CHAPTER 1

# Literature Review

## Overview of HIV/AIDS

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

## Discovery and characterization of HIV

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

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

## Origin and evolution of HIV

Exploration of the retrovirus led researchers to identify similarities between HIV and a retrovirus in African non-human primates that were then known as Simian Immunodeficiency Virus (SIV) [Gao, 1994 #656]. About 40 different primates, in Africa, were infected with SIV with some harboring multiple strain of SIVs [Apetrei, 2004 #737]. Phylogenetic analysis of SIV from African non-human primates and HIV in human provided remarkable understanding of viral transmission as zoonotic [Bailes, 2002 #659] and evolution of the virus in human after transmission [Gao, 1999 #738]. HIV is divided into two groups – HIV-1 and HIV-2 [Gao, 1999 #738]. Each group resulted from an independent cross species transmission from different African non-human primates to human [Sharp, 2010 #646]. HIV-2 was discovered in 1986. This group was transmitted from sooty mangabey monkeys (Cercocebus atys) [Hirsch, 1989 #657] and its prevalence was also high in the geographical location of these monkeys in West Africa [Santiago, 2005 #658]. Sooty mangabey monkeys were naturally infected by a strain of SIV [Hirsch, 1989 #657]. The phylogenetic analysis of HIV-2 strains showed that they closely group with the SIVsmm strain [Hirsch, 1989 #657] that were non-pathogentic to its host monkeys [Gao, 1992 #542]. SIVsmm evolved in its host to produce multiple strains and subsequent multiple zoonotic transmissions from sooty mangabey monkeys to human [Hahn, 2000 #655] gave rise to different subtypes of HIV-2. Although HIV-2 subtypes A to G were identified in human, it was assumed that more subtypes were introduced into human [Gürtler, 2004 #654] but were lost for low adaptation fitness [Damond, 2004 #653].

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

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

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

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

## HIV-1 Diversity

### HIV-1 subtypes

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

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

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

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

### HIV-1 recombination

Initially, HIV-1 group M subtypes E [Artenstein, 1995 #931;Tovanabutra, 2002 #933;Wasi, 1995 #932] and I (Figure 1.5) were also classified (reviewed in [McCutchan, 2006 #493]). With availability of complete HIV genome sequence and phylogenetic analysis from it, the subtypes E and I were reclassified as circular recombinant forms CRF01\_AE (recombinant form of subtype A and E) and CRF04\_cpx (recombinant form of more than two subtypes, designated by “cpx”) respectively (reviewed in [McCutchan, 2006 #493]). The same criterion of epidemiological unlinked isolates from three or more people applies for classification as a circular recombinant form (CRF) [Robertson, 2000 #505]. A recombinant isolate that is discovered in single patient is termed as Unique Recombinant Form (URF) (reviewed in [McCutchan, 2006 #493]). There are 55 CRFs listed in Los Alamos National Laboratory database for HIV sequences (http://www.hiv.lanl.gov/content/sequence/ HIV/CRFs/CRFs.html) as on July 22, 2013. The recombination breakpoints to shuffle HIV genome [Archer, 2008 #11;McCutchan, 2002 #10;Zhang, 2010 #891;Salminen, 1995 #971] from different strains of the virus are listed in Los Alamos National Laboratory website (http://www.hiv.lanl.gov/content/sequence/HIV/CRFs/ breakpoints.html), each appeared in a publication. CRFs account for at least 20% of all the HIV infections [Robertson, 2000 #505;Robertson, 1995 #7;Sharp, 1995 #518]. CRF02\_AG is the most prevalent circulating recombinant form infecting over 9 million people on the whole [McCutchan, 2000 #939] and geographically epidemic in the West African region (Figure 1.5). CRF01\_AE is the dominant circulating recombinant form in South-East Asia (Figure 1.5) [Taylor, 2008 #214].

### Intra-patient HIV diversity

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

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

## HIV genome and proteins – structures and functions

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

### Accessory genes

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

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

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

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

### Structural genes and proteins

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

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

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

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

### Regulator Genes

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

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

## HIV replication

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

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

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

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

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

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

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

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

## Antiretroviral Drugs

### Reverse Transcriptase Inhibitors

#### Nucleoside Reverse Transcriptase Inhibitors (NRTIs)

NRTIs are analogs of nucleotides but without 3’ hydroxyl group (reviewed in [Sarafianos, 2004 #293]). The drug is taken in unphosphorylated form, which cytokinases phosphorylates to form 5’ triophosphates (reviewed in [De Clercq, 2002 #414;Ilina, 2008 #413]). This then, leads to the incomplete termination of HIV-1 cDNA synthesis [Sluis-Cremer, 2000 #5]. The action of the drugs is shown in **Figure 1.8** and the list of approved NRTI drugs shown in **Table 1.1**.

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

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

### Protease Inhibitors (PI)

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

### Integrase Inhibitors

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

### Cell entry inhibitors

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

#### CCR5 co receptor antagonist:

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

#### Fusion Inhibitors

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

## HIV Treatment

### Brief history of antiretroviral treatment

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

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

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

### Treatment guideline

The World Health Organization (WHO) has produced the clinical guideline (http://www.who.int/hiv/pub/guidelines/arv2013/art/en/index.html) for HIV treatment. From the treatment point of view, the guideline has grouped HIV infected individuals as adult, pregnant and breast feeding women, children less than 3 years old, children 3 or more years old and HIV infected individuals co-infected with other diseases. Regardless of the grouping, the guideline recommends treatment initiation to all confirmed HIV infected people with CD4+ cell count less than or equal to 350 cells/mm3 in resource poor countries and less or equal to 500 cells/mm3 in resource rich countries. However, where laboratory tests for CD4+ count may not be feasible, the WHO clinical stages should be used as a guide for treatment initiation [Weinberg, #1161].

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

Patients undergoing ART therapy are monitored for effectiveness of the treatment at a defined interval. Laboratory tests for viral load should be done for monitoring the treatment response. A viral load of greater than 1000 viral RNA copies/ml blood sample indicates virologic failure to the treatment and the patient is recommended to switch to new drug regimen (**Figure 1.11**). If a viral load test is not feasible routinely, CD4+ count and clinical monitoring need to be used [Weinberg, 2010 #1161].

## HIV Drug Resistance

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

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

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

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

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

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

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

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

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

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

## HIV drug resistance genotyping

### Conventional Population Based HIV Drug Resistance Genotyping

The presence/absence of certain drug resistant mutations (**Figure 1.13** and **Figure 1.14**) in HIV has strong relation with virologic failure [Lorenzi, 1999 #1045;Zolopa, 1999 #1046;van Leeuwen, 1995 #1047;Larder, 1989 #1049;Condra, 1995 #1050;Molla, 1996 #1051;Larder, 1989 #1052] and characterization of these drug resistance mutations (DRMs) can be used to optimize the antiretroviral therapy [Baxter, 2000 #1053;Durant, 1999 #1055;Van Vaerenbergh, 2001 #229]. Conventional HIV genotyping involves Sanger dideoxy termination based population sequencing that produces consensus sequence of the viral population in a sample [Ewing, 1998 #1114;Sanger, 1977 #1115;Metzker, 2005 #1116].

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

## Next Generation sequencing technologies

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

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

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

Ion torrent technology implements sequencing by synthesis method and electronic sensors connected to complementary metal-oxide-semiconductor integrated circuit are used with a microprocessor for signal processing [Jakobson, 2002 #148;Milgrew, 2004 #149]. The sequencing step is similar to Roche/454 homopolymer sequencing but the base detection is completely electronic, and that reduces the ion torrent cost relative to other systems [Glenn, 2011 #1411]. During DNA sequencing, a base incorporation releases a hydroxyl ion (H+) that shifts the pH of the surrounding solution and this correlates directly to the number of nucleotides incorporated in that particular base flow cycle (reviewed in [Niedringhaus, #1138]. This change in pH is detected by a sensor at the bottom of each well, converted to a voltage and digitalized by semi conductor CMOS integrated circuits [Pennisi, #1139]. Signal processing software is used to convert the data for measurement of base incorporations in that flow using a physical model [Rothberg, 2011 #174]. The final sequences generated, after processing, have the read length up to 200 bases (lower than Roche/454) but like Roche/454, Ion torrent sequences have indel errors at homopolymer regions at rate of 1.5 per 100 bases [Loman, 2012 #1410].

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

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

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

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

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

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

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

In a study by Hedskog et al, UDPS was used to explore the dynamics of the HIV quasispecies using longitudinal samples collected before and after treatment [Hedskog, 2010 #386]. The study showed that the prevalence of drug resistant variants was high and wild-type variants was low when the HIV infected individuals were undergoing treatment. After treatment interruption, the author also detected drug sensitive HIV variants that were not present before treatment suggesting that the sensitive variants emerged through continued evolution of the drug resistant variants [Hedskog, 2010 #386].

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

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

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

Because of the PCR and sequencing steps a HIV variant can be sequenced multiple times. However identifying all of the sequence reads generated from the same virion would enable much more accurate characterization of viral diversity and the true prevalence of resistant variants. A unique identifier can be annealed to cDNA of a HIV genome during the reverse transcription step, which gets copied to subsequent amplicons of the cDNA during PCR amplification [Jabara, 2011 #1188]. While each individual sequence read from UDPS might contain an error, generation of a consensus sequence from reads from the same viral template would result in removal of these PCR adn sequencing-induced errors thereby capturing the true viral diversity. Based on this method, a Primer ID technology has been introduced for accurate sampling and genotyping of HIV variants from the viral quasispecies [Jabara, 2011 #1188]. With the application of the Primer ID technology, Jabara et al [Jabara, 2011 #1188] were able to to resolve 80% of the unique sequence polymorphisms that were different than consensus sequence from conventional Sanger genotype method.

## Thesis Outline

Chapter 2: The chapter introduces and describes, in detail, a novel algorithm QTrim for quality trimming of Roche/454 ultra-deep high throughput sequence data. QTrim is optimized for both poor and high quality data. HIV-1 resistance test requires high quality genotypic data and QTrim is designed to quality control the data for the test. QTrim is compared to other widely used quality trimming tools and the comparative analysis result is also presented. The entire chapter is produced as a paper for publication entitled “QTrim: A novel tool for the quality trimming of sequence reads generated using the Roche/454 sequencing platform” (Shrestha, RK and Travers, Simon; being reviewed).

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

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

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

CHAPTER 2

# QTrim – A Novel Algorithm for Quality Trimming HTS Sequence Data

## Introduction

The invention of high-throughput sequencing (HTS) technologies, such as Roche/454 pyrosequencing, has revolutionized the field of virology (Antonelli, 2013). The current Roche/454 pyrosequencing technology allows for the generation of as many as one million high quality sequence reads with read lengths of up to 1000 base pairs (<http://www.454.com>). This technology therefore provides unprecedented sampling depth to study highly heterogeneous HIV quasispecies (Beerenwinkel and Zagordi, 2011). Since large volumes of sequence data are produced, the data quality has to be high because the manual curation of quality and sequencing errors, as could be done with traditional Sanger sequencing, is no longer feasible. One way to ensure high quality data is repetitive HTS of a genomic region generating large amount of data, resulting in higher coverage per base that compensates for the lower quality bases to a certain level. However, sequencing from a sample of HIV quasispecies would mean that every sequence read could represent a unique variant. Therefore, each sequence read has to be quality controlled, independent of other sequence reads from same genomic region.

One of the major limitations of pyrosequencing is that sequence quality is not consistent, either within a read or between reads generated in the same sequencing run (Huse et al., 2007) and, thus, downstream analysis of such data may be compromised as a result of low quality data (Mardis, 2008). The quality scores for the current generation 454 sequencing platforms are similar to PHRED scores (Ewing and Green, 1998) and represent the probability of a base call error at each individually base in a read (Brockman et al., 2008). These quality scores range from 0 to 40 and are log-scaled (Cock et al., 2010), meaning that scores of 30 and 40 represent a probability of an incorrect base call of 1 in 1000 and 1 in 10000 respectively. As with most sequencing approaches, the quality of sequence data generated using 454 pyrosequencing decreases linearly across a sequence read (Gilles et al., 2011; Suzuki et al., 2011). The identification of a true base with a high quality score is pertinent, particularly in HIV drug resistance studies where low quality sequence data might represent artificial viral mutations (Kunin et al., 2009) affecting resistance test on the whole. Thus, in many instances it is imperative to undertake quality filtering of 454 sequence data to remove those low scored bases prior to subsequent analysis. Quality trimming generally entails some form of iterative removal from one or both ends of a sequence read with the primary goal to ensure that the resultant read is of high quality. Quality trimming methods range from strict approaches that have zero tolerance of low quality base calls in the output reads (Delport et al.; Gianella et al., 2011) through to averaging approaches that allow the inclusion of a proportion of low quality base calls within an output read (Chou and Holmes, 2001; Schmieder and Edwards, 2011). Algorithms like PRINSEQ (Schmieder and Edwards, 2011), Geneious (Kearse et al., 2012) and LUCY (Li and Chou, 2004) that use averaging approach are available but the output reads have large number of poor quality bases or large number of reads are discarded. We have developed a quality trimming algorithm (QTrim) that uses a novel averaging approach to minimize poor quality bases and maximize the output of high quality reads from 454 sequence data. To enable its use by a broad range of researchers, QTrim is available as a standalone python executable script for individuals with computational expertise and as a web-interface for individuals with little, or no, bioinformatics experience.

## Methods and Materials

QTrim is a python quality trimming bioinformatics tool and takes as input a fastq file or a fasta file with its associated quality (.qual) file. If a combination of fasta and qual files are submitted, they are converted to fastq format. QTrim reads sequences from the input file one at a time using biopython package ([www.biopython.org](http://www.biopython.org)) (Cock et al., 2009). QTrim trims the input sequences based on the nucleotide quality score. The first step in QTrim is removal of any ambiguous base (Ns) that have quality score of zero at the 3’ end of sequences. It is then, followed by three sequential trimming steps detailed as below before final output of the clean reads (Figure 2.1):

1. QTrim checks if a sequence read is greater or equal to the minimum required sequence read length. If this criteria is satisfied then the mean quality of the bases across a read is checked. If the mean quality is less than threshold quality, a single base is trimmed out from 3’ end. This process is looped until the mean quality score across the read satisfies the quality threshold (Figure 2.2 A). The resulting read is discarded if it does not satisfy the required minimum sequence read length.
2. For a sequence read that satisfies both minimum read length and mean quality in step 1 above, further trimming is done using a sliding window approach that evaluates the mean quality score of the last N number of nucleotides at 3’ end (N is equal to the window size). If the mean quality score of the bases within the window is less than the required mean quality, a single base is trimmed out from the 3’ end and the window is reset (Figure 2.2 B). The process is repeated until the mean quality in the window is satisfied.

The sequence is discarded if the read length is below the minimum read length although there can be further trimming.

1. In the third and the last trimming step, QTrim checks the quality score of the last nucleotides from the 3’ end. QTrim iteratively trims out the last nucleotide until the quality score of the last nucleotide in the sequence is greater or equal to quality threshold value (Figure 2.2 C). Finally, the read is saved in an output file.

For desired optimal quality trimming, the user is required to set three parameters – mean quality across a sequence read, minimum output sequence read length and sliding window size. The quality threshold is the mean quality that each trimmed read must satisfy, the second defines the minimum allowed read length (base pairs) a read can reach during trimming before being discarded, while the final parameter (optional) defines the window size to be used during trimming. If no sliding window size is defined at input the default value is set to the user-defined minimum allowed read length.

The default mode of QTrim execution trims the poor quality bases from the 3’ end of sequence reads, ignoring any ambiguous bases (Ns) interspersed among the high quality bases in the reads. Depending upon simultaneously trimming from 5’ and 3’ ends, trimming only from 3’ end and ignoring or removal of interspersed ambiguous Ns in the reads, there are four modes of QTrim execution.

1. Mode 1: Trimming from 3’ end with removal of interspersed ambiguous bases (Ns) in the reads.
2. Mode 2: Trimming from 3’ end without removal of interspersed ambiguous bases (Ns) in the reads. This is the default mode.
3. Mode 3: Trimming from 5’ and 3’ ends with removal of interspersed ambiguous bases (Ns) in the reads.
4. Mode 4: Trimming from 5’ and 3’ ends without removal of interspersed ambiguous bases (Ns) in the reads.

### Graphical plots in QTrim

QTrim uses matplotlib (Hunter, 2007) and numpy to generate the following analytical plots; each plot is produced for both the raw and trimmed data

1. **Distribution of number of reads by mean quality across the sequence read**

QTrim calculates the average quality of every sequence read. The number of sequences representing each possible quality score is plotted as shown in figure 2.3. The plot gives an overview of the data quality with the spread of mean quality. Before trimming good quality data generally has a large number of sequence reads with high mean quality whereas poor quality data will have large number of sequence reads with low mean quality.

1. **Distribution of number of reads by sequence read length**

QTrim calculates the length of every sequence read in a dataset and the number of sequences in each group of sequence lengths is then plotted as shown in figure 2.4. The plot gives an overview of the data quality with respect to sequence length. In general the read lengths for a high quality dataset will all be approximately the same length whereas the read lenths for poor quality data are generally very variable.

1. **The trend of the mean quality score at an interval of 10 bases**

In order to show the range average quality scores across every read in a dataset the mean quality score is calculated from every 10th base position of every read. A box and whisker plot showing this information is plotted as shown in figure 2.. The plot shows variation in the quality score of nucleotides with increase in the sequence read length. The good quality data will have consistent high mean quality before the quality drops off (at around 400 bases for a Roche/454 Junior plate and 1000 bases for a FLX+ plate) while poor quality data will have steep drop in mean quality.

### Test Data

Two previously sequenced Roche/454 datasets (A and B) were chosen to assess QTrim’s quality trimming approach and compare it with other widely used methods. Dataset A (high quality) has only 9% of sequences with an average quality score below 20 while the majority of sequence reads in dataset B (poor quality) have a mean quality score less than 20 (Figure 2.6 A). Further, dataset A exhibits a small range of untrimmed sequence read lengths (Figure 2.7A) while the wide range of untrimmed sequence lengths is further evidence of the poor quality data present in dataset B (Figure 2.7B).

The poor quality data was obtained from amplicon based ultra deep sequencing of the HIV reverse transcriptase gene. The original study aimed to characterize the emergence and persistence of drug resistant mutations in HIV-1 subtype C infected individuals from the Karonga district in Malawi (Bansode et al., 2013). The good quality 454 sequence data originates from a metagenomic project by collaborators sequencing bacterial 16s genes from seawater sponges (manuscript in prep).

Both the good and poor quality test datasets were quality trimmed using QTrim at two quality threshold levels: mean quality in sequence reads of 20 (Q20) and 30 (Q30) with a minimum read length of 50 set for both runs. The datasets were also trimmed using other widely used methods, including PRINSEQ (Schmieder and Edwards, 2011), the Modified-Mott algorithm implemented in Geneious (Kearse et al., 2012), Roche/454 Newbler v2.6, FASTX (Blankenberg et al., 2010) and the Lucy algorithm (Chou and Holmes, 2001; Li and Chou, 2004) implemented in clean\_reads (Blanca et al., 2011). Apart from Newbler v2.6, which is preset to trim at Q20, all other methods were executed at Q20 and Q30 with a constant minimum read length of 50. The performance of QTrim was compared to the above mentioned tools on the basis of the total number of reads in the output, longest average read length in the output, number of poor quality bases in the output, and time of execution. The best tool in the comparison should generate the highest number of trimmed sequencing reads that satisfy the quality threshold with the longest average read length.

## Results

When applied to the good quality dataset, QTrim and PRINSEQ performed at an equivalent level (Figure 2.7 A and C; Table 2.2 Q20 and Table 2.3) and outperformed all the other methods (Figure 2.7 A and C), with 15829 trimmed reads with a mean length of 448 nucleotides output by QTrim and 15825 trimmed reads with a mean length of 450 nucleotides output by PRINSEQ in the Q20 threshold analysis. In terms of the percentage of total bases in the output, again QTrim and PRINSEQ outperformed the other methods while they both allowed some poor quality bases (**Table 2.1**).

In the more stringent Q30 analysis, the number of produced reads remained similar to that of the Q20 analysis (**Table 2.1**) however the mean read length reduced to 422 and 426 nucleotides for QTrim and PRINSEQ respectively. The percentage of bases and poor quality bases in QTrim and PRINSEQ output are reduced as well; PRINSEQ has one percent bases greater than QTrim while QTrim has half percent poor quality bases less than PRINSEQ (**Table 2.1**). For both the Q20 and Q30 analysis all of the other approaches produced a comparable number of trimmed reads to QTrim and PRINSEQ, however the average read lengths were significantly shorter (**Figure 2.8 A and C**).

When applied to the poor quality data, PRINSEQ and QTrim were, by far, the two best performing approaches (**Figure 2.8 B and D**). A total of 32818 trimmed reads with a mean length of 273 nucleotides was produced by QTrim and 32381 trimmed reads with a mean length of 282 nucleotides was produced by PRINSEQ in the Q20 threshold analyses. The lower quality of this data is reflected in the much shorter trimmed reads produced and over 50% of bases trimmed out by all methods (**Table 2.2**) from this analysis when compared to the trimmed read lengths produced during the analysis of the good quality data. The performance was further evident when the stringent Q30 analysis of the poor quality data was undertaken. The average trimmed read length reduced from 273 nucleotides (Q20) to 162 nucleotides (Q30) for QTrim and from 282 nucleotides (Q20) to 176 nucleotides (Q30) for PRINSEQ. PRINSEQ has the highest percentage of poor quality bases (**Table 2.2**). Further, the dramatic reduction in the number of bases produced for all methods in the Q30 analysis (ranging from a 80% reduction in the number of bases Q30 analysis in QTrim, to 99.8% reduction in FASTX (**Table 2.2**)), indicates that, for many reads, the sequences were of too low quality to pass the minimum read length threshold.

Finally, when compared by execution time, QTrim is twice as fast as PRINSEQ (379372 versus 189966 bases trimmed per second) while most other methods a faster (Figure 2.9) on a standard desktop computer with a 2 GHz Intel**®** Core**™** Duo CPU and 2GB of RAM.

### QTrim Web Service

A QTrim web service has been developed to facilitate quick and easy quality trimming of HTS sequence data for researchers with little knowledge of command line tool execution. The web service is available at <http://hiv.sanbi.ac.za/tools#/qtrim> (Figure 2.10). The web service users have an option to register and create an account. Registration is free for academic users whereas commercial users need to pay. Account holders can login using their userid/email and password after registration while non-account holders can also use the service as normal but the results are not saved in the server, which means only the account holders can retrieve the interested results in the future.

On the QTrim job submission page (Figure 2.11), users can upload the HTS sequence data files, provide a job name and submit the job straight away with default parameter settings or change the parameter settings like mean quality, minimum read length, mode of trimming in the advance settings and then submit the job for quality trimming.

Users can view their job details using the job name. The plots for range of quality scores across reads, mean quality scores of reads and read length distribution generated for the uploaded raw data and trimmed result data will be displayed for the selected job. Users can also download the trimmed results files including the plots from links on the result web page of the selected job. The web service users who don’t create account

## Discussion and Conclusion

QTrim is a novel algorithm implementing an averaging approach for sensitive quality trimming of 454 sequence data. The algorithms used by averaging approaches differ greatly, ranging from approaches such as clean\_reads (Blanca et al., 2011) and PRINSEQ (Schmieder and Edwards, 2011) that use a window-based approach to iteratively trim sequence reads until the user-defined quality threshold is satisfied within the window, to FASTX (Blankenberg et al., 2010) that iteratively trims nucleotides from a sequence read until the percentage of low quality bases in a read satisfies a user-defined threshold. While all of the reads in such approaches will satisfy the mean quality score threshold, the algorithms used can result in tools that ‘over-trim’ reads resulting in the loss of data that, if included, would be both high quality and informative.

Upon comparison with the other approaches, QTrim performs equally as well as the best of these methods (PRINSEQ (Schmieder and Edwards, 2011)) on the basis of total output reads and the mean output read length. The trimmed reads produced by PRINSEQ are, on average, slightly longer than those from QTrim. Upon further examination, however, this is as a result of PRINSEQ allowing a higher number of low quality bases (quality score < 20) at the 3’ end of its trimmed read. For example, PRINSEQ generates 8% more low quality bases than QTrim in the Q20 trimming of both datasets tested here, and 17% and 25% more low quality bases in the Q30 trimming analyses of the poor quality and good quality datasets respectively. We find that this is the case in all of the methods that use an averaging approach for quality trimming. As soon as the minimum quality score in a read satisfies the quality threshold, the read is defined as trimmed without any further analysis. In QTrim, however, we employ two further steps, which ensure that low quality bases at the 3’ end of quality trimmed reads are removed. Thus, while the reads may be slightly shorter than those produced by PRINSEQ, users can be confident that the quality of the generated reads is consistent across the length of the quality trimmed data produced by QTrim.

QTrim performance comparison with other approaches is distinctive when the trimming is done on poor quality data. While most other tools discard almost all reads, QTrim is able to “rescue” high quality nucleotides more than PRINSEQ, which has comparable total reads output and mean read length. This means QTrim enables sensitive trimming of sub-optimal sequence data thereby enabling researchers to undertake downstream analysis on lesser quality sequence data that otherwise may have been discarded.

Fastx (Blankenberg et al., 2010) exhibits the highest speed of execution and this is because it trims out all nucleotides once the required percentage of high quality nucleotides is obtained. Other methods, including QTrim that check the mean quality of nucleotides trim single base at each time. Although QTrim is slower than Geneious (Kearse et al., 2012), it is comparable with Clean\_reads (Blanca et al., 2011) and Roche’s Newbler v2.6 and as much as twice the speed of PRINSEQ.

From users prospective, QTrim is simple to use with less parameters to define while other methods require complex combinations of parameters to be defined. QTrim is available as standalone executable file with required library files of Biopython (Cock et al., 2009), matplotlib (Hunter, 2007) and numpy, which allows users to extract and execute straight away without installation of any secondary software. The executables are available for Linux and MacOSX and are downloadable from <https://hiv.sanbi.ac.za/software/qtrim>. This makes QTrim to be easily integrated into next generation sequence analysis pipeline.

**Chapter 4**

# Seq2Res: A computational tool to facilitate HIV drug resistance genotyping using high-throughput sequencing

* 1. Introduction

UNAIDS estimates that approximately 34 million individuals - including 2.6 million newly infected - were living with HIV and as many as 1.6 million died of HIV-related illnesses in the year 2011 alone (UNAIDS, 2012). The Sub-Saharan African region has the highest prevalence of HIV infections (Asamoah-Odei et al., 2004). Antiretroviral therapy (ART) programmes have been established over the entire region (Blower et al., 2005; Herbst et al., 2009; Nattrass, 2006; Stringer et al., 2006) with the aim of suppressing viral replication, resulting in a lower viral load (Autran et al., 1997; Li et al., 1998; Mocroft et al., 2010) and thereby extending the life expectancy of HIV positive individuals (Fang et al., 2007; Harrison et al., 2010; Mills et al., 2011). By the end of 2011, 8 million people from low and middle-income countries were receiving the life saving drugs (20 times more than in 2003) (UNAIDS, 2012; WHO, 2011).

In the order of 1010 new viruses are produced per replication cycle with a mutation rate in the order of 10-5 per nucleotide per cycle (Ho et al., 1995; Mansky, 1996a, b, 1998; Mansky and Temin, 1995). The high mutation rate is due to the error prone reverse transcriptase enzyme that transcribes viral RNA to cDNA during the replication (Bebenek et al., 1989; Berkhout et al., 2001; Preston et al., 1988; Roberts et al., 1988). The accumulation of random mutations in HIV leads to development of drug resistance (Johnson et al., 2008).

With the scale up of antiretroviral (ARV) drugs there is growing evidence suggesting that drug resistant HIV can exist as minor variants in individuals undergoing treatment (Aghokeng et al., 2011; Dudley et al., 2012; Gupta et al., 2012; Lataillade et al., 2010; Li et al., 2011; Little et al., 2002; Simen et al., 2009; Yebra et al., 2011). This observation is also supported by several research studies on population-level surveillance of HIV drug resistant variants (Dudley et al., 2012; Hamers et al., 2012; Hamers et al., 2011a; Hamers et al., 2011b; Lataillade et al., 2010; Li et al., 2011; Little et al., 2002; Simen et al., 2009). The most likely reasons for the increase in drug resistant viral variants in the general HIV-infected population is poor adherence during therapy (Bangsberg et al., 2003; Golin et al., 2002; Low-Beer et al., 2000) and an increased rate of transmission of resistant viral variants (Hamers et al., 2011a; Jakobsen et al., 2010; Supervie et al., 2010; Wittkop et al., 2011; Yerly et al., 1999). Thus, ARV treatment, management and surveillance of drug resistant HIV variants (Bennett et al., 2009) is essential for prolonging the usage of, and optimizing the outcome from, a particular drug cocktail (Adetunji et al., 2012). Therefore, the World Health Organization (WHO) guideline ([http://www.**who**.int/**hiv**/drugresistance/](http://www.who.int/hiv/drugresistance/)) recommends that a pretreatment HIV drug resistance test is carried out, but this has only been possible in resource-rich countries (Aberg et al., 2009; Clumeck et al., 2008; Gazzard et al., 2008; Kaplan et al., 2009; Masur and Kaplan, 2009; Perfect et al., 2010).

Both individual and population-level screening of resistant HIV variants must be done routinely for tracking of resistant viruses and to prevent the exhaustion of ART treatment options. However, due to the cost factor, regular pretreatment resistance testing in low and middle-income countries, was not possible until the advent of ultra deep pyrosequencing (UDPS). UDPS is a robust, reliable and affordable way to explore clinically relevant low abundance (< 1%) resistant HIV variants (Dudley et al., 2012; Ji et al., 2012; Tsibris et al., 2009; Wang et al., 2007). However, the huge volume of sequence data from UDPS is a challenge for data analysis and management.

This chapter describes the development and testing of a computational tool designed to facilitate low cost HIV drug resistance test using UDPS technologies.

* 1. Methods and Materials

While UDPS approaches provide an exciting prospect for enabling high-throughput, low-cost HIV drug resistance genotyping, the sheer volume of data generated by such sequencing platforms means that the bioinformatics requirements for the management, analysis and interpretation of the data is immense. The use of UDPS for routine HIV drug resistance genotyping requires a bioinformatics platform that can facilitate fast and sensitive analysis of data by individuals, such as clinicians and wet-lab researchers, with little, or no, bioinformatics expertise.

Further, the rapidly evolving field of high-throughput sequencing means that any bioinformatics platform must be robust and easy to adapt to analyze data from new sequencing platforms.

Thus, Seq2Res has been developed (mostly using the Python high level programming language) in a modularized format, keeping each distinct analysis block independent of the other, thereby enabling easy insertion of new modules to allow the analysis of new data-types. Given that the vast majority of HIV resistance data generated to date has been produced using the Roche/454 platform, Seq2Res has been initially developed to analyze such data.

* + 1. Structure of UDPS raw sequence reads

Seq2Res analyses the raw sequence reads generated by the sequencing instrument before any preprocessing of the data is done. All raw sequence reads in a file are in the 5’ to 3’ end orientation. A raw UDPS sequence read consists of key sequence, an MID (Multiplex Identifier) sequence to label a sample, a primer sequence to identify a specific amplicon region and the actual amplicon region sequence from 5’ to the 3’ end (**Figure 4.1**).

The key sequence consists of four nucleotides that are used by Roche/454 high throughput sequencing platforms to calibrate the measurement of optical emission to count the number of nucleotides added to the growing strand during sequencing. The key sequence is removed by Seq2Res before beginning subsequent analysis.

On a raw sequence read the key sequence is followed directly by the sequence pertaining to the Multiplex Identifier (MID) that facilitates the identification of a sample. The Roche/454 platform has approximately 150 standard MID sequences that can be used. The MID sequences are usually ten nucleotides long, however, MIDs can be designed as any unique custom sequence.

The MID sequence is followed by a short sequence to identify the genomic amplicon region called a primer. A genomic region can be sequenced in both directions and therefore the primer can be either in a forward or reverse orientation. Primer sequences are designed to bind specifically at the 5’ region of its associated genomic region’s forward (forward primer) and reverse (reverse primer) strands.

The primer sequence is followed by a genomic amplicon sequence. This is the sequence that is of interest and is analyzed. The primer in forward or reverse orientation is used to know that the genomic amplicon sequence read is from the forward strand or the reverse strand. An amplicon sequence reads from reverse strand are reverse complemented to convert to its complementary forward sequence. Therefore, all the subsequent analyses are done on the forward strand sequence reads.

* + 1. Seq2Res: Required Data

One of the primary focuses in the development of Seq2Res was to make usage as easy as possible by keeping the required number of files and parameters to a minimum. Users are therefore only required to provide the raw sequence data file, the primer information (multiple amplicons can be analyzed concurrently) and, if present, the MID used together with their associated patient identifier.

* + - 1. Raw sequence reads file

The raw data file can be in a number of formats:

1. The standard flowgram format (sff) file.
2. FASTQ format
3. FASTA format together with the associated QUAL file.
   * + 1. Primer file containing amplicon primers

The primer file is a five column, tab-delimited file containing the name of each amplicon in the first column, the forward and reverse primer sequences in the second and third column, and the amplicon start and end positions relative to the HXB2 HIV *pol* reference sequence in the fourth and fifth columns, respectively (**Figure 4.2 A**).

* + - 1. MID file containing sample identifiers

The MID file contains two tab-delimited columns of data of which the first column contains the Roche/454 standard MID name (or the actual nucleotide sequence of the MID used) with the second column containing the user-defined sample identifier associated with that MID (**Figure 4.2 B**).

* + - 1. Threshold Prevalence cutoff

The threshold prevalence cutoff is the minimum percentage of the amplified and sequenced viral sequences pertaining to an individual (MID) that are required for the prediction of resistance or intermediate resistance to a drug. If the percentage of sequence reads predicted as resistant to a drug is greater or equal to the cutoff, the viral population in the sample is called as resistant to the drug. If it is less than the cutoff, then the percentage of sequence reads predicted intermediate resistance to a drug is checked. If this is greater or equal to the cutoff value, then the viral population in a sample is called as intermediate resistant. If the percentage of intermediate resistant sequences is also less than cutoff value, the viral population in a sample is called as susceptible to the drug. The default prevalence cutoff is 15%.

In the event that multiple amplicons cover the same gene region (e.g. reverse transcriptase), the amplicon with the highest prevalence of predicted resistance (or intermediate resistance if there is no resistance predicted over the threshold) is output as the significant result for each drug.

* + 1. Advanced parameters

While, in most cases, the three required files are sufficient to run analysis in Seq2Res, advanced end-users or individuals with non-conventional data are provided with the following advanced parameters that relate to preprocessing of the data:

* + - 1. Quality-trimming parameters

These parameters are supplied for the trimming step. QTrim iteratively trims out poor quality nucleotides from the ends of a sequence, based on the mean of quality scores across the sequence. The quality trimming parameters that end-users can set are:

1. Mean quality: This is a minimum mean quality score across a sequence read. Every sequence read must satisfy this criterion for further analysis to take place. If a sequence does not satisfy the mean quality criteria, it is removed. The default mean quality score is 20.
2. Minimum read length: This is the minimum length, counted in base pairs, required for a sequence read in order for further analysis to take place. If a sequence does not meet the minimum read length cutoff, it is removed. The default minimum read length is 50 base pairs.
3. Mode of trimming: Users can set two modes of trimming, which are a) trimming from 3’ end, or b) trimming from 5’ and 3’ ends. The default is trimming from 3’ end
   * + 1. Demultiplex parameters

The demultiplex parameters refer to the primer and MID tolerance, as well as the key sequence length.

1. Primer tolerance refers to the number of nucleotide mismatches between the user-supplied primer and the primer in a sequence read. For an amplicon of a sequence to be identified, a primer has to appear in the sequence. But due to sequencing errors, the primers may not appear exactly as supplied in the primer file. The primer tolerance allows the amplicon of the sequence to be identified although there are some errors, less or equal to, the primer tolerance. The default primer tolerance is 3.
2. MID tolerance refers to the number of nucleotide mismatches between the user supplied MID and the MID in sequence read. Similar to the primer tolerance, some errors in an MID sequence can be tolerated as Roche/454 have designed their MIDs to have a tolerance of 2 (the defauls setting)
3. Key sequence length: This is usually a nucleotide sequence of 4 unique bases at the 5’ end of a raw sequence read. The key sequence is not a part of sequence reads and is removed. The default value is 4. If no key sequence is present in sequence reads, a value of zero must be supplied.
   * 1. Development and processing of data using Seq2Res

The steps undertaken by each module of Seq2Res are described below. The output of each module serves as the input for the next and, thus, makes for easy swapping/replacement of modules in future versions (**Figure 4.3**).

1. **Pre-processing of submitted files**

Depending on the format of the input file, Seq2Res may need to first pre-process the file. Since the subsequent steps in the Seq2Res pipeline work only with FASTQ files, the other formats are converted to FASTQ format during the pre-processing. An SFF file is converted to a FASTQ file with a tool called sff\_extractor (<http://bioinf.comav.upv.es/sff_extract/index.htm>) while a FASTA file with a paired quality scores file is merged to a FASTQ file.

1. **Processing optimal full length positions**

The start and end nucleotide positions that defines the full length of the amplicons are supplied in the primer file. Seq2Res processes these supplied positions to associate the start and end codon positions with the standard HIV *pol* reference sequence. Because Seq2Res only considers codon positions that contribute to drug resistance in the amplicons, the start and end codon positions for full length are processed to find the first and last drug resistant codon positions in the amplicons (**Figure 4.4**). The start and end codon positions for the amplicons are redefined by the first and last drug resistant codon positions in the amplicons respectively and this new start and end codon positions are defined as the optimal full-length positions. Seq2Res considers all the amplicon sequence reads covering optimal full-length positions for downstream analysis. This is a critical step as there may be a large number of sequence reads that are not full length according to the amplicon borders but satisfy the optimal full-length criteria.

1. **Demultiplexing**

Seq2Res utilizes Biopython ([www.biopython.org](http://www.biopython.org)) (Cock et al., 2009) packages to read individual sequence reads from the FASTQ file and removes any key sequence, if present, from the sequence reads. For the first step of the demultiplexing Seq2Res searches for the MID in every sequence read which is located at the 5’ region of a sequence read. A subsequence, of length that is equal to MID length, is obtained from 5’ end of the sequence read. The subsequence is then pair-wise global aligned with every MIDs in the input list of MIDs. The MID with the number of mismatches less or equal to the MID tolerance in the alignment (default MID tolerance 2) identifies the sample of the sequence. The MID is added at the sequence identifier and the MID subsequence is deleted from the sequence read. Thus, every sequence read is either tagged with the details of the MID in the sequence identifier or is discarded from subsequent analysis.

Next, Seq2Res searches for the primer in each sequence read in a similar way to identify the MID. A subsequence equal to a primer length is obtained from 5’ end of the sequence read. The subsequence is then pair-wise globally aligned with every primer in the input list of forward and reverse primers. The primer, either forward or reverse, with the number of mismatches less or equal to user threshold of primer tolerance (the default primer tolerance is 3) is selected and added in the sequence ID along with the strand forward (+) or strand reverse (-).

If the sequence reads are generated using Primer ID technology (Jabara et al., 2011), Seq2Res searches the primer, MID and Primer ID using the PIDA algorithm. At this stage the PIDA algorithm also generates the consensus sequences representing each primer ID and it is only these consensus sequences that are passed through for subsequent analysis.

1. **Advanced Sequence Reads Quality Control**

The non-discarded sequences are quality trimmed in the next step. During sequencing, Roche/454 UDPS calls a base with a certain probability (Margulies et al., 2005), and this score for each base is saved in the quality file. The scores range from minimum of zero to a maximum of 40. A score of 10 would mean that the probability of a wrong base call is 1 in 10, 20 means 1 in a 100, and 40 means 1 in a 10000 (Cock et al., 2010). In general, the quality scores of a base decrease as the sequence length increases. A lower quality score indicates that a base has a lower probability of being correct and including these poor quality bases would compromise the data analysis. Seq2Res uses the QTrim quality-trimming tool (<https://hiv.sanbi.ac.za/tools/qtrim>) (Shrestha et.al 2014) for removing these lower quality bases. Default parameter settings in QTrim are: mean quality score of 20, minimum read length of 50 nucleotides and trimming from 3’ end of a sequence read. Seq2Res allows flexibility to control these default QTrim parameter settings.

1. **Binning of sequence reads by amplicon and MID**

Once sequence reads have been quality trimmed, the information contained in the sequence IDs is used to bin groups of sequences on the basis of sample/MID, and then by amplicon. All subsequent analysis is performed on each ‘bin’ of sequences concurrently using a high-performance computing cluster.

1. **Reference Mapping**

The high quality sequence reads in each bin are mapped to the full HXB2 *pol* gene reference sequence. This mapping step is one of the most critical steps of the entire pipeline as it is here that we are able to correct both PCR- and sequencing-induced errors in sequence reads. RAMICS (Rapid Amplicon Mapping In Codon Space) is a tool that has been developed by Imogen Wright (a PhD student in the research group of Prof Simon Travers, manuscript currently under review in Nucleic Acids Research) that maps sequence reads to a reference sequence using hidden Markov models in “codon-space” and is capable of identifying, and accounting for, PCR- and sequencing-induced errors in sequence reads during the mapping process. RAMICS is particularly powerful at homopolymer regions, accurately identifying the over- or under-call of nucleotides. Further, RAMICS is able to identify whether each sequence read is full length and maps to the entire amplicon.

In Seq2Res, RAMICS maps amplicon sequence reads to a subsequence of the reference sequence that correspond to amplicon specific optimal full-length sequence. For example, if an amplicon optimal start and end codon positions are 66 and 100 respectively, which correspond to the HIV *pol* reference sequence, RAMICS copies the nucleotide sequence from codon 66 to 100 of the reference sequence and maps the amplicon sequence reads to it.

RAMICS provides a novel ‘fasta-like’ (fastm) output file (**Figure 4.5**) that, for each read, details what amino acid is present at each position in the read relative to the HXB2 reference sequence. RAMICS also accounts for low quality base calls when generating the fastm format. If the quality score of one or more of the bases in a DRM codon is less than the cutoff (default is 20) then that base is flagged as not sequenced (i.e. no coverage). This means that the level of resistance is not artificially inflated/deflated as a result of sequencing error. Any single or double nucleotide insertions resulting from PCR or sequencing error are, thus, excluded from subsequent analysis. Any full codon sized insertions or deletions are identified and documented in the fastm output file.

1. **Codon position translation**

For every sequence read written to the fastm file, Seq2Res translates each codon position that was produced with respect to the reference sequence (the HXB2 full *pol sequence*) into codon positions relative to the gene(s) that the amplicon covers (protease, reverse transcriptase or integrase). For example, the codon positions 57 and 156 in HIV *pol* sequence are translated to the protease codon position 1 and reverse transcriptase codon position 1 respectively.

For each read, the amino acid present at all positions evaluated as part of the Stanford HIV resistance testing algorithm (Liu and Shafer, 2006) are extracted and saved to a file in the format required for submission to the Stanford algorithm (**Figure 4.6**).

1. **Resistance Testing**

For resistance testing we use a locally installed version of the latest version (7.0) of the Stanford HIVdb resistance-testing algorithm (Sierra – kindly provided by Tommy Liu and Robert Schafer) (Liu and Shafer, 2006). For each sequence read contained in the submitted file, the Sierra algorithm produces the drug resistance result in a XML file detailing their resistance level to all drugs– Resistant (R), Intermediate (I) or Susceptible (S) as well as the relevant drug resistant mutations (DRMs) present.

1. **Resistance report presentation**

Seq2Res processes the XML file output from Sierra to obtain sequence name, DRMs and resistance level of the DRMs in each sequence. The information obtained per sequence is stored in a row in a tabulated tab-delimited file. The first ten columns of each row in the file contain user and sequence information. The order of the information in the first ten columns are USERID, JOBID, input file name, sample name, the MID of the sample, a sequence read ID, number of the sequences with that sequence read ID, the amplicon name of the sequence, strand orientation (forward or reverse) of the sequence and the length of the sequence. From column 11, a resistance level (R, I or S) to a specific drug appears at odd columns and the pertinent DRMs to that resistance call follow in each of the even columns. For example, a resistance call “R” to Abacavir (ABC) may appear at column 11 and DRM K65R from a sequence that is resistant (R) to the first drug ABC may appear in column 12. Similarly, the resistance level and DRMs for the drug didanosine (d4T) appear in column 13 and 14 respectively. This is followed for a defined order of antiretroviral drugs (**Table 4.1**). If a sequence read does not cover DRM codon positions associated with a drug, a “ dash ( - )” is output to the columns for that drug. For example, if a sequence covers only protease gene region and not reverse transcriptase gene, the table lists out resistance levels and the associated DRMs for PI drugs and “-“ for reverse transcriptase inhibitor drugs. Thus, this step generates an easily searchable table from which all other results and conclusions are generated.

From the table, the number of sequence reads that are predicted to have high resistance (R), intermediate resistance (I) or susceptibility (S) by local Sierra to a particular antiretroviral drug is calculated. The algorithm classifies an antiretroviral drug as either resistant, intermediate or susceptible to the viral population in a sample using the following conditions:

1. If an amplicon covers one gene, the percentage of sequence reads predicted resistant, intermediately resistant and sensitive are calculated. If the prevalence of resistant sequence reads to a drug is greater or equal to the prevalence cutoff, the viral population in the sample is called as resistant to the drug. If the viral sample is not resistant and the prevalence of predicted intermediate sequences reads is greater or equal to the prevalence cutoff, the viral population in a sample is reported as intermediately resistant to the drug. If the viral sample is neither resistant nor intermediate resistant, it is called as susceptible to the associated drug.
2. If more than one amplicon covers one gene, for example RT1 and RT2 amplicons for the RT gene, the following conditions are applied (Figure 4.7):
   1. If the percentage resistance for either RT1 or RT2 is above the prevalence cutoff, the sample is considered resistant to the associated drugs. The reported prevalence is equal to the prevalence of the amplicon with the highest number of resistant calls.
   2. If the percentage resistance for both RT1 and RT2 are less than the cutoff, the percentage intermediate resistance for either amplicon above the cutoff is reported (if observed above the user-defined threshold). As with the resistance prevalence, the reported intermediate resistance is equal to the prevalence of the amplicon with the highest number of intermediate calls.
   3. If the percentage resistance and intermediate resistance for RT1 and RT2 are both less than the cutoff, the sample is called as susceptible to the associated drug.

The antiretroviral drug classes (NNRTI, NRTI, PI, and IN), the drugs, the number of sequence reads showing resistance, the predicted resistance and intermediate resistance levels of the sample to each drug, as well as the drug resistant mutations associated with each drug are presented in the drug resistance report (Figure 4.8). Each row in the drug resistant report is color coded by either Red or Orange or Green. For the viral population in the sample, Red color indicates **highly resistant**, the orange color indicates **intermediately resistant** and the green color indicates **susceptible** to the associated drug. The three-color codes are on the basis of three-step resistance level presentation as used by Stanford HIV database (http://hivdb.stanford.edu/DR/asi/releaseNotes/index.html#hivalg).

* + 1. Quality Analysis in Seq2Res

Seq2Res inherits all the features of the QTrim quality-trimming tool. All the trimming results are available in Seq2Res. The demultiplexed sequence reads from every sample in a single file is trimmed and generates analytical graphical plots of the untrimmed and trimmed data for direct comparison. The plots show the comparison of trimmed and untrimmed data on mean quality of sequence reads against number of sequence reads (**Figure 4.9**), sequence read length verses number of sequence reads (**Figure 4.10**) and the trend of quality scores across sequence reads (**Figure 4.11**).

* + 1. Graphical analysis of DRM prevalence

From the RAMICS produced fastm files for every sample, the percentage of mutations at known standard drug resistant codon positions (**Table 4.2**) are calculated. The mutations that confer viral resistance to drugs are then grouped by the drug class – PI, NRTI, NNRTI and IN. A bar plot is generated for each drug class, showing the prevalence of each DRM and a red horizontal line that cuts through the plot representing the user defined prevalence cutoff for quick observation of DRMs with prevalence below or above or on the red line (**Figure 4.12**).

* + 1. Evaluating the sensitivity of Seq2Res.
       1. Confirmation of the accuracy of the locally installed version of Sierra

The first step of evaluating the sensitivity of Seq2Res was to ensure that the locally installed version of the Stanford resistance testing algorithm (Sierra) was fully functioning and sensitive on sequences of lengths that are comparable to 454 sequence reads.

To achieve this, we acquired two datasets of sample data from the Stanford Database website (<http://hivdb.stanford.edu/DR/asi/releaseNotes/index.html#sampledata>). Both datasets consisted of sequences, generated using population-based Sanger-style sequencing, with an associated genotyping result. The first dataset consisted of 2055 sequences (Dataset A) while the second consisted of 5838 sequences (Dataset B)

In order to analyze these data using Seq2Res, the sequences in both the test data were fragmented into three ‘amplicons’ per sequence with some overlap between the adjacent subsequences. The fragments corresponded to HXB2 positions 55 to 159 (PR), 50 to 298 (RT1) and 290 to 399 (RT2).

The resulting amplicons were individually mapped to the HXB2 *pol* sequence using RAMICS and then submitted to the locally installed version of Sierra. The resulting resistance calls for each linked set of amplicons were then compared with the known genotypic call for each ‘parental’ sequence.

* + 1. Test Data for simulation

While the above approach is appropriate for ascertaining if the drug resistance calls on consensus sequences with a known genotype are correct, it does not fully evaluate a pipeline developed to undertake resistance genotyping on high throughput sequencing data. Thus, we undertook a comprehensive analysis of simulated UDPS sequence data to fully test the performance of Seq2Res.

Sequences covering the entire *pol* gene from five individuals were selected from dataset B (above). The selection criteria required that each sequence had to have at least the K65R, K103N mutations in reverse transcriptase (in order to evaluate the capacity of Seq2Res to call the correct DRM at homopolymer regions) as well as at least one other reverse transcriptase mutations and one or more protease DRMs. These were saved as resistant sequences.

A ‘susceptible’ sequence was generated for each resistant sequence by reverting each of the DRMs back to wild type. Thus, the final dataset that was used to simulate the UDPS data contained ten sequences in total – five resistant and five susceptible sequences. The resistance profile of each of these ‘seed’ sequences was evaluated using the Stanford HIV resistance algorithm.

* + 1. Simulation of high throughput sequencing amplicons

Each of the ‘seed’ sequences was then fragmented into three subsequences representing an individual amplicon. These fragments corresponded to HXB2 positions 169 – 469 (PR), 466 – 812 (RT1) and 672 – 1017 (RT2)

In order to simulate the fragmented PR, RT1 and RT2, we chose a next generation sequence simulator called ART (Huang et al., 2012) as the features of tool include functions to produce sequencing-like errors. ART can emulate the Roche/454 sequencing process to simulate high throughput sequence data and include Roche/454 specific error models in homopolymer, substitution and insertion-deletion errors, as well as flexible base quality profiles. These are all customizable parameters and ART allows user-supplied sequence read profiles. A real Roche/454 sequence dataset was provided to ART to generate sequence read profiles. The profile was then used to simulate 20,000 Roche/454 sequence reads for each PR, RT1 and RT2 fragment (amplicon sequences) in FASTQ format.

* + - 1. Generation of different known prevalence of DRM data

For each amplicon from each patient the simulated sensitive and resistant sequences were randomly sampled to generate UDPS datasets containing 10,000 sequence reads with a known mix of resistant and sensitive sequences. Seven datasets were generated for each amplicon in each patient representing 0.1%, 1%, 5%, 10%, 15%, 20% and 50% prevalence of resistant sequences in the dataset (**Table 4.3**). The final simulated dataset comprised of 105 fastq files corresponding to 7 prevalence levels x 3 amplicon x 5 patients. Each of these fastq files was submitted to Seq2Res and the observed results compared with the expected results.

* + 1. Computational Resources

Seq2Res is currently housed at the South African National Bioinformatics Institute (SANBI). The computing infrastructure used to run Seq2Res consists of two Blade Servers (PowerEdge M610x) each comprising 12 core processors (Intel(R) Xeon(R) CPU X5675 @ 3.07GHz), 32-gigabyte memory and a graphical processing unit (GPUs) (NVIDIA®Tesla™ M2090).

An application programming interface (API) has been developed to run Seq2Res over the internet. A web front end that uses the API has been developed for biologists and clinicians to upload the HIV sequence data and other required files and parameters for the viral drug resistance genotyping of sequenced patients. The API can also be executed from command line executable scripts. Baruch Lubinsky, a software developer employed in the research group of Prof Simon Travers, developed the API.

* 1. Results
     1. Seq2Res running time

The running time of Seq2Res depends on several factors, including the Internet bandwith SANBI in order to download the uploaded sequence and parameter files, the number of raw sequence reads, amplicons and samples in the uploaded sequence file, and the number of jobs running on the available servers. Therefore, it is challenging to provide an exact running time for a Seq2Res process.

Nonetheless, we tested Seq2Res at the current Internet speed at SANBI (2 megabytes/second) with no other processes running on the servers, using 119 simulated files. Each file contained one sample, one amplicon and 10,000 sequence reads and we allowed a two-minute time interval between each submission using the API with command line scripting. After the completion of each job, an email is sent to the user. The start time was noted as the time of job submission and the time of job completion was noted as the time recorded in the email of each process. We observed that the average time for each job execution was three minutes. However, the execution time increased with an increase in samples and amplicons in the input raw sequence file. We also executed Seq2Res locally in the SANBI computational infrastructure on Roche/454 Junior and FLX plates, bypassing the internet connection. The raw sequence files from the Junior plates had 48 samples, with each sample containing three amplicons. The FLX plate consists of eight sections and a raw sequence file from each section had eight samples, with each of the samples containing three amplicons. The execution time for a file from Roche/454 Junior plate required on average of 30 minutes and a file from a section of Roche/454 FLX plate required on average of 15 minutes.

* + 1. Comparison of Mutation and Resistance Level Calls using the Sierra web service and Seq2Res

For every sequence in dataset A and dataset B \*containg consensus sequences retrieved from the Stanford database), the DRMs reported using the Sierra web service (Liu and Shafer, 2006) were compared to those reported for the same sequence analyzed by Seq2Res. The one to one comparison of DRMs in every sequence reported by Sierra web service and RAMICS as part of Seq2Res, showed that except for a single drug resistant mutation in reverse transcriptase codon position 236 in a sequence from dataset A, there was a 100% congruency in DRMs calls. In the sequence from dataset A, Sierra web service reported amino acid Leucine (236L) while RAMICS reported wild type amino acid Proline (236P) at the discordant codon position.

* + 1. Drug Resistant Mutations in the selected test sequences for simulation

The ten ‘seed’ sequences used for simulation were analyzed using web sierra. The five ‘resistant’ sequences were confirmed to have multiple DRMs, many of which convey resistance to various drugs DRMs (**Table 4.4**). The five ‘sensitive’ sequences, on the other hand, were confirmed to contain no DRMs (**Table 4.4**).

The five resistant sequences with DRMs were observed to be highly resistant to most of the antiretroviral drugs while the five sequences without any DRMs were observed to be sensitive to the antiretroviral drugs (**Table 4.5**).

* + 1. Quality Trim analysis of simulated data

The quality trimming report for the 105 simulated datasets showed that there was no sequence reads discarded in any dataset as a result of poor quality. Because the same simulator ART using the same quality profile produced all the simulated datasets, we analyzed the quality of one of the 105 simulated datasets. The selected dataset had 2985786 bases before trimming and 2979626 bases after trimming. 6160 bases were trimmed out of 10,000 sequences. On average, less than a base (0.6162) was trimmed per sequence. The mean quality of the sequence reads in the dataset did not seem to change before trimming and after trimming. The read mean quality score was observed to be between 30 and 32 in both untrimmed and trimmed state (**Figure 4.13**). The median quality score at every 10th base position from all sequence reads was observed to be above 30 across the sequence reads (**Figure 4.14**). Similar quality scores in sequence reads were observed in the other 104 simulated datasets.

* + 1. Optimal codon positions of the amplicons in the simulated datasets

The codon positions of the nucleotide start and end positions corresponding to the reference sequence for PR were 57 and 157, for RT1 were 156 and 271 and for RT2 were 224 and 339. Seq2Res processed these start and end codon positions of the full-length amplicons to get the first and last DRM codon positions in the amplicons. These first and last DRM codon positions in the amplicons are the optimal full-length codon positions and the sequence in between the positions covers all DRMs of interest in that amplicon. The optimal full-length start and end codon positions obtained for PR was 66 and 149, for RT1 was 195 and 270 and for RT2 was 224 and 336 (**Table 4.6**). In further downstream processing, the amplicon sequence reads that extend from optimal start to end codon positions are considered although they are not necessarily the full-length amplicon.

* + 1. Prevalence of known drug resistant mutations

Each simulated dataset had a defined proportion of resistant and sensitive sequences and therefore the prevalence of DRMs in the dataset is known prior to the analyses. The analyses of the simulated data using Seq2Res showed that the observed prevalences of the DRMs were, in all cases, almost identical to that of the expected prevalence (**Figure 4.15 and Table 4.7**).

* + 1. Resistance calls for HIV sequences to antiretroviral drugs

All the simulated datasets (105 datasets), each containing one amplicon, were analyzed with Seq2Res using the default prevalence cutoff of 15%. As expected, when using the default prevalence cutoff of 15% we observed that all the viral sequences were predicted as sensitive to all drugs for the simulated datasets containing DRMs at a prevalence below 15% (the 0.1%, 1%, 5% and 10% datasets) (**Table 4.8**),

In the simulated datasets containing resistant viral sequences at a prevalence of 15% all sequences were predicted as sensitive to PIs for samples 56252, 63377, and 4134 (**Table 4.9**) while intermediate and resistant drug resistance calls were observed for samples 21354 and 2368 (**Table 4.9**). Most viral sequences were predicted as resistant or intermediate resistant to NRTIs, resistant to EFV and NVP (NNRTIs) and sensitive to ETR and RPV (**Table 4.9**).

In the simulated datasets containing resistant viruses at a prevalence of 20% or 50% we find that the vast majority of the genotyping calls for all drugs showed a prediction of either resistance or intermediate resistant (**Table 4.10**) significantly correlating with the known resistance profile of the data (**Figure 4.15**).

* + 1. Seq2Res web Application Programming Interface (API) and web outputs

The Seq2Res web interface (available at <http://hiv.sanbi.ac.za/tools/#/seq2res>) (**Figure 4.16**) has been created to enable easy execution of the pipeline for HIV drug resistance genotyping for users with little or no bioinformatics experience.

A click on “Submit job” takes users to the Seq2Res job submission page (**Figure 4.17**). Users can provide a job name and upload a raw sequence file, primer file containing the forward and reverse primer details and MID file containing sample specific tags sequences. While the required parameters are kept to the bare minimum on the initial website to avoid confusion, users can also set a number of other parameters for the analysis in the advanced options. Users are informed by email about the completion of their job.

Clicking the “My Jobs” button at Seq2Res homepage takes users to a page containing the list of all the jobs that the user has submitted (**Figure 4.18**). Job details like the name of job, the date of job submission and the status of the job processing - “complete” or “pending” or “error” while processing.

A click on a job from the list of jobs initially displays two plots that show the overall sequence analysis results of all samples in the input file for that job. The first plot in the result page shows the number of sequences discarded in each step of Seq2Res processing and the number of sequences that are passed in downstream analysis for making final result (**Figure 4.19**). The second plot shows the number of sequences in forward and reverse strands per amplicon per sample on which HIV drug resistance genotyping was performed (**Figure 4.20**).

A list of individual patient results in the form of links are listed down the results page, below the above mentioned two plots. A click on a green “show” button at the left of each patient and MID name takes users to a drug resistance result page of the selected patient. A table at the top of the sample specific result page shows the drug resistant report of the selected patient (**Figure 4.21**). The susceptibility of drugs for the patient is shown by color code – Red: Resistant, Orange: Intermediate resistant, Green: Susceptible (**Figure 4.21**).

Below the table in a patient specific result page, there are four bar plots – one plot for each drug class – NRTI, NNRTI, PI and Integrase Inhibitors (INs). Each bar plot shows the prevalence of drug class specific DRMs, the DRMs and the coverage (number of reads) at each DRM codon position (**Figure 4.22**).

* 1. Discussion and Conclusion

HIV drug resistance testing is essential to characterize the viral population (Baba et al., 2005; Simen et al., 2009) and to treat HIV infected individuals with the correct combination of antiretroviral drugs to suppress the viral replication for longer time periods and thereby increase life expectancy (Harrison et al., 2010; van Sighem et al., 2010). The Roche/454 UDPS technology has shown great potential to genotype even the minor HIV variants that are clinically relevant (Lataillade et al., 2010; Simen et al., 2007; Simen et al., 2009; Varghese et al., 2009). However, the Roche/454 UDPS platform currently generates up to a million sequences and, thus, manual analysis at any processing step is not feasible. Seq2Res computational pipeline is designed to analyze this huge amount of data for HIV drug resistance testing and produce sample specific drug resistance genotyping reports.

* + 1. Optimal full-length

HIV drug resistance is strongly associated with the pattern of DRMs (Bennett et al., 2009; Johnson et al., 2009; Johnson et al., 2010; Zhang et al., 2010). Thus, it is required to sequence the entirety of HIV genes associated with drug resistance, to reveal the DRMs patterns. Among the current high throughput sequencing technologies, Roche/454 has the ability to generate sequence reads of up to 1000 nucleotides length ([www.454.com](http://www.454.com)). However, the reverse transcriptase gene covers over a thousand bases and therefore needs to be sequenced using at least two amplicons. In initial analyses we found that a lot of sequences were being discarded for being not full length even though they covered all of the DRMs of interest within the amplicon. In many cases, people design amplicons that contain DRM codons that are not directly proximal to the 5’ and 3’ ends in the amplicon. Therefore, to ensure maximal coverage and to avoid unnecessary discarding of sequence reads Seq2Res finds the first and last DRM codon positions in the amplicon and considers the sequences covering those positions as optimal full-length.

* + 1. Sensitivity test of reference mapping and resistance call by Local Sierra

We tested the sensitivity of the RAMICS mapping tool and the accuracy of the locally installed version Sierra, using two datasets A and B retrieved from the Stanford HIV drug resistance database. The sequence reads in each dataset were fragmented into 3 subsequences with some overlaps, generating three amplicon datasets from each dataset. The sequence reads in each dataset were mapped to the reference sequence with RAMICS and the mutation calls at DRM codon positions output by RAMICS, were submitted to local Sierra.

The comparison of DRM calls and drug resistance calls obtained from Seq2Res to that from the Sierra web service showed that there was only one mutation call in one sequence that was different between the two approaches.

In order to ascertain why the mutation call did not correspond between Seq2Res and the Stanford database, we manually reviewed the alignment generated by RAMICS and that of the Sierra web service. RAMICS identified a deletion of nucleotide at the second position of codon 234 relative to the HXB2 reference sequence, which translates to an unknown amino acid at that position (**Figure 4.23 1A and B**). The alignment downstream from codon 234 exhibits wild type amino acids in the query sequence (**Figure 4.23 1B**).

The Sierra web service alignment, on the other hand, did not identify the potential deletion at codon 234 and, thus, the alignment of the discordant sequence to the reference sequence is incorrectly out of reading frame from codon 234 (**Figure 4.23 C and D**)

This shows the strength of the RAMICS mapping tool used in Seq2Res at correctly mapping sequence reads in the correct open reading frame. While most of the mapping/alignment tools including muscle (Edgar, 2004a, b), Clustalw (Larkin et al., 2007), MAFFT (Katoh et al., 2009; Katoh et al., 2002), Mosaik (Lee et al., 2013), T-coffee (Notredame et al., 2000), Kalign (Lassmann and Sonnhammer, 2005) align sequences at a nucleotide level, RAMICS maps in “codon-space” and thus is able to identify PCR sequencing error and genuine indels and call DRMs correctly.

* + 1. Homopolymer errors in simulated data

After confirming the sensitivity of the RAMICS and local Sierra tools used in Seq2Res, we evaluated the complete Seq2Res pipeline using simulated datasets. The simulated datasets were generated using the ART simulation tool (Huang et al., 2012).

The high insertion and deletion error rate at homopolymer regions in Roche/454 data is well known (Gilles et al., 2011; Huse et al., 2007; Kunin et al., 2009; Margulies et al., 2005; Moore et al., 2006; Vera et al., 2008; Wang et al., 2007) and, thus, we initially evaluated the simulated data to ensure that ART was introducing errors at homopolymer regions.

Quality trimming of the ART (Huang et al., 2012) simulated dataset showed that less than a base was trimmed per sequence. This is not entirely unexpected as ART does not simulate the drop off of base quality towards the 3’ end of sequence reads but only simulates the insertion and deletion errors seen in Roche/454 data mainly at homopolymer regions.

The “seed” sequences for the generation of the simulated data had all been specifically chosen on the basis that they contained at least one DRM at, or adjacent to, homopolymer regions (**Table 4.11**).

For example, in the simulated datasets generated from both the resistant and sensitive ‘seed’ sequences from sample 2368, we observed as many as four nucleotide insertions and three nucleotide deletions at homopolymer regions in the simulated data (**Table 4.12**). Generally, the most frequent error was a one nucleotide insertion (+1) or one nucleotide deletion (-1) with as many as 24.24% of sequences having a single nucleotide insertion at a homopolymer region (**Table 4.12**). Similar error profiles were observed in the simulated datasets for the other seed sequences.

We further compared the error rates at homopolymer regions in the simulated data with published reports. Gilles et al reported a 5.97% ±1.33 homopolymer related error of the total 45,895 erroneous bases in the 454 GS-FLX Titanium bases (Gilles et al., 2011) while Huse et.al reported that 39% of the errors were related to homopolymer out of 159,981 total errors (Huse et al., 2007). It appears that the error rate of indels at homopolymer regions is not consistent and, thus, the inconsistency in our simulated data is an error profile that is similar to real data.

While indel errors at homopolyer region exist in the sequence data, they are accurately identified in Seq2Res using the mapping tool RAMICS. Most mapping tools map the sequences to the reference at nucleotide space and tend to put lots of insertions and deletions in order to find the matching base. On the other hand, RAMICS maps at codon space, which mean a combination of three bases are translated to amino acid in the sequence reads and reference sequence and are aligned together although the nucleotide bases in the codons of aligned sequences are not the same (**Figure 4.24**). RAMICS is also capable of identifying a insertion or deletion of a base or codon in the sequence reads (**Figure 4.24**).

* + 1. Seq2Res sensitivity test with simulated data

The simulated sequence reads derived from resistant and sensitive “seed” sequences were pooled together at various fixed proportions to generate datasets with varying known prevalences of resistance in the resulting datasets (0.1%, 1%, 5%, 10%, 15%, 20% and 50%). The simulated datasets were analyzed using Seq2Res and the subsequent results were analyzed on the basis of the identification of drug resistant mutations and the prediction of resistance to ARVs.

In all cases we found that the observed prevalence of DRMs in the simulated data was significantly comparable to the expected frequencies indicating that all of the steps used in Seq2Res are successful at identifying PCR and sequencing error from genuine drug resistance mutations in the dataset simulated with error profile from real Roche/454 dataset.

As expected, in the simulated datasets with resistant variants present in the population at a prevalence level below the 15% threshold, the samples were called as susceptible to all drugs in each drug class.

In the simulated amplicon datasets containing 15% resistant sequences, we expected that for a particular antiretroviral drug the dataset of viral sequence reads with DRMs against the associated drug are called resistant. However, resistance call was not observed for some simulated amplicon datasets containing 15% resistant sequences. This was because the observed prevalence of DRMs in the amplicons was marginally lower than the expected prevalence. The reason for the lower observed prevalence was further investigated.

In the simulated amplicon datasets containing resistant sequences above 15% (20% and 50%), as expected, resistant/intermediate resistant call was made to all the antiretroviral drugs associated to the DRMs in the sequence reads.

We further investigated the differences in the observed prevalence of DRMs across same amplicons (PR, RT1 and RT2) and samples.

The mean observed prevalence of DRMs across the amplicons from a sample at similar expected prevalence showed that the observed prevalence was marginally greater or lower than the expected prevalence. The mean across samples for the same expected prevalence also showed that the observed prevalence was marginally greater or lower than the expected prevalence.

Furthermore, we investigated the reason for the observed prevalence to be marginally greater or lower than the expected prevalence. There are two reasons that deviated the observed prevalence from the expected prevalence.

The first reason was that the number of simulated sequence reads both with and without DRMs was discarded as non-optimal full length (**Table 4.13, 4.14**). The proportion of the discarded simulated sequences with and without DRMs was not same as the proportion that they were in the dataset before analysis. For example, if the dataset before analysis had 1000 sequences with DRMs and 9000 sequences without DRMs, the proportion was 1:10. If 100 sequences with DRMs and 50 sequences without DRMs were discarded as non-optimal full length, their proportion of discarded sequences was 2:1. If the number of sequences with DRMs discarded were higher than those without DRMs in the discarded proportion, the observed prevalence was lower than the expected prevalence. Similarly, if the number of sequences without DRMs discarded were higher than those with DRMs in the discarded proportion, the observed prevalence was higher than the expected prevalence.

The second reason was the increase in the number of sequences with DRMs due to substitution errors generating DRMs at DRM positions while simulation. This increase in the sequences with DRMs due to simulation errors increased the observed prevalence over the expected prevalence. This large effect of the simulation errors at DRM codon positions was seen only at low expected prevalence level 0.1% where the observed prevalence was higher than the expected prevalence.

Thus, evaluation of the Seq2Res pipeline for resistance genotyping on high throughput simulated sequence showed that the pipeline is able to correctly account for PCR and sequencing induced errors that affect DRM and resistance call on UDPS HIV drug resistance testing data.