chemistry and sequencing cost (reviewed in [Shendure, 2008 #1132;Metzker, 2009 #185]).

Roche/454 and Illumina implement a ‘Sequencing by synthesis’ (SBS) technique for DNA sequencing [Margulies, 2005 #1135]. DNA fragments are PCR amplified to millions of copies such that while sequencing, simultaneous addition of million bases, one to each growing strand of template fragment, emits detectable fluorescent light [Margulies, 2005 #177]. A defined order of free nucleotide molecules are flowed in the reaction plate, nucleotides are allowed to incorporate, fluorescent light is detected and any unincorporated nucleotides are washed off 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 is proportional to the number of bases subsequently added, as a homopolymer run, in a particular reaction cycle [Margulies, 2005 #1135]. In the homopolymer region (repetition of a base over 3 times) the light intensity and the bases added can be disproportionate, generating high insertion or deletion (indel) errors [Loman, 2012 #1410;Luo, 2012 #1405] at the rate of 0.38 per 100 bases [Loman, 2012 #1410]. Illumina, on the other hand, stops the reaction after single nucleotide addition, detects the color of light emission that depends on a base [Bentley, 2008 #1136] but has base calling biases, generating substitution errors [Luo, 2012 #1405]. The major advantage of Illumina over other systems is that it has the highest throughput (**Table 1.2**). The sequencing chemistry of the systems impacts on sequence read length. Roche/454 yields a lower number of sequences but the longest read length (up to 800 bases) (**Table 1.2**).

Applied Biosystems SOLiD implements a ‘Sequencing by ligation’ technique for DNA sequencing, thus bypassing any DNA polymerase related sequencing errors [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 a 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 only 35 bases. In the next cycle, a new universal primer is hybridized at an offset position of one base (n-1) to the previous primer position followed by a ligation sequencing process. The primer resetting cycle is repeated five times providing dual measurements of each base and the final sequence is decoded from color code information using 4 by 4 color code (reviewed in [Mardis, 2008 #1126]).

Ion torrent technology implements sequencing by synthesis method and electronic sensors connected to complementary metal-oxide-semiconductor integrated circuit are used 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, and that reduces the ion torrent cost relative to other systems [Glenn, 2011 #1411]. During DNA sequencing, a base incorporation releases a hydroxyl ion (H+) that shifts the 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 a 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 base incorporations in that flow using a physical model [Rothberg, 2011 #174]. The final sequences generated, after processing, have the read length up to 200 bases (lower than Roche/454) but like Roche/454, Ion torrent sequences have indel errors at homopolymer regions at rate of 1.5 per 100 bases [Loman, 2012 #1410].

Pacific Bioscience’s Single Molecule Real Time technology is considered the third generation technology available in the market now ([www.pacificbiosciences.com](http://www.pacificbiosciences.com)). The technology does not involve PCR amplification of the template DNA; instead the base sequencing is done on single molecule of a DNA, enabling the detection of variation at molecule level. Besides this advantage, it provides the read length of about 10,000 bases [McCarthy, 2010 #1679;Eid, 2009 #1683].

DNA sequencing takes place in the zero mode waveguide (ZMW) [Levene, 2003 #1676]. ZMW is a nano-size chamber that is 7 nanometer in diameter and 10 nanometer in depth [McCarthy, 2010 #1679]. A DNA template and polymerase complex is immobilized at the base of a ZMW and different color flurophore labeled nucleotides are added into a ZMW chamber. During base incorporation at sequencing step, nucleotide fluorescence is detected with the light that illuminates the ZMW chamber, followed by cleavage of the flurophore.

A comparative study of the sequencing platforms Ion torrent, illumina and pacfic bioscience SMRT by Quail et al [Quail, 2012 #528] showed that error rate of the SMRT technology was 13%. The number of sequence reads without any error was 0% [Quail, 2012 #528]. The accuracy of pacbio SMRT sequence reads is the least (~85%) among the sequencing platforms (reviewed in [Kumar, 2012 #2961]).

## HIV-1 Drug resistance Genotyping in the era of high throughput sequencing (HTS)

As conventional Sanger-based genotyping is unable to characterize the HIV viral quasispecies at less than 20% prevalence [Gunthard, 1998 #2965;Van Laethem, 1999 #2966], a true HIV diversity cannot be ascertained [Korn, 2003 #1685;Schuurman, 2002 #1637]. Beside that the method is also expensive and labour intensive. An alternative low-cost genotyping method is required that has the ability to sequence the HIV population to “deeper” level and characterize the overall spectrum of true viral diversity in the viral quasispecies.

Different approaches like sensitive real time PCR drug resistant test [Johnson, 2008 #387], allele-specific RT-PCR [Palmer, 2006 #1608] and single genome sequencing [Palmer, 2005 #1118] were developed to genotype HIV variants in the viral quasispeices to infer drug resistant mutations but were highly expensive and difficult to implement as a HIV drug resistant diagnostic tool.

Roche/454 Ultra Deep pyrosequencing (UDPS) technology is able to generate up to 10,00,000 sequence reads of up to 1,000 base pairs read length per sequencing run (www.454.com). Such a sequencing profile means that an unprecedented range of viral variants including the drug resistant minor variants can be explored in the HIV quasispecies of a infected individual [Bimber, 2010 #1684;Wang, 2007 #1103;Jabara, 2011 #1188;Hoffmann, 2007 #1141]. For example, Wang and colleagues identified 58 viral variants per sample in average using UDPS, while only eight viral variants were identified using conventional Sanger sequencing method [Wang, 2007 #1103]. In a study by Le tal, the conventional Sanger genotyping method was unable to identify a massive 95% of mutations that were detected by UDPS method [Le, 2009 #1546]. In another similar study, Hoffmann and colleagues identified four additional minor drug resistant mutations with UDPS [Hoffmann, 2007 #1141]. In addition to this, Johnson et al revealed that minor HIV variants were present in treatment naïve individuals using UDPS method and that they were associated with reduced efficacy of the drug cocktails in the treatment [Johnson, 2008 #387]. Le et al suggest that the low abundance drug resistant HIV variants, observed using UDPS, provide the information on drugs involved in historical antiretroviral therapy [Le, 2009 #1686]. In a study by Simen et. al, UDPS revealed 28% of the treatment naïve individuals exhibited HIV variants with NNRTI resistant mutations that correlated with treatment failure while only 14% individuals had HIV variants with NNRTI resistant mutations as revealed by conventional Sanger method [Simen, 2009 #383].

UDPS has been used to explore the dynamics of the HIV quasispecies using longitudinal samples collected before and after treatment [Hedskog, 2010 #386]. A study by Hedskog et al [Hedskog, 2010 #386] showed that the prevalence of drug resistant variants was high and wild variants was low when the HIV infected individuals were in treatment. After treatment interruption, the author also detected drug sensitive HIV variants that were not present before treatment suggesting that the sensitive variants emerged out from drug resistant variants [Hedskog, 2010 #386].

These evidences suggest that UDPS can be used as a clinical tool for HIV drug resistance genotyping of wide range of viral variants including the minor variants of prevalence 1% or below [Gibson, 2014 #1735]. Although wang et al [Wang, 2007 #1103] observed that UDPS genotyping had on average 0.0098% error rate (six times more errors in homopolymeric region than in non-homopolymeric region [Wang, 2007 #1103][Brodin, 2013 #1687]), the authors observed 72 and 392 HIV variants at prevalence greater or equal to 20% and less than 20% respectively after authenticating the variants at P-value less than 0.001. In addition to this, UDPS using Roche/454 Junior system allows at least 48 samples to be genotyped in a single run (four times the number of samples sequencing conventional Sanger genotyping method), thus, enabling low-cost drug resistance genotyping per sample in low and middle income countries like sub-Saharan African countries [Dudley, 2012 #1209]. Each sequence of a sample is tagged with a specific multiplex identifier (MID) sequence that can be used to pool the sequence reads from a sample together [Hamady, 2008 #1638]. Dudley et al showed that the cost of Roche/454 UDPS genotyping was pushed down to $20 per sample that was up to five times cheaper than conventional Sanger genotyping method [Dudley, 2012 #1209].

The genotyping step in UDPS method is precedented by PCR DNA amplification that could introduces errors such as DNA recombination [Kanagawa, 2003 #1179;Meyerhans, 1990 #1181;Yang, 1996 #1182], DNA synthesis errors [Hughes, 2003 #1178;Mansky, 1995 #18] and DNA re-sampling errors [Liu, 1996 #1183]. These errors add artificial variation in the HIV-1 population, confounding the real ones.

Furthermore, the absence of a terminal signal at every sequencing cycle of DNA synthesis in Roche/454 pyrosequencing method adds series of similar bases at homopolymer site and the light intensity required for quantification of total bases added become smaller with increasing homopolymer length [Margulies, 2005 #1135;Shendure, 2008 #168]. Thus, Roche/454 pyrosequencd genotypic data has insertion/deletion errors high at homopolymer regions [Huse, 2007 #107][Wang, 2007 #1103]. HIV drug resistant mutations (DRMs) are present at a number of homopolymer regions within the HIV genome (**Figure 1.15**). Studies have shown that a drug resistant mutations at homopolymer region like codon positions 65 (K65R) and 103 (K103N) are present at high frequency during virologic failure [Geretti, 2009 #2964][Hosseinipour, 2009 #2962][Brenner, 2006 #1688;Doualla-Bell, 2006 #1690].

Sequencing a HIV variant multiple times and generating a consensus sequence from the reads could generate a highly accurate genotypic data of the variant. A unique identifier can be annealed to cDNA of a HIV genome, which gets copied to subsequent amplicons of the cDNA in PCR amplification. Each individual sequence read from UDPS might contain an error, however, the consensus sequence of those reads with same unique identifier would represent their template viral sequence. Based on this method, a Primer ID technology has been introduced for accurate sampling and genotyping of HIV variants from the viral quasispecies [Jabara, 2011 #1188]. With the application of the Primer ID technology, Jabara et al [Jabara, 2011 #1188] were able to to resolve 80% of the unique sequence polymorphisms that were different than consensus sequence from conventional Sanger genotype method.

## Thesis Outline

Chapter 2: The chapter introduces and describes, in detail, a novel algorithm QTrim for quality trimming of Roche/454 ultra-deep high throughput sequence data. QTrim is optimized for both poor and high quality data. HIV-1 resistance test requires high quality genotypic data and QTrim is designed to quality control the data for the test. 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 impact of the technology 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 antiretroviral 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.

Bibliography