# Literature Review

## CHAPTER 1

### Global epidemiology Report of HIV

United Nations Acquired Immune Deficiency Syndrome (UNAIDS) global report 2012 approximates 34 million people have been infected by Human Immune deficiency Virus (HIV) by the end of 2011 [UNAIDS 2012]. Though the report shows that the global trend of new HIV infections (figure 2) and HIV-related deaths (figure 3) per year is declining, the current number of HIV infections is the highest since 1990 (figure 1). There is a huge difference in the HIV prevalence by age and sex, among the countries around the global. Figure 1 shows the distribution of HIV prevalence by regions all over the world. Sub-Saharan Africa region is the most HIV aggravated region; there are 23.5 million HIV positive people. UNAIDS estimates that approximately 1 in every 20 adults are HIV infected in this region (UNAIDS). 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 (figure 5). The next severely affected regions are Caribbean and Eastern Europe, Central Asia, South Asia, South-East Asia and East Asia. UNAIDS targets to decline the HIV transmission by 50%, provide anti-retroviral drugs to 15 million HIV positive people and increase the annual global investment to US $ 24 billion in low and middle-income countries by 2015.

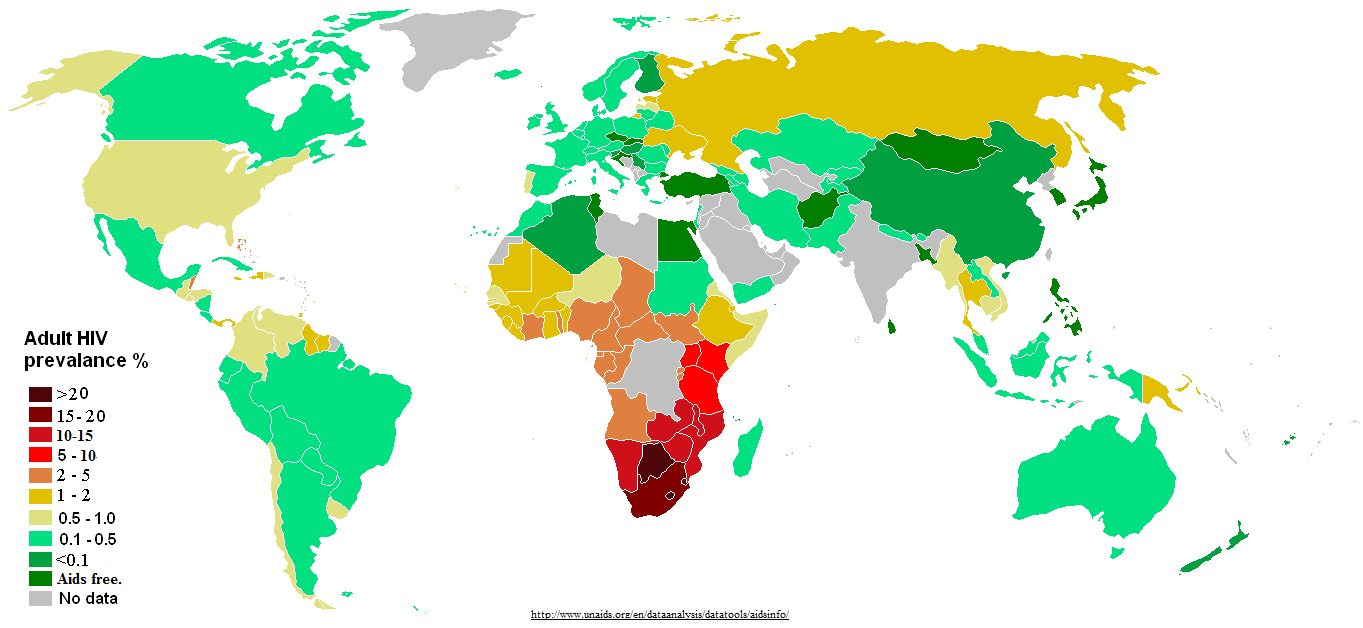


Figure 1: Distribution of HIV prevalence in adults around the world.

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

### Introduction to HIV

HIV is a retrovirus, which has two strands of Ribo Nucleic Acid (RNA) as the genetic material. The most common modes of the virus infection are sex transmission, intravenous drug usage, transfusion of infected blood and mother to child transmission at birth. HIV enters CD4+ expressing immune cells like T-lymphocytes and, very quickly, replicates in it. The immune cells burst, releasing thousands of new HIV viruses; each one capable of targeting new immune cell to begin life cycle again. The HIV patient’s immune system is disabled to defend foreign particles when the virus destroys huge amount of immune cells. Opportunistic infections by other diseases take the chance of week immune system to attack the human body and eventually kill the patient.

### Origin and Evolution of HIV

In 1981, the first patient of HIV AIDS was identified in the USA. In 1983, the virus was first isolated at “Institut Pasteur” in Paris (Alvine, Kien). Researches on the causative agent of the disease – HIV, implied that it is a retrovirus. Exploration of the retrovirus led researchers to find similar type of retrovirus in non-human primates that were then called as Simian Immunodeficiency Virus (SIV). Researches have established that HIV in human was a result of zoonotic transmission of SIV from non-human African wild primates (Bailes et al. 2002). About 40 different primates, in Africa, are infected with SIV and some are harboring some than one strain of SIVs. Phylogenetic analysis of SIVs from African non-human primates and two HIVs (HIV -1 and HIV -2) in human provided remarkable understanding of viral transmission and evolution from non-human primates to human.

Researchers have now established that HIV-1 and HIV-2 transmission to human are independent and their source are different. Discovered in 1986 AD, HIV-2 is transmitted from Sooty mangabey monkeys (Cercocebus atys) and its prevalence is also high in the geographical location of these monkeys in West Africa (Santiago et al., 2005). Sooty mangabey monkeys are naturally infected by a strain of SIV that is very close to HIV-2 (Hirsch et al. 1989) and the phylogenetic analysis of all HIV-2 strains show that it closely groups with the SIVsmm strains (Geo et al 1992, 1994) (Figure 2). SIVsmm does no harm to its host monkeys and must have modified to produce multiple strain and subsequent multiple zoonotic transmissions from sooty mangabey monkeys to human (Hahn et al 2000). Although HIV-2 subtypes A to G are identified in human, it is assumed that more subtypes were introduced into human (Gurtler 2004) but are lost for low adaptation fitness (Damond et.al 2004).

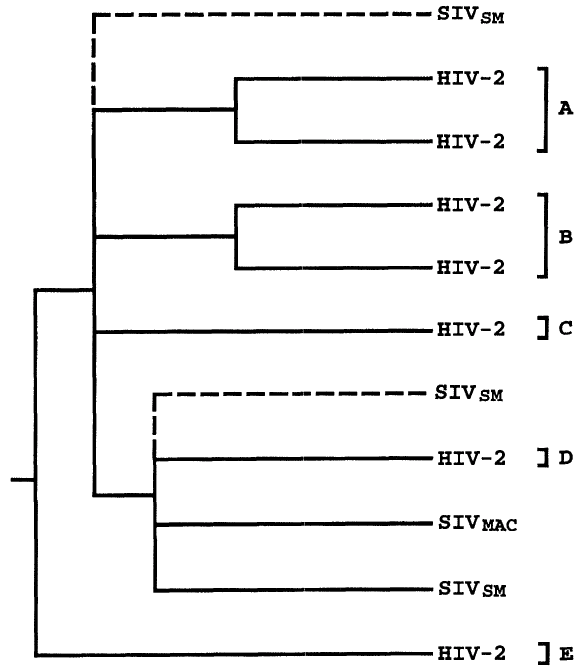


Figure 2: Phylogenetic tree showing close relationship between HIV-2 and SIV from Sooty mangabey monkeys. Source: Adapted from Gao et al 1994

Initial researches show that chimpanzees (Pan troglodytes) are the source of HIV-1 infection to human (Peeters et al. 1989; Huet et al. 1990) but the lack of enough evidence from vast number of other chimpanzees being test showed negative results and then subsequently discarded the idea (Vanden Haesevelde et al. 1996). Years later, in 1999, another chimpanzee was tested positive for SIV close to HIV-1. Four chimpanzee subspecies were analyzed with mitochondrial DNA (Groves 2001). The subspecies were: western (Pan troglodytes verus), Nigerian (Pan t. ellioti), central (P. t. troglodytes) and eastern (Pan t. schweinfurthii) chimpanzees (Gagneux et al. 1999). Retrospective research showed that all chimpanzees that are tested negative were subspecies *P.t versus* (Prince et al. 2002; Switzer et al. 2005) and those tested positive were *P.t troglodytes or P.t. schweinfurthii* (Geo et.al 1999, Corbet et al 2000, Santiago et al. 2003; Worobey et al. 2004; Keele et al. 2006, 2009; Van Heuverswyn et al. 2007). The phylogenetic analysis from these studies and further evidences from faecal samples (Keele et al 2006) confirmed the source of HIV-1 as *Pan troglodytes* (Figure 3).

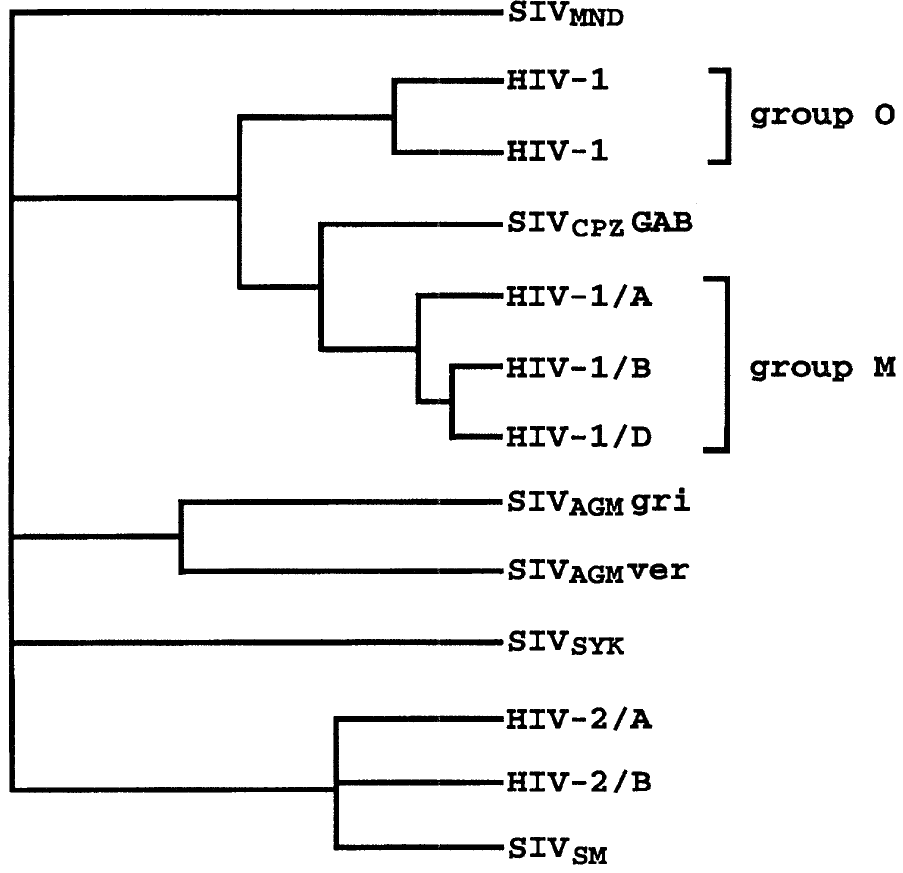


Figure 3: Phylogenetic tree showing HIV-1 evolving from Chimpanzee SIV.

Source: sharp et al 1995

### Structure of HIV

The structure of HIV is shown in Figure 4. HIV is spherical in shape and the size is approximately 0.0001mm or 1/12th the size of *Escherichia coli*. The viral structure can be divided into outer viral envelope and the inner viral core.

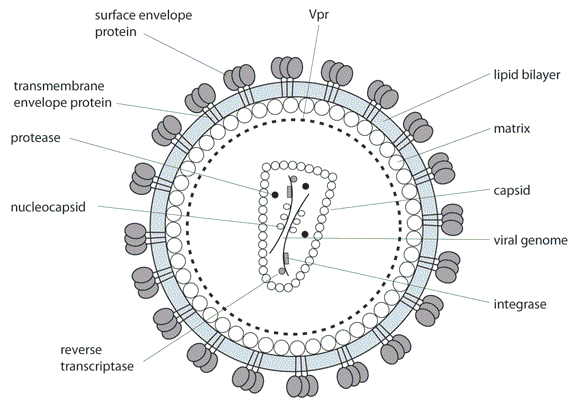


Figure 4: Structure of HIV

The viral envelope consists of the outermost lipid bilayer membrane, derived from the host immune cell membrane during budding out from the cell. HIV has genetic code for all its structure parts, but the lipid bilayer. Several host protein molecules are embedded on the membrane. Besides these proteins, the viral protein “Env” are also present on the surface. Env proteins form spike like structure that emerges out from the viral lipid membrane. Env consists of three glycoprotein-120 (gp-120) that make trimeric structure with three gp41 stems that anchor the viral envelope (Zhu et al. 2006, Subramaniam et al. 2006, Zanetti et al. 2006, Zhu et al. 2008).

The inner core consists of matrix, capsid (p24 viral proteins) and nucleocapsid. Inside the capsid, there are two single strands of viral RNA molecules that codes for all the viral proteins and structures.

### HIV genome structure and organization

The length of HIV-1 HXB2 (HIV reference sequence) genome is about 9719 bps and HIV-2 BEN genome is about 10359 bps (figure 5). Overall, HIV genome produces 15 proteins. The 5’ and 3’ ends consist of long terminal repeats (LTR) or also known as Untranslated Terminal Repeats (UTR). The 5’ – gag – pol – env – 3 ’ are the three genes that code for most of the structural proteins and necessary enzymes for viral replication. The structural elements and genes of HIV genome and their function or product are described below as they appear from 5’ to 3’.

**LTR or UTR**: The long terminal repeat is the most conserved region in the HIV -1 genome [Pillart et al 2004, Kuiken et al 2009]. It promotes viral packing before budding out, regulates and promotes transcriptional activation, binds primer for reverse transcription and dimerizes two single strands of RNA [Lu et al 2011]. It has splice donor site contributing to the packing of the full length viral RNAs [Mann et al 1985, McBridge et al 1996]. It interacts with the nucleocapsid domains to incorporate the full-length genome during the assembly of new virions [Berkhout 1996]. Researchers have proposed that the 5’-UTR adopts multiple secondary structures and

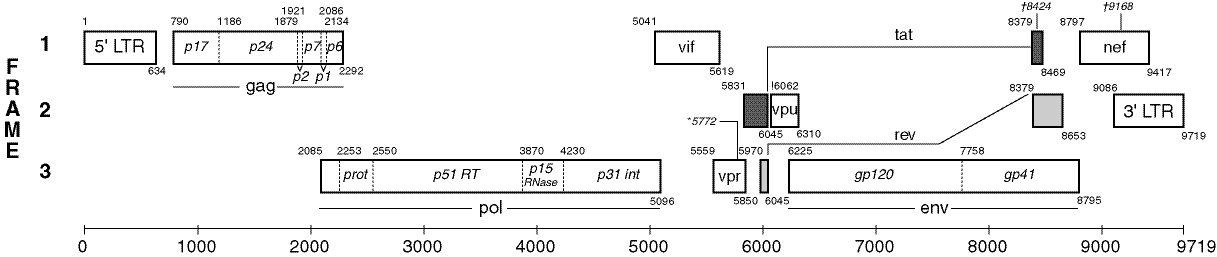


Figure 5: The landscape of HXB2 genome. Source: www.hiv.lanl.gov

regulates multiple RNA-dependent functions during viral replication. The nucleotides from 105-115 of 5’-UTR region can bind to oligoribonucleotides having *gag* start codon (AUG) to disrupt pre-existing stem loop and form a heteroduplex [Spriggs 2008].

**GAG**: The gag region spans from 790 to 2292 bp numbering in HXB2 genome at reading frame 1. The translational product of Gag is a precursor Pr55 that has all the building blocks to form a fully infectious virion, even in the absence of other viral products [Wang et al 1993]. The viral protease enzyme cleaves the gag precursor to yield the necessary structural proteins of mature virion [Gheysen et al 1989, Hunter 1994, Coffin, Swanstrom and Wills 1997, Freed 1998,]. The protease enzyme is packed in to nascent virions as a component of the Gag-pol polyprotein, which is, later, produced by ribosomal frameshifting between the overlapping gag and pol genes [Lillehoj et al 1988, Shioda et al 1990, Schneider et al 1997]. Protease cleaves gag precursor forming three structural proteins: matrix, capsid and nucleocapsid. The matrix protein is always at the N-terminal and p6 at the C-terminal of gag precursor with two spacer regions –p1 and p2 (Figure 5) which separate capsid from nucleocapsid and nucleocapsid from p6 respectively [Mervis et al 1988, Henderson et al 1992].

**POL**: The pol region spans from 2085 to 5096 bps numbering in HXB2 genome at reading frame 3 (figure 8). Pol produces the most necessary enzymes - protease, reverse transcriptase and integrase for viral replication. These enzymes are the targets for drug development against HIV. Many drugs have been developed to interfere the functions of protease and reverse transcriptase. The protease enzyme processes the gag-pol precursor polyprotein to produce these enzymes. The protease enzyme cleaves the gag polyprotein to form the viral structure proteins. The reverse transcriptase enzyme produces the cDNA from the viral RNA. The integrase enzyme integrates the proviral cDNA into the host genome.

**ENV**: Env spans from 6225 to 8795 bps numbering in HXB2 at reading frame 3. It produces viral glycoprotein as a precursor (gp160), which is processed to produce glycoprotein 120 (gp120) and glycoprotein 41 (gp41). Gp120 is a non-covalent complex of external protein and gp41 is a trans-membrane protein. They are organized to form trimeric complexes on the surface of HIV and mediate the HIV entry into the host cell [Allan et al 1985, Robey et al. 1985, Veronese et al 1985]. The exposed external complex gp120 binds to the CD4 receptor on the host immune cell. This triggers the conformational change in it and binds to one of the two chemokine receptors, CCR5 or CXCR4, to fuse with the host cell [Dalgleish eta al 1985, Feng et al 1996, Deng et al 1996, Choe et al 1996, Dragic et al 1996].

**TAT**: Tat is a trans-activating factor for HIV gene expression. After infection of host cell, HIV proviral genome is integrated in to the host genome, which is regulated by cellular as well as the viral transcription regulatory factors. Tat is the primary transcriptional regulatory factor. An example of Tat action is the control of RNA polymerase II elongation during transcription. In the absence of Tat, polymerase II disengages from the template DNA strand, terminating the transcription prematurely (Kao *et al*., 1987; Kessler & Mathews, 1992; Ratnasabapathy *et al*., 1990; Toohey & Jones, 1989). Sodroski *et al*. (1985) first explained the function of Tat. There are two forms of Tat – Tat-1 and Tat-2. Tat-1 is a minor form, which is 72 amino acids long; Tat-2 is the major form, which is 86 amino acids long. Tat is found at least at the lower level in the infected host cell, and located in the nucleus.

**REV**: Rev is another trans-activating factor for HIV gene expression. It is 19-kD phosphoprotein. Like Tat, it is also mainly localized in the nucleus of host cell, but cycles rapidly between the nucleus and cytoplasm as it promotes nuclear export of the transcriptional products. Rev binds at the Rev Responsive Element (RRE), which is an RNA element encoded within the env region of the virus.

**Virion Infectivity Factor (VIF)**: This gene is encoded at region 5041 to 5619 bp numbered in HXB2 at reading frame 1. It produces a protein of 220 amino acids and 23 kiloDalton. This protein promotes the viral infectivity to the host, but has no role in viral production. The protein is produced in the late stage of viral production [shreehy et al. 2002, Von Schwedler et al 1993] to suppress the innate antiviral activity of human immune cells [simon et al. 1998, Madani et al 1998]. Researcher show that human APOBEC3 family members – APOBEC3G and APOBEC3F are the two potent cytidine deaminases that attributes to introduction of excessive G → A mutations in the minus strand reverse transcripts [Harris et al 2003, Mangeat et al 2003, Lecossier et al 2003, Zhang et al 2003]. Vif protein prevents APOBEC3 members from hypermutating HIV reverse transcripts by inducing proteasomal degradation and exclusion from virions through recruitment of a cullin5 ECS E3 ubiquitin ligase complex [Conticello et al 2003, Marin et al 2003, Sheehy et al 2003, Stopak et al 2003, Mariani et al 2003, Mehle et al 2004, Wiegand et al 2004].

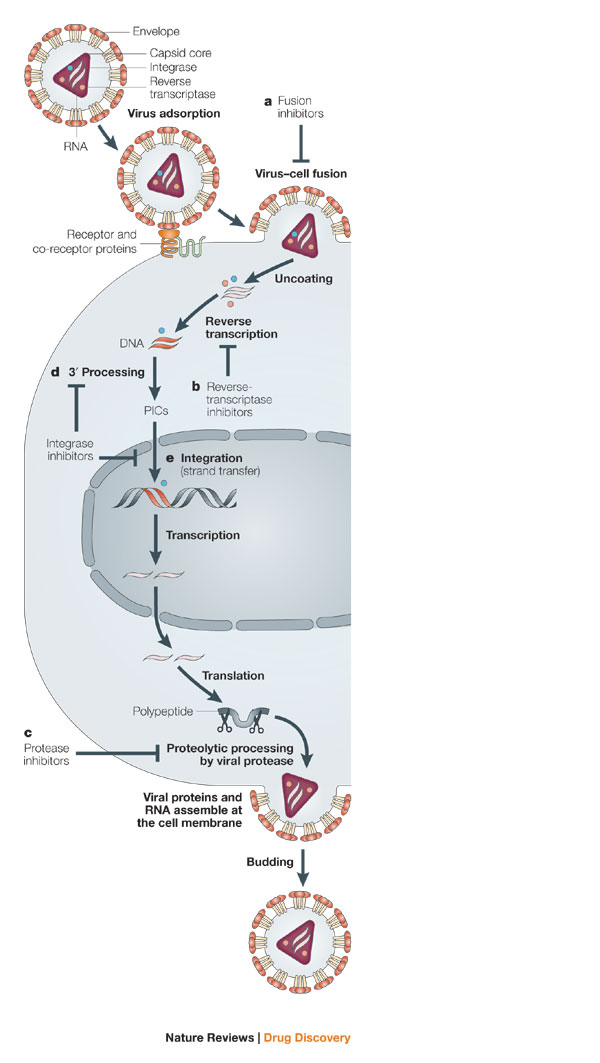
**Viral Protein R or VPR:** It is 96 amino acids long protein (14 kiloDaltons). This protein is packed into the nascent virions during budding out. Vpr plays role in entry of viral core into the host cell nucleus and thus localizes in the host cell nucleus after infecting the cell. Vpr arrests the cell cycle during the transfer from G2 to M phase [Jowett et al 1995, Rogel et all 1995] by preventing the activation of the p34cdc2/cyclin B complex [He et al 1995]. One research shows that vpr is also important for efficient viral replication in monocyte or macrophage cells, but in T-cells [Connor et al 1995].

**VPU or Viral Protein Unit:** This is the unique protein to HIV-1 and some other closely related simian immunodeficiency viruses (SIVcpz, SIV-GSN, SIV-MUS, Siv-MON, SIV-DEN). It is a 16 kiloDalton, 81 amino acids long protein. The Env precursor protein, gp160, makes stable intracellular complex with host cell CD4, reducing the CD4 expression at cell surface and processing of gp160 to gp120 and gp41. HIV protein vpu prevents the formation of Env-CD4 complex; increasing the gp160 processing but rapidly degrades CD4 in endoplasmic reticulum [Willey et al 1992]. Researchers have showed that vpu action of degrading CD4 could be blocked with peptide aldehyde or lactacystin, by interfering proteasome function of the protein [Schubert et al 1998].

**NEF:** This gene was formerly named F, 3’-ORF, or B-ORF. This accessory regulatory protein does not help the virus to increase infectivity but plays a role during the biogenesis of viral particles [Laguette et al 2009] and virulence [Piguet et al 1999, Prince et al 2002, Kirchhoff et al 2008]. The function of Nef is that it down regulates the production of major histocompatibility complex type 1 (MHC type 1) in the host cell [Schwartz et al 1996, Collin et al 1998, Cohen et al 1999]. This impairs the function of cytotoxic T lymphocyte cells to clear the infected cells [Collins et al 1998, Tomiyama et al 2002, Yang, et al 2002, Adnan et al 2006]. Nef also down regulates CD4 on host cell surface [Garcia et al 1991, Lama et al 1999] and modulates cellular activation [Baur et al 1994, Sawai et al 1994, Bodeus et al 1995, Saksela et al 1995, Smith et al 1996]. This shows that Nef plays role in host immune evasion.

### HIV life cycle

The life cycle of HIV starts from entry into the host immune cell to budding out nascent HIV virions for further infection to new cells or to new host. The mechanism of the viral entry into the immune cell involves HIV gp120 binding to the CD4 receptor of the host cell [Dalgeish et al 1984, Maddon et al 1986, McDougal et al 1986]. This binding brings conformational changes in gp120/gp41 complex [Sullivan et al 1998], to enable binding to a chemokine receptor either CCR5 or CXCR4 [Sattentau et al 1991, Moore et al 1992, Sattentau and Moore 1993, Berger et al 1999]. These binding assist the viral membrane fusion with the host cell membrane.

Researches show that a seven –trans membrane protein domain in T-cells fusion [Berson et al 1996] and CC CKR5 in Macrophages cells [Alkhatib et al 1996] serves as an accessory fusin factor for the viral fusion.

FFigure 6: The pathway of the HIV life cycle. Important events labeled a – e are the possible target points for disrupting HIV life cycle with development of antiviral drugs. Source: Pommier, Yves, Allison A. Johnson, and Christophe Marchand. “Integrase Inhibitors to Treat HIV/Aids.” *Nature Reviews Drug Discovery* 4, no. 3 (March 2005): 236–248. doi:10.1038/nrd1660.

HIV fusion to the host cell and successfully release genome and integrate it to cell genome has always been the subject of interest. The technology of live cell imaging has substantially improved our understanding on the viral process of internalization, its steps of intracellular pathways through the dense cell cytoplasm to reach the nucleus and release the genome [Sun et al 2013]. Unlike most other viruses, HIV has pH-independent gp41 fusion protein [Wilen et al 2012], forms fusion pore [Melikyan et al 2008] and directly fuse with the plasma membrane [Marsh et al 2006, Wilen et al 2012]. The fusion process is initiated after coreceptor binding induction and subsequent exposure of hydrophobic gp41 and insertion of it into the host cell membrane [Wilen et al 2012].

Upon fusion, there are many events occurring in the journey of HIV to the cell nucleus. From cell entry to nucleus, HIV loses some of its proteins, exploits cellular factors, reverse transcribe RNA genome to DNA preintegration form and ultimately integrate it into the cell genome [Cullen et al 2001]. McDonald et al [2002] were able to track the HIV virion pathway to the nucleus using GFP-labeled particles. The researched showed that HIV uses cytoplasmic dynein and the microtubules network to accumulate at the perinuclear region, often near the microtubules-organizing center. The viral movement is curvilinear in the cytoplasm. The same study showed that the reverse transcription process starts in the intact capsid and the required deoxynucleotides are imported to build up the DNA. The capsid proteins dissociates but the reverse transcription complex remains intact interaction with microtubules, on the way to microtubules organizing center, close to nucleus. Ultimately, the reverse transcription complex reach close to the nucleus; enters through nuclear pore and the integrase enzyme integrates the preintegration DNA into the host DNA [McDonald et al 2002].

HIV uses the cell mechanism of gene expression to transcribe the viral genes. HIV gene Tat promotes the transcription of the viral DNA [Ott et al 2011, Razooky and Weinberger 2011]. HIV viral production from the infected cell can be broadly explained in three steps: assembly, budding and maturation.

Upon production of all necessary viral particles, initiation of HIV virion assembly occurs at the plasma membrane. The Gag polyprotein mediates all the essential events in the assembly, including the binding to the plasma membrane. It also makes protein-protein interactions to create spherical particles at the membrane, concentrates the viral Env protein and packages two copies of capped and polyadenlylated full length viral genomic RNA [Johnson et al 22010]. The complete assembled virion particles are released from the plasma membrane by the host Endosomal Sorting Complexes Required for Transport (ESCRT) machinery [Morita and Sundquist 2004, Bieniasz 2009, Carlton et al 2009, Usami et al 2009, Hurley and Hanson 2010, Peel et al 2011]. The maturation of the nascent HIV virions begin concomitantly with budding. Maturation is driven by the viral PR enzyme cleavage of the Gag and Gag-Pro-Pol polyproteins. The processed polyproteins produces matrix, capsid, nucleocapsid, p6, protease, reverse transcriptase and integrase proteins [Hill et al 2005]. These processed proteins get rearranged to create the mature infectious virion, ready to next infection to start new life cycle [Mariani et al 2003].

### HIV types and subtypes, Subsubtypes and Circulating Recombinant Forms

HIV is hugely diverse virus. Its diversity is obtained from phylogenetic analysis of genomic region to genome wide sequence. It creates its diversity to adapt in different environment like host immune system and drug pressure.

Classification of HIV by group, subtype and subsubtype require a reference sequence with the criteria: a sequenced full-length genome, no recombination history, HIV genome sequence published in peer reviewed citation, isolated from recent samples, HIV is covered in major geographical distribution, HIV has no sign of hypermutation, HIV genome sequence is real sequence from a patient, HIV genome has no extreme indels, and the virus must be viable and intact (Leitner et al 2005).

There are two distinct types of HIV: HIV-1 and HIV-2. These viruses can be differentiated by their genome organization and phylogenetic relationship (Hahn et al 2000), pathogenesis, transmissibility and pattern of spread (De cock et al 1993, Kanki and De cock, 1994). A notable difference is the source of the infection to human. Evidences show that HIV-1 is zoonotic transmission to human from chimpanzee *(Pan troglodytes*) and HIV-2 from sooty mangabey (*Cercocebus atys*). The genome wide sequence of both HIV show that HIV-2 has an extra gene “vpx” which is lacking in HIV-1 genome (Henderson et al 1988, Kappes et al 1988, Tristem et al 1992, Bergamaschi et al 2009). Drug resistance clinical results show that non-nucleoside reverse transcriptase inhibitors (NNRTIs) anti-retroviral drugs are effective against HIV-1, but non-effective against HIV-2 reverse transcriptase (Hizi et al 1993).

Isolation, characterization and sequence analysis of each type of HIV show that there exist genomic heterogeneity and variability among different isolates (Benn et al 1985, Hahn et al 1985, Wong-Stall et al 1985). Analysis of isolates within a patient and between patients demonstrates that intra-patient isolates are more related than inter-patient isolates (Hahn et al 1986). HIV-1 isolates from around the world and their phylogenetic sequence analysis reveals distinct subtypes that cluster together to form three groups: M (Main), O (Outlier) and N (non-M and non-O) (Robertson et al 2000, Leitner et al 2005), which represent three independent zoonotic transfection of SIV from chimpanzee to human (Hahn et al 2000, Sharp et al 2001).

Within each group, phylogeny of the genomic region with most variability allows classification of the virus to subtypes. An unrooted phylogenetic analysis of the HIV-1 isolates from group M using complete gag, pol, env and nef sequences (Figure 7) or full length genome sequence analysis show distinct subtypes A – D, F – K (figure Buonaguro et al 2007). The phylogenetic clades of subtypes can be constructed from any part of the HIV-1 genome when the alignment is at least 300-500 bases long (Leitner et al 2005). Sequence length shorter than threshold from genomic regions under higher evolutionary pressure e.g. env V3 can reconstruct the phylogenetic clades distinguishing the subtypes, while regions under slower evolutionary change need long sequences to give reliable results (Leitner et al 2005). Further phylogenetic structure for subtype A and F have been identified leading to the classification of subsubtypes A1 and A2 for subtype A and F1 and F2 for subtype F (Gao et al 2001, Lietner et al 2005).

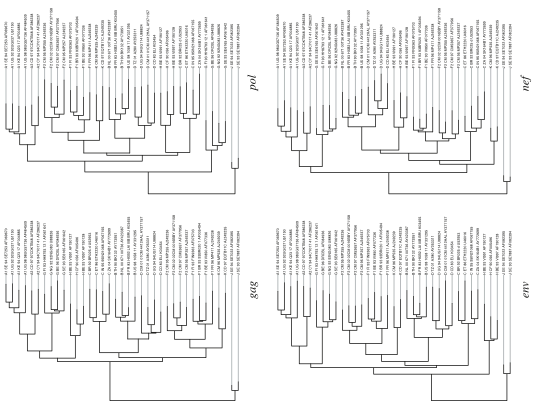


Figure 7: Phylogenetic tree showing HIV-1 Group M subtypes using gag, pol, env and nef sequences. Source: Leitner, Thomas, Bette Korber, Marcus Daniels, Charles Calef, and Brian Foley. “HIV-1 Subtype and Circulating Recombinant Form (CRF) Reference Sequences, 2005.” *HIV Sequence Compendium* 2005 (2005): 41–48.

Two or more HIV subtypes infecting a single patient create inter-subtype recombinant forms called “unique recombinant forms” (McCutchan 2006). Isolation and identification of unique recombinant form from at least three epidemiologically unlinked patients and characterized by full-length genome sequencing is designed as circulating recombinant forms (CRFs) (Los Alamos Laboratory [[http://www.hiv.lanl.gov/content/hiv-db/CRFs/CRFs.html](http://www.hiv.lanl.gov/content/hiv-db/CRFs/CRFs.html" \t "_blank)]). There are huge numbers of CRFs known; the most prevalent are CRF01\_AG, CRF02\_AG in West Africa and CRF01\_AE in Southeast Asia. Country specific prevalent CRF may be different e.g. CRF01\_AG in Nepal and India (Shahid et al 2011). All discovered CRFs are documented in [http://www.hiv.lanl.gov/content/hiv-db/CRFs/CRFs.html](http://www.hiv.lanl.gov/content/hiv-db/CRFs/CRFs.html" \t "_blank).

Unlike group M, group O is endemic and largely confined to certain geographical region of Cameroon and neighboring West Central African region; group O only represents a small minority of HIV-1 strain in the region (Peeters et al 1997, Jaffe and Schochetman 1998, Janssens et al 1999). Group N is also confined to Cameroon and is found in limited isolates only (Simon et al 1998).

HIV-2, first isolated from patients in West Africa, exhibits HIV-1 AIDS like symptoms but shows seronegative to HIV-1 assay (Clavel et al 1986, Clavel et al 1987, Leys et al 1990, Gao et al 1992). Both viruses exhibit significant similarity at genome level (guyader et al 1987).

HIV-2 is classified to epidemic subtypes (A, B) and non-epidemic subtypes (C, D, E, F) [Hahn et al 2000, Tebit et al 2007]. Researchers expect that HIV-2 subtypes are independent cross-species transmissions of SIV to human population, very much like HIV-1 groups [Hahn et al]. Aguchi et al [2000] suggested a new HIV-2 subtype G. Some researches are done on subtypes A and B (Gao et al 1994, Chen et al 1997, Damond et al 2001, Damond et al 2002, Pieniazek et al 2004, Ishikawa et al 2004, Tebit et al 2007) and very little is known about the other subtypes. Widely spread subtypes A and B have high chance co-infecting a single person. The first HIV-2 CRF reported was CRF01\_AB, which is estimated to be recombined as between 1964 to 1973 [Iba et al 2010]. The discovery of the CRF01\_AB outside of subtype A and B epidemic area shows that it is ongoing global spread [Iba et al 2010].

### Anti retroviral drugs and HIV treatment

HIV infection can be confirmed using a serological assays or antigen-antibody assays [Laperche et al 2012, Rosenberg et al 2012, Workman et al 2009, Schupbach et al 2006, Fox et al 2011, Fiscus et al 2007, Pilcher et al 2005]. Upon HIV infection, the virus replicates very quick, destroying the host immune cells; and the patient has to undergo HIV treatment with the only available method, which is anti retroviral therapy (ART) [Gilks et al 2006].

The decision to start ART should depend on clinical, immunological and viral assessment of HIV in adults and adolescents [Gilks et al 2006]. Different parameters are checked to decide the ART treatment of a patient: CD4 cell count, pregnancy, co-infections with other viruses (HBV, HCV) and bacteria (TB), plasma HIV RNA level, body weight etc. However, International AIDS Society-USA recommends ART initiation regardless of CD4 count, in resource limited settings without readily access to proper clinical facilities [Thomson et al 2010]. The scale-up of ART program in the past decade has emerged drug resistant HIV virus [Boden et al 1999] circulating in the human population. The test for drug resistant HIV needs to be prioritized for the selection of drug regiment in ART [Hirsch et al 2000, Grant et al 2002, Hirsch et al 2003, Bansi et al 2010]. A person starting ART must consider important factors like adherence to the therapy for life long, toxicity, emergence of HIV drug resistance and subsequent change of drug regimen, drug interactions, the risk of viral transmission to uninfected healthy people etc [Thompson et al 2010].

The initial ART drug regimen (first line therapy) has a combination of two Nonnucleoside Reverse Transcriptase Inhibitors (NRTIs) and one Nucleoside and Nucleotide Reverse Transcriptase Inhibitors (NNRTIs). Upon first line ART drug failure, second line ART is recommended, which consists of change in a drug from NRTIs class and addition of a drug from Protease Inhibitor drug class.

**Nucleotide and Nucleoside Reverse Transcriptase Inhibitors (NNRTIs)**

NNRTIs are highly HIV-1 reverse transcriptase selective drug, and ineffective against HIV-2 reverse transcriptase. Three-dimensional structural information of reverse transcriptase is the center for the development of NNRTIs drugs [Ren et al 1995, Sarafianos et al 2009]. The drugs in this class, bind at an allosteric site 10 [Å](http://en.wikipedia.org/wiki/%C3%85) from polymerase active site of the HIV-1 reverse transcriptase [Kohlstaedt et al 1992, Esnouf et al 1995, Ding et al 1995], close to substrate-binding site. The binding induces conformational change in the enzyme, which distorts the catalytic aspartate triad and inhibits the function of the enzyme to convert RNA to cDNA. [Esnouf et al 1995, Balzarini 2004]. NNRTIs drug are administered along with other drugs in ART at the concentration lower than required to affect human cells [De Clercq 1996] leading to non-toxicity. NNRTIs drugs from first generation have inherent disadvantageous characteristics like cross-resistance and low genetic barrier [Adams et al2010] as they easily select for mutant viral strain with several degrees of drug resistance [Balzarini 2004]. Low genetic barrier indicates that the virus with one or two mutations can easily escape the drug. High genetic barrier indicates many mutations are required for the virus to escape the drug. Nevirapine and delavirdine are first generation NNRTIs drugs that easily loose their inhibitory potential against mutant viral strains. This has limited the use of first generation NNRTIs in the future. Etravirine [MInuto and Haubrich 2009], Efavirenz [Ren et al 2000], capravirine [Gewurz et al 2004] and dapivirine [Fletcher et al 2009] are second generation NNRTIs with higher genetic barrier and requires a markedly longer time period to obtain significant resistance [Ghosn et al 2009Adams et al 2010].

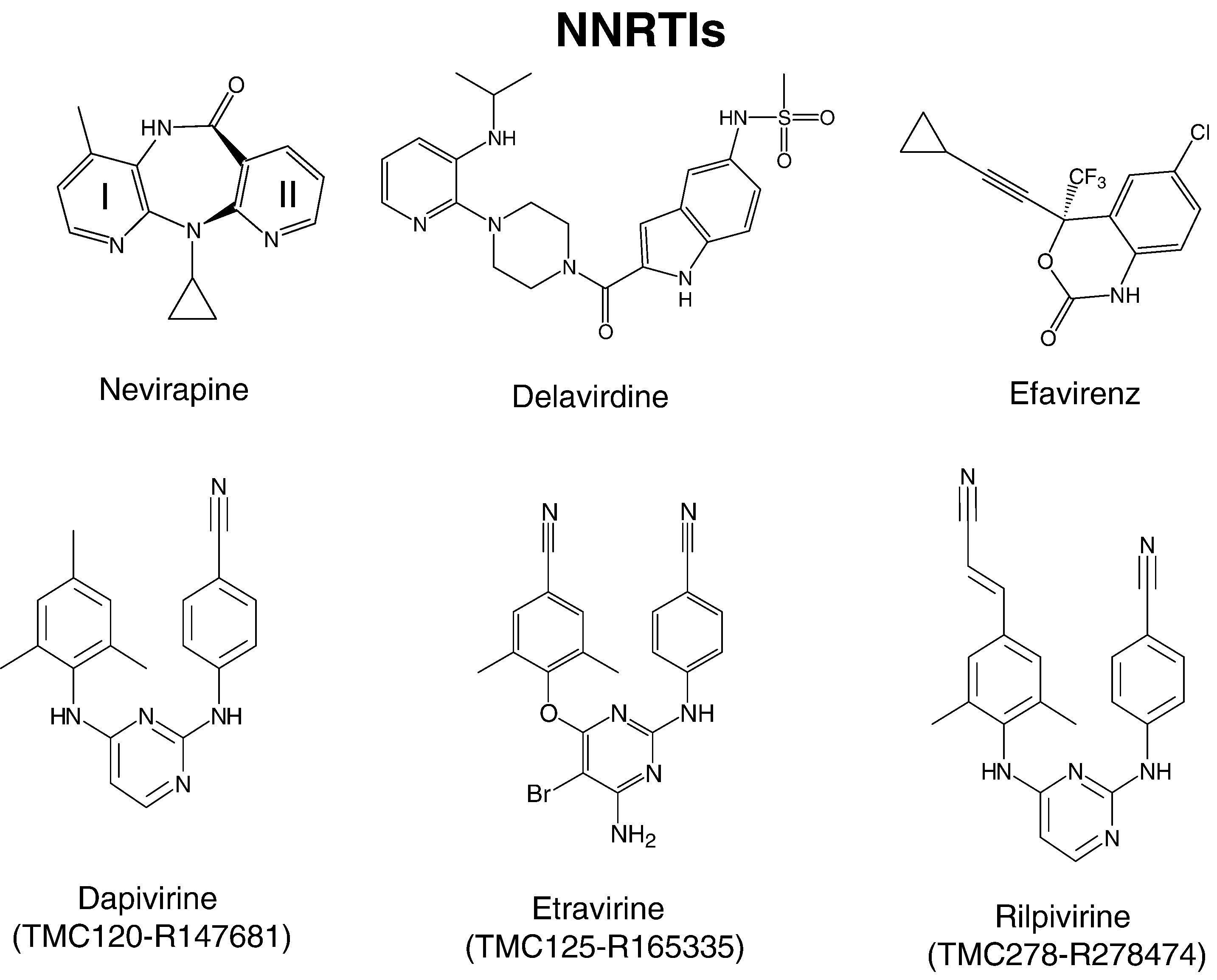


Figure: Chemical structures of NNRTI molecules. Source: Sarafianos, Stefan G., Bruno Marchand, Kalyan Das, Daniel M. Himmel, Michael A. Parniak, Stephen H. Hughes, and Eddy Arnold. “Structure and Function of HIV-1 Reverse Transcriptase: Molecular Mechanisms of Polymerization and Inhibition.” *Journal of Molecular Biology* 385, no. 3 (January 23, 2009): 693–713. doi:10.1016/j.jmb.2008.10.071.

**Nonnucleoside Reverse Transcriptase Inhibitors (NRTIs)**

Reverse Transcriptase is the major target enzyme for the development of anti-AIDS drugs [Sarafianos et al 2004]. The enzyme produces cDNA from the viral template RNA. Nucleoside Reverse Transcriptase Inhibitors (NRTIs) are analogs of normal nucleotides but without 3’hydroxyl group. The drug is taken as unphosphorylated form, which cytokinases phosphorylates to form 5’ triophosphates [De Clarcq 2002, Ilina and Parniak 2008]. It leads to incomplete termination of cDNA synthesis and are the major anti retroviral drugs in ART regimen available in clinics. Reverse transcriptase possesses two characteristics that provide resistance to NRTIs: hydrolytic excision of nucleotide (primer-unblocking) blocking DNA synthesis (Arion et al 1998, Meyer et al 1999) and dynamic copy choice that is switching of the template RNA [Svarovskaia et al 200, Nikolenko et al 2004] while generating cDNA. There exists equilibrium between the rates of DNA synthesis and RNA degradation, which determines the template switching. The presence of NRTIs drugs terminates the DNA synthesis chain, breaking the equilibrium and increasing template switching. RNaseH degrades the viral RNA once the DNA synthesis stops; this leads to dissociation of template RNA and synthesized strand, terminating HIV-1 replication [Nikolenko et al 2004]. Researchers [Nikolenko et al 2004] postulated that reverse transcriptase has property to excise an incorporated NRTI and resume the DNA synthesis and subsequent RNA degradation by RNaseH. For e.g. NRTI drug 3′-azido-3′-deoxythymidine (AZT) related mutation in reverse transcriptase increases the rate of AZT excision allowing the continuation of DNA synthesis and RNA degradation at the end. Mutation in RNaseH leads to delay in RNA degradation and reverse transcriptase gets more time to excise the incorporated NRTI and resume DNA synthesis. Mutation in RNaseH gives the virus resistance to AZT and Stavudine/ 2,3-didehydro-2,3-dideoxythymidine(d4T) drugs but not to EFV and 3TC [Nikolenko et al 2004].



Figure: Chemical structures of NRTI molecules. Source: Sarafianos, Stefan G., Bruno Marchand, Kalyan Das, Daniel M. Himmel, Michael A. Parniak, Stephen H. Hughes, and Eddy Arnold. “Structure and Function of HIV-1 Reverse Transcriptase: Molecular Mechanisms of Polymerization and Inhibition.” *Journal of Molecular Biology* 385, no. 3 (January 23, 2009): 693–713. doi:10.1016/j.jmb.2008.10.071.

**Protease Inhibitors (PR)**

HIV protease enzyme mediates the maturation of newly formed HIV particles by proteolytic cleavage of gag and gag-pol precursors [Kohl et al 1988]. Protease is a logical target for the treatment of HIV infection. Inhibition of the enzyme function could stop the formation of new virus progenies. The detail knowledge of protease function and its three-dimensional structure has opened numerous approaches to rational design of inhibitors. Protease inhibitors are designed to bind specifically to the virus protease with high affinity although they occupy more space than the natural substrates [Wensing et al 2010]. There are nine protease inhibitors approved for clinical use: amprenavir, lopinavir, atazanavir, tipranavir, darunavir, indinavir, nelfinavir, saquinavir and ritonavir (Figure 9). The chemical structures of them are shown in figure 8. Except tipranavir, all protease inhibitors are competitive peptidomimetic inhibitors. They have hydroxyethylene core, which prohibits cleavage activity of the HIV-1 protease [Craig et al 1991, Vacca et al 1994, Kempf et al 1995, Sham et al 1998, Koh et al 2003, Partaledise et al 1995, Robinson et al 2000]. Tipranavir has dihydropyrone ring instead of peptidomimetic hydroxyethylene core [Turner et a 1998] and can inhibit HIV isolates replication that are resistant to other protease inhibitors [Larder et al 2000, Back et al 2000].

The first generation protease inhibitors used are Saquinavir, Ritonavir, Indinavir and Nelfinavir (Figure 9). The limitation of first generation protease inhibitors was low bioavailability and short half-life, often required to multiple daily dose in-take. This reduced the patient adherence to the therapy, and subsequently increased drug failure [Bartlett et al 2001, 2006]. Two approaches were developed to fight the resistance selection by first generation inhibitors: first, increase the level of protease inhibitors in the plasma in combination with low-dose ritonavir and second, develop a novel protease inhibitor with high potency against known protease inhibitor resistant HIV variants [Wensing et al 2010]. The second generation of protease inhibitor therapy is the combination of first generation protease inhibitors with ritonavir. Ritonavir reduced the metabolism of first generation inhibitors, thus, increasing their level in the plasma [Zeldin and Petrus 2004, Youle 2007]. Patients with extensive drug resistant HIV have limited therapeutic options. They are put under double-boosted protease inhibitor therapy to increase the genetic barrier of HIV to prevent drug failure [Ribera et al 2006, Gilliam et al 2006, Chetchotisakd et al 2007, Petersen et al 2007, Von Hentig et al 2007, Monosuthi et al 2008, Stebbing et al 2009].

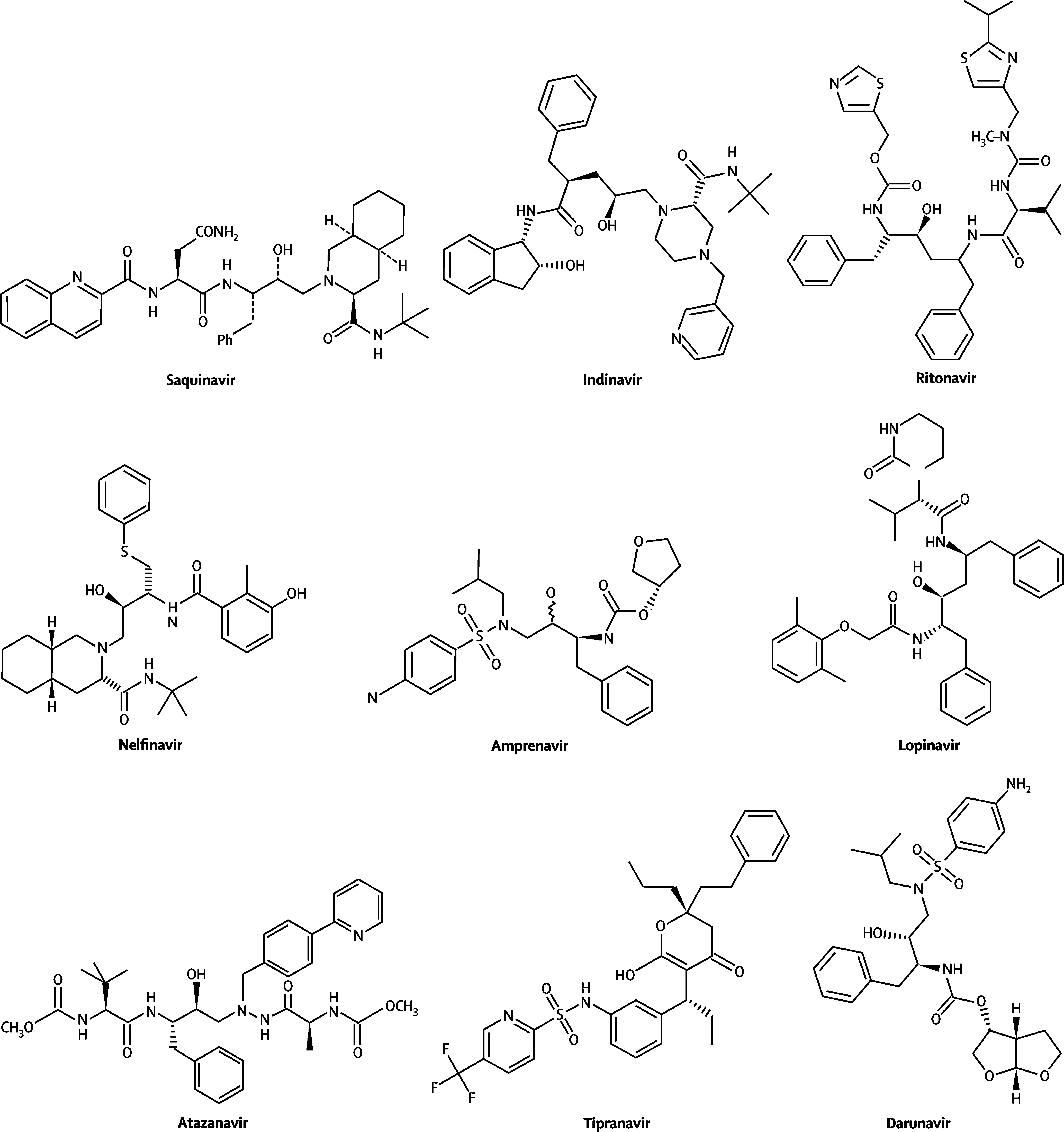


Figure 9: Molecular structures of Protease Inhibitors. Source: Chow, Warren A, Chunling Jiang, and Min Guan. “Anti-HIV Drugs for Cancer Therapeutics: Back to the Future?” *The Lancet Oncology* 10, no. 1 (January 2009): 61–71. doi:10.1016/S1470-2045(08)70334-6.

Incorporation of protease inhibitor in ART has resolves many AIDS defining illness. However, patients consuming the inhibitors have developed lipodystrophy and hyperlipidemia. This is the side effect of the inhibitor as, besides inhibiting protease enzyme, it also inhibits degradation and secretion of apolipoprotein B; but in the presence of oleic acid, protease inhibitor increases the secretion of apolipoprotein B [Liang et al 2001].

**Integrase Inhibitors**

HIV integrase catalyzes the function of integrating the reverse transcribed DNA to the host genome. The enzyme cleaves two nucleotides from 3’ end of the DNA, soon after the reverse transcription step [Sherman and Fyfe 1990, LaFemina et al 1991, Bushman and Craigie 1991]. The RT/viral DNA preintegration complex, formed after reverse transcription, is transported to the nucleus; integrase catalyzes the strand transfer process, in which it creates staggered nick in host chromosome and joins 3’ end of viral DNA of 5’ end of host chromosome [Grobler et al 2002].

Although the antiretroviral drug regimen, consisting of two NRTIs along with a NNRTIs or protease inhibitors (also called Highly Active Anti Retroviral Therapy - HAART) could suppress the viral replication, it is difficult for the patients to well-accept the drug regimen. It is expensive, leads to multiple drug resistance and requires high adherence [Cohen 2002]. This has diverted the search for drugs that target third enzyme, the integrase. Researchers have demonstrated the integrase inhibitors feasibility and efficacy in Rhesus Macaques [Hazuda et al 2004].

Integrase inhibitors usually targets the strand transfer complex of the enzyme bound to the viral DNA, with the 3'-end [dinucleotide](http://europepmc.org/abstract/MED/18565342/?whatizit_url_Chemicals=http://www.ebi.ac.uk/chebi/searchId.do?chebiId=CHEBI%3A47885" \t "_blank) already cleaved. X-ray structure of the integrase enzyme has revealed the active site model of the enzyme complexed with the DNA [Chen et al 2008]. This model has application for designing wide range of potential integrase inhibitors. Diketo acids (DKA) and its derivatives (like L-708,906, S-1360, and V-165 (Fikkert et al 2003)) are the highly potent integrase inhibitors for strand transfer inhibition [Hazuda et al 2000]. A very small concentration of DKA is required for effective integrase inhibitor. A drug 5-CITEP, derived from DKA, can inhibit integrase function at micromolecular concentration [Pluymers et al 2002]. Another compound, 4-Aryl-2, 4-dioxobutanoic acids, inhibits integrase activity at nanomolar concentration range [Wai et al 2000,]. Integrase inhibitor L-870810 showed high antiviral activity against different clinical isolates with multiple drug resistance ability and HIV from different subtypes [Hazuda et al 2004].



Figure 10: Chemical structure of different integrase inhibitors. Structures with similar chemical properties are grouped together. Source: Bera, Sibes, Krishan K. Pandey, Ajaykumar C. Vora, and Duane P. Grandgenett. “HIV-1 Integrase Strand Transfer Inhibitors Stabilize an Integrase–Single Blunt-Ended DNA Complex.” *Journal of Molecular Biology* 410, no. 5 (July 29, 2011): 831–846. doi:10.1016/j.jmb.2011.01.043.

Some integrase inhibitors with weak antiviral activity have been reported. Styrylquinolines inhibit 3′ processing capacity of integrase [Zouhiri et al 2000, Pommier et al 2005]. Another compound, thiazolothiazepine, also possesses weak antiviral activity [Neamati et al 1999, Pommier et al 2005].

Integrase inhibitors are added in the drug regimen for patients with viral quasispecies highly resistant to the most of the reverse transcriptase and protease inhibitors.

**Entry Inhibitors**

HIV gp120 binds to the CD4 receptor [Dalgleish et al 1984, Sattentau and Weiss 1988] and CCR5/CXCR4 co-receptors [Moore et al 2004] to enter into the host cells. HIV entry inhibitors are designed either to inhibit the binding of gp120 to CD4 or co-receptors CCR5 or CXCR4.

HIV gp120 binding to CD4 forms an unglycolylated cavity in the gp120. CD4 phenylalanine residue 43 is the only residue that binds to this cavity and provides significant energy of about 23% of the total energy of CD4-gp120 binding [Kwong et al 1998, Madani et al 2004]. This cavity has been the primary target for designing small molecules that could bind it and inhibit HIV entry [Kwong et al 1998, 2000, Wyatt et al 1998]. Some of the molecules that inhibit gp120-CD4 binding are: PRP-542 [Jacobson et al 2000], TNX-355 [Moore et al 1992, Kuritzkes et al 2004], CADA [Vermiere et al 2002, 2003], BMS-806 [Madani et al 2004, Lin et al 2003, Guo et al 2003].

Another target of entry inhibitor is the gp120 conformational change to bind to co-receptors CCR5 or CXCR4. For CCR5 using virus, gp120 recognizes the N-terminal domain or the second extracellular loop of CCR5 co receptor [Wu et al 1997, Dragic et al 1998, Zhang et al 2007]; whereas for CXCR4 using virus, the gp120 only recognizes the second extracellular loop [Picard et al 1997]. Unavailability of co receptors for gp120 binding prevents HIV entry into host cell. Co receptor antagonists are designed that binds specifically to targeted co receptors, making it unavailable to HIV gp120 binding.

Some of the CCR5 antagonists developed are TAK-779 [Baba et al 1999], TAK-652 [Baba et al 2005], vicriviroc [Strizki et al 2005, Schurmann et al 2007, Gulick et al 2007], AD101 (SCH-350581) [Tsamis et al 2003], Maravirok (UK-427857) [Rosario et al 2005, De Clercq 2005, Fatkenheuer et al 2005, Rosario et al 2006, Wheeler et al 2007], Aplaviroc (GW-873140) [Nichols et al 2007, Latinovic et al 2009], PRO-140 [Trkola et al 2001].

It has been observed that HIV is capable of co receptor switch from CCR5 to inclusion or exclusively CXCR4 [Esbjornsson et al 2010] in late-stage of disease progression. HIV can use CXCR4 co receptor as the next door to enter the host cell. Therefore, the development of CXCR4 antagonists is essential.

Some of the CXCR4 antagonists either for clinical use or in clinical trial stage are AMD3100 [Donzella et al 1998], AMD070 [Schols et al 2003, Stone et al 2007], KRH-1636 [Ichiyama et al 2003] and KRH-2731 [Murakami et al 2004].

HIV envelope protein gp41 is also a possible target for drug design. Three different strategies for develop anti HIV molecules have been reported based on gp41 coiled coil region structural information. In near future, drugs targeting gp41 might be available [Jiang and Debnath et al 2000].

**Fusion Inhibitors**

Binding of gp120-CD4 and subsequently to co receptor change the confirmation in the viral envelope that shifts the gp41 from a non-fusogenic to a fusogenic state. This change in state drives HIV and host cell fusion. The exposed gp41 N-terminal domain gets inserted into cellular membrane through fusion peptide. The heptad regions HR1 and HR2 changes the free energy associated with six-helix bundle to make it thermo stable and this change in free energy is necessary for the fusion pore formation [Weiss 2003, Briz et al 2006] and entry into the host cell.

The design of fusion inhibitors is based on the targeting the heptad regions HR1 or HR2. Binding of the inhibitor to either HR1 or HR2 of gp41 makes it unable to make fusion pore. Enfuvirtide [Duffalo and James 2003, Poveda et al 2005] is a synthetic peptide, approved for clinical use in 2003 (commercial name-Fuzeon) [Robertson 2003], which can bind to HR1 region of gp41 [Wild et al 1993]. T-1249 is a second-generation fusion inhibitor drug that target different HR1 region of gp41 than Enfuvirtide [Kilby and Eron 2003]. T-1249 is active against Enfuvirtide resistance HIV strains, HIV-2 and SIV. However, it is discontinued in 2004 for clinical use [Briz et al 2006, Melby et al 2007]. Sifuvirtide is another HIV fusion inhibitor peptide under research [Wang et al 2009].

### Structures of Protease and Reverse Transcriptase

Most of the antiviral drugs target either HIV protease and/or reverse transcriptase enzyme. The three dimensional structure of these proteins are vital for the development of inhibitor drugs against them.

**Structure of Protease**

HIV protease enzyme was crystallized and its three dimensional structure was determined to boost antiviral drug development targeting the enzyme [Navia et al 1989]. The structure of protease is stored in protein data bank with ID number [2HVP](http://pdb.org/pdb/explore/explore.do?structureId=2HVP). The enzyme is a homodimer of two identical 99 amino acid chains, each monomer containing active site conserved region of amino acid sequence Asp-Thr-Gly (position 25 to 27) (Fig). The conserved active site motifs are located in loops that approach the center of the dimers. The two dimers are connected with four stranded anti parallel beta sheets involving both C and N terminals of the subunits. Both monomer subunits together form a long cleft where the catalytically important residue asparatic acid from both subunits are located in a coplanar configuration. Each monomer subunit has distinct structure called “flap structure” containing anti parallel beta-sheets and a beta loop in between the sheets, extending over the substrate binding active site [Ho et al 1994, Hosur et al 1994].

The amino acids towards the end of the N-terminus and C- terminus of each subunits are conventionally named P1, P2, P3 and P1’, P2’, P3’ etc respectively [Jack et al 1988]. At the active site, the subsites that interacts with the corresponding substrate or inhibitor side chains are named as S1, S2, S3 etc (from N terminus) and S1’, S2’, S3’ etc (from C – terminus) respectively from central asparatic acid residue. The seven amino acids at major processing subsites S4-S3’ are the most efficient at cleaving peptide substrates.

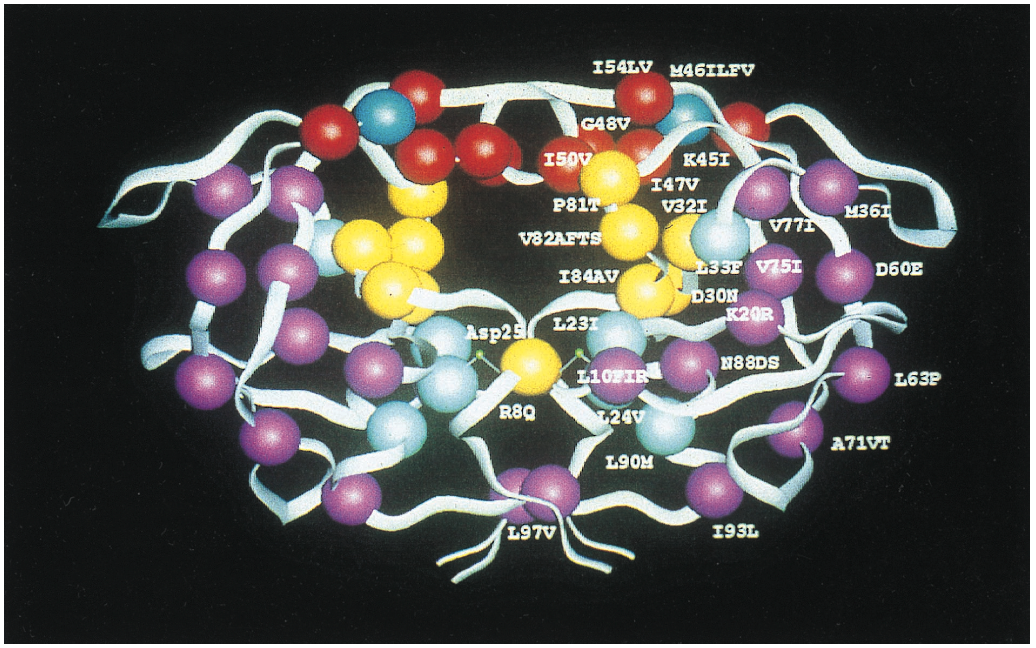


Figure: A detailed molecular structure of HIV-1 protease. Different regions of protease are denoted by color. Yellow – Active site region; Red – Flap region; Blue - flap hinge residue; Light blue - residues adjacent to the active site; Purple - residues distant from the active site of the enzyme. Source: Boden, Daniel, and Martin Markowitz. “Resistance to Human Immunodeficiency Virus Type 1 Protease Inhibitors.” *Antimicrobial Agents and Chemotherapy* 42, no. 11 (November 1, 1998): 2775–2783.

**Structure of reverse transcriptase**

HIV-1 reverse transcriptase is a heterodimer of two subunits: p66 (66 kilo Daltons) and p51 (51 kilo Daltons). The catalytic polymerase domain is located at 440 amino acids at N – terminal of p66 subunit and the RNaseH domain is located at 120 amino acids at C – terminal of the same subunit [Jacob and Molina et al 1993]. The p51 subunit is 440 amino acids long and is identical to the first 440 residues of p66 subunit from N – terminal but has significant structural conformation. The identical regions of p66 and p51 subunits have four sub domains: fingers (residue 1-85, and 118-155), palm (residues 86-117 and 156-236), thumb (residue 237-318) and connection (319-426) [Kohlstaedt et al 1992, Jacobo and Molina et al 1993]. However, the relative positions of the four subdomains at identical regions of p66 and p51 are different (fig below). Reverse transcriptase functions two enzymatic activities: DNA polymerase and RNase H, both required for copying single stranded RNA to double stranded DNA. Both the enzymatic catalytic functions occur in the active site of p66. The polymerase and RNase H has spatially distinct domains in p66. The four subdomains: fingers, palm, thumb and connections in p66 discussed above occur in polymerase domain (Figure).

The DNA binding cleft is formed in the subdomains of polymerase and RNase H. The p51 sub domains “thumb” and “connection” form the floor of DNA binding cleft. The DNA, bound to the cleft, makes contact with active sites of both polymerase and RNase H. The p66 thumb positions the DNA at the cleft through interactions that involve primer and template strands.

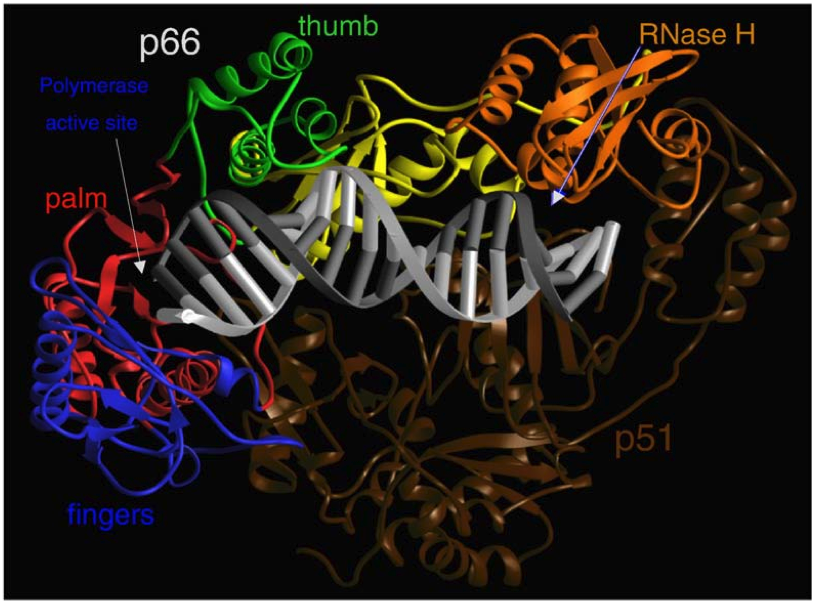


Figure: Ribbon representation of HIV-1 Reverse Transcriptase complexed with nucleic acid. The fingers, palm, thumb, connection, and RNase H sub domains are encoded with colors blue, red, green, yellow, and orange, respectively. The p51 sub - unit is in dark brown. The template and primer DNA strands are shown in light gray and dark gray, respectively. Source: Huang, Huifang, Rajiv Chopra, Gregory L. Verdine, and Stephen C. Harrison. “Structure of a Covalently Trapped Catalytic Complex of HIV-1 Reverse Transcriptase: Implications for Drug Resistance.” *Science* 282, no. 5394 (November 27, 1998): 1669–1675. doi:10.1126/science.282.5394.1669.

The polymerase active site of p66 has three catalytic carboxylates: D110, D185 and D189. They bind two divalent ions (Mg++ or Mn++) that are required for catalysis [Larder et al 1987].

Some conserved residues that are involved in DNA polymerization are: K65, R72, Y115 and Q151 [Huang et al 1998,Martin-Hernandez et al 1996, Gao et al 1997, Boyer et al 2000].

### Drug Resistance in HIV

Lots of researches have been conducted for elucidating the details of HIV life cycle [Frankel and Young 1998]. HIV proteins with active participation in the viral replication have been identified with detailed information and their three-dimensional structures. Different strategies based on the function and three-dimensional structures of the viral proteins are applied for anti retroviral drug development, to arrest HIV replication. Many anti retroviral drugs have been developed to date; eventually all failed to arrest HIV replication. The reason for the lost of therapeutic drug effect is the development of HIV resistance to those drugs.

HIV develops resistance to drugs with suitable mutations to counteract the drug effect. HIV has extremely high replication rate of around 1010 virus produced per day [Coffin at 1995, Ho et al 1995]. The error-pruned reverse transcriptase adds a wrong base per 18000 nucleotides during cDNA elongation [Roberts et al 1988]. This indicates that every base position in HIV genome of around 10,000 nucleotides has high probability of wrong base addition during polymerization. This creates a high genetic variability in HIV population. This pool of genetically variable HIV variants is called HIV viral quasispecies.

Larder et al first observed resistant HIV in 1989 for the drug Zidovudine (AZT; 3′-azido-3′-deoxythymidine). Since then, many other drugs have been developed that target HIV proteins crucial at different stages of its life cycle. HIV has been successful to resist any antiretroviral drugs (provide list of drugs at annex) designed to date. HIV acquires mutations at different positions of its genes to resist the drugs. Some drugs require one mutation for HIV to be resistance (low genetic barrier) whereas others require two or more (high genetic barrier). Clinical monotherapy leads to drug failure in short time. Power combinations of triple and quadruple mixture of protease and reverse transcriptase inhibitor have superseded monotherapy, to increase the genetic barrier for HIV [Gulick et al 1997, Hammer et al 1997, Perelson et al 1997]. Combined antiretroviral therapy has served successful reduction in viral load for certain time period. However, this does not expunge the virus from the patient, and eventually it replicates actively from the reservoir [Wong et al 1997]. Therefore, at the current context of our successfully designed drugs, HIV drug resistance is inevitable; sustained suppression of the virus is not possible.

**Mechanism of HIV acquiring resistance to drugs**

**Resistance to reverse transcriptase inhibitors**

Reverse transcriptase inhibitors include nucleoside analogue inhibitors (NRTIs), non-nucleoside inhibitors (NNRTIs), acyclic nucleoside phosphonates and one pyrophosphate analogue.

Nucleotide analogue inhibitors (NRTIs) are catalysed, inside the cell, to triphosphate form lacking 3’ hydroxyl at the ribose moiety [Arts and Wainberg 1996]. These NRTIs compete for incorporation with other normal naturally present nucleosides during polymerization. Addition of a defective nucleotide analog terminates the polymerization elongation process due to lack of 3’ hydroxyl group [Sluis and Cremer 2000]. One mechanism that HIV can resist NRTIs is that HIV can discriminate against the drugs via mutations. Reverse transcriptase obtains resistance to these nucleoside analogs with substitution mutations close to the nucleotide-binding site. The substituted amino acids (K65R, D67N, T69D, K70R and L74V) either fold over the triphosphate group of the incoming nucleotide analog [Huang et al 1998] or (mutation: M184V) exerts steric clash with the oxathiolane ring of the inhibitors and interferes with its binding [Sarafianos et al 1999, Gao et al 2000]. Two inhibitors zidovudine and stavudine have their triphosphate at distant from dNTP-binding site of RT. Another mechanism of gaining resistance to NRTIs is hydrolytic removal of analog inhibitors mediated by nucleotide excision mutations. Mutations in RT (M41L, L210W, T215F/Y and K219Q/E) removes NRTIs from the elongating nucleotide chain through phosphorolysis mediated either ATP or pyrophosphate [Boyer et al 2002, Arion and Parniak 1999, Meyer 1999, Mas et al 2000, Boyer et al 2001, Meyer et al 2000, Naeger, Margot, Miller 2002]. Mutation in RNasH lowers the affinity for degradation of the template RNA while cDNA synthesis. This delay in degradation allows more time for excision of the incorporated NRTI [Sluis and Cremer 2000, Nikolenko et al 2004].

Non-nucleoside inhibitors (NNRTIs) bind to the hydrophobic pocket at proximity of active polymerization site [De Clercq 1999, Sarafianos et al 2009]. In the wild type HIV, NNRTIs binding to the hydrophobic pocket is stabilized by interactions of aromatic rings in inhibitors with lipophilic pocket residues’ side-chains in Tyr 181, tyr 188 [Ren J et al 2001], Trp229 and Tyr318, electrostatic forces with resides Lys101, Lys103 and glu138, van der Waals forces from residues Leu100, Val106, Val179, Tyr181, gly190, Trp229, Leu234 and Tyr318; and hydrogen bonds between NNRTIs and main chains of the RT enzyme [Menendez and Arias 2002]. The range of mutations involved at three different clustering regions: 98 – 108, 179 – 190 and 225 – 236 residue number of p66 subunit and Glu138 of p51 subunit of RT hinder the formation of those forces and are responsible for resisting NNRTIs. The single amino acid mutations causing high level of resistance to low genetic barrier NNRTI drugs nevirapine and delavirdine are K103N, V106A, Y181C, Y188H, G190A and P236L [Hsiou et al 2001]. Efavirenz [Ren J et al 2001], capravirine [Gewurz et al 2004] and HBY097 are higher genetic barrier NNRTIs drugs requiring two mutations.

Pyrophosphate analogue like foscarnet inhibits RT by binding to the putative pyrophosphate-binding site to prevent chain elongation [Crumpacker 1992] but the mechanism of resistance to this drug has not been elucidated although mutations responsible for resistant have been identified [Mellors J et al 1995].

**HIV PR inhibitor resistance**

Protease is an essential enzyme that catalyses the proteolysis breaks down of gag and gag-pol polyprotein to functional proteins to build up new viral particles [Boden and Markowitz 1998, Huff and Kahn 2001]. Protease inhibitor (PI) has been the key component to the success of the combined antiretroviral therapy. PIs are substrate-based compounds that act as competitive inhibitors of the proteolytic breakdown of the polyproteins. Low pharmacokinetics of drugs in plasma is responsible for emergence of resistance HIV. Suboptimal therapy in which a PI is combined with RT inhibitors suffers low pharmacokinetics as PIs are eliminated by cytochrome P450 from plasma. PI doses are kept high in plasma, with adminstration of low dose of ritonavir that suppress cytochrome P450 activity [Sham et al 1998], to set high genetic barrier that ensures HIV has to incorporate many mutations in protease to be resistance. This delays the emergence of resistant viruses [Molla et al 1996, Zhang et al 1997]. HIV could achieve cross-resistance to the structurally similar PIs (Figure 9 : structures of PIs) [Condra et al 1995, Hertogs et al 2000].

Protease enzyme is a dimer of two non-covalently associated polypeptides. Each polypeptide has conserved region of Asp-Thr-Gly sequence; Aspartyl group of this region is the key for the enzyme catalysis function. Substitution mutations like D30N, G48V, V82A and I84V are prevalent in PI resistant viruses; the mutations are usually found in substrate binding pocket of the protease enzyme (Figure). The mutations reduce the overall viral fitness with reduction of the catalytic activity and viral replication. Although the protease enzyme activity is reduced with high dose PIs, HIV compensates the reduced viral fitness with additional mutations at cleavage site p7(NC)/P1 or p1/p6 of gag polyprotein region [Zhang et al 1997, Berkout 1999, Menendez, Arias 2002, Nijhius et al 2007]

**Resistance to Integrase inhibitors**

Integrase catalyses two main Mg++ or Mn++ dependend functions: 3’ processing and strand transfer of RT/viral DNA pre-complex. Currently available integrase inhibitors target the inhibition of strand-transfer. The foremost bad side of the strand transfer inhibitors is that they have low genetic barrier that quickly confer substantial low antiviral activity. The accumulation of mutations in HIV exhibits high-level resistance to inhibitors [Fikkert et al 2003]. The mutations observed that exhibit reduced susceptibility to inhibitor L-870810 are V72I, F121Y, T125K and V151I; similarly mutations selected by diketo acids are T66I, L74M, S153Y, Q148HR, M154I, N155S and S230R. Mutations observed for widely used drugs raltegravir and elvitegravir are (show the mutations to the drugs). The mutation positions show that L-870810 and diketo acids mutation patterns do not overlap. Susceptibility test of resistant virus from one integrase inhibitor to another inhibitor shows resistant virus is not cross-resistant [Hazuda et al 2004] but HIV, resistant to a diketo acid derivative may exhibit cross resistance to another diketo acid derivative [Fikkert et al 2003]. The associated resistant mutations map to the active site of integrase enzyme (show structure of integrase enzyme).

Integrase inhibitors L-870810 and diketo acids engage the divalent metal ions Mg++ and Mn++ at the integrase active site. The spatial confirmation of the chelating moieties in the inhibitors is observed to be consistent with spacing between the two active site metals [Grobler et al 2002]. Mutations at the active site of integrase interfere the available space for chelating moieties rendering unsuccessful engagement of the metals [Hazuda et al 2004]. However, some mutations are distal to the metal binding active site and resistance exhibited by these mutations suggests inhibitor-enzyme interactions may extend beyond the active site of integrase enzyme [Hazuda et al 2004]. HIV is able to switch between the resistance pathways (N155H pathway to the Q148H/G140S pathway in case of drug raltegravir) to escape the drug [Mukherjee et al 2011].

**Resistance to Entry inhibitors**

HIV use CCR5 as the primary co receptor to gain entry into the host cell [Morris et al 2004]. Maraviroc (Pfizer Inc.), which has been approved for clinical prescription to antiviral therapy experienced patients and vicriviroc (Schering-Plough Research Institute), which is in preclinical trial are two widely discussed entry inhibitors [Tsibris and Kuritzkes 2007, Kuhmann and Hartley 2008]. Maraviroc binds to CCR5 co receptor, blocking HIV entry to host cell. Many researches have found the mode of developing resistance by HIV to CCR5 binding inhibitors is the switch of co receptor to CXCR4 [Regoes and Bonhoeffer 2005, Moncunill et al 2008] at later stage of infection. However, research on Indian HIV population shows no co receptor switch [Cecilia et al 2000].

The resistance mechanism of HIV against CCR5 mechanism is explained as either competitive or non-competitive [Trkola et al 2002, Farger and Berger et al 2002, Kuhmann et al 2004, Pugach et al 2007, Moncunill et al 2008]. In competitive resistance mechanism explains that HIV gp120 changes the confirmation that has high affinity for inhibitor-free CCR5 co receptors. HIV gp120 can only bind to CCR5 co receptor at non-saturating concentration of the inhibitor [Gorry et al 2002, Koning et al 2003]. Increase in the drug concentration overcomes the competitive resistance. In non-competitive resistance, HIV gp120 acquires the ability to bind to inhibitor bound co receptor but less efficiently than free co receptor [Pugach et al 2007]. Between the two resistance mechanisms, the non-competitive method is suggested to be predominant in resistant HIV. The stepwise accumulation of mutations at V3 region of gp120, explained in these researches [Trkola et al 2002, Kuhmann et al 2004], are likely to be involved in both competitive and non competitive resistance mechanisms. Researchers are still unable to uncover the resistance mechanism of non-competitive resistance. The general accepted model is that bridging sheet and V3 base of gp120 binds to the tyrosine-sulfated N-terminus and V3 interaction with ECL2 of CCR5 [Cromier and Dragic 2002, Tsamis et al 2003, Huang et al 2007]. The binding of CCR5 antagonist alters the confirmations of two gp120 binding site in the co receptor. The change in the confirmation is not recognized by the incoming HIV gp120. Accumulation of mutations in resistant HIV gp120 changes its confirmation, which now can recognize the new confirmation of inhibitor bound co receptor.

**Resistance to Fusion inhibitors**

Perhaps the lowest number of antiviral drugs being developed falls under fusion inhibitor. The first fusion inhibitor Enfuvirtide was approved for clinical use a decade ago and within a year, HIV drug resistance against the drug was observed. Enfuvirtide binds to the heptad region of gp41, which is critical for formation of fusion pore in the host cell membrane. The genotypic analysis of sensitive and resistant HIV virus shows the resistant mutations at positions G36, I37, and V38 of heptad region towards N-terminal [Rimsky et al 1998, Wei et al 2002]. The mutations affects binding of the inhibitor at the heptad region and HIV replication is not interrupted. Some researches show highly variable gp120/gp41 is also responsible for resistance against fusion inhibitor [Blair et al 2000, Miller and Hazuda 2004]. Viruses displaying fast fusion kinetics to CD4 and higher affinity to co receptor show reduced susceptibility to fusion inhibitor than the viruses with lower fusion kinetics to CD4 and lower affinity to co receptor [Reeves et al 2002, Miller and Hazuda 2004].

### HIV Drug Resistance Testing

After high scale up of HIV antiviral drugs in low and middle-income countries, public health experts are concerned for potential emergence of drugs resistant virus and transmission of drug resistant virus in the population under ART and/or to non-infected people. In 2004, WHO and HIV ResNet collaborated to develop a **Global strategy for prevention and assessment of HIV drug resistance** to be implemented in low and middle-income countries*.* The global strategy aimed for surveillance and monitoring HIV drug resistance and to be prepared for drug resistant HIV epidemic emergence. The strategy for surveillance and monitoring HIV drug resistance aims are:

* To select the first and second line regimens of ART, as well as antiretroviral drugs for prevention of mother-to-child transmission (PMTCT), at population level
* To support national HIV programs in minimizing the emergence and transmission of HIVDR.

World Health Organization (WHO) recommends that first and second line antiviral drug regimens should be selected based on HIV drug resistance test report. This is supported by the fact that increasing number of researches shows the presence of drug resistant virus before starting a new drug regimen [Baxter et al 2000, Demeter and Haubrich 2001, Sebastian and Faruki 2003, Haubrich and Demeter 2001]. Also the success of virologic suppression is impacted by the choice of prescribed antiviral drug regimen [Matthews et al 2002]. Both phenotypic and genotypic assays for drug resistance testing are widely available to clinicians to measure the susceptibility of HIV to drug regimen. The assays help clinicians for long term antiretroviral treatment planning. Phenotypic assays use viral sample from infected patient to assess its response to individual drugs whereas genotypic assays use the actual viral DNA sequence, to assess drug resistance on the basis of the established HIV genetic mutational patterns. Many studies have shown that patients who undergo genotypic assays respond to the antiviral therapy better than control patients who are not under genotypic assays [Baxter et al 2000, Cingolani et al 2002, Cohen et al 2002, Durant et al 1999, Haubrich, Keiser et al 2001, Maynard et al 2002, Tural et al 2002]. Genotypic assay has advantages like detection of HIV variant at very low prevalence, provides insight into the potential emergence of drug resistant variants and can detect transitional mutations that do not cause drug resistance by themselves but indicate the presence of selective drug pressure. For treatment-experienced patients, the next choice of drug regimen based on genotypic assays is virologically superior to treatment history as well [Haupts et al 2003]. Genotypic assay is the most helpful for clinical management of patient with primary HIV infection consisting highly drug resistant viral variants [Narciso and Lazzarin 2003, Visco-Comandini and Balotta 2003, Riera-Jaume et al 2006]. Therefore, genotypic assays are widely used for detecting resistant viral isolates [Shafer et al 2002]. Genotypic assay steps include sequencing RNA from a sample of viral swarm; compare the sequence to a standard HIV reference sequence and use mutations in sequence as query to well-documented HIV mutation database. For this purpose, all the drug resistant mutations observed in research experiments are well documented in databases like Stanford HIV mutations database (<http://hiv.stanford.edu>), Rega (<http://rega.kuleuven.be/>) and ANRS ([www.medpocket.com](http://www.medpocket.com)) with comparable decision support systems to interpret any change in the HIV sequence to drug resistant mutations [Frentz et al 2010]. The online database of RT and protease sequence at Stanford University (<http://hiv.stanford.edu>) has two sequence analysis programs HIVseq (<http://hivdb.stanford.edu/DR/asi/releaseNotes/index.html#hivseq>) and HIValg (<http://hivdb.stanford.edu/DR/asi/releaseNotes/index.html#hivalg>) for interrogating and interpreting the mutations in the input sequence to the sequence databases (Stanford HIVdb, Rega or ANRS).

With the advent of ultra deep high throughput sequencing technology, we are able to sequence even the minor HIV variants, which are not detected with early traditional sequencing methods. Ultra deep high throughput sequence technology is able to sequence 1% or low prevalence HIV variants [Johnson et al 2008, Hedskog et al 2010, Siemen et al 2007, Siemen et al 2009, Balduin et al 2009]. Different sequencing technologies are discussed in the next section.

## Next Generation Sequencing Technologies

DNA sequencing technology has made lips and bound advancements and played a pivotal role in all the research areas that touch genomics [Gilbert 1981]. Since the first DNA sequencing technology developed by Sanger using dye-labeled dideoxynucleotide chain termination method [Sanger et al 1977], many sequencing technologies have been developed that revolutionized genomic fields. The automated Sanger sequencing technology had ruled the sequencing market for over two decades accomplishing milestones like completion of First grade Human Genome Sequence [Collins et al 2004]. This method is considered as the ‘First Generation sequencing technology’. Despite some technical improvements, the first generation technology could not meet the huge demand in the market. This directed to the development of new methods with high throughput, much cheaper at cost and less technical hassles. The newer technologies are considered as Next Generation Sequencing (NGS). NGS has changed the scientific approaches in basic, applied and clinical research. All NGS technologies constitute various combined strategies of template preparation, sequencing and imaging [Metzker 2009]. The booming use of NGS is the ability to produce an enormous volume of data at very cheap price in comparison to the First generation method. Various NGS products coexist in market with some having advantage to other in a particular application. Current NGS technologies in the market are Roche/454 [Margulis et al 2005] (<http://454.com/products/technology.asp>), Illumina/Solexa [Bentley et al 2008, Li et al 2009] (<http://www.illumina.com/systems.ilmn>), Applied Biosystems SOLiDTM system. The following technologies are consideresed “next-next generation sequencing technologies” or “third generation sequencing technologies” : Life/APG and Helicos Biosciences ([www.helicosbio.com](http://www.helicosbio.com)), Pacific Biosciences Single Molecule Real Time sequencing ([www.pacificbiosciences.com](http://www.pacificbiosciences.com) ) [Chin et al 2013], Ion Torrent ([www.iontorrent.com](http://www.iontorrent.com) ) [Rothberg et al 2011] and Nanopore sequencing [Branton et al 2008].These technologies have enabled high throughput sequencing from single molecule DNA and differ from NGS technologies at DNA synthesis and scanning.

**Roche 454 Genome Sequencer FLX system**

It is the first NGS system launched into the market in late 2005. The technology is based on sequencing by synthesis method with emission of light from pyrophosphate chemical reaction [Margulies et al 2005].

DNA library preparation for sequencing consists of shearing the sample DNA to fragments, ligation of standard 454 specific A and B adaptors to the ends of fragments. The purpose of adaptors is to provide priming site for amplification and sequencing. B adaptors contain 5’-biotin tag for immobilization of library DNA fragments onto magnetic beads coated with streptavidin. The fragmented DNA is denatured to release the non-biotinylated strand while biotinylated strand form template DNA library (Figure). The imaging system in GS is not designed to detect fluorescent from single base addition. Therefore the single strand DNA requires amplification. Roche system employs emulsion based clonal PCR amplification (emPCR) for library amplification [Dressman et al 2003]. In the emPCR, the single strand DNA is immobilized by hybridization of adaptor and primer, which is coated on the capture beads under a condition optimized to capture one DNA strand per bead [Metzker 2009]. The beads containing DNA are emulsified along with amplification reagents in water-in-oil mixture that forms microreactor. Each DNA containing microbead and necessary reagents are capture in a microreactor, where clonal amplification of the template DNA occurs. Upon amplification, the microreactors are broken, releasing the beads. In the next step of pyrosequencing, the microbeads are packed in PicoTiterPlate device. The PicoTiterPlate has nearly 2 millions of wells for simultaneous reaction in each well. The device enables high throughput sequencing. The device is designed in such a way that each well can accommodate only one microbead. The enzymes sulfurylase and luciferase attached to enzyme beads are added to the wells. The two enzymes are the key components for identifying the number of base addition in sequencing. The packing beads are added in the well to ensure the beads are positioned in the well. The fluidic system flows buffers and nucleotides across the wells. Nucleotides are flown in predetermined base order such that any base addition detected during sequencing would mean the base added is the flowing base. DNA polymerase adds nucleotide to the growing strand if it is complementary to the template strand base. A series of chemical reactions detects the base(s) added to the growing strand. Nucleotide addition releases a pyrophosphate group, which is converted into Adenosine TriPhosphate (ATP) by sulphurylase enzyme using adenosine phosphosulfate. Luciferase enzyme uses luciferin to hydrolyze ATP to oxyluciferin that emits light. The light is detected by CCD camera. The light detected from a particular well denotes the nucleotide incorporation in the growing strand of the well. There is no terminator in the incorporated nucleotide, which means DNA polymerase go on incorporating nucleotides until the template strand has complementary base. The cycle is repeated and the pattern of light emission and detection reveals the actual sequence of template DNA. This indicates that the number of bases addition across the PicoTitrePlate is asynchronous.

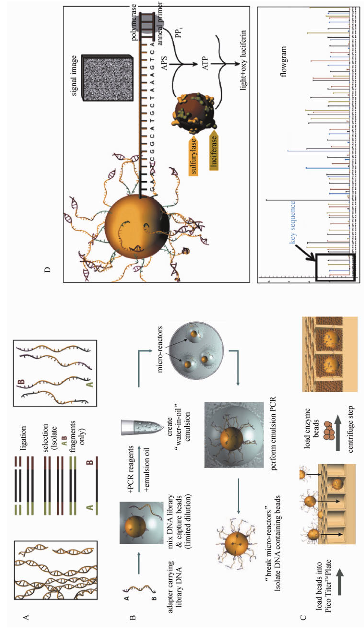


Figure: The Roche 454 GS working principle. (A) Prepare adaptor ligated single strand DNA. (B) Emulsion based clonal amplification. (C) Depositing DNA beads into the picotiter TM plate. (D) Sequencing and base calling.

This depends on the base order of template strands across the wells. The template strands with more homopolymer sequences are leading the strands with less homopolymers. Finally the raw reads are processed for poor quality sequences and checked for sequences without key sequences and mixed sequences (two or more sequences in a single microbead).

Homopolymer sequencing is the major limitation of roche 454 technology (Ruthberg and Laemon 2008). This is because there is no terminator group in an added nucleotide; multiple bases are incorporated consecutively at homopolymer site. The intensity/magnitude of light emitted depends on multiple repetitive bases added and this determines the sequence of template. Roche 454 technology is prone to insertion – deletion error while determining the number of bases added if homopolymer is longer. The next disadvantage is the cost per base addition is higher than other NGS technology [Ruthberg and Laemon 2008]. One advantage of this technology is the read length. Roche has marketed different models of platforms characterized with different read lengths and total reads outputs [Table].

|  |  |  |
| --- | --- | --- |
| **Platform** | **Read length (bp)** | **Reads output/Run** |
| GS Junior | 400 | 70,000 |
| GS FLX titanium XLR70 | 600 | 7,00,000 |
| GS FLX titanium XL+ | 1000 | 1,000,000 |

Table: Roche 454 platforms and the average read length and Reads output per Run. Source: www.454.com

The best application of Roche 454 technology regarding its long read length is de novo assembly and metagenomics [Zhou et al 2010].

**Illumina/Solexa**

Illumina/Solexa came into the market as a competitor to Roche 454 in 2006. Though it has similar basic working principle – sequencing-by-synthesis, as in Roche 454 technology, there are differences in library preparation, nucleotide incorporation and light imaging. Sample DNA is fragmented to required size and a single ‘A’ base is added to 3’-ends of the fragment DNA. Adaptors, with single ‘T’ overhangs, are added to both ends (5’ and 3’) of the DNA. The DNA is then denatured and a single strand is immobilized on the surface of solid support flow cell. The flow cell surface is coated with adapters and complementary adapters; single strand DNA gets immobilized when adaptor at one end of it anneals with its complementary adaptor on the flow cell. The adaptor at the other end is free to hybridize with its complementary adaptor on the flow cell and this creates a ‘bridge’ structure (Figure). These flow cell adapters functions as primer as well for PCR reaction. The single strand DNA is amplified at its spatial location by ‘bridge PCR’ method. This forms a cluster of clones of original template single strand DNA [Adessi et al 2000, Fudurco et al 2006]. Millions of clusters are formed from spatially separated millions of single strand template DNA (Figure).

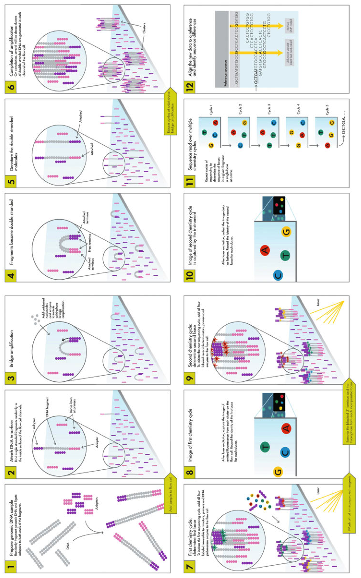


Figure: The working principle of Solexa platform.

Sequencing step follows PCR amplification. The sequencing reagents and all four nucleotides labeled with different fluorescent dyes are supplied on to the surface of the flow cell. Each nucleotide also has a terminator. DNA polymerase incorporates the complementary nucleotide and terminator group blocks further incorporation. The CCD camera detects and identifies terminal nucleotide and the position of its fluorescent dye on the solid support. The fluorescent dye and terminator group are removed from the 3’-end of the terminal nucleotide for next round of sequencing cycle. This ensures base-by-base sequencing. The cycle is repeated for predetermined number of times. A base-calling algorithm assigns sequences and associated quality values to each read and a quality-checking pipeline evaluates the illumina data from each run, removing poor-quality sequences. The current illumina platform in the market and their performance comparison is shown in the table.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Illumina Platforms** | **HiSeq 2500/1500** | **HiSeq 2000/100** | **HiScanSQ** | **Genome Analyzer IIx** | **MiSeq** |
| **Output (GB Max)** | 600 | 300 | 150 | 95 | 7.5 – 8.5 |
| **Single Reads** | 3 billion total  187 millions/lane | 1.5 billion total  187 millions/lane | 750 million total  94 million/lane | 320 million total 40 million/lane | 15 - 17 million total 15 - 17 million/lane |
| **Paired end reads** | 6 billion 374 million/lane | 3 billion 374 million/lane | 1.5 billion 188 million/lane | 640 million 80 million/lane | 30 - 34 million total 30 - 34 million/lane |
| **Required input** | 50 ng with Nextera 100 ng - 1 µg with TruSeq | 50 ng with Nextera 100 ng - 1 µg with TruSeq | 50 ng with Nextera 100 ng - 1 µg with TruSeq | 50 ng with Nextera 100 ng - 1 µg with TruSeq | 50 ng with Nextera 100 ng - 1 µg with TruSeq |
| **Read length (bp)** | 2 x 100 | 2 x 100 | 2 x 100 | 2 x 150 | 2 x 250 |
| **Percentage of Bases > Q30** | > 85% (2 x 50 bp) > 80% (2 x 100 bp) | > 85% (2 x 50 bp) > 80% (2 x 100 bp) | > 85% (2 x 50 bp) > 80% (2 x 100 bp) | > 85% (2 x 50 bp) > 80% (2 x 100 bp) | > 85% (2 x 100 bp) > 80% (2 x 150 bp) > 70% (2 x 250 bp) |

Table: Comparison of different illumina platform performances.

(Source: <http://www.illumina.com/systems/sequencing.ilmn>)

Illumina/Solexa produces huge amount of data, upto 600 gigabtyes (Table). The time and cost per base is much reduced in comparison to first generation technology, but the reads are shorter and is prone to high error rates often resulting to false positive sequence variations [Metzker 2010]. Because millions of reads are generated, erroneous base is often corrected by coverage but unveiling the complete sequence contiguity is often limited by short read length [Lander and Waterman 1988, Arratia et al 1991].

**Applied Biosystems SOLiD technology**

Support Oligonucleotide Ligation Detection (SOLiD) sequencing technology is based on sequencing-by-ligation technology. Alike Roche 454, SOLiD uses adapter-ligated fragment library and emulsion PCR to amplify the input DNA fragments for sequencing, on small 1-micrometer magnetic beads [Mardis 2008]. The template sequences are denatured and mangnetic bead enrichment is performed to selected beads with highly amplified template DNA. 3’ modification is done on the template sequences to allow covalent bonding to the solid support. The magnetic beads are then deposited on a flow cell slide.

SOLiD uses DNA ligation technique for sequencing, unlike any other NGS systems. This means, unlike other NGS, there is no any DNA polymerase generated errors during sequencing. During sequencing, a universal sequencing primer is hybridized to the amplified template sequences and a library of 1,2-probes (or dibase probe) along with DNA ligase enzyme is added. The probes are 8mers and are designed to interrogate 1st and 2nd position bases (Figure) such that the 16 dinucleotides (4 bases X 4 bases) are encoded by four fluorescent dyes (Figure). The fluorescent dye is attached at the end of octamer probe. Upon addition of probes, they hybridize with the complementary sequence on the templates followed by ligation by ligase enzyme. Unlike polymerization from at 3’-OH end, probes ligation can be bi-directional (5’-PO4 or 3’-OH ends). The fluorescence is measured and the dye is cleaved off from the fifth base leaving free end for the next probe ligation. The next cycle of probe hybridation, ligation, fluorescence measurement and dye cleavage is repeated for seven cycles, extending the sequence read length of 35 bases. The synthesized DNA strand is removed and a new universal primer is hybridized at offset position of one base (n-1) than the previous primer position. Again the seven cycles of probe hybridization, ligation, fluorescence measurement and dye cleavage is repeated. The primer resetting is repeated 5 round followed by 7 probe-ligation cycles providing dual measurement of each base (Figure).

After completion of 5 round primer reset sequencing, color calls are ordered in linear sequence and each base is interrogated from two different primers sequencing reaction [Figure a(8)]. This is Two Base encoding method. This method is able to discriminate measurement errors and contributes to low error rate and reduced systemic noise [Zhou et al 2010].

The output read length is short (35 bp) and the file size is over 60 GB (over 1 billion reads). Like Illumina, the short read length is a limitation in this technology as well [Metzker 2010]. Substitution error is the most common error type along with underrepresentation of AT-rich and GC-rich region [Harismendy et al 2009]. SOLiD data can be undercalling true variants, according to Shen et al 2008.

Some platforms from Applied Biosystems available in market are 5500 series SOLiDTM, SOLiDTM 3 plus and SOLiDTM 4.

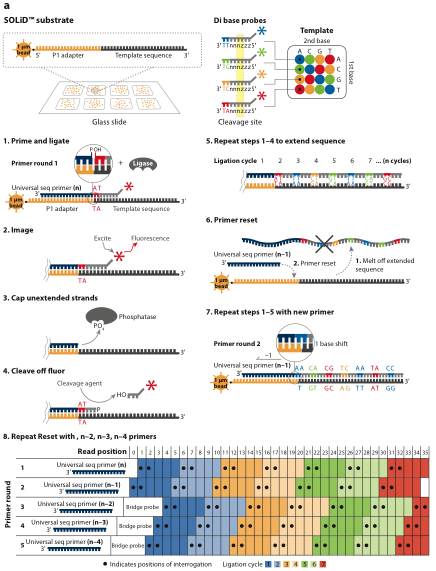


Figure: The workflow of SOLiDTM technology sequencing system. Source: Mardis 2008

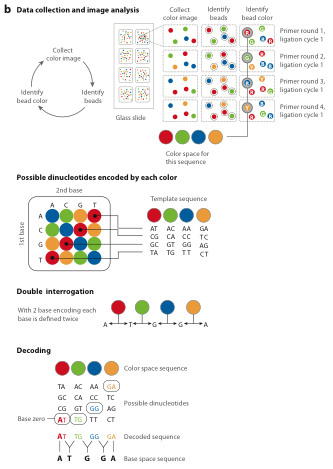


Figure: SOLiDTM system two base color coding system for double interrogating to get base order of template sequence. Source: Mardis 2008

**Helicos Bioscience**

Helicos Bioscience sequencing technology uses a single DNA molecule template. The technology requires very less amount of starting DNA (<1 microgram). This rules out any PCR related errors, as this technology bypasses PCR amplification step.

The single molecule templates are immobilized on a solid support by one of the two approaches. In the first approach (Figure), which is one pass sequencing method, primers are spatially distributed and covalently attached on the solid support [Harris et al 2008]. The DNA is sheared to small fragments of size ~ 200 to 250 bps. Adaptors, usually known sequence of poly A, are attached to the ends of the fragmented DNA; the DNA is then hybridized with the immobilized primers, usually poly T sequence, on the solid support by the adaptors. In second approach (Figure B), which is two pass sequencing method, the single molecule DNA is spatially distributed and covalently immobilized on the solid support by priming and extending single molecule templates from immobilized primers.



Figure: Two approaches of template DNA immobilization in Helicos sequencing technology.

The dye labeled single nucleotides mixed with DNA polymerase is flown across the surface. Polymerase incorporates labeled nucleotide in all fragments with complementary base at the first free position. The synthesis is asynchronous like in Roche 454; some strands grow ahead, fall behind or not extend at all [Zhou et al 2010]. The inbuilt camera captures all the fluorescence from entire support surface in series of images. For better resolution of the image, the captured DNA fragment molecules on the support surface have to be at least few hundred nanometers apart [Wash and Image 2008]. The labeled dye is cleaved off from the incorporated base allowing new base to add on the growing strand. The cycle of four labeled nucleotides is flown across the surface in predetermined order for 25 to 45 rounds, capturing images from the entire surface in each nucleotide flow.

Except the PCR step, this technology is similar to Roche 454, showing the similar problem of inaccurate sequencing at homopolymer region. The polymerase enzyme kinetics rate of base incorporation can be slowed down so that the free nucleotides could be washed off after a nucleotide addition on growing strand [Harris et al 2008].

One advantage of Helicos single molecule sequence is much reduced cost associated to sample preparation and amplification as in other NGS. Since no PCR amplification step, the sequencing has fewer errors. The major limitation is low accuracy at detection of single sequencing reaction, unlike light emission from huge number of clonally amplified templates in Roche 454. This low accuracy detection leads to deletion errors in sequencing. Another advantage of Helicos single molecule sequencing is ability to sequence the same template sequence by removing the newly synthesized strand. The sequencing accuracy is greatly increased sequencing the same template over and over [Zhou et al 2010].

**Pacific Biosciences Single Molecule Real Time (SMRT) Sequencing Technology**

Pacific Biosciences has also brought single molecule sequencing with real time detection [Eid et al 2009]. This sequencing technology is a revolution of nanotechnology. Unlike Helicos, DNA polymerase is attached to a support. The support has nanopore structured zero mode waveguide (ZMW) [Levene et al 2005, Samiee et al 2005], for real time detection of sequencing reaction [Levene et al 2003]. ZMW consists of huge number of sub-wavelength hole, each few nanometers in diameter, fabricated on a thin metal sheet supported by transparent substrate. An engineered φ29 DNA polymerase enzyme is attached to the bottom of these holes. These polymerase enzymes have high efficiency at incorporating phospholinked nucleotides and enables resequencing of close circular templates [Eid et al 2009, Korlach et al 2010].

During sequencing, DNA polymerase polymerizes the phospholinked hexaphosphate nucleotides; each nucleotide incorporated creates a burst of fluorescent light. The burst of light does not pass through the nano-sized hole, but leaves an exponentially decayed evanescence light wave at the bottom of the hole. This gives very small volume of fluorescence detection [Zhou et al 2010]. The fluorescence detection reveals the identity of the nucleotide. The phospholinked fluorescent group is removed as part of pyrophosphate. DNA synthesis does not stop after each nucleotide addition. A continuous burst of fluorescence is detected at each waveguide hole and the sequence of fluorescence determined the template DNA sequence.

The test of this sequencing technology was tested that showed only 83% accuracy []. A short interval between two consecutive nucleotide incorporation events and binding and releasing of nucleotide in the active site before incorporation to the growing strand are responsible for sequencing errors and thus low accuracy [Metzker 2010]. However, repeated sequencing of same template DNA molecule up to 15 times increases the accuracy to >99.999% [Eid et al 2009] with average read length 964 [Metzker 2010]. The low yield ratio (~30%) of useable DNA polymerase in the waveguide wells is another limitation [Korlach et al 2008].

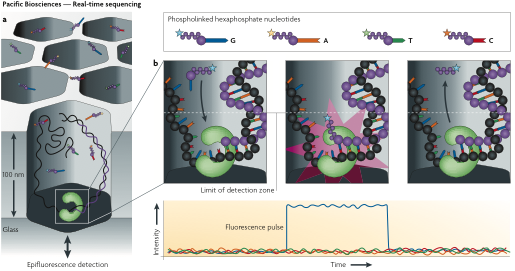


Figure: Sequencing single DNA molecule at zero mode waveguide. Source: Metzker 2010

**Ion torrent for DNA sequencing**

Ion torrent DNA sequencing is based on the detection of hydrogen ions released by DNA polymerase upon base incorporation in the growing strand. Ion torrent has overcome problem related to imaging technology, electromagnetic intermediates like light emission [Smith et al 1986] and preparation of specialized labeled nucleotide and reagents for sequencing. Ion torrent has shifted the sequencing chemistry to low cost ubiquitous electronic integrated circuits [Ruthberg et al 2011]. The integrated circuits have ion-sensitive field-effect transistor (ISFET) [Bergveld 1970, 2003] that is sensitive to hydrogen ions and compatible to CMOS processors [Mausells et al 1999, Jakobson et al 2002, Milgrew et al 2003]. The circuit has a large array of sensor elements, each with a single floating gate connected to an underlying ISFET. Therefore, the heart of ion torrent sequencing is CMOS integrate circuit.

Sequencing reactions are confined in a 3.5-μm-diameter well formed by adding a 3-μm-thick dielectric layer over the electronics and imprinted to the sensor plate. A tantalum oxide layer on the sensor plate detects the hydrogen ion. The semiconductor electronics connected to sensor array carries the sequencing information at high speed to the electronic reader board. A microprocessor on the board processes the incoming signal form sensor array [Ruthberg et al 2011].

DNA sequencing step including fragmenting the input DNA, ligating the DNA fragments to ligators and clonally amplifying the DNA onto 2 μm acrylamide bead. The amplification allows sensor well to achieve a high signal-to-noise ratio required for accurately detect the number of hydrogen ions releases. Beads with amplified DNA are selected by magnetic bead based process. Sequencing primers and polymerase are added to the template DNA and loaded on to the chip. The chip is centrifuged such that each sensor well contains only one bead as well depth is chosen to accommodate only a bead.

During sequencing, fluidic system controls all four differently labeled nucleotides in a stepwise fashion one at a time. The base is added by the polymerase, if it is complementary to the base in the template strand. DNA polymerase adds go on adding if the template has homopolymer sequence. Every nucleotide incorporated gets hydrolysed, releasing a single proton. Protons cause shift in the pH of the surrounding solution (0.2 pH change per single proton) proportional to the number of nucleotides incorporated in that flow. This change in pH is detected by sensor at the bottom of each well, converted to a voltage and digitalized by semi conductor integrated circuits. Signal processing softwares are used to convert the data for measurement of incorporation in that flow using a physical model. The model considers nucleotide diffusion rate, polymerase rate and buffer effectes. A base calling software takes account of all the signals change and corrected base calls for each flow in each well generating a sequence from each well. All this processes takes up to 4 seconds. The unincorporated nucleotides are washed off for next nucleotide flow. Washing off does not involve any enzymatic reagents as the nano-sized wells allow diffusion in and out in 1/10th seconds.

Each sequence read is quality filtered using two signal-based filters, removing low accuracy reads. Lastly, each base in the quality passed reads is assigned a quality value using Phred method [Ewing and Green 1998].

Ion torrent DNA sequencing is also limited by homopolymer. The quality of bases decreases sharply when the read length goes over 50 bases. The mappable reads percentage is also low. Factors like incomplete loading of beads to chip and poor clonal DNA amplification leads to poor sequencing error.

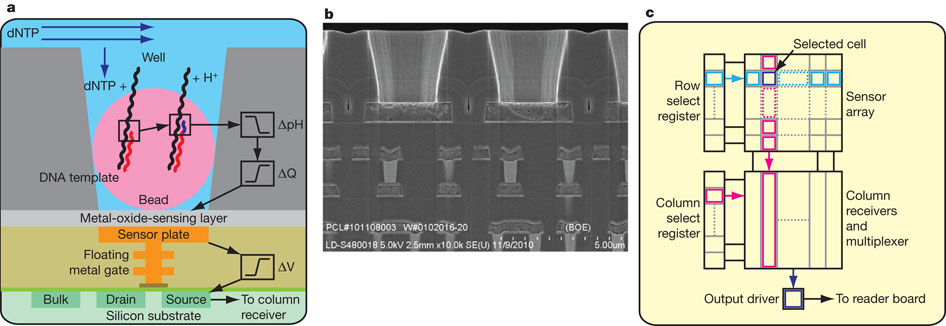
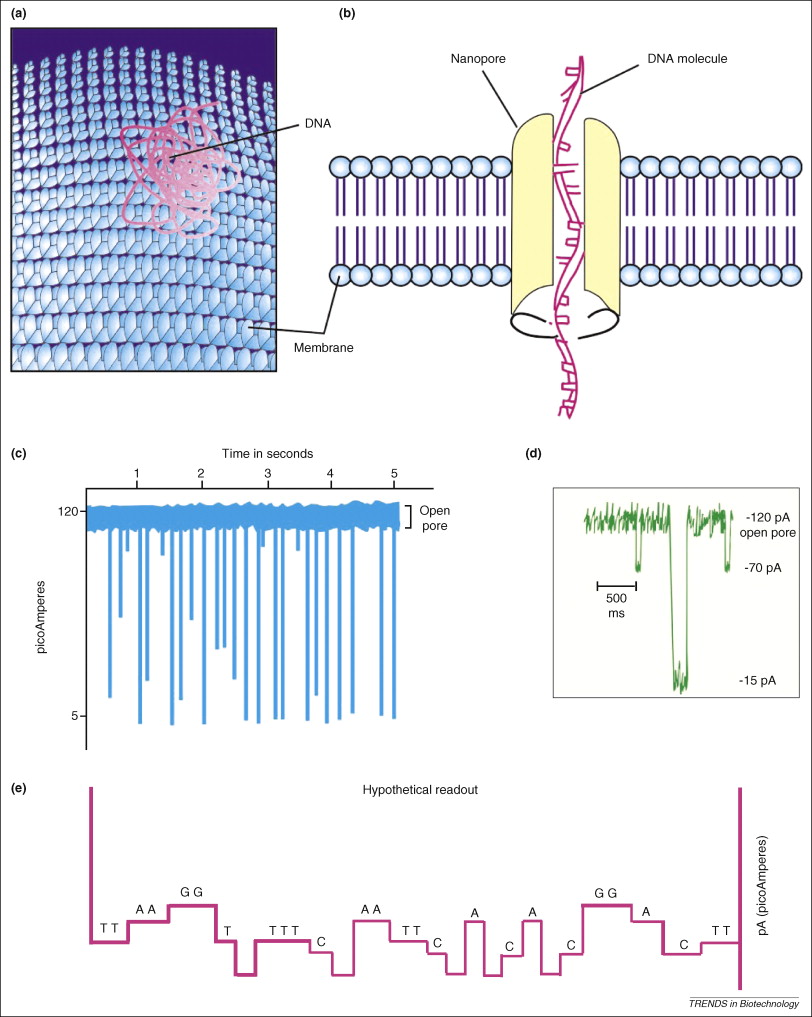


Figure: Ion torrent a) Sensor b) Well and c) chip architecture

Source : JM Rothberg *et al.* *Nature* **475**, 348-352 (2011)

**Nanopore Technology**



# Figure: Systematic representation of a general nanopore sequencing method. A) Membrane surface with nanopores B) DNA transiently passing through a nanopore C) peaks showing nucleotide base detection in DNA D) A graph showing change in current alteration when DNA pass through nanopore E) A hypothetical read out graph showing DNA sequence. Source: Gupta 2008, Single-molecule DNA sequencing technologies for future genomics research, Trends in Biotechnology, vol 26 (11)

In 1989, David Deamer from University of California (Santa Cruz, CA, USA) had first produced the idea of nanopore sequencing. The idea is developed from translocation of biopolymer in some natural events like HIV virus injecting its RNA through pores of 1-10 nm size. The technique is based on transition of a polynucleotide molecule through a hole of few nanometers size and detecting the bases by their effect on change in electric current [Kasianowicz et al 1996, Howorka, Cheley and Bayley 2001]. Kasianowicz et al 1996 first studied this in vitro using a α-hemolysin pore of few nanometer diameters in a lipid bilayer membrane. A current is applied across the lipid membrane. Negatively charged DNA or RNA can traverse through the nanopore to the opposite side with positive charge. Double stranded DNA [Fologea et al 2005] or ssDNA [Keyser et al 2006] can translocate through a solid-state nanopore; and ssDNA or RNA can pass through biological protein nanopore [Storm et al 2005]. This translocation event through nanopore brings change in the electrical conductance of the membrane [Vestgarden et al 2008] at the range of picoamperes (pA), measured using electric circuit [Friedman et al 2005]. Different nucleotides change magnitude of electrical conductance at different pA [Venkatesan and Bashir 2011] (Figure). This nanopore technology approach is able to distinguish DNA of similar length, DNA regions with difference in GC:AT ratio and nucleotide sequence of DNA [Rhee and Burns 2006, Ryan et al 2007]. Despite huge ongoing researches, this technology is at the initial proof of concept stage and in the estimated to be commercialized in near future [Gupta 2008]. Research on biological nanopore like alpha hemolysin and synthetic nanopore like carbon nanotube [Lu et al 2004, Fologea et al 2005, Lui et al 2010], atomically thin sheets of grapheme with nanopores [bayley 2010] are under development.

The technical difficulty in nanopore sequencing is translocation of DNA through nanopore at a speed too fast to enable the resolution of individual bases. Therefore, two different research groups apply some modifications in basic nanopore technology to develop a slightly modified nano sequencing technologies.

**NABsys Hybridization assisted nanopore sequencing** (HANS): NABsys (<https://nabsys.com>) in joint ventured with Brown University are developing this nanopore sequencing technology. This technique of sequencing combines nanopore sequencing with hybridization of each DNA molecule with library of individual 6-mers nucleotide probes. The probes library consists of all possible combinations of nucleotide bases. The hybridized DNA, now in the duplex form, is passed through the nanopore, and the change in electrical conductance is measured to detect the region of hybridization by each probe. The hybridized duplex region of DNA shows different electrical conductance than the single strand DNA. Each probe with known sequences can be hybridized to full length DNA and map them after passing through the nanopore. This approach determines the full length DNA sequence.

Complete Genomics (Mountain View, CA, USA) in collaboration with BioNanomatrix (Philadelphia, PA, USA; <http://www.technologyreview.com/news/409919/the-100-genome/>) is developing a slightly modified nanopore sequencing technology to NABsys HANS. Several differently labeled 5-mer probes are hybridized to the template full length DNA. The duplex DNA when transient through 100nm diameter channels, the fluorescence signals is captured by camera that records the hybridization sites of the probes. It is then repeated with different sets of 5-mer probes to hybridize at different locations of DNA so that at the end, the entire DNA is covered. The complete DNA sequence can be eluded using this technique and the read length can be several thousands.

**Direct, electrical detection of single DNA molecules**: Oxford is developing a nanopore sequencing method that uses three biological molecules - modified alpha hemolysin, exonuclease and synthetic cyclodextrin - all work in a system [Howorka et al 2001, Stoddart et al 2009, Clarke et al 2009]. The modified alpha hemolysin has a pore for polynucleotide to pass through. Exonuclease is attached to the extracellular face and synthetic cyclodextrin attached at inner surface of the hemolysis nanopore. The combinations of these molecules are placed in synthetic lipid membrane. The voltage is applied across the membrane changing the salt concentration and DNA is loaded into the nanopore. The exonuclease cleaves off each individual nucleotide from DNA. Each nucleotide is detected by the change in ionic current in the pore. This technique has the advantage of producing long read length, high scalability and low cost.

**Nanopore DNA sequencing with MspA**: This nanopore technique uses *Mycobacterium smegmatis* Porin A (MspA) protein as a pore as opposed to synthetic α-haemolysin in oxford nanopore. MspA is an octameric protein with pore size of ~1.2 nm diameter and ~0.5 nm length. Negatively charged amino acids obstruct the transition of negatively charged DNA through its pore. In MspA, negatively charged asparagines are replaced with neutral asparagines with site directed mutagenesis [Butler et al 2008, Manrao et al 2012]. Single stranded DNA transiting through the MspA pore is sequenced with changing current. The rapid transition of ssDNA through the pore [Li et al 2003, Strom et al 2005] lowers the efficiency of DNA sequencing. To slow down the translocation [Fologea et al 2005], a region of dsDNA is introduced in the MspA pore. The dsDNA region is an obstacle for significant large scale DNA sequencing [Schadt et al 2009].

**Nanopore DNA sequencing with optical readout**: This technology has emerged as a need for ultrafast approach for DNA sequencing using single molecule of polynucleotide. In this technology, nucleotides in DNA fluorescently labeled with markers that report its nucleotide sequence. Design DNA polymer (DDP) technology is used for this purpose. Then, the fluorescent-tagged oligonucleotide molecular beacons that is complementary to the DDP converted DNA is hybridized. The hybridized molecule is then passed through the nanopore. The molecular beacons are dehybridized in the nanopore, one at a time, and the flash of light from fluorophore of each beacon is detected. The dehybridization step limits the speed of DNA translocation through the nanopore. The contrasting multicolored bases are discriminated by the optical probes [McNally et al 2010], which is huge advantage over electronic nanopore sequencing method. Parallel use of high-density nanopore arrays [Meller et al 2011] can increase the throughput to 1 megabases per second [Soni and Meller 2007].

**Some final notes on NGS and raw sequence data**

All the next generation sequencing technologies have one or other outstanding issues related to short read length, higher error rate, technical difficulty in sample preparation, cost and the volume of the data generated. Bioinformatics tools are being developed to deal with all the issues from quality control to finish polished final data analysis. Quality control is the first step on after DNA sequencing. Different quality trimming tools are available online. There are different aspects for quality trimming of the raw data but for a specific purpose of the data analysis, it is not necessary that quality-trimming tools cover all the aspects. Therefore, often quality trimming tools are developed with some specific requirements for downstream data analysis. We developed a quality-trimming tool – QTrim, (discussed in Chapter 2) specifically for quality control of Roche 454 raw sequence data.