Management and analysis of HIV-1 ultra-deep sequence data

Thesis submitted in partial fulfillment of the
requirements of the Degree of
Doctor of Philosophy (PhD)
2014

Submitted by:
Ram Krishna Shrestha
Student Registration Number: 3107977

Supervisor: Prof Dr Simon A Travers

South African National Bioinformatics Institute
University of the Western Cape
Private Bag X17, Bellville 7535
Republic of South Africa

Acknowledgement

I would like to acknowledge that I was able to do this project because I had the great supervisor Prof Dr Simon A Travers. This project would not have been complete without his guidance. He had been very helpful throughout my PhD. He always answered my questions though I asked repeatedly. He always said to ask him for any confusion I had. I am very thankful to him for providing comments as soon as possible to at least six drafts of every chapter in the thesis.

I thank to post doctorate fellow Dr. Natasha Wood, PhD fellow colleague Imogen Wright and web developer Baruch Lubinsky for proof reading my initial drafts of the every chapter and providing comments and correcting English grammar. They all, as research teammates, were really great and provided me the right research environment.

I would like to thank Prof Dr Grace McCormack, Dr Vijay Bansode and Dr Mónica BJ Moinz for providing the sequence data that used in the development of QTrim

I would like to thank Prof Dr Carolyn Williamson (University of Cape Town) and her research team members - post doctorate fellow Cecilia Rademeyer and PhD fellow Murray Logan for providing me the biological data generated using primer ID technology. The data was used as a sample data for the development of Primer ID Algorithm (PIDA) module in Seq2Res pipeline.

I would like to thank Prof Maria Papathanasopoulos, Irene Kets (Wits Medical School) and to all involved in CIPRA-SA study for providing a huge amount of data

generating using Roche/454 Junior and FLX high throughput sequencing technology. The data was used for in the application of Seq2Res pipeline.

I would like to thank Dr Robert Shafer and Dr Tommy Fuliswa Liu from Stanford University for generous help providing a copy of Sierra drug resistance interpretation algorithm, explaining the method to execute it and providing its update.

I would like to thank to Baruch Lubinsky for developing a web base API for Seq2Res to enable online sequence data submission and to execute Seq2Res pipeline remotely.

I would like to thank to Dr. Lynn Morris from National Institute for Communicable Diseases (NICD) and Prof Maria Papathanasopoulos (Wits Medical School) for their suggestions in development of Seq2Res.

I would like to thank to Dr Gert Van Zyl (NHLS Tygerberg and Stellenbosch University) for providing valuable suggestions to improve Seq2Res.

I would like to thank Dr Miguel Lacerda (University of Cape Town) for providing statistics help while analyzing the results.

I acknowledge that “The Atlantis philanthropies” and Department of Science and Technology (DST, South Africa) for funding my PhD.

Lastly but not the least I thank my wife Rajani Awal for supporting throughout my PhD and thesis writing. She encouraged me to do my best and always remind me to complete my thesis.

Ram Krishna Shrestha

Table of Contents

List of Figures

List of Tables

CHAPTER 1

Literature Review

1.1. 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). 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 and Bhamarapravati, 2000). 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). UNAIDS estimates that approximately 1 in every 20 adults is HIV infected in this region (UNAIDS). This is 25 or more times the HIV prevalence in any other region of the world (UNAIDS). Countries in Sub-Saharan Africa also have varying HIV prevalence with South Africa at the top followed by Nigeria

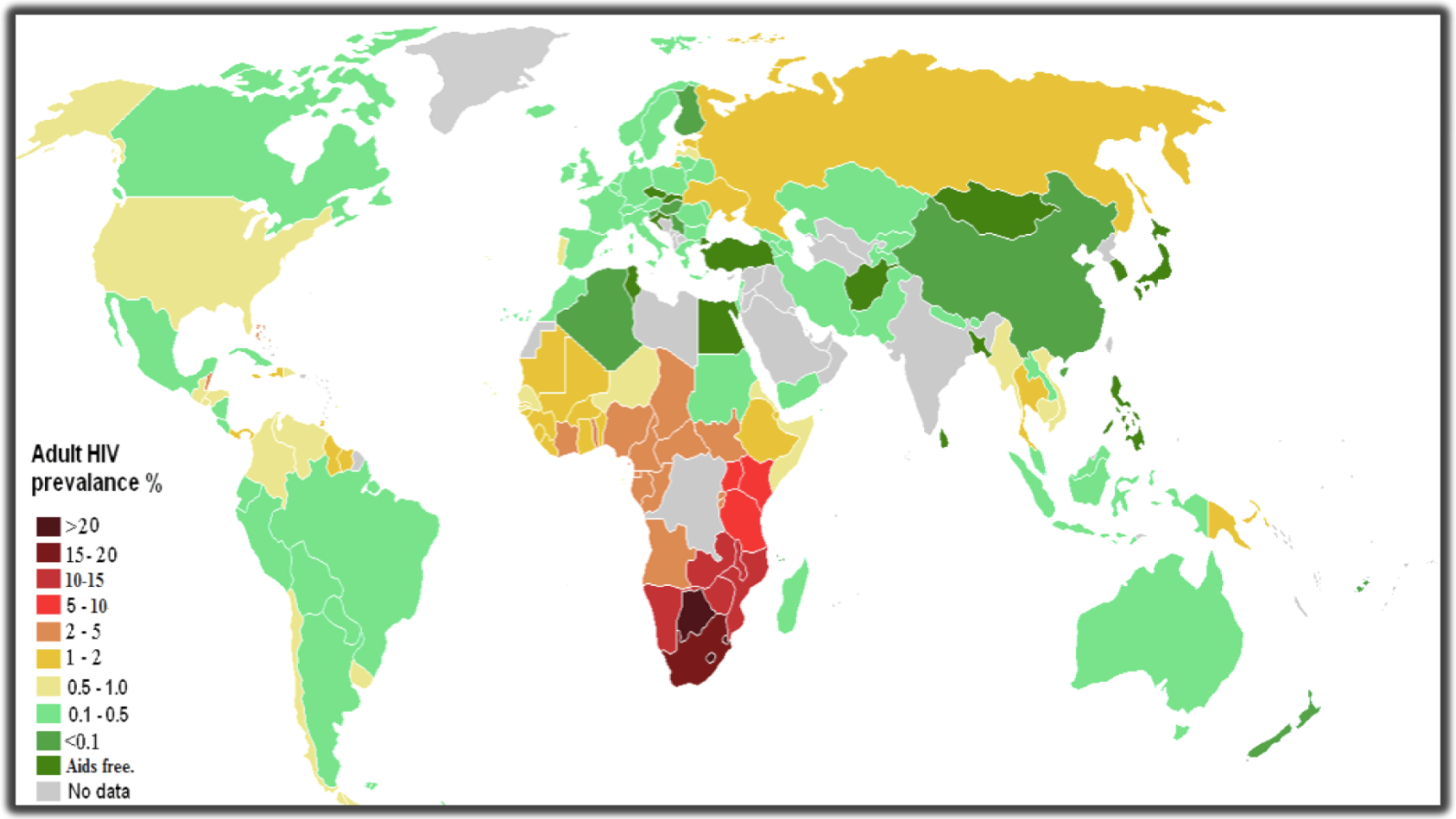


Figure 1. 1: Distribution of HIV prevalence in adults around the world.

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

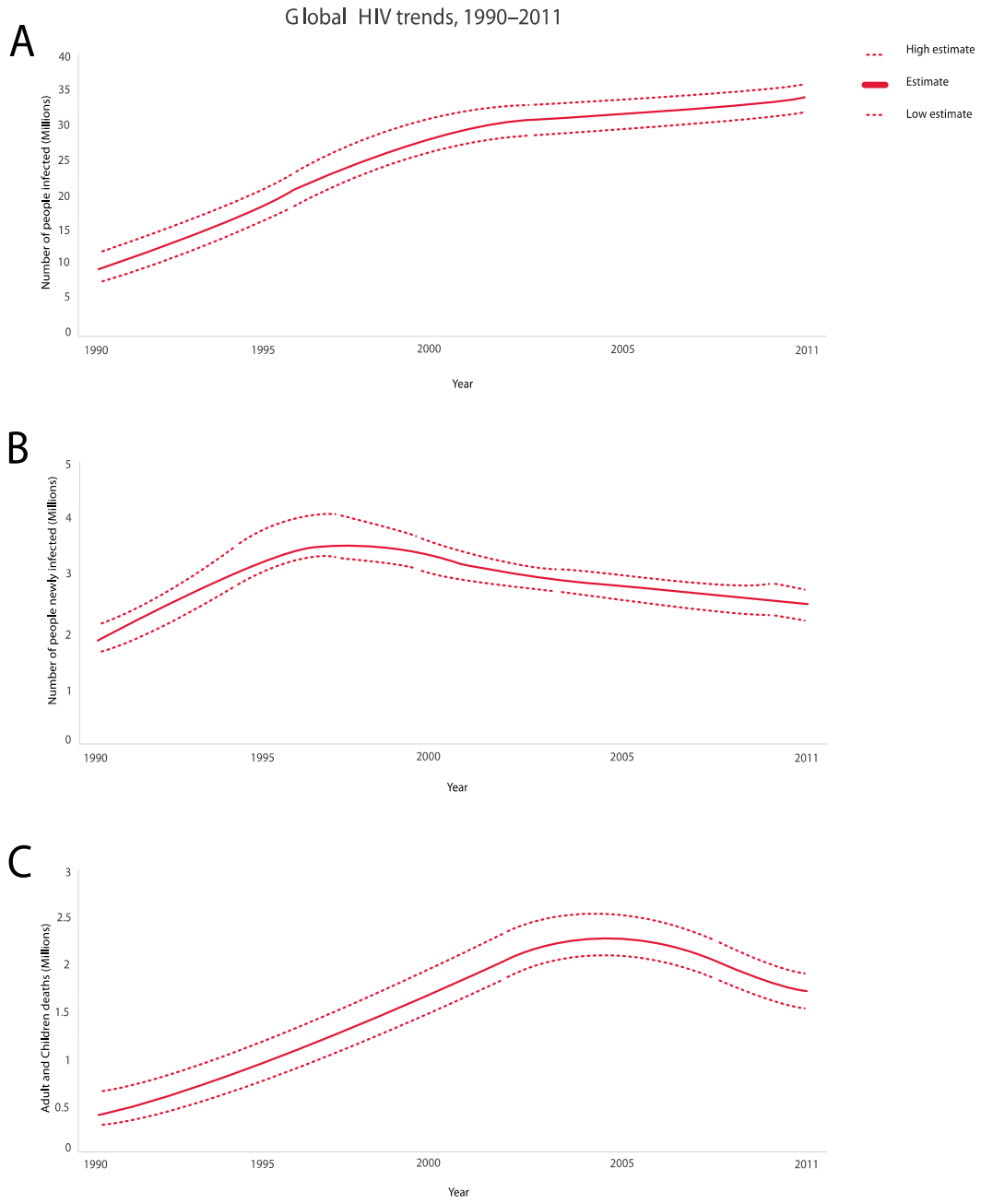


Figure 1. 2: Global trend of new HIV infections from 1990 to 2011. The number of people living with HIV globally is increasing (A) while the number of people newly infected with HIV (B) and the number of Adults and child deaths due to HIV are decreasing (C) globally in the time period. This scenario can be attributed to global scale up of drugs while infected people continue transmitting the virus to uninfected people (Zaidi et al.). (Source: modified from UNAIDS 2012)

(Esparza and Bhamarapravati, 2000). The next severely affected regions, besides African continent, are Asia (China, Thailand, Indonesia), Caribbean and Eastern Europe, North America, western and central Europe (Esparza and Bhamarapravati, 2000).

1.2. Discovery and characterization of HIV

As early as 1959, HIV infection cases had been documented but were unreported (Corbitt et al., 1990; Nahmias et al., 1986). Curious doctors at that time collected patient blood samples and kept frozen, which were later shown to have HIV antibodies (Zhu et al., 1998). In June 1981, a case of acute immune depletion associated secondary infection was reported in some homosexuals in the USA (Friedman-Kien, 1981; Friedman-Kien et al., 1981). Their infection was coupled with no lymphocyte proliferation (Gottlieb et al., 1981). Until 1983, the causative agent responsible for the severe immune depletion, named Acquired Immunodeficiency Syndrome (AIDS), was unknown (Francis et al., 1983; Gallo et al., 1983), 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 et al., 1986). Jay Levy's group in San Francisco, USA also subsequently found the virus confirming the discovery in Paris (Levy et al., 1984). They found HIV was a lentivirus from Group VI retrovirus with two single strand RNA molecules (Baltimore, 1971); unique to any other previously isolated viruses and the virus can transmit from infected to healthy people (Rogers et

al., 1987; Wofsy et al., 1986), mother to child through umbilical cord (Gallo et al., 1983; Ziegler et al., 1985).

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

1.3. Origin and evolution of HIV

Exploration of the retrovirus led researchers to identify similarities between HIV and a retrovirus in African non-human primates that were then known as Simian Immunodeficiency Virus (SIV) (Gao et al., 1994). About 40 different primates, in Africa, were infected with SIV with some harboring multiple strain of SIVs (Apetrei et al., 2004). Phylogenetic analysis of SIV from African non-human primates and HIV in human provided remarkable understanding of viral transmission as zoonotic (Bailes et al., 2002) and evolution of the virus in human after transmission (Gao et al., 1999). HIV is divided into two groups – HIV-1 and HIV-2 (Gao et al., 1999). Each group resulted from an independent cross species transmission from different African non-human primates to human (Sharp and Hahn, 2010). HIV-2 was discovered in 1986. This group was transmitted from sooty mangabey monkeys (Cercocebus atys) (Hirsch et al., 1989) and its prevalence was also high in the geographical location of these monkeys in West Africa (Santiago et al., 2005). Sooty mangabey monkeys were naturally infected by a strain of SIV (Hirsch et al., 1989). The phylogenetic analysis

of HIV-2 strains showed that they group closely with the SIVsmm strain (Hirsch et al., 1989) that was non-pathogentic to its host monkeys (Gao et al., 1992). SIVsmm evolved in its host to produce multiple strains and subsequent multiple zoonotic transmissions from sooty mangabey monkeys to human (Hahn et al., 2000) 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) but were lost for low adaptation fitness (Damond et al., 2004).

HIV-1 is the result of at least three cross species transmission events from chimpanzees (Pan troglodytes troglodytes (Ptt)) to human (Huet et al., 1990; Peeters et al., 1989). 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) (Hahn et al., 2000; Keele et al., 2006). 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 et al., 2009).

Group M is the most prevalent and accounts for 98% of all infections (reviewed in (Sharp and Hahn, 2010)). Its epicenter is thought to be Kinshasha of present day Democratic Republic of Congo (Sharp and Hahn; Sharp and Hahn, 2010). 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 et al., 2001) whereas molecular clock analysis of group M shows that the origin of it's most recent common ancestor dates back to late 1920s (Korber et al., 2000). By 1960, long before

human discovered its presence, HIV-1 group M had already diversified substantially (Worobey et al., 2008).

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

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

1.4. HIV-1 Diversity

1.4.1HIV-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 et al., 2000b). The subtype classification is based on the phylogenetic and sequence distance analyses using gene sequence data forming major clades (Robertson et al., 2000b). "At least three epidemiologically unlinked sequences are required for defining a subtype" (Robertson et al., 2000a).

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 et al., 2000b). The predominating group M subtypes are A, B, C and D (reviewed in (McCutchan, 2006)). The range of amino acid variation at gene level within a subtype and between subtypes differs from 15%– 20% and 25% - 35% respectively (Korber et al., 2001). 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 et al., 2006). Subtypes F, H, J and K infections are rare and collectively account for only 0.94% infections (Hemelaar et al., 2006).

Subtypes can be further classified to sub-subtype based on a distinct sister clade formation (Gao et al., 2001) within a clade with the same rule of "phylogenetic and

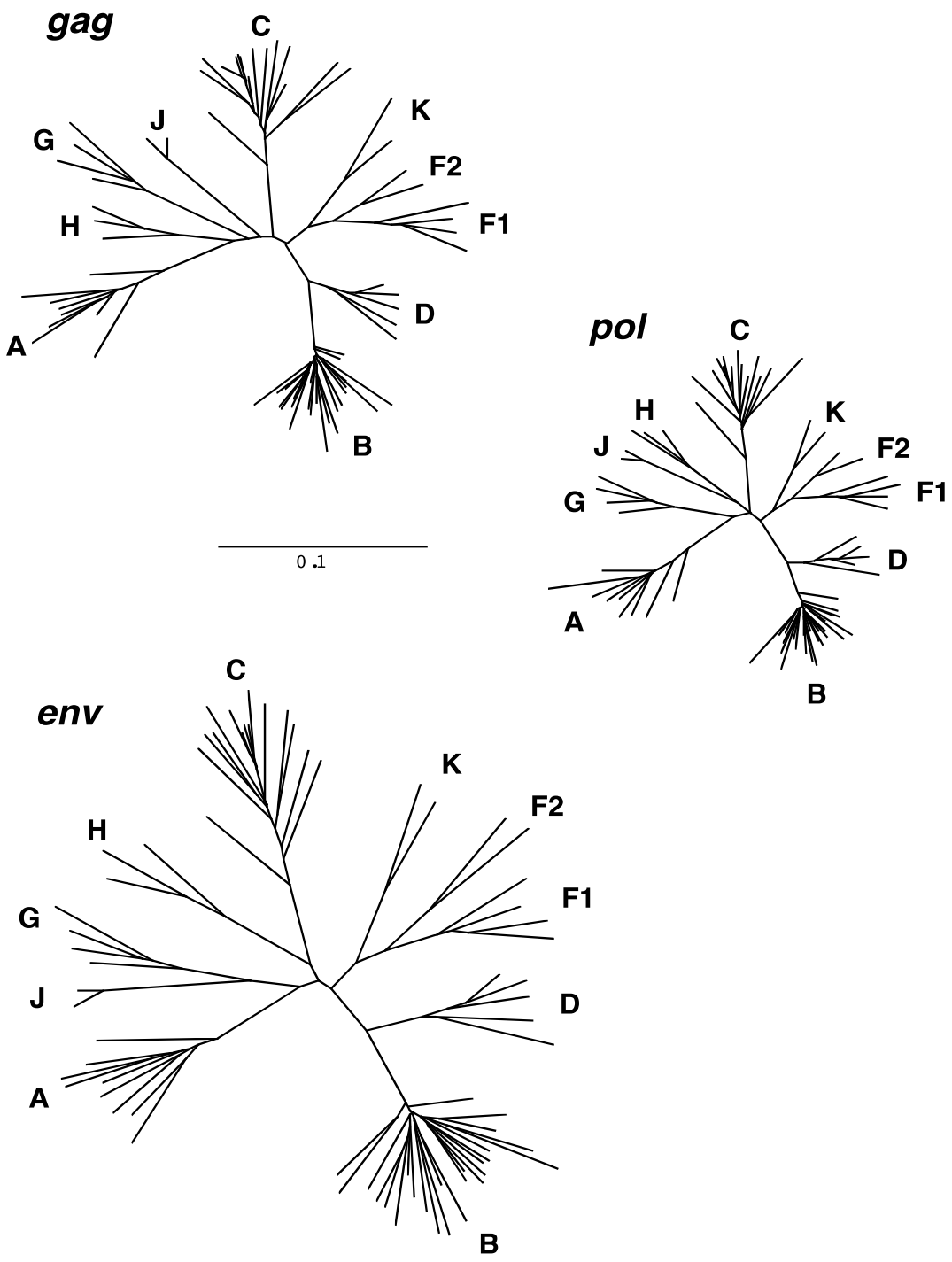


Figure 1. 3: Phylogenetic tree showing HIV-1 group M diversification to subtypes A-D, F-H, J and K, inferred from nucleotide sequence alignments of *gag*, *pol* and *env* genes. Source: Robertson et al 2000 (Robertson et al., 2000a)

HIV-1[a]	Sub-classification[b]	Description	Main geographic concentrations	Global prevalence[c]	Compete genome sequences[d]
Group M					
Subtype	A	Includes sub-subtypes A1, A2, and A3	East and West Africa, former Soviet Republics	High	62
	B	B prime (B') variant in Southeast Asia	America, Western Europe, Australia	High	129
	C	50% of all HIV-1infections worldwide	Southern Africa, East Africa, India	High	209
	D	East and West African variants	East and West Africa	High	47
	F	F1 and F2 sub-subtypes	West Central Africa	Very low	10
	G		West Central Africa	Very low	8
	H		West Central Africa	Very low	3
	J		West Central Africa	Very low	2
	K		West Central Africa	Very low	2
CRF	CRF01_AE	Former "E"; subtypes A, E	Southeast Asia	High	52
	CRF02_AG	Subtypes A, G	West Africa	High	38
	CRF03_AB	Subtypes A, B	Former Soviet Republics	Very low	3
	CRF04_cpx	Former "I"; subtypes A,G,H,K		Very low	3
	CRF05_DF	Subtypes D, F	Central Africa	Very low	3
	CRF06_cpx	Subtypes A, G, J, K	West Africa	Very low	5
	CRF07_BC	Subtypes B', C	China	Very low	3
	CRF08_BC	Subtypes B', C; related to CRF07_BC	China	Very low	4
	CRF09_cpx	Subtypes A, F, G, related to CRF02_AG	West Africa	Very low	3
	CRF10_CD	Subtypes C, D	Tanzania	Very low	3
	CRF11_cpx	Subtypes G, J	West Central Africa	Very low	10
	CRF12_BF	Subtypes B, F	Argentina	Very low	7
	CRF13_cpx	Subtypes A, G, J, and CRF01_AE	West Central Africa	Very low	3
	CRF14_BG	Subtypes B, G	Spain, Portugal	Very low	6
	CRF15_01E	CRF01_AE, subtype B	Thailand	Very low	4
	CRF16_A2D	Sub-subtype A2, subtype D	Kenya	Very low	3
URF	AC, AD, CD, ACD	Estimated 30% of strains in East Africa	Kenya, Rwanda, Tanzania, Uganda	Moderate	44
	Complex	Common in West/West Central Africa	Nigeria, Cameroon	Moderate	33
	BF	Abundant in South America	Brazil, Argentina	Moderate	31
	CRF01_AE/B	Emerging in Southeast Asia	Thailand, Myanmar,	Very low	13
Group O		West Central Africa	Cameroon, Senegal	Very low	18
Group N		West Central Africa	Cameroon	Very low	3
HIV-2	Subtypes A and B		West Africa	Moderate	22

[a]HIV Type 1 and HIV Type 2; HIV-1 strains are classified into groups M, N, and O, respectively. Group M is subdivided into subtypes, circulating recombinant forms (CRF), and unique recombinant forms (URF).
[b]Subtypes are assigned letters in order of their discovery; CRF are numbered sequentially in order of discovery with the component subtypes after the underscore. "cpx" (complex) indicates that more than three subtypes are present. Some CRF also contain unclassified regions. URF are grouped according to their geographic origin for this analysis.
[c]Estimated from degree of geographic spread, prevalence of HIV in countries and regions, and proportion of total HIV infections in the country or region. Dark grey fill indicates the six most globally prevalent strains; light grey fill indicated strains of moderate abundance.
[d]Based on the HIV sequence database at http://hiv-web.lanl.gov. Because of recombination, only complete or virtually complete genomes of HIV provide unequivocal classification. Partial genome sequences and multiple sequences from a single individual, were excluded. CRF16_A2D includes two complete genomes from our unpublished data. At least five additional CRF have been described but sequences are not yet publicly available.

Figure 1. 4: HIV diversity around the globe, its level of prevalence in the area and number of genome sequenced. Source: McCutchan 2006

distant analyses but not justifiable to call a subtype due to low genetic distance" (Robertson et al., 2000b). Only subtypes A and F exhibit distinct sister clades (Gao et al., 2001). Subtype A has sub-subtypes A1 and A2 (A3 and A4 are mentioned by Taylor et al (Taylor et al., 2008)); Subtype F has sub-subtypes F1 and F2 (**Figure 1.3** ***gag***); sub-subtype F3 mentioned by Taylor et al (Taylor et al., 2008).

The lower diversity observed in Group N (Ayouba et al., 2000), O (Lemey et al., 2004b) and P (Vallari et al., 2011) 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 et al., 2002). Group N as well does not show distinct sub clade within itself (**Figure 1.5**).

1.4.2 HIV-1 recombination

Initially, HIV-1 group M subtypes E (Artenstein et al., 1995; Tovanabutra et al., 2002; Wasi et al., 1995) and I (**Figure 1.5**) were also classified (reviewed in (McCutchan, 2006)). 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)). The same criterion of epidemiological unlinked isolates from three or more people applies for classification as a circular recombinant form (CRF) (Robertson et al., 2000a). A recombinant isolate that is discovered in single patient is termed as Unique Recombinant Form (URF) (reviewed in (McCutchan, 2006)). There are 55 CRFs listed in Los Alamos National Laboratory

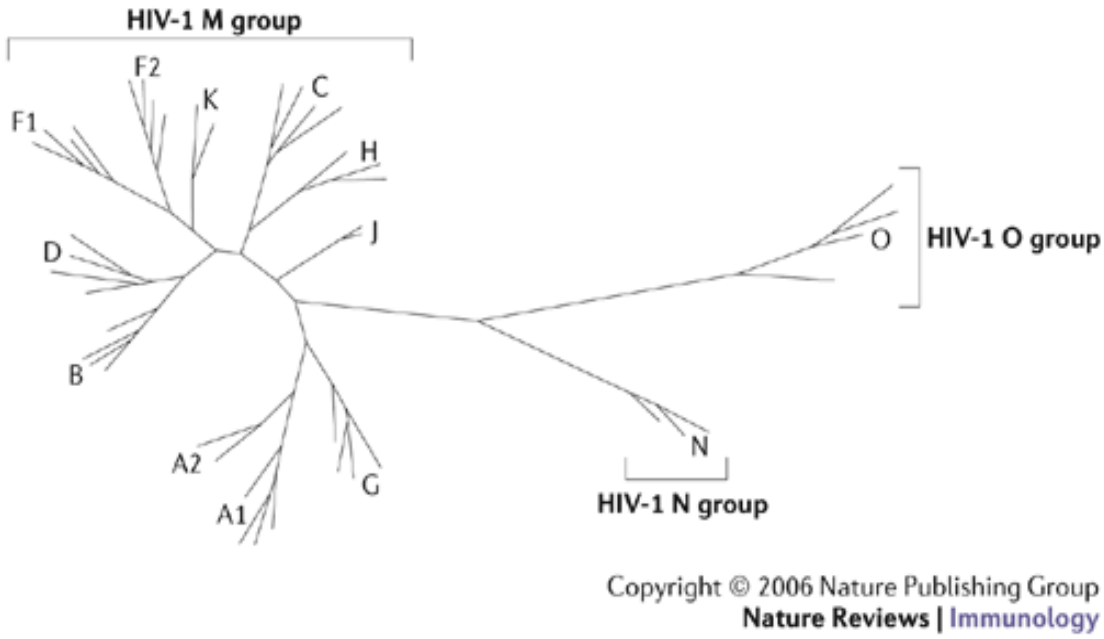


Figure 1. 5: Neighboring joining phylogenetic tree showing HIV-1 group M, N and O. Group M shows distinct nine subtypes A-D, F-H, J, K while no specific subtype is observed in group N and O. Source: Letvin 2006 (Letvin, 2006)

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 et al., 2008; McCutchan et al., 2002; Salminen et al., 1995; Zhang et al., 2010b) 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 et al., 2000a; Robertson et al., 1995; Sharp et al., 1995). CRF02_AG is the most prevalent circulating recombinant form infecting over 9 million people on the whole (McCutchan, 2000) 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 et al., 2008).

1.4.3 Intra-patient HIV diversity

HIV infection initiates mostly with a single virion (Fischer et al.; Keele et al., 2008). Evidences of multiple HIV variants transmission are also recorded (Long et al., 2000; Ping et al., 2000). Generally, HIV is genetically homogenous for a short post infection time (Delwart et al., 2002; Haase). In the long-term post infection period, virus replicates rapidly to produce genetically heterogeneous population (Long et al., 2000). 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)). It is observed that the diversity at a gene, for example *env,* in viral quasispecies can be approximately 30% (Fouchier et al., 1992). Factors that contribute to high genetic heterogeneity in viral quasispecies are high replication rate and turnover (Carpenter et al., 2000), viral genome recombination (Fang et al., 2004;

Gu et al., 1995; Lole et al., 1999), higher mutation rate by erroneous reverse transcriptase (Bebenek et al., 1989; Roberts et al., 1988), and host immune selection (Borrow et al., 1997; Price et al., 1997). On the whole, HIV replication is the overall source of genetic heterogeneity in the viral population (reviewed in (Smyth et al.)).

Intra patient HIV genome recombination is a common event (Fang et al., 2004; Neher and Leitner). Two genomes from different viral strains from same subtype or different subtypes can be co-packed into single virion during replication (Stuhlmann and Berg, 1992). 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 et al., 1996; Kostrikis et al., 2002). 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 et al.)). 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 et al., 2003)).

1.5. HIV genome and proteins – structures and functions

HIV has nine genes and produces 15 proteins (Frankel and Young, 1998). 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 and Young, 1998).

1.5.1. Accessory genes

Vif promotes the viral infectivity to the host, but has no role in viral production (Jager et al.). Vif is produced in the late stage of viral production (Sheehy et al., 2002; von Schwedler et al., 1993) to suppress the innate antiviral immunity of host (Madani and Kabat, 1998; Simon et al., 1998b). It is observed that vif is expressed only when the virus infects immune cells that express cytidine deaminase APOBEC3G (Navarro and Landau, 2004). The reason is that Vif protein prevents APOBEC3 proteins from hyper mutating HIV reverse transcripts as a mechanism of defense (Conticello et al., 2003; Mangeat et al., 2003; Mariani et al., 2003; Marin et al., 2003; Mehle et al., 2004; Schafer et al., 2004; Sheehy et al., 2003; Simon et al., 2005; Stopak et al., 2003; Wiegand et al., 2004; Zhang et al., 2003).

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

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

Nef has no role in viral infectivity but plays a role during the biogenesis of viral particles (Laguette et al., 2009) and virulence (Lenassi et al.; Simmons et al., 2001). Nef down regulates the production of major histocompatibility complex type I (MHC type I) in the host cell (Blagoveshchenskaya et al., 2002; Greenberg et al., 1998; Lewis et al., 2012). This impairs the function of cytotoxic T lymphocyte cells to clear the infected cells (Adnan et al., 2006; Baur et al., 1994; Collins et al., 1998; Couillin et al., 1994; Sawai et al., 1994). Nef also down regulates CD4 on host cell surface (Garcia and Miller, 1991; Lama et al., 1999) and modulates cellular activation to evade host immune system (Baur et al., 1994; Sawai et al., 1994).

1.5.2. Structural genes and proteins

The *Gag* gene produces a precursor polyprotein (pr55gag) of ~ 500 amino acids long and weighs 55 kilo Dalton (Briggs et al., 2004). The *Gag* precursor has all the building blocks to form a fully infectious virion, even in the absence of other viral products (Wang and Barklis, 1993). A proteolytic cleavage of *gag* precursor yields the structural proteins – matrix, capsid, nucleocapsid and p6 (**Figure 1.6**) (Wiegers et al., 1998). The cleavage takes place in the nascent virus after budding out from host cell

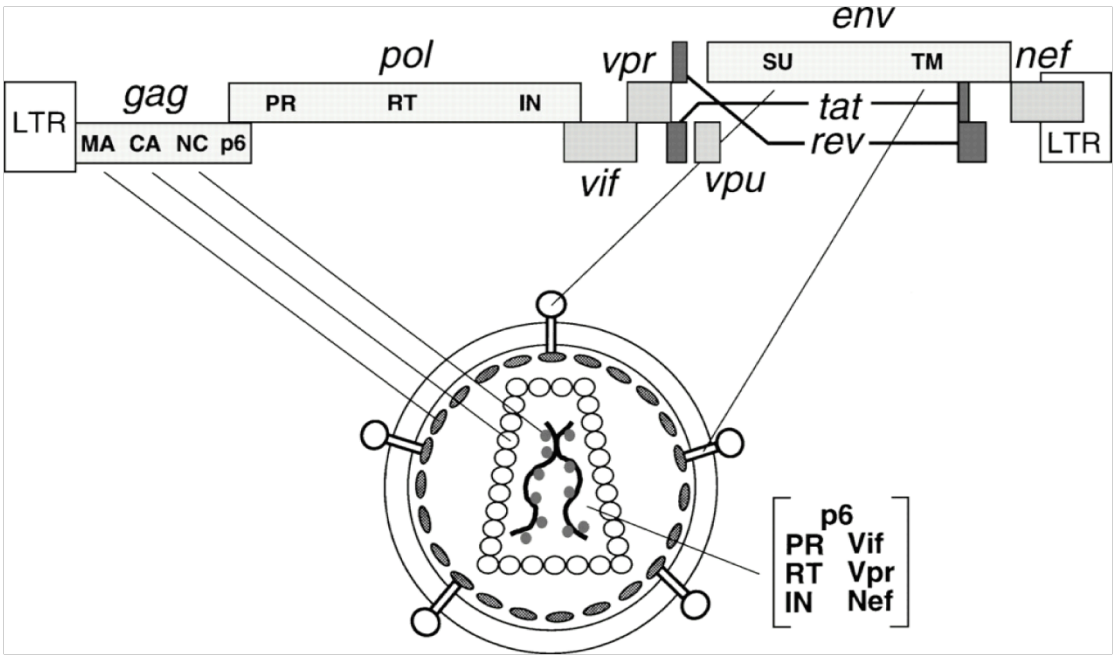


Figure 1. 6: HIV genes and proteins positions in the viral genome and their viral parts.

Source: Frankel and Young 1998 (Frankel and Young, 1998)

(Göttlinger et al., 1989). The matrix protein is at the N-terminal and p6 at the C-terminal of *gag* precursor (Borsetti et al., 1998; Wiegers et al., 1998).

All *gag* proteins play role at "post assembly and post processing stages in viral infectivity" (Wang and Barklis, 1993). 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 (Dawson and Yu, 1998; Gheysen et al., 1989; Wang and Barklis, 1993; Zhang et al., 1998). The highly basic region in matrix mediates electrostatic association with phospholipids in plasma membrane during assembly (Chukkapalli et al.; Zhou and Resh, 1996). The capsid is a curved and closed shell consisting 250 hexamers and 12 pentamers (reviewed in (Briggs and Krausslich)). Capsid packs viral proteins, nucleocapsid and viral genome during assembly to pass on to new HIV particles (Ganser-Pornillos et al., 2004). Nucleocapsid plays role in efficient viral assembly by making pr55–pr55 inter-protein contacts (Dawson and Yu, 1998; Zhang et al., 1998) and localization of viral proteins (Larsen et al., 2008). P6 protein plays role in detaching and releasing the newly formed HIV particles (Demirov et al., 2002).

The *pol* polyprotein is produced by translational frame shift (Karacostas et al., 1993), such that, as much as 241 nucleotides at 5' region of the gene overlaps with 3' region of gag gene (Jacks et al., 1988; Ratner et al., 1985). 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 et al., 1988b; ERICKSON-VIITANEN et al., 1989; Nutt et al., 1988). The

reverse transcriptase enzyme reverse transcribes the viral RNA to produce a cDNA molecule after infecting host cell (Jacobo-Molina and Arnold, 1991; Sarafianos et al., 2009). The RNase H domain in RT degrades the viral RNA molecule following cDNA production (Davies et al., 1991). 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 et al., 1994).

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 et al., 1994). 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 et al., 1997). Three gp120 molecules bound with three gp41 molecules to form envelop spikes (Pancera et al.). They are organized to form trimeric complexes on the surface of HIV and mediate HIV entry into the host cell (Liu et al., 2008). The exposed external complex gp120 binds to the CD4 receptor on the host immune cell (Rizzuto et al., 1998). 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 et al., 2007; Rizzuto et al., 1998; Wu et al., 1996; Wu et al., 1997). The gp120-CD4 complex also triggers conformational change in gp41 trans-membrane protein from native non-fusion state to fusion state (Chan et al., 1997; Kliger et al., 1997). Gp41 plays role in the viral fusion and release of viral contents in to the host cell (Furuta et al., 1998; Melikyan, 2008),. The gp41 consists of heptad repeats - HR1 and HR2 that play role in fusion process (Furuta et al., 1998; Tan et al., 1997). HR1 is a bundle of three helical motifs

and HR2 is trimeric coiled coil structure (Dwyer et al., 2003). During fusion process, HR2 makes numerous contacts with HR1 to form stable six helical bundles (Melikyan et al., 2000).

1.5.3. Regulator Genes

Tat is a trans-activating factor localized in the nucleus for HIV gene expression (Rosen and Pavlakis, 1990b; Roy et al., 1990). The HIV proviral genome integrated in to the host genome is regulated by cellular as well as the viral transcription regulatory factors (Cullen, 1991; Gaynor, 1992). Tat is the primary transcriptional regulatory factor (Marcello et al., 2001). 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 et al., 2002; Chou et al.; He and Zhou). Sodroski et al. (1985) first explained the function of Tat (Sodroski et al., 1985).

Rev is a 19 kilo Dalton phosphoprotein (Malim et al., 1989a) trans-activating factor for HIV gene expression (Rosen and Pavlakis, 1990b). Like Tat, it is also mainly localized in the nucleus of host cell (Rosen and Pavlakis, 1990a), but cycles rapidly between the nucleus and cytoplasm as it promotes nuclear export of the transcriptional products (Fischer et al., 1995; Fischer et al., 1994; Henderson and Percipalle, 1997; Malim et al., 1989b). Rev binds at the Rev Responsive Element (RRE), which is an RNA element encoded within the env region of the virus (Daly et al., 1989; Malim and Cullen, 1991).

1.6. HIV replication

There are 11 major events in HIV's replication cycle (**Figure 1.7**). The initial step of viral entry in to a host immune cell includes HIV gp120 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.

Following the fusion, the viral core enters in to the cytoplasm of host cell. The reverse transcriptase enzyme reverse transcribes the RNA molecule to in the intact capsid of the viral core (McDonald et al., 2002). The reverse transcriptase enzyme is not perfect at copying mRNA molecule to cDNA and has no capability for error correction (Bebenek et al., 1989; Bebenek et al., 1993; Preston et al., 1988; Roberts et al., 1988). The rate of errors generated by reverse transcriptase is in the order of 10^{-5} per base per replication cycle (Mansky and Temin, 1995). This is a crucial step as it contributes to generation of variations in the viral quasispecies (reviewed in (Goodenow et al., 1989; Nowak et al., 1990)).

Subsequently the capsid is dissembled, termed as uncoating (McDonald et al., 2002; Shah et al., 2013), releasing the ribonucleoprotein complex in to the cytosol (Dismuke and Aiken, 2006). The process can take an hour or less since time of post infection (Hulme et al., 2011). 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 et al., 1993; Farnet and Haseltine, 1991; Miller et al., 1997). The PIC protects cDNA from endonuclease degradation (Miller et al., 1997).

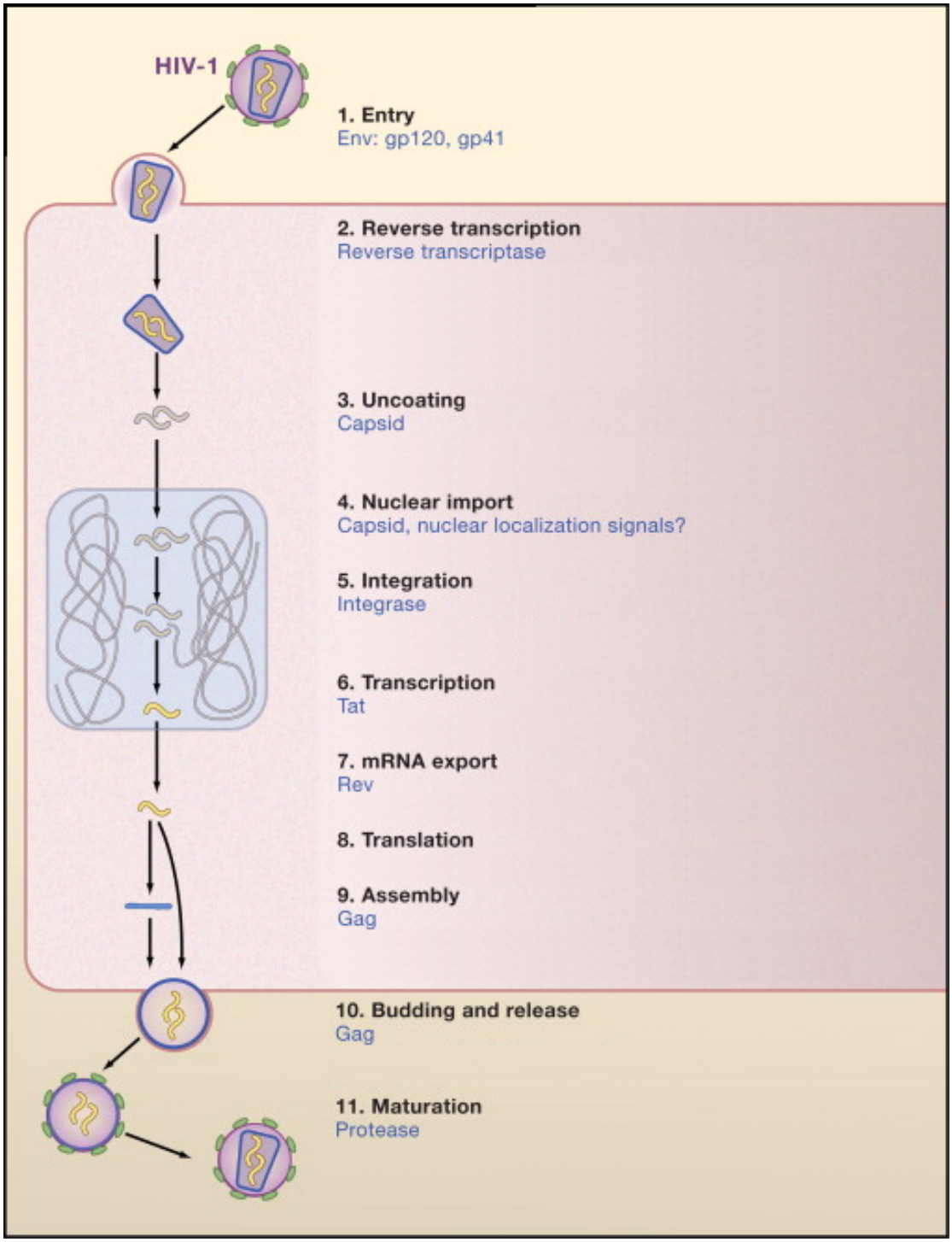


Figure 1. 7: The HIV replication cycle showing major stages. Viral proteins that play role in each event are colored blue. (Source: modified from Ho and Bieniasz 2008) (Ho and Bieniasz, 2008)

In an ATP dependent process (Bukrinsky et al., 1992), PIC is transported on host microtubules towards the nuclear membrane (McDonald et al., 2002). Integrase assists in nuclear import in association with nuclear import machinery like importin (Fassati et al., 2003) and transportin-SR2 (Christ et al., 2008). It is now established that central polypurine tract-central termination sequence (cPPT-CTS) plays role in kinetics of nuclear import (Riviere et al.).

Post-nuclear entry, integrase processes the viral DNA for integration (LaFemina et al., 1992) into host genome. Host transcriptional co-factor LEDGF/p75 and HIV integrase interact to tether to the host chromosome during the integration process (Emiliani et al., 2005).

The integrated HIV provirus hijacks the host cell transcriptional machinery for viral genes to transcribe (Davey et al., 2011). HIV protein Tat promotes the transcription of the viral DNA (Ott et al., 2011; Razooky and Weinberger, 2011). The viral transcriptome encodes structural proteins, accessory proteins and viral enzymes necessary for a complete functional HIV (reviewed in (Karn and Stoltzfus)).
The viral Rev protein facilitates exporting of the unprocessed viral transcriptome to cytoplasm for translation (Malim et al., 1989b). HIV has no translation system of its own; the host translational machinery is exploited for translation of the viral transcriptome to its proteome (Cherry et al., 2005; Thompson and Sarnow, 2000).

Upon translation of all viral proteins, viral *gag* initiates virion assembly at the cell membrane (Dong et al., 2005; Nermut et al., 1998; Saad et al., 2006). The complete assembled virion particles bud out and are released from the plasma membrane by the

host ESCRT machinery involving Tsg101 and ALIX regulatory proteins (Fujii et al., 2007; Garrus et al., 2001; Saksena et al., 2007). The maturation of the nascent HIV virions begins concomitantly with budding out (Klimkait et al., 1990; Schubert et al., 2000). *Gag* and *Pol* polyproteins are proteolytically cleaved by protease enzyme in the maturation step (Darke et al., 1988; Pettit et al., 2005).

Each HIV replication cycle releases new infectious virions in the order of 10^9 per day (Carpenter et al., 2000). The number of new infecting HIV determines the replication rate of the virus (Tersmette et al., 1989). A long post infection period shows higher turnover rate associated with CD4+ cell depletion and viral population expansion (Carpenter et al., 2000).

1.7. Antiretroviral Drugs

1.7.1. Reverse Transcriptase Inhibitors

1.7.1.1 Nucleoside Reverse Transcriptase Inhibitors (NRTIs)

NRTIs are analogs of nucleotides but without 3' hydroxyl group (reviewed in (Sarafianos et al., 2004)). The drug is taken in unphosphorylated form, which cytokinases phosphorylates to form 5' triophosphates (reviewed in (De Clercq, 2002; Ilina and Parniak, 2008)). This then, leads to the incomplete termination of HIV-1 cDNA synthesis (Sluis-Cremer et al., 2000). The action of the drugs is shown in **Figure 1.8**.

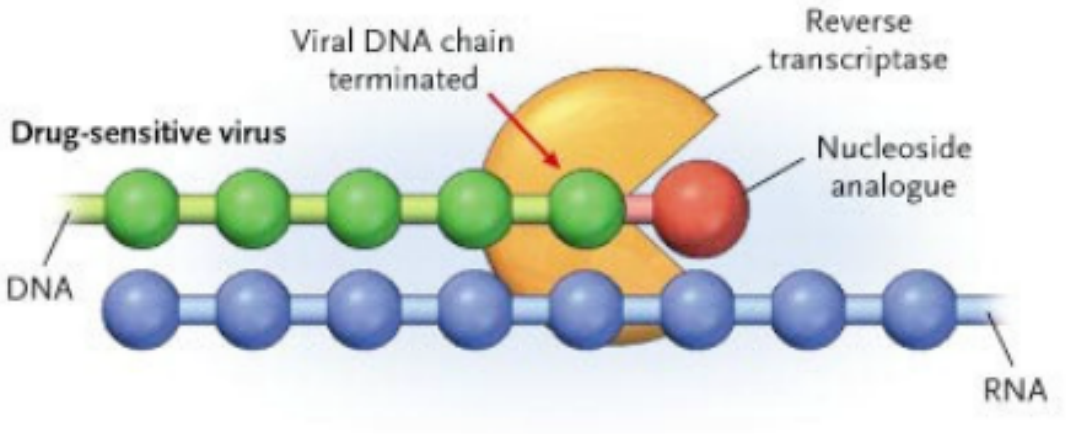


Figure 1. 8: Mechanisms of NRTI drug to inhibit reverse transcriptase. Nucleoside analogs lacking 3' hydroxyl group is incorporated in growing chain of drug sensitive virus resulting in incomplete termination of viral cDNA. Source: Adapted from (Clavel and Hance, 2004)

1.7.4.2 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 et al., 2006; Sarafianos et al., 2009), 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**) (Balzarini, 2004; Esnouf et al., 1995). The list of approved NNRTI drugs is shown in **Table 1.1,**

1.7.2 Protease Inhibitors (PI)

Protease Inhibitors interfere with the cleavage of the *gag-pol* polypeptide (Seelmeier et al., 1988) as competitive peptidomimetic inhibitors. The hydroxyethylene core in the inhibitors prohibits the cleavage action of the HIV protease enzyme (Vacca et al., 1994; Vacca et al., 1991). However, as an adverse side effect, patients that have used these inhibitors have developed lipodystrophy and hyperlipidemia (Carr et al., 2000; Carr et al., 1998a; Carr et al., 1998b, c; Liang et al., 2001; Miller et al., 2000; Tsiodras et al., 2000).

1.7.3 Integrase Inhibitors

The feasibility and efficacy of integrase inhibitors have been tested in Rhesus Macaques (Hazuda et al., 2004). Most of the integrase inhibitors target the strand transfer function of the enzyme (Bera et al., 2011; Espeseth et al., 2000; Hazuda et al., 2000; McColl and Chen, 2010; Pannecouque et al., 2002). An X-ray structure of the

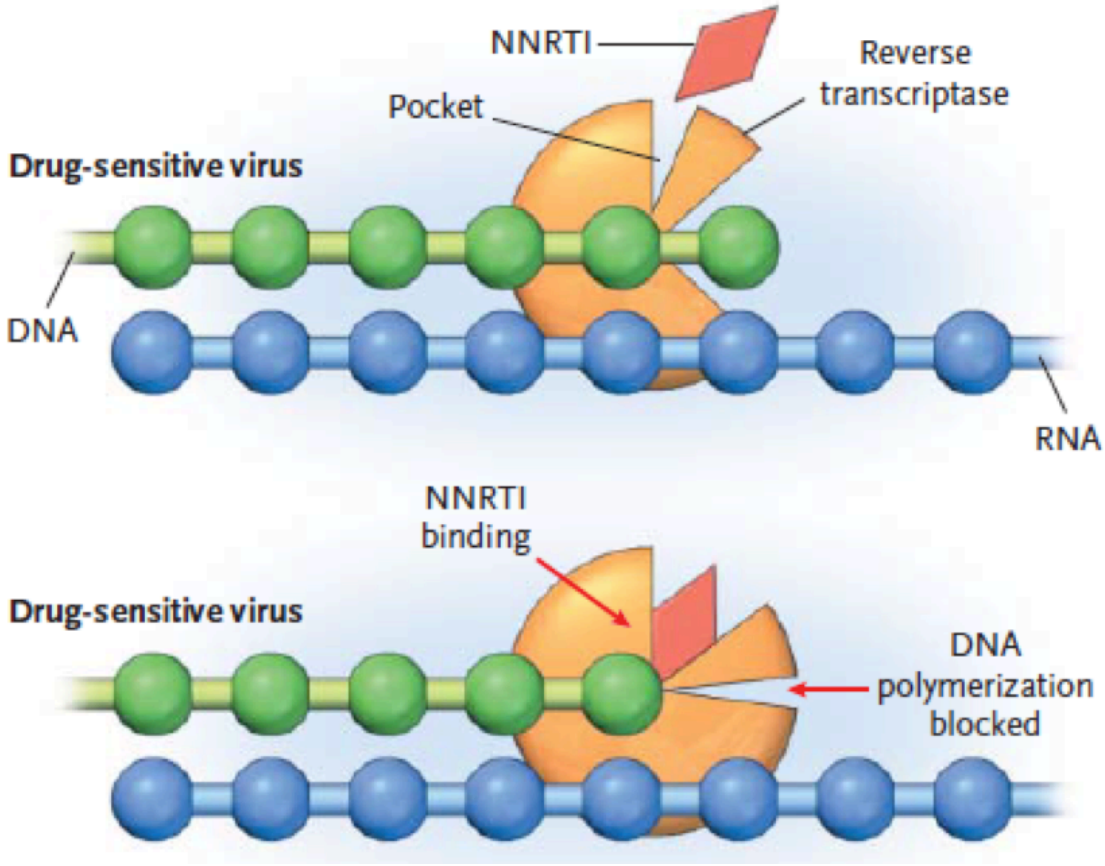


Figure 1. 9: Mechanism of NNRTI drugs to inhibit reverse transcriptase. The drugs bind to drug sensitive viral reverse transcriptase disabling its function. Source: Adapted from (Clavel and Hance, 2004)

Table 1. 1: Antiretroviral drugs used in the HIV treatment, their mechanisms of action and mechanisms of resistance. Source: Adapted from (Clavel and Hance, 2004) and (Ammaranond and Sanguansittianan, 2012; Colin et al., 2013)

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

integrase enzyme has revealed the active site model of the enzyme complexes with the DNA (Chen et al., 2008). 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 and Sanguansittianan, 2012)).

1.7.4 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:

1.7.4.1 CCR5 co receptor antagonist:

Maraviroc is the only CCR5 antagonist in clinical use (De Clercq, 2005a, b; Fätkenheuer et al., 2005; Rosario et al., 2005; Rosario et al., 2006; Wheeler et al., 2007). 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 and van der Ryst, 2005). This binding prevents HIV gp120 binding to the co-receptor, thereby disabling the viral entry in to the cell (Fätkenheuer et al., 2005). 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 et al., 2010).

1.7.4.2 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 et al., 2003)). Enfuvirtide (Duffalo and James, 2003; Poveda et al., 2005) is a synthetic peptide, approved for clinical use in 2003 (Robertson, 2003), which can bind to the gp41 HR1 region (Wild et al., 1993). However the emergence of Enfuvirtide resistant viral strains lead to its discontinuation for clinical use in 2004 (Briz et al., 2006). Sifuvirtide is another HIV fusion inhibitor peptide under research (Wang et al., 2009).

1.8. HIV Treatment

1.8.1 Brief history of antiretroviral treatment

The treatment of HIV infection has been a great challenge and still remains as an unsolved problem (Sandstrom and Kaplan, 1987). In 1985, an assay for diagnosis of HIV antibody was developed for the confirmation of HIV infection (Ward et al., 1986). 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 (Koch et al., 1992; Richman et al., 1987). 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 et al., 1993; Fischl et al., 1990; Lundgren et al., 1994; Volberding et al., 1995; Volberding et al., 1990). 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

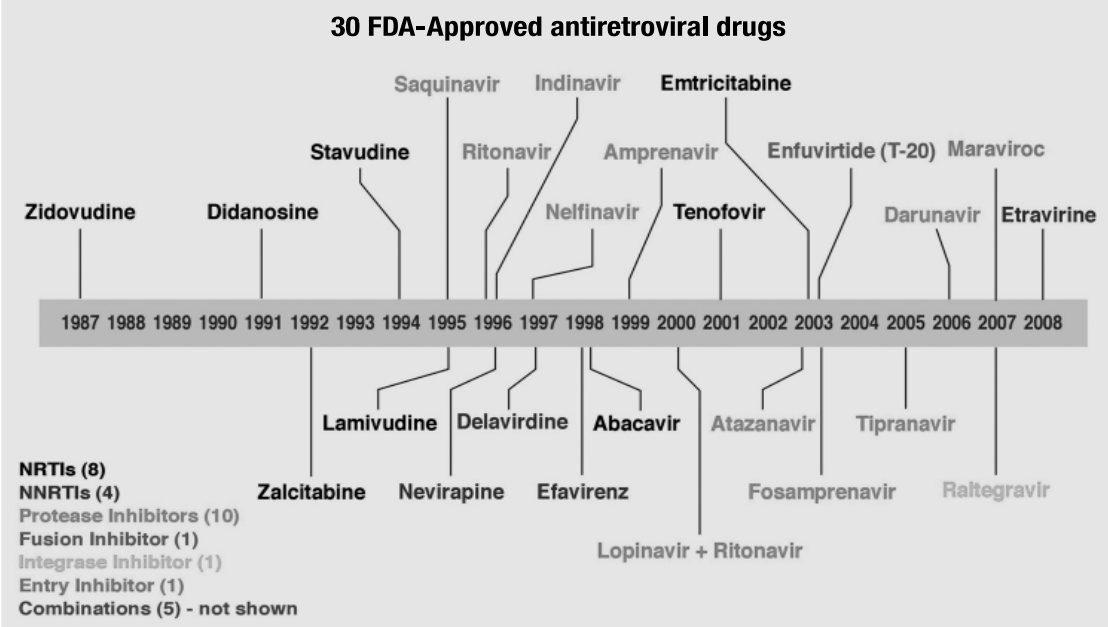


Figure 1. 10: The time line of approved HIV antiretroviral drugs. Source: (Palmisano and Vella, 2011)

were toxic as well. The administration of the drugs was altered to reduce the toxicity of each drug but the approach remained ineffective (Skowron et al., 1993). Then, a combination therapy containing two NRTI drugs (Saravolatz et al., 1996), 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 et al., 1996). Triple NRTI combination therapy containing 3TC, ZDV and d4T was better tolerated but could not control HIV reproduction (Kuritzkes et al., 1999). A good result obtained from using those NRTI drugs was the substantial reduction in HIV transmission from mother to child at birth (Connor et al., 1994a; McGowan and Shah, 2000; McIntyre et al., 2009).

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 (Montaner et al., 1998a; Montaner et al., 1998b; Staszewski et al., 1999b). Besides antiretroviral activity, combination therapy was also studied for toxicity and tolerability (Montaner et al., 1998b; Staszewski et al., 1999a; Staszewski et al., 1999b). The triple combination therapy of Nevirapine/efavirenz (NNRTI drug) with two NRTI drugs showed a good viral suppressing result (Staszewski et al., 1999a; Staszewski et al., 1999b) and was superior to monotherapy and dual therapy (Robbins et al., 2003a).

A drug cocktail with 2 NRTIs and a protease inhibitor showed highly effective result (Cameron et al., 1999; Merry et al., 1997) with viral suppression time longer than the

study period (Gulick et al., 2000; Hammer et al., 1997). The concept of highly active antiretroviral therapy was conceived after the cocktail of three drugs from different classes showed effective results (Gulick et al., 1998; Gulick et al., 1997; Hammer et al., 1997). 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 et al., 1997). 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 et al., 2012).

1.9. 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 and Kovarik).

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 and Kovarik, 2010).

1.10. 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 (Shirasaka et al., 1995; Simen et al., 2009) that could lead to virologic failure (**Figure 1.11**).

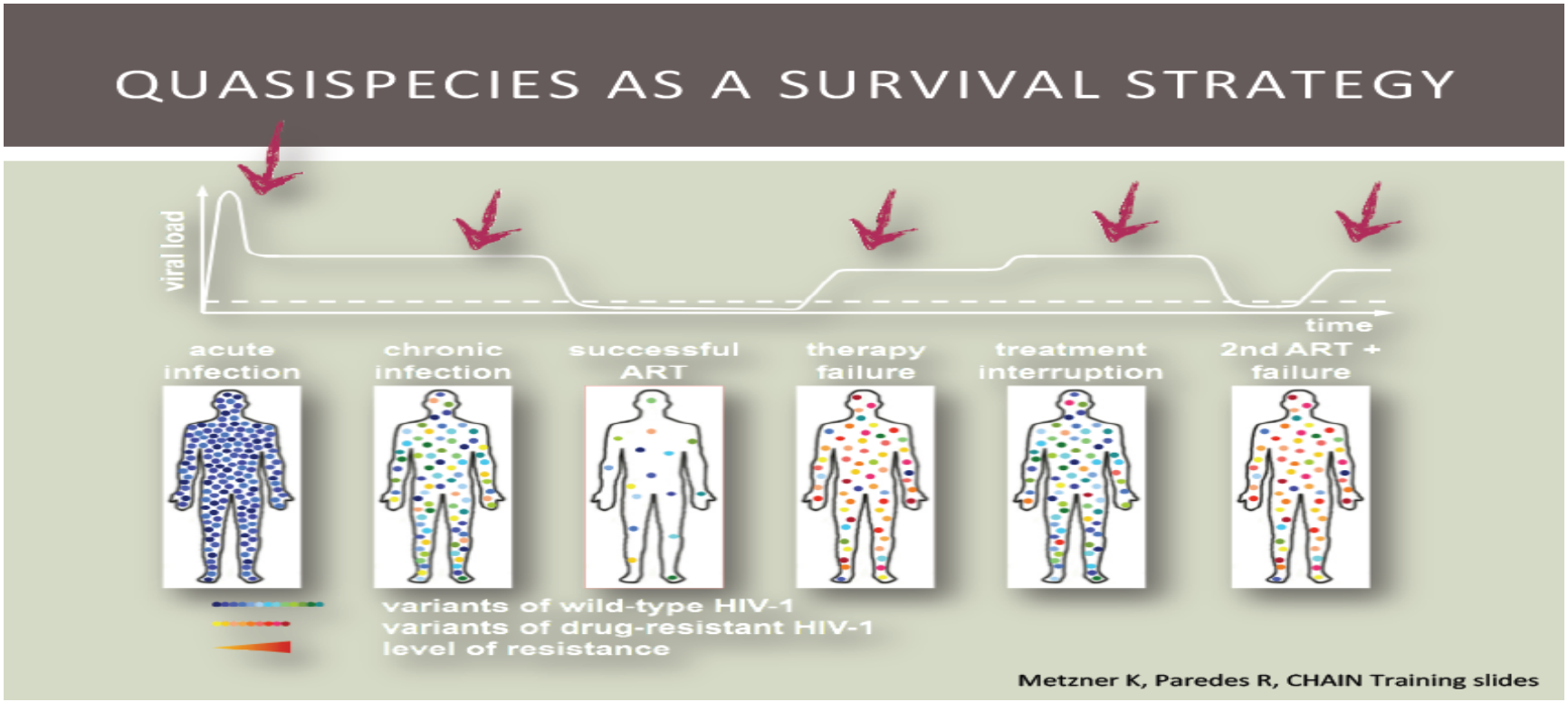


Figure 1. 11: Dynamics of HIV viral load within an infected patient. The viral load peaks at acute infection followed by a significant drop due to a lower CD4+ count. The introduction of highly active cocktail of anti retroviral drugs arrest the viral replication cycle and decreases the viral load to safe level. Spontaneous mutations give rise to resistant viruses that replicate with high turnover increasing the viral load and the therapy fails. The interruption of drugs contributes to a higher viral load. The introduction of second line anti retroviral drugs suppresses the viral load back to safe levels again and the cycle of viral load repeats itself. Source: Roger Paredes personal communication

A study by Richman et al (Richman et al., 1994a) 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 et al., 1994a).

Gupta et al (Gupta et al., 2008) 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 et al., 2008).

The error prone nature of the reverse transcriptase (di Marzo Veronese et al., 1993; Dumonceaux et al., 1998) and high turnover per replication per viral cycle (Mansky, 1996a; Mansky and Temin, 1995) are two major driving forces that result in multiple mutations conferring resistance to the drug (Kellam et al., 1994; Larder et al., 1991; Larder and Kemp, 1989; Tisdale et al., 1993).

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 (Collier et al., 1996; Montaner et al., 1998b; Staszewski et al., 1999b). 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 et al., 2000).

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 et al., 2008). 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 et al., 1999). 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 and D'Aquila, 2001).

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 and D'Aquila, 2001).

The limitations led to development of HIV resistance assays based on recombination of the virus from plasma samples (Hertogs et al., 1998; Kellam and Larder, 1994; Martinez-Picado et al., 1999; Petropoulos et al., 2000; Shi and Mellors, 1997). 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

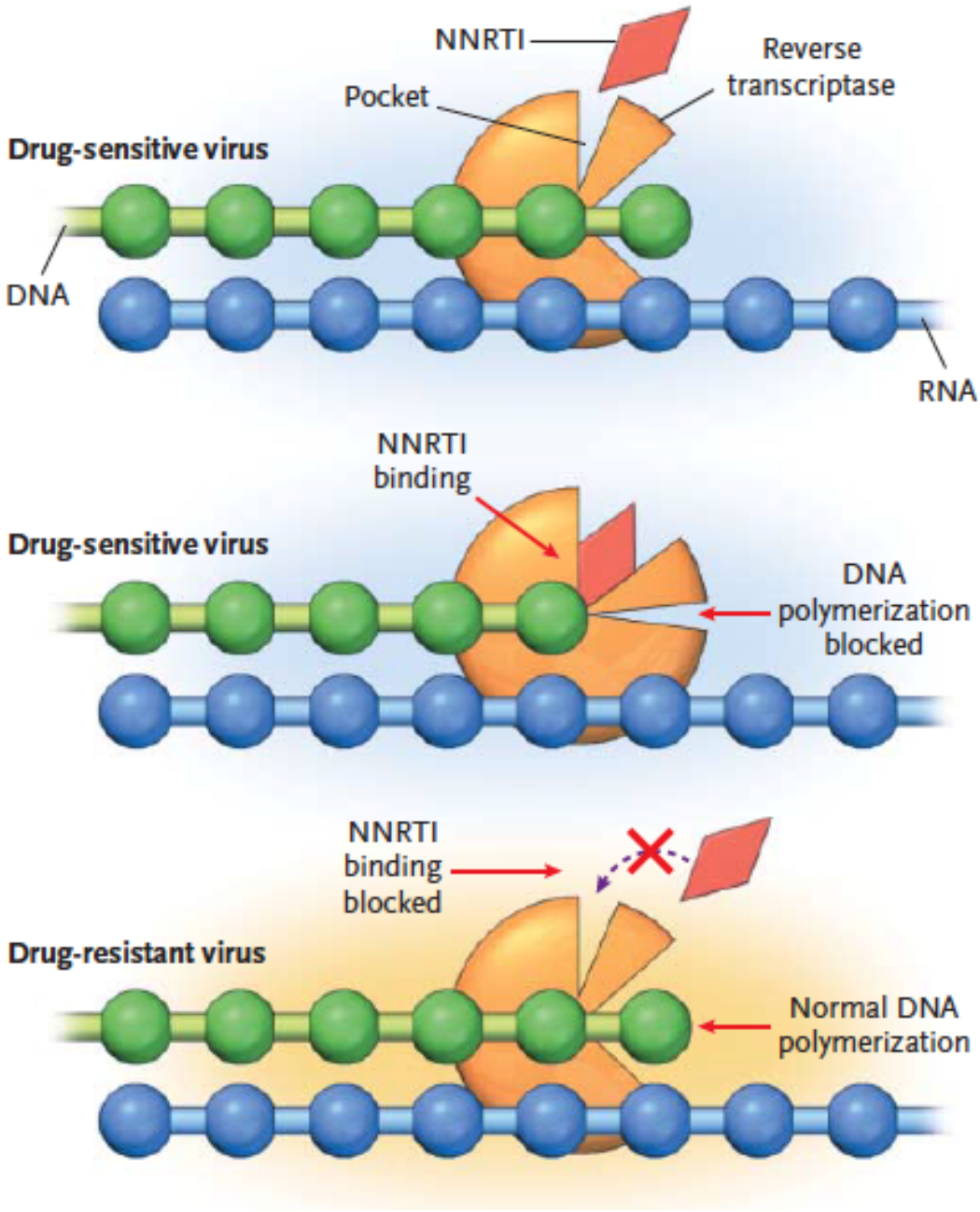


Figure 1. 12: Development of HIV drug resistance against NNRTI. The accumulation of drug resistant mutations changes three-dimensional confirmation of a viral protein, disabling drug binding and carrying out its normal function. Source: (Clavel and Hance, 2004)

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 et al., 1998) and PhenoSense assay (ViroLogic, South San Francisco, California) (Petropoulos et al., 2000) 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 et al., 2009). The assays involve complexities and are expensive (reviewed in (Hirsch MS, 2000)).

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 et al., 2003)) of known drug resistant mutations using a genotypic interpretation algorithm (Rhee et al., 2009). 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 (Larder et al., 1999; Mayer et al., 2001).

Sanger based technology (Sanger et al., 1977a) has been the standard for sequencing HIV-1 genes for drug resistance genotyping. Oligonucleotide hybridization based genotypic assays, as in GeneChip (Affymetrix) (Kozal et al., 1996) and LiPA

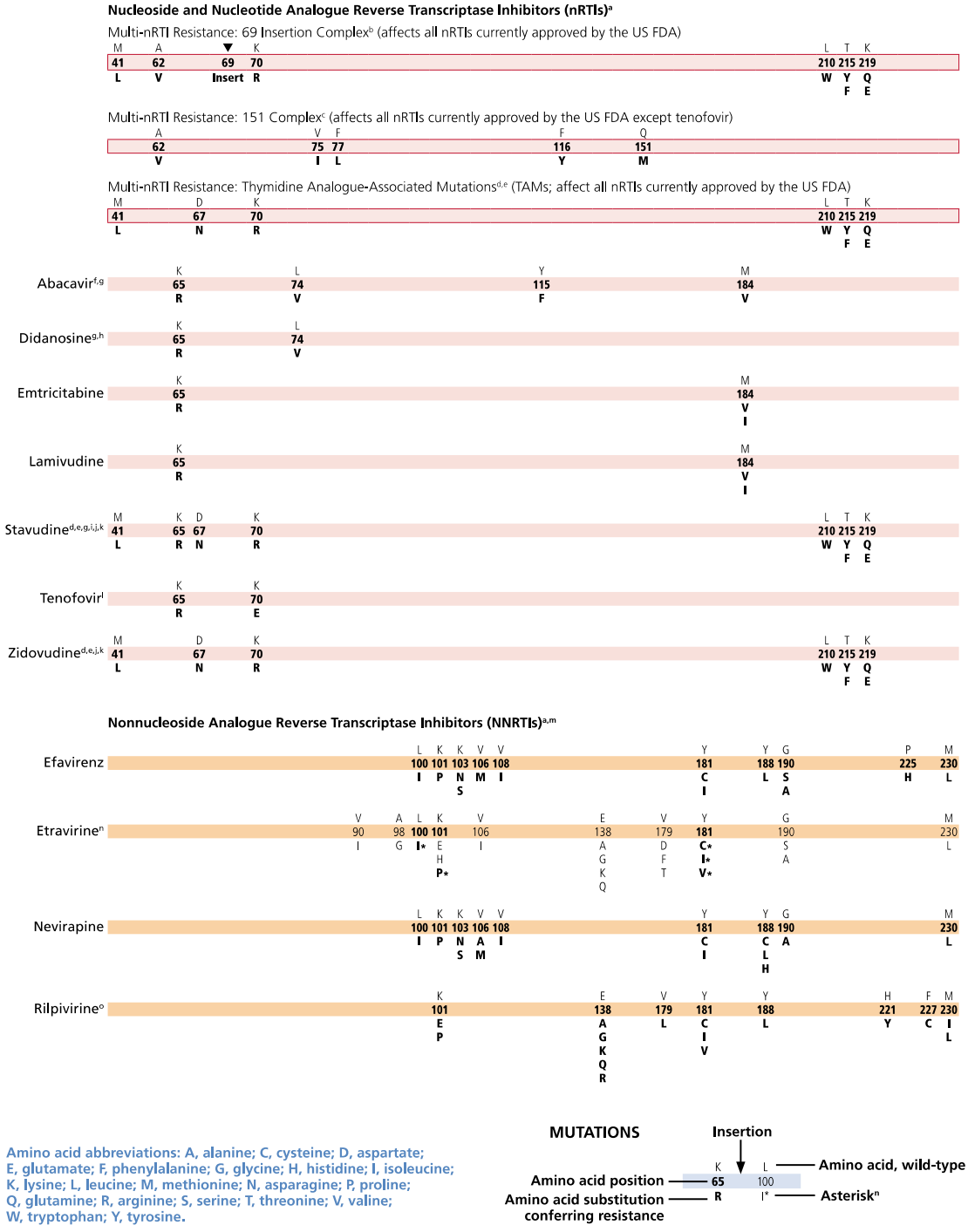


Figure 1. 13: Mutations in HIV-1 reverse transcriptase gene by codon positions that is associated with resistance to reverse transcriptase inhibitors. Source: Adapted from Johnson et al 2013 (Johnson et al., 2013)

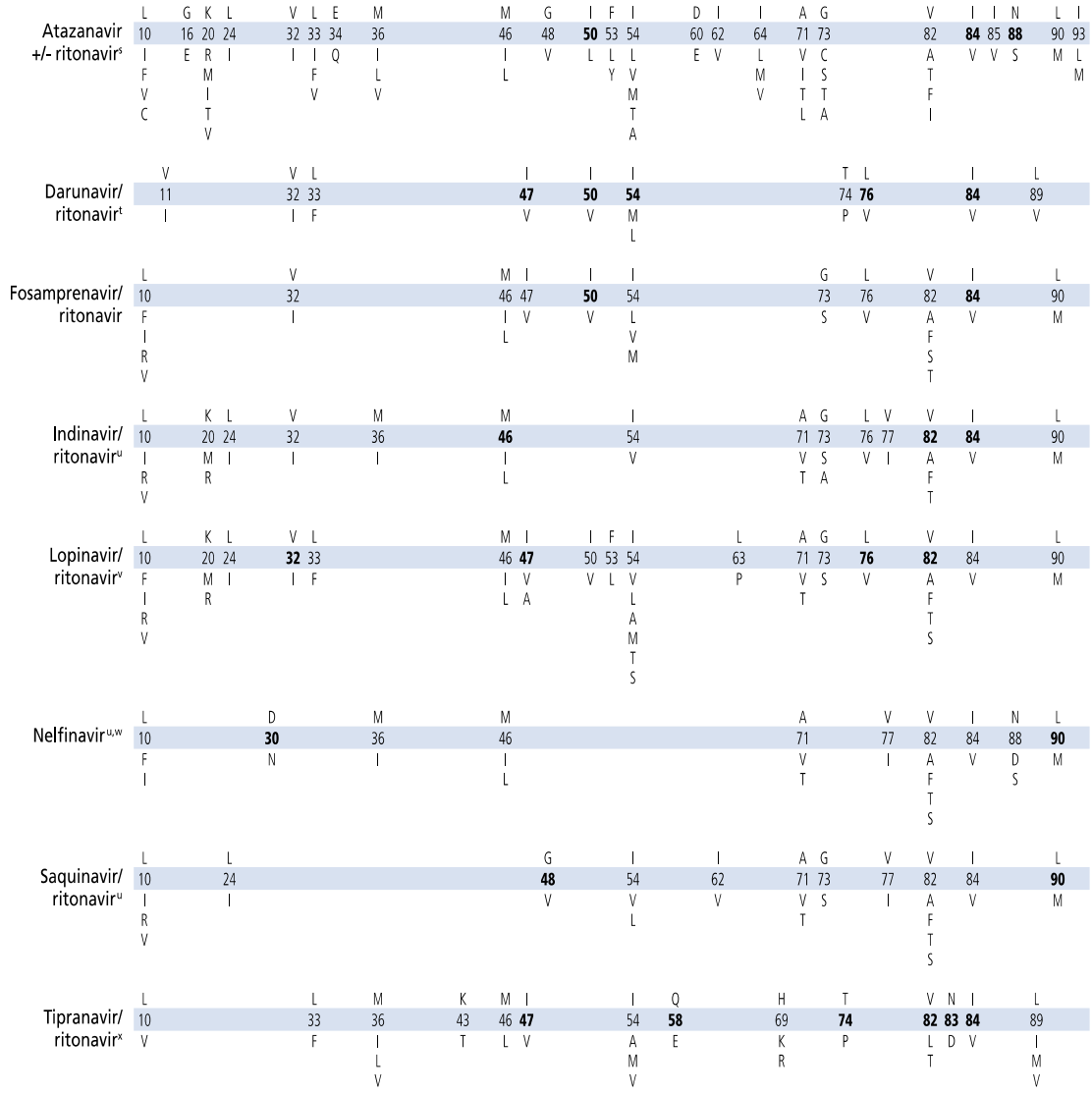


Figure 1. 14: Figure 1.14: Mutations in HIV-1 protease gene by codon positions that is associated with resistance to protease inhibitors. Source: Adapted from Johnson et al 2013 (Johnson et al., 2013)

(InnoGenetics) (Stuyver et al., 1997), were in used but limited to preselected drug resistant mutation codons.

1.11. HIV drug resistance genotyping

1.11.1 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 (Condra et al., 1995; Larder et al., 1989a; Larder and Kemp, 1989; Lorenzi et al., 1999; Molla et al., 1996; van Leeuwen et al., 1995; Zolopa et al., 1999) and characterization of these drug resistance mutations (DRMs) can be used to optimize the antiretroviral therapy (Baxter et al., 2000; Durant et al., 1999; Van Vaerenbergh, 2001). Conventional HIV genotyping involves Sanger dideoxy termination based population sequencing that produces consensus sequence of the viral population in a sample (Ewing et al., 1998; Metzker, 2005; Sanger et al., 1977b).

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 et al.). 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 et al., 2012). Thus, conventional population based sequencing method has limited

sensitivity; the low frequency variants below 20% prevalence are not detected reliably (Ji et al.; Johnson and Geretti; Palmer et al., 2005); and it underestimates the total number of variants in the viral population (Palmer et al., 2005). Undetected low frequency HIV variants have been shown to be clinically significant (Paredes et al., 2010; Rowley et al., 2010; Simen et al., 2007; Simen et al., 2009). These minor variants rebound in the presence of drugs, leading to ultimate virologic failure (Paredes et al.; Rowley et al.). Besides this sensitivity limitation, the conventional population based genotyping method is highly expensive (Dias-Neto et al., 2009; Liu et al., 2012) 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 et al., 2012; Wang et al., 2007).

1.11.2 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 (Metzker, 2009; Shendure and Ji, 2008a)).

Table 1. 2: Comparative analysis of different NGS systems. Source: adapted from Shokralla et al 2012 (Shokralla et al.), Niedringhaus et al 2011 (Niedringhaus et al.) and Glenn 2011 (Glenn, 2011)

Platform	Read length (bp)	Reads/run	Sequencing output/run	Run time	Advantages	Primary applications
Roche/454 GS FLX	400–500	1 x 10^6	≤500 Mb	10 h	Longest read lengths among 2nd generation, high throughput compared to 1st generation sequencing	1*, 2, 3*, 4, 7, 8*
Roche/454 GS FLX+	600–800	1 x 10^6	≤700 Mb	23 h		
Roche/454 GS Junior	400–450	1 x 10^5	~35 Mb	10 h		
Illumina HiSeq 2000	100–200	6 x 10^9	≤540–600 Gb	11 d	Very high throughput	1*, 2, 3*, 4, 5, 6, 7, 8
Illumina HiSeq 1000	100–200	3 x 10^9	≤270–300 Gb	8.5 d		
Illumina GAIIx	50–75	6.4 x 10^8	≤95 Gb	7.5–14.5 d		
Illumina MiSeq	100–150	7 x 10^6	≤1–2 Gb	19–27 h		
AB SOLiD 5500 system	35–75	2.4 x 10^9	~100 Gb	4 d	Very high throughput; lowest reagent cost needed to reassemble a human genome among the widely accepted 2nd generation platforms, lower error rate	3*, 5, 6, 8
AB SOLiD 5500 xl system	35–75	6 x 10^9	~250 Gb	7–8 d		
Ion Torrent -314 chip	100–200	1 x 10^6	≥10 Mb	3.5 h	Direct measurement of nucleotide incorporation events; DNA synthesis reaction operates under natural conditions (no need for modified DNA bases)	1, 2, 3, 4, 8
Ion Torrent -316 chip	100–200	6 x 10^6	≥100 Mb	4.7 h		
Ion Torrent -318 chip	100–200	11 x 10^6	≥1 Gb	5.5 h		

Bold indicates applications that are most often used, economical or growing
1 = de novo BACs, plastids, microbial genomes. 2 = transcriptome characterization.
3 = targeted re-sequencing. 4 = de novo plant and animal genomes.
5 = re-sequencing and transcript counting. 6 = mutation detection.
7 = metagenomics. 8 = other
*Pooling multiple samples with sequence tags (i.e. MIDs) is required for efficient use of this application

Roche/454 and Illumina implement a 'Sequencing by synthesis' (SBS) technique for DNA sequencing (Margulies et al., 2005b). 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 et al., 2005a). 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 et al., 2005b). 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 et al., 2012; Luo et al., 2012) at the rate of 0.38 per 100 bases (Loman et al., 2012). Illumina, on the other hand, stops the reaction after single nucleotide addition, detects the color of light emission that depends on a base (Bentley et al., 2008) but has base calling biases, generating substitution errors (Luo et al., 2012). 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 et al., 2008). 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, 2008b)).

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 et al., 2002; Milgrew et al., 2004). 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). 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 et al.). 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). Signal processing software is used to convert the data for measurement of base incorporations in that flow using a physical model (Rothberg et al., 2011). 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 et al., 2012).

Pacific Bioscience's Single Molecule Real Time technology is considered the third generation technology available in the market now (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 (Eid et al., 2009; McCarthy, 2010).

DNA sequencing takes place in the zero mode waveguide (ZMW) (Levene et al., 2003). ZMW is a nano-size chamber that is 7 nanometer in diameter and 10 nanometer in depth (McCarthy, 2010). 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 et al., 2012) showed that error rate of the SMRT technology was 13%. The number of sequence reads without any error was 0% (Quail et al., 2012). The accuracy of PacBio SMRT sequence reads is the least (~85%) among the sequencing platforms (reviewed in (Kumar et al., 2012)).

1.11.3 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 et al., 1998; Van Laethem et al., 1999), a true HIV diversity cannot be ascertained (Korn et al., 2003; Schuurman et al., 2002). Beside that the method is also expensive and labor 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 et al., 2008), allele-specific RT-PCR (Palmer et al., 2006) and single genome sequencing (Palmer et al., 2005) were developed to genotype HIV variants in the viral quasispecies 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 et al., 2010; Hoffmann et al., 2007; Jabara et al., 2011; Wang et al., 2007), 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 et al., 2007). 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 et al., 2009). In another similar study, Hoffmann and colleagues identified four additional minor drug resistant mutations with UDPS (Hoffmann et al., 2007). 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 et al., 2008). 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 et al., 2009). 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 et al., 2009).

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 et al., 2010). 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 et al., 2010).

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 et al., 2014). Although Wang et al (Wang et al., 2007) observed that UDPS genotyping had on average 0.0098% error rate (six times more errors in homopolymeric region than in non-homopolymeric region (Brodin et al., 2013; Wang et al., 2007)), 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 et al., 2012). 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 et al., 2008). 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 et al., 2012).

The genotyping step in UDPS method is preceded by PCR DNA amplification that could introduces errors such as DNA recombination (Kanagawa, 2003; Meyerhans et al., 1990; Yang et al., 1996), DNA synthesis errors (Hughes and Totten, 2003; Mansky and Temin, 1995) and DNA re-sampling errors (Liu et al., 1996). 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 et al., 2005b; Shendure and Ji, 2008). Thus, Roche/454 pyrosequencing genotypic data has insertion/deletion errors high at homopolymer regions (Huse et al., 2007; Wang et al., 2007). 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 (Brenner et al., 2006; Doualla-Bell et al., 2006; Geretti et al., 2009; Hosseinipour et al., 2009).

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 et al., 2011). 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 and 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 et al., 2011). With the application of the Primer ID technology, Jabara et al (Jabara et al.,

2011) were able to resolve 80% of the unique sequence polymorphisms that were different than consensus sequence from conventional Sanger genotype method.

1.12. Chapter Outlines

Chapter 2: This chapter introduces and describes a novel algorithm, QTrim, designed for the quality trimming of UDPS sequence data with Phred-like quality scores (e.g. Roche/454, Illumina and Ion Torrent) and the evaluation of its performance in comparison to other widely used tools. The results show that QTrim is comparable to the next best tool while quality trimming a good quality data set and outperforms all the tools while trimming a poor quality data set. The tool has been published in BMC Bioinformatics (Shrestha, RK, Lubinsky, B, Bansode, VB, Moinz, MB, McCormack, GP, Travers, SA (2014) QTrim: a novel tool for the quality trimming of sequence reads generated using the Roche/454 sequencing platform. *BMC Bioinformatics* **15**: 33).

Chapter 3: This chapter introduces and discusses the application of primer ID technology developed by Jabara and colleagues (Jabara et al., 2011) to reduce the PCR and sequencing errors. We developed a module to facilitate the analysis of HIV drug resistance genotyping data generated using the primer ID technology in the Seq2Res drug resistance-testing pipeline. We describe the workflow of the primer ID module and discuss the limitations of the technology including primer ID collision and underrepresentation of the HIV variants in the viral population.

Chapter 4: This chapter introduces and describes the Seq2Res computational pipeline that facilitates low cost HIV drug resistance through easy analysis of HIV drug resistance genotyping data generated using UDPS sequencing technologies. The

chapter describes a workflow of the pipeline, the requirements of the pipeline, HIV drug resistance output files and plots that summaries overall analysis. We evaluated and validated drug resistant mutation calls in the Seq2Res pipeline in comparison with the Stanford HIV drug resistance interpretation algorithm using two biological datasets downloaded from the Stanford Database. We validated the prevalence calls of drug resistant mutations in the Seq2Res pipeline using five simulation datasets with known prevalence of known drug resistant mutations. We observed that regardless of the prevalence level of the drug resistant mutations in the dataset, Seq2Res is capable of accurately identifying their presence at the correct prevalence level.

Chapter 5: In this chapter we present the application of Seq2Res to analyze HIV drug resistance genotyping sequence data generated using the Roche/454 Junior platform and the Roche/454 FLX platform from HIV infected individuals sampled as part of CIPRA-SA (Comprehensive International Program for Research in AIDS in South Africa) study. We compared the results with those obtained using Standard population based consensus sequencing and observed that:

1. The sensitivity of the Roche/454 Junior and Roche/454 FLX platforms are comparable for HIV drug resistance genotyping.
2. Ultra deep pyrosequencing is, at least, comparable to conventional population based Sanger method at HIV drug resistance genotyping
3. Resistance to nevirapine is significantly more likely to be observed in individuals previously exposed to ARVs through PMTCT than in drug naïve individuals.
4. At 15% and below (both FLX and Junior), the prediction of NVP resistance significantly correlates ($p < 0.05$) with time since NVP exposure

Chapter 6: This chapter summary the significance of the work, development of Seq2Res pipeline, validation of the pipeline and the application of the pipeline

CHAPTER 2

QTrim – A Novel Algorithm for Quality Trimming high throughput Sequence Data

2.1 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, 2008a). 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.

2.2 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) (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 has 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 these criteria are 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.

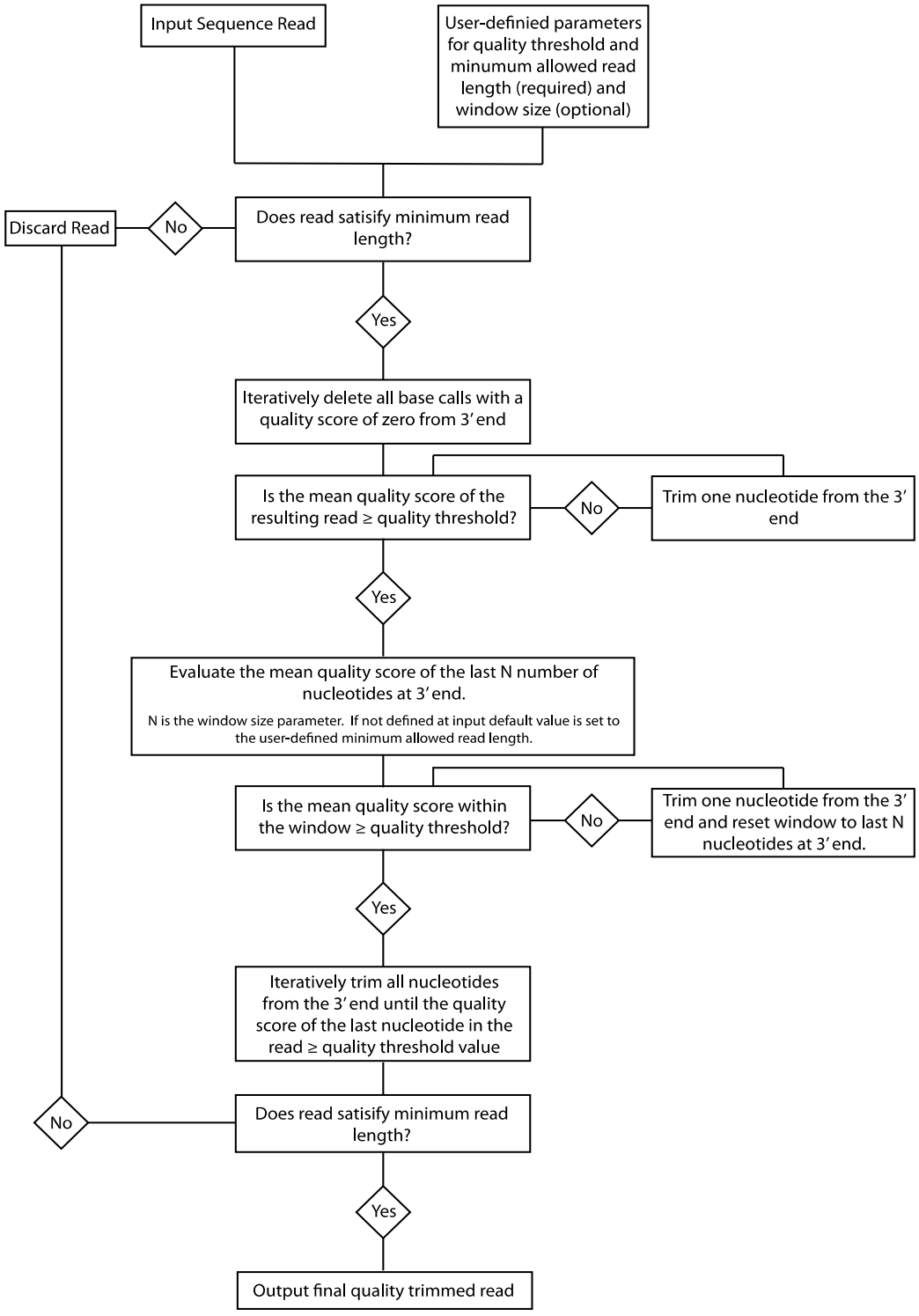


Figure 2. 1: Systematic work flow of QTrim

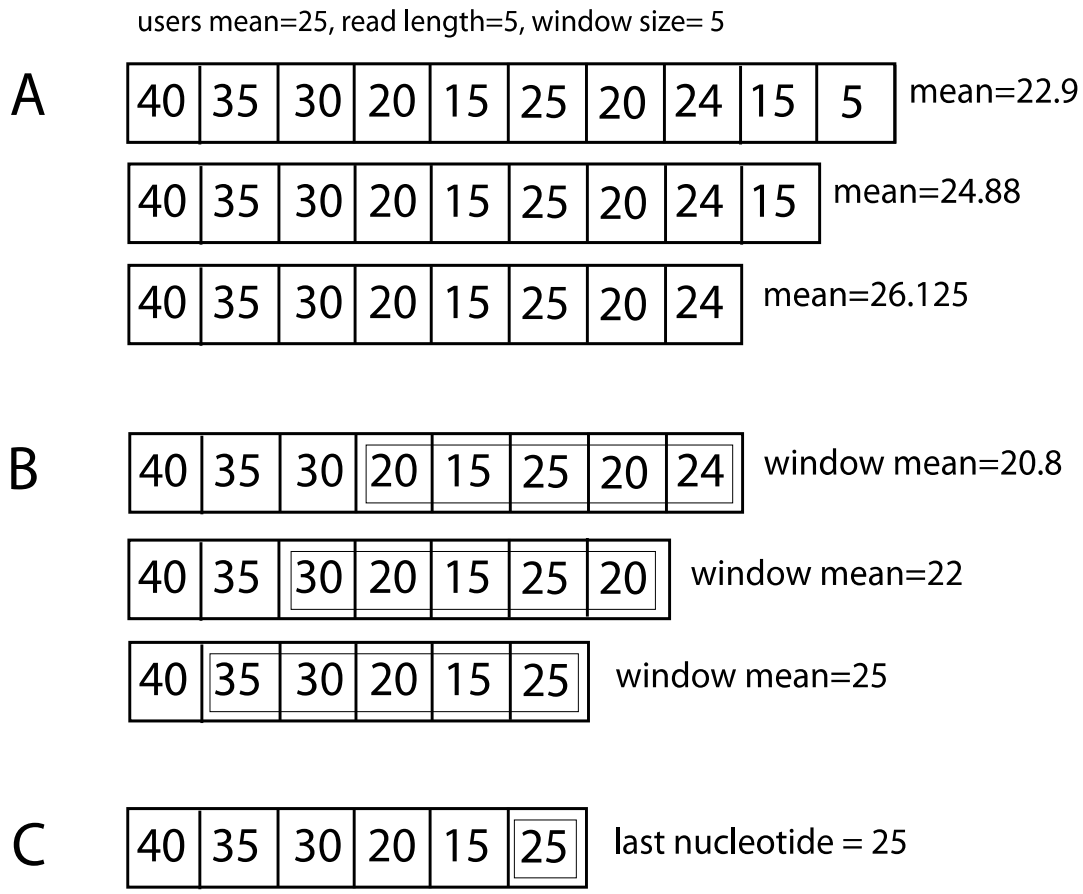


Figure 2. 2: Sequential steps describing the trimming process in QTrim. A) Iterative trimming with mean across sequence. B) Iterative trimming with mean in window. C) Iterative trimming with mean in last nucleotide. At step C, the last nucleotide has a score of 25, which satisfies the user mean and QTrim therefore does no further trimming. The final read length is also greater than the user-specified read length. Therefore, the sequence is written to an output file.

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.

3. 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.

(A) Mode 1: Trimming from 3' end with removal of interspersed ambiguous bases (Ns) in the reads.

(B) Mode 2: Trimming from 3' end without removal of interspersed ambiguous bases (Ns) in the reads. This is the default mode.

(C) Mode 3: Trimming from 5' and 3' ends with removal of interspersed ambiguous bases (Ns) in the reads.

(D) Mode 4: Trimming from 5' and 3' ends without removal of interspersed ambiguous bases (Ns) in the reads.

2.2.1 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.

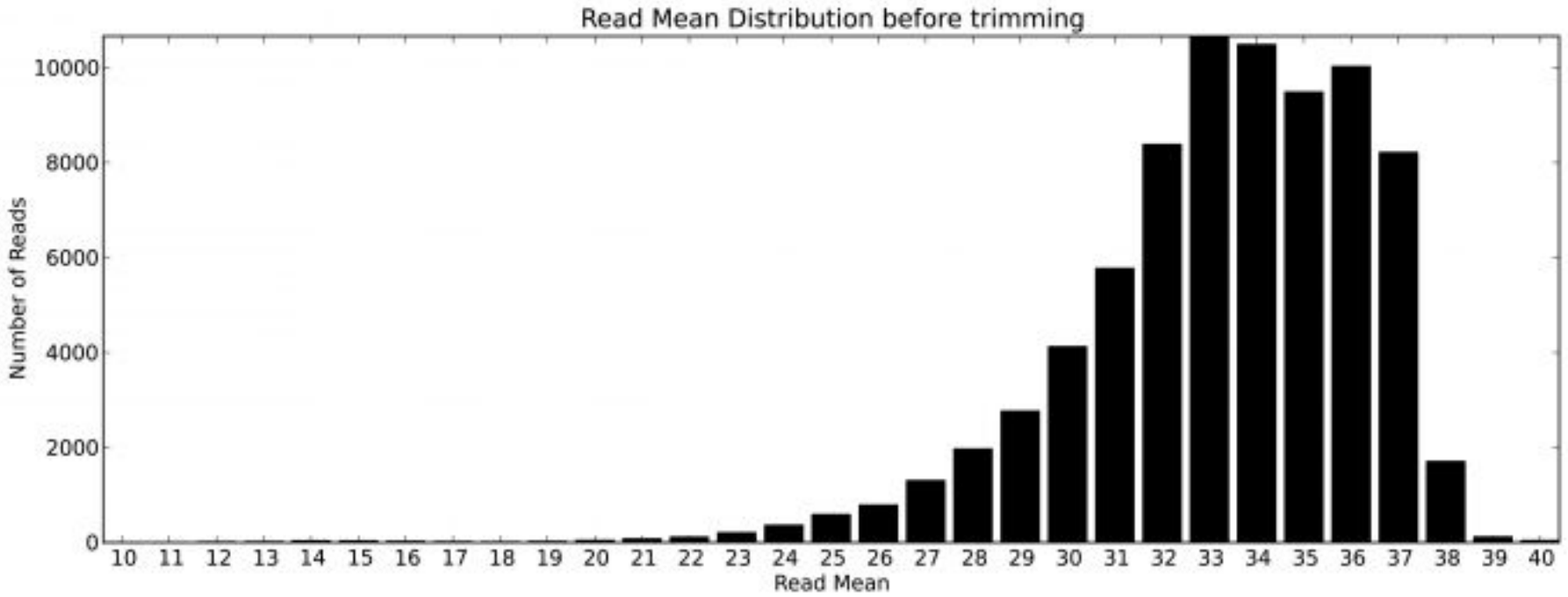


Figure 2. 3: QTrim generated plot showing the distribution of sequence reads with specific mean score.

2. **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 lengths for poor quality data are generally very variable.

3. **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 10^{th} base position of every read. A box and whisker plot showing this information is plotted as shown in **Figure 2.5**. 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.

2.2.2 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

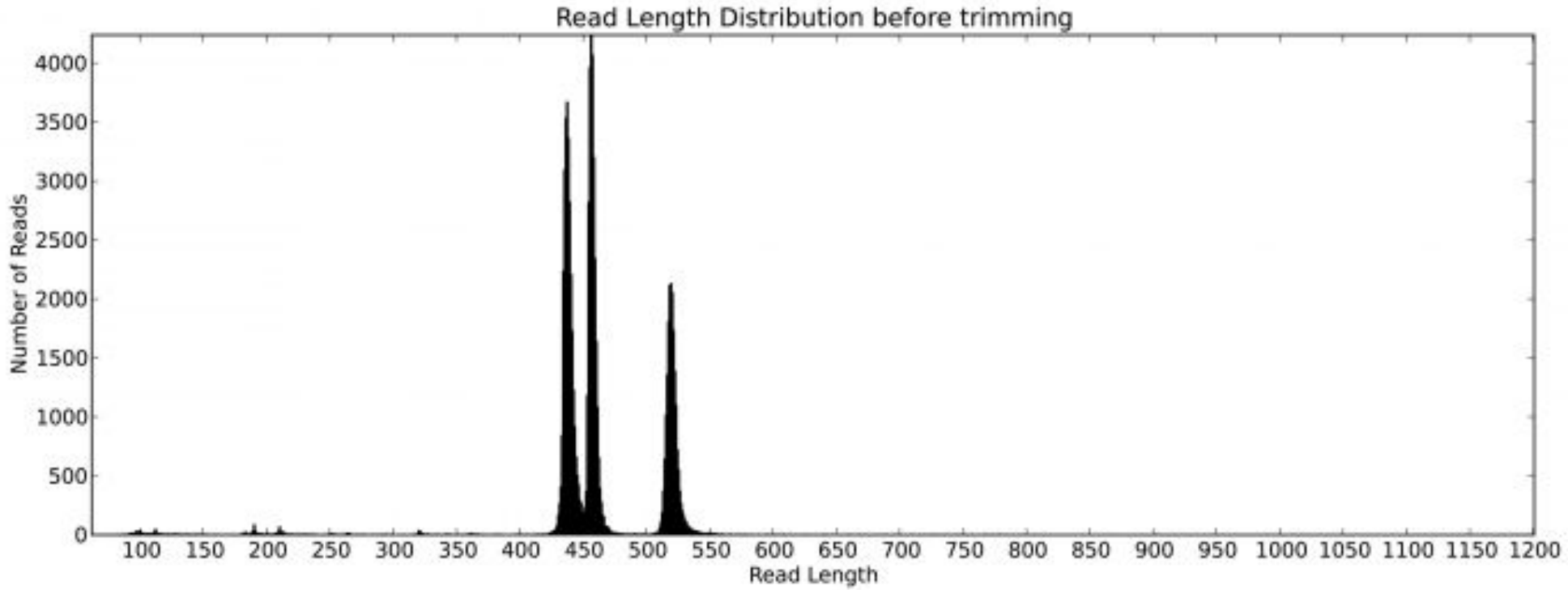


Figure 2. 4: QTrim generated plot showing the distribution of sequence reads with read lengths.

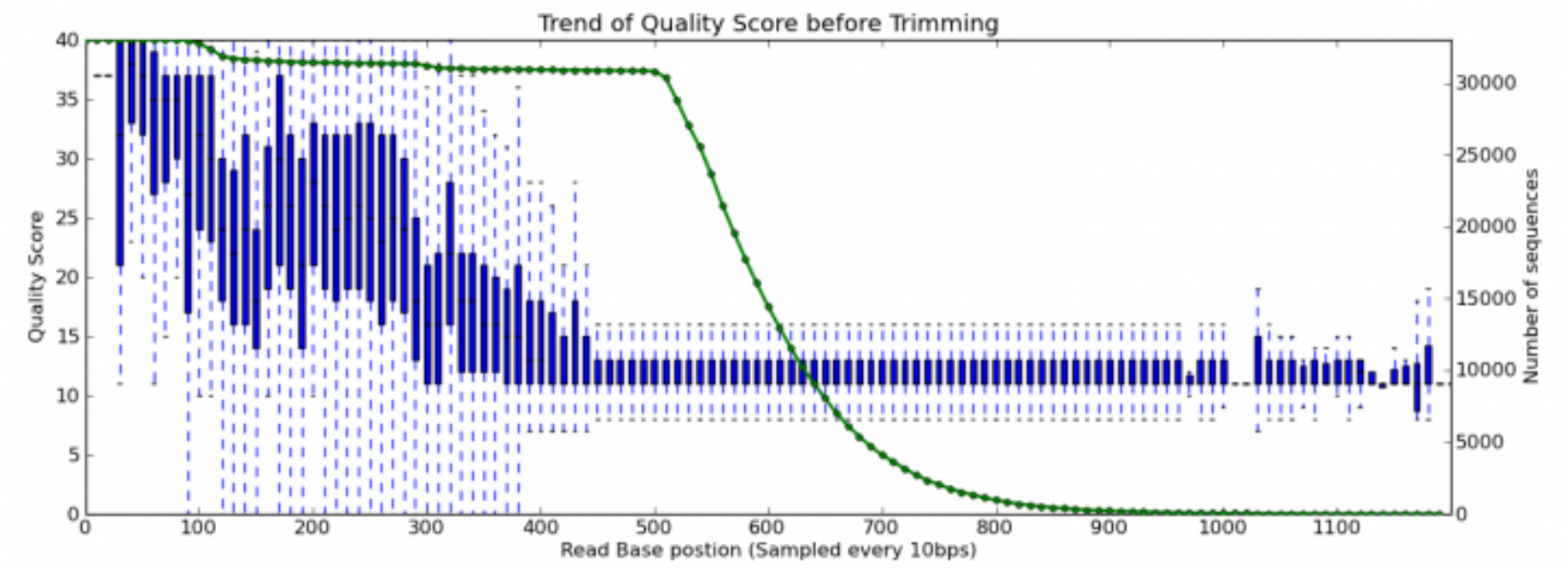


Figure 2. 5: QTrim generated box plot showing the average quality score across the sequence reads and the number of sequences contributing to the calculation of the average quality score at every 10^{th} base (Bin size: 10). The green line represents the total number of sequences (represented at the secondary Y axis at the right side) for evaluating the average quality score.

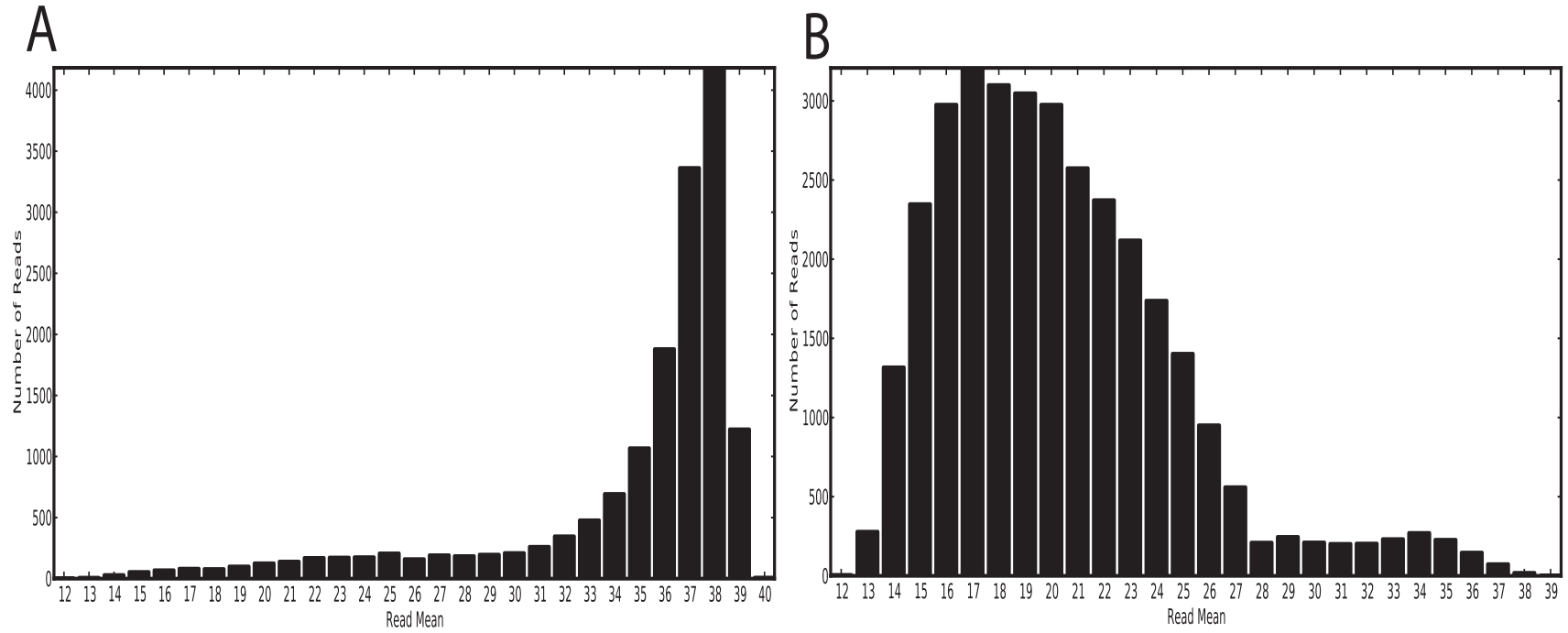


Figure 2. 6: Distribution of test data sequence reads by mean quality. A) Large number of sequences in test data A have higher mean quality score B) Large number of sequences in test data B have lower mean quality score

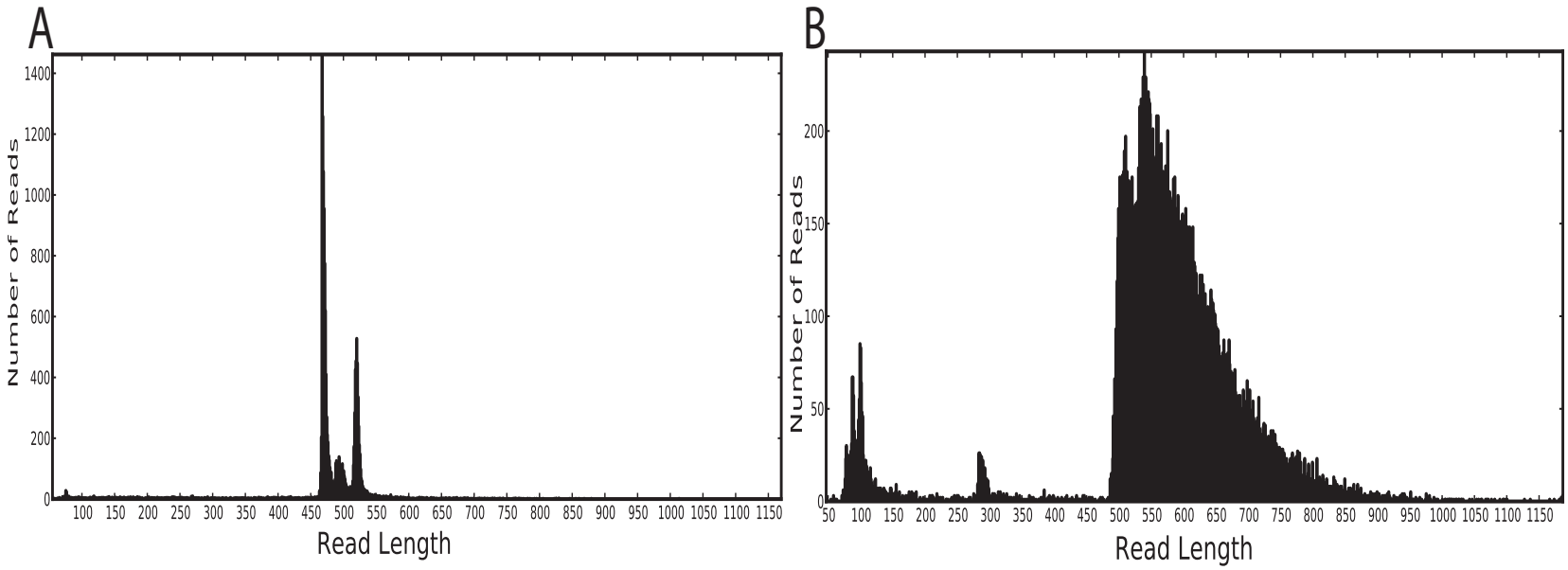


Figure 2. 7: Distribution of test data sequence reads by read length. A) Close range of from 450 to 550 read length in test data A and the plot shows very close distribution of sequences by read length. B) Wide range of read length in test data B and the plot shows wide distribution of sequences by read length

untrimmed sequence lengths is further evidence of the poor quality data present in dataset B (**Figure 2.7 B**).

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.

2.3 Results

When applied to the good quality dataset, QTrim and PRINSEQ performed at an equivalent level (**Figure 2.8** A and C; **Table 2.2** Q20 and Table 2.3) and outperformed all the other methods (**Figure 2.8** 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

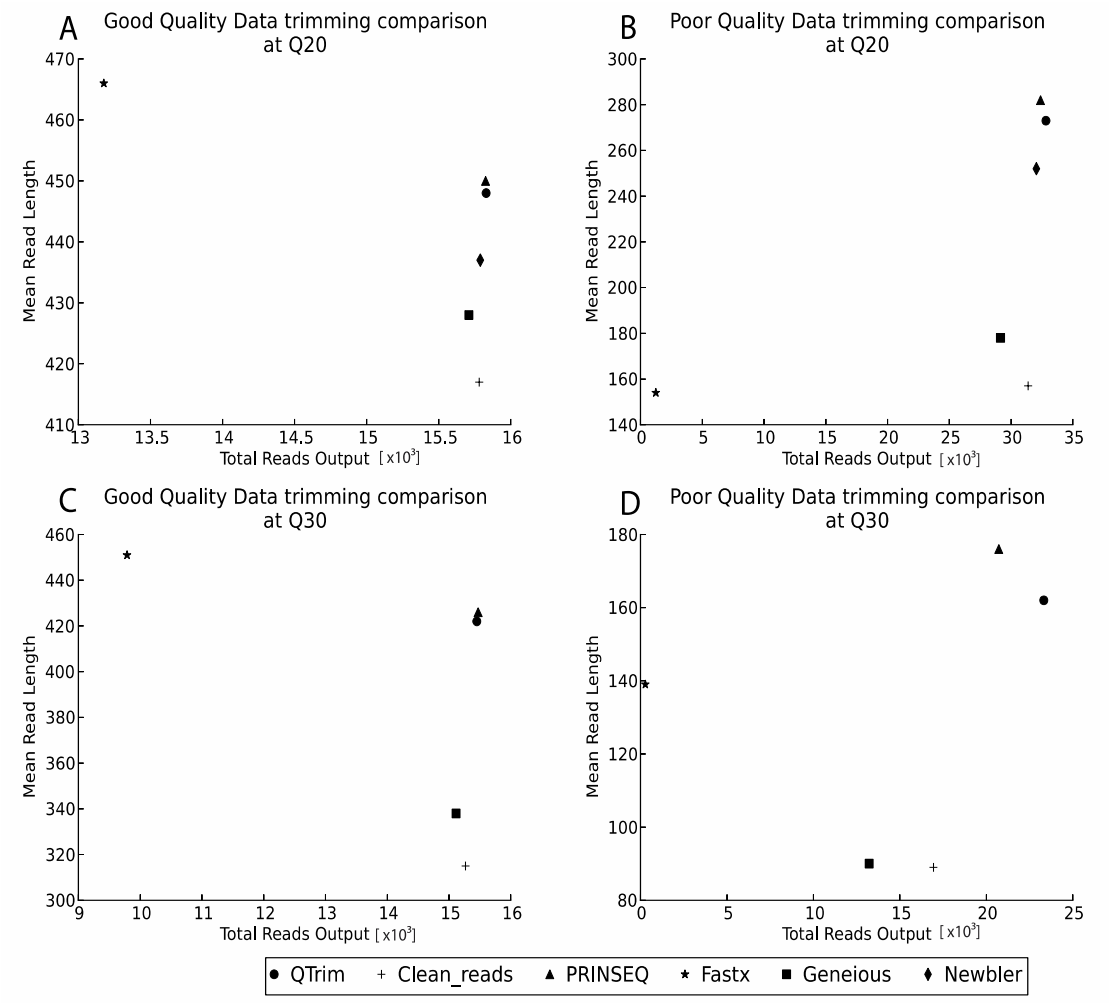


Figure 2. 8:Comparison of QTrim with other methods taking into account the total number of produced reads and mean read length of produced reads using two datasets. The good quality data is presented in A and C, and the poor quality data in B and D. The top panel (A and B) displays the results at a mean threshold quality score of 20, whereas the bottom panel shows the results at a mean threshold of 30.

Table 2. 1: Quality trimming of the good quality data with QTrim and other tested methods at a mean quality of 20 (Q20) and 30 (Q30) and minimum read length of 50. The table shows the total sequence reads, mean read length, percentage of total bases in output and percentage of poor quality bases in the output file by all methods.

	Total output reads		Mean read length		Post trimming % of bases		Post trimming % of poor quality bases	
	Q20	Q30	Q20	Q30	Q20	Q30	Q20	Q30
QTrim	15829	15450	448	422	94.305	86.754	4.227	1.723
Clean_reads	15781	15267	417	315	87.582	64.025	2.127	0.030
PRINSEQ	15825	15471	450	426	94.613	87.744	4.562	2.129
FASTX	13176	9787	466	450	81.561	58.555	2.113	0.841
Geneious	15709	15113	428	338	89.335	68.032	2.003	0.102
Roche's Newbler v2.6	15788	NA	437	NA	91.763	NA	2.831	NA

NA: The data is not available as Roche's Newbler v.26 is preset to trim at Q20 only. Users do not have option to trim at Q30

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.

2.3.1 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

Table 2. 2: Quality trimming of the poor quality data with QTrim and other tested methods at a mean quality of 20 (Q20) and 30 (Q30) and minimum read length of 50. The table shows the total sequence reads, mean read length, percentage of total bases in output and percentage of poor quality bases in the output file by all methods.

	Total output reads		Mean read length		Post trimming % of bases		Post trimming % of poor quality bases	
	Q20	Q30	Q20	Q30	Q20	Q30	Q20	Q30
QTrim	32818	23321	273	162	47.788	20.118	23.445	5.849
Clean_reads	31379	16940	157	89	26.340	8.031	10.822	0.036
PRINSEQ	32381	20717	282	176	48.570	19.437	24.969	7.096
FASTX	1242	279	154	139	1.020	0.207	5.970	1.318
Geneious	29142	13218	178	90	27.592	6.343	8.417	0.017
Roche's Newbler v2.6	32047	NA	252	NA	42.964	NA	21.573	NA

NA: The data is not available as Roche's Newbler v.26 is preset to trim at Q20 only. Users do not have option to trim at Q30

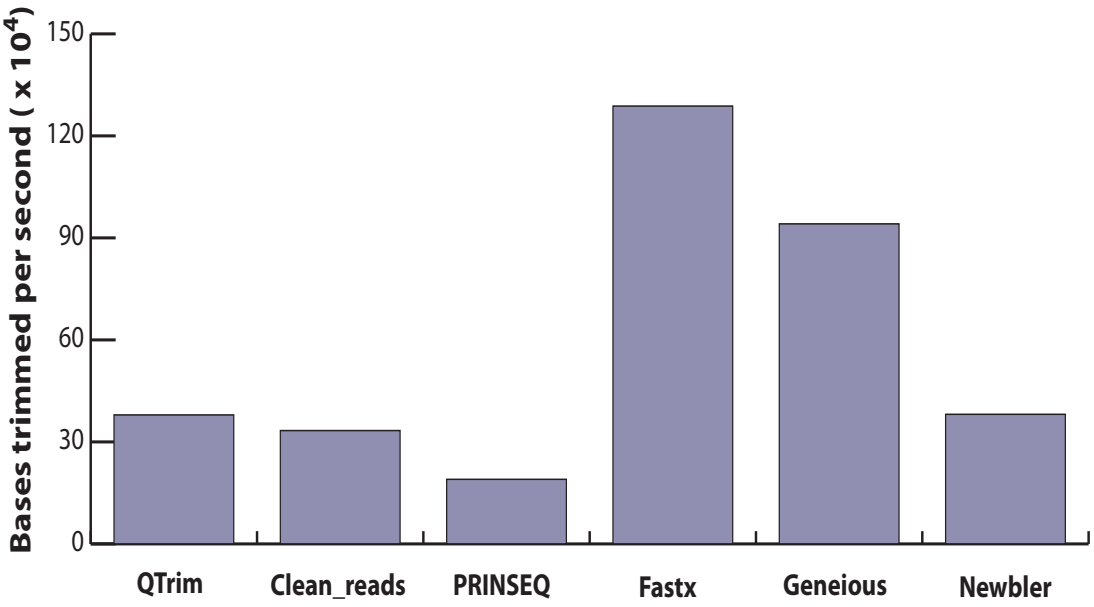


Figure 2. 9: Speed comparison of all the tools with the number of bases trimmed per second. Geneious and Newbler v2.6 are graphical tools and their time of execution is done manually with stopwatch. Other tools are command line and the time is obtained using command line “time” function.

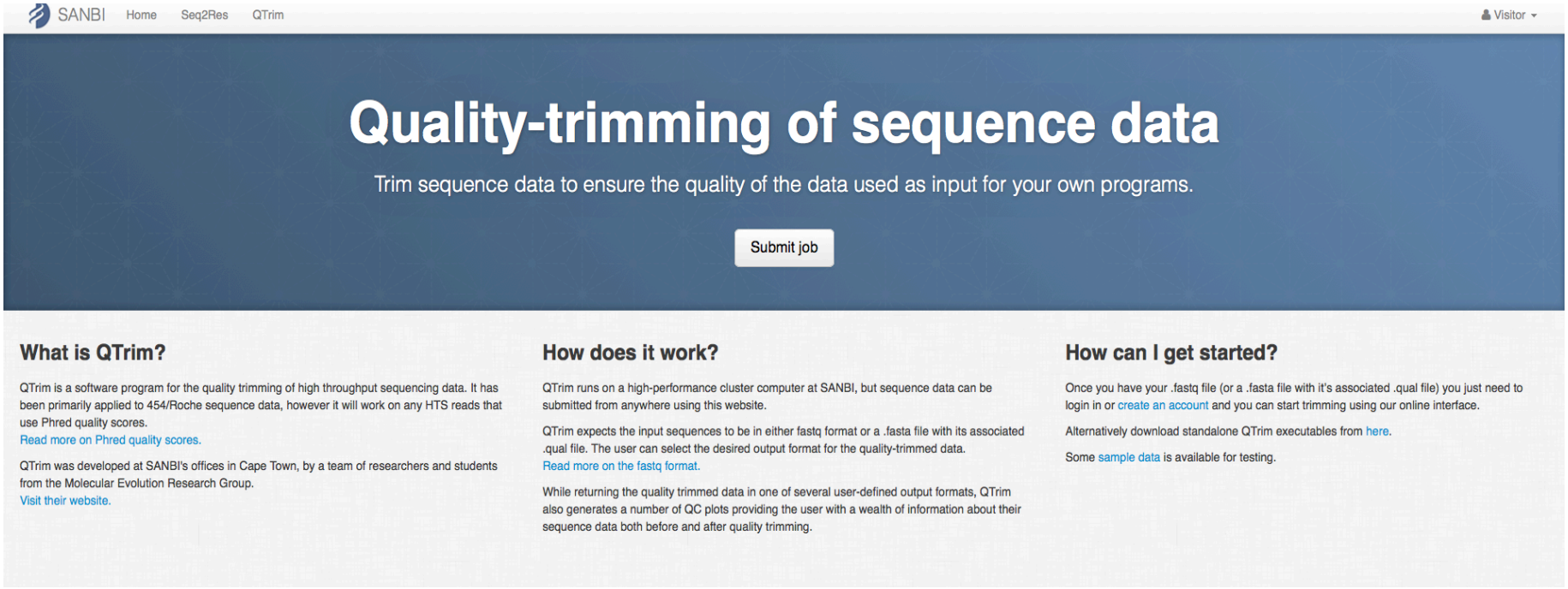


Figure 2. 10: Online QTrim home page. Users need to be registered to access and submit sequence file for quality trimming with QTrim.

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

2.4 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

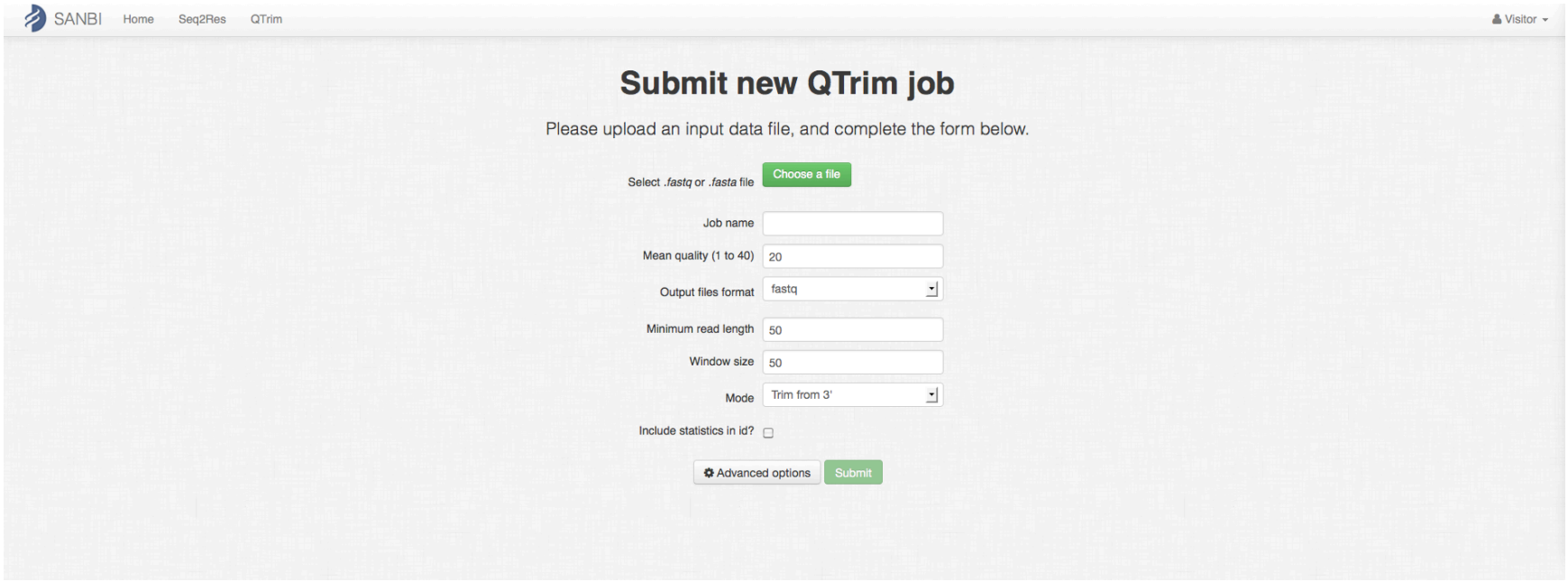


Figure 2. 11: Online QTrim job submission page. Sequence file can be uploaded and required parameters can be set before running QTrim online.

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 3

Primer ID Algorithm PIDA – Algorithm for processing Ultra-Deep High Throughput Sequence Data generated using Primer ID technology

3.1 Introduction

Ultra Deep PyroSequencing (UDPS) platforms are capable of generating as much as millions of sequence reads from DNA fragments at low cost and in less time than other sequencing approaches. This capability enables the potential to fully characterize viral quasispecies including the low frequency variants (Fischer et al., 2010; Hoffmann et al., 2007; Mitsuya et al., 2008; Rozera et al., 2009; Varghese et al., 2009; Wang et al., 2007) However, a high rate of sequencing errors are incorporated and accumulated at the PCR amplification step (Kanagawa, 2003) and by instrumental/hardware error and sequencing errors like nucleotide insertion and deletion errors (reviewed in (Metzker, 2009)). These errors inflate and confound the real genetic diversity in the population (Kunin et al., 2009; Zagordi et al.). The errors generated at PCR step are: **1)** incorporation of wrong nucleotide by polymerase

enzyme during many cycles of amplification (Hughes and Totten, 2003; Kanagawa, 2003) **2)** recombination of two DNA fragments producing a new chimeric DNA (Judo et al., 1998; Meyerhans et al., 1990; Yang et al., 1996) **3)** differential amplification of DNA fragments change the ratio before and after PCR step, obscuring true original sample diversity (Liu et al., 1996; Polz and Cavanaugh, 1998). In order to avoid modification and over diversification of the original sample and to ensure downstream results are truly reflective of the actual viral diversity, it is essential to correct those inevitable errors.

The accurate quantification of low abundance drug resistant HIV viruses, in particular, may be substantially improved by the implementation of the primer ID approach as described by Jabara and colleagues (Jabara et al., 2011). Primer ID is a unique identifier that is annealed to a viral cDNA during the reverse transcription step from RNA to DNA to enable identification of the viral template from which each individual sequence reads originated. A number of studies have already used this approach for such purposes (Beerenwinkel et al., 2012; Eisele and Siliciano, 2012; Jabara et al., 2011; Schmitt et al., 2012). While the original publication describes the development of an algorithm to analyze the complex data output from primer ID-based sequencing (Jabara et al., 2011), this code has not been made available to the public and is unlikely to be made so in the near future (Cassandra B. Jabara, personal communication). Thus, to facilitate the analysis of HIV drug resistance sequence data generated using the primer ID approach in the Seq2Res resistance testing computational pipeline, this chapter describes the development and application of a such a tool.

3.2 Methods and Materials

3.2.1 Raw sequence reads containing primer ID

The structure of a raw sequence read containing a primer ID is shown in Figure 3. A set of primer ID sequences of custom length (usually 8) is generated randomly. The number of primer IDs depends on the length of primer ID. A set of primer ID of length 8 has 65536 (4^8) unique nucleotide combinations. A single primer ID is embedded within the primer used in the cDNA synthesis. The set of primer ID and primers creates random library of sequences. A cDNA primer binds to a viral RNA and extends from 3' end to generate a cDNA that now contains a primer ID tag. A unique multiplex identifier (MID) sequence per sample with a spacer sequence at both ends of the MID is prepared. The cDNA with primer ID at 5' end is then attached to the 3' end of the MID with spacer sequences. A PCR priming site sequence is then added to the 5' end.

The cDNA sequence with primer ID, MID and PCR priming site is then PCR amplified to produce millions of sequences. The primer ID is copied through the PCR steps thus enabling tracking of the viral template from which each sequence derived. During PCR amplification a forward read is extended through the target sequence followed by the primer ID, MID and PCR priming site. In reverse sequences, however, the primer is extended the target sequence is the last part to be extended following the primer ID, MID and PCR priming site. Correct assignment of forward sequences requires that the full fragment has been amplified and sequenced in order to cover the primer ID sequence at the end of the read (**Figure 3.1 B**) however reverse sequences

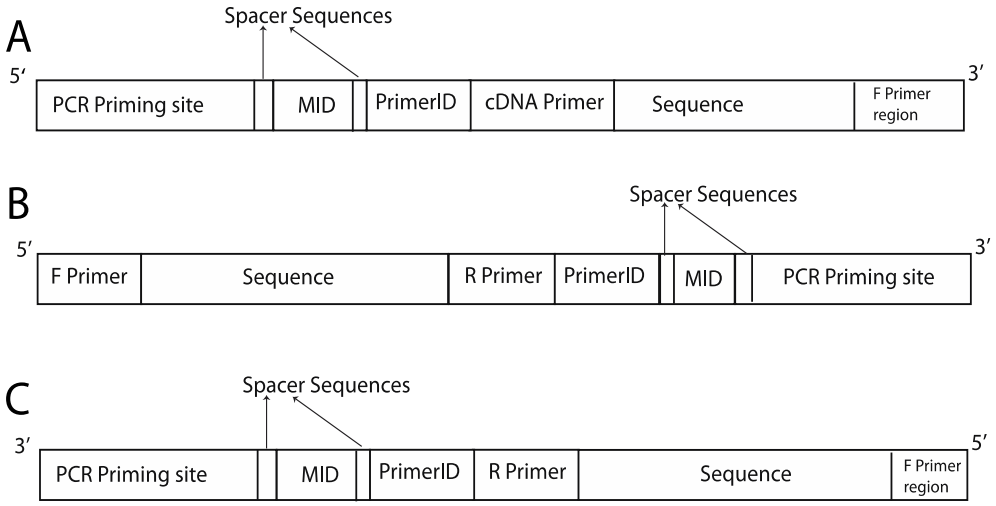


Figure 3. 1: The structure of Primer ID raw sequence data. A) cDNA structure with locations of PCR priming site, spacer sequences, MID and Primer ID. B) A forward strand Primer ID raw sequence read produced in PCR step. C) A reverse sequence produced in the PCR step. The primers tag the amplicon region, which is being amplified, MID tags the specific individual from whom the sample is obtained and Primer ID tags the template RNA sequence. Two spacer sequences are used to separate the MID sequence from Primer ID and PCR priming site. The nucleotide sequence and length of spacers are customary and may vary between the projects.

do not need to be full length in order to assign the read to a particular template on the basis of the primer ID (**Figure 3.1 C**).

3.2.2 Processing primer ID data using PIDA

A novel algorithm - Primer ID algorithm (PIDA), was developed for integration into the Seq2Res pipeline to facilitate fast and accurate processing of sequence reads generated using the primer ID approach. The algorithm requires the raw sequence reads input in FASTQ format while other required files contain information about the primers used, the multiplex identifiers used (if present) and the minimum allowed read lengths.

The primer file is a five column tab delimited file containing the amplicon name in the first column followed by the forward and reverse primer sequences in the second and third column while the fourth and fifth columns contain start and end nucleotide positions set by first nucleotide position of forward primer and last nucleotide position of reverse primer relative to the standard HIV *pol* reference sequence. (**Figure 3.2 A**).

In instances when multiple samples have been sequenced together on the same sequencing plate, each sample is tagged with a unique MID sequence. In order to interpret these, the user must provide a tab-delimited file with the MID name in the first column and a unique patient identifier in the second one (**Figure 3.2 B**). When the standard Roche MIDs are not used then the MID number can be replaced with the MID sequence.

A

Amplicon specific Primer file

Name	Forward	Reverse	start	end
GeneA	ATAGAG	GGATGA	x1	y1
GeneB	ATGGAT	ACACTAG	x2	y2

B

Sample ID/MID file

Name	patientname
MID1	PatientA
MID2	PatientB

C

Gene file

Name	Forward	Reverse
GeneA	10	50
GeneB	10	30

Figure 3. 2:Different files required as input for the PIDA algorithm. Each column in all files has to be separated by a single tab. A) The amplicon specific primer file contains gene names that are amplified; the forward and reverse primers used for each amplicon; and the start and end nucleotide positions of the targeted amplicon region relative to standard HIV *pol* reference sequence. B) The MID file contains names of MID sequence used to tag the samples and sample names for identification. C) The gene file contains the amplicon names and minimum required sequence length for forward and reverse sequences. The information displayed in the files is just for the purpose and can be changed as required.

Note: x1, y1, x2 and y2 can be replaced with numbers that represent the amplicon start and end positions.

In some instances the end-user may only be interested in subsequent analysis of a short amplicon fragment located within an amplicon thereby enabling non full-length sequences to be analyzed. Thus, we allow the user to define the minimum read length required for both the forward and reverse sequences for each amplicon. The gene file details these lengths with the amplicon name in the first column followed by the forward and reverse sequence minimum read lengths in columns two and three respectively (**Figure 3.2 C**).

The other information that the end-users are required to supply are the universal PCR primer sequence and the format of the sequence containing primer ID, spacers, MID and PCR primer that was prepared for cDNA production. For example, a user may input the format as primerid8.cg.mid5.tga.primingsite, which indicates a primer ID of length 8 nucleotides, a spacer sequences 'cg', MID sequence of length 5 nucleotides, another spacer sequence 'tga' and followed by the word 'primingsite'. Users also have options to choose:

A) Threshold number of sequences required to generate consensus sequence
B) Maximum mismatches allowed between a user supplied primer and primer region in a sequence read. This is defined as primer tolerance.
C) Maximum mismatches allowed between a user-supplied MID and the sequence in the MID region of a sequence read. This is defined as MID tolerance.

The steps of processing the raw data into consensus sequences in the algorithm are discussed in detail below:

3.2.3 Sequence Demultiplex using tag sequences

For each sequence read, the information provided in primer file and MID file is used to identify the amplicon and sample to which the read belongs (**Figure 3.3**, red text). The algorithm begins with a search for the forward primer at 5' end of the read. A subsequence of length equal to forward primer is obtained from forward primer region at 5' end, which is then pair-wise aligned with all forward primers one at a time. If the number of mismatches in pair-wise alignment is less or equal to, the primer tolerance, the sequence read amplicon is designated as being identified with the aligned forward primer and the search for reverse primer is skipped. If none of the forward primers are identified within the read the algorithm searches for the presence of each of the reverse primers. A subsequence of length equal to the reverse primer is obtained from the reverse primer region and is pair-wise aligned with every reverse primer one at a time. If a reverse primer that is aligned with the number of mismatches less or equal to primer tolerance, the sequence read amplicon is identified with the aligned reverse primer. All sequences in which a reverse primer is identified are reverse complemented to ensure all subsequent analysis is performed on sequences in the same strand orientation. If neither forward nor reverse primers are found, sequence reads are discarded.

Following identification of a sequence read's source amplicon, PIDA identifies the MID associated with that read and bins all reads with the same primer and MID together for downstream analysis.

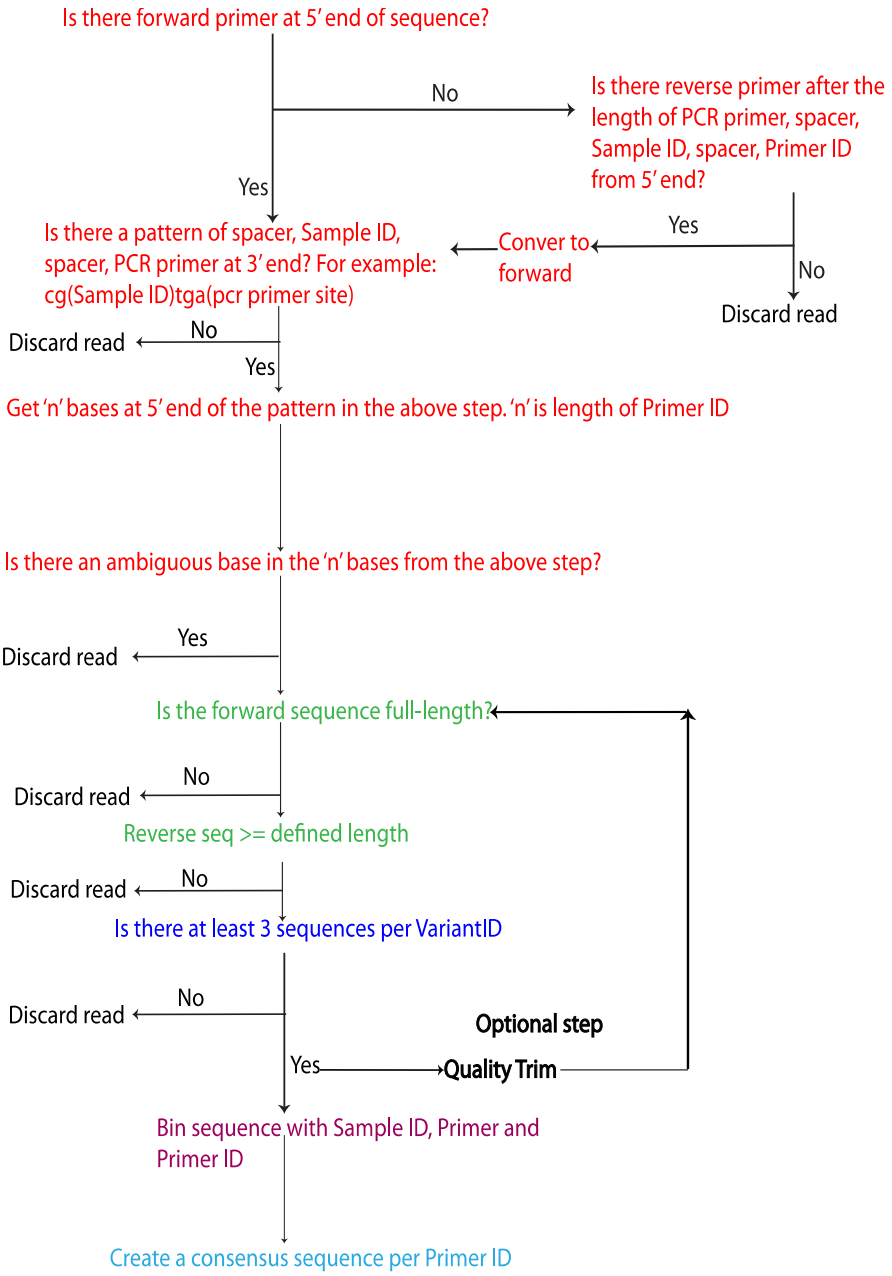


Figure 3. 3: The logical flow of the PIDA algorithm to process raw sequence data generated with Primer ID technology into generation of consensus sequence. The colors denote PIDA processing steps. Red: sequence demultiplexing, Green: Sequence filter by length, Blue: sequence filter by number of sequence represent by each Primer ID, Black: Quality trimming, Purple: Sequencing binning, Sky blue: Consensus sequence generation

A subsequence from MID region of a sequence read is extracted and pair-wise aligned with list of supplied MID sequences one at a time. If the number of mismatches between a MID and the subsequence is less or equal to defined tolerance, the sample for the associated MID is identified for the sequence read or discarded if none of the MIDs match with the obtained MID from the sequence.

Once a matching MID is found, a sequence of 'n' nucleotides is obtained as an primer ID sequence, where 'n' is the length of primer ID, from the sequence read region at the 5' end of MID and spacer sequence. The read is discarded if there is a presence of an ambiguous base in the obtained 'n' nucleotides. The tags - Primer, MID and primer ID - are then added at sequence id for further downstream processing.

3.2.4 Selection of sequences with threshold length

With sequence reads containing the primer ID sequence information at the 5' end (Figure 3.1C) the entire target sequence plus the downstream information must be sequenced in order for the important MID and primer ID information to be identified and the sequence read retained for further analysis. Conversely, when the primer ID information is contained at the 3' end, the entire target sequence does not need to be sequenced to identify the read based upon the primer ID and MID motifs. Therefore, if the entire query sequence is not required for downstream analysis, the user can set a parameter to define the minimum length of query sequence that must sequenced to retain a non full-length reverse read. (**Figure 3.3** green text).

3.2.5 Selection of Primer IDs with minimum number of sequences

The default number of representative sequences required for a single primer ID to generate a consensus sequence is three as recommended in the original publication to avoid ambiguous bases (Jabara et al., 2011). However, end-users have the option to set this value as required. Only the primer IDs with a number of representative sequences greater than, or equal to, the defined threshold number is passed through for subsequent analysis (**Figure 3.3** blue text).

3.2.6 Quality trimming

Quality trimming is an optional step in the algorithm. If the user selects the trimming option, the algorithm uses QTrim (Shrestha et al., 2014) to quality trim the non-discarded sequence reads. As quality trimming can change the sequence length once it is complete the PIDA algorithm repeats the previous two steps of the process to ensure that the quality trimmed reads are of a sufficient length and quantity for consensus sequence generation (**Figure 3.3** black text).

3.2.7 Generating a Consensus Sequence

For each primer ID the representative sequences are binned together (**Figure 3.3** purple text) and are aligned to each other using MAFFT (Katoh et al., 2005; Katoh et al., 2002; Katoh and Toh, 2008, 2010). A consensus sequence is generated from the resulting alignment by calling the most common nucleotide that at each position in the

alignment **(Figure 3.3** Sky blue text). In the case of ties in frequency between two or more bases at a position, an ambiguous base representing the bases is added to the consensus sequence. To avoid large numbers of ambiguous bases in a consensus sequence users can define an odd number as the minimum number of sequences.

3.2.8 Test datasets

Two datasets (described here as Run1 and Run2) were generated by our collaborators (Prof Carolyn Williamson's research group, University of Cape Town) using the Primer ID approach and were used here to evaluate PIDA. Each dataset comprised one sequencing run (Roche/454 Junior plate) containing data from four HIV infected patients from a study to analyze vaccine response in HIV subtype C (the results from this study do not comprise part of this thesis and will be published elsewhere). For each patient, four amplicons covering one region in the envelope gene (*env)*, two regions of gag (gag54, *gag*472) and one region in the nef gene (*nef*23) were amplified using subtype C specific primers with each primer used in the cDNA generation step associated with a unique primer ID. Amplicons for each patient were tagged using a unique MID and sequence data was generated using one Roche/454 Junior system for each dataset.

3.3 Results

3.3.1 Initial demultiplexing

Runs 1 and Run2 were analyzed independently of each other using PIDA. Before any processing of the data was undertaken the total number of raw sequence reads for each dataset was 125,865 and 40,544 for Runs 1 and 2 respectively.

The initial demultiplexing step involved identifying sequence reads that had incomplete/missing primer or MID motifs and discarding them from subsequent analysis. For Run 1 the primer was not found in 2962 (2.4%) reads while 7557 (6%) had no MID present. Similarly, for Run 2 the primer was not found in 751 (1.9%) of reads while 1109 (2.74%) had no MID present (**Table 3.1**).

For the remaining reads that contained complete primers and MIDs, we examined the sequence of the primer ID and retained reads that did not contain ambiguous bases at primer ID region. There were 257 (0.2%) and 37 (0.1%) reads with ambiguous bases in primer ID in Run1 and Run2 respectively and were discarded (**Table 3.1**)

Thus, following initial demultiplexing, 91% of sequences from Run1 were retained, while 95% of the sequences from Run2 were passed through for subsequent analysis.

3.3.2 Quality trimming of sequence data and sequence length evaluation

All sequence reads were quality trimmed using QTrim (Shrestha et al., 2014) with a mean quality score of 20, resulting in a loss of 2950 (2.34%) and 14 (0.034%) sequences from Run1 and Run2 respectively as a result of poor quality. For forward reads the MID and primer ID sequences are located at the 3' end and, thus, the sequenced amplicons must be full length in order to extract all necessary information

Table 3. 1: Analysis of raw sequence data before accounting Primer ID for downstream analysis.

	Run1	Run2
Total raw reads	125865	40544
Total read discarded for no primer	2962	751
Total reads discarded for no MID	7557	1109
Total reads discarded for ambiguous base in Primer ID	257	37
Total reads after demultiplexing	115089	38647

for downstream analysis. For reverse reads, however, the required information is located at the 5' end and, thus, full-length sequence reads are not always necessary as the region of interest may be close to the 5' end. Therefore, the required read lengths to define reads that could be passed through to the next stage of analysis (i.e. they cover the region of interest in the query sequence) were different for forward and reverse sequence reads (**Table 3.2**)

Using these parameters forward and reverse sequence read lengths were analyzed independently of each other and those sequences that were shorter than the required read length were removed from the analysis. 32 (0.03%) and 23 (0.06%) sequences from Runs1 and Run2 respectively were discarded as short reads in the trimming step (**Table 3.3**).

For all subsequent processing and analysis, sequence reads from each amplicon and patient were binned together resulting in 16 unique datasets generated from each sequencing run (**Table 3.4**). Each of these datasets was subsequently analyzed independently of all others.

3.3.3 Characterization of primer IDs

For each dataset, the number of unique primer ID tags contained in the data was counted. A wide range of unique primer ID tags was observed between the various datasets ranging from 110 for the gag472 amplicon in the patient E to 4193 for the *env* amplicon in the patient B (**Table 3.5**)

Table 3. 2 : Amplicon specific forward and reverse sequence read length of interest

ENV		GAG54		GAG472		NEF23	
Forward	Reverse	Forward	Reverse	Forward	Reverse	Forward	Reverse
377	160	315	162	318	160	364	177

Table 3. 3: Number of reads discarded for not being full length

		ENV		GAG54		GAG472		NEF23	
		F	R	F	R	F	R	F	R
Run 1	Patient A	0	0	0	0	0	0	0	0
	Patient B	0	10	0	2	0	7	0	13
	Patient C	0	0	0	0	0	0	0	0
	Patient D	0	0	0	0	0	0	0	0
Run 1	Patient E	0	0	0	3	0	0	0	2
	Patient F	0	4	0	0	0	0	0	7
	Patient G	0	1	0	2	0	1	0	0
	Patient H	0	0	0	1	0	0	0	2

Table 3. 4: Number of reads retained for being full length

		ENV		GAG54		GAG472		NEF23	
		F	R	F	R	F	R	F	R
Run 1	Patient A	7239	13422	946	3036	1197	2988	1350	1940
	Patient B	3780	6159	1490	2493	2326	4272	2119	2846
	Patient C	5158	7556	2687	4675	2054	2865	2112	2999
	Patient D	4038	5709	1861	3380	2544	3782	2354	3589
Run 2	Patient E	862	1915	457	741	524	1997	22	2024
	Patient F	1118	2374	658	243	312	1023	1357	1320
	Patient G	1445	4020	671	726	363	1756	1123	1998
	Patient H	2066	3603	83	712	211	1394	8	1498

Table 3. 5: Total number of unique Primer ID tags in each dataset.

		ENV	GAG54	GAG472	NEF23
Run 1	Patient A	1064	1051	232	282
	Patient B	4193	3031	4097	470
	Patient C	750	422	228	479
	Patient D	1030	1710	145	374
Run 2	Patient E	253	590	110	546
	Patient F	515	424	115	126
	Patient G	2105	1048	660	930
	Patient H	607	279	229	157

For each dataset the number of sequence reads tagged with each unique primer ID in that dataset were grouped together and counted. We observed that the number of representative sequences for each primer ID tag ranged from one to 4144 (**Figure 3.4-3.7**). The minimum number of sequences required to generate a consensus sequence representing a primer ID tag is three and, thus, for each dataset we separated primer ID tags with two or less sequences from those with three or more representative sequences. We observed that the percentage of primer ID tags with less than three representative sequences was, on average, three times higher than those with three or more representative sequences (**Table 3.6**).

While only three sequences are required to generate a consensus sequence, we find that for each dataset the average number of sequence reads for each primer ID is significantly greater than three (**Table 3.7**). Patient G contained an average of six sequences per primer ID for each amplicon with the remainder of patients averaging between 18 and 88 representative sequences per primer ID.

3.3.4 Generation of consensus sequences

Consensus sequences were only generated for those primer ID tags with three or more representative sequences (**Figure 3.8**). Generation of consensus sequences showed that while there may be a large number of sequences representing a particular amplicon for a patient (**Figure 3.8 A and B**), this could comprise data representing a small number of primer IDs meaning that the resulting number of consensus sequences was, in fact, quite low (**Figure 3.8 C and D**). For example, Patient A's *env* amplicon in Run1 had the highest total number of sequence reads (19700,

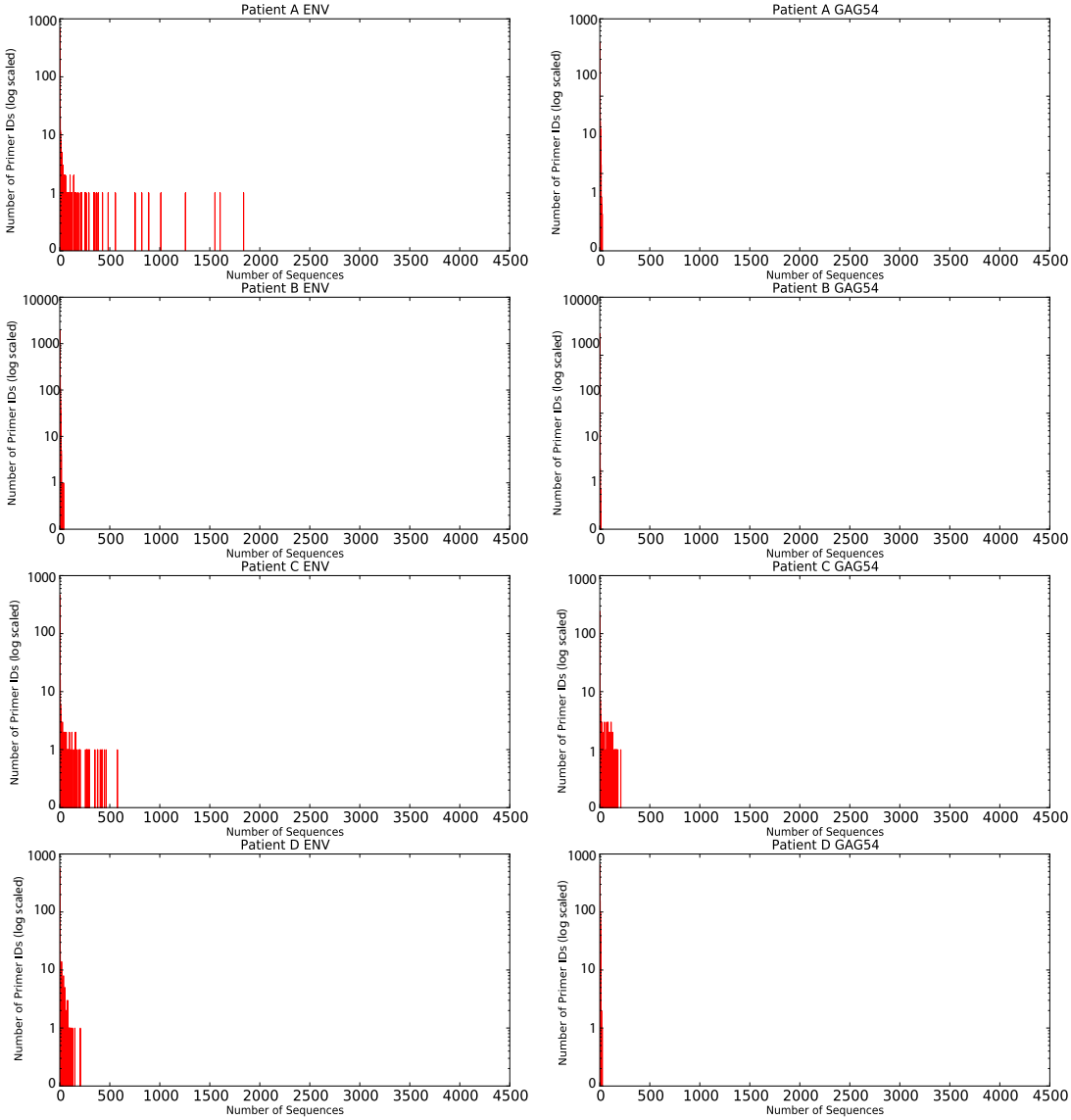


Figure 3. 4: Count of Primer IDs representing specific number of sequence reads in *ENV* and GAG54 amplicons of Run1 Patients.

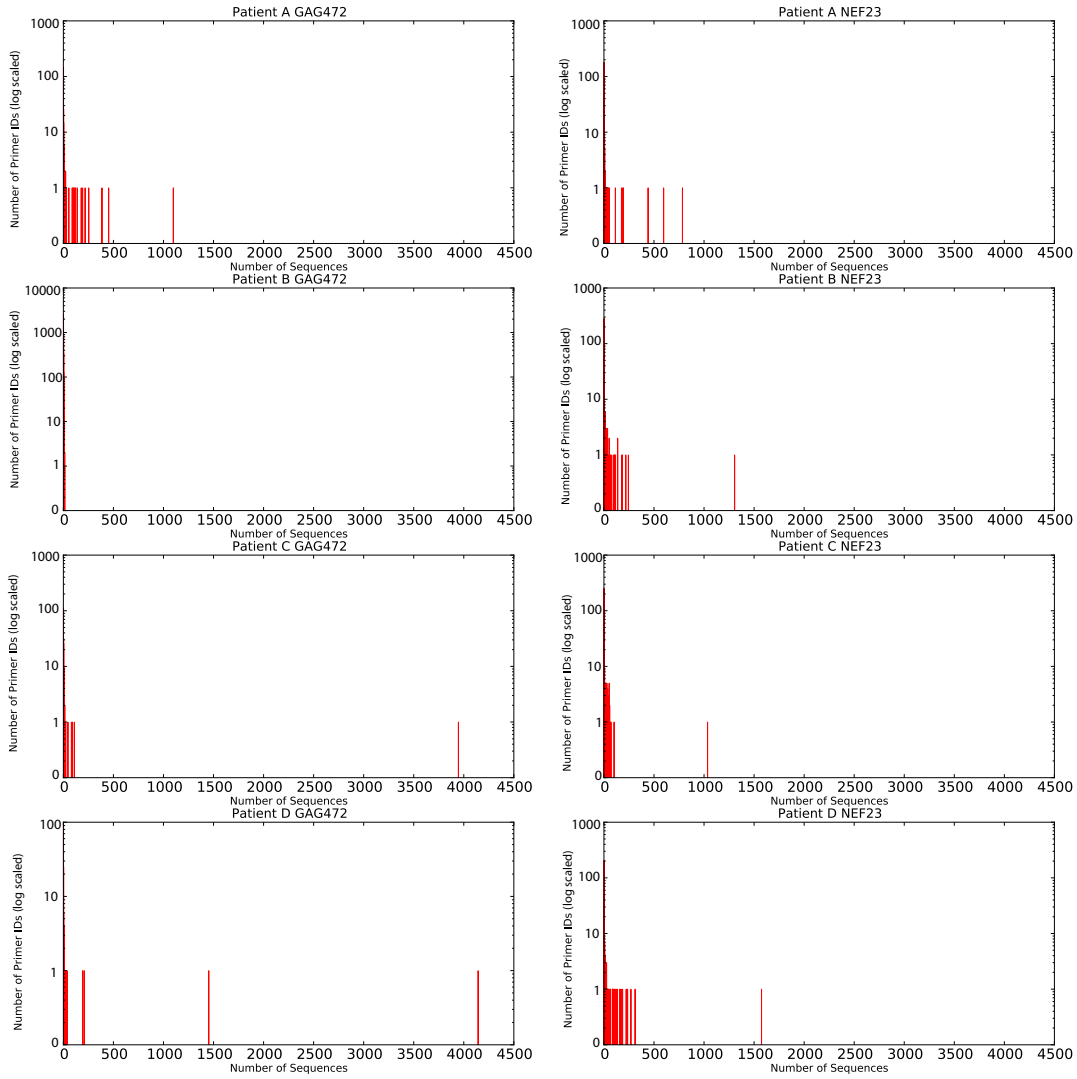


Figure 3. 5 : Count of Primer IDs representing specific number of sequence reads in GAG472 and NEF23 amplicons of Run1 Patients

.

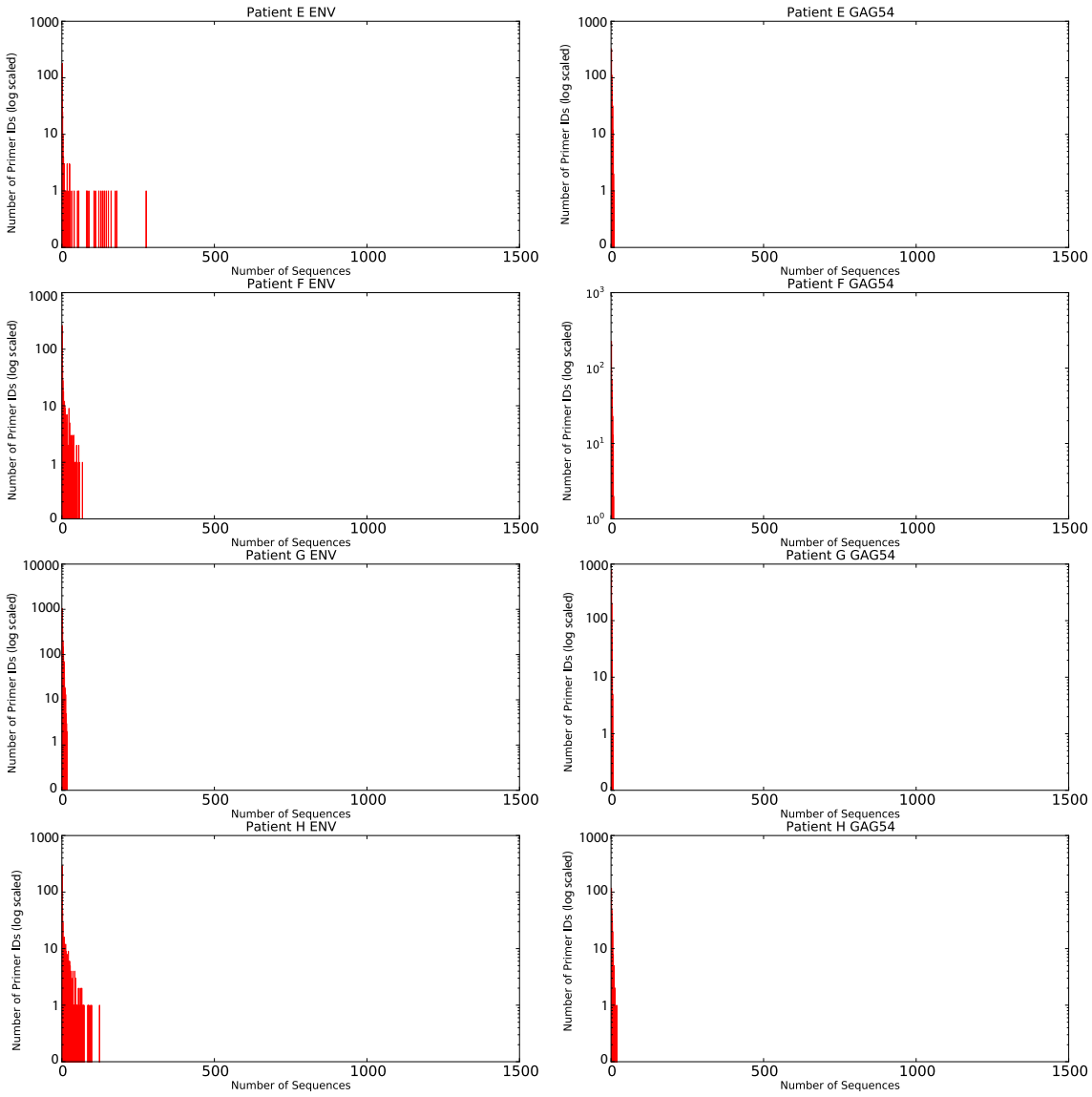


Figure 3. 6 : Count of Primer IDs representing specific number of sequence reads in *ENV* and GAG54 amplicons of Run2 Patients.

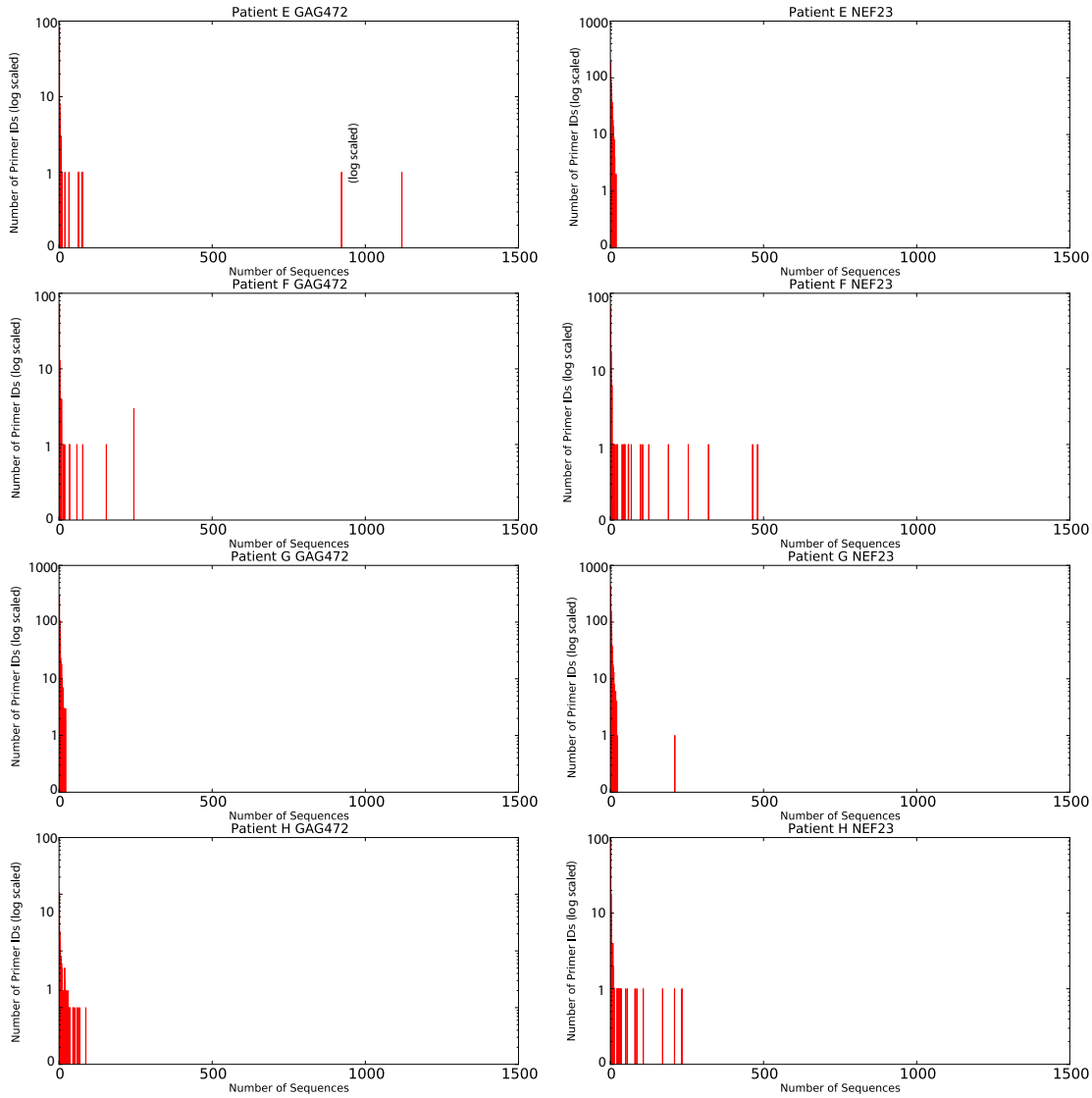


Figure 3. 7: Count of Primer IDs representing specific number of sequence reads in GAG472 and NEF23 amplicons of Run1 Patients.

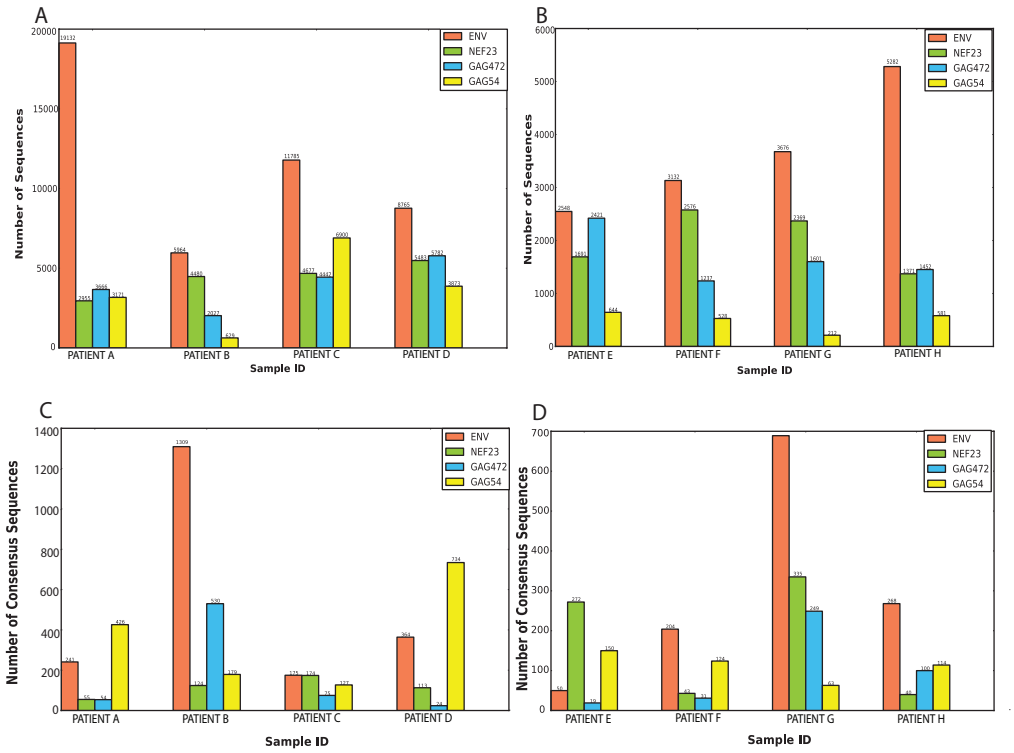


Figure 3. 8: The distribution of final total number of sequences, each represented by Primer ID with greater or equal to threshold number of sequence, per amplicon per sample before consensus sequence generation in (A) Run1 and (B) Run2. The total number of consensus sequences generated per amplicon per sample in (C) Run1 and (D) Run2. Each consensus sequence is generated by collapsing the sequences represented by a unique Primer ID in an amplicon of a sample.

Table 3. 6: Breakdown of the percentage of Primer IDs tags for each unique dataset with two or less and three or more representative sequences.

		ENV		GAG54		GAG472		NEF23	
		<3	≥3	<3	≥3	<3	≥3	<3	≥3
Run 1	Patient A	77.35	22.65	59.47	40.53	76.72	23.28	80.50	19.50
	Patient B	68.78	31.22	94.09	5.91	87.06	12.94	73.62	26.38
	Patient C	76.67	23.33	69.91	30.09	67.11	32.89	63.67	36.33
	Patient D	64.66	35.34	57.08	42.92	83.45	16.55	69.79	30.21
Run 2	Patient E	80.24	19.76	74.58	25.42	82.73	17.27	50.18	49.82
	Patient F	60.39	39.61	70.75	29.25	73.04	26.96	65.87	34.13
	Patient G	67.27	32.73	93.99	6.01	62.27	37.73	63.98	36.02
	Patient H	55.85	44.15	59.14	40.86	56.33	43.67	83.12	16.88

Table 3. 7: Average number of representative sequence per Primer ID tag for each of the unique datasets. The average values were round to the nearest integer value.

		ENV	GAG54	GAG472	NEF23	Average
Run1	Patient A	76.9	4.60	65.82	24.35	43
	Patient B	7.50	3.54	54.01	5.30	18
	Patient C	64.33	3.86	56.46	229.0	89
	Patient D	54.87	35.93	27.04	47.77	42
Run2	Patient E	50.98	4.3	127.47	6.21	48
	Patient F	15.35	4.29	39.93	59.93	30
	Patient G	5.34	3.36	6.43	7.071	6
	Patient H	19.71	5.10	14.54	34.27	19
	Average	37	9	49	52	

Figure 3.8A) but the number of primer IDs that they represent was only 241 (**Figure 3.8C**). On the other hand, Patient B's *env* amplicon had a lower number of reads (**Figure 3.8A**) but had more than six times the number of primer ID tags than patient A had for the same amplicon (**Figure 3.8C**).

3.4 Discussion

Since the introduction of Roche/454 and other ultra deep pyrosequencing technologies, the sequencing of the full spectrum of the diversity of the HIV quasispecies within a HIV infected individual has become possible (Abbate et al., 2011; Beerenwinkel and Zagordi, 2011; Mild et al., 2011; Rozera et al., 2009; Zagordi et al., 2010). However, there are a number of potential biases that must be considered in the experimental protocol and subsequent data analysis.

Current pyrosequencing technologies require a high volume of input DNA and therefore the sample DNA need to be PCR amplified prior to sequencing. The PCR amplification step can introduce various errors into the amplified data including 1) in-vitro recombination of the viral DNA (Judo et al., 1998; Meyerhans et al., 1990; Yang et al., 1996), 2) misincorporation of a base at one or more positions in the growing strand (Hughes and Totten, 2003; Kanagawa, 2003), 3) differential amplification of two different viral templates thereby artificially inflating the prevalence of one viral variant relative to the others (Liu et al., 1996; Polz and Cavanaugh, 1998), and 4) amplification of small amount of DNA obscure the true original sampling and then the independent observation of single genome is lost (Eckert and Kunkel, 1991; Kanagawa, 2003; Liu et al., 1996). These PCR errors introduce allelic skewing,

artificial homogeneity, artificial diversity, and inflate genetic variation (Kanagawa, 2003).

Jabara et al (Jabara et al., 2011) introduced a novel strategy that tags each cDNA-generating primer with a unique primer identifier (primer ID) sequence thereby allowing the identification of sequence reads that originate from the same viral template. As sequences with the same primer ID indicate that they originate from the same template viral RNA sequence, any difference in one sequence relative to all others from the same template can be attributed to PCR or sequencing error and can be removed by the generation of a consensus sequence from all of the sequences representing that primer ID. The technology can be anticipated to be widely accepted in future UDPS involving highly heterogeneous population sample (Sheward et al., 2012). Because the technology developers and publishers have not made the code available for public use, the PIDA algorithm has been developed to facilitate the analysis of HIV drug resistance genotyping data generated using the primer ID approach.

Primer ID is an elegant strategy to reduce errors in the final data for analysis. The primer ID is added to the cDNA primer for reverse transcription process, not to the viral RNA. Therefore primer ID tag cannot address the errors introduced in to the cDNA in the reverse transcription (Boyer et al., 1992). The primer ID that is added to the cDNA is copied through the PCR amplification and ultra deep sequencing protocols. The PCR amplification may introduce errors in every cycle of DNA synthesis. Although primer ID strategy can reduce PCR errors, the errors introduced in the first cycle of PCR amplification cannot be addressed due to lack of enough

amplification of the template sequences (Horton, 1995). On total, the mutations from reverse transcription and first cycle of PCR amplification can range from 0.01 to 0.1% ((Barnes, 1992; Mansky and Temin, 1995), reviewed in (Beerenwinkel et al., 2012)). This would mean that there are unavoidable errors in the experimental analysis that include reverse transcription and PCR in the experimental protocols. The primer ID strategy also fails to address these errors. Despite this, a number of studies have shown that using the primer ID approach significantly reduces artificial PCR and sequencing induced errors generated in a sequence dataset ((Hiatt et al., 2013), reviewed in (Beerenwinkel et al., 2012)). With Jabara and colleagues (Jabara et al., 2011) showing that 80% of the unique sequence polymorphisms were corrected after creating consensus sequences. Further, Kinde et al also showed that errors were reduced by approximately 20 fold using the primer ID technology (Kinde et al., 2011).

We counted the number of primer IDs by pooling the sequences with same primer ID in to a bin. There were 19558 and 8694 primer IDs in Run1 and Run2 respectively. The number of representative sequences for each primer ID ranged from one to 4144 in Run1 and one to 1119 in Run2, while Jabara et al reported the small range of one to 96 representative sequences per primer ID (Jabara et al., 2011). The variation in number of sequences reads in each primer ID bin was over thousand folds (Figure 3.4 – 3.7), while Jabara et al reported only 100 folds (Jabara et al., 2011). The median sequence reads per primer ID in Run1 and Run2 was three, while Jabara et al (Jabara et al., 2011) reported the median of six. The wide range in the number of representative sequence per primer ID is clearly the poor feature of PCR as indicated by Jabara et al (Jabara et al., 2011). Polz and Cavanough also indicated PCR has amplification bias (Polz and Cavanaugh, 1998).

3.4.1 Lost of HIV variants due to PCR artifacts

In our study, primers IDs with one or two representative sequence reads were discarded as the number of sequence reads is below the threshold required for consensus sequence generation. This resulted in a total of 73% and 93% of primer IDs in Run1 and Run2 respectively being discarded. This loss was much higher than 38% that was observed by Jabara et al (Jabara et al., 2011). A primer ID tags a single viral cDNA and the removal of primer IDs with one or two representative sequences would mean likely removal of true HIV variants from the dataset. This would mean that the sequence final dataset probably under represents the true diversity of the viral variants present in the viral quasispecies of that HIV-infected individual.. Therefore, while the primer ID approach undoubtedly reduces artificial diversity by removing many of the PCR and sequencing-induced errors from a dataset, it appears to be too stringent resulting in an underrepresentation of the true viral diversity. Thus the primer ID approach should be further optimized and extended to improve its accuracy.

For example, Schmitt and colleagues have developed an approach to further correct any form of sequencing errors including the errors in reverse transcription and early PCR cycle errors by tagging both strands of duplex DNA (Schmitt et al., 2012). This approach compares one strand with the other strand of the DNA for error correction (Schmitt et al., 2012). The authors showed that the error could be reduced to less than one per billion bases sequenced. However, the approach has not been applied at UDPS of HIV quasispecies with primer ID strategy. While this doesn't fix the problems in relation to underrepresentation of the data as a result of low numbers of sequences for some primer IDs, it is certainly a step in the right direction.

The large numbers of primer IDs with one or two representative sequence reads perhaps indicate the poor PCR protocol. A study by Polz and Cavanough (Polz and Cavanaugh, 1998) also showed that it is the PCR artifact to considerably and reproducibly over amplify specific template DNA. Their study also showed that GC rich priming site amplified with higher efficiency indicating that primer binding energies might play role in over amplification (Polz and Cavanaugh, 1998). These might be the possible reasons for the huge difference in the amplification of the viral templates. Christopherson et al (Christopherson et al., 1997) suggested that the amplification efficiency could be improved up to 10 fold by adjusting the PCR annealing temperature and gradually increasing the temperature during cDNA synthesis. All these suggest that the PCR protocol needs to be improved for better amplification of DNA templates to avoid the lost of viral depth.

3.4.2 Loss of HIV variants due to Primer ID collision

Sheward et al (Sheward et al., 2012) suggested that the generation of truly unique primer IDs is extremely unlikely. The generation of only 10,000 primer IDs of 8 bases length, could generate ~726 duplicated primer IDs called primer ID collision. The result of primer ID collision could tag multiple templates with the same random primer ID. Of the two HIV variants tagged with same primer ID, the variant with low representative sequences gets lost during consensus sequence generation. The authors calculated that 30 of 2000 consensus sequences would be the result of primer ID collision. We also observed over 1000 primer IDs collision in our study (**Table 3.8**). This could lose viral variants depth in the experimental sample.

Table 3. 8: The number of duplicated Primer IDs in Run1 and Run2.

	Sample	Number of Duplicated Primer IDs
Run1	Patient A	77
	Patient B	1103
	Patient C	44
	Patient D	75
Run2	Patient E	32
	Patient F	15
	Patient G	242
	Patient H	42

3.5 Conclusion

Primer ID is a novel technology for correcting PCR and sequencing errors in UDPS data. Primer ID error correction requires at least three representative sequences for a primer ID to generate consensus sequence. Therefore the PCR protocol needs to be optimized for templates amplification. The improved PCR protocol for the high number of template amplification would increase the generation of a consensus sequence per primer ID. The high number of consensus sequences increases the depth of HIV variants in the analysis.

We have developed an algorithm called PIDA for the analysis of raw datasets generated using this technology. We have tested the algorithm in two real primer ID raw sequence read data sets and the comparative results were observed in both the datasets. PIDA algorithm showed that primer ID technology has ability to reduce errors through consensus sequence generation.

Chapter 4

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

4.1 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) programs 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 10^{10} 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., 2011; 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/) 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.

4.2 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.

4.3 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.

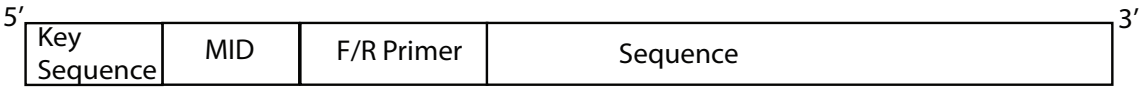


Figure 4. 1: A HTS raw sequence read showing the location of a key sequence, Multiple Identifier (MID), forward or reverses primer and a sequence of targeted genomic region. A key sequence is a sequence of four base in any order (TGCA) that is used by Roche/454 high throughput sequencing platform for calibrating the number of base call while sequencing. The MID sequence identifies the sample from where the sequences derived, the primers (forward or reverse) identifies the genomic amplicon region that was resequenced. The sequence is the actual sequence of the amplicon region.

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.

4.4 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.

4.4.1 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.

4.4.1.1 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**).

4.4.1.2 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**).

A

Amplicon specific Primer file

Name	Forward	Reverse	start	end
GeneA	ATAGAG	GGATGA	x1	y1
GeneB	ATGGAT	ACACTAG	x2	y2

B

Sample ID/MID file

Name	patientname
MID1	PatientA
MID2	PatientB

Figure 4. 2: A) Tab delimited primer file containing five columns – name of amplicon, forward and primers used for resequencing the amplicon and start and end positions of the amplicons set by first nucleotide position of forward primer and last nucleotide position of reverse primer relative to a standard HIV *pol* reference sequence. B) Tab delimited two column MID file containing the name of Roche/454 standard MID in first column and the patient name or the sample name in the second column. In an instance, when no Roche/454 MID is used, the actual nucleotide sequence used as MID can be supplied in the first column.

Note: x1, x2, y1, y2 can be replaced with the actual numbers corresponding to reference sequence

4.4.1.3 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.

4.4.1.4 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:

4.4.1.5 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

4.4.1.6 Demultiplex parameters

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

A) 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.

B) 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 default setting)

C) 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.

4.4.2 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.

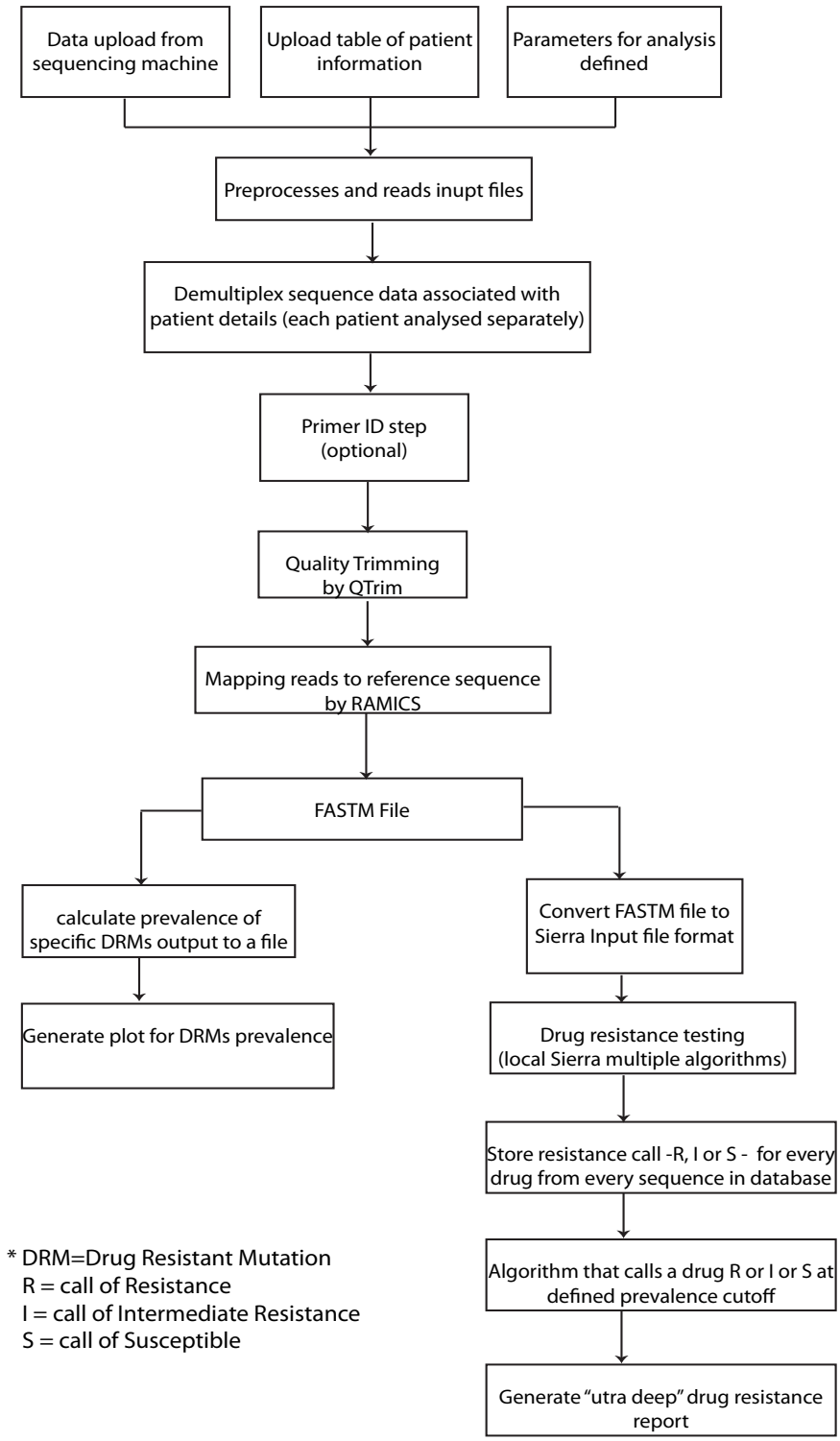


Figure 4. 3: Seq2Res pipeline workflow

2. 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.

3. Demultiplexing

Seq2Res utilizes Biopython (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

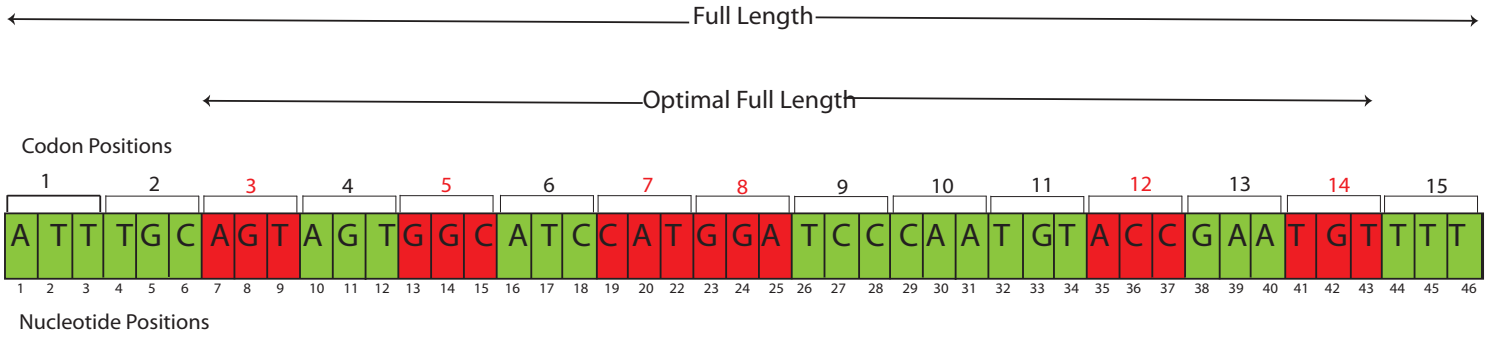


Figure 4. 4: An example of an amplicon showing optimal full length and full length. The amplicon is 46 nucleotides in length starting from 1 to 46 and has 15 codon positions (each three nucleotide compose one codon). The green codons are non drug resistant codon positions and the red codon positions are drug resistant codon positions. A Seq2Res user has to input start and end position in nucleotide numbering as 1 and 46 respectively for the amplicon. Seq2Res processes that the start and end nucleotide positions fall on codon position 1 and codon position 15 respectively, which is the full length. For the known drug resistant codon positions (in this example: 3, 5, 7, 8, 12, 14), Seq2Res processes the start and end codon positions to get the first and last drug resistant codon positions that is 3 and 14 as the new start and end positions. Positions 3 to 14 define the optimal full length.

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.

4. 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., 2005a), 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.

5. 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.

6. 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.

7. 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*

```
>SIMULATED-1
66L, 67V, 68T, 69I, 70K, 71L, 72G, 73G, 74Q, 75I, 76R, 77E, 78A, 79L, 80L, 81D, 82T, 83G, 84A
>SIMULATED-2
66L, 67V, 68T, 69I, 70K, 71L, 72G, 73G, 74Q, 75I, 76R, 77E, 78A, 79L, 80L, 81D, 82T, 83G, 84A
>SIMULATED-3
66L, 67V, 68T, 69I, 70K, 71L, 72G, 73G, 74Q, 75I, 76R, 77E, 78A, 79L, 80L, 81D, 82T, 83G, 84A
>SIMULATED-4
66L, 67V, 68T, 69I, 70K, 71L, 72G, 73G, 74Q, 75I, 76R, 77E, 78A, 79L, 80L, 81D, 82T, 83G, 84A
>SIMULATED-5
66L, 67V, 68T, 69I, 70K, 71L, 72G, 73G, 74Q, 75I, 76R, 77E, 78A, 79L, 80L, 81D, 82T, 83G, 84A
```

Figure 4. 5: A FASTA format like FASTM file. The FASTM format begins with sequence ID followed by list of codon positions that it covers with respected to the reference sequence and the single letter denotation of an amino acid at the codon positions.

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**).

8. 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.

9. 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

&name=SIMULATED-1&**PR**=10L,11RV,13I,20R,23L,24L&**RT**=40E,41M,44E,62A,65R,66RK,67D,68G,**69iT**,70R,71W

&name=SIMULATED-2&**PR**=10L,11MV,13I,20R,23L,24L&**RT**=40E,41M,44E,62A,65R,66RK,67D,68G,**69i**,70R,71W

&name=SIMULATED-3&**PR**=10L,11V,13I,20R,23L,24L&**RT**=40E,41M,44E,62A,65R,66K,67D,**68d**,69dT,70R,71W

&name=SIMULATED-4&**PR**=10L,11V,13I,20R,23L,24L

&name=SIMULATED-5&**RT**=40E,41M,44E,62A,65R,66K,67D,68G,69T,70R,71W

Figure 4. 6: Seq2Res generated file after translation codon positions of sequence reads from FASTM file. Each line in the file contains all information of a sequence like the complete sequence ID as the name, codon positions and their amino acids including insertion and deletions from PR and RT amplicons.

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

Table 4. 1: Generic names of antiretroviral drugs, their drug class and abbreviations.

Drug Class	**Generic Drug Name**	**Abbreviation**
Nucleoside Reverse Integrase Inhibitors (NRTIs)	1. Abacavir	ABC
	2. didanosine	DDI
	3. Emtricitabine	FTC
	4. Lamivudine	3TC
	5. Stavudine	d4T
	6. Tenofovir	TDF
	7. Zidovudine	AZT
Non-Nucleoside Reverse Integrase Inhibitors (NNRTIs)	8. Efavirenz	EFV
	9. Etravirine	ETR
	10. Nevirapine	NVP
	11. Rilpivirine	RPV
Protease Inhibitors (PIs)	12. Atazanavir/r	ATV/r
	13. Darunavir/r	DRV/r
	14. Fosamprenavir/r	FPV/r
	15. Indinavir/r	IDV/r
	16. Lopinavir/r	LPV/r
	17. Nelfinavir	NFV
	18. Saquinavir/r	SQV/r
	19. Tipranavir/r	TPV/r
Integrase Inhibitors (IN)	20. Raltegravir	RAL
	21. Elvitegravir	EVG
	22. Dolutegravir	DTG

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**):
    a. 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.
    b. 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.
    c. 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

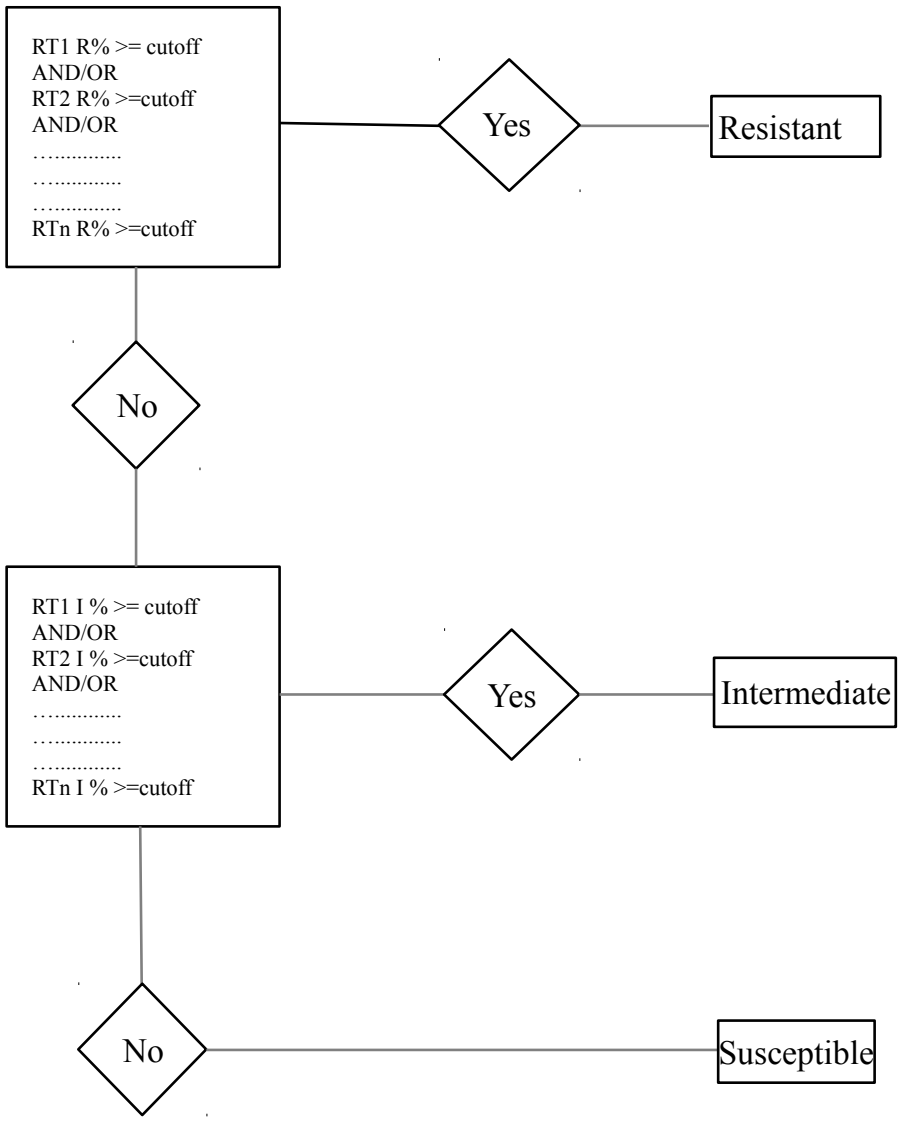


Figure 4. 7: Conditions applied in Seq2Res for drug susceptibility calls for a viral population when multiple amplicons are sequenced from a gene usually Reverse Transcriptase (RT). RT1, RT2, …. RTn are the total 'n' amplicons from the gene RT. R% indicates the percentage of sequences predicted as resistant for a drug. I% indicates the percentage of sequence predicted as Intermediate Resistant for the same drug. The viral population is called as susceptible, if it is predicted neither resistant nor intermediate resistant.

Class	Drug	# Sequences	% Resistant	% Intermediate	Drug Class Mutations
NRTI	ABC	7475	51.52	0.95	M41L,K65R,V75I,A62V
NRTI	DDI	7475	51.52	0.83	M41L,K65R,V75I,A62V
NRTI	FTC	7475	0.0	52.35	M41L,K65R,V75I,A62V
NRTI	3TC	7475	0.0	52.35	M41L,K65R,V75I,A62V
NRTI	D4T	7475	51.12	1.34	M41L,K65R,V75I,A62V
NRTI	TDF	7475	51.52	0.95	M41L,K65R,V75I,A62V
NRTI	AZT	7475	0.0	1.27	M41L,K65R,V75I,A62V
NNRTI	EFV	7475	49.75	0.05	K103N
NNRTI	ETR	7475	0.0	0.0	K103N
NNRTI	NVP	7475	49.81	0.05	K103N
NNRTI	RPV	7475	0.0	0.0	K103N
PI	ATV/r	LC	-	-	
PI	DRV/r	LC	-	-	
PI	FPV/r	LC	-	-	
PI	IDV/r	LC	-	-	
PI	LPV/r	LC	-	-	
PI	NFV	LC	-	-	
PI	SQV/r	LC	-	-	
PI	TPV/r	LC	-	-	
INI	RAL	LC	-	-	
INI	ELV	LC	-	-	
INI	DTG	LC	-	-	

Figure 4. 8: An example of a drug resistance report. The columns from left to right in the report shows the drug class, the drug in the drug class, the number of sequence reads showing resistance to the associated drug, the percentage of sequence reads showing high level resistance to the associated drug, the percentage of sequence reads showing intermediate resistance to the associated drug and the drug resistant mutations in the observed sequence reads shows the resistance to the associated drug.

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).

4.4.3 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**).

4.4.4 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**) is 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**).

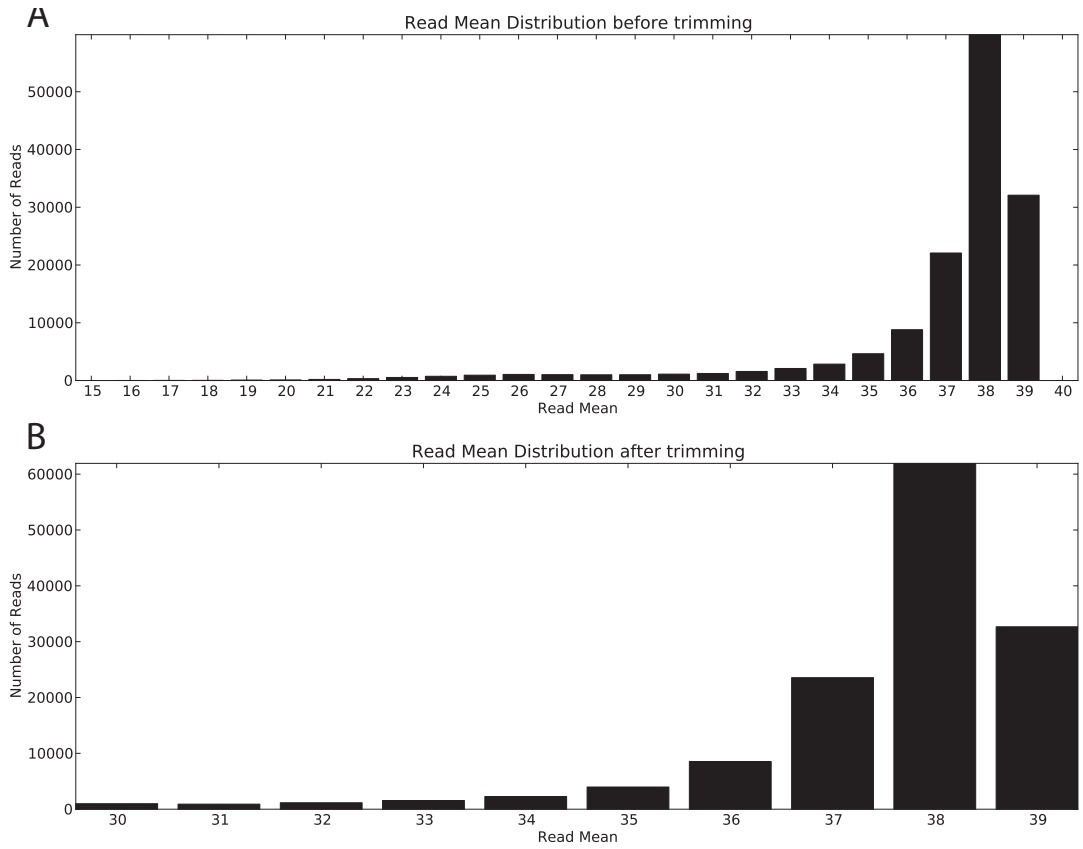


Figure 4. 9: QTrim analysis on number of sequence reads and read mean in both untrimmed and trimmed data in Seq2Res.

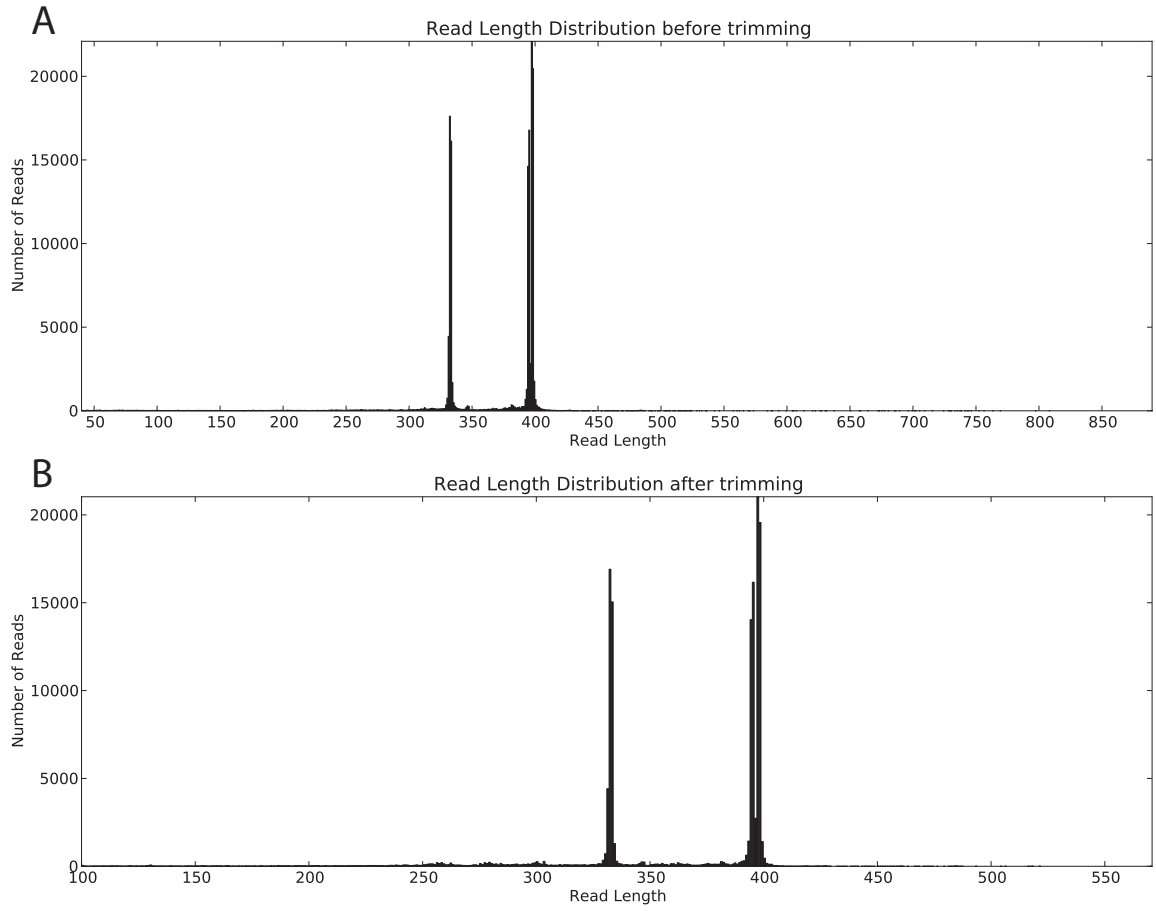


Figure 4. 10: QTrim analysis showing number of sequence reads with same read length in both untrimmed and trimmed data in Seq2Res.

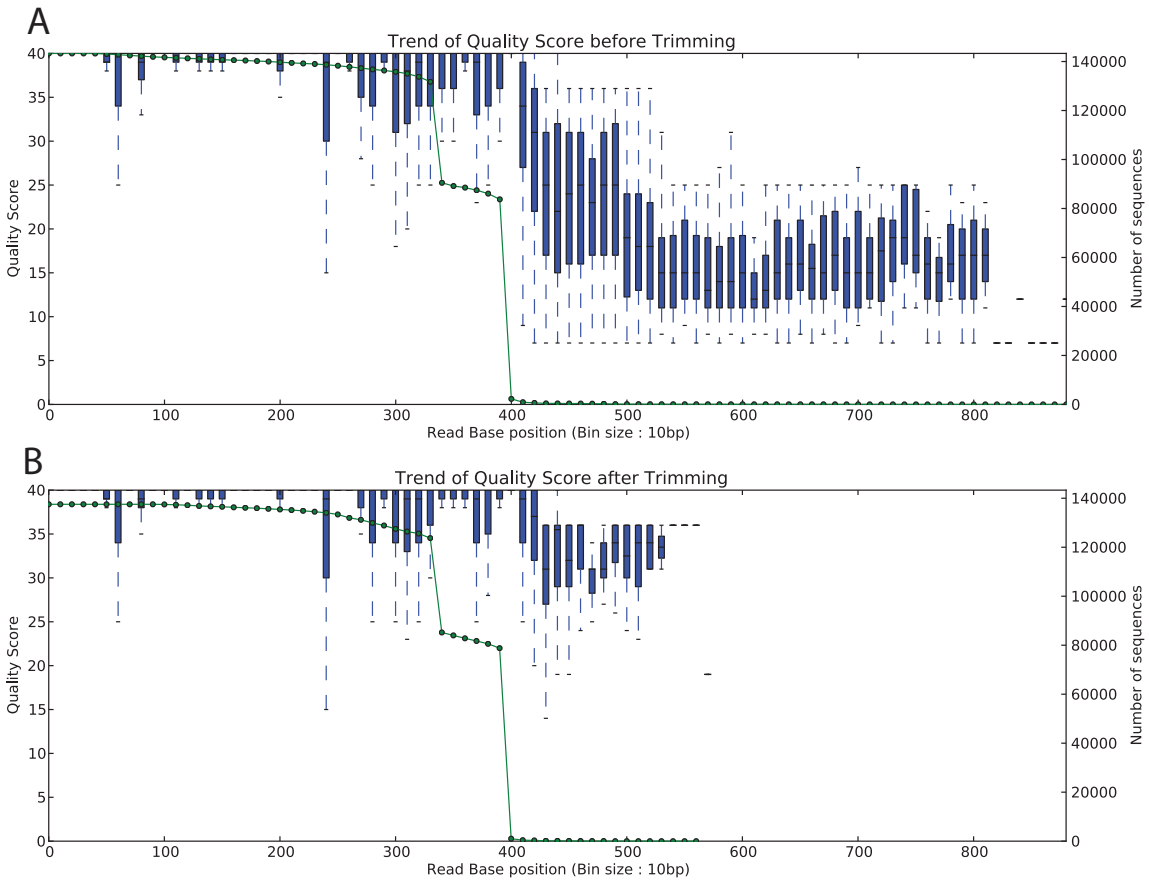


Figure 4. 11: QTrim analysis showing the trend of average quality score with increase in the read length of sequence reads at every 10 base pairs. The green dotted line shows the number of sequences that contribute in the average quality score.

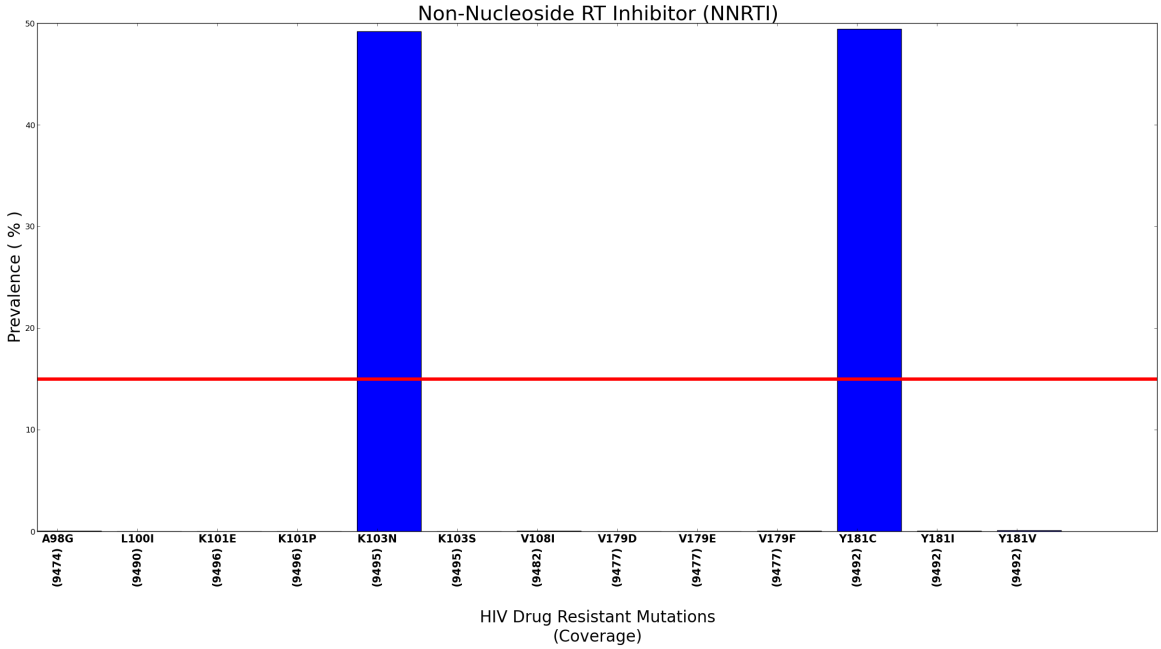


Figure 4. 12: An example of DRM prevalence plot in Seq2Res. The plot shows the prevalence of K103N and Y181C DRMs with prevalence of 49.50% and 49.85% respectively. The red line represents the user defined prevalence cutoff.

Table 4. 2: The standard list of HIV drug resistant mutation codon positions with respect to the genes

HIV gene	Standard drug resistant mutation codon positions in the gene
PR	10,11,13,20,23,24,30,32,33,34,35,36,43,46,47,48,50,53,54,55,58,60,62, 63,64,71,73,74,76,77,82,83,84,85,88,89,90,93
RT	40,41,44,62,65,66,67,68,69,70,71,74,75,77,90,98,100,101,103,106,108, 115,116,118,138,151,179,181,184,188,190,210,215,219,221,225,227, 230,236,238,318,333,348
IN	51,66,68,72,74,92,95,97,114,121,125,128,138,140,143,145,146,147,148, 151,153,154,155,157,163,201,203,206,230,263

4.4.5 Evaluating the sensitivity of Seq2Res.

4.4.5.1 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.

4.4.6 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.

4.4.7 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.

4.4.8 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.

Table 4. 3: Number of sequences mixed to generate simulated sequence data with different percentage of resistant sequences.

Percentage of resistant sequences	Number of resistant sequences	Number of sensitive sequences
0.1	10	9990
1	100	9900
5	500	9500
10	1000	9000
15	1500	8500
20	2000	8000
50	5000	5000

4.4.9 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.

4.5 Results

4.5.1 Seq2Res running time

The running time of Seq2Res depends on several factors, including the Internet bandwidth 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.

4.5.2 Comparison of Mutation and Resistance Level Calls using the Sierra web service and Seq2Res

For every sequence in dataset A and dataset B (containing 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.

4.5.3 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**).

4.5.4 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

Table 4. 4: Drug resistant mutations grouped by drug class to which the mutation is highly resistant. The drug resistant mutations are present in five resistant sequences (indicated by '_R' in sequence names) and none are present in five susceptible sequences (indicated by '_S' in sequence names) as reported from web Sierra.

Sequences names	NRTIs	NNRTIs	PIs
56252_R	K65R, K70R, V75I, F77L, Y115F, F116Y, Q151M	K103N	A71V, L90M
56252_S	None	None	None
21354_R	K65R, D67N, K70G	K103N	L10I, G48V, I50V, I54V, A71V, V82A
21354_S	None	None	None
63377_R	K65R, D67N, K70G	K103N	L10I, G48V, I50V, I54V, A71V, V82T
63377_S	None	None	None
4134_R	K65R, D67N, K70G	K103N, Y181C	L10I, G48V, I54V, V82A
4134_S	None	None	None
2368_R	M41L, A62V, K65R, V75I	K103N, Y181C	A71V, G73S, L90M
2368_S	None	None	None

Table 4. 5: Resistance Calls of five resistant sequences (indicated by '_R' in sequence names) and five susceptible sequences (indicated by '_S' in sequence names) for simulated data using web Sierra

Drugs	56252_R	56252_S	21354_R	21354_S	63377_R	63377_S	4134_R	4134_S	2368_R	2368_S
3TC	R	S	R	S	R	S	R	S	R	S
ABC	S	S	S	S	S	S	I	S	I	S
ATV/r	R	S	R	S	R	S	R	S	R	S
AZT	S	S	S	S	S	S	I	S	I	S
D4T	R	S	R	S	R	S	R	S	R	S
DDI	R	S	R	S	R	S	R	S	R	S
DRV/r	R	S	I	S	I	S	I	S	I	S
EFV	R	S	I	S	I	S	I	S	I	S
ETR	R	S	R	S	R	S	R	S	R	S
FPV/r	R	S	R	S	R	S	R	S	R	S
FTC	R	S	S	S	S	S	S	S	S	S
IDV/r	I	S	R	S	R	S	R	S	I	S
LPV/r	S	S	I	S	I	S	S	S	S	S
NFV	I	S	R	S	R	S	I	S	I	S
NVP	I	S	R	S	R	S	R	S	I	S
RPV	S	S	R	S	R	S	I	S	I	S
SQV/r	R	S	R	S	R	S	R	S	R	S
TDF	I	S	R	S	R	S	R	S	R	S
TPV/r	S	S	S	S	I	S	I	S	S	S
R: Resistant, I: Intermediate, S: Susceptible										

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 10^{th} 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.

4.5.5 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.

4.5.6 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.

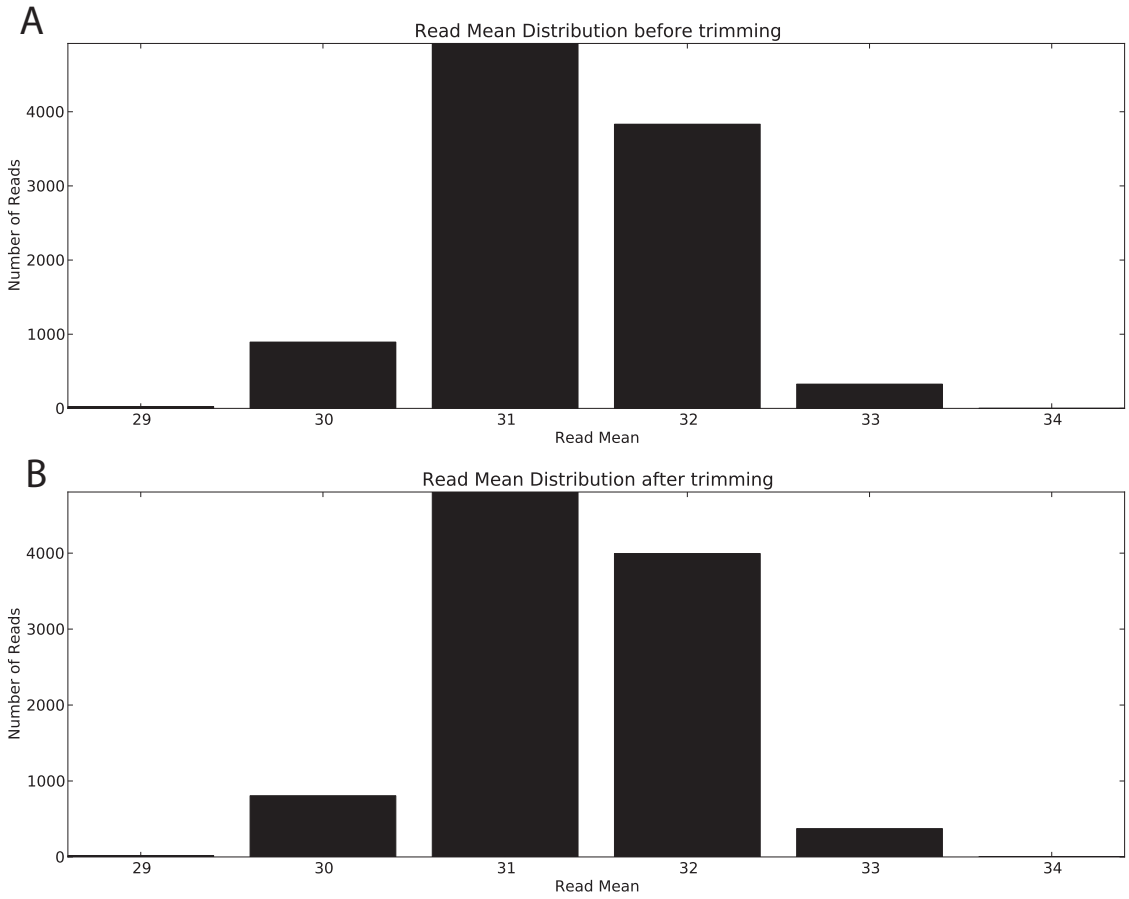


Figure 4. 13: Distribution of sequence reads by mean quality score in (A) untrimmed and (B) trimmed simulated data

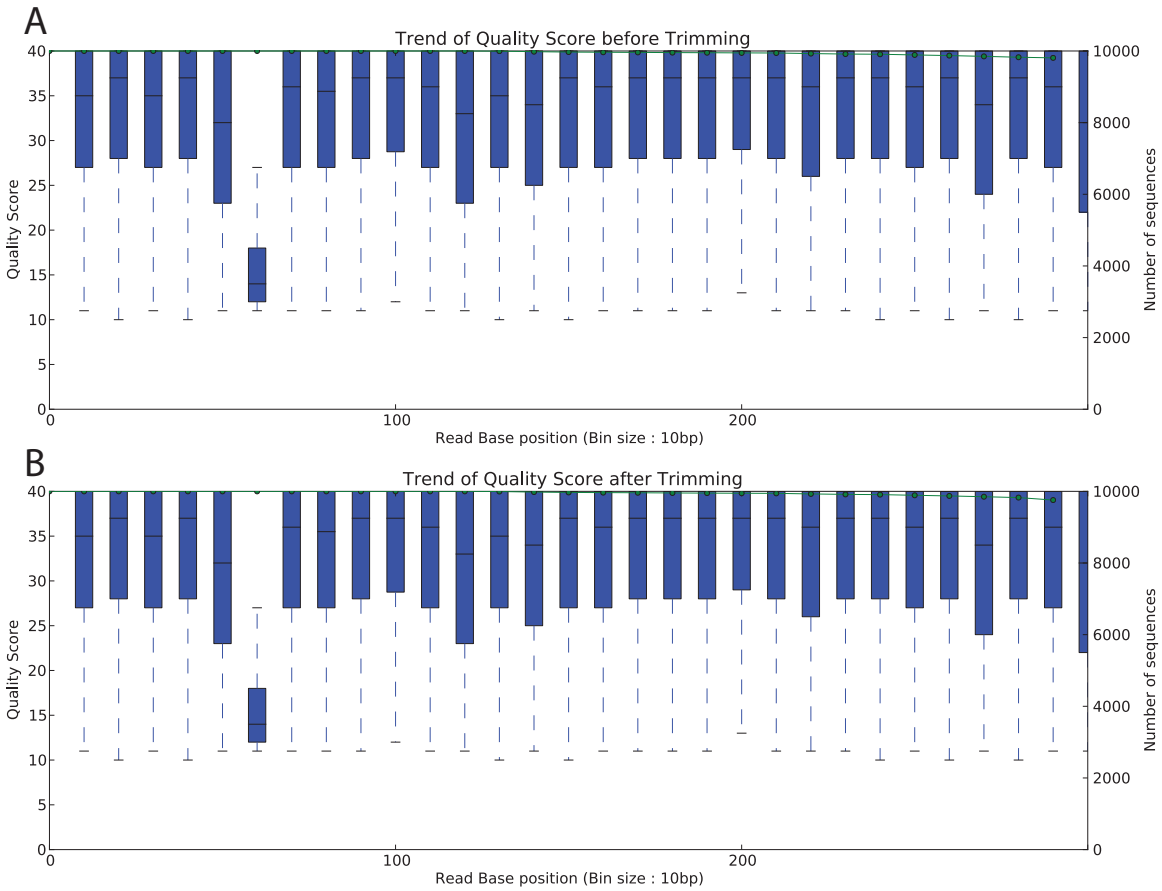


Figure 4. 14: Trend of average quality score at every 10th base pair across sequence reads in (A) an untrimmed and (B) a trimmed simulated data. The median quality score was increasing and decreasing across the sequence reads in both A and B.

Table 4. 6: The start and end nucleotide positions of full-length amplicons, codon positions of full-length amplicons and codon positions of optimal full-length amplicons in the simulated datasets

Simulated Amplicon name	Full length amplicon Nucleotide position		Full length amplicon Codon positions		Optimal full length amplicon codon positions	
	Start	End	Start	End	Start	End
PR	169	469	57	157	66	149
RT1	466	812	156	271	195	270
RT2	672	1017	224	339	224	336

The analyses of the simulated data using Seq2Res showed that the observed prevalence of the DRMs were, in all cases, almost identical to that of the expected prevalence (**Figure 4.15 and Table 4.7**).

4.5.7 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**).

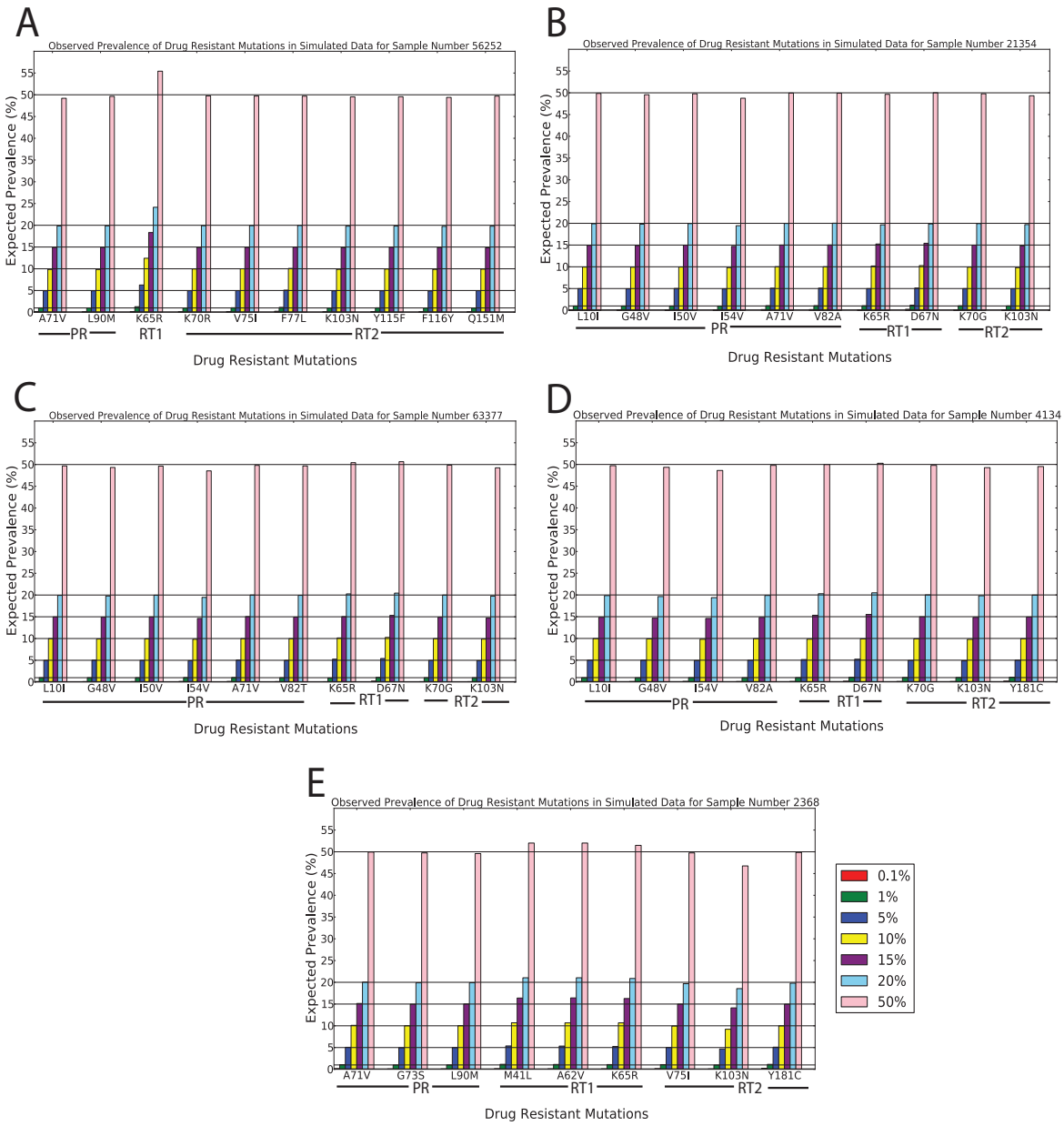


Figure 4. 15: Observed prevalence of the drug resistance mutations (DRMs) in the simulated data of samples. The horizontal lines show the expected prevalence of the DRMs while the colored bars show the observed prevalence.

Table 4. 7: The mean observed prevalence of all DRMs from all three simulated amplicons datasets - PR, RT1 and RT2 in all samples. The mean across samples calculated from the mean observed prevalence is shown at the last column.

Expected Prevalence (%)	56252	21354	63377	4134	2368	**Mean across samples**
0.1	0.11	0.13	0.13	0.13	0.14	**0.128**
1	0.99	1.04	1.05	1.00	1.07	**1.03**
5	5.03	5.00	5.10	5.00	5.09	**5.044**
10	9.94	9.95	10.18	9.86	10.16	**10.018**
15	14.86	14.94	15.24	14.88	15.37	**15.058**
20	19.86	19.75	20.28	19.83	20.10	**19.964**
50	49.50	49.46	50.16	49.36	50.13	**49.722**

Table 4. 8: Resistance calls to PI, NRTI and NNRTI drugs in PR, RT1 and RT2 amplicons simulated with the expected DRM prevalence of 0.1%, 1%, 5% and 10% in all samples. The green color denoted that drugs are sensitive to the samples.

Drug class	Sample/Drug	56252	21354	63377	4134	2368
PI	ATV/r	S	S	S	S	S
	DRV/r	S	S	S	S	S
	FPV/r	S	S	S	S	S
	IDV/r	S	S	S	S	S
	LPV/r	S	S	S	S	S
	NFV	S	S	S	S	S
	SQV/r	S	S	S	S	S
	TPV/r	S	S	S	S	S
NRTI	ABC	S	S	S	S	S
	DDI	S	S	S	S	S
	FTC	S	S	S	S	S
	3TC	S	S	S	S	S
	D4T	S	S	S	S	S
	TDF	S	S	S	S	S
	AZT	S	S	S	S	S
NNRTI	EFV	S	S	S	S	S
	ETR	S	S	S	S	S
	NVP	S	S	S	S	S
	RPV	S	S	S	S	S

Table 4. 9: Resistance calls to PI, NRTI and NNRTI drugs in RT1 amplicon simulated datasets with the expected DRM prevalence of 15% from all samples.

R – Resistant (colored Red), I – Intermediate Resistant (colored Orange), S - Susceptible (colored Green)

Drug class	Sample/Drug	56252	21354	63377	4134	2368
PI	ATV/r	S	I	S	S	I
	DRV/r	S	S	S	S	S
	FPV/r	S	S	S	S	I
	IDV/r	S	I	S	S	I
	LPV/r	S	I	S	S	S
	NFV	S	I	S	S	R
	SQV/r	S	S	S	S	I
	TPV/r	S	S	S	S	S
NRTI	ABC	R	R	R	R	R
	DDI	R	R	R	R	R
	FTC	I	I	I	I	I
	3TC	I	I	I	I	I
	D4T	R	R	R	R	R
	TDF	R	R	R	R	R
	AZT	I	I	S	I	I
NNRTI	EFV	R	R	R	R	R
	ETR	S	S	S	S	S
	NVP	R	R	R	R	R
	RPV	S	S	S	S	S

Table 4. 10: Resistance calls to PI, NRTI and NNRTI drugs in PR, RT1 and RT2 amplicons simulated datasets with the expected DRM prevalence of 20% and 50% from all samples.

R – Resistant (colored Red), I – Intermediate Resistant (colored Orange), S - Susceptible (colored Green)

Drug class	Sample/Drug	56252	21354	63377	4134	2368
PI	ATV/r	I	R	R	R	I
	DRV/r	S	I	I	S	S
	FPV/r	I	R	R	I	I
	IDV/r	I	R	R	R	I
	LPV/r	S	R	R	I	I
	NFV	R	R	R	R	R
	SQV/r	I	R	R	R	R
	TPV/r	S	I	I	I	S
NRTI	ABC	R	R	R	R	R
	DDI	R	R	R	R	R
	FTC	I	I	I	I	I
	3TC	I	I	I	I	I
	D4T	R	R	R	R	R
	TDF	R	R	R	R	R
	AZT	I	I	I	I	I
NNRTI	EFV	R	R	R	R	R
	ETR	S	S	S	I	I
	NVP	R	R	R	R	R
	RPV	S	S	S	I	I

4.5.8 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

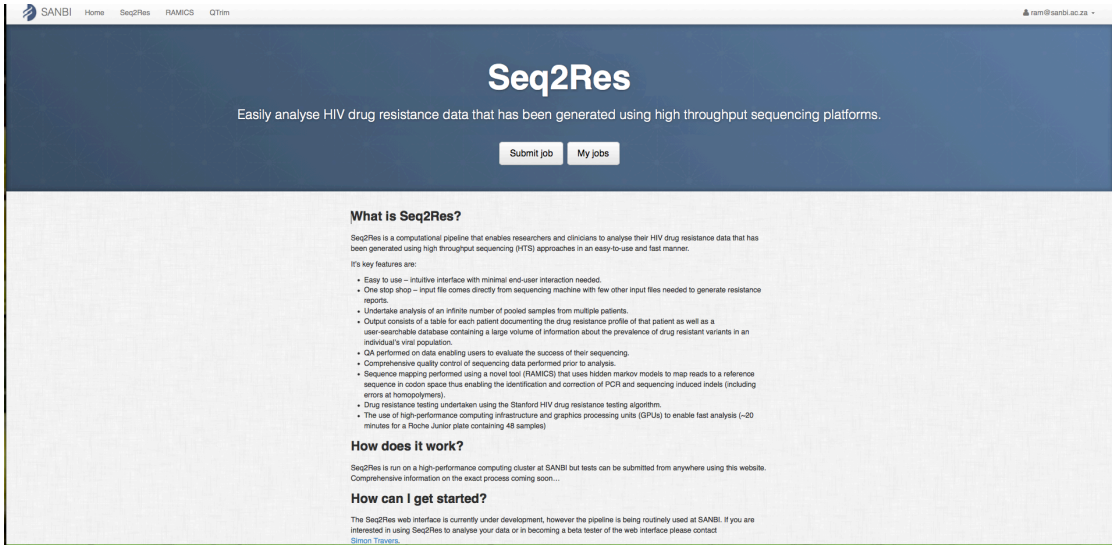


Figure 4. 16: Seq2Res homepage. Users are required to register before Seq2Res access. New users have to contact Prof Simon Travers at simon@sanbi.ac.za for registration. Academic users are registered for free whereas business users have to purchase license. Once registered, users get a login ID and password, which can be used to log in to get Seq2Res access. With "Submit Job" button, users can submit new job and with "My Jobs" button, users can view previously submitted jobs that are both completed or in process.

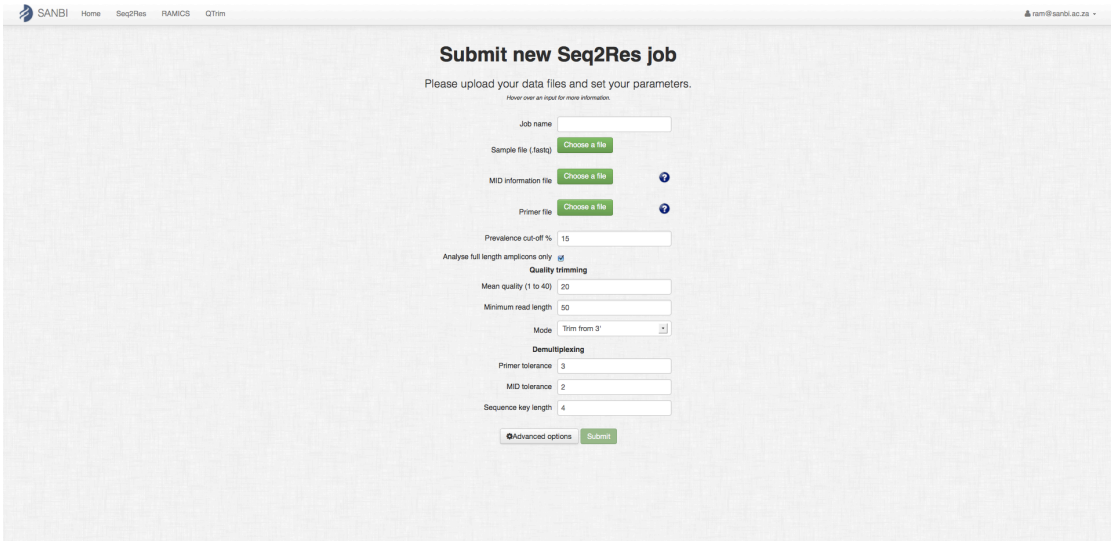


Figure 4. 17: Seq2Res job submission homepage. Users can input a raw sequence file, primer file containing primers for the raw sequence file and sample specific MID file by exploring the files in their computer file system. Users can click on help button (with symbol ?) for any confusion in the file format of the primer and MID files. In advance options, users can change the analysis parameters. Users can hover the mouse cursor point on the fill up area to get help on the parameters. The "submit" button will light up after entering all the required fields. Users can then click the "submit" button to submit their job. Users will automatically get an email if Seq2Res outputs error while processing or the job is processed successfully.

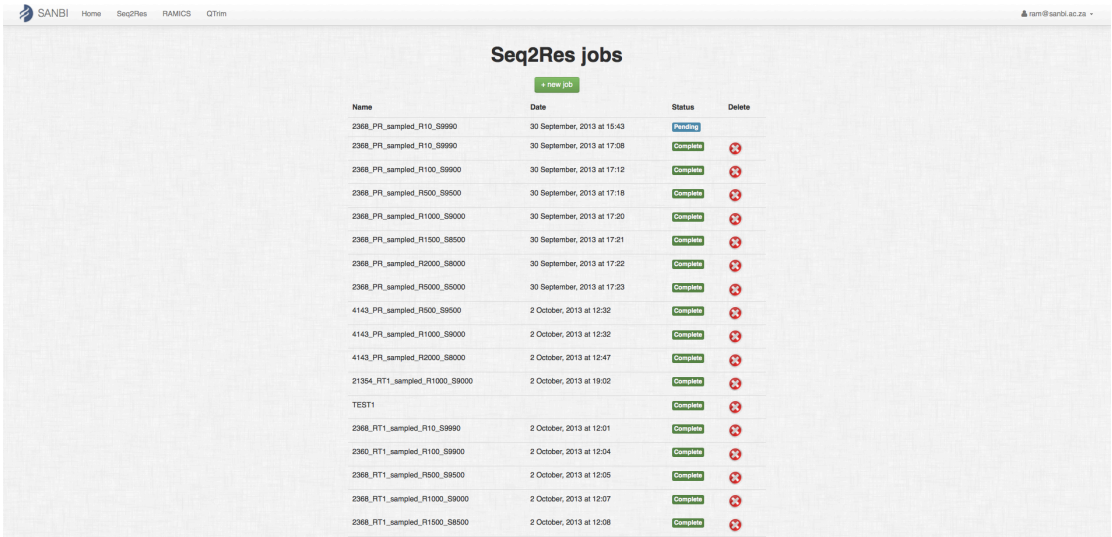


Figure 4. 18: Seq2Res page for viewing user submitted list of jobs. The page shows job specific details like name of the job, the date when the job was submitted, the status of the job either completed process or in pending and an option to delete the job. Users can sort the jobs by name and by date, clicking at “name” and “date” respectively

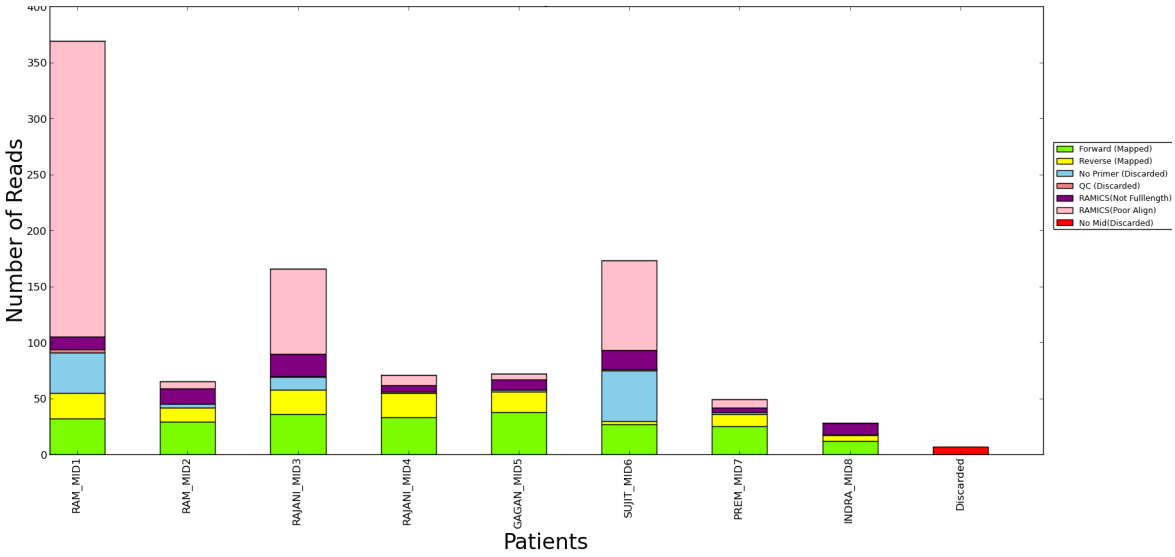


Figure 4. 19: Number of sequences mapped that went to final result (grass green and bright yellow color) and the number of sequences that are discarded (other colors) per patient at different processing steps of Seq2Res.

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**).

4.6 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

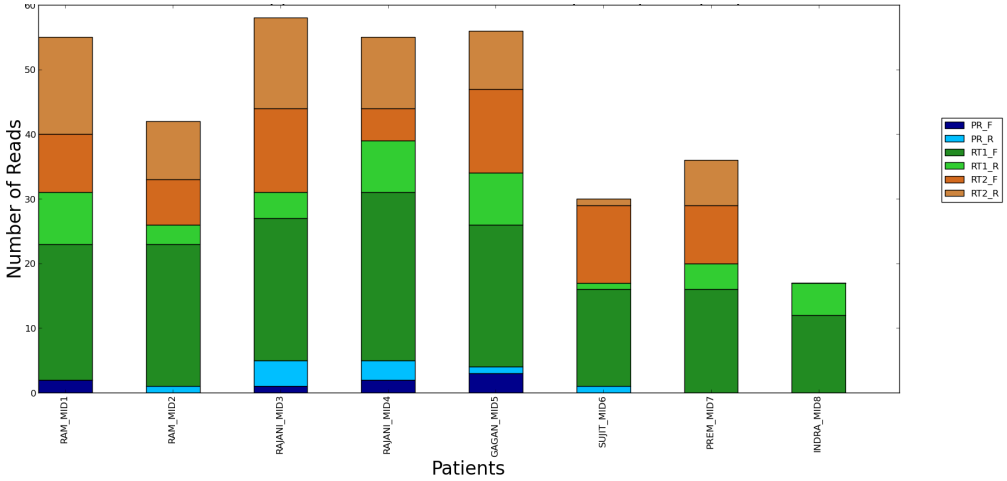


Figure 4. 20: Number of sequences in forward (with "_F" in legend) and reverse (with "_R" in legend) strands per amplicon per sample that were mapped to reference sequences and went to final result.

Figure 4. 21: Patient specific drug resistant result page showing the drug resistant report of the sample at the top in the page.

Color code: Red – highly resistant, Orange – intermediate resistant, Green - Susceptible

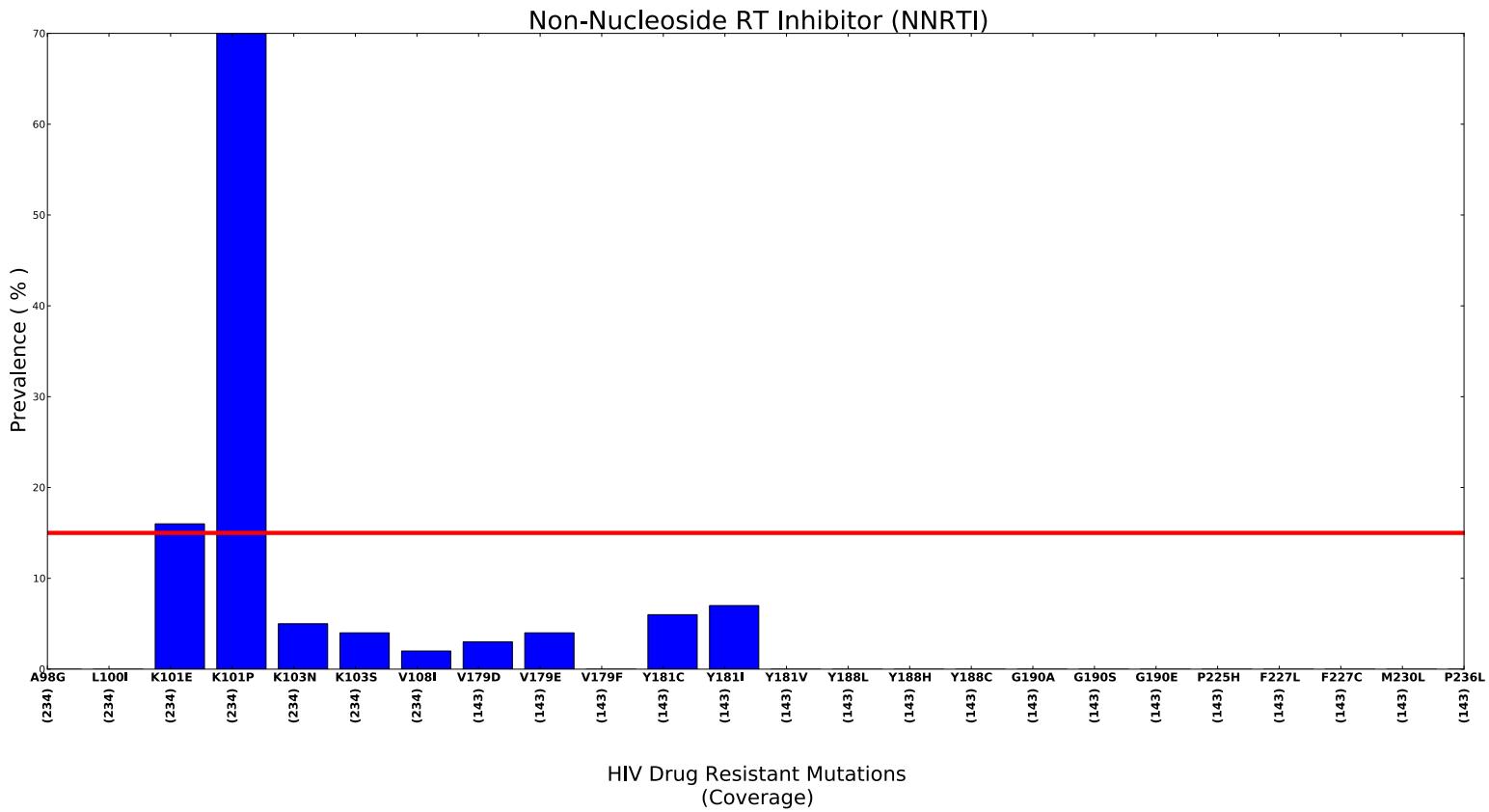


Figure 4. 22: A prevalence plot showing the observed prevalence of mutations that are resistant to NNRTI drugs. A number (in vertical orientation) shows the coverage of the drug resistant mutations. A horizontal red line shows the cutoff prevalence.

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.

4.6.1 Optimal full-length

HIV drug resistance is strongly associated with the pattern of DRMs (Bennett et al., 2009; Johnson et al., 2009; Zhang et al., 2010a). 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). 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.

4.6.2 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

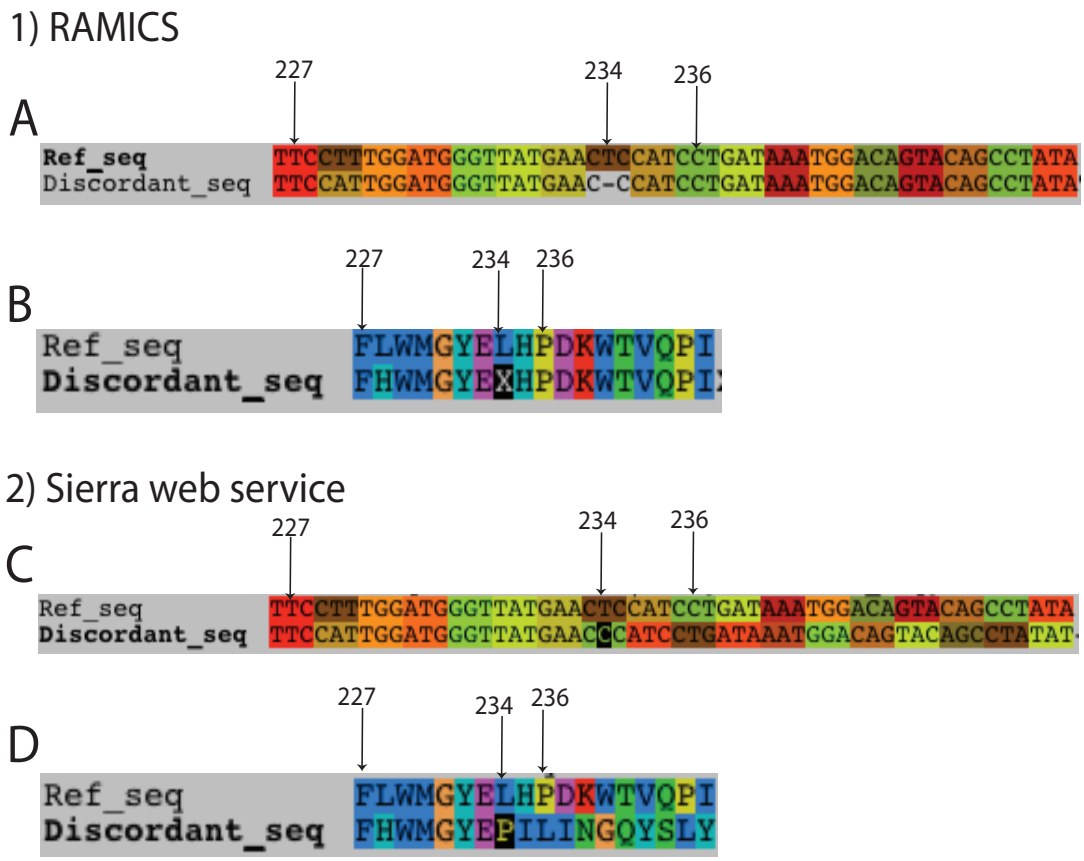


Figure 4. 23: Difference in nucleotide sequence alignment (A and C) and the corresponding amino acid sequence alignment (B and D) as obtained from RAMICS (A and B) and Sierra web service (C and D). The numbers indicate the codon position corresponding to the HIV *pol* reference sequence.

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.

4.6.3 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., 2005a; 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

Table 4. 11: The selected drug resistant mutations in the simulated datasets within or adjacent to the homopolymer region in HXB2 reference sequence, the wild type sequence and mutated sequence of the drug resistant mutations.

Drug Class	Drug Resistant Mutation (DRM)	Wild type codon sequence in Homopolymer Region of DRM (bold and underlined)	Mutated codon sequence in Homopolymer Region of DRM (bold and italic)
PI	G48V	**GGG**GG	***GUG***GG
PI	I54V	TTTT**AUC**AAA	TTTT***GUC***AAA
NRTI	K65R	AAAG**AAA**AAAG	AAAG***AGA***AAAG
NNRTI	K103N	AAAAAAG**AAA**AAA	AAAAAAG***AGA***AAA

Table 4. 12: The percentage of simulated sequence reads with insertions and deletions in the simulated datasets from Sample 2368.

Drug Resistant Mutation	Number of sequences analyzed	Derived from Sensitive (S) or Resistant (R) "seed" sequence	Number of Insertions (+)/deletions (-)	% of sequences with indels
K103N	9990	S	+4	0.1
	9990	S	+3	0.6
	9990	S	+2	2.47
	9990	S	+1	24.24
	9990	S	0	70.07
	9990	S	-1	3.06
	9990	S	-2	0.08
	5000	**R**	**+2**	**0.12**
	5000	**R**	**+1**	**6.64**
	5000	**R**	**0**	**90.94**
	5000	**R**	**-1**	**2.3**
K65R	9990	S	+2	0.35
	9990	S	+1	14.02
	9990	S	0	83.01
	9990	S	-1	2.09
	9990	S	-2	0.01
	9990	S	-3	0.03
	5000	**R**	**+1**	**2.34**
	5000	**R**	**0**	**95.54**
	5000	**R**	**-1**	**1.68**
	5000	**R**	**-2**	**0.06**
	5000	**R**	**-3**	**0.02**
G48V	9990	S	+2	0.02
	9990	S	+1	5.21
	9990	S	0	92.5
	9990	S	-1	1.95
	9990	S	-3	0.03
	5000	**R**	**+2**	**0.06**
	5000	**R**	**+1**	**5.28**
	5000	**R**	**0**	**92.58**
	5000	**R**	**-1**	**1.74**
	5000	**R**	**-2**	**0.02**
I54V	9990	S	+1	2.21
	9990	S	0	95.76
	9990	S	-1	1.57
	9990	S	-2	0.01
	5000	**R**	**+2**	**0.02**
	5000	**R**	**+1**	**2.16**
	5000	**R**	**0**	**95.96**
	5000	**R**	**-1**	**1.46**

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**).

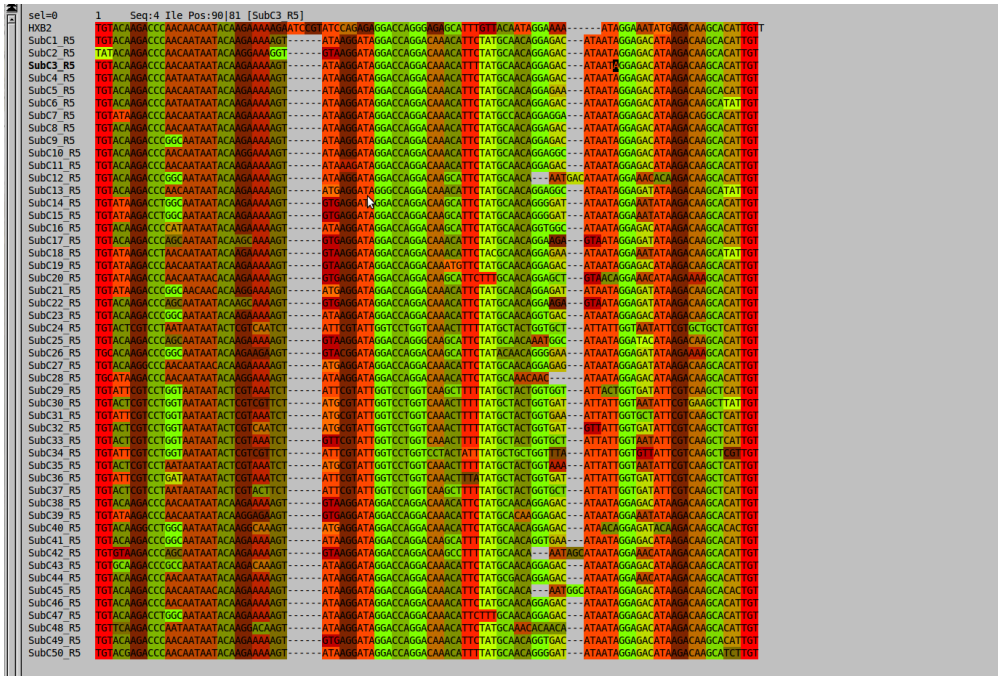
4.6.4 Seq2Res sensitivity test with simulated data

The simulated sequence reads derived from resistant and sensitive "seed" sequences were pooled together at various fixed proportions to generate datasets with varying known prevalence 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.

A)



B)

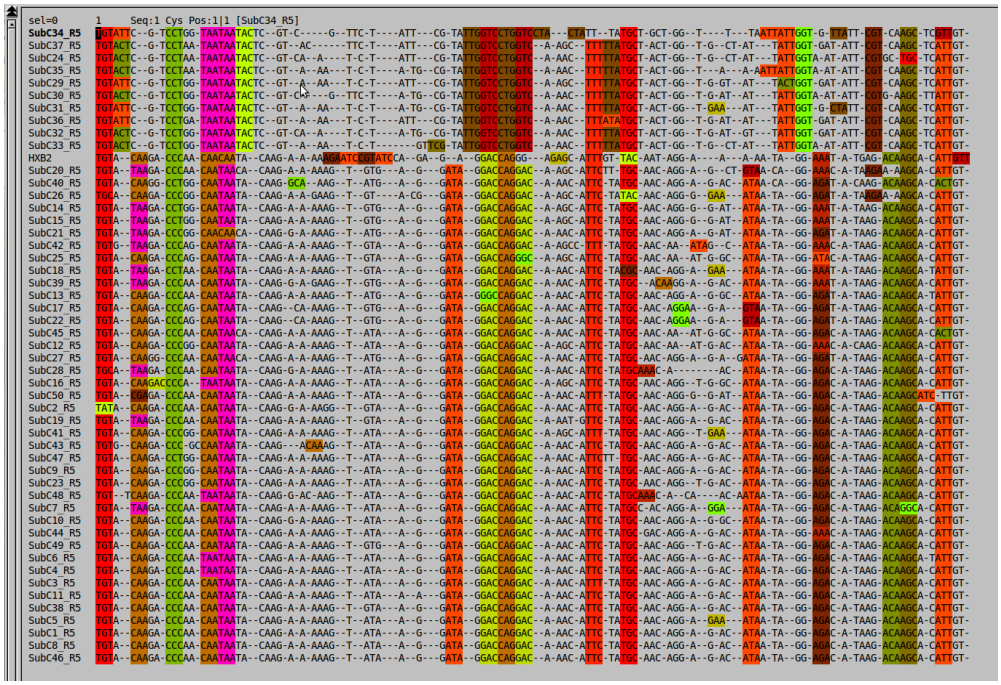


Figure 4. 24: Mapping of HIV subtype C V3 sequence reads to the HIV HXB2 reference sequence using A) RAMICS that maps at codon space and B) Muscle that maps at nucleotide space. Each color stripe represents an amino acid. Both the alignments are shown using alignment viewer called seaview (Gouy et al., 2010).

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.

Table 4. 13: Total number of sequences discarded as non-optimal full-length in the simulated datasets for every sample at all expected prevalence for each DRMs.

Sample	Expected Prevalence (%)	L90M	A71V	K65R	K70R	V75I	F77L	Y115F	F116Y	Q151M
56252	0.1	0	13	0	0	15	24	13	0	0
	1	1	0	2	0	104	103	103	1	2
	5	15	15	15	3	2	-9	6	9	3
	10	35	34	23	4	0	-8	6	14	8
	15	39	39	36	10	10	2	6	13	18
	20	48	51	51	10	8	3	18	25	21
	50	111	155	119	26	29	32	45	61	31

Sample	Expected Prevalence (%)	G48V	I50V	I54V	V82A	K65R	D67N	K70G	K103N
21354	0.1	0	0	11	17	2	17	0	0
	1	2	2	4	105	25	15	2	3
	5	15	7	17	0	143	133	27	28
	10	28	24	37	17	266	260	58	68
	15	41	32	58	26	404	392	76	94
	20	58	44	92	35	592	577	102	124
	50	128	108	208	95	1406	1379	269	313

Sample	Expected Prevalence (%)	G48V	I50V	I54V	V82T	L10I	A71V	K65R	D67N	K70G	K103N
63377	0.1	0	0	14	0	0	14	4	16	1	1
	1	2	2	101	3	3	102	34	23	5	6
	5	6	10	18	10	10	3	120	109	27	33
	10	28	24	34	20	19	19	270	261	50	61
	15	39	32	65	36	31	23	413	393	85	103
	20	54	35	84	38	39	31	529	516	96	120
	50	153	121	229	117	117	106	1322	1306	252	310

Sample	Expected Prevalence (%)	G48V	I54V	V82A	L10I	K65R	D67N	K70G	K103N	Y181C
4134	0.1	0	14	0	0	1	14	0	0	19
	1	7	4	6	6	24	18	3	3	105
	5	9	17	10	9	103	97	30	38	24
	10	30	38	23	22	244	238	56	68	54
	15	63	68	46	44	313	298	73	87	79
	20	75	100	51	54	429	413	91	11	95
	50	161	232	118	123	1138	1117	273	322	298

Sample	Expected Prevalence (%)	L90M	A71V	G73S	M41L	A62V	K65R	V75I	K103N	Y181C
2368	0.1	0	16	1	11	2	2	15	1	22
	1	2	104	3	18	20	22	102	7	108
	5	10	3	13	115	117	124	24	57	17
	10	16	7	17	230	232	231	57	127	53
	15	23	10	28	311	309	319	81	162	82
	20	47	33	48	472	472	484	138	245	127
	50	122	87	106	1112	1112	1152	275	560	267

Table 4. 14: Total number of sequence reads discarded as non full-length after reference mapping in the amplicons of the samples at all expected prevalence.

Amplicon	Prevalence (%)	56252	21354	63377	4134	2368
PR	0.1	172	192	178	186	183
	1	172	193	180	192	182
	5	175	188	170	188	181
	10	180	187	170	195	177
	15	179	190	185	195	160
	20	174	192	155	196	183
	50	150	170	172	192	163
RT1	0.1	2378	2838	2802	2309	2854
	1	2356	2830	2804	2312	2846
	5	2261	2837	2775	2295	2826
	10	2136	2800	2799	2306	2800
	15	2014	2809	2772	2255	2742
	20	1925	2834	2739	2254	2739
	50	1192	2761	2707	2275	2525
RT2	0.1	0	501	478	479	516
	1	0	494	479	475	514
	5	0	501	485	482	515
	10	0	503	484	480	524
	15	0	501	489	476	525
	20	0	483	480	463	539
	50	0	496	475	504	503

Chapter 5

The application of Seq2Res to evaluate ultra deep pyrosequencing as a large-scale, cost-effective alternative to conventional HIV resistance genotyping

5.1 Introduction

HIV exists in an infected individual as a complex heterogeneous population called quasispecies (Yin et al., 2012) primarily arising as a result of HIV's high replication rate and the high error rate of the viral reverse transcriptase (Bebenek et al., 1993; Ji and Loeb, 1992; Preston et al., 1988). During the replication process, HIV develops random mutations (Bebenek et al., 1989; Bebenek et al., 1993; Berkhout et al., 2001; Roberts et al., 1988) in its genes that can result in viral resistance against one or more antiretroviral drugs (Clavel and Hance, 2004; D'Aquila et al., 2003; Kantor and Katzenstein, 2004; Sebastian and Faruki, 2004). Viral variants containing DRMs can be present at varying levels in the viral quasispecies (Devereux et al., 1999; Johnson et al., 2008; Metzner et al., 2009) with these variants emerging to dominate the viral population in response to treatment (Adje et al., 2001; Adje-Toure et al., 2003; Johnson et al., 2008; Marconi et al., 2008).

Approximately 8 million HIV infected individuals in resource-limited countries are receiving antiretroviral therapy by the end of 2012 (UNAIDS, 2012) following the scale-up of treatment programs in 2002 (Beck et al., 2006; Ferradini et al., 2006; Gilks et al., 2006; Stringer et al., 2006). The number of HIV infected individuals with transmitted drug resistant mutations (DRMs) is increasing (Aghokeng et al., 2011; Zaidi et al., 2013) with studies showing that while antiretroviral (ART) drugs increase life expectancy of infected individuals, this increase life expectancy increases the risk of transmission of drug resistant HIV variants to uninfected individuals (Zaidi et al., 2013).

The history of HIV treatment in 1980s has shown that the therapy with a single dose antiretroviral drug or a combination of drugs from a single drug class usually results in treatment failure (Kellam et al., 1994; Larder et al., 1989b; Larder et al., 1991; Larder and Kemp, 1989; Larder et al., 1987). This has shown that administration of a single antiretroviral drug selects the drug resistant variants and increases the chance of rapid drug failure (Hamers et al., 2012; Jackson et al., 2000; Partaledis et al., 1995; Tisdale et al., 1993). A perfect example of this is the emergence of resistance in single dose nevirapine (NVP) programs. NVP, an NNRTI, is prescribed to HIOV positive pregnant woman in order to prevent HIV transmission from mother to child in resource poor settings (Audureau et al., 2013; Chi et al., 2013; Shapiro et al., 2010; Stringer et al., 2010a; Zolfo et al., 2010). The drug is effective in reducing the viral transmission as has been reported in several research reports (Connor et al., 1994a; Connor et al., 1994b; Guay et al., 1999; Jackson et al., 2003; Lallemant et al., 2004). However, studies have shown that the use of single dose NVP to prevent mother to child transmission of the virus could show nevirapine associated resistant mutations (Coovadia et al., 2009; Eshleman et al., 2004b; Eshleman et al., 2005; Eshleman et al.,

2001; Flys et al., 2005; Havlir et al., 1996; Jackson et al., 2000; Loubser et al., 2006; Martinson et al., 2007; Richman et al., 1994b; Tisdale et al., 1993). The persistence of NVP resistant virus in the mothers and children treated with single dose NVP (Hauser et al., 2011) compromises the treatment with subsequent NVP containing highly active antiretroviral therapy (HAART) (Arrive et al., 2007; Chi et al., 2007; Lehman et al., 2012; Martinson et al., 2007). Thus, studies have shown that the first line therapy, which is a combination of at least three fully active ART drugs from different drug classes – Non-Nucleotide Reverse Transciptase Inhibitors (NNRTIs) and Nucleotide Reverse Transcriptase Inhibitors (NRTIs) are necessary for optimum suppression of HIV from replication and resistance development (Gupta et al., 2009; Hamers et al., 2012; Robbins et al., 2003b; Shafer et al., 2003; van Leeuwen et al., 2003; van Leth et al., 2004). For this, the World Health Organization (WHO) recommends drug resistance testing before prescribing ART drugs.

Resistance testing reveals the drug resistance mutations in the HIV quasispecies. HIV with DRMs is present at varying prevalence levels in the quasispecies. The conventional HIV resistance genotyping is limited to detecting the mutations in HIV with prevalence of 20% or greater (Booth and Geretti, 2007; Liang et al., 2011; Wang et al., 2007). Ultra deep pyrosequencing (UDPS) technology has ability to detect HIV with prevalence to as low as 1% or below (Archer et al., 2009; Balduin M, 2011; Bansode et al., 2013; Dudley et al., 2012; Gilles et al., 2011; Hedskog et al., 2010; Hoffmann et al., 2007; Huse et al., 2007; Ji et al., 2012; Ji et al., 2010; Lataillade et al., 2010; Le et al., 2009; Liang et al., 2011; Wang et al., 2007).

Here, we describe the application of the computational tool Seq2Res to HIV resistance testing to a significant dataset generated using UDPS.

5.2 Methods and Materials

The datasets used in this study had been generated as part of the CIPRA-SA study (Comprehensive International Program for Research in AIDS in South Africa) which was a prospective, unblinded, randomized controlled trial of comparing "doctor-initiative-doctor monitored" and "doctor-initiative-nurse-monitored" strategies for antiretroviral drug monitoring in resource poor setting (Sanne et al., 2010). The study population consisted of 831 HIV infected individuals with a CD4+ count less than 350-cells/mm^3 or AIDS-defining illness were enrolled on the study. HIV positive mothers with previous exposure of single dose nevirapine (NVP) drug for prevention of viral transmission from mother to child (PMTCT) during their pregnancy were also included in the study.

562 patients were followed up with the remainder not included in the study for reasons such as drug toxicity, death, and withdrawal of consent or loss to follow-up. Baseline blood samples were retrieved from all 562 patients (sampled from 2005 – 2006). In this instance, baseline describes samples obtained from individuals immediately before initiation of first line antiretroviral therapy (ART). 71% of these patients received the drug combination D4T-3TC-EFV, 20% received D4T-3TC-NVP, 8% received D4T-3TC-LPV/r and 1% received D4T-3TC-NLF.

Virologic failure (VF) to the treatment was defined as decline of viral load less than 1.5 $\log_{10}$ from baseline to 12 weeks of treatment or two consecutive samples from a patient taken four weeks apart have viral load greater than 1000 RNA copies/ml (Sanne et al., 2010). VF to first line ART was identified in 79 patients, with 15 patients failing second-line therapy (Sanne et al., 2010). Blood samples had been retrieved for all of these individuals upon failure detection.

From all the obtained samples, the entire protease and reverse transcriptase genes of HIV had been amplified as three fragments - PR (HXB2 *pol* position 169 - 480) RT1 (HXB2 *pol* position 466 – 795) and RT2 (HXB2 pol position 796 – 1185) using HIV subtype C specific primers. Ten UDPS sequencing runs using the Roche/454 Junior platform (hereafter only Junior) platform had been attempted for 471 samples using MID tags to pool 48 samples per sequencing plate. Further, sequencing was attempted for 630 samples using the FLX platform (hereafter only FLX). 12 FLX runs were undertaken, dividing each plate into 8 distinct sections with 8 MID tagged samples per section for each sequencing run.

Conventional genotyping results were also available for 349 of the samples. All of the sequence data had been generated by our collaborators in the laboratory of Prof Maria Papathanasopoulos at the University of the Witwatersrand Medical School, South Africa.

Sequence data from all samples (baseline and first line VF samples) were analyzed using Seq2Res. To facilitate direct comparisons with the Sanger data the prevalence cutoff was set to 20%, consistent with the reported ability of Sanger-based sequencing to detect resistant variants to a level of 20% in the viral population (Hudelson et al., 2010; Larder et al., 1993; Leitner et al., 1993; Schuurman et al., 1999; Van Laethem et al., 1999). The presence of resistance in the UDPS genotypic data was further explored at prevalence levels of 15%, 10%, 5% and 1% of the amplified and sequenced viral population.

Every sequence read was tested for drug resistance using the Stanford HIV drug resistance database resistance interpretation algorithm (Liu and Shafer, 2006; Rhee et al., 2003; Shafer, 2006). The percentage of sequences that were predicted as resistant for a particular antiretroviral drug by the algorithm was calculated. Using a number of prevalence cutoffs (20%, 15%, 10%, 5% and 1%), if the percentage of sequences predicted as resistant to a drug from the total number of sequences analyzed for a sample, was greater or equal to the cutoff, the sample was predicted resistance for that drug.

As the baseline regimens of each of the individuals was known, resistance was defined as predicted resistance to one or more of the drugs in the individuals regimen. In the very rare event of an individual's regimen not being recorded, resistance was defined as predicted resistance to one or more of the entire spectrum of drugs used as first line therapy in the study (D4T, 3TC, EFV, NVP, LPV/r or NLF).

In the case of VF, samples that failed the first line therapy, resistance was defined as predicted resistance to one or more of the drugs in the first line therapy. Similarly, individuals that failed second line therapy were predicted resistant if one or more of the drugs in the second line therapy were predicted to be resistant.

The number of samples with and without predicted resistance was obtained and statistical significance was calculated using Fisher's exact test.

5.3 Results

In the preliminary assessment of the sample's sequence data from FLX and Junior, the samples in which protease (PR) or reverse transcriptase (RT) or both were not amplified were not considered for subsequent analysis. A total of 599 samples and 468 samples that were sequenced in FLX and Junior platform respectively from both baseline and first line VFs had both PR and RT sequences (**Table 5.1**) and were considered for analysis.

Out of the samples that were eligible for analysis, 464 samples were sequenced using both the FLX and junior platforms, 327 samples using both the FLX and population based Sanger genotyping method and 257 samples with both the junior platform and population based Sanger genotyping method (**Table 5.2**).

5.3.1 Analysis of baseline samples

5.3.1.1 Genotyping of baseline samples using the Roche/454 FLX sequencing platform

FLX sequencing was successful for baseline samples from a total of 526 patients of which 187 samples had previous ARV exposure as a result of PMTCT therapy while the remaining 339 had no previous exposure to ARVs. The eventual clinical outcome of all of these individuals was known and showed that out of the 339 no-PMTCT patients, 50 had exhibited VF while 289 exhibited virologic success (VS) during the course of follow-up. On the other hand, out of 187 PMTCT exposed patients, 25 exhibited VF and 162 exhibited VS.

Table 5. 1: Total number of samples attempted to be sequenced and total samples considered for downstream analysis

Sequencing Method	Total attempted to be sequence	Total samples with PR and RT
Consensus	349	349
FLX	630	599
Junior	471	468

Table 5. 2: Number of samples with and without previous PMTCT exposure that are sequenced in two or more sequencing methods

	Consensus and FLX	Consensus and Junior	FLX and Junior
Number of PMTCT samples	126	104	146
Number of non-PMTCT samples	201	153	318
Total	327	257	464

The number of samples with and without predicted drug resistant HIV in the PMTCT and no-PMTCT groups at varying prevalence cutoffs (1%, 5%, 10%, 15% and 20%) that were predicted drug resistant to at least one baseline drug increased when the prevalence cutoff was decreased to 1% (**Figure 5.1**). In the no-PMTCT VF group of 50, there was only one sample exhibiting resistant virus to at least one baseline drug at prevalence level range 5% to 20% but this increased to five samples (10%) at prevalence level of 1%. Similar increments in the number of individuals predicted as resistant as the prevalence levels decreased were observed in other groups as well (**Figure 5.1**).

At the 1% prevalence cutoff in the no-PMTCT group there was a significant difference between the number of clinical viral failures predicted as resistant when compared with the number of clinical viral successes predicted as resistant ($p < 0.05$, Fisher's exact test) while a similar observation was observed at the 15% prevalence cutoff in the PMTCT exposed group (**Figure 5.1**).

5.3.1.2 Genotyping of baseline samples using the Roche/454 Junior platform sequencing platform

407 patients were sampled at baseline and sequenced using Junior, 250 patients had no previous ARV exposure through PMTCT and 147 patients had previous PMTCT therapy. The clinical outcome showed that out of 250 non-PMTCT patients, 40 had VF and 210 had VS. On the other hand, out of 147 PMTCT exposed patients, 21 had VF and 136 had VS (**Figure 5.2**).

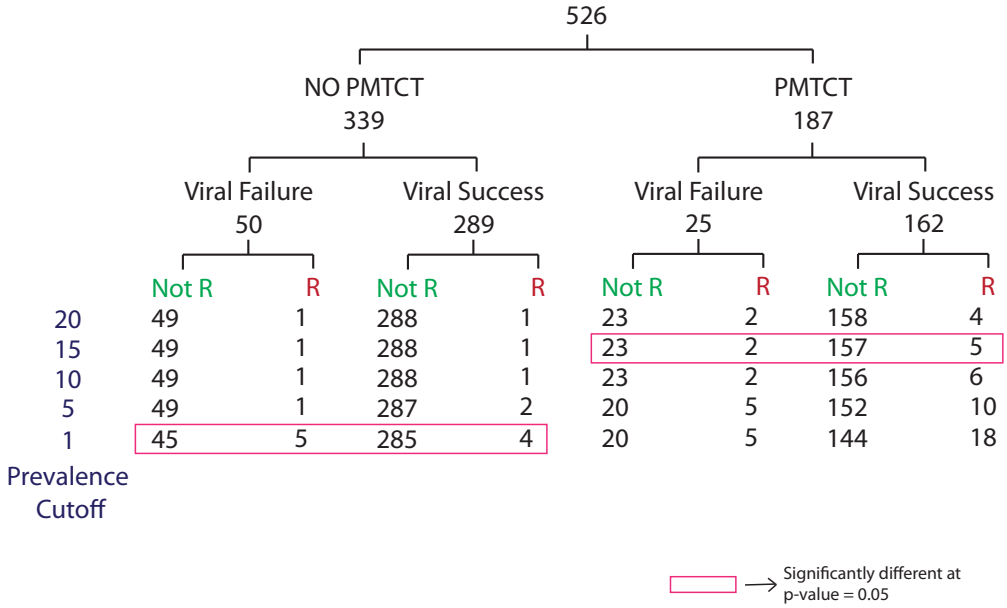


Figure 5.1: Resistance call to at least one drug in baseline regimen at different prevalence levels for the PMTCT and Non-PMTCT exposed baseline samples sequenced using FLX technology. The data in red rectangle showed significant difference (p-value <0.05) using two-tailed t test.

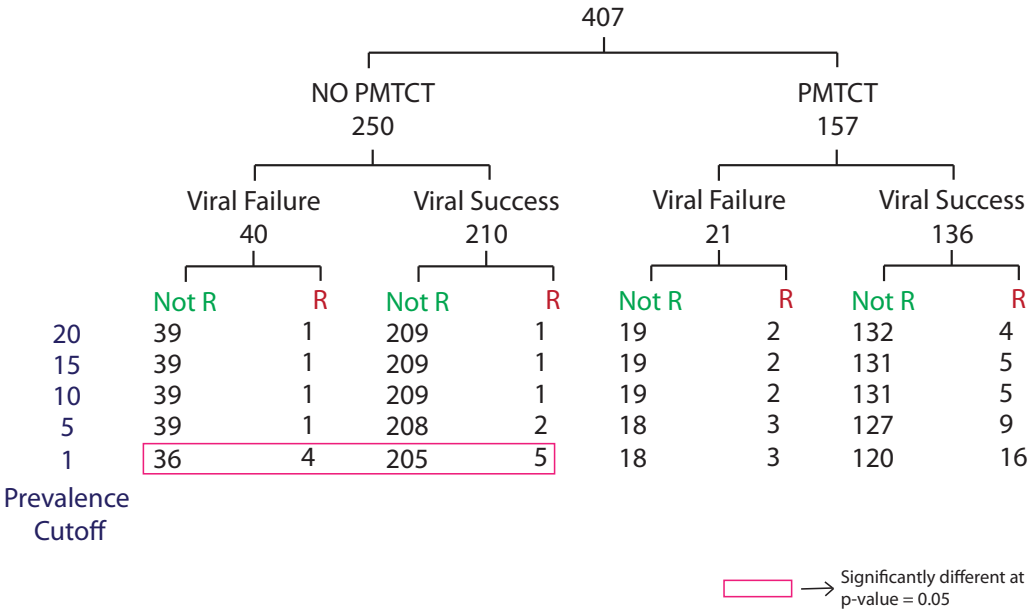


Figure 5.2: Resistance call to at least one drug in baseline regimen at different prevalence cutoffs for PMTCT and non-PMTCT exposed baseline samples sequenced using junior technology. The data in red rectangle showed significant difference (p-value <0.005) using two-tailed T test.

The obtained baseline blood samples were sequenced using junior platform and again analyzed using Seq2Res. The observation on the number of samples with and without predicted resistant HIV showed that in the no-PMTCT VF group of 40, there was only one individual predicted resistant to at least one drug in baseline at prevalence cutoff 20% and that increased to four individuals at prevalence cutoff 1% (**Figure 5.2**). Similar increments in the number of samples with predicted resistant HIV was observed in the no-PMTCT VS group as well as in PMTCT VF and VS groups (**Figure 5.2**). The highest number of samples (16 samples) with predicted resistant HIV was observed in PMTCT VS group at a prevalence cutoff 1%. Further, in the No-PMTCT group there was a significant difference between the numbers of clinical viral failures predicted as resistant when compared with the number of clinical viral successes predicted as resistant ($p < 0.05$, Fisher's exact test, **Figure 5.2**).

5.3.1.3 Comparison of number of sequence reads per baseline sample generated by Roche/454 FLX and Roche/454 Junior

Sequencing had been successful on both FLX and junior platforms for 464 samples. The initial analysis focused on comparing the number of sequence reads generated by each platform for each sample and identifying if 'deeper' sequencing coverage resulted in more accurate prediction of resistance. We saw that the FLX platform generated on average 6412 sequence reads per sample (standard deviation 2297) while the Junior platform generated an average 1903 sequence reads per sample (Standard deviation 595, **Figure 5.3**). Thus, it is clear that the FLX platform produced significantly (P-value $< 2.2^{-16}$) more reads per sample than the junior platform.

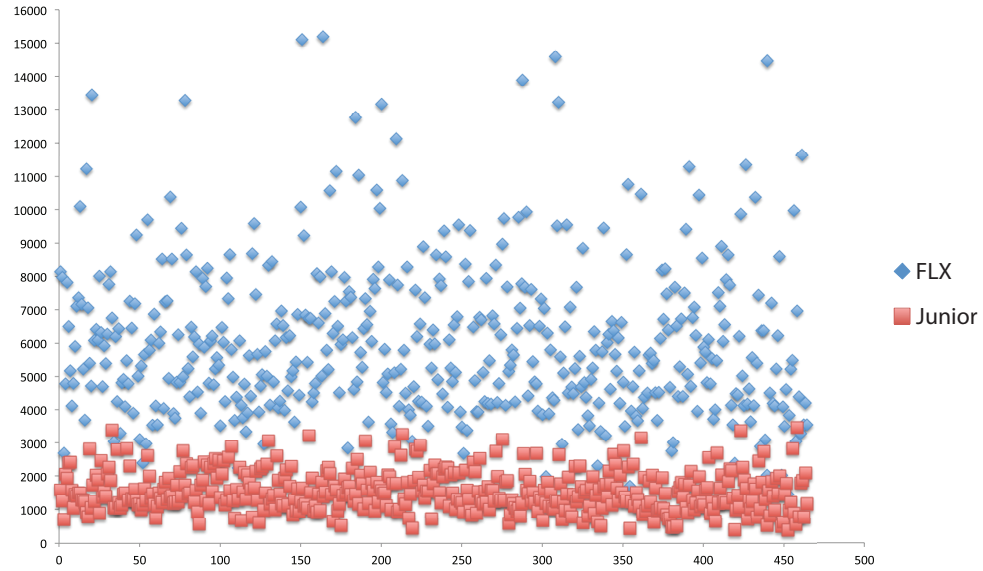


Figure 5. 3: Comparison of the number of reads generated for the baseline samples sequenced on both FLX and Junior 454 pyrosequencing.

5.3.1.4 Comparison of genotyping results between Roche/454 FLX and Roche/454 Junior platforms on baseline samples

Baseline samples from 405 individuals had been sequenced using both the FLX and Junior platforms. 249 had no previous PMTCT therapy while 156 had previous exposure to ARVs as a result of PMTCT therapy. Of these 249 patients, 40 exhibited VF and 209 exhibited VS to the first line antiretroviral therapy regimen. Of the 146 PMTCT exposed patients, 21 exhibited VF and 135 exhibited VS in first line antiretroviral therapy (**Figure 5.4**).

Our results show that the number of samples with and without predicted resistance in both the PMTCT and no-PMTCT groups is not significantly different between the FLX and Junior sequencing platforms at all prevalence cutoffs (**Figure 5.4**).

Thus, despite the significantly higher numbers of sequence reads generated per individual for the FLX data, both the genotyping results from both platforms were completely comparable.

5.3.1.5 Comparison of ultra deep pyrosequencing and population based Sanger method for resistance prediction using baseline samples

In section 3.1.4 we showed that there was no significant difference between the FLX and Junior platforms for the prediction of resistance. While genotyping results from the Junior and FLX sequencing platforms are comparable between each other, this is essentially meaningless unless these results are comparable to that of the current "gold-standard" of population based Sanger genotyping method, to be used as a replacement. Thus, we compared the junior platform (now also referred to as UDPS) results with those from the Sanger-based genotyping.

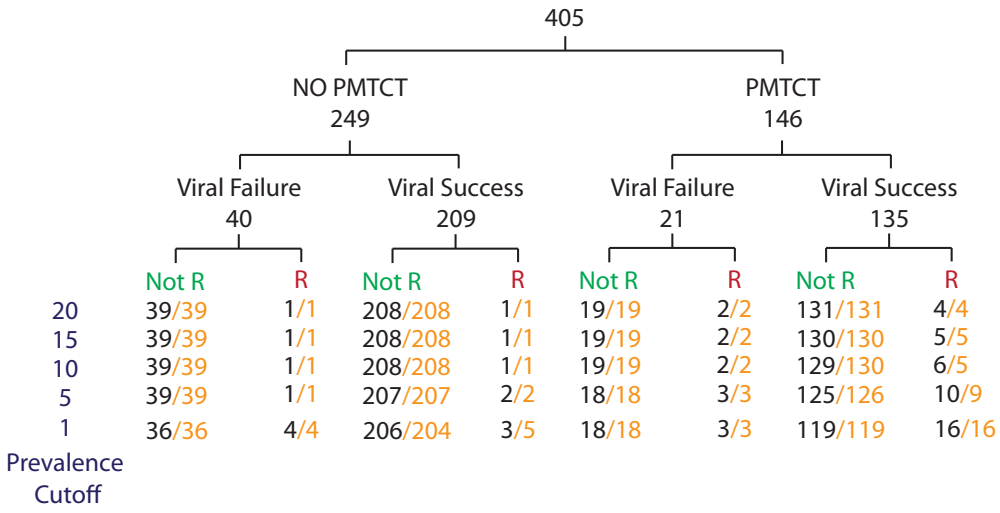


Figure 5.4: Resistance call to at least one drug in baseline regimen at different prevalence cutoffs for PMTCT and non-PMTCT exposed baseline samples sequenced using both FLX and Junior. The sequenced viral population in the samples sequenced using FLX and Junior that are called as resistant to at least one drug in baseline regimen are shown in Black and Orange respectively.

239 of 302 baseline samples were sequenced using both junior platform and conventional population based Sanger genotyping technology. 128 of them had no previous PMTCT therapy exposure and 111 had previous PMTCT therapy exposure. Out of 128 no-PMTCT patients, 15 exhibited VF and 113 exhibited VS in first-line antiretroviral therapy. Similarly, out of 111 previously PMTCT exposed patients, 10 exhibited VF and 101 exhibited VS (**Figure 5.5**).

The results from the junior platform showed identical numbers of individuals predicted as resistant at the 20%, 15% and 10% prevalence levels with now significant difference from the resistance predictions from the Sanger-based genotyping (Figure 5.5). The numbers of individuals predicted as resistant by the Junior platform increased at the lower prevalence levels (Figure 5.5), however these observations were still not significantly different from the predictions of the Sanger-based genotyping.

Thus, it appears for baseline samples at least, the UDPS resistance genotyping approaches employed here are directly comparable to that of Sanger-based resistance genotyping.

5.3.2 Analysis of virologic failure samples

Using the baseline samples, we showed that there was no significant difference between FLX and Junior platform and between UDPS and population based Sanger method. We repeated the platforms comparative analysis test using samples collected following clinical evidence of viral failure in individuals on 1st line therapy (either a

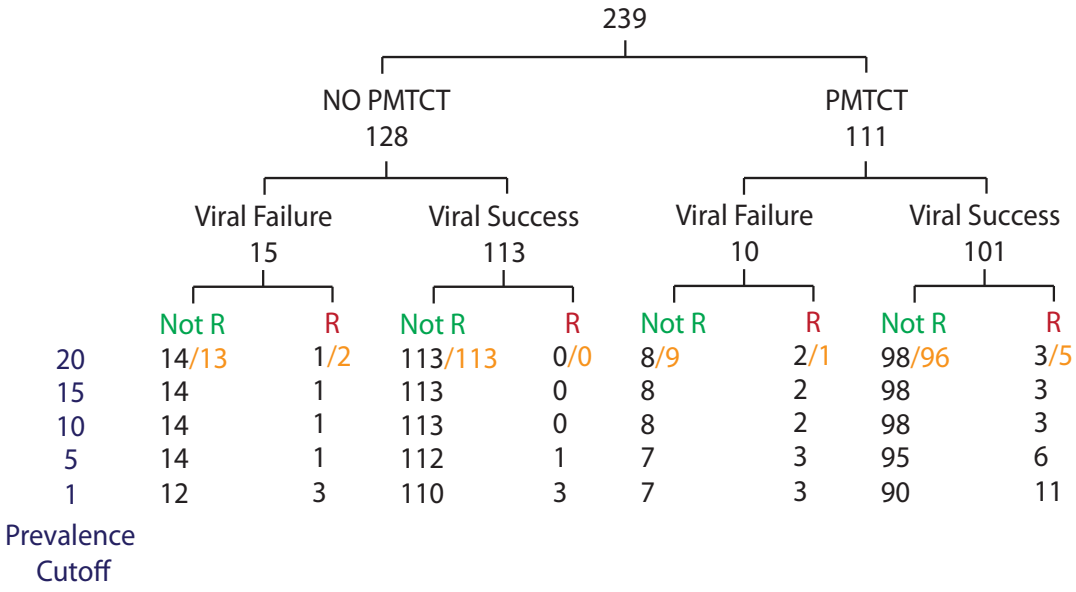


Figure 5.5: Comparison of the number of samples with the amplified and sequenced viral population predicted as resistant or non-resistant sequenced using UDPS and population based Sanger method at the prevalence cutoff 20%. A sequenced viral population in a sample was called resistant if any one drug in baseline regimen was resistant to it. Samples called as resistant to at least a baseline drug by HTS and population based Sanger method are shown in Black and orange respectively.

decline of < 1.5 $\log_{10}$ in viral load from baseline to 12 weeks of treatment or two consecutive viral loads 4 weeks apart of >1000 RNA copies/ml).

5.3.2.1 Resistance genotyping of samples collected from individuals at first-line virologic failure (VF1) using Roche/454 FLX platform

51 of the first line ART VF1 samples had been sequenced using FLX technology. 15 of these had previous ARV exposure through PMTCT while 36 had no previous exposure through PMTCT.

Genotyping using the FLX platform predicted resistance to at least one of the first line drugs at all prevalence levels in 14 out of 15 the PMTCT samples (**Figure 5.6**). On the other hand, in the no-PMTCT sample, 23 out of 36 had predicted resistance to at least one of the first line drugs at all prevalence levels while 13 had no predicted resistance (**Figure 5.6**).

The observation of the number of samples with and without predicted resistance showed that there was a significant difference between the PMTCT and no-PMTCT groups at all prevalence cutoffs. The observation also showed that the viral resistance prediction in the samples from PMTCT group was more than in no-PMTCT group at prevalence cutoffs using FLX system.

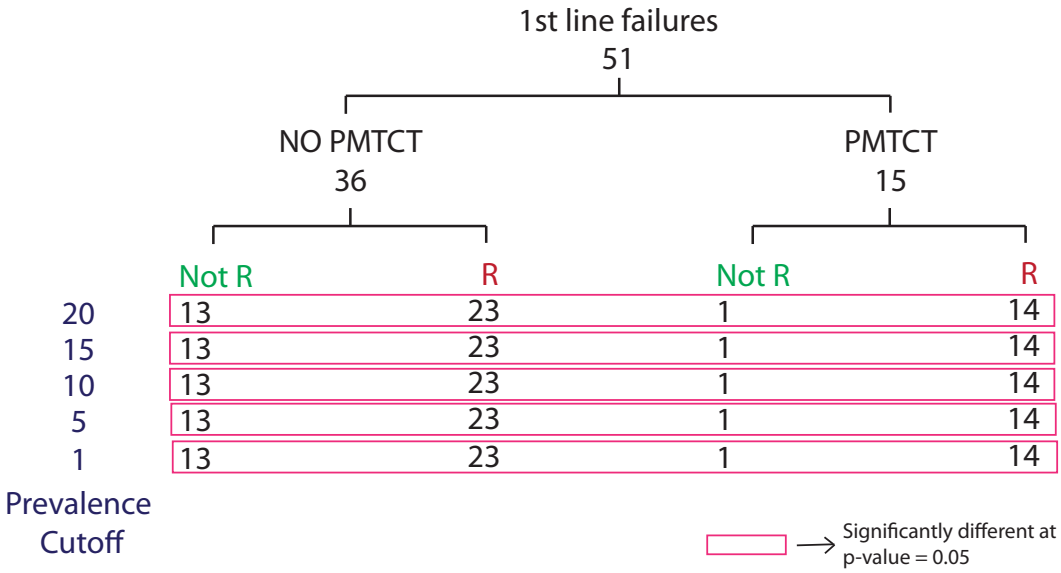


Figure 5.6: Number of samples with sequenced viral population showing predicted resistance and non-resistance to a drug using Roche/454 FLX platform, in no previous PMTCT therapy and with PMTCT therapy that had virologic failure at first line ART. Significant difference (p-value <0.05) was observed at all prevalence cutoffs.

5.3.2.2 Resistance genotyping of samples collected from individuals at first line virologic failure (VF1) using Roche/454 Junior

Out of the 36 first-line therapy failure samples sequenced using the junior platform, 23 had no previous PMTCT therapy exposure while 13 had previous PMTCT therapy exposure. The numbers of predicted resistant and non-resistant viral samples were calculated at all prevalence cutoffs (**Figure 5.7**).

We observed that the amplified and UDPS sequenced viral population in all 13 VF samples in PMTCT group were predicted resistant when the prevalence cutoff was below 20% (**Figure 5.7**). At the 20% cutoff, 12 out of 13 PMTCT samples (92.3%) were predicted resistant (**Figure 5.7**). Similarly in no-PMTCT group, 14 out of 23 samples (60.86%) were predicted resistant at prevalence cutoffs 20% down to 5%. The number of samples with predicted resistance increased to 15 out of 23 (65%) at a prevalence cutoff of 1%. The observation of the number of samples that were predicted drug resistant showed that there were significant differences between the PMTCT and no-PMTCT at all prevalence cutoffs. The result obtained was similar to the result in VF1 samples sequenced using FLX platform, which indicated that the likelihood of predicting resistance in PMTCT group is more than in no-PMTCT group.

Genotypic results from samples sequenced using Roche/454 FLX and Roche/454 Junior showed that there was no significant difference in the number of samples predicted as resistant at all prevalence cutoffs. We further studied the genotypic results from samples sequenced in both Roche/454 FLX and Roche/454 Junior to validate the above result.

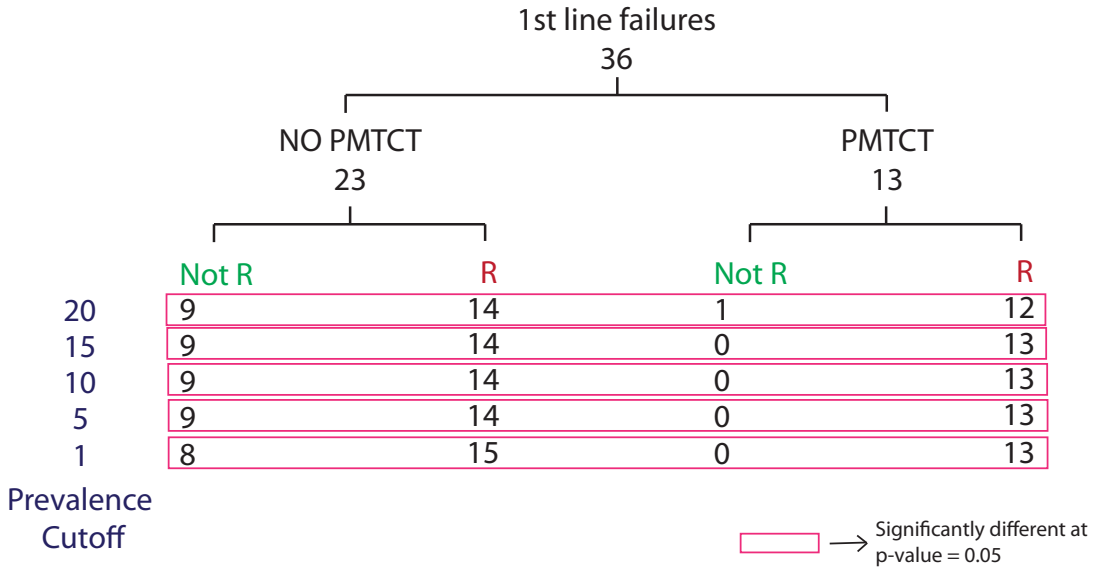


Figure 5.7: Number of samples with sequenced viral population showing predicted resistance and non-resistance to a drug using Roche/454 Junior platform, in no previous PMTCT therapy and with PMTCT therapy that had virologic failure at first line ART. Significant difference (p-value <0.05) was observed at all prevalence cutoffs.

5.3.2.3 Comparison of genotyping results between Roche/454 FLX and Roche/454 Junior platforms on first line virologic failure samples

50 VF1 samples sequenced using both the FLX and junior platforms were available. 36 of them had no previous exposure to the drug NVP and 14 had previous exposure to the drug NVP through PMTCT. The number of VF1 samples on both no-PMTCT and PMTCT groups that were predicted resistance were calculated at different prevalence cutoffs (**Figure 5.8**).

We observed the number of VF1 samples that were predicted resistant was consistent from 20% to 5% for both FLX and junior platforms but increased by one at 1% cutoff (**Figure 5.8**). We did not observe significant difference in the number of samples predicted as resistant at all prevalence cutoffs. 13 of 14 (93%) PMTCT samples and 69% or above no-PMTCT samples were predicted resistant by both FLX and Junior. The observation showed that both Junior and FLX are comparable for HIV drug resistance genotyping of viral failure samples.

5.3.2.4 Comparison of the genotyping performance of the Roche/454 Junior platform and conventional population based Sanger genotyping method using first line virologic failure samples

Genotypic data from both junior platform and conventional population based Sanger genotyping method were available from 13 VF1 samples. Out of the 13 patients, 6 had no previous PMTCT therapy and 7 had PMTCT therapy exposure. The observation of the number of VF1 samples that were predicted with or without resistant for both approaches at 20% prevalence cutoff showed that there was 100%

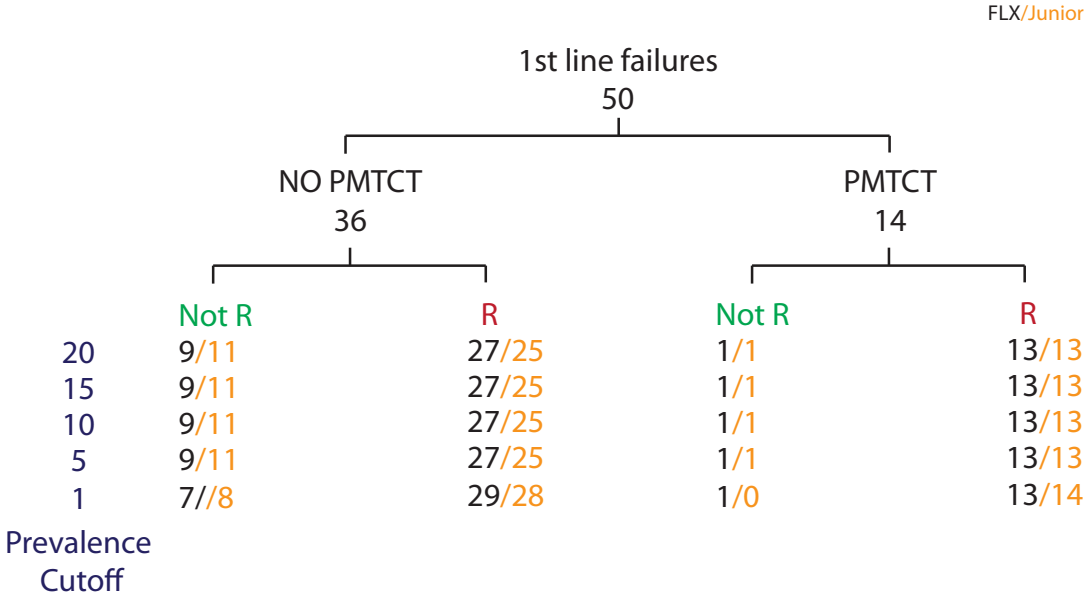


Figure 5. 8: Number of samples with sequenced viral population showing predicted resistance and non-resistance to a drug using Roche/454 FLX and Roche/454 Junior platforms, in no previous PMTCT therapy and with PMTCT therapy that had virologic failure at first line ART. No Significant difference (p-value <0.05) was observed at all prevalence cutoffs.

concordance across all clinical outcome categories between the Sanger-based and UDPS-based resistance genotyping approaches (**Figure 5.9**).

5.3.2.5 Resistance to nevirapine is more likely to be present at baseline in individuals previously exposed to nevirapine through PMTCT

We compared the resistance predictions for PMTCT versus no-PMTCT therapy exposed individuals and identified the percentage of individuals with predicted resistance to NVP at baseline for conventional Sanger genotyping method, FLX and Junior platform (**Figure 5.10**). In all comparisons we found that the percentage of individuals with predicted resistance to NVP was always significantly higher in the PMTCT therapy exposed group when compared with the no-PMTCT therapy exposed group. This discordance became more evident at the 'deeper' prevalence cutoff (**Figure 5.10**), suggesting a large number of PMTCT-exposed individuals were harboring low-abundance NVP resistant viruses.

To ascertain whether the prediction of NVP resistance in PMTCT exposed individuals correlates with the time since NVP exposure, we compared the time since NVP exposure in baseline PMTCT samples with predicted NVP resistance and those predicted as susceptible to NVP. At prevalence thresholds of 15% and below (for both FLX and Junior platforms) we found that the prediction of NVP resistance significantly correlates with time since NVP exposure (**Table 5.3**). The median number of days since PMTCT exposure was observed to be 674 days for those individuals predicted as susceptible to NVP and 172 days for those predicted as resistant.

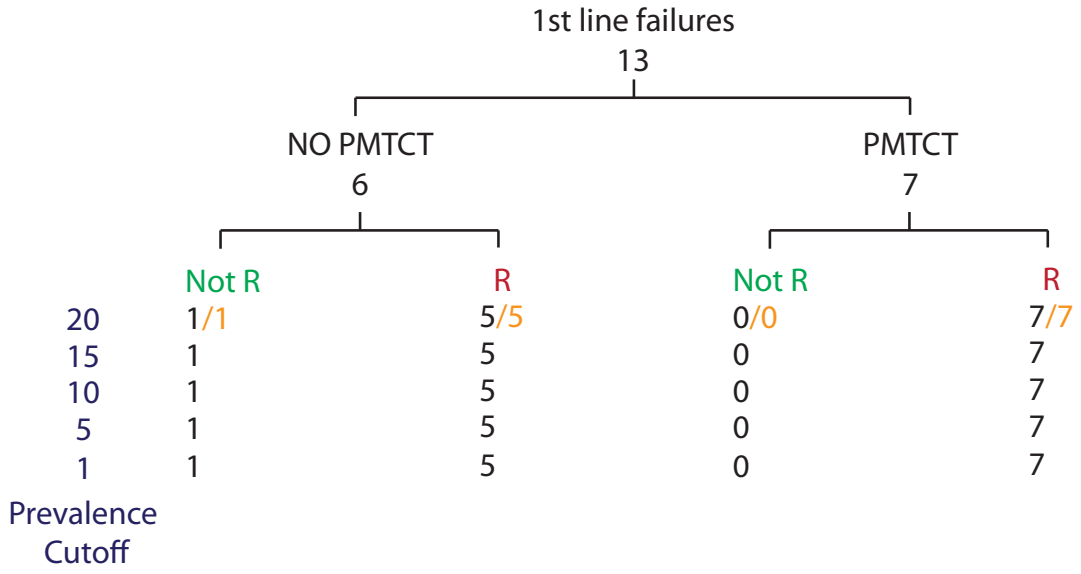


Figure 5.9: Comparison of the number of resistant and non-resistant samples that had first line ART failure, sequenced using Roche/454 Junior and population based Sanger method at the prevalence cutoff 20%. The sequenced viral population in a sample was called resistant if a drug in baseline regimen was resistant to it. Samples called as resistant to at least a baseline drug by FLX and population based Sanger method are shown in Black and orange respectively.

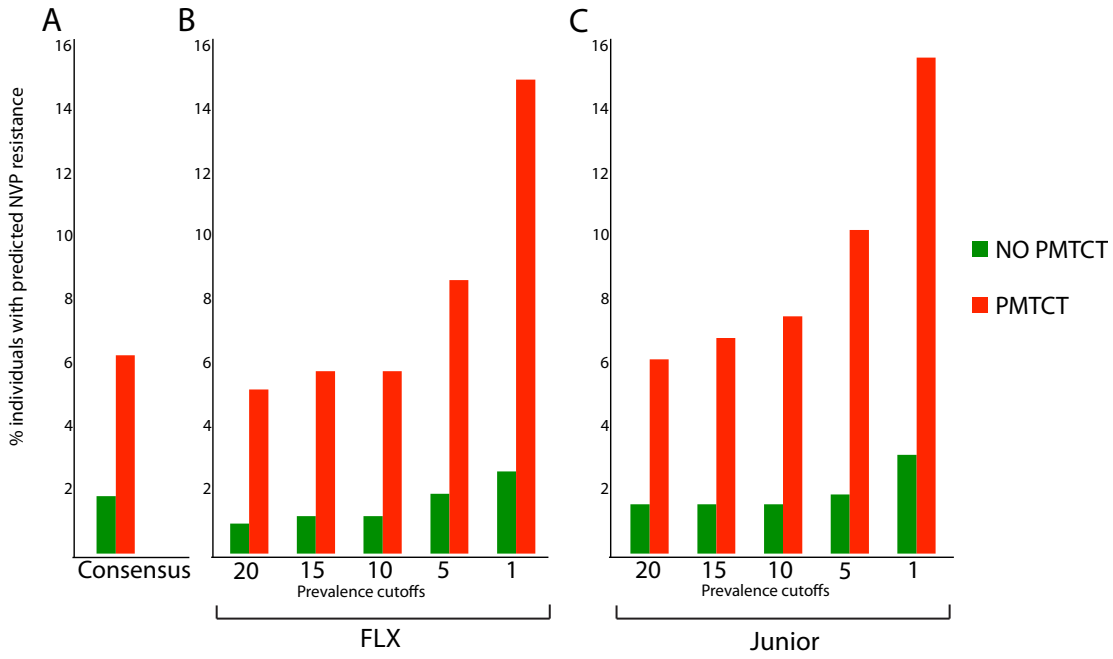


Figure 5.10: The percentage of non-PMTCT exposed and PMTCT exposed baseline samples from patients with predicted NVP resistance sequenced using A) conventional method B) Roche/454 FLX at prevalence cutoffs 20%, 15%, 10%, 5% and 1% C) Roche/454 Junior at prevalence cutoffs 20%, 15%, 10%, 5%and 1%..

Table 5.3: Prediction of nevirapine resistance significantly correlates with the time of nevirapine exposure

Median days since NVP exposure			
Prevalence	***Sensitive***	***Resistant***	***p-value***
20	661.5	178	0.1711
15	661.5	136	**0.04492**
10	660	136	**0.03447**
5	663	193	**0.006021**
1	711	221.5	**0.0005366**

5.4 Discussion and Conclusions

We have analyzed 562 baseline samples and 79 first line ART viral failure (VF1) samples using UDPS (both FLX and Junior platforms) and conventional population based Sanger genotyping method. The baseline samples were collected in 2005 – 2006. The samples were grouped as – individuals with prior exposure to ARVs through PMTCT therapy and individuals without prior exposure to ARVs.

5.4.1 Genotyping results from the Roche/454 Junior platform are comparable to the Roche/454 FLX platform

The junior platform is a desktop sequencing platform with a single sequencing plate that has the capacity of generating up to 100,000 reads per run (70,000 for amplicon data) and a throughput of ~35 megabyte data (www.454.com). The FLX platform has similar sequencing chemistry like Junior platform and is a larger sequencing platform that has the capacity of generating up to 1,000,000 reads per run and the throughput of ~450 megabytes (www.454.com). The number of reads per run and the depth of sequencing per sample limit the number of samples that can be sequenced in Junior and FLX platforms. As the junior platform has lower throughput than the FLX platform, more samples can be pooled for sequencing on the FLX platform than in the junior platform for the same sequencing coverage per sample.

In our analysis the FLX data was not generated to achieve equal coverage as the Junior data but was, in fact, generated to explore whether 'deeper' sequencing (i.e. higher coverage) is a more sensitive approach for HIV drug resistance genotyping.

Despite the FLX platform producing significantly higher number of sequence reads per sample than Junior (almost three times as many reads per individual), there was no significant difference in the resistance prediction in 405 baseline samples at all prevalence levels (20%, 15%, 10%, 5% and 1%). This implies that, in our data at least, the FLX approach is not sequencing at a 'deeper' level and, thus, finding very lowly abundant variants but is more likely sequencing re-sampled data generated by the PCR step. This is further supported by the observation that the prevalence of many DRMs is equivalent between the FLX and Junior data. Thus, the higher coverage of the FLX approach doesn't give any significant advantage over the lower number of reads generated by the Junior platform and, in a resource limited setting at least, the lower coverage from the Junior platform is sufficient for HIV resistance genotyping. Whether this holds true in populations with a longer history of ARV programs and a wider selection of available ARVs is not known, and should be explored further in a similar manner and depth of sampling to our study.

Given its easier setup and usability and lower setup costs the junior platform is undoubtedly preferred over the FLX platform for HIV resistance genotyping in resource limited settings. The price of a junior platform (~$125,000) is much cheaper than FLX (~$500,000). Two publications have previously evaluated the Junior platform in terms of cost in relation to HIV resistance genotyping with Dudley et al obtaining high quality HIV genotypic data sequencing 48 samples in a single Roche/454 Junior platform at the cost of $20 per sample (Dudley et al., 2012) while Hezhao Ji et al sequenced only HIV PR gene from 96 samples using 1/16th of the full PicoTiterPlate at the cost of $32.46 per sample and HIV PR and RT genes at same sequencing capacity for $52.75 per sample (Ji et al., 2010). The cost per sample included both labor cost and material cost. While the Roche/454 platform has

historically been used for HIV drug resistance genotyping because of its longer read lengths, the read lengths of a number of other platforms such as the Illumina and Ion Torrent platforms are increasing thus reducing Roche/454's advantage. This is further compounded by the fact that Roche have recently announced that they are discontinuing the 454 sequencing platform by 2016 to focus on the development of a desktop PacBio system. Further, the significantly higher levels of coverage on other approaches means that there can potentially be even further cost savings using these sequencing approaches for HIV drug resistance genotyping in resource limited settings with a high burden of HIV.

5.4.2 Evidence of minor drug resistant HIV variants in baseline samples

Numerous studies have shown that HIV infected patients that are drug naïve could be harboring viral variants containing primary DRMs (Balduin M, 2011; Bansode et al., 2011; Hamers et al., 2011a; Kozal et al., 2011; Metzner et al., 2011; Simen et al., 2007; Simen et al., 2009; Varghese et al., 2009). Simen et al (Simen et al., 2009) studied the risk of VF in the subsequent ART treatment due to minor drug resistant variants at baseline using conventional Sanger genotyping and UDPS methods. The authors observed 113 mutations below 20% prevalence in the sample; 45 of them (39.8%) were found at 5% or lower prevalence in the viral population. 7 of 84 (8.33%) patients having low abundance NNRTI resistant mutations (<20% mutation prevalence) at baseline experienced VF at the first line ART. Consistent to this report, we also observed 5 of 50 (10%) of the baseline no-PMTCT samples that were predicted resistant at 1% prevalence cutoff indicating that they harbored the low abundant drug resistant HIV variants. Kozal et al (Kozal et al., 2011) also reported

that 147 patients in the cohort of 411 had DRMs when sequenced to 0.4% prevalence level in the viral population. Similarly, Metzner et al (Metzner et al., 2011) reported 20 of 246 (13.7%) drug naïve patients had minor drug resistant viral variants.

Thus, the studies described above showed that primary DRMs in baseline samples from drug naïve patients could be expected. In the high HIV burden regions like sub-Saharan African, the drug resistant viral variants at baseline could be expected even higher. This is supported by Hamers et al (Hamers et al., 2011) whose study on PASER-M cohort findings showed that high prevalence of primary DRMs in treatment naïve individuals can be expected in ART rolled out regions. The authors had pretreatment genotypic data for 2436 baseline samples collected from different African countries. They observed primary DRMs in 5.6% of 2436 patients (range: 1.1% of 176 in South Africa to 12.3% of 179 in Uganda). According to the authors, the high prevalence of primary DRMs in Uganda could related to earlier roll out of ART in the country than other countries (South Africa, Nigeria, Kenya, Zambia and Zimbabwe). Transmission of drug resistant HIV variants to healthy individuals in the ART scaled up region could have lead to high prevalence of primary DRMs (Varghese et al., 2009). More supportive data came from the ART scaled up country – Thailand where a study showed 4.9% of 499 patients had primary HIV-1 drug resistance (Sungkanuparph et al., 2012). Gupta et al (Gupta et al., 2012) and Aghokeng et al (Aghokeng et al., 2011) observed high rate of increase in prevalence of DRMs in ART scaled up regions around the world. The drug resistant HIV variants at baseline have necessitated the surveillance of primary DRMs in the ART scaled up region and it is recommended (Hamers et al., 2011; Sungkanuparph et al., 2012).

The use of NVP in PMTCT therapy has been observed to result in viral variants with the drug-associated mutations in the HIV quasispecies (Arrive et al., 2007; Flys et al., 2005; Johnson et al., 2005; Martinson et al., 2007). Thus, as expected we observed a higher number of samples with predicted resistance to at least one drug in the PMTCT-exposed group than in the no-PMTCT group at 1% prevalence cutoff. We observe 2.6% of no-PMTCT baseline samples had predicted resistance to at least one drug while 12.2% of PMTCT had predicted resistance to at least one drug using genotypic data from FLX. Similarly, a higher number of baseline samples were predicted resistant from the PMTCT group than the no-PMTCT group using genotypic data from junior platform. Various studies also showed that high number of individuals treated with sdNVP in PMTCT therapy had NVP resistance. For example Arrive et al (Arrive et al., 2007) studied NVP resistance 4-8 weeks after receiving sdNVP and they observed 35.7% of PMTCT mothers has the drug-associated resistance. Flys et al (Flys et al., 2005) studied NVP resistance in individuals that received sdNVP after 1 year or more. The authors observed high level NVP resistant mutation K103N in 8 of 9 women and 4 of 5 infants 6-8 weeks after sdNVP. The mutation persisted in 3 of 9 women and 1 of 5 infants after 12-24 months of sdNVP administration. Johnson et al (Johnson et al., 2005) studied genotypic data from 50 South African women before and after sdNVP administration using conventional population based Sanger genotyping method. They found that the NVP resistance emerged in at least 65% of them. They expected the prevalence of NVP resistance would be more using UDPS method. The prevalence of NVP resistance in these studies varied from our observation due to variation in the sample size. But these all studies including our finding showed that NVP resistance is high in sdNVP exposed individuals.

5.4.3 The presence of NVP resistance correlates with the time since sdNVP exposure

The World Health Organization (WHO) recommends the use of a sdNVP for HIV infected pregnant mother for HIV prevention from mother to child (WHO, 2008). However, a number of research publications have shown that the sdNVP rapidly develops HIV variants with NVP resistant mutations (Eshleman et al., 2001; Hudelson et al., 2010; Jackson et al., 2000; Palmer et al., 2006). The single dose NVP is provided one time to the pregnant women before labor. The effect of NVP decreases in those women in PMTCT therapy as the time goes on and thus NVP resistant viral variants decline in them (Eshleman et al., 2001; Kassaye et al., 2007). However, studies have shown the persistence of the minor NVP resistant viral variants (Flys et al., 2005; Flys et al., 2006; Loubser et al., 2006; Palmer et al., 2006; Rowley et al., 2010) and may compromise the subsequent first line therapy that contains the NNRTI drug (Boltz et al., 2011; Ciaranello et al., 2011; Jourdain et al., 2004; Lockman et al., 2007; Moorthy et al., 2011). A study by Chi et al (Chi et al., 2007) showed the time correlation (less than six months) before initiation of ART treatment in NVP exposed individuals as a risk factor that might produce poor treatment outcome. We then assessed the time correlation of NVP exposure with resistance prediction at different prevalence cutoffs. We knew the ART treatment initiation date and the date of PMTCT for the individuals in the study. We then calculated the time of sdNVP exposure before ART initiation for the individuals. We observed that at 15% and lower prevalence cutoff, the prediction of NVP resistance significantly correlates ($p < 0.05$) with time since NVP exposure (**Table 5.3**). According to our finding, for median time of 671 days NVP exposure, the samples were predicted NVP sensitive and for median time of 174 days NVP exposure, the

samples were predicted NVP resistant. Our finding is supported by various earlier studies. Coovadia et al (Coovadia et al., 2009) studied the effect of sdNVP exposure to the virologic response to NVP based first line ART and observed that women who received sdNVP 18-36 months prior to NVP based first line ART initiation had likelihood of sustained virologic suppression. A study by Stringer et al (Stringer et al., 2010b) provides more support for our observation on correlation between time and resistance prediction. The authors studied the prevalence of VF after NVP containing first line therapy in the patients experienced with single dose NVP. They also observed that the time elapse between NVP exposure and initiation of NVP containing ART therapy were correlated. The authors observed VF in 47 of 116 (40%) of women with six or less months of time elapse, 25 of 67 (37%) of women with seven to 12 months of time elapse and 42 of 172 (24%) of women with more than 12 months of time elapse between NVP exposure and initiation of NVP containing ART therapy. Their observation showed that as the time elapsed increased, the rate of VF in NVP containing ART was declined. The authors concluded that risk of VF in recent drug exposed patients was high and suggested that NVP should not be included in the subsequent first line therapy for the drug exposed patients before 12 months of the therapy. The authors' conclusion was highly consistent with our observation of time elapse and resistant prediction.

Thus, over the time of sdNVP exposure in PMTCT therapy, the prevalence of the drug resistant variants decreases in the viral population (Eshleman et al., 2004b) until eventually they are no longer present. When the selection pressure of the drug NVP is removed, the resistant variants get less fit in the viral quasispecies (Mammano et al., 2000; Quinones-Mateu and Arts, 2002) and the sensitive wild type viruses reemerge to dominate in the viral population. Therefore, we hypothesize that long term

virologic suppress can be achieved with the time exposure to the sdNVP drug over six months before initiating first line ART.

5.4.4 First line therapy failure correlates to historical antiretroviral drug use

Pregnant mothers who were administered sdNVP as a HIV prophylaxis to prevent the viral transmission from mother to child (Guay et al., 1999; Jackson et al., 2003) harbored NVP resistant mutations after the therapy (Eshleman et al., 2004a; Eshleman et al., 2004b; Eshleman et al., 2001; Johnson et al., 2005). Studies have showed that the presence of the drug resistant viral variants could relate to the poor clinical outcomes in the first line ART (Casado et al., 2000; Chi et al., 2007; Johnson et al., 2008; Lecossier et al., 2005; Metzner et al., 2009). We studied the resistance in first line antiretroviral treatment failure samples using UDPS genotypic data.

Our finding showed that the frequency of samples with predicted resistance in VF1 PMTCT samples was significantly higher than VF1 no-PMTCT samples (93.33% vs. 63.88%) using FLX platform. The finding was consistent from prevalence cutoff 20% to 1%. Lockman et al (Lockman et al., 2007) had also studied the response of NVP based first line ART on the samples exposed to the drug through PMTCT therapy. They observed that 5% women that received placebo and 18.4% women that received sdNVP experienced first line VF in the first six month of the initiation of the ART treatment. They defined NVP resistance as the presence of any high level NVP resistant mutations: 100I, 103N, 106A/M, 108I, 181C/I, 188L/C/H, or 190A. Genotypic data from 16 of 20 (80%) mothers who received sdNVP and experienced VF within six months of first line ART showed that 12 had NVP resistance at the time

of failure, one had baseline NVP resistance but no NVP resistance was detected at the time of VF and three had no detectable NVP resistance before ART and at the time of VF. Genotyping using conventional Sanger method could the reason of undetectable NVP resistance in the some VF mother.

In more support to our finding, Geretti et al (Geretti et al., 2009) also showed detection of resistance before pretreatment was highly associated with virologic failure of drug regiment. The authors studied 93 HIV infected patients diagnosed in the median year 2000, received drug cocktail after 1 median year and had median CD4 count of 218. Before the highly active antiretroviral treatment, they were under nevirapine or efaverinz treatment. In the pretreatment samples, the common mutation found was K103N that provide high-level resistance to NVP and EFV. Using population based sequencing, four of 18 virologic failure samples harbored K103N mutation while none of 75 virological success samples had the mutation. Using sensitive genotyping method, the association of resistance and VF was increased. Seven of 18 VF sample had the mutation while none of 75 virologic success samples had the mutation.

Chapter 6

Final Summary

In countries with a high burden of HIV, drug resistance testing is not routinely performed in a clinical setting due to the cost. The advent of Roche/454, and other, pyrosequencing platforms holds great promise in the development of a high-throughput, robust, reliable and affordable HIV drug resistance test. Further, the depth of coverage obtained using ultra deep pyrosequencing means that the presence of clinically relevant low abundance drug resistant viral variants within an individual's viral population can also be explored.

The sheer volume of data generated by such an approach means that there is a need for powerful, sensitive and user-friendly bioinformatics applications for management and analysis of the data. We have developed a bioinformatics pipeline (Seq2Res) that takes sequence data directly from the Roche/454 sequencing platform and outputs drug resistance information for each sample.

As part of this pipeline we have developed a novel approach (QTrim) (Shrestha et al., 2014) for quality trimming of the Roche/454 sequence reads. UDPS data outputs a quality score associated with every nucleotide and these must be accounted for prior to downstream analysis of the data, to ensure accurate drug resistance results and avoid false positives and false negatives. We compared the performance of QTrim (Shrestha et al., 2014) with other widely used algorithms on both good and poor quality Roche/454 sequence data. We evaluated the methods based on mean read length, total number of reads and the percentage of poor quality bases in the resulting

output. We found that our approach performed marginally better than the next best method for good quality data and significantly outperformed all methods when poor quality data is analyzed.

UDPS reads often contain PCR and sequencing induced errors which can artificially inflate viral diversity and may result in the generation of false positives in HIV resistance genotyping. These errors can potentially be corrected if all of the sequence reads originating from the same viral template. Recent studies have described the use of a primer ID approach that tags each viral template with a unique identifier during the cDNA generation step (Jabara et al., 2011). Thus sequence reads deriving from the same viral template sequence can be identified and the generation of a consensus sequence from all of the reads originating from the same template can accurately reduce the presence of PCR and sequencing induced errors in the viral genotyping data. The analysis of this data is complex with no know algorithms currently published for analysis of such data. We have developed a module for the Seq2Res resistance testing platform that is capable of processing sample specific sequence data produced using the Primer ID approach. Application of this approach to primer ID data showed that while the primer ID approach can potentially reduce the presence of errors, it also has the potential for under representing viral diversity by removing viral sequences from subsequent analysis because there are insufficient reads for a template to generate a consensus sequence.

We evaluated and verified the sensitivity of the resistance testing module of Seq2Res at accurately identifying drug resistant mutations using the two real consensus sequence datasets downloaded from Stanford resistance database. We found that Seq2Res is at least as sensitive as the Stanford database as it identified all the drug

resistant mutations that were reported by the Web Sierra. In addition, Seq2Res was also able to identify a drug resistant mutation that was reported by Web Sierra incorrectly.

We also tested the ability of Seq2Res to report the prevalence of the drug resistant mutations correctly using five simulated datasets and showed that regardless of the prevalence level of the DRMs in the dataset, Seq2Res is capable of accurately identifying their presence at the correct prevalence level.

Finally, we applied Seq2Res for drug resistance genotyping on real biological datasets generated as part of the CIPRA-SA (Comprehensive International Program for Research in AIDS in South Africa) study. 471 samples were genotyped using Roche/454 Junior and 630 samples were genotyped using Roche/454 FLX. Both datasets were evaluated using the Seq2Res resistance testing platform. Consensus sequences for 349 samples from conventional genotyping were also available and were used for comparison with the UDPS data.

Roche/454 FLX and Roche/454 Junior genotypic data for baseline samples showed that the number of samples that were predicted as resistant increased when the prevalence of resistance was decreased to 1% cutoff. There was no significant difference in the number of samples predicted as resistant, showing that FLX and Junior were comparable at HIV drug resistance genotyping. We further compared the sensitivity of HIV drug resistance genotyping of both UDPS and population based Sanger genotyping at prevalence cutoff 20%, which showed that UDPS and population based Sanger genotyping were comparable. This showed that UDPS is more sensitive than population bases Sanger method.

In our analysis, the number of baseline samples predicted as resistant, using UDPS, was higher in the PMTCT group than in the no-PMTCT group at the prevalence cutoff 15% and below. It showed that NVP resistance was more likely to develop in NVP exposed patients in PMTCT group than in the drug naïve group. Further, we analyzed the correlation of time since NVP exposure and virologic clinical outcome to first line ART. The time (in days) between sdNVP exposure and ART initiation was used and these results showed that the presence of NVP resistance in an individual's viral population significantly correlates ($p < 0.05$) with time since NVP exposure. The samples from patients that receive sdNVP at median time of 174 days before ART initiation were identified as having resistant variants present in their viral population while the samples from patients that receive sdNPV at median time of 631 days were predicted as susceptible to NVP. Our finding correlates with previous studies that showed that ART initiation before six months of NVP exposure could be a risk factor for clinical poor outcome (Chi et al., 2007). Coovadia et al (Coovadia et al., 2009) observed that NVP based ART initiation after 18-36 months of NVP exposure had sustained likelihood of virologic suppression. Similarly, Stringer et al (Stringer et al., 2010b) observed as the time since NVP exposure increased, the rate of VF in NVP containing ART was declined.

Roche/454 FLX and Roche/454 Junior genotypic data from first line virologic failure (VF1) samples were studied for resistance. Seq2Res reported that up to 100% of the VF1 samples from PMTCT and ~65% of the VF1 samples from no-PMTCT were resistant. It showed that drug exposed individuals are more likely to experience first line VF than the drug naive individuals.

In conclusion, UDPS is capable of massive parallel pyrosequencing to explore an unprecedented range of HIV variants, including the clinically relevant low abundant drug resistant variants missed by conventional population based Sanger genotyping method. In addition, UDPS is much cheaper per sample meaning that it can be applied in a cost-effective manner for routine resistance genotyping in resource-limited settings with a high burden of IV infections such as sub-Saharan African.

Bibliography

Abbate, I, Vlassi, C, Rozera, G, Bruselles, A, Bartolini, B, Giombini, E, Corpolongo, A, D'Offizi, G, Narciso, P, Desideri, A, Ippolito, G, Capobianchi, MR (2011) Detection of quasispecies variants predicted to use CXCR4 by ultra-deep pyrosequencing during early HIV infection. *AIDS* **25**: 611-617.

Aberg, JA, Kaplan, JE, Libman, H, Emmanuel, P, Anderson, JR, Stone, VE, Oleske, JM, Currier, JS, Gallant, JE (2009) Primary care guidelines for the management of persons infected with human immunodeficiency virus: 2009 update by the HIV medicine Association of the Infectious Diseases Society of America. *Clin Infect Dis* **49**: 651-681.

Adetunji, AA, Achenbach, C, Feinglass, J, Darin, KM, Scarsi, KK, Ekong, E, Taiwo, BO, Adewole, IF, Murphy, R (2012) Optimizing Treatment Switch for Virologic Failure during First-Line Antiretroviral Therapy in Resource-Limited Settings. *J Int Assoc Provid AIDS Care* **12**: 236-240.

Aghokeng, AF, Kouanfack, C, Laurent, C, Ebong, E, Atem-Tambe, A, Butel, C, Montavon, C, Mpoudi-Ngole, E, Delaporte, E, Peeters, M (2011) Scale-up of antiretroviral treatment in sub-Saharan Africa is accompanied by increasing HIV-1 drug resistance mutations in drug-naive patients. *AIDS* **25**: 2183-2188.

Ammaranond, P, Sanguansittianan, S (2012) Mechanism of HIV antiretroviral drugs progress toward drug resistance. *Fundamental & Clinical Pharmacology* **26**: 146–161.

Antonelli, G (2013) Emerging new technologies in clinical virology. *Clin Microbiol Infect* **19**: 8-9.

Apetrei, C, Robertson, DL, Marx, PA (2004) The history of SIVS and AIDS: epidemiology, phylogeny and biology of isolates from naturally SIV infected non-human primates (NHP) in Africa. *Frontiers in bioscience: a journal and virtual library* **9**: 225-254.

Archer, J, Braverman, MS, Taillon, BE, Desany, B, James, I, Harrigan, PR, Lewis, M, Robertson, DL (2009) Detection of low-frequency pretherapy chemokine (CXC motif) receptor 4 (CXCR4)-using HIV-1 with ultra-deep pyrosequencing. *AIDS* **23**: 1209-1218.

Archer, J, Pinney, JW, Fan, J, Simon-Loriere, E, Arts, EJ, Negroni, M, Robertson, DL (2008) Identifying the important HIV-1 recombination breakpoints. *PLoS computational biology* **4**: e1000178.

Ariën, KK, Abraha, A, Quiñones-Mateu, ME, Kestens, L, Vanham, G, Arts, EJ (2005) The Replicative Fitness of Primary Human Immunodeficiency Virus Type 1 (HIV-1) Group M, HIV-1 Group O, and HIV-2 Isolates. *Journal of Virology* **79**: 8979-8990.

Arrive, E, Newell, ML, Ekouevi, DK, Chaix, ML, Thiebaut, R, Masquelier, B, Leroy, V, Perre, PV, Rouzioux, C, Dabis, F (2007) Prevalence of resistance to nevirapine in mothers and children after single-dose exposure to prevent vertical transmission of HIV-1: a meta-analysis. *Int J Epidemiol* **36**: 1009-1021.

Artenstein, AW, Coppola, J, Brown, AE, Carr, JK, Sanders-Buell, E, Galbarini, E, Mascola, JR, VanCott, TC, Schonbrood, P, McCutchan, FE, et al. (1995) Multiple introductions of HIV-1 subtype E into the western hemisphere. *Lancet* **346**: 1197-1198.

Asamoah-Odei, E, Garcia Calleja, JM, Boerma, JT (2004) HIV prevalence and trends in sub-Saharan Africa: no decline and large subregional differences. *Lancet* **364**: 35-40.
Audureau, E, Kahn, JG, Besson, MH, Saba, J, Ladner, J (2013) Scaling up prevention of mother-to-child HIV transmission programs in sub-Saharan African countries: a multilevel assessment of site-, program- and country-level determinants of performance. *BMC Public Health* **13**: 286.
Autran, B, Carcelain, G, Li, TS, Blanc, C, Mathez, D, Tubiana, R, Katlama, C, Debre, P, Leibowitch, J (1997) Positive effects of combined antiretroviral therapy on CD4+ T cell homeostasis and function in advanced HIV disease. *Science* **277**: 112-116.
Ayouba, A, Souquieres, S, Njinku, B, Martin, PM, Muller-Trutwin, MC, Roques, P, Barre-Sinoussi, F, Mauclere, P, Simon, F, Nerrienet, E (2000) HIV-1 group N among HIV-1-seropositive individuals in Cameroon. *AIDS* **14**: 2623-2625.
Bailes, E, Chaudhuri, RR, Santiago, ML, Bibollet-Ruche, F, Hahn, BH, Sharp, PM (2002) The Evolution of Primate Lentiviruses and the Origins of AIDS. In: The Molecular Epidemiology of Human Viruses, Springer US, pp. 65-96.
Balduin M, KD (2011) Low-frequency hiv-1 drug resistance mutations and risk of nnrti-based antiretroviral treatment failure: A systematic review and pooled analysis. *JAMA* **305**: 1327-1335.
Baldwin, CE, Sanders, RW, Berkhout, B (2003) Inhibiting HIV-1 entry with fusion inhibitors. *Curr Med Chem* **10**: 1633-1642.
Baltimore, D (1971) Expression of animal virus genomes. *Bacteriological Reviews* **35**: 235.
Balzarini, J (2004) Current Status of the Non-nucleoside Reverse Transcriptase Inhibitors of Human Immunodeficiency Virus Type 1. *Current Topics in Medicinal Chemistry* **4**: 921-944.
Bangsberg, DR, Charlebois, ED, Grant, RM, Holodniy, M, Deeks, SG, Perry, S, Conroy, KN, Clark, R, Guzman, D, Zolopa, A (2003) High levels of adherence do not prevent accumulation of HIV drug resistance mutations. *Aids* **17**: 1925.
Bansode, V, McCormack, GP, Crampin, AC, Ngwira, B, Shrestha, RK, French, N, Glynn, JR, Travers, SA (2013) Characterizing the emergence and persistence of drug resistant mutations in HIV-1 subtype C infections using 454 ultra deep pyrosequencing. *BMC Infect Dis* **13**: 52.
Barnes, WM (1992) The fidelity of Taq polymerase catalyzing PCR is improved by an N-terminal deletion. *Gene* **112**: 29-35.
Baur, AS, Sawai, ET, Dazin, P, Fantl, WJ, Cheng-Mayer, C, Peterlin, BM (1994) HIV-1 Nef leads to inhibition or activation of T cells depending on its intracellular localization. *Immunity* **1**: 373-384.
Baxter, JD, Mayers, DL, Wentworth, DN, Neaton, JD, Hoover, ML, Winters, MA, Mannheimer, SB, Thompson, MA, Abrams, DI, Brizz, BJ, Ioannidis, JP, Merigan, TC (2000) A randomized study of antiretroviral management based on plasma genotypic antiretroviral resistance testing in patients failing therapy. CPCRA 046 Study Team for the Terry Beirn Community Programs for Clinical Research on AIDS. *AIDS* **14**: F83-93.
Bebenek, K, Abbotts, J, Roberts, JD, Wilson, SH, Kunkel, TA (1989) Specificity and mechanism of error-prone replication by human immunodeficiency virus-1 reverse transcriptase. *J Biol Chem* **264**: 16948-16956.
Bebenek, K, Abbotts, J, Wilson, SH, Kunkel, TA (1993) Error-prone polymerization by HIV-1 reverse transcriptase. Contribution of template-primer misalignment, miscoding, and termination probability to mutational hot spots. *J Biol Chem* **268**: 10324-10334.

Beck, EJ, Vitoria, M, Mandalia, S, Crowley, S, Gilks, CF, Souteyrand, Y (2006) National adult antiretroviral therapy guidelines in resource-limited countries: concordance with 2003 WHO guidelines? *AIDS* **20**: 1497-1502.
Beerenwinkel, N, Gunthard, HF, Roth, V, Metzner, KJ (2012) Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. *Front Microbiol* **3**: 329.
Beerenwinkel, N, Zagordi, O (2011) Ultra-deep sequencing for the analysis of viral populations. *Curr Opin Virol* **1**: 413-418.
Ben-Artzi, H, Shemesh, J, Zeelon, E, Amit, B, Kleiman, L, Gorecki, M, Panet, A (1996) Molecular analysis of the second template switch during reverse transcription of the HIV RNA template. *Biochemistry* **35**: 10549-10557.
Bennett, DE, Camacho, RJ, Otelea, D, Kuritzkes, DR, Fleury, H, Kiuchi, M, Heneine, W, Kantor, R, Jordan, MR, Schapiro, JM, Vandamme, AM, Sandstrom, P, Boucher, CA, van de Vijver, D, Rhee, SY, Liu, TF, Pillay, D, Shafer, RW (2009) Drug resistance mutations for surveillance of transmitted HIV-1 drug-resistance: 2009 update. *PLoS One* **4**: e4724.
Bentley, DR, Balasubramanian, S, Swerdlow, HP, Smith, GP, Milton, J, Brown, CG, Hall, KP, Evers, DJ, Barnes, CL, Bignell, HR, Boutell, JM, Bryant, J, Carter, RJ, Keira Cheetham, R, Cox, AJ, Ellis, DJ, Flatbush, MR, Gormley, NA, Humphray, SJ, Irving, LJ, Karbelashvili, MS, Kirk, SM, Li, H, Liu, X, Maisinger, KS, Murray, LJ, Obradovic, B, Ost, T, Parkinson, ML, Pratt, MR, Rasolonjatovo, IM, Reed, MT, Rigatti, R, Rodighiero, C, Ross, MT, Sabot, A, Sankar, SV, Scally, A, Schroth, GP, Smith, ME, Smith, VP, Spiridou, A, Torrance, PE, Tzonev, SS, Vermaas, EH, Walter, K, Wu, X, Zhang, L, Alam, MD, Anastasi, C, Aniebo, IC, Bailey, DM, Bancarz, IR, Banerjee, S, Barbour, SG, Baybayan, PA, Benoit, VA, Benson, KF, Bevis, C, Black, PJ, Boodhun, A, Brennan, JS, Bridgham, JA, Brown, RC, Brown, AA, Buermann, DH, Bundu, AA, Burrows, JC, Carter, NP, Castillo, N, Chiara, ECM, Chang, S, Neil Cooley, R, Crake, NR, Dada, OO, Diakoumakos, KD, Dominguez-Fernandez, B, Earnshaw, DJ, Egbujor, UC, Elmore, DW, Etchin, SS, Ewan, MR, Fedurco, M, Fraser, LJ, Fuentes Fajardo, KV, Scott Furey, W, George, D, Gietzen, KJ, Goddard, CP, Golda, GS, Granieri, PA, Green, DE, Gustafson, DL, Hansen, NF, Harnish, K, Haudenschild, CD, Heyer, NI, Hims, MM, Ho, JT, Horgan, AM, Hoschler, K, Hurwitz, S, Ivanov, DV, Johnson, MQ, James, T, Huw Jones, TA, Kang, GD, Kerelska, TH, Kersey, AD, Khrebtukova, I, Kindwall, AP, Kingsbury, Z, Kokko-Gonzales, PI, Kumar, A, Laurent, MA, Lawley, CT, Lee, SE, Lee, X, Liao, AK, Loch, JA, Lok, M, Luo, S, Mammen, RM, Martin, JW, McCauley, PG, McNitt, P, Mehta, P, Moon, KW, Mullens, JW, Newington, T, Ning, Z, Ling Ng, B, Novo, SM, O'Neill, MJ, Osborne, MA, Osnowski, A, Ostadan, O, Paraschos, LL, Pickering, L, Pike, AC, Chris Pinkard, D, Pliskin, DP, Podhasky, J, Quijano, VJ, Raczy, C, Rae, VH, Rawlings, SR, Chiva Rodriguez, A, Roe, PM, Rogers, J, Rogert Bacigalupo, MC, Romanov, N, Romieu, A, Roth, RK, Rourke, NJ, Ruediger, ST, Rusman, E, Sanches-Kuiper, RM, Schenker, MR, Seoane, JM, Shaw, RJ, Shiver, MK, Short, SW, Sizto, NL, Sluis, JP, Smith, MA, Ernest Sohna Sohna, J, Spence, EJ, Stevens, K, Sutton, N, Szajkowski, L, Tregidgo, CL, Turcatti, G, Vandevondele, S, Verhovsky, Y, Virk, SM, Wakelin, S, Walcott, GC, Wang, J, Worsley, GJ, Yan, J, Yau, L, Zuerlein, M, Mullikin, JC, Hurles, ME, McCooke, NJ, West, JS, Oaks, FL, Lundberg, PL, Klenerman, D, Durbin, R, Smith, AJ (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**: 53-59.
Berkhout, B, Das, AT, Beerens, N (2001) HIV-1 RNA editing, hypermutation, and error-prone reverse transcription. *Science* **292**: 7.
Bimber, BN, Dudley, DM, Lauck, M, Becker, EA, Chin, EN, Lank, SM, Grunenwald, HL, Caruccio, NC, Maffitt, M, Wilson, NA, Reed, JS, Sosman, JM, Tarosso, LF,

Sanabani, S, Kallas, EG, Hughes, AL, O'Connor, DH (2010) Whole-genome characterization of human and simian immunodeficiency virus intrahost diversity by ultradeep pyrosequencing. *J Virol* **84**: 12087-12092.
Blagoveshchenskaya, AD, Thomas, L, Feliciangeli, SF, Hung, CH, Thomas, G (2002) HIV-1 Nef downregulates MHC-I by a PACS-1- and PI3K-regulated ARF6 endocytic pathway. *Cell* **111**: 853-866.
Blanca, JM, Pascual, L, Ziarsolo, P, Nuez, F, Cañizares, J (2011) ngs_backbone: a pipeline for read cleaning, mapping and SNP calling using Next Generation Sequence. *BMC Genomics* **12**: 285.
Blankenberg, D, Gordon, A, Kuster, GV, Coraor, N, Taylor, J, Nekrutenko, A (2010) Manipulation of FASTQ data with Galaxy. *Bioinformatics* **26**: 1783-1785.
Blower, S, Bodine, E, Kahn, J, McFarland, W (2005) The antiretroviral rollout and drug-resistant HIV in Africa: insights from empirical data and theoretical models. *AIDS* **19**: 1-14.
Boltz, VF, Zheng, Y, Lockman, S, Hong, F, Halvas, EK, McIntyre, J, Currier, JS, Chibowa, MC, Kanyama, C, Nair, A, Owino-Ong'or, W, Hughes, M, Coffin, JM, Mellors, JW (2011) Role of low-frequency HIV-1 variants in failure of nevirapine-containing antiviral therapy in women previously exposed to single-dose nevirapine. *Proc Natl Acad Sci U S A* **108**: 9202-9207.
Booth, CL, Geretti, AM (2007) Prevalence and determinants of transmitted antiretroviral drug resistance in HIV-1 infection. *J Antimicrob Chemother* **59**: 1047-1056.
Borrow, P, Lewicki, H, Wei, X, Horwitz, MS, Peffer, N, Meyers, H, Nelson, JA, Gairin, JE, Hahn, BH, Oldstone, MB, Shaw, GM (1997) Antiviral pressure exerted by HIV-1-specific cytotoxic T lymphocytes (CTLs) during primary infection demonstrated by rapid selection of CTL escape virus. *Nat Med* **3**: 205-211.
Borsetti, A, Ohagen, A, Gottlinger, HG (1998) The C-terminal half of the human immunodeficiency virus type 1 Gag precursor is sufficient for efficient particle assembly. *J Virol* **72**: 9313-9317.
Bourgeois, CF, Kim, YK, Churcher, MJ, West, MJ, Karn, J (2002) Spt5 cooperates with human immunodeficiency virus type 1 Tat by preventing premature RNA release at terminator sequences. *Mol Cell Biol* **22**: 1079-1093.
Boyer, JC, Bebenek, K, Kunkel, TA (1992) Unequal human immunodeficiency virus type 1 reverse transcriptase error rates with RNA and DNA templates. *Proc Natl Acad Sci U S A* **89**: 6919-6923.
Brenner, BG, Oliveira, M, Doualla-Bell, F, Moisi, DD, Ntemgwa, M, Frankel, F, Essex, M, Wainberg, MA (2006) HIV-1 subtype C viruses rapidly develop K65R resistance to tenofovir in cell culture. *AIDS* **20**: F9-13.
Briggs, JA, Krausslich, HG The molecular architecture of HIV. *J Mol Biol* **410**: 491-500.
Briggs, JA, Simon, MN, Gross, I, Krausslich, HG, Fuller, SD, Vogt, VM, Johnson, MC (2004) The stoichiometry of Gag protein in HIV-1. *Nat Struct Mol Biol* **11**: 672-675.
Briz, V, Poveda, E, Soriano, V (2006) HIV entry inhibitors: mechanisms of action and resistance pathways. *Journal of Antimicrobial Chemotherapy* **57**: 619-627.
Brockman, W, Alvarez, P, Young, S, Garber, M, Giannoukos, G, Lee, WL, Russ, C, Lander, ES, Nusbaum, C, Jaffe, DB (2008) Quality scores and SNP detection in sequencing-by-synthesis systems. *Genome Res* **18**: 763-770.
Brodin, J, Mild, M, Hedskog, C, Sherwood, E, Leitner, T, Andersson, B, Albert, J (2013) PCR-induced transitions are the major source of error in cleaned ultra-deep pyrosequencing data. *PLoS One* **8**: e70388.

Bukrinsky, MI, Sharova, N, Dempsey, MP, Stanwick, TL, Bukrinskaya, AG, Haggerty, S, Stevenson, M (1992) Active nuclear import of human immunodeficiency virus type 1 preintegration complexes. *Proceedings of the National Academy of Sciences* **89**: 6580-6584.
Bukrinsky, MI, Sharova, N, McDonald, TL, Pushkarskaya, T, Tarpley, WG, Stevenson, M (1993) Association of integrase, matrix, and reverse transcriptase antigens of human immunodeficiency virus type 1 with viral nucleic acids following acute infection. *Proc Natl Acad Sci U S A* **90**: 6125-6129.
Cameron, W, Japour, AJ, Xu, Y, Hsu, A, Mellors, J, Farthing, C, Cohen, C, Poretz, D, Markowitz, M, Follansbee, S (1999) Ritonavir and saquinavir combination therapy for the treatment of HIV infection. *AIDS* **13**: 213-224.
Carpenter, CC, Cooper, DA, Fischl, MA, Gatell, JM, Gazzard, BG, Hammer, SM, Hirsch, MS, Jacobsen, DM, Katzenstein, DA, Montaner, JS, Richman, DD, Saag, MS, Schechter, M, Schooley, RT, Thompson, MA, Vella, S, Yeni, PG, Volberding, PA (2000) Antiretroviral therapy in adults: updated recommendations of the International AIDS Society-USA Panel. *JAMA* **283**: 381-390.
Carpenter, CC, Fischl, MA, Hammer, SM, Hirsch, MS, Jacobsen, DM, Katzenstein, DA, Montaner, JS, Richman, DD, Saag, MS, Schooley, RT (1997) Antiretroviral therapy for HIV infection in 1997: updated recommendations of the International AIDS Society-USA panel. *JAMA, the journal of the American Medical Association* **277**: 1962-1969.
Carr, A, Miller, J, Law, M, Cooper, DA (2000) A syndrome of lipoatrophy, lactic acidaemia and liver dysfunction associated with HIV nucleoside analogue therapy: contribution to protease inhibitor-related lipodystrophy syndrome. *AIDS* **14**: F25-32.
Carr, A, Samaras, K, Burton, S, Law, M, Freund, J, Chisholm, DJ, Cooper, DA (1998a) A syndrome of peripheral lipodystrophy, hyperlipidaemia and insulin resistance in patients receiving HIV protease inhibitors. *AIDS* **12**: F51-58.
Carr, A, Samaras, K, Chisholm, DJ, Cooper, DA (1998b) Abnormal fat distribution and use of protease inhibitors. *Lancet* **351**: 1736.
Carr, A, Samaras, K, Chisholm, DJ, Cooper, DA (1998c) Pathogenesis of HIV-1-protease inhibitor-associated peripheral lipodystrophy, hyperlipidaemia, and insulin resistance. *Lancet* **351**: 1881-1883.
Chan, DC, Fass, D, Berger, JM, Kim, PS (1997) Core Structure of gp41 from the HIV Envelope Glycoprotein. *Cell* **89**: 263-273.
Chen, X, Tsiang, M, Yu, F, Hung, M, Jones, GS, Zeynalzadegan, A, Qi, X, Jin, H, Kim, CU, Swaminathan, S, Chen, JM (2008) Modeling, Analysis, and Validation of a Novel HIV Integrase Structure Provide Insights into the Binding Modes of Potent Integrase Inhibitors. *Journal of Molecular Biology* **380**: 504-519.
Cherry, S, Doukas, T, Armknecht, S, Whelan, S, Wang, H, Sarnow, P, Perrimon, N (2005) Genome-wide RNAi screen reveals a specific sensitivity of IRES-containing RNA viruses to host translation inhibition. *Genes Dev* **19**: 445-452.
Chi, BH, Sinkala, M, Stringer, EM, Cantrell, RA, Mtonga, V, Bulterys, M, Zulu, I, Kankasa, C, Wilfert, C, Weidle, PJ, Vermund, SH, Stringer, JS (2007) Early clinical and immune response to NNRTI-based antiretroviral therapy among women with prior exposure to single-dose nevirapine. *AIDS* **21**: 957-964.
Chi, BH, Stringer, JS, Moodley, D (2013) Antiretroviral drug regimens to prevent mother-to-child transmission of HIV: a review of scientific, program, and policy advances for sub-Saharan Africa. *Curr HIV/AIDS Rep* **10**: 124-133.
Chou, H-H, Holmes, MH (2001) DNA sequence quality trimming and vector removal. *Bioinformatics* **17**: 1093–1104.

Chou, S, Upton, H, Bao, K, Schulze-Gahmen, U, Samelson, AJ, He, N, Nowak, A, Lu, H, Krogan, NJ, Zhou, Q, Alber, T HIV-1 Tat recruits transcription elongation factors dispersed along a flexible AFF4 scaffold. *Proc Natl Acad Sci U S A* **110**: E123-131.
Christ, F, Thys, W, De Rijck, J, Gijsbers, R, Albanese, A, Arosio, D, Emiliani, S, Rain, JC, Benarous, R, Cereseto, A, Debyser, Z (2008) Transportin-SR2 imports HIV into the nucleus. *Curr Biol* **18**: 1192-1202.
Christopherson, C, Sninsky, J, Kwok, S (1997) The effects of internal primer-template mismatches on RT-PCR: HIV-1 model studies. *Nucleic Acids Res* **25**: 654-658.
Chukkapalli, V, Oh, SJ, Ono, A Opposing mechanisms involving RNA and lipids regulate HIV-1 Gag membrane binding through the highly basic region of the matrix domain. *Proc Natl Acad Sci U S A* **107**: 1600-1605.
Ciaranello, AL, Lockman, S, Freedberg, KA, Hughes, M, Chu, J, Currier, J, Wood, R, Holmes, CB, Pillay, S, Conradie, F, McIntyre, J, Losina, E, Walensky, RP (2011) First-line antiretroviral therapy after single-dose nevirapine exposure in South Africa: a cost-effectiveness analysis of the OCTANE trial. *AIDS* **25**: 479-492.
Clavel, F, Hance, AJ (2004) HIV Drug Resistance. *New England Journal of Medicine* **350**: 1023-1035.
Clumeck, N, Pozniak, A, Raffi, F (2008) European AIDS Clinical Society (EACS) guidelines for the clinical management and treatment of HIV-infected adults. *HIV Med* **9**: 65-71.
Cock, PJ, Antao, T, Chang, JT, Chapman, BA, Cox, CJ, Dalke, A, Friedberg, I, Hamelryck, T, Kauff, F, Wilczynski, B (2009) Biopython:freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**: 1422-1423.
Cock, PJA, Fields, CJ, Goto, N, Heuer, ML, Rice, PM (2010) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research* **38**: 1767-1771.
Cohen, EA, Subbramanian, RA, Gottlinger, HG (1996) Role of auxiliary proteins in retroviral morphogenesis. *Curr Top Microbiol Immunol* **214**: 219-235.
Cohen, EA, Terwilliger, EF, Sodroski, JG, Haseltine, WA (1988) Identification of a protein encoded by the vpu gene of HIV-1. *Nature* **334**: 532-534.
Collier, AC, Coombs, RW, Schoenfeld, DA, Bassett, RL, Timpone, J, Baruch, A, Jones, M, Facey, K, Whitacre, C, McAuliffe, VJ, Friedman, HM, Merigan, TC, Reichman, RC, Hooper, C, Corey, L (1996) Treatment of human immunodeficiency virus infection with saquinavir, zidovudine, and zalcitabine. AIDS Clinical Trials Group. *N Engl J Med* **334**: 1011-1017.
Condra, JH, Schleif, WA, Blahy, OM, Gabryelski, LJ, Graham, DJ, Quintero, JC, Rhodes, A, Robbins, HL, Roth, E, Shivaprakash, M, et al. (1995) In vivo emergence of HIV-1 variants resistant to multiple protease inhibitors. *Nature* **374**: 569-571.
Connor, EM, Sperling, RS, Gelber, R, Kiselev, P, Scott, G, O'Sullivan, MJ, VanDyke, R, Bey, M, Shearer, W, Jacobson, RL (1994a) Reduction of maternal-infant transmission of human immunodeficiency virus type 1 with zidovudine treatment. *New England Journal of Medicine* **331**: 1173-1180.
Connor, EM, Sperling, RS, Gelber, R, Kiselev, P, Scott, G, O'Sullivan, MJ, VanDyke, R, Bey, M, Shearer, W, Jacobson, RL, et al. (1994b) Reduction of maternal-infant transmission of human immunodeficiency virus type 1 with zidovudine treatment. Pediatric AIDS Clinical Trials Group Protocol 076 Study Group. *N Engl J Med* **331**: 1173-1180.
Connor, RI, Chen, BK, Choe, S, Landau, NR (1995) Vpr Is Required for Efficient Replication of Human Immunodeficiency Virus Type-1 in Mononuclear Phagocytes. *Virology* **206**: 935-944.

Coovadia, A, Hunt, G, Abrams, EJ, Sherman, G, Meyers, T, Barry, G, Malan, E, Marais, B, Stehlau, R, Ledwaba, J, Hammer, SM, Morris, L, Kuhn, L (2009) Persistent minority K103N mutations among women exposed to single-dose nevirapine and virologic response to nonnucleoside reverse-transcriptase inhibitor-based therapy. *Clin Infect Dis* **48**: 462-472.
Corbitt, G, Bailey, A, Williams, G (1990) HIV infection in Manchester, 1959. *The Lancet* **336**: 51.
Cullen, BR (1991) Regulation of HIV-1 gene expression. *FASEB J* **5**: 2361-2368.
Daly, TJ, Cook, KS, Gray, GS, Maione, TE, Rusche, JR (1989) Specific binding of HIV-1 recombinant Rev protein to the Rev-responsive element in vitro. *Nature* **342**: 816-819.
Damond, F, Worobey, M, Campa, P, Farfara, I, Colin, G, Matheron, S, Brun-Vézinet, Ft, Robertson, DL, Simon, Ft (2004) Identification of a highly divergent HIV type 2 and proposal for a change in HIV type 2 classification. *AIDS research and human retroviruses* **20**: 666–672.
Darke, PL, Nutt, RF, Brady, SF, Garsky, VM, Ciccarone, TM, Leu, C-T, Lumma, PK, Freidinger, RM, Veber, DF, Sigal, IS (1988) HIV-1 protease specificity of peptide cleavage is sufficient for processing of gag and pol polyproteins. *Biochemical and Biophysical Research Communications* **156**: 297-303.
Davey, NE, Travé, G, Gibson, TJ (2011) How viruses hijack cell regulation. *Trends in Biochemical Sciences* **36**: 159-169.
Davies, JF, Hostomska, Z, Hostomsky, Z, Jordan, Matthews, DA (1991) Crystal structure of the ribonuclease H domain of HIV-1 reverse transcriptase. *Science* **252**: 88-95.
Dawson, L, Yu, X-F (1998) The Role of Nucleocapsid of HIV-1 in Virus Assembly. *Virology* **251**: 141-157.
De Clercq, E (2002) Strategies in the design of antiviral drugs. *Nature Reviews Drug Discovery* **1**: 13-25.
Decroly, E, Vandenbranden, M, Ruysschaert, JM, Cogniaux, J, Jacob, GS, Howard, SC, Marshall, G, Kompelli, A, Basak, A, Jean, F (1994) The convertases furin and PC1 can both cleave the human immunodeficiency virus (HIV)-1 envelope glycoprotein gp160 into gp120 (HIV-1 SU) and gp41 (HIV-I TM). *Journal of Biological Chemistry* **269**: 12240-12247.
Delport, W, Young, JA, Poon, AF, Pond, SLK A bioinformatics pipeline for the analysis and interpretation of HIV-1 ultradeep sequence data.
Delwart, E, Magierowska, M, Royz, M, Foley, B, Peddada, L, Smith, R, Heldebrant, C, Conrad, A, Busch, M (2002) Homogeneous quasispecies in 16 out of 17 individuals during very early HIV-1 primary infection. *AIDS* **16**: 189-195.
Demirov, DG, Orenstein, JM, Freed, EO (2002) The late domain of human immunodeficiency virus type 1 p6 promotes virus release in a cell type-dependent manner. *J Virol* **76**: 105-117.
di Marzo Veronese, F, Reitz, MS, Jr., Gupta, G, Robert-Guroff, M, Boyer-Thompson, C, Louie, A, Gallo, RC, Lusso, P (1993) Loss of a neutralizing epitope by a spontaneous point mutation in the V3 loop of HIV-1 isolated from an infected laboratory worker. *J Biol Chem* **268**: 25894-25901.
Dias-Neto, E, Nunes, DN, Giordano, RJ, Sun, J, Botz, GH, Yang, K, Setubal, JC, Pasqualini, R, Arap, W (2009) Next-generation phage display: integrating and comparing available molecular tools to enable cost-effective high-throughput analysis. *PLoS One* **4**: e8338.
Dismuke, DJ, Aiken, C (2006) Evidence for a functional link between uncoating of the human immunodeficiency virus type 1 core and nuclear import of the viral preintegration complex. *J Virol* **80**: 3712-3720.

Dong, X, Li, H, Derdowski, A, Ding, L, Burnett, A, Chen, X, Peters, TR, Dermody, TS, Woodruff, E, Wang, JJ, Spearman, P (2005) AP-3 directs the intracellular trafficking of HIV-1 Gag and plays a key role in particle assembly. *Cell* **120**: 663-674.
Doualla-Bell, F, Avalos, A, Brenner, B, Gaolathe, T, Mine, M, Gaseitsiwe, S, Oliveira, M, Moisi, D, Ndwapi, N, Moffat, H, Essex, M, Wainberg, MA (2006) High prevalence of the K65R mutation in human immunodeficiency virus type 1 subtype C isolates from infected patients in Botswana treated with didanosine-based regimens. *Antimicrob Agents Chemother* **50**: 4182-4185.
Dudley, DM, Chin, EN, Bimber, BN, Sanabani, SS, Tarosso, LF, Costa, PR, Sauer, MM, Kallas, EG, O'Connor, DH (2012) Low-cost ultra-wide genotyping using Roche/454 pyrosequencing for surveillance of HIV drug resistance. *PLoS One* **7**: e36494.
Duffalo, ML, James, CW (2003) Enfuvirtide: A Novel Agent for the Treatment of HIV-1 Infection. *The Annals of Pharmacotherapy* **37**: 1448-1456.
Dumonceaux, J, Nisole, S, Chanel, C, Quivet, L, Amara, A, Baleux, F, Briand, P, Hazan, U (1998) Spontaneous mutations in the env gene of the human immunodeficiency virus type 1 NDK isolate are associated with a CD4-independent entry phenotype. *J Virol* **72**: 512-519.
Durant, J, Clevenbergh, P, Halfon, P, Delgiudice, P, Porsin, S, Simonet, P, Montagne, N, Boucher, CA, Schapiro, JM, Dellamonica, P (1999) Drug-resistance genotyping in HIV-1 therapy: the VIRADAPT randomised controlled trial. *Lancet* **353**: 2195-2199.
Dwyer, JJ, Hasan, A, Wilson, KL, White, JM, Matthews, TJ, Delmedico, MK (2003) The hydrophobic pocket contributes to the structural stability of the N-terminal coiled coil of HIV gp41 but is not required for six-helix bundle formation. *Biochemistry* **42**: 4945-4953.
Eckert, KA, Kunkel, TA (1991) DNA polymerase fidelity and the polymerase chain reaction. *PCR Methods Appl* **1**: 17-24.
Edgar, RC (2004a) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**: 113.
Edgar, RC (2004b) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**: 1792-1797.
Eid, J, Fehr, A, Gray, J, Luong, K, Lyle, J, Otto, G, Peluso, P, Rank, D, Baybayan, P, Bettman, B, Bibillo, A, Bjornson, K, Chaudhuri, B, Christians, F, Cicero, R, Clark, S, Dalal, R, Dewinter, A, Dixon, J, Foquet, M, Gaertner, A, Hardenbol, P, Heiner, C, Hester, K, Holden, D, Kearns, G, Kong, X, Kuse, R, Lacroix, Y, Lin, S, Lundquist, P, Ma, C, Marks, P, Maxham, M, Murphy, D, Park, I, Pham, T, Phillips, M, Roy, J, Sebra, R, Shen, G, Sorenson, J, Tomaney, A, Travers, K, Trulson, M, Vieceli, J, Wegener, J, Wu, D, Yang, A, Zaccarin, D, Zhao, P, Zhong, F, Korlach, J, Turner, S (2009) Real-time DNA sequencing from single polymerase molecules. *Science* **323**: 133-138.
Eisele, E, Siliciano, RF (2012) Redefining the viral reservoirs that prevent HIV-1 eradication. *Immunity* **37**: 377-388.
Emiliani, S, Mousnier, A, Busschots, K, Maroun, M, Maele, BV, Tempé, D, Vandekerckhove, L, Moisant, F, Ben-Slama, L, Witvrouw, M, Christ, F, Rain, J-C, Dargemont, C, Debyser, Z, Benarous, R (2005) Integrase Mutants Defective for Interaction with LEDGF/p75 Are Impaired in Chromosome Tethering and HIV-1 Replication. *Journal of Biological Chemistry* **280**: 25517-25523.
Eshleman, SH, Guay, LA, Mwatha, A, Brown, ER, Cunningham, SP, Musoke, P, Mmiro, F, Jackson, JB (2004a) Characterization of nevirapine resistance mutations in women with subtype A vs. D HIV-1 6-8 weeks after single-dose nevirapine (HIVNET 012). *J Acquir Immune Defic Syndr* **35**: 126-130.

Eshleman, SH, Guay, LA, Mwatha, A, Cunningham, SP, Brown, ER, Musoke, P, Mmiro, F, Jackson, JB (2004b) Comparison of nevirapine (NVP) resistance in Ugandan women 7 days vs. 6-8 weeks after single-dose nvp prophylaxis: HIVNET 012. *AIDS Res Hum Retroviruses* **20**: 595-599.
Eshleman, SH, Hoover, DR, Chen, S, Hudelson, SE, Guay, LA, Mwatha, A, Fiscus, SA, Mmiro, F, Musoke, P, Jackson, JB, Kumwenda, N, Taha, T (2005) Nevirapine (NVP) resistance in women with HIV-1 subtype C, compared with subtypes A and D, after the administration of single-dose NVP. *J Infect Dis* **192**: 30-36.
Eshleman, SH, Mracna, M, Guay, LA, Deseyve, M, Cunningham, S, Mirochnick, M, Musoke, P, Fleming, T, Glenn Fowler, M, Mofenson, LM, Mmiro, F, Jackson, JB (2001) Selection and fading of resistance mutations in women and infants receiving nevirapine to prevent HIV-1 vertical transmission (HIVNET 012). *AIDS* **15**: 1951-1957.
Esnouf, R, Ren, J, Ross, C, Jones, Y, Stammers, D, Stuart, D (1995) Mechanism of inhibition of HIV-1 reverse transcriptase by non-nucleoside inhibitors. *Nature Structural & Molecular Biology* **2**: 303–308.
Esparza, J, Bhamarapravati, N (2000) Accelerating the development and future availability of HIV-1 vaccines: why, when, where, and how? *Lancet* **355**: 2061-2066.
Ewing, B, Green, P (1998) Base-Calling of Automated Sequencer Traces UsingPhred. II. Error Probabilities. *Genome Research* **8**: 186-194.
Ewing, B, Hillier, L, Wendl, MC, Green, P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* **8**: 175-185.
Fang, CT, Chang, YY, Hsu, HM, Twu, SJ, Chen, KT, Lin, CC, Huang, LY, Chen, MY, Hwang, JS, Wang, JD, Chuang, CY (2007) Life expectancy of patients with newly-diagnosed HIV infection in the era of highly active antiretroviral therapy. *QJM* **100**: 97-105.
Fang, G, Weiser, B, Kuiken, C, Philpott, SM, Rowland-Jones, S, Plummer, F, Kimani, J, Shi, B, Kaul, R, Bwayo, J, Anzala, O, Burger, H (2004) Recombination following superinfection by HIV-1. *AIDS* **18**: 153-159.
Farnet, CM, Haseltine, WA (1991) Determination of viral proteins present in the human immunodeficiency virus type 1 preintegration complex. *J Virol* **65**: 1910-1915.
Fassati, A, Gorlich, D, Harrison, I, Zaytseva, L, Mingot, JM (2003) Nuclear import of HIV-1 intracellular reverse transcription complexes is mediated by importin 7. *EMBO J* **22**: 3675-3685.
Fätkenheuer, G, Pozniak, AL, Johnson, MA, Plettenberg, A, Staszewski, S, Hoepelman, AIM, Saag, MS, Goebel, FD, Rockstroh, JK, Dezube, BJ, Jenkins, TM, Medhurst, C, Sullivan, JF, Ridgway, C, Abel, S, James, IT, Youle, M, van der Ryst, E (2005) Efficacy of short-term monotherapy with maraviroc, a new CCR5 antagonist, in patients infected with HIV-1. *Nature Medicine* **11**: 1170-1172.
Ferradini, L, Jeannin, A, Pinoges, L, Izopet, J, Odhiambo, D, Mankhambo, L, Karungi, G, Szumilin, E, Balandine, S, Fedida, G, Carrieri, MP, Spire, B, Ford, N, Tassie, JM, Guerin, PJ, Brasher, C (2006) Scaling up of highly active antiretroviral therapy in a rural district of Malawi: an effectiveness assessment. *Lancet* **367**: 1335-1342.
Fischer, W, Ganusov, VV, Giorgi, EE, Hraber, PT, Keele, BF, Leitner, T, Han, CS, Gleasner, CD, Green, L, Lo, CC, Nag, A, Wallstrom, TC, Wang, S, McMichael, AJ, Haynes, BF, Hahn, BH, Perelson, AS, Borrow, P, Shaw, GM, Bhattacharya, T, Korber, BT (2010) Transmission of single HIV-1 genomes and dynamics of early immune escape revealed by ultra-deep sequencing. *PLoS One* **5**: e12303.
Fischl, MA, Olson, RM, Follansbee, SE, Lalezari, JP, Henry, DH, Frame, PT, Remick, SC, Salgo, MP, Lin, AH, Nauss-Karol, C, Lieberman, J, Soo, W (1993) Zalcitabine

compared with zidovudine in patients with advanced HIV-1 infection who received previous zidovudine therapy. *Ann Intern Med* **118**: 762-769.
Fischl, MA, Richman, DD, Hansen, N, Collier, AC, Carey, JT, Para, MF, Hardy, WD, Dolin, R, Powderly, WG, Allan, JD, et al. (1990) The safety and efficacy of zidovudine (AZT) in the treatment of subjects with mildly symptomatic human immunodeficiency virus type 1 (HIV) infection. A double-blind, placebo-controlled trial. The AIDS Clinical Trials Group. *Ann Intern Med* **112**: 727-737.
Flys, T, Nissley, DV, Claasen, CW, Jones, D, Shi, C, Guay, LA, Musoke, P, Mmiro, F, Strathern, JN, Jackson, JB, Eshleman, JR, Eshleman, SH (2005) Sensitive drug-resistance assays reveal long-term persistence of HIV-1 variants with the K103N nevirapine (NVP) resistance mutation in some women and infants after the administration of single-dose NVP: HIVNET 012. *J Infect Dis* **192**: 24-29.
Flys, TS, Chen, S, Jones, DC, Hoover, DR, Church, JD, Fiscus, SA, Mwatha, A, Guay, LA, Mmiro, F, Musoke, P, Kumwenda, N, Taha, TE, Jackson, JB, Eshleman, SH (2006) Quantitative analysis of HIV-1 variants with the K103N resistance mutation after single-dose nevirapine in women with HIV-1 subtypes A, C, and D. *J Acquir Immune Defic Syndr* **42**: 610-613.
Fouchier, RA, Groenink, M, Kootstra, NA, Tersmette, M, Huisman, HG, Miedema, F, Schuitemaker, H (1992) Phenotype-associated sequence variation in the third variable domain of the human immunodeficiency virus type 1 gp120 molecule. *J Virol* **66**: 3183-3187.
Francis, DP, Curran, JW, Essex, M (1983) Epidemic acquired immune deficiency syndrome: epidemiologic evidence for a transmissible agent. *Journal of the National Cancer Institute* **71**: 5–9.
Frankel, AD, Young, JAT (1998) HIV-1: Fifteen Proteins and an RNA. *Annual Review of Biochemistry* **67**: 1-25.
Friedman-Kien, AE (1981) Disseminated Kaposi's sarcoma syndrome in young homosexual men. *Journal of the American Academy of Dermatology* **5**: 468–471.
Friedman-Kien, AE, Laubenstein, L, Marmor, M, Hymes, K, Green, J, Ragaz, A, Gottleib, J, Muggia, F, Demopoulos, R, Weintraub, M (1981) Kaposi's sarcoma and Pneumocystis pneumonia among homosexual men—New York City and California. *MMWR* **30**: 305–308.
Furuta, RA, Wild, CT, Weng, Y, Weiss, CD (1998) Capture of an early fusion-active conformation of HIV-1 gp41. *Nature Structural & Molecular Biology* **5**: 276-279.
Gallo, RC, Sarin, PS, Gelmann, EP, Robert-Guroff, M, Richardson, E, Kalyanaraman, VS, Mann, D, Sidhu, GD, Stahl, RE, Zolla-Pazner, S, Leibowitch, J, Popovic, M (1983) Isolation of human T-cell leukemia virus in acquired immune deficiency syndrome (AIDS). *Science (New York, NY)* **220**: 865-867.
Ganser-Pornillos, BK, von Schwedler, UK, Stray, KM, Aiken, C, Sundquist, WI (2004) Assembly properties of the human immunodeficiency virus type 1 CA protein. *J Virol* **78**: 2545-2552.
Gao, F, Bailes, E, Robertson, DL, Chen, Y, Rodenburg, CM, Michael, SF, Cummins, LB, Arthur, LO, Peeters, M, Shaw, GM (1999) Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. *Nature* **397**: 436-441.
Gao, F, Vidal, N, Li, Y, Trask, SA, Chen, Y, Kostrikis, LG, Ho, DD, Kim, J, Oh, M-D, Choe, K, Salminen, M, Robertson, DL, Shaw, GM, Hahn, BH, Peeters, M (2001) Evidence of Two Distinct Subsubtypes within the HIV-1 Subtype A Radiation. *AIDS Research and Human Retroviruses* **17**: 675-688.
Gao, F, Yue, L, Robertson, DL, Hill, SC, Hui, H, Biggar, RJ, Neequaye, AE, Whelan, TM, Ho, DD, Shaw, GM (1994) Genetic diversity of human immunodeficiency virus type 2: evidence for distinct sequence subtypes with differences in virus biology. *Journal of virology* **68**: 7433–7447.

Gao, F, Yue, L, White, AT, Pappas, PG, Barchue, J, Hanson, AP, Greene, BM, Sharp, PM, Shaw, GM, Hahn, BH (1992) Human infection by genetically diverse SIVSM-related HIV-2 in West Africa. *Nature* **358**: 495-499.
Garcia, JV, Miller, AD (1991) Serine phosphorylation-independent downregulation of cell-surface CD4 by nef. *Nature* **350**: 508–511.
Gaynor, R (1992) Cellular transcription factors involved in the regulation of HIV-1 gene expression. *AIDS* **6**: 347-363.
Gazzard, B, Clumeck, N, d'Arminio Monforte, A, Lundgren, JD (2008) Indicator disease-guided testing for HIV--the next step for Europe? *HIV Med* **9 Suppl 2**: 34-40.
Geretti, AM, Fox, ZV, Booth, CL, Smith, CJ, Phillips, AN, Johnson, M, Li, JF, Heneine, W, Johnson, JA (2009) Low-frequency K103N strengthens the impact of transmitted drug resistance on virologic responses to first-line efavirenz or nevirapine-based highly active antiretroviral therapy. *J Acquir Immune Defic Syndr* **52**: 569-573.
Gheysen, D, Jacobs, E, de Foresta, F, Thiriart, C, Francotte, M, Thines, D, De Wilde, M (1989) Assembly and release of HIV-1 precursor Pr55gag virus-like particles from recombinant baculovirus-infected insect cells. *Cell* **59**: 103-112.
Gianella, S, Delport, W, Pacold, ME, Young, JA, Choi, JY, Little, SJ, Richman, DD, Kosakovsky Pond, SL, Smith, DM (2011) Detection of minority resistance during early HIV-1 infection: natural variation and spurious detection rather than transmission and evolution of multiple viral variants. *J Virol* **85**: 8359-8367.
Gibson, RM, Meyer, AM, Winner, D, Archer, J, Feyertag, F, Ruiz-Mateos, E, Leal, M, Robertson, DL, Schmotzer, CL, Quinones-Mateu, ME (2014) Sensitive Deep-Sequencing-Based HIV-1 Genotyping Assay To Simultaneously Determine Susceptibility to Protease, Reverse Transcriptase, Integrase, and Maturation Inhibitors, as Well as HIV-1 Coreceptor Tropism. *Antimicrob Agents Chemother* **58**: 2167-2185.
Gilks, CF, Crowley, S, Ekpini, R, Gove, S, Perriens, J, Souteyrand, Y, Sutherland, D, Vitoria, M, Guerma, T, De Cock, K (2006) The WHO public-health approach to antiretroviral treatment against HIV in resource-limited settings. *Lancet* **368**: 505-510.
Gilles, A, Meglecz, E, Pech, N, Ferreira, S, Malausa, T, Martin, JF (2011) Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. *BMC Genomics* **12**: 245.
Glenn, TC (2011) Field guide to next-generation DNA sequencers. *Mol Ecol Resour* **11**: 759-769.
Golin, CE, Liu, H, Hays, RD, Miller, LG, Beck, CK, Ickovics, J, Kaplan, AH, Wenger, NS (2002) A prospective study of predictors of adherence to combination antiretroviral medication. *J Gen Intern Med* **17**: 756-765.
Goodenow, M, Huet, T, Saurin, W, Kwok, S, Sninsky, J, Wain-Hobson, S (1989) HIV-1 isolates are rapidly evolving quasispecies: evidence for viral mixtures and preferred nucleotide substitutions. *J Acquir Immune Defic Syndr* **2**: 344-352.
Gottlieb, MS, Schroff, R, Schanker, HM, Weisman, JD, Fan, PT, Wolf, RA, Saxon, A (1981) \textitPneumocystis carinii Pneumonia and Mucosal Candidiasis in Previously Healthy Homosexual Men. *New England Journal of Medicine* **305**: 1425-1431.
Göttlinger, HG, Sodroski, JG, Haseltine, WA (1989) Role of capsid precursor processing and myristoylation in morphogenesis and infectivity of human immunodeficiency virus type 1. *Proceedings of the National Academy of Sciences* **86**: 5781-5785.
Gouy, M, Guindon, S, Gascuel, O (2010) SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol* **27**: 221-224.
Grabar, S, Pradier, C, Le Corfec, E, Lancar, R, Allavena, C, Bentata, M, Berlureau, P, Dupont, C, Fabbro-Peray, P, Poizot-Martin, I, Costagliola, D (2000) Factors

associated with clinical and virological failure in patients receiving a triple therapy including a protease inhibitor. *AIDS* **14**: 141-149.
Greenberg, ME, Iafrate, AJ, Skowronski, J (1998) The SH3 domain-binding surface and an acidic motif in HIV-1 Nef regulate trafficking of class I MHC complexes. *EMBO J* **17**: 2777-2789.
Gu, Z, Gao, Q, Faust, EA, Wainberg, MA (1995) Possible involvement of cell fusion and viral recombination in generation of human immunodeficiency virus variants that display dual resistance to AZT and 3TC. *J Gen Virol* **76 (Pt 10)**: 2601-2605.
Guay, LA, Musoke, P, Fleming, T, Bagenda, D, Allen, M, Nakabiito, C, Sherman, J, Bakaki, P, Ducar, C, Deseyve, M, Emel, L, Mirochnick, M, Fowler, MG, Mofenson, L, Miotti, P, Dransfield, K, Bray, D, Mmiro, F, Jackson, JB (1999) Intrapartum and neonatal single-dose nevirapine compared with zidovudine for prevention of mother-to-child transmission of HIV-1 in Kampala, Uganda: HIVNET 012 randomised trial. *Lancet* **354**: 795-802.
Gulick, RM, Mellors, JW, Havlir, D, Eron, JJ, Gonzalez, C, McMahon, D, Jonas, L, Meibohm, A, Holder, D, Schleif, WA, Condra, JH, Emini, EA, Isaacs, R, Chodakewitz, JA, Richman, DD (1998) Simultaneous vs sequential initiation of therapy with indinavir, zidovudine, and lamivudine for HIV-1 infection: 100-week follow-up. *JAMA* **280**: 35-41.
Gulick, RM, Mellors, JW, Havlir, D, Eron, JJ, Gonzalez, C, McMahon, D, Richman, DD, Valentine, FT, Jonas, L, Meibohm, A, Emini, EA, Chodakewitz, JA, Deutsch, P, Holder, D, Schleif, WA, Condra, JH (1997) Treatment with Indinavir, Zidovudine, and Lamivudine in Adults with Human Immunodeficiency Virus Infection and Prior Antiretroviral Therapy. *New England Journal of Medicine* **337**: 734-739.
Gulick, RM, Mellors, JW, Havlir, D, Eron, JJ, Meibohm, A, Condra, JH, Valentine, FT, McMahon, D, Gonzalez, C, Jonas, L (2000) 3-year suppression of HIV viremia with indinavir, zidovudine, and lamivudine. *Annals of internal medicine* **133**: 35-39.
Gunthard, HF, Wong, JK, Ignacio, CC, Havlir, DV, Richman, DD (1998) Comparative performance of high-density oligonucleotide sequencing and dideoxynucleotide sequencing of HIV type 1 pol from clinical samples. *AIDS Res Hum Retroviruses* **14**: 869-876.
Gupta, R, Hill, A, Sawyer, AW, Pillay, D (2008) Emergence of drug resistance in HIV type 1-infected patients after receipt of first-line highly active antiretroviral therapy: a systematic review of clinical trials. *Clin Infect Dis* **47**: 712-722.
Gupta, RK, Hill, A, Sawyer, AW, Cozzi-Lepri, A, von Wyl, V, Yerly, S, Lima, VD, Gunthard, HF, Gilks, C, Pillay, D (2009) Virological monitoring and resistance to first-line highly active antiretroviral therapy in adults infected with HIV-1 treated under WHO guidelines: a systematic review and meta-analysis. *Lancet Infect Dis* **9**: 409-417.
Gupta, RK, Jordan, MR, Sultan, BJ, Hill, A, Davis, DH, Gregson, J, Sawyer, AW, Hamers, RL, Ndembi, N, Pillay, D, Bertagnolio, S (2012) Global trends in antiretroviral resistance in treatment-naive individuals with HIV after rollout of antiretroviral treatment in resource-limited settings: a global collaborative study and meta-regression analysis. *Lancet* **380**: 1250-1258.
Gürtler, L (2004) [Zoonotic infections stimulation]. *Bundesgesundheitsblatt, Gesundheitsforschung, Gesundheitsschutz* **47**: 609-610.
Haase, AT Targeting early infection to prevent HIV-1 mucosal transmission. *Nature* **464**: 217-223.
Hahn, BH, Shaw, GM, De, KM, Sharp, PM (2000) AIDS as a zoonosis: scientific and public health implications. *Science* **287**: 607–614.

Hamady, M, Walker, JJ, Harris, JK, Gold, NJ, Knight, R (2008) Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat Methods* **5**: 235-237.
Hamers, RL, Schuurman, R, Sigaloff, KC, Wallis, CL, Kityo, C, Siwale, M, Mandaliya, K, Ive, P, Botes, ME, Wellington, M, Osibogun, A, Wit, FW, van Vugt, M, Stevens, WS, de Wit, TF (2012) Effect of pretreatment HIV-1 drug resistance on immunological, virological, and drug-resistance outcomes of first-line antiretroviral treatment in sub-Saharan Africa: a multicentre cohort study. *Lancet Infect Dis* **12**: 307-317.
Hamers, RL, Wallis, CL, Kityo, C, Siwale, M, Mandaliya, K, Conradie, F, Botes, ME, Wellington, M, Osibogun, A, Sigaloff, KC, Nankya, I, Schuurman, R, Wit, FW, Stevens, WS, van Vugt, M, de Wit, TF (2011) HIV-1 drug resistance in antiretroviral-naive individuals in sub-Saharan Africa after rollout of antiretroviral therapy: a multicentre observational study. *Lancet Infect Dis* **11**: 750-759.
Hammer, SM, Eron, JJ, Jr., Reiss, P, Schooley, RT, Thompson, MA, Walmsley, S, Cahn, P, Fischl, MA, Gatell, JM, Hirsch, MS, Jacobsen, DM, Montaner, JS, Richman, DD, Yeni, PG, Volberding, PA (2008) Antiretroviral treatment of adult HIV infection: 2008 recommendations of the International AIDS Society-USA panel. *JAMA* **300**: 555-570.
Hammer, SM, Katzenstein, DA, Hughes, MD, Gundacker, H, Schooley, RT, Haubrich, RH, Henry, WK, Lederman, MM, Phair, JP, Niu, M, Hirsch, MS, Merigan, TC (1996) A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. AIDS Clinical Trials Group Study 175 Study Team. *N Engl J Med* **335**: 1081-1090.
Hammer, SM, Squires, KE, Hughes, MD, Grimes, JM, Demeter, LM, Currier, JS, Eron, JJ, Feinberg, JE, Balfour, HH, Deyton, LR, Chodakewitz, JA, Fischl, MA, Phair, JP, Pedneault, L, Nguyen, B-Y, Cook, JC (1997) A Controlled Trial of Two Nucleoside Analogues plus Indinavir in Persons with Human Immunodeficiency Virus Infection and CD4 Cell Counts of 200 per Cubic Millimeter or Less. *New England Journal of Medicine* **337**: 725-733.
Hanna, GJ, D'Aquila, RT (2001) Clinical use of genotypic and phenotypic drug resistance testing to monitor antiretroviral chemotherapy. *Clin Infect Dis* **32**: 774-782.
Harrison, KM, Song, R, Zhang, X (2010) Life expectancy after HIV diagnosis based on national HIV surveillance data from 25 states, United States. *J Acquir Immune Defic Syndr* **53**: 124-130.
Hauser, A, Mugenyi, K, Kabasinguzi, R, Kuecherer, C, Harms, G, Kunz, A (2011) Emergence and persistence of minor drug-resistant HIV-1 variants in Ugandan women after nevirapine single-dose prophylaxis. *PLoS One* **6**: e20357.
Havlir, DV, Eastman, S, Gamst, A, Richman, DD (1996) Nevirapine-resistant human immunodeficiency virus: kinetics of replication and estimated prevalence in untreated patients. *J Virol* **70**: 7894-7899.
Hazuda, DJ, Anthony, NJ, Gomez, RP, Jolly, SM, Wai, JS, Zhuang, L, Fisher, TE, Embrey, M, Guare, JP, Egbertson, MS, Vacca, JP, Huff, JR, Felock, PJ, Witmer, MV, Stillmock, KA, Danovich, R, Grobler, J, Miller, MD, Espeseth, AS, Jin, L, Chen, I-W, Lin, JH, Kassahun, K, Ellis, JD, Wong, BK, Xu, W, Pearson, PG, Schleif, WA, Cortese, R, Emini, E, Summa, V, Holloway, MK, Young, SD (2004) A naphthyridine carboxamide provides evidence for discordant resistance between mechanistically identical inhibitors of HIV-1 integrase. *Proceedings of the National Academy of Sciences of the United States of America* **101**: 11233-11238.
He, J, Choe, S, Walker, R, Marzio, PD, Morgan, DO, Landau, NR (1995) Human immunodeficiency virus type 1 viral protein R (Vpr) arrests cells in the G2 phase of the cell cycle by inhibiting p34cdc2 activity. *Journal of Virology* **69**: 6705-6711.

He, N, Zhou, Q New insights into the control of HIV-1 transcription: when Tat meets the 7SK snRNP and super elongation complex (SEC). *J Neuroimmune Pharmacol* **6**: 260-268.
Hedskog, C, Mild, M, Jernberg, J, Sherwood, E, Bratt, G, Leitner, T, Lundeberg, J, Andersson, B, Albert, J (2010) Dynamics of HIV-1 quasispecies during antiviral treatment dissected using ultra-deep pyrosequencing. *PLoS One* **5**: e11345.
Hemelaar, J, Gouws, E, Ghys, PD, Osmanov, S (2006) Global and regional distribution of HIV-1 genetic subtypes and recombinants in 2004. *AIDS* **20**: W13-W23.
Herbst, AJ, Cooke, GS, Barnighausen, T, KanyKany, A, Tanser, F, Newell, ML (2009) Adult mortality and antiretroviral treatment roll-out in rural KwaZulu-Natal, South Africa. *Bull World Health Organ* **87**: 754-762.
Hertogs, K, de Bethune, MP, Miller, V, Ivens, T, Schel, P, Van Cauwenberge, A, Van Den Eynde, C, Van Gerwen, V, Azijn, H, Van Houtte, M, Peeters, F, Staszewski, S, Conant, M, Bloor, S, Kemp, S, Larder, B, Pauwels, R (1998) A rapid method for simultaneous detection of phenotypic resistance to inhibitors of protease and reverse transcriptase in recombinant human immunodeficiency virus type 1 isolates from patients treated with antiretroviral drugs. *Antimicrob Agents Chemother* **42**: 269-276.
Hiatt, JB, Pritchard, CC, Salipante, SJ, O'Roak, BJ, Shendure, J (2013) Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. *Genome Res* **23**: 843-854.
Himmel, DM, Sarafianos, SG, Dharmasena, S, Hossain, MM, McCoy-Simandle, K, Ilina, T, Clark, AD, Jr., Knight, JL, Julias, JG, Clark, PK, Krogh-Jespersen, K, Levy, RM, Hughes, SH, Parniak, MA, Arnold, E (2006) HIV-1 reverse transcriptase structure with RNase H inhibitor dihydroxy benzoyl naphthyl hydrazone bound at a novel site. *ACS Chem Biol* **1**: 702-712.
Hirsch MS, B-VF (2000) Antiretroviral drug resistance testing in adult hiv-1 infection: Recommendations of an international aids society–usa panel. *JAMA* **283**: 2417-2426.
Hirsch, VM, Olmsted, RA, Murphey-Corb, M, Purcell, RH, Johnson, PR (1989) An African primate lentivirus (SIVsmclosely related to HIV-2.
Ho, DD, Bieniasz, PD (2008) HIV-1 at 25. *Cell* **133**: 561-565.
Ho, DD, Neumann, AU, Perelson, AS, Chen, W, Leonard, JM, Markowitz, M (1995) Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. *Nature* **373**: 123-126.
Hoffmann, C, Minkah, N, Leipzig, J, Wang, G, Arens, MQ, Tebas, P, Bushman, FD (2007) DNA bar coding and pyrosequencing to identify rare HIV drug resistance mutations. *Nucleic Acids Res* **35**: e91.
Horton, RM (1995) PCR-mediated recombination and mutagenesis. SOEing together tailor-made genes. *Mol Biotechnol* **3**: 93-99.
Hosseinipour, MC, van Oosterhout, JJ, Weigel, R, Phiri, S, Kamwendo, D, Parkin, N, Fiscus, SA, Nelson, JA, Eron, JJ, Kumwenda, J (2009) The public health approach to identify antiretroviral therapy failure: high-level nucleoside reverse transcriptase inhibitor resistance among Malawians failing first-line antiretroviral therapy. *AIDS* **23**: 1127-1134.
Huang, C-c, Lam, SN, Acharya, P, Tang, M, Xiang, S-H, Hussan, SS-u, Stanfield, RL, Robinson, J, Sodroski, J, Wilson, IA, Wyatt, R, Bewley, CA, Kwong, PD (2007) Structures of the CCR5 N Terminus and of a Tyrosine-Sulfated Antibody with HIV-1 gp120 and CD4. *Science* **317**: 1930-1934.
Huang, W, Li, L, Myers, JR, Marth, GT (2012) ART: a next-generation sequencing read simulator. *Bioinformatics* **28**: 593-594.

Hudelson, SE, McConnell, MS, Bagenda, D, Piwowar-Manning, E, Parsons, TL, Nolan, ML, Bakaki, PM, Thigpen, MC, Mubiru, M, Fowler, MG, Eshleman, SH (2010) Emergence and persistence of nevirapine resistance in breast milk after single-dose nevirapine administration. *AIDS* **24**: 557-561.
Huet, T, Cheynier, R, Meyerhans, A, Roelants, G, Wain-Hobson, S (1990) Genetic organization of a chimpanzee lentivirus related to HIV-1.
Hughes, JP, Totten, P (2003) Estimating the accuracy of polymerase chain reaction-based tests using endpoint dilution. *Biometrics* **59**: 505-511.
Hulme, AE, Perez, O, Hope, TJ (2011) Complementary assays reveal a relationship between HIV-1 uncoating and reverse transcription. *Proc Natl Acad Sci U S A* **108**: 9975-9980.
Hunter, JD (2007) Matplotlib: A 2D graphics environment. *Computing In Science and Engineering* **9**: 90-95.
Huse, SM, Huber, JA, Morrison, HG, Sogin, ML, Welch, DM (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome biol* **8**: R143.
Hussain, A, Wesley, C, Khalid, M, Chaudhry, A, Jameel, S (2008) Human immunodeficiency virus type 1 Vpu protein interacts with CD74 and modulates major histocompatibility complex class II presentation. *Journal of virology* **82**: 893–902.
Ilina, T, Parniak, MA (2008) Inhibitors of HIV‐1 Reverse Transcriptase. In: Advances in Pharmacology, Academic Press, pp. 121-167.
Jabara, CB, Jones, CD, Roach, J, Anderson, JA, Swanstrom, R (2011) Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. *Proc Natl Acad Sci U S A* **108**: 20166-20171.
Jacks, T, Power, MD, Masiarz, FR, Luciw, PA, Barr, PJ, Varmus, HE (1988) Characterization of ribosomal frameshifting in HIV-1 gag-pol expression. *Nature* **331**: 280-283.
Jackson, JB, Becker-Pergola, G, Guay, LA, Musoke, P, Mracna, M, Fowler, MG, Mofenson, LM, Mirochnick, M, Mmiro, F, Eshleman, SH (2000) Identification of the K103N resistance mutation in Ugandan women receiving nevirapine to prevent HIV-1 vertical transmission. *AIDS* **14**: F111-115.
Jackson, JB, Musoke, P, Fleming, T, Guay, LA, Bagenda, D, Allen, M, Nakabiito, C, Sherman, J, Bakaki, P, Owor, M, Ducar, C, Deseyve, M, Mwatha, A, Emel, L, Duefield, C, Mirochnick, M, Fowler, MG, Mofenson, L, Miotti, P, Gigliotti, M, Bray, D, Mmiro, F (2003) Intrapartum and neonatal single-dose nevirapine compared with zidovudine for prevention of mother-to-child transmission of HIV-1 in Kampala, Uganda: 18-month follow-up of the HIVNET 012 randomised trial. *Lancet* **362**: 859-868.
Jacobo-Molina, A, Arnold, E (1991) HIV reverse transcriptase structure-function relationships. *Biochemistry* **30**: 6351–6361.
Jager, S, Kim, DY, Hultquist, JF, Shindo, K, LaRue, RS, Kwon, E, Li, M, Anderson, BD, Yen, L, Stanley, D, Mahon, C, Kane, J, Franks-Skiba, K, Cimermancic, P, Burlingame, A, Sali, A, Craik, CS, Harris, RS, Gross, JD, Krogan, NJ Vif hijacks CBF-beta to degrade APOBEC3G and promote HIV-1 infection. *Nature* **481**: 371-375.
Jakobsen, MR, Tolstrup, M, Sogaard, OS, Jorgensen, LB, Gorry, PR, Laursen, A, Ostergaard, L (2010) Transmission of HIV-1 drug-resistant variants: prevalence and effect on treatment outcome. *Clin Infect Dis* **50**: 566-573.
Jakobson, CG, Dinnar, U, Feinsod, M, Nemirovsky, Y (2002) Ion-sensitive field-effect transistors in standard CMOS fabricated by post processing. *IEEE Sensors Journal* **2**: 279-287.
Ji, H, Li, Y, Graham, M, Liang, BB, Pilon, R, Tyson, S, Peters, G, Tyler, S, Merks, H, Bertagnolio, S, Soto-Ramirez, L, Sandstrom, P, Brooks, J (2011) Next-generation

sequencing of dried blood spot specimens: a novel approach to HIV drug-resistance surveillance. *Antivir Ther* **16**: 871-878.
Ji, H, Liang, B, Li, Y, Van Domselaar, G, Graham, M, Tyler, S, Merks, H, Sandstrom, P, Brooks, J (2012) Low abundance drug resistance variants in transmitted HIV drug resistance surveillance specimens identified using tagged pooled pyrosequencing. *J Virol Methods* **187**: 314-320.
Ji, H, Masse, N, Tyler, S, Liang, B, Li, Y, Merks, H, Graham, M, Sandstrom, P, Brooks, J (2010) HIV drug resistance surveillance using pooled pyrosequencing. *PLoS One* **5**: e9263.
Ji, JP, Loeb, LA (1992) Fidelity of HIV-1 reverse transcriptase copying RNA in vitro. *Biochemistry* **31**: 954-958.
Johnson, JA, Geretti, AM (2010) Low-frequency HIV-1 drug resistance mutations can be clinically significant but must be interpreted with caution. *J Antimicrob Chemother* **65**: 1322-1326.
Johnson, JA, Li, J-F, Wei, X, Lipscomb, J, Irlbeck, D, Craig, C, Smith, A, Bennett, DE, Monsour, M, Sandstrom, P, Lanier, ER, Heneine, W (2008) Minority HIV-1 Drug Resistance Mutations Are Present in Antiretroviral Treatment–Naïve Populations and Associate with Reduced Treatment Efficacy. *PLoS Med* **5**: e158.
Johnson, JA, Li, JF, Morris, L, Martinson, N, Gray, G, McIntyre, J, Heneine, W (2005) Emergence of drug-resistant HIV-1 after intrapartum administration of single-dose nevirapine is substantially underestimated. *J Infect Dis* **192**: 16-23.
Johnson, VA, Brun-Vézinet, Ft, Clotet, B, Gunthard, HF, Kuritzkes, DR, Pillay, D, Schapiro, JM, Richman, DD (2009) Update of the drug resistance mutations in HIV-1: December 2009. *Top HIV Med* **17**: 138–145.
Johnson, VA, Calvez, V, Gunthard, HF, Paredes, R, Pillay, D, Shafer, RW, Wensing, AM, Richman, DD (2013) Update of the drug resistance mutations in HIV-1: March 2013. *Top Antivir Med* **21**: 6-14.
Jourdain, G, Ngo-Giang-Huong, N, Le Coeur, S, Bowonwatanuwong, C, Kantipong, P, Leechanachai, P, Ariyadej, S, Leenasirimakul, P, Hammer, S, Lallemant, M (2004) Intrapartum exposure to nevirapine and subsequent maternal responses to nevirapine-based antiretroviral therapy. *N Engl J Med* **351**: 229-240.
Jowett, JB, Planelles, V, Poon, B, Shah, NP, Chen, M-L, Chen, IS (1995) The human immunodeficiency virus type 1 vpr gene arrests infected T cells in the G2+ M phase of the cell cycle. *Journal of virology* **69**: 6304–6313.
Judo, MS, Wedel, AB, Wilson, C (1998) Stimulation and suppression of PCR-mediated recombination. *Nucleic Acids Res* **26**: 1819-1825.
Kanagawa, T (2003) Bias and artifacts in multitemplate polymerase chain reactions (PCR). *J Biosci Bioeng* **96**: 317-323.
Kaplan, JE, Benson, C, Holmes, KH, Brooks, JT, Pau, A, Masur, H (2009) Guidelines for prevention and treatment of opportunistic infections in HIV-infected adults and adolescents: recommendations from CDC, the National Institutes of Health, and the HIV Medicine Association of the Infectious Diseases Society of America. *MMWR Recomm Rep* **58**: 1-207; quiz CE201-204.
Karacostas, V, Wolffe, EJ, Nagashima, K, Gonda, MA, Moss, B (1993) Overexpression of the HIV-1 gag-pol polyprotein results in intracellular activation of HIV-1 protease and inhibition of assembly and budding of virus-like particles. *Virology* **193**: 661–671.
Karn, J, Stoltzfus, CM Transcriptional and posttranscriptional regulation of HIV-1 gene expression. *Cold Spring Harbor Perspectives in Medicine* **2**.
Kassaye, S, Lee, E, Kantor, R, Johnston, E, Winters, M, Zijenah, L, Mateta, P, Katzenstein, D (2007) Drug resistance in plasma and breast milk after single-dose

nevirapine in subtype C HIV type 1: population and clonal sequence analysis. *AIDS Res Hum Retroviruses* **23**: 1055-1061.
Katoh, K, Asimenos, G, Toh, H (2009) Multiple alignment of DNA sequences with MAFFT. *Methods Mol Biol* **537**: 39-64.
Katoh, K, Kuma, K, Toh, H, Miyata, T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* **33**: 511-518.
Katoh, K, Misawa, K, Kuma, K, Miyata, T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**: 3059-3066.
Katoh, K, Toh, H (2008) Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform* **9**: 286-298.
Katoh, K, Toh, H (2010) Parallelization of the MAFFT multiple sequence alignment program. *Bioinformatics* **26**: 1899-1900.
Kearse, M, Moir, R, Wilson, A, Stones-Havas, S, Cheung, M, Sturrock, S, Buxton, S, Cooper, A, Markowitz, S, Duran, C, Thierer, T, Ashton, B, Meintjes, P, Drummond, A (2012) Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**: 1647-1649.
Keele, BF, Giorgi, EE, Salazar-Gonzalez, JF, Decker, JM, Pham, KT, Salazar, MG, Sun, C, Grayson, T, Wang, S, Li, H, Wei, X, Jiang, C, Kirchherr, JL, Gao, F, Anderson, JA, Ping, LH, Swanstrom, R, Tomaras, GD, Blattner, WA, Goepfert, PA, Kilby, JM, Saag, MS, Delwart, EL, Busch, MP, Cohen, MS, Montefiori, DC, Haynes, BF, Gaschen, B, Athreya, GS, Lee, HY, Wood, N, Seoighe, C, Perelson, AS, Bhattacharya, T, Korber, BT, Hahn, BH, Shaw, GM (2008) Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proc Natl Acad Sci U S A* **105**: 7552-7557.
Kellam, P, Boucher, CA, Tijnagel, JM, Larder, BA (1994) Zidovudine treatment results in the selection of human immunodeficiency virus type 1 variants whose genotypes confer increasing levels of drug resistance. *J Gen Virol* **75 (Pt 2)**: 341-351.
Kellam, P, Larder, BA (1994) Recombinant virus assay: a rapid, phenotypic assay for assessment of drug susceptibility of human immunodeficiency virus type 1 isolates. *Antimicrob Agents Chemother* **38**: 23-30.
Kinde, I, Wu, J, Papadopoulos, N, Kinzler, KW, Vogelstein, B (2011) Detection and quantification of rare mutations with massively parallel sequencing. *Proc Natl Acad Sci U S A* **108**: 9530-9535.
Kliger, Y, Aharoni, A, Rapaport, D, Jones, P, Blumenthal, R, Shai, Y (1997) Fusion peptides derived from the HIV type 1 glycoprotein 41 associate within phospholipid membranes and inhibit cell-cell Fusion. Structure-function study. *J Biol Chem* **272**: 13496-13505.
Klimkait, T, Strebel, K, Hoggan, MD, Martin, MA, Orenstein, JM (1990) The human immunodeficiency virus type 1-specific protein vpu is required for efficient virus maturation and release. *Journal of Virology* **64**: 621-629.
Koch, MA, Volberding, PA, Lagakos, SW, Booth, DK, Pettinelli, C, Myers, MW (1992) Toxic effects of zidovudine in asymptomatic human immunodeficiency virus-infected individuals with CD4+ cell counts of 0.50 x 10(9)/L or less. Detailed and updated results from protocol 019 of the AIDS Clinical Trials Group. *Arch Intern Med* **152**: 2286-2292.
Korber, B, Gaschen, B, Yusim, K, Thakallapally, R, Kesmir, C, Detours, V (2001) Evolutionary and immunological implications of contemporary HIV-1 variation. *British Medical Bulletin* **58**: 19-42.
Korber, B, Muldoon, M, Theiler, J, Gao, F, Gupta, R, Lapedes, A, Hahn, BH, Wolinsky, S, Bhattacharya, T (2000) Timing the ancestor of the HIV-1 pandemic strains. *Science* **288**: 1789-1796.

Korn, K, Reil, H, Walter, H, Schmidt, B (2003) Quality control trial for human immunodeficiency virus type 1 drug resistance testing using clinical samples reveals problems with detecting minority species and interpretation of test results. *J Clin Microbiol* **41**: 3559-3565.
Kostrikis, LG, Touloumi, G, Karanicolas, R, Pantazis, N, Anastassopoulou, C, Karafoulidou, A, Goedert, JJ, Hatzakis, A (2002) Quantitation of human immunodeficiency virus type 1 DNA forms with the second template switch in peripheral blood cells predicts disease progression independently of plasma RNA load. *J Virol* **76**: 10099-10108.
Kozal, MJ, Chiarella, J, St John, EP, Moreno, EA, Simen, BB, Arnold, TE, Lataillade, M (2011) Prevalence of low-level HIV-1 variants with reverse transcriptase mutation K65R and the effect of antiretroviral drug exposure on variant levels. *Antivir Ther* **16**: 925-929.
Kozal, MJ, Shah, N, Shen, N, Yang, R, Fucini, R, Merigan, TC, Richman, DD, Morris, D, Hubbell, E, Chee, M, Gingeras, TR (1996) Extensive polymorphisms observed in HIV-1 clade B protease gene using high-density oligonucleotide arrays. *Nat Med* **2**: 753-759.
Kumar, S, Banks, TW, Cloutier, S (2012) SNP Discovery through Next-Generation Sequencing and Its Applications. *Int J Plant Genomics* **2012**: 831460.
Kunin, V, Engelbrektson, A, Ochman, H, Hugenholtz, P (2009) Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ Microbiol* **12**: 118-123.
Kuritzkes, DR, Marschner, I, Johnson, VA, Bassett, R, Eron, JJ, Fischl, MA, Murphy, RL, Fife, K, Maenza, J, Rosandich, ME (1999) Lamivudine in combination with zidovudine, stavudine, or didanosine in patients with HIV-1 infection. A randomized, double-blind, placebo-controlled trial. *AIDS* **13**: 685-694.
LaFemina, RL, Schneider, CL, Robbins, HL, Callahan, PL, LeGrow, K, Roth, E, Schleif, WA, Emini, EA (1992) Requirement of active human immunodeficiency virus type 1 integrase enzyme for productive infection of human T-lymphoid cells. *Journal of Virology* **66**: 7414-7419.
Laguette, N, Benichou, S, Basmaciogullari, S (2009) Human Immunodeficiency Virus Type 1 Nef Incorporation into Virions Does Not Increase Infectivity. *Journal of Virology* **83**: 1093-1104.
Lallemant, M, Jourdain, G, Le Coeur, S, Mary, JY, Ngo-Giang-Huong, N, Koetsawang, S, Kanshana, S, McIntosh, K, Thaineua, V (2004) Single-dose perinatal nevirapine plus standard zidovudine to prevent mother-to-child transmission of HIV-1 in Thailand. *N Engl J Med* **351**: 217-228.
Lama, J, Mangasarian, A, Trono, D (1999) Cell-surface expression of CD4 reduces HIV-1 infectivity by blocking Env incorporation in a Nef- and Vpu-inhibitable manner. *Current biology: CB* **9**: 622-631.
Larder, B, De Vroey, V, Dehertogh, P (1999) Abstracts of the 3d International Workshop on HIV Drug Resistance and Treatment Strategies (San Diego). London: International Medical Press; 1999. Predicting HIV-1 phenotypic resistance from genotype using a large phenotype-genotype relational database [abstract 59]. 41-42.
Larder, BA, Darby, G, Richman, DD (1989a) HIV with reduced sensitivity to zidovudine (AZT) isolated during prolonged therapy. *Science* **243**: 1731-1734.
Larder, BA, Darby, G, Richman, DD (1989b) HIV with reduced sensitivity to zidovudine (AZT) isolated during prolonged therapy. *Science* **243**: 1731-1734.
Larder, BA, Kellam, P, Kemp, SD (1991) Zidovudine resistance predicted by direct detection of mutations in DNA from HIV-infected lymphocytes. *AIDS* **5**: 137-144.
Larder, BA, Kemp, SD (1989) Multiple mutations in HIV-1 reverse transcriptase confer high-level resistance to zidovudine (AZT). *Science* **246**: 1155-1158.

Larder, BA, Kohli, A, Kellam, P, Kemp, SD, Kronick, M, Henfrey, RD (1993) Quantitative detection of HIV-1 drug resistance mutations by automated DNA sequencing. *Nature* **365**: 671-673.
Larder, BA, Purifoy, DJM, Powell, KL, Darby, G (1987) Site-specific mutagenesis of AIDS virus reverse transcriptase. **327**: 716-717.
Larkin, MA, Blackshields, G, Brown, NP, Chenna, R, McGettigan, PA, McWilliam, H, Valentin, F, Wallace, IM, Wilm, A, Lopez, R, Thompson, JD, Gibson, TJ, Higgins, DG (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* **23**: 2947-2948.
Larsen, LS, Beliakova-Bethell, N, Bilanchone, V, Zhang, M, Lamsa, A, Dasilva, R, Hatfield, GW, Nagashima, K, Sandmeyer, S (2008) Ty3 nucleocapsid controls localization of particle assembly. *J Virol* **82**: 2501-2514.
Lassmann, T, Sonnhammer, EL (2005) Kalign--an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics* **6**: 298.
Lataillade, M, Chiarella, J, Yang, R, Schnittman, S, Wirtz, V, Uy, J, Seekins, D, Krystal, M, Mancini, M, McGrath, D, Simen, B, Egholm, M, Kozal, M (2010) Prevalence and clinical significance of HIV drug resistance mutations by ultra-deep sequencing in antiretroviral-naive subjects in the CASTLE study. *PLoS One* **5**: e10952.
Le, T, Chiarella, J, Simen, BB, Hanczaruk, B, Egholm, M, Landry, ML, Dieckhaus, K, Rosen, MI, Kozal, MJ (2009) Low-abundance HIV drug-resistant viral variants in treatment-experienced persons correlate with historical antiretroviral use. *PLoS One* **4**: e6079.
Lee, W-P, Stromberg, M, Ward, A, Stewart, C, Garrison, E, Marth, GT (2013) MOSAIK: A hash-based algorithm for accurate next-generation sequencing read mapping. *arXiv preprint arXiv:13091149*.
Lehman, DA, Wamalwa, DC, McCoy, CO, Matsen, FA, Langat, A, Chohan, BH, Benki-Nugent, S, Custers-Allen, R, Bushman, FD, John-Stewart, GC, Overbaugh, J (2012) Low-frequency nevirapine resistance at multiple sites may predict treatment failure in infants on nevirapine-based treatment. *J Acquir Immune Defic Syndr* **60**: 225-233.
Leitner, T, Halapi, E, Scarlatti, G, Rossi, P, Albert, J, Fenyo, EM, Uhlen, M (1993) Analysis of heterogeneous viral populations by direct DNA sequencing. *Biotechniques* **15**: 120-127.
Lemey, P, Pybus, OG, Rambaut, A, Drummond, AJ, Robertson, DL, Roques, P, Worobey, M, Vandamme, A-M (2004a) The Molecular Population Genetics of HIV-1 Group O. *Genetics* **167**: 1059-1068.
Lemey, P, Pybus, OG, Rambaut, A, Drummond, AJ, Robertson, DL, Roques, P, Worobey, M, Vandamme, AM (2004b) The molecular population genetics of HIV-1 group O. *Genetics* **167**: 1059-1068.
Lenassi, M, Cagney, G, Liao, M, Vaupotic, T, Bartholomeeusen, K, Cheng, Y, Krogan, NJ, Plemenitas, A, Peterlin, BM HIV Nef is secreted in exosomes and triggers apoptosis in bystander CD4+ T cells. *Traffic* **11**: 110-122.
Letvin, NL (2006) Progress and obstacles in the development of an AIDS vaccine. *Nature Reviews Immunology* **6**: 930-939.
Levene, MJ, Korlach, J, Turner, SW, Foquet, M, Craighead, HG, Webb, WW (2003) Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* **299**: 682-686.
Levy, JA, Hoffman, AD, Kramer, SM, Landis, JA, Shimabukuro, JM, Oshiro, LS (1984) Isolation of lymphocytopathic retroviruses from San Francisco patients with AIDS. *Science* **225**: 840-842.

Lewis, MJ, Lee, P, Ng, HL, Yang, OO (2012) Immune Selection In Vitro Reveals Human Immunodeficiency Virus Type 1 Nef Sequence Motifs Important for Its Immune Evasion Function In Vivo. *Journal of Virology* **86**: 7126-7135.
Li, S, Chou, H-H (2004) LUCY2: an interactive DNA sequence quality trimming and vector removal tool. *Bioinformatics* **20**: 2865–2866.
Li, TS, Tubiana, R, Katlama, C, Calvez, V, Ait Mohand, H, Autran, B (1998) Long-lasting recovery in CD4 T-cell function and viral-load reduction after highly active antiretroviral therapy in advanced HIV-1 disease. *Lancet* **351**: 1682-1686.
Liang, B, Luo, M, Scott-Herridge, J, Semeniuk, C, Mendoza, M, Capina, R, Sheardown, B, Ji, H, Kimani, J, Ball, BT, Van Domselaar, G, Graham, M, Tyler, S, Jones, SJ, Plummer, FA (2011) A comparison of parallel pyrosequencing and sanger clone-based sequencing and its impact on the characterization of the genetic diversity of HIV-1. *PLoS One* **6**: e26745.
Liang, J-S, Distler, O, Cooper, DA, Jamil, H, Deckelbaum, RJ, Ginsberg, HN, Sturley, SL (2001) HIV protease inhibitors protect apolipoprotein B from degradation by the proteasome: A potential mechanism for protease inhibitor-induced hyperlipidemia. *Nature Medicine* **7**: 1327-1331.
Liu, J, Bartesaghi, A, Borgnia, MJ, Sapiro, G, Subramaniam, S (2008) Molecular architecture of native HIV-1 gp120 trimers. *Nature* **455**: 109-113.
Liu, L, Li, Y, Li, S, Hu, N, He, Y, Pong, R, Lin, D, Lu, L, Law, M (2012) Comparison of next-generation sequencing systems. *J Biomed Biotechnol* **2012**: 251364.
Liu, SL, Rodrigo, AG, Shankarappa, R, Learn, GH, Hsu, L, Davidov, O, Zhao, LP, Mullins, JI (1996) HIV quasispecies and resampling. *Science* **273**: 415-416.
Liu, TF, Shafer, RW (2006) Web resources for HIV type 1 genotypic-resistance test interpretation. *Clin Infect Dis* **42**: 1608-1618.
Lockman, S, Shapiro, RL, Smeaton, LM, Wester, C, Thior, I, Stevens, L, Chand, F, Makhema, J, Moffat, C, Asmelash, A, Ndase, P, Arimi, P, van Widenfelt, E, Mazhani, L, Novitsky, V, Lagakos, S, Essex, M (2007) Response to antiretroviral therapy after a single, peripartum dose of nevirapine. *N Engl J Med* **356**: 135-147.
Lole, KS, Bollinger, RC, Paranjape, RS, Gadkari, D, Kulkarni, SS, Novak, NG, Ingersoll, R, Sheppard, HW, Ray, SC (1999) Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. *J Virol* **73**: 152-160.
Loman, NJ, Misra, RV, Dallman, TJ, Constantinidou, C, Gharbia, SE, Wain, J, Pallen, MJ (2012) Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol* **30**: 434-439.
Long, EM, Martin, HL, Jr., Kreiss, JK, Rainwater, SM, Lavreys, L, Jackson, DJ, Rakwar, J, Mandaliya, K, Overbaugh, J (2000) Gender differences in HIV-1 diversity at time of infection. *Nat Med* **6**: 71-75.
Lorenzi, P, Opravil, M, Hirschel, B, Chave, JP, Furrer, HJ, Sax, H, Perneger, TV, Perrin, L, Kaiser, L, Yerly, S (1999) Impact of drug resistance mutations on virologic response to salvage therapy. Swiss HIV Cohort Study. *AIDS* **13**: F17-21.
Loubser, S, Balfe, P, Sherman, G, Hammer, S, Kuhn, L, Morris, L (2006) Decay of K103N mutants in cellular DNA and plasma RNA after single-dose nevirapine to reduce mother-to-child HIV transmission. *AIDS* **20**: 995-1002.
Low-Beer, S, Yip, B, O'Shaughnessy, MV, Hogg, RS, Montaner, JS (2000) Adherence to triple therapy and viral load response. *J Acquir Immune Defic Syndr* **23**: 360-361.
Lundgren, JD, Phillips, AN, Pedersen, C, Clumeck, N, Gatell, JM, Johnson, AM, Ledergerber, B, Vella, S, Nielsen, JO (1994) Comparison of long-term prognosis of

patients with AIDS treated and not treated with zidovudine. AIDS in Europe Study Group. *JAMA* **271**: 1088-1092.
Luo, C, Tsementzi, D, Kyrpides, N, Read, T, Konstantinidis, KT (2012) Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. *PLoS One* **7**: e30087.
Madani, N, Kabat, D (1998) An endogenous inhibitor of human immunodeficiency virus in human lymphocytes is overcome by the viral Vif protein. *Journal of virology* **72**: 10251-10255.
Malim, MH, Bohnlein, S, Hauber, J, Cullen, BR (1989a) Functional dissection of the HIV-1 Rev trans-activator--derivation of a trans-dominant repressor of Rev function. *Cell* **58**: 205-214.
Malim, MH, Cullen, BR (1991) HIV-1 structural gene expression requires the binding of multiple Rev monomers to the viral RRE: implications for HIV-1 latency. *Cell* **65**: 241-248.
Malim, MH, Hauber, J, Le, S-Y, Maizel, JV, Cullen, BR (1989b) The HIV-1 rev trans-activator acts through a structured target sequence to activate nuclear export of unspliced viral mRNA. *Nature* **338**: 254–257.
Mammano, F, Trouplin, V, Zennou, V, Clavel, F (2000) Retracing the evolutionary pathways of human immunodeficiency virus type 1 resistance to protease inhibitors: virus fitness in the absence and in the presence of drug. *J Virol* **74**: 8524-8531.
Mansky, LM (1996a) Forward mutation rate of human immunodeficiency virus type 1 in a T lymphoid cell line. *AIDS Res Hum Retroviruses* **12**: 307-314.
Mansky, LM (1996b) The mutation rate of human immunodeficiency virus type 1 is influenced by the vpr gene. *Virology* **222**: 391-400.
Mansky, LM (1998) Retrovirus mutation rates and their role in genetic variation. *J Gen Virol* **79 (Pt 6)**: 1337-1345.
Mansky, LM, Temin, HM (1995) Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. *Journal of Virology* **69**: 5087-5094.
Marcello, A, Zoppe, M, Giacca, M (2001) Multiple modes of transcriptional regulation by the HIV-1 Tat transactivator. *IUBMB Life* **51**: 175-181.
Mardis, ER (2008a) The impact of next-generation sequencing technology on genetics. *Trends in genetics* **24**: 133.
Mardis, ER (2008b) The impact of next-generation sequencing technology on genetics. *Trends Genet* **24**: 133-141.
Margulies, M, Egholm, M, Altman, WE, Attiya, S, Bader, JS, Bemben, LA, Berka, J, Braverman, MS, Chen, Y-J, Chen, Z, Dewell, SB, Du, L, Fierro, JM, Gomes, XV, Godwin, BC, He, W, Helgesen, S, Ho, CH, Irzyk, GP, Jando, SC, Alenquer, MLI, Jarvie, TP, Jirage, KB, Kim, J-B, Knight, JR, Lanza, JR, Leamon, JH, Lefkowitz, SM, Lei, M, Li, J, Lohman, KL, Lu, H, Makhijani, VB, McDade, KE, McKenna, MP, Myers, EW, Nickerson, E, Nobile, JR, Plant, R, Puc, BP, Ronan, MT, Roth, GT, Sarkis, GJ, Simons, JF, Simpson, JW, Srinivasan, M, Tartaro, KR, Tomasz, A, Vogt, KA, Volkmer, GA, Wang, SH, Wang, Y, Weiner, MP, Yu, P, Begley, RF, Rothberg, JM (2005a) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**: 376-380.
Margulies, M, Egholm, M, Altman, WE, Attiya, S, Bader, JS, Bemben, LA, Berka, J, Braverman, MS, Chen, YJ, Chen, Z, Dewell, SB, Du, L, Fierro, JM, Gomes, XV, Godwin, BC, He, W, Helgesen, S, Ho, CH, Irzyk, GP, Jando, SC, Alenquer, ML, Jarvie, TP, Jirage, KB, Kim, JB, Knight, JR, Lanza, JR, Leamon, JH, Lefkowitz, SM, Lei, M, Li, J, Lohman, KL, Lu, H, Makhijani, VB, McDade, KE, McKenna, MP, Myers, EW, Nickerson, E, Nobile, JR, Plant, R, Puc, BP, Ronan, MT, Roth, GT, Sarkis, GJ, Simons, JF, Simpson, JW, Srinivasan, M, Tartaro, KR, Tomasz, A, Vogt,

KA, Volkmer, GA, Wang, SH, Wang, Y, Weiner, MP, Yu, P, Begley, RF, Rothberg, JM (2005b) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**: 376-380.
Martinez-Picado, J, Sutton, L, De Pasquale, MP, Savara, AV, D'Aquila, RT (1999) Human immunodeficiency virus type 1 cloning vectors for antiretroviral resistance testing. *J Clin Microbiol* **37**: 2943-2951.
Martinson, NA, Morris, L, Gray, G, Moodley, D, Pillay, V, Cohen, S, Dhlamini, P, Puren, A, Bhayroo, S, Steyn, J, McIntyre, JA (2007) Selection and persistence of viral resistance in HIV-infected children after exposure to single-dose nevirapine. *J Acquir Immune Defic Syndr* **44**: 148-153.
Masur, H, Kaplan, JE (2009) New guidelines for the management of HIV-related opportunistic infections. *JAMA* **301**: 2378-2380.
Mayer, KH, Hanna, GJ, Richard, T (2001) Clinical use of genotypic and phenotypic drug resistance testing to monitor antiretroviral chemotherapy. *Clinical Infectious Diseases* **32**: 774-782.
McCarthy, A (2010) Third generation DNA sequencing: pacific biosciences' single molecule real time technology. *Chem Biol* **17**: 675-676.
McCutchan, FE (2000) Understanding the genetic diversity of HIV-1. *AIDS* **14 Suppl 3**: S31-44.
McCutchan, FE (2006) Global epidemiology of HIV. *Journal of Medical Virology* **78**: S7–S12.
McCutchan, FE, Carr, JK, Murphy, D, Piyasirisilp, S, Gao, F, Hahn, B, Yu, X-F, Beyrer, C, Birx, DL (2002) Precise mapping of recombination breakpoints suggests a common parent of two BC recombinant HIV type 1 strains circulating in China. *AIDS research and human retroviruses* **18**: 1135–1140.
McDonald, D, Vodicka, MA, Lucero, G, Svitkina, TM, Borisy, GG, Emerman, M, Hope, TJ (2002) Visualization of the intracellular behavior of HIV in living cells. *The Journal of Cell Biology* **159**: 441-452.
McGowan, JP, Shah, SS (2000) Prevention of perinatal HIV transmission during pregnancy. *Journal of Antimicrobial Chemotherapy* **46**: 657-668.
McIntyre, JA, Hopley, M, Moodley, D, Eklund, M, Gray, GE, Hall, DB, Robinson, P, Mayers, D, Martinson, NA (2009) Efficacy of short-course AZT plus 3TC to reduce nevirapine resistance in the prevention of mother-to-child HIV transmission: a randomized clinical trial. *PLoS medicine* **6**: e1000172.
Melikyan, GB (2008) Common principles and intermediates of viral protein-mediated fusion: the HIV-1 paradigm. *Retrovirology* **5**: 111.
Melikyan, GB, Markosyan, RM, Hemmati, H, Delmedico, MK, Lambert, DM, Cohen, FS (2000) Evidence That the Transition of HIV-1 Gp41 into a Six-Helix Bundle, Not the Bundle Configuration, Induces Membrane Fusion. *The Journal of Cell Biology* **151**: 413-424.
Merry, C, Barry, MG, Mulcahy, F, Ryan, M, Heavey, J, Tjia, JF, Gibbons, SE, Breckenridge, AM, Back, DJ (1997) Saquinavir pharmacokinetics alone and in combination with ritonavir in HIV-infected patients. *AIDS* **11**: F29-F33.
Metzker, ML (2005) Emerging technologies in DNA sequencing. *Genome Res* **15**: 1767-1776.
Metzker, ML (2009) Sequencing technologies — the next generation. *Nature Reviews Genetics* **11**: 31-46.
Metzner, KJ, Rauch, P, Braun, P, Knechten, H, Ehret, R, Korn, K, Kaiser, R, Sichtig, N, Ranneberg, B, van Lunzen, J, Walter, H (2011) Prevalence of key resistance mutations K65R, K103N, and M184V as minority HIV-1 variants in chronically HIV-1 infected, treatment-naive patients. *J Clin Virol* **50**: 156-161.

Meyerhans, A, Vartanian, JP, Wain-Hobson, S (1990) DNA recombination during PCR. *Nucleic Acids Res* **18**: 1687-1691.
Mild, M, Hedskog, C, Jernberg, J, Albert, J (2011) Performance of ultra-deep pyrosequencing in analysis of HIV-1 pol gene variation. *PLoS One* **6**: e22741.
Milgrew, MJ, Hammond, PA, Cumming, DRS (2004) The development of scalable sensor arrays using standard CMOS technology. *Sensors and Actuators B: Chemical* **103**: 37-42.
Miller, J, Carr, A, Smith, D, Emery, S, Law, MG, Grey, P, Cooper, DA (2000) Lipodystrophy following antiretroviral therapy of primary HIV infection. *AIDS* **14**: 2406-2407.
Miller, MD, Farnet, CM, Bushman, FD (1997) Human immunodeficiency virus type 1 preintegration complexes: studies of organization and composition. *Journal of Virology* **71**: 5382-5390.
Mills, EJ, Bakanda, C, Birungi, J, Chan, K, Ford, N, Cooper, CL, Nachega, JB, Dybul, M, Hogg, RS (2011) Life expectancy of persons receiving combination antiretroviral therapy in low-income countries: a cohort analysis from Uganda. *Ann Intern Med* **155**: 209-216.
Mitsuya, Y, Varghese, V, Wang, C, Liu, TF, Holmes, SP, Jayakumar, P, Gharizadeh, B, Ronaghi, M, Klein, D, Fessel, WJ, Shafer, RW (2008) Minority human immunodeficiency virus type 1 variants in antiretroviral-naive persons with reverse transcriptase codon 215 revertant mutations. *J Virol* **82**: 10747-10755.
Mocroft, A, Phillips, AN, Ledergerber, B, Smith, C, Bogner, JR, Lacombe, K, Wiercinska-Drapalo, A, Reiss, P, Kirk, O, Lundgren, JD (2010) Estimated average annual rate of change of CD4(+) T-cell counts in patients on combination antiretroviral therapy. *Antivir Ther* **15**: 563-570.
Molla, A, Korneyeva, M, Gao, Q, Vasavanonda, S, Schipper, PJ, Mo, HM, Markowitz, M, Chernyavskiy, T, Niu, P, Lyons, N, Hsu, A, Granneman, GR, Ho, DD, Boucher, CA, Leonard, JM, Norbeck, DW, Kempf, DJ (1996) Ordered accumulation of mutations in HIV protease confers resistance to ritonavir. *Nat Med* **2**: 760-766.
Montaner, JS, Reiss, P, Cooper, D, Vella, S, Harris, M, Conway, B, Wainberg, MA, Smith, D, Robinson, P, Hall, D (1998a) A randomized, double-blind trial comparing combinations of nevirapine, didanosine, and zidovudine for HIV-infected patients. *JAMA: the journal of the American Medical Association* **279**: 930-937.
Montaner, JS, Reiss, P, Cooper, D, Vella, S, Harris, M, Conway, B, Wainberg, MA, Smith, D, Robinson, P, Hall, D, Myers, M, Lange, JM (1998b) A randomized, double-blind trial comparing combinations of nevirapine, didanosine, and zidovudine for HIV-infected patients: the INCAS Trial. Italy, The Netherlands, Canada and Australia Study. *JAMA* **279**: 930-937.
Moore, MJ, Dhingra, A, Soltis, PS, Shaw, R, Farmerie, WG, Folta, KM, Soltis, DE (2006) Rapid and accurate pyrosequencing of angiosperm plastid genomes. *BMC Plant Biol* **6**: 17.
Moorthy, A, Kuhn, L, Coovadia, A, Meyers, T, Strehlau, R, Sherman, G, Tsai, WY, Chen, YH, Abrams, EJ, Persaud, D (2011) Induction therapy with protease-inhibitors modifies the effect of nevirapine resistance on virologic response to nevirapine-based HAART in children. *Clin Infect Dis* **52**: 514-521.
Nahmias, A, Weiss, J, Yao, X, Lee, F, Kodsi, R, Schanfield, M, Matthews, T, Bolognesi, D, Durack, D, Motulsky, A (1986) Evidence for human infection with an HTLV III/LAV-like virus in Central Africa, 1959. *The Lancet* **327**: 1279-1280.
Nattrass, N (2006) South Africa's "rollout" of highly active antiretroviral therapy: a critical assessment. *J Acquir Immune Defic Syndr* **43**: 618-623.
Navarro, F, Landau, NR (2004) Recent insights into HIV-1 Vif. *Current Opinion in Immunology* **16**: 477-482.

Neher, RA, Leitner, T Recombination rate and selection strength in HIV intra-patient evolution. *PLoS Comput Biol* **6**: e1000660.
Nermut, MV, Hockley, DJ, Bron, P, Thomas, D, Zhang, WH, Jones, IM (1998) Further evidence for hexagonal organization of HIV gag protein in prebudding assemblies and immature virus-like particles. *J Struct Biol* **123**: 143-149.
Niedringhaus, TP, Milanova, D, Kerby, MB, Snyder, MP, Barron, AE (2011) Landscape of next-generation sequencing technologies. *Anal Chem* **83**: 4327-4341.
Nomaguchi, M, Fujita, M, Adachi, A (2008) Role of HIV-1 Vpu protein for virus spread and pathogenesis. *Microbes and Infection* **10**: 960–967.
Notredame, C, Higgins, DG, Heringa, J (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302**: 205-217.
Nowak, MA, May, RM, Anderson, RM (1990) The evolutionary dynamics of HIV-1 quasispecies and the development of immunodeficiency disease. *AIDS* **4**: 1095-1103.
Ott, M, Geyer, M, Zhou, Q (2011) The Control of HIV Transcription: Keeping RNA Polymerase II on Track. *Cell Host & Microbe* **10**: 426-435.
Palmer, S, Boltz, V, Martinson, N, Maldarelli, F, Gray, G, McIntyre, J, Mellors, J, Morris, L, Coffin, J (2006) Persistence of nevirapine-resistant HIV-1 in women after single-dose nevirapine therapy for prevention of maternal-to-fetal HIV-1 transmission. *Proc Natl Acad Sci U S A* **103**: 7094-7099.
Palmer, S, Kearney, M, Maldarelli, F, Halvas, EK, Bixby, CJ, Bazmi, H, Rock, D, Falloon, J, Davey, RT, Jr., Dewar, RL, Metcalf, JA, Hammer, S, Mellors, JW, Coffin, JM (2005) Multiple, linked human immunodeficiency virus type 1 drug resistance mutations in treatment-experienced patients are missed by standard genotype analysis. *J Clin Microbiol* **43**: 406-413.
Palmisano, L, Vella, S (2011) A brief history of antiretroviral therapy of HIV infection: success and challenges. *Ann Ist Super Sanita* **47**: 44-48.
Pancera, M, Majeed, S, Ban, YE, Chen, L, Huang, CC, Kong, L, Kwon, YD, Stuckey, J, Zhou, T, Robinson, JE, Schief, WR, Sodroski, J, Wyatt, R, Kwong, PD Structure of HIV-1 gp120 with gp41-interactive region reveals layered envelope architecture and basis of conformational mobility. *Proc Natl Acad Sci U S A* **107**: 1166-1171.
Pandey, V, Nutter, RC, Prediger, E (2008) Applied Biosystems SOLiD™ System: Ligation - Based Sequencing. *Next Generation Genome Sequencing: Towards Personalized Medicine*: 29-42.
Paredes, R, Lalama, CM, Ribaudo, HJ, Schackman, BR, Shikuma, C, Giguel, F, Meyer, WA, 3rd, Johnson, VA, Fiscus, SA, D'Aquila, RT, Gulick, RM, Kuritzkes, DR (2010) Pre-existing minority drug-resistant HIV-1 variants, adherence, and risk of antiretroviral treatment failure. *J Infect Dis* **201**: 662-671.
Partaledis, JA, Yamaguchi, K, Tisdale, M, Blair, EE, Falcione, C, Maschera, B, Myers, RE, Pazhanisamy, S, Futer, O, Cullinan, AB (1995) In vitro selection and characterization of human immunodeficiency virus type 1 (HIV-1) isolates with reduced sensitivity to hydroxyethylamino sulfonamide inhibitors of HIV-1 aspartyl protease. *Journal of Virology* **69**: 5228-5235.
Peeters, M, Honoré, C, Huet, T, Bedjabaga, L, Ossari, S, Bussi, P, Cooper, RW, Delaporte, E (1989) Isolation and partial characterization of an HIV-related virus occurring naturally in chimpanzees in Gabon. *Aids* **3**: 625–630.
Pennisi, E (2010) Genomics. Semiconductors inspire new sequencing technologies. *Science* **327**: 1190.
Perfect, JR, Dismukes, WE, Dromer, F, Goldman, DL, Graybill, JR, Hamill, RJ, Harrison, TS, Larsen, RA, Lortholary, O, Nguyen, MH, Pappas, PG, Powderly, WG, Singh, N, Sobel, JD, Sorrell, TC (2010) Clinical practice guidelines for the

management of cryptococcal disease: 2010 update by the infectious diseases society of america. *Clin Infect Dis* **50**: 291-322.
Perrin, L, Kaiser, L, Yerly, S (2003) Travel and the spread of HIV-1 genetic variants. *Lancet Infect Dis* **3**: 22-27.
Petropoulos, CJ, Parkin, NT, Limoli, KL, Lie, YS, Wrin, T, Huang, W, Tian, H, Smith, D, Winslow, GA, Capon, DJ, Whitcomb, JM (2000) A novel phenotypic drug susceptibility assay for human immunodeficiency virus type 1. *Antimicrob Agents Chemother* **44**: 920-928.
Pettit, SC, Lindquist, JN, Kaplan, AH, Swanstrom, R (2005) Processing sites in the human immunodeficiency virus type 1 (HIV-1) Gag-Pro-Pol precursor are cleaved by the viral protease at different rates. *Retrovirology* **2**: 66.
Piketty, C, Race, E, Castiel, P, Belec, L, Peytavin, G, Si-Mohamed, A, Gonzalez-Canali, G, Weiss, L, Clavel, F, Kazatchkine, MD (1999) Efficacy of a five-drug combination including ritonavir, saquinavir and efavirenz in patients who failed on a conventional triple-drug regimen: phenotypic resistance to protease inhibitors predicts outcome of therapy. *AIDS* **13**: F71-77.
Ping, LH, Cohen, MS, Hoffman, I, Vernazza, P, Seillier-Moiseiwitsch, F, Chakraborty, H, Kazembe, P, Zimba, D, Maida, M, Fiscus, SA, Eron, JJ, Swanstrom, R, Nelson, JA (2000) Effects of genital tract inflammation on human immunodeficiency virus type 1 V3 populations in blood and semen. *J Virol* **74**: 8946-8952.
Plantier, JC, Leoz, M, Dickerson, JE, De Oliveira, F, Cordonnier, F, Lemee, V, Damond, F, Robertson, DL, Simon, F (2009) A new human immunodeficiency virus derived from gorillas. *Nat Med* **15**: 871-872.
Polz, MF, Cavanaugh, CM (1998) Bias in template-to-product ratios in multitemplate PCR. *Appl Environ Microbiol* **64**: 3724-3730.
Poveda, E, Briz, V, Soriano, V (2005) Enfuvirtide, the first fusion inhibitor to treat HIV infection. *Aids Rev* **7**: 139–147.
Preston, BD, Poiesz, BJ, Loeb, LA (1988) Fidelity of HIV-1 reverse transcriptase. *Science* **242**: 1168-1171.
Price, DA, Goulder, PJ, Klenerman, P, Sewell, AK, Easterbrook, PJ, Troop, M, Bangham, CR, Phillips, RE (1997) Positive selection of HIV-1 cytotoxic T lymphocyte escape variants during primary infection. *Proc Natl Acad Sci U S A* **94**: 1890-1895.
Pruss, D, Reeves, R, Bushman, FD, Wolffe, AP (1994) The influence of DNA and nucleosome structure on integration events directed by HIV integrase. *Journal of Biological Chemistry* **269**: 25031-25041.
Quail, MA, Smith, M, Coupland, P, Otto, TD, Harris, SR, Connor, TR, Bertoni, A, Swerdlow, HP, Gu, Y (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* **13**: 341.
Quinn, TC (1996) Global burden of the HIV pandemic. *Lancet* **348**: 99-106.
Quinones-Mateu, ME, Arts, EJ (2002) Fitness of drug resistant HIV-1: methodology and clinical implications. *Drug Resist Updat* **5**: 224-233.
Ratner, L, Haseltine, W, Patarca, R, Livak, KJ, Starcich, B, Josephs, SF, Doran, ER, Rafalski, JA, Whitehorn, EA, Baumeister, K (1985) Complete nucleotide sequence of the AIDS virus, HTLV-III.
Raymond, S, Delobel, P, Mavigner, M, Cazabat, M, Encinas, S, Souyris, C, Bruel, P, Sandres-Saune, K, Marchou, B, Massip, P, Izopet, J (2010) CXCR4-using viruses in plasma and peripheral blood mononuclear cells during primary HIV-1 infection and impact on disease progression. *AIDS* **24**: 2305-2312.

Razooky, BS, Weinberger, LS (2011) Mapping the architecture of the HIV-1 Tat circuit: A decision-making circuit that lacks bistability and exploits stochastic noise. *Methods* **53**: 68-77.
Rhee, SY, Fessel, WJ, Liu, TF, Marlowe, NM, Rowland, CM, Rode, RA, Vandamme, AM, Van Laethem, K, Brun-Vezinet, F, Calvez, V, Taylor, J, Hurley, L, Horberg, M, Shafer, RW (2009) Predictive value of HIV-1 genotypic resistance test interpretation algorithms. *J Infect Dis* **200**: 453-463.
Rhee, SY, Gonzales, MJ, Kantor, R, Betts, BJ, Ravela, J, Shafer, RW (2003) Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res* **31**: 298-303.
Richman, DD, Fischl, MA, Grieco, MH, Gottlieb, MS, Volberding, PA, Laskin, OL, Leedom, JM, Groopman, JE, Mildvan, D, Hirsch, MS, et al. (1987) The toxicity of azidothymidine (AZT) in the treatment of patients with AIDS and AIDS-related complex. A double-blind, placebo-controlled trial. *N Engl J Med* **317**: 192-197.
Richman, DD, Havlir, D, Corbeil, J, Looney, D, Ignacio, C, Spector, SA, Sullivan, J, Cheeseman, S, Barringer, K, Pauletti, D (1994a) Nevirapine resistance mutations of human immunodeficiency virus type 1 selected during therapy. *Journal of Virology* **68**: 1660-1666.
Richman, DD, Havlir, D, Corbeil, J, Looney, D, Ignacio, C, Spector, SA, Sullivan, J, Cheeseman, S, Barringer, K, Pauletti, D, et al. (1994b) Nevirapine resistance mutations of human immunodeficiency virus type 1 selected during therapy. *J Virol* **68**: 1660-1666.
Riviere, L, Darlix, JL, Cimarelli, A (2010) Analysis of the viral elements required in the nuclear import of HIV-1 DNA. *J Virol* **84**: 729-739.
Rizzuto, CD, Wyatt, R, Hernandez-Ramos, N, Sun, Y, Kwong, PD, Hendrickson, WA, Sodroski, J (1998) A conserved HIV gp120 glycoprotein structure involved in chemokine receptor binding. *Science* **280**: 1949-1953.
Robbins, GK, De Gruttola, V, Shafer, RW, Smeaton, LM, Snyder, SW, Pettinelli, C, Dubé, MP, Fischl, MA, Pollard, RB, Delapenha, R (2003a) Comparison of sequential three-drug regimens as initial therapy for HIV-1 infection. *New England Journal of Medicine* **349**: 2293-2303.
Robbins, GK, De Gruttola, V, Shafer, RW, Smeaton, LM, Snyder, SW, Pettinelli, C, Dube, MP, Fischl, MA, Pollard, RB, Delapenha, R, Gedeon, L, van der Horst, C, Murphy, RL, Becker, MI, D'Aquila, RT, Vella, S, Merigan, TC, Hirsch, MS (2003b) Comparison of sequential three-drug regimens as initial therapy for HIV-1 infection. *N Engl J Med* **349**: 2293-2303.
Roberts, JD, Bebenek, K, Kunkel, TA (1988) The accuracy of reverse transcriptase from HIV-1. *Science* **242**: 1171-1173.
Robertson, D (2003) US FDA approves new class of HIV therapeutics. *Nature Biotechnology* **21**: 470-471.
Robertson, DL, Anderson, JP, Bradac, JA, Carr, JK, Foley, B, Funkhouser, RK, Gao, F, Hahn, BH, Kalish, ML, Kuiken, C (2000a) HIV-1 nomenclature proposal. *Science* **288**: 55–55.
Robertson, DL, Anderson, JP, Bradac, JA, Carr, JK, Foley, B, Funkhouser, RK, Gao, F, Hahn, BH, Kalish, ML, Kuiken, C, Learn, GH, Leitner, T, McCutchan, F, Osmanov, S, Peeters, M, Pieniazek, D, Salminen, M, Sharp, PM, Wolinsky, S, Korber, B (2000b) HIV-1 nomenclature proposal. *Science* **288**: 55-56.
Rogel, ME, Wu, LI, Emerman, M (1995) The human immunodeficiency virus type 1 vpr gene prevents cell proliferation during chronic infection. *Journal of virology* **69**: 882–888.

Rogers, MF, Thomas, PA, Starcher, ET, Noa, MC, Bush, TJ, Jaffe, HW (1987) Acquired Immunodeficiency Syndrome in Children: Report of the Centers for Disease Control National Surveillance, 1982 to 1985. *Pediatrics* **79**: 1008-1014.
Roques, P, Robertson, DL, Souquière, S, Damond, F, Ayouba, A, Farfara, I, Depienne, C, Nerrienet, E, Dormont, D, Brun-Vézinet, F, Simon, F, Mauclère, P (2002) Phylogenetic Analysis of 49 Newly Derived HIV-1 Group O Strains: High Viral Diversity but No Group M-like Subtype Structure. *Virology* **302**: 259-273.
Rosen, CA, Pavlakis, GN (1990a) Tat and Rev: positive regulators of HIV gene expression. *AIDS* **4**: A51.
Rosen, CA, Pavlakis, GN (1990b) Tat and Rev: positive regulators of HIV gene expression. *AIDS* **4**: 499-509.
Rothberg, JM, Hinz, W, Rearick, TM, Schultz, J, Mileski, W, Davey, M, Leamon, JH, Johnson, K, Milgrew, MJ, Edwards, M, Hoon, J, Simons, JF, Marran, D, Myers, JW, Davidson, JF, Branting, A, Nobile, JR, Puc, BP, Light, D, Clark, TA, Huber, M, Branciforte, JT, Stoner, IB, Cawley, SE, Lyons, M, Fu, Y, Homer, N, Sedova, M, Miao, X, Reed, B, Sabina, J, Feierstein, E, Schorn, M, Alanjary, M, Dimalanta, E, Dressman, D, Kasinskas, R, Sokolsky, T, Fidanza, JA, Namsaraev, E, McKernan, KJ, Williams, A, Roth, GT, Bustillo, J (2011) An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**: 348-352.
Rowley, CF, Boutwell, CL, Lee, EJ, MacLeod, IJ, Ribaudo, HJ, Essex, M, Lockman, S (2010) Ultrasensitive detection of minor drug-resistant variants for HIV after nevirapine exposure using allele-specific PCR: clinical significance. *AIDS Res Hum Retroviruses* **26**: 293-300.
Roy, S, Delling, U, Chen, CH, Rosen, CA, Sonenberg, N (1990) A bulge structure in HIV-1 TAR RNA is required for Tat binding and Tat-mediated trans-activation. *Genes Dev* **4**: 1365-1373.
Rozera, G, Abbate, I, Bruselles, A, Vlassi, C, D'Offizi, G, Narciso, P, Chillemi, G, Prosperi, M, Ippolito, G, Capobianchi, MR (2009) Massively parallel pyrosequencing highlights minority variants in the HIV-1 env quasispecies deriving from lymphomonocyte sub-populations. *Retrovirology* **6**: 15.
Saad, JS, Miller, J, Tai, J, Kim, A, Ghanam, RH, Summers, MF (2006) Structural basis for targeting HIV-1 Gag proteins to the plasma membrane for virus assembly. *Proc Natl Acad Sci U S A* **103**: 11364-11369.
Salemi, M, Strimmer, K, Hall, WW, Duffy, M, Delaporte, E, Mboup, S, Peeters, M, Vandamme, AM (2001) Dating the common ancestor of SIVcpz and HIV-1 group M and the origin of HIV-1 subtypes using a new method to uncover clock-like molecular evolution. *FASEB J* **15**: 276-278.
Salminen, MO, Carr, JK, Burke, DS, McCutchan, FE (1995) Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning. *AIDS Res Hum Retroviruses* **11**: 1423-1425.
Sandstrom, EG, Kaplan, JC (1987) Antiviral therapy in AIDS. Clinical pharmacological properties and therapeutic experience to date. *Drugs* **34**: 372-390.
Sanger, F, Nicklen, S, Coulson, AR (1977a) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences* **74**: 5463-5467.
Sanger, F, Nicklen, S, Coulson, AR (1977b) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**: 5463-5467.
Sanne, I, Orrell, C, Fox, MP, Conradie, F, Ive, P, Zeinecker, J, Cornell, M, Heiberg, C, Ingram, C, Panchia, R, Rassool, M, Gonin, R, Stevens, W, Truter, H, Dehlinger, M, van der Horst, C, McIntyre, J, Wood, R (2010) Nurse versus doctor management of HIV-infected patients receiving antiretroviral therapy (CIPRA-SA): a randomised non-inferiority trial. *Lancet* **376**: 33-40.

Santiago, ML, Range, F, Keele, BF, Li, Y, Bailes, E, Bibollet-Ruche, F, Fruteau, C, Noë, R, Peeters, M, Brookfield, JF (2005) Simian immunodeficiency virus infection in free-ranging sooty mangabeys (Cercocebus atys atys) from the Tai Forest, Cote d'Ivoire: implications for the origin of epidemic human immunodeficiency virus type 2. *Journal of virology* **79**: 12515–12527.
Sarafianos, SG, Hughes, SH, Arnold, E (2004) Designing anti-AIDS drugs targeting the major mechanism of HIV-1 RT resistance to nucleoside analog drugs. *The International Journal of Biochemistry & Cell Biology* **36**: 1706-1715.
Sarafianos, SG, Marchand, B, Das, K, Himmel, DM, Parniak, MA, Hughes, SH, Arnold, E (2009) Structure and Function of HIV-1 Reverse Transcriptase: Molecular Mechanisms of Polymerization and Inhibition. *Journal of Molecular Biology* **385**: 693-713.
Saravolatz, LD, Winslow, DL, Collins, G, Hodges, JS, Pettinelli, C, Stein, DS, Markowitz, N, Reves, R, Loveless, MO, Crane, L (1996) Zidovudine alone or in combination with didanosine or zalcitabine in HIV-infected patients with the acquired immunodeficiency syndrome or fewer than 200 CD4 cells per cubic millimeter. *New England Journal of Medicine* **335**: 1099-1106.
Sawai, ET, Baur, A, Struble, H, Peterlin, BM, Levy, JA, Cheng-Mayer, C (1994) Human immunodeficiency virus type 1 Nef associates with a cellular serine kinase in T lymphocytes. *Proceedings of the National Academy of Sciences* **91**: 1539-1543.
Schmieder, R, Edwards, R (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**: 863-864.
Schmitt, MW, Kennedy, SR, Salk, JJ, Fox, EJ, Hiatt, JB, Loeb, LA (2012) Detection of ultra-rare mutations by next-generation sequencing. *Proc Natl Acad Sci U S A* **109**: 14508-14513.
Schubert, U, Ott, DE, Chertova, EN, Welker, R, Tessmer, U, Princiotta, MF, Bennink, JR, Krausslich, HG, Yewdell, JW (2000) Proteasome inhibition interferes with gag polyprotein processing, release, and maturation of HIV-1 and HIV-2. *Proc Natl Acad Sci U S A* **97**: 13057-13062.
Schuurman, R, Brambilla, D, de Groot, T, Huang, D, Land, S, Bremer, J, Benders, I, Boucher, CA (2002) Underestimation of HIV type 1 drug resistance mutations: results from the ENVA-2 genotyping proficiency program. *AIDS Res Hum Retroviruses* **18**: 243-248.
Schuurman, R, Demeter, L, Reichelderfer, P, Tijnagel, J, de Groot, T, Boucher, C (1999) Worldwide evaluation of DNA sequencing approaches for identification of drug resistance mutations in the human immunodeficiency virus type 1 reverse transcriptase. *J Clin Microbiol* **37**: 2291-2296.
Seelmeier, S, Schmidt, H, Turk, V, von der Helm, K (1988) Human immunodeficiency virus has an aspartic-type protease that can be inhibited by pepstatin A. *Proc Natl Acad Sci U S A* **85**: 6612-6616.
Shafer, RW (2006) Rationale and uses of a public HIV drug-resistance database. *J Infect Dis* **194 Suppl 1**: S51-58.
Shafer, RW, Smeaton, LM, Robbins, GK, De Gruttola, V, Snyder, SW, D'Aquila, RT, Johnson, VA, Morse, GD, Nokta, MA, Martinez, AI, Gripshover, BM, Kaul, P, Haubrich, R, Swingle, M, McCarty, SD, Vella, S, Hirsch, MS, Merigan, TC (2003) Comparison of four-drug regimens and pairs of sequential three-drug regimens as initial therapy for HIV-1 infection. *N Engl J Med* **349**: 2304-2315.
Shah, VB, Shi, J, Hout, DR, Oztop, I, Krishnan, L, Ahn, J, Shotwell, MS, Engelman, A, Aiken, C (2013) The host proteins transportin SR2/TNPO3 and cyclophilin A exert opposing effects on HIV-1 uncoating. *J Virol* **87**: 422-432.
Shapiro, RL, Hughes, MD, Ogwu, A, Kitch, D, Lockman, S, Moffat, C, Makhema, J, Moyo, S, Thior, I, McIntosh, K, van Widenfelt, E, Leidner, J, Powis, K, Asmelash, A,

Tumbare, E, Zwerski, S, Sharma, U, Handelsman, E, Mburu, K, Jayeoba, O, Moko, E, Souda, S, Lubega, E, Akhtar, M, Wester, C, Tuomola, R, Snowden, W, Martinez-Tristani, M, Mazhani, L, Essex, M (2010) Antiretroviral regimens in pregnancy and breast-feeding in Botswana. *N Engl J Med* **362**: 2282-2294.
Sharp, PM, Hahn, BH The evolution of HIV-1 and the origin of AIDS. *Philosophical Transactions of the Royal Society B: Biological Sciences* **365**: 2487-2494.
Sharp, PM, Hahn, BH (2010) The evolution of HIV-1 and the origin of AIDS. *Philosophical Transactions of the Royal Society B: Biological Sciences* **365**: 2487-2494.
Sheehy, AM, Gaddis, NC, Choi, JD, Malim, MH (2002) Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. *Nature* **418**: 646-650.
Shendure, J, Ji, H (2008) Next-generation DNA sequencing. *Nature Biotechnology* **26**: 1135-1145.
Sheward, DJ, Murrell, B, Williamson, C (2012) Degenerate Primer IDs and the birthday problem. *Proc Natl Acad Sci U S A* **109**: E1330; author reply E1331.
Shi, C, Mellors, JW (1997) A recombinant retroviral system for rapid in vivo analysis of human immunodeficiency virus type 1 susceptibility to reverse transcriptase inhibitors. *Antimicrob Agents Chemother* **41**: 2781-2785.
Shokralla, S, Spall, JL, Gibson, JF, Hajibabaei, M (2012) Next-generation sequencing technologies for environmental DNA research. *Mol Ecol* **21**: 1794-1805.
Shrestha, RK, Lubinsky, B, Bansode, VB, Moinz, MB, McCormack, GP, Travers, SA (2014) QTrim: a novel tool for the quality trimming of sequence reads generated using the Roche/454 sequencing platform. *BMC Bioinformatics* **15**: 33.
Simen, BB, Simons, JF, Hullsiek, KH, Novak, RM, MacArthur, RD, Baxter, JD, Huang, C, Lubeski, C, Turenchalk, GS, Braverman, MS, Desany, B, Rothberg, JM, Egholm, M (2009) Low-Abundance Drug-Resistant Viral Variants in Chronically HIV-Infected, Antiretroviral Treatment–Naive Patients Significantly Impact Treatment Outcomes. *Journal of Infectious Diseases* **199**: 693-701.
Simmons, A, Aluvihare, V, McMichael, A (2001) Nef triggers a transcriptional program in T cells imitating single-signal T cell activation and inducing HIV virulence mediators. *Immunity* **14**: 763-777.
Simon, F, Mauclère, P, Roques, P, Loussert-Ajaka, I, Müller-Trutwin, MC, Saragosti, S, Georges-Courbot, MC, Barré-Sinoussi, F, Brun-Vézinet, F (1998a) Identification of a new human immunodeficiency virus type 1 distinct from group M and group O. *Nature Medicine* **4**: 1032-1037.
Simon, JHM, Gaddis, NC, Fouchier, RAM, Malim, MH (1998b) Evidence for a newly discovered cellular anti-HIV-1 phenotype. *Nature Medicine* **4**: 1397-1400.
Skowron, G, Bozzette, SA, Lim, L, Pettinelli, CB, Schaumburg, HH, Arezzo, J, Fischl, MA, Powderly, WG, Gocke, DJ, Richman, DD, Pottage, JC, Antoniskis, D, McKinley, GF, Hyslop, NE, Ray, G, Simon, G, Reed, N, LoFaro, ML, Uttamchandani, RB, Gelb, LD, Sperber, SJ, Murphy, RL, Leedom, JM, Grieco, MH, Zachary, J, Hirsch, MS, Spector, SA, Bigley, J, Soo, W, Merigan, TC (1993) Alternating and intermittent regimens of zidovudine and dideoxycytidine in patients with AIDS or AIDS-related complex. *Ann Intern Med* **118**: 321-330.
Sluis-Cremer, N, Arion, D, Parniak*, MA (2000) Molecular mechanisms of HIV-1 resistance to nucleoside reverse transcriptase inhibitors (NRTIs). *Cellular and Molecular Life Sciences CMLS* **57**: 1408-1422.
Smyth, RP, Davenport, MP, Mak, J (2012) The origin of genetic diversity in HIV-1. *Virus Res* **169**: 415-429.

Sodroski, J, Rosen, C, Wong-Staal, F, Salahuddin, SZ, Popovic, M, Arya, S, Gallo, RC, Haseltine, WA (1985) Trans-acting transcriptional regulation of human T-cell leukemia virus type III long terminal repeat. *Science* **227**: 171-173.
Staszewski, S, Morales-Ramirez, J, Tashima, KT, Rachlis, A, Skiest, D, Stanford, J, Stryker, R, Johnson, P, Labriola, DF, Farina, D (1999a) Efavirenz plus zidovudine and lamivudine, efavirenz plus indinavir, and indinavir plus zidovudine and lamivudine in the treatment of HIV-1 infection in adults. *New England Journal of Medicine* **341**: 1865-1873.
Staszewski, S, Morales-Ramirez, J, Tashima, KT, Rachlis, A, Skiest, D, Stanford, J, Stryker, R, Johnson, P, Labriola, DF, Farina, D, Manion, DJ, Ruiz, NM (1999b) Efavirenz plus zidovudine and lamivudine, efavirenz plus indinavir, and indinavir plus zidovudine and lamivudine in the treatment of HIV-1 infection in adults. Study 006 Team. *N Engl J Med* **341**: 1865-1873.
Strebel, K, Klimkait, T, Martin, MA (1988) A novel gene of HIV-1, vpu, and its 16-kilodalton product. *Science* **241**: 1221-1223.
Stringer, EM, Ekouevi, DK, Coetzee, D, Tih, PM, Creek, TL, Stinson, K, Giganti, MJ, Welty, TK, Chintu, N, Chi, BH, Wilfert, CM, Shaffer, N, Dabis, F, Stringer, JS (2010a) Coverage of nevirapine-based services to prevent mother-to-child HIV transmission in 4 African countries. *JAMA* **304**: 293-302.
Stringer, JS, McConnell, MS, Kiarie, J, Bolu, O, Anekthananon, T, Jariyasethpong, T, Potter, D, Mutsotso, W, Borkowf, CB, Mbori-Ngacha, D, Muiruri, P, Ong'ech, JO, Zulu, I, Njobvu, L, Jetsawang, B, Pathak, S, Bulterys, M, Shaffer, N, Weidle, PJ (2010b) Effectiveness of non-nucleoside reverse-transcriptase inhibitor-based antiretroviral therapy in women previously exposed to a single intrapartum dose of nevirapine: a multi-country, prospective cohort study. *PLoS Med* **7**: e1000233.
Stringer, JS, Zulu, I, Levy, J, Stringer, EM, Mwango, A, Chi, BH, Mtonga, V, Reid, S, Cantrell, RA, Bulterys, M, Saag, MS, Marlink, RG, Mwinga, A, Ellerbrock, TV, Sinkala, M (2006) Rapid scale-up of antiretroviral therapy at primary care sites in Zambia: feasibility and early outcomes. *JAMA* **296**: 782-793.
Struck, D, Wallis, CL, Denisov, G, Lambert, C, Servais, JY, Viana, RV, Letsoalo, E, Bronze, M, Aitken, SC, Schuurman, R, Stevens, W, Schmit, JC, Rinke de Wit, T, Perez Bercoff, D (2012) Automated sequence analysis and editing software for HIV drug resistance testing. *J Clin Virol* **54**: 30-35.
Stuhlmann, H, Berg, P (1992) Homologous recombination of copackaged retrovirus RNAs during reverse transcription. *J Virol* **66**: 2378-2388.
Stuyver, L, Wyseur, A, Rombout, A, Louwagie, J, Scarcez, T, Verhofstede, C, Rimland, D, Schinazi, RF, Rossau, R (1997) Line probe assay for rapid detection of drug-selected mutations in the human immunodeficiency virus type 1 reverse transcriptase gene. *Antimicrob Agents Chemother* **41**: 284-291.
Sungkanuparph, S, Sukasem, C, Kiertiburanakul, S, Pasomsub, E, Chantratita, W (2012) Emergence of HIV-1 drug resistance mutations among antiretroviral-naive HIV-1-infected patients after rapid scaling up of antiretroviral therapy in Thailand. *J Int AIDS Soc* **15**: 12.
Supervie, V, Garcia-Lerma, JG, Heneine, W, Blower, S (2010) HIV, transmitted drug resistance, and the paradox of preexposure prophylaxis. *Proc Natl Acad Sci U S A* **107**: 12381-12386.
Suzuki, S, Ono, N, Furusawa, C, Ying, B-W, Yomo, T (2011) Comparison of Sequence Reads Obtained from Three Next-Generation Sequencing Platforms. *PLoS ONE* **6**: e19534.
Tan, K, Liu, J-h, Wang, J-h, Shen, S, Lu, M (1997) Atomic structure of a thermostable subdomain of HIV-1 gp41. *Proceedings of the National Academy of Sciences* **94**: 12303-12308.

Taylor, BS, Sobieszczyk, ME, McCutchan, FE, Hammer, SM (2008) The Challenge of HIV-1 Subtype Diversity. *New England Journal of Medicine* **358**: 1590-1602.
Tersmette, M, Gruters, RA, Wolf, Fd, Goede, REd, Lange, JM, Schellekens, PT, Goudsmit, J, Huisman, HG, Miedema, F (1989) Evidence for a role of virulent human immunodeficiency virus (HIV) variants in the pathogenesis of acquired immunodeficiency syndrome: studies on sequential HIV isolates. *Journal of Virology* **63**: 2118-2125.
Thompson, SR, Sarnow, P (2000) Regulation of host cell translation by viruses and effects on cell function. *Curr Opin Microbiol* **3**: 366-370.
Tisdale, M, Kemp, SD, Parry, NR, Larder, BA (1993) Rapid in vitro selection of human immunodeficiency virus type 1 resistant to 3'-thiacytidine inhibitors due to a mutation in the YMDD region of reverse transcriptase. *Proc Natl Acad Sci U S A* **90**: 5653-5656.
Tovanabutra, S, Robison, V, Wongtrakul, J, Sennum, S, Suriyanon, V, Kingkeow, D, Kawichai, S, Tanan, P, Duerr, A, Nelson, KE (2002) Male viral load and heterosexual transmission of HIV-1 subtype E in northern Thailand. *J Acquir Immune Defic Syndr* **29**: 275-283.
Tsibris, AM, Korber, B, Arnaout, R, Russ, C, Lo, CC, Leitner, T, Gaschen, B, Theiler, J, Paredes, R, Su, Z, Hughes, MD, Gulick, RM, Greaves, W, Coakley, E, Flexner, C, Nusbaum, C, Kuritzkes, DR (2009) Quantitative deep sequencing reveals dynamic HIV-1 escape and large population shifts during CCR5 antagonist therapy in vivo. *PLoS One* **4**: e5683.
Tsiodras, S, Mantzoros, C, Hammer, S, Samore, M (2000) Effects of protease inhibitors on hyperglycemia, hyperlipidemia, and lipodystrophy: a 5-year cohort study. *Arch Intern Med* **160**: 2050-2056.
UNAIDS (2012) Global Report 2012: UNAIDS Report on the Global AIDS Epidemic. ebookpartnership. com.
Vacca, JP, Dorsey, BD, Schleif, WA, Levin, RB, McDaniel, SL, Darke, PL, Zugay, J, Quintero, JC, Blahy, OM, Roth, E (1994) L-735,524: an orally bioavailable human immunodeficiency virus type 1 protease inhibitor. *Proceedings of the National Academy of Sciences* **91**: 4096-4100.
Vacca, JP, Guare, JP, deSolms, SJ, Sanders, WM, Giuliani, EA, Young, SD, Darke, PL, Zugay, J, Sigal, IS, Schleif, WA, et al. (1991) L-687,908, a potent hydroxyethylene-containing HIV protease inhibitor. *J Med Chem* **34**: 1225-1228.
Vallari, A, Holzmayer, V, Harris, B, Yamaguchi, J, Ngansop, C, Makamche, F, Mbanya, D, Kaptué, L, Ndembi, N, Gürtler, L, Devare, S, Brennan, CA (2011) Confirmation of Putative HIV-1 Group P in Cameroon. *Journal of Virology* **85**: 1403-1407.
Van Laethem, K, Van Vaerenbergh, K, Schmit, JC, Sprecher, S, Hermans, P, De Vroey, V, Schuurman, R, Harrer, T, Witvrouw, M, Van Wijngaerden, E, Stuyver, L, Van Ranst, M, Desmyter, J, De Clercq, E, Vandamme, AM (1999) Phenotypic assays and sequencing are less sensitive than point mutation assays for detection of resistance in mixed HIV-1 genotypic populations. *J Acquir Immune Defic Syndr* **22**: 107-118.
van Leeuwen, R, Katlama, C, Kitchen, V, Boucher, CA, Tubiana, R, McBride, M, Ingrand, D, Weber, J, Hill, A, McDade, H, et al. (1995) Evaluation of safety and efficacy of 3TC (lamivudine) in patients with asymptomatic or mildly symptomatic human immunodeficiency virus infection: a phase I/II study. *J Infect Dis* **171**: 1166-1171.
van Leeuwen, R, Katlama, C, Murphy, RL, Squires, K, Gatell, J, Horban, A, Clotet, B, Staszewski, S, van Eeden, A, Clumeck, N, Moroni, M, Pavia, AT, Schmidt, RE, Gonzalez-Lahoz, J, Montaner, J, Antunes, F, Gulick, R, Banhegyi, D, van der Valk,

M, Reiss, P, van Weert, L, van Leth, F, Johnson, VA, Sommadossi, JP, Lange, JM (2003) A randomized trial to study first-line combination therapy with or without a protease inhibitor in HIV-1-infected patients. *AIDS* **17**: 987-999.
van Leth, F, Phanuphak, P, Ruxrungtham, K, Baraldi, E, Miller, S, Gazzard, B, Cahn, P, Lalloo, UG, van der Westhuizen, IP, Malan, DR, Johnson, MA, Santos, BR, Mulcahy, F, Wood, R, Levi, GC, Reboredo, G, Squires, K, Cassetti, I, Petit, D, Raffi, F, Katlama, C, Murphy, RL, Horban, A, Dam, JP, Hassink, E, van Leeuwen, R, Robinson, P, Wit, FW, Lange, JM (2004) Comparison of first-line antiretroviral therapy with regimens including nevirapine, efavirenz, or both drugs, plus stavudine and lamivudine: a randomised open-label trial, the 2NN Study. *Lancet* **363**: 1253-1263.
van Sighem, AI, Gras, LA, Reiss, P, Brinkman, K, de Wolf, F (2010) Life expectancy of recently diagnosed asymptomatic HIV-infected patients approaches that of uninfected individuals. *AIDS* **24**: 1527-1535.
Van Vaerenbergh, K (2001) Study of the impact of HIV genotypic drug resistance testing on therapy efficacy. *Verhandelingen - Koninklijke Academie voor Geneeskunde van België* **63**: 447-473.
VANDEN HAESEVELDE, MM, Peeters, M, JANNES, G, JANSSENS, W, VAN DER GROEN, G, SHARP, PM, SAMAN, E (1996) Sequence analysis of a highly divergent HIV-1-related lentivirus isolated from a wild captured chimpanzee. *Virology* **221**: 346–350.
Varghese, V, Shahriar, R, Rhee, SY, Liu, T, Simen, BB, Egholm, M, Hanczaruk, B, Blake, LA, Gharizadeh, B, Babrzadeh, F, Bachmann, MH, Fessel, WJ, Shafer, RW (2009) Minority variants associated with transmitted and acquired HIV-1 nonnucleoside reverse transcriptase inhibitor resistance: implications for the use of second-generation nonnucleoside reverse transcriptase inhibitors. *J Acquir Immune Defic Syndr* **52**: 309-315.
Vella, S, Schwartlander, B, Sow, SP, Eholie, SP, Murphy, RL (2012) The history of antiretroviral therapy and of its implementation in resource-limited areas of the world. *AIDS* **26**: 1231-1241.
Vera, JC, Wheat, CW, Fescemyer, HW, Frilander, MJ, Crawford, DL, Hanski, I, Marden, JH (2008) Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Mol Ecol* **17**: 1636-1647.
Volberding, PA, Lagakos, SW, Grimes, JM, Stein, DS, Rooney, J, Meng, TC, Fischl, MA, Collier, AC, Phair, JP, Hirsch, MS, et al. (1995) A comparison of immediate with deferred zidovudine therapy for asymptomatic HIV-infected adults with CD4 cell counts of 500 or more per cubic millimeter. AIDS Clinical Trials Group. *N Engl J Med* **333**: 401-407.
Volberding, PA, Lagakos, SW, Koch, MA, Pettinelli, C, Myers, MW, Booth, DK, Balfour, HH, Jr., Reichman, RC, Bartlett, JA, Hirsch, MS, et al. (1990) Zidovudine in asymptomatic human immunodeficiency virus infection. A controlled trial in persons with fewer than 500 CD4-positive cells per cubic millimeter. The AIDS Clinical Trials Group of the National Institute of Allergy and Infectious Diseases. *N Engl J Med* **322**: 941-949.
von Schwedler, U, Song, J, Aiken, C, Trono, D (1993) Vif is crucial for human immunodeficiency virus type 1 proviral DNA synthesis in infected cells. *Journal of virology* **67**: 4945-4955.
Wain-Hobson, S, Sonigo, P, Danos, O, Cole, S, Alizon, M (1985) Nucleotide sequence of the AIDS virus, LAV. *Cell* **40**: 9–17.
Wang, C, Mitsuya, Y, Gharizadeh, B, Ronaghi, M, Shafer, RW (2007) Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. *Genome Res* **17**: 1195-1201.

Wang, CT, Barklis, E (1993) Assembly, processing, and infectivity of human immunodeficiency virus type 1 gag mutants. *J Virol* **67**: 4264-4273.
Wang, R-R, Yang, L-M, Wang, Y-H, Pang, W, Tam, S-C, Tien, P, Zheng, Y-T (2009) Sifuvirtide, a potent HIV fusion inhibitor peptide. *Biochemical and Biophysical Research Communications* **382**: 540-544.
Ward, JW, Grindon, AJ, Feorino, PM, Schable, C, Parvin, M, Allen, JR (1986) Laboratory and epidemiologic evaluation of an enzyme immunoassay for antibodies to HTLV-III. *JAMA* **256**: 357-361.
Wasi, C, Herring, B, Raktham, S, Vanichseni, S, Mastro, TD, Young, NL, Rubsamen-Waigmann, H, von Briesen, H, Kalish, ML, Luo, CC, et al. (1995) Determination of HIV-1 subtypes in injecting drug users in Bangkok, Thailand, using peptide-binding enzyme immunoassay and heteroduplex mobility assay: evidence of increasing infection with HIV-1 subtype E. *AIDS* **9**: 843-849.
Weinberg, JL, Kovarik, CL (2010) The WHO Clinical Staging System for HIV/AIDS. *Virtual Mentor* **12**: 202-206.
Westby, M, van der Ryst, E (2005) CCR5 antagonists: host-targeted antivirals for the treatment of HIV infection. *Antivir Chem Chemother* **16**: 339-354.
WHO (2008) Towards universal access: scaling up priority HIV/AIDS interventions in the health sector: progress report 2008.
WHO, U (2011) UNAIDS. 2011. Global HIV/AIDS response: Epidemic update and health sector progress towards universal access (Progress Report 2011). *World Health Organization, Geneva, Switzerland*.
Wiegers, K, Rutter, G, Kottler, H, Tessmer, U, Hohenberg, H, Krausslich, HG (1998) Sequential steps in human immunodeficiency virus particle maturation revealed by alterations of individual Gag polyprotein cleavage sites. *J Virol* **72**: 2846-2854.
Wild, C, Greenwell, T, Matthews, T (1993) A synthetic peptide from HIV-1 gp41 is a potent inhibitor of virus-mediated cell-cell fusion. *AIDS research and human retroviruses* **9**: 1051-1053.
Willey, RL, Maldarelli, F, Martin, MA, Strebel, K (1992) Human immunodeficiency virus type 1 Vpu protein induces rapid degradation of CD4. *Journal of Virology* **66**: 7193-7200.
Wittkop, L, Gunthard, HF, de Wolf, F, Dunn, D, Cozzi-Lepri, A, de Luca, A, Kucherer, C, Obel, N, von Wyl, V, Masquelier, B, Stephan, C, Torti, C, Antinori, A, Garcia, F, Judd, A, Porter, K, Thiebaut, R, Castro, H, van Sighem, AI, Colin, C, Kjaer, J, Lundgren, JD, Paredes, R, Pozniak, A, Clotet, B, Phillips, A, Pillay, D, Chene, G (2011) Effect of transmitted drug resistance on virological and immunological response to initial combination antiretroviral therapy for HIV (EuroCoord-CHAIN joint project): a European multicohort study. *Lancet Infect Dis* **11**: 363-371.
Wofsy, C, Hauer, L, Michaelis, B, Cohen, J, Padian, N, Evans, L, Levy, J (1986) Isolation of AIDS-associated retrovirus from genital secretions of women with antibodies to the virus. *The Lancet* **327**: 527–529.
Woods, CK, Brumme, CJ, Liu, TF, Chui, CK, Chu, AL, Wynhoven, B, Hall, TA, Trevino, C, Shafer, RW, Harrigan, PR (2012) Automating HIV drug resistance genotyping with RECall, a freely accessible sequence analysis tool. *J Clin Microbiol* **50**: 1936-1942.
Worobey, M, Gemmel, M, Teuwen, DE, Haselkorn, T, Kunstman, K, Bunce, M, Muyembe, JJ, Kabongo, JM, Kalengayi, RM, Van Marck, E, Gilbert, MT, Wolinsky, SM (2008) Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature* **455**: 661-664.
Wu, L, Gerard, NP, Wyatt, R, Choe, H, Parolin, C, Ruffing, N, Borsetti, A, Cardoso, AA, Desjardin, E, Newman, W, Gerard, C, Sodroski, J (1996) CD4-induced

interaction of primary HIV-1 gp120 glycoproteins with the chemokine receptor CCR-5. *Nature* **384**: 179-183.
Wu, L, LaRosa, G, Kassam, N, Gordon, CJ, Heath, H, Ruffing, N, Chen, H, Humblias, J, Samson, M, Parmentier, M, Moore, JP, Mackay, CR (1997) Interaction of Chemokine Receptor CCR5 with its Ligands: Multiple Domains for HIV-1 gp120 Binding and a Single Domain for Chemokine Binding. *The Journal of Experimental Medicine* **186**: 1373-1381.
Yang, YL, Wang, G, Dorman, K, Kaplan, AH (1996) Long polymerase chain reaction amplification of heterogeneous HIV type 1 templates produces recombination at a relatively high frequency. *AIDS Res Hum Retroviruses* **12**: 303-306.
Yerly, S, Kaiser, L, Race, E, Bru, J-P, Clavel, F, Perrin, L (1999) Transmission of antiretroviral-drug-resistant HIV-1 variants. *The Lancet* **354**: 729.
Yin, L, Liu, L, Sun, Y, Hou, W, Lowe, AC, Gardner, BP, Salemi, M, Williams, WB, Farmerie, WG, Sleasman, JW, Goodenow, MM (2012) High-resolution deep sequencing reveals biodiversity, population structure, and persistence of HIV-1 quasispecies within host ecosystems. *Retrovirology* **9**: 108.
Zagordi, O, Klein, R, Daumer, M, Beerenwinkel, N (2010) Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. *Nucleic Acids Res* **38**: 7400-7409.
Zaidi, J, Grapsa, E, Tanser, F, Newell, M-L, Bärnighausen, T (2013) Dramatic increases in HIV prevalence after scale-up of antiretroviral treatment: a longitudinal population-based HIV surveillance study in rural kwazulu-natal. *AIDS* **27**: 000-000.
Zhang, J, Hou, T, Wang, W, Liu, JS (2010a) Detecting and understanding combinatorial mutation patterns responsible for HIV drug resistance. *Proc Natl Acad Sci U S A* **107**: 1321-1326.
Zhang, M, Foley, B, Schultz, AK, Macke, JP, Bulla, I, Stanke, M, Morgenstern, B, Korber, B, Leitner, T (2010b) The role of recombination in the emergence of a complex and dynamic HIV epidemic. *Retrovirology* **7**: 25.
Zhang, Y, Qian, H, Love, Z, Barklis, E (1998) Analysis of the Assembly Function of the Human Immunodeficiency Virus Type 1 Gag Protein Nucleocapsid Domain. *Journal of Virology* **72**: 1782-1789.
Zhou, W, Resh, MD (1996) Differential membrane binding of the human immunodeficiency virus type 1 matrix protein. *J Virol* **70**: 8540-8548.
Zhu, T, Korber, BT, Nahmias, AJ, Hooper, E, Sharp, PM, Ho, DD (1998) An African HIV-1 sequence from 1959 and implications for the origin of the epidemic. *Nature* **391**: 594-597.
Ziegler, J, Johnson, R, Cooper, D, Gold, J (1985) Postnatal transmission of AIDS-associated retrovirus from mother to infant. *The Lancet* **325**: 896–898.
Zolfo, M, De Weggheleire, A, Schouten, E, Lynen, L (2010) Time for "test and treat" in prevention of mother-to-child transmission programs in low- and middle-income countries. *J Acquir Immune Defic Syndr* **55**: 287-289.
Zolopa, AR, Shafer, RW, Warford, A, Montoya, JG, Hsu, P, Katzenstein, D, Merigan, TC, Efron, B (1999) HIV-1 genotypic resistance patterns predict response to saquinavir-ritonavir therapy in patients in whom previous protease inhibitor therapy had failed. *Ann Intern Med* **131**: 813-821.