From Department of Microbiology, Tumor and Cell Biology
Karolinska Institutet, Stockholm, Sweden

DECIPHERING HIV GENETIC VARIABILITY AND EVOLUTION BY MASSIVE PARALLEL PYROSEQUENCING AND BIOINFORMATICS

Johanna Brodin

Karolinska Institutet

Stockholm 2014

Published by Karolinska Institutet. Printed by åtta.45

ISBN 978-91-7549-562-0

ABSTRACT

HIV-1 is a virus with a very variable genome and therefore has the ability to adapt to new environments which include escape from immune pressure and suboptimal antiretroviral treatment. Next-generation sequencing (NGS), especially ultra-deep pyrosequencing (UDPS), has enabled in-depth sequencing studies with a previously unattainable resolution. However, the technology is more error prone than traditional sequencing which makes it challenging to interpret UDPS results. In this thesis we carried out comprehensive work to identify, characterize and reduce errors as well as investigate the UDPS performance (**Papers II**, **III** and **IV).** In **Papers IV** and **V** we used UDPS to study HIV-1 minority variants. Novel primer design software was developed in **Paper I** and a new method to tag molecules was developed and evaluated in **Paper VI.** The design of primers is of special importance in NGS to avoid selective amplification which may skew estimates of variant frequencies. We developed a computer program, PrimerDesign, to meet the changed requirements for primer design. PrimerDesign is tailored to design primers from a multiple alignment and is suitable for all types of NGS that is preceded by amplification. The new Primer ID methodology has the potential to provide highly accurate deep sequencing. We identified three major challenges; a skewed resampling of Primer IDs, low recovery of templates and erroneous consensus sequences. Undetected this would lead to an underestimation in diversity of the quasispecies and cause a skewed and incorrect results. As many of our other findings, the methodology is not limited to HIV or virology.

The resolution of UDPS analysis is primarily determined by the number of input DNA templates, the error frequency of the method and the efficiency of data cleaning. In **Papers II** and **IV** we therefore optimized the pre-UDPS protocol and investigated the characteristics and sources of errors that occurred when UDPS was used to sequence a fragment of the HIV-1 *pol* gene. UDPS introduced indel errors located in homopolymeric regions that were removed by our in-house data cleaning software. The remaining errors were primarily substitution errors that were introduced in the PCR that preceded UDPS. Transitions were significantly more frequent than transversions, which will limit detection of minor variants and mutations in HIV-1 as well as other species. Further, we evaluated the quality and reproducibility of the UDPS technology in analysis of the same *pol* gene fragment. We concluded that the UDPS repeatability was good for both major and minor variants. In our experimental settings, *in vitro* recombination and sequencing directions posed a minor problem, but still needs to be considered especially for studies of minor viral variants and linkage between mutations.

Minority resistance mutations have been shown to impact the clinical outcome in treated patients. We examined the presence of pre-existing drug resistance mutations in treatment-naïve HIV-1 infected individuals and found very low levels of M184I, T215A and T215I, but no presence of M184V, Y181C, Y188C or T215Y/F. This indicates that the natural occurrence of these mutations is very low. When the same individuals experienced treatment failure or interruption, almost 100 % of the wild-type virus respective drug resistance variants were replaced. Other patients were followed from primary HIV infection (PHI) until their virus switched coreceptor use from CCR5 (R5) to CXCR4 (X4). We did not find any X4-using virus present as a minority population during PHI. The results indicate that the X4-using population most probably evolved in stepwise fashion from the R5-using populations in each of the three patients.

In conclusion, we have developed and used new NGS and bioinformatic methods to study HIV-1 genetic variation. We have shown that UDPS can be used to gain new insights in HIV evolution and to detect minority drug resistance mutations as well as minority variants.

LIST OF PUBLICATIONS

I. **Brodin J**, Krishnamoorthy M, Athreya G, Fischer W, Hraber P, Gleasner C, Green L, Korber B, Leitner T. A multiple-alignment based primer design algorithm for genetically highly variable DNA targets. BMC Bioinformatics. 2013 Aug 21;14:255.

II. **Brodin J**, Mild M, Hedskog C, Sherwood E, Leitner T, Andersson B, Albert J. PCR-induced transitions are the major source of error in cleaned ultra-deep pyrosequencing data. PLoS One. 2013 Jul 23;8(7):e70388.

III. Hedskog C, **Brodin J**, Heddini A, Bratt G, Albert J, Mild M. Longitudinal ultradeep characterization of HIV type 1 R5 and X4 subpopulations in patients followed from primary infection to coreceptor switch. AIDS Res Hum Retroviruses. 2013 Sep;29(9):1237-44.

IV. Mild M, Hedskog C, **Jernberg J**, Albert J. Performance of ultra-deep pyrosequencing in analysis of HIV-1 pol gene variation. PLoS One. 2011;6(7):e22741.

V. Hedskog C, Mild M, **Jernberg J**, Sherwood E, Bratt G, Leitner T, Lundeberg J, Andersson B, Albert J. Dynamics of HIV-1 quasispecies during antiviral treatment dissected using ultra-deep pyrosequencing. PLoS One. 2010 Jul 7;5(7):e11345.

VI. **Brodin J***, Hedskog C*, Heddini A, Benard E, Neher R, Mild M, Albert J. Challenges with using Primer IDs to improve accuracy of next generation sequencing. Manuscript.

CONTENTS

LIST OF ABBREVIATIONS

3TC	Lamivudine
AIDS	acquired immunodeficiency syndrome
APOBEC	Apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like
ART	antiretroviral therapy
ARVs	antiretrovirals
AZT	Azidothymidine
cART	combinational antiretroviral therapy
CCR5	C-C chemokine receptor type 5
CD4 cell	CD4+ T-lymphocyte
cDNA	complementary DNA
CRF	circulating recombinant form
CXCR4	C-X-C chemokine receptor type 4
DNA	deoxyribonucleic acid
EMA	European Medicines Agency
emPCR	emulsion based PCR
Env	Envelope
FDA	Food and Drug association
FPR	false positive rate
Gag	Group specific antigen
GALT	gut-associated lymphoid tissue
Gp	glycoprotein
HAART	highly active antiretroviral therapy
HCV	hepatitis C virus
HIV	human immunodeficiency virus
HIV-1	HIV type 1
HIV-2	HIV type 2
IN	integrase (IN)
Indel	insertion and deletion
LTR	long terminal repeat
MSM	men who have sex with men
Nef	negative factor
NGS	next-generation sequencing
NJ	neighbor joining
NNRTI	non-nucleoside reverse transcriptase inhibitor
NRTI	nucleoside reverse transcriptase inhibitor
PCR	polymerase chain reaction
PHI	primary HIV infection
PIs	protease inhibitors
Pol	Polymerase
PR	protease (PR)
R5-using	HIV variant using the CCR5 coreceptor
RNA	ribonucleic acid

ROI	region of interest
RRE	rev responsible element
RT	reverse transcription (RT)
SGS	single genome sequencing
SIVs	simian immunodeficiency viruses
ssRNA	single stranded RNA molecule
TAR	transactivation response element
TDR	transmitted drug resistance
UDPS	ultra-deep pyrosequencing
V3	variable loop 3
Vif	virion infectivity factor
Vpr	viral protein R
Vpu	viral protein U
X4-using	HIV variant using the CXCR4 coreceptor

1 INTRODUCTION

1.1 HUMAN IMMUNODEFICIENCY VIRUS

1.1.1 History

In 1981 came the first alarming reports of young men experiencing unusual opportunistic infections and rare malignancies [1] [2]. Over three decades has now passed since these first reports. The causative agent leading to acquired immunodeficiency syndrome (AIDS) has been found [3-5] and is referred to as human immunodeficiency virus (HIV). The virus is globally spread and has to date infected in excess of 60 million individuals and caused over 25 million deaths. The pandemic and the disease is far from over and more than 35 million people are living with HIV infection today [6],

1.1.2 Origin and classification

AIDS is in fact caused by two viruses, HIV type 1 (HIV-1) and HIV type 2 (HIV-2). These are morphologically similar but genetically and antigenically distinct [7]. HIV-1 is much more widespread, more infectious and causes a faster progression to AIDS than HIV-2 [8, 9]. This thesis is primarily focused on HIV-1. HIV-1 is a part of the lentivirus genus and belongs to the *Retroviridae* family. It comprises four distinct lineages, termed groups (M) (main), N (non-M-non-O), O (Outlier), and P. Each group is the result of an independent cross-species transmission event of simian immunodeficiency viruses (SIVs) [7]. SIVs are naturally infecting African primates [10]. Group M originates from SIVcpz, a virus that infects two of four subspecies of chimpanzees. Group M is by far the most prevalent group and completely dominates the global pandemic. The transmission event that founded the M group is estimated to have occurred in southeastern Cameroon around 1910 [11]. As Group M spread, time and geographical dispersal caused the virus to evolve into different lineages. Group M is therefore divided into nine pure subtypes (A, B, C, D, F, G, H, J and K) and many (currently 58 known) circulating recombinant forms (CRFs) [12, 13]. Groups N, O and P represent less than 1 % of the infections and are very regionally located [14-18]. The same is true for HIV-2.

1.1.3 The current HIV epidemic

According to WHO´s and UNAIDS´ estimations, between 32.2 and 38.8 million individuals were living with HIV infections in 2012. In 2012, 2.3 million individuals became infected by HIV and 1.6 million individuals died due to AIDS. The infection is not evenly distributed over the world. In Sub-Saharan Africa, 25 million individuals are estimated to live with HIV and in some countries like Botswana and Swaziland the HIV prevalence is well over 20 % in the adult population [6].

Figure 1. Adults and children estimated to be living with HIV globally in 2012. Adapted from [6].

In 1983, the first known HIV-infection in Sweden was reported. Until June 2013, approximately 10,500 individuals have been diagnosed with the infection [19]. At that time approximately 6,200 HIV infected individuals were known to be living in Sweden, which corresponds to a prevalence of 0.06 %. In 2012, 441 new infections were reported. This corresponds roughly to the average incidence of newly infected patients per year in the preceding decade. The majority of infections (51 %) were heterosexual acquired and 70 % stated that they had been infected abroad [20]. Of patients infected abroad, 79 % were born abroad and infected prior to the first arrival to Sweden. Of the 117 domestic transmissions, 56 % occurred between men who have sex with men (MSM). The number of MSM transmissions has increased significant since 2003 [20].

1.2 HIV-1 VIROLOGY

1.2.1 Structure, genes and regulatory enzymes

The HIV-1 particle is enveloped, spherical with a diameter of approximately 120 nm. The envelope is obtained when the virus buds from the host cell and consists of a lipid layer derived from the cell membrane and viral trimeric transmembrane glycoprotein (gp41) linked to the outer trimeric glycoprotein (gp120). The viral envelope surrounds the nucleocapsid, which contains the viral enzymes: reverse transcription (RT), protease (PR) and integrase (IN), as well as two positive sensed single stranded RNA molecules (ssRNA). Each of the RNA strands consists of approximately 9,700 bases.

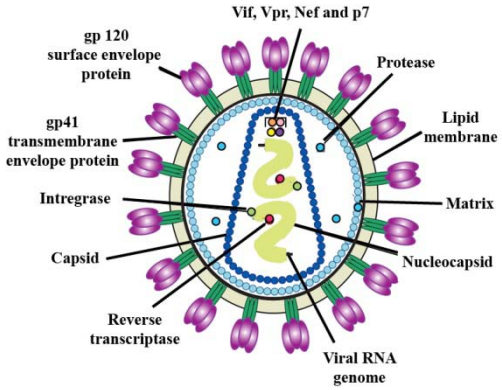

Figure 2. Schematic structure of the HIV-1 virion.

Like all retroviruses, the HIV genome contains the *gag*, *env* and *pol* genes. They encode the major structural and enzymatic proteins; group specific antigen (Gag), polymerase (Pol), and envelope (Env). The *gag*-gene encodes the capsid proteins. The Gag precursor is the p55 protein which is processed to p17 (matrix), p24 (capsid), p7 (nucleocapsid), and p6 proteins, by the viral protease. The genomic region of *pol* encodes the viral enzymes RT, PR and IN. The *env*-gene encodes the polyprotein which is cleaved into the outer gp120 and the transmembrane gp41. The genome also codes for two regulatory proteins, Tat and Rev and four accessory proteins Vif, Vpr, Vpu and Nef.

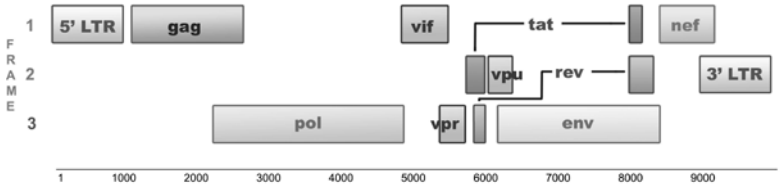

Figure 3. The genome organization of HIV-1

1.2.2 Replication

The HIV-1 replication cycle begins with the virus attaching to the target cell via binding of the virus envelope protein gp120 to the primary cellular receptor, the CD4 protein [21, 22]. CD4 is found on CD4+ T-lymphocytes (CD4 cells), macrophages, monocytes, dendritic cells and brain microglia. The envelope protein binding induces a conformational change allowing the envelope to bind to a coreceptor, which is either of the chemokine receptors CCR5 (R5-using virus) or CXCR4 (X4-using virus). Some viruses, however, use both coreceptors (R5X4-using or dual tropic virus) [23] (more on this in section 1.4.2). Thereafter, fusion of the host membrane and the viral envelope is mediated by a second conformational change which is unlocked by the coreceptor binding whereby the viral nucleocapsid is delivered into the host cell cytoplasm [24].

Following the host cell's reception of the viral contents, the capsid is partially opened and the enzyme RT starts the reverse transcription of one ssRNA strand and generates a cDNA strand with its reverse transcription activity. The RNase H activity of the RT degrades the viral RNA template at the same time. Both RNA strands are needed to complete the cDNA synthesis, in part because the long terminal repeats (LTRs) at both ends of the genome are extended. Genetic variability in form of mutations occurs during the reverse transcription since the RT enzyme is error prone and lacks a proof reading mechanism. Another factor that contributes to the genetic variability is RT's ability to switch between the two RNA strands which create a hybrid cDNA strand if the virus particle contains two genetically distinct RNA molecules as a result of dual infection of the cell from which the virus was produced. A complementary DNA strand is synthesized by the DNA polymerase activity of RT. A pre-integration complex termed PIC is created and transports the dsDNA molecule into the nucleus where IN catalyzes the integration of the viral genome into the host genome. At this stage of the process, the viral DNA genome is referred to as a provirus which can either directly continue the replication cycle or (much more rarely) enter a latent stage. In case the provirus stays active, the next step is transcription performed by the RNA polymerase II of the host cell. The first viral transcript is a full length RNA copy which is spliced into small mRNAs and translated to the early viral proteins Nef, Tat and Rev. Tat interacts with the transactivation response element (TAR) in 5'-end of the HIV mRNA

to promote efficient viral mRNA elongation. Rev binds to the rev responsible element (RRE) in the *env*-region of the viral mRNA, which induces a switch from synthesis of early to late viral proteins by promoting transport of unspliced and partially spliced RNA from the nucleus into the cytoplasm.

The late transcription involves production of longer mRNAs by alternative splicing. The proteins Gag, Gag-Pol, Env, Vif, Vpr and Vpu are transcribed together with full-length mRNA. All mRNAs are translated by the cellular host translation processes in the cytoplasm. The assembly of the components of new virus particles, i.e. structural proteins, viral enzymes and genomic RNA, takes place at the cellular membrane. The new viruses then bud from the cell taking a part of the host cell's lipid layer with it to form the envelope. When the immature virus particle has left the cell, it matures after PR cleaves the Gag and Gag-Pol polyproteins into functional proteins forming the matrix, capsid and nucleocapsid proteins (Gag) as well as the viral enzymes (Gag-Pol). Following these last steps, the virus is ready to infect new cells.

1.2.3 Genetic variability

HIV-1 displays very high genetic variability and is ranked one of the most rapidly evolving organisms known [25]. The genetic diversity found at a single time point in a single infected individual exceeds the global variation in influenza isolates in an entire season [26]. This enormous variation allows the virus to evolve and escape both the immune pressure and suboptimal antiretroviral therapy. The genetic viral variants constituting the populations are called haplotypes, and these haplotypes form a viral quasispecies [27-29]. Several factors contribute to this effect, for example a high turn-over rate, the error-prone reverse transcriptase and high potential for recombination.

In the chronic stage of infection, in an untreated individual, one ml of plasma contains on average 10^4-10^5 or more HIV-particles. The generation time is short (the average replication time is ~1-2 days [30]) and the production rate of new virions is high which results in the production of approximately 10^{10} new virions per day in patients who are not receiving antiretroviral therapy (ART) [31-33].

Single nucleotide substitutions (point mutations) are spontaneously generated as the virus replicates. These are primarily caused by the error-prone reverse transcription process. Mutations also occur when the DNA is transcribed by the host RNA polymerase II and when G-to-A mutations are mediated by the cellular antiretroviral enzyme APOBEC3G (or APOBEC3F). There is no consensus on the relative contribution of these processes, but together they generate an average of 3.4×10^{-5} mutations per nucleotide synthesized [34-37]. Since the HIV-1 genome is approximately 10,000 nucleotides long, this means that every third newly synthesized HIV genome contains a point mutation. Furthermore, the combination of the high mutation rate and the high virus production rate means that every possible single point mutation in the HIV genome arises spontaneously many times every day. These point mutations occur more or less randomly [38], but with a transition vs. transversion bias. G-to-A transitions are especially common, possibly as a result of APOBEC editing. Insertions and deletions (indels) of one or several nucleotides are also created during reverse transcription and contribute to the genetic variation [39]. Finally, recombination is a third source of genetic variation. Recombination arises because the RT enzyme switches between the two ssRNA molecules in the virus particle when the DNA copy is created in the newly infected cell. Thus, the DNA copy will always be a recombinant, but this usually has little consequence because the two RNA molecules in the incoming

virus particle usually are nearly identical. However, if two genetically distinct HIV variants infect the same cell, the viruses that are produced from this cell may be "heterozygous", i.e. contain two genetically distinct RNA copies. If such a virus infects a new cell the DNA copy will be a mosaic with bits from both RNA variants. The effective recombination rate (e.g. the creation of genetically distinct variant) has been estimated to 1.4×10^{-5} recombination per site and generation [40].

New viral variants that arise through point mutations, indels and recombination are continuously screened for their fitness. There exist several different definitions of viral fitness, but here I refer to fitness as the virus ability to produce progeny, which depends on the ability of the virus to perform all steps in the replication cycle as well as the ability to adapt to the surrounding environment such as challenges posed by the immune system of the host and ART. The process by which mutations are maintained to the next replication cycle, and eventually becomes fixed, is a combination of selection and chance. Viral variants that carry advantageous mutations, i.e. mutations that increase virus fitness, will tend to increase in frequency. This is referred to as positive or Darwinian selection. The frequency of variants with disadvantageous mutations will tend to decrease and some mutations may even be directly lethal. This is referred to as negative or purifying selection. Neutral mutations will either become fixed or disappear depending on chance. However, chance will also affect the fate of moderately advantageous and disadvantageous mutations as described by Kimura in his model of neutral evolution [41].

1.3 HIV-1 INFECTION

1.3.1 Pathogenesis

The clinical stages of the infection can be monitored through clinical symptoms and the levels of CD4 cells and virus particles in the blood as well as through many other clinical, immunological and viral markers.

The period after the virus has been transmitted and the infection has been established is referred to as the eclipse period. The eclipse period lasts until the virus can be detected in blood which usually takes 7-21 days. During the eclipse period, the infected individual is asymptomatic and the virus spreads from the initial sites of replication in mucosa and local lymphatic tissue to other replication sites, primarily lymphatic tissue throughout the body.

About 50 -70 % of the infected individuals experience clinical manifestations in the acute phase [42]. The phase is also referred to as primary HIV infection (PHI) (~2-4 weeks). Symptomatic patients suffer from a flulike illness characterized by fever, sore throat, lymphadenopathy and rash [43]. Once the virus becomes detectable in blood plasma, it increases exponentially reaching 10^7 or more copies of viral RNA/ml blood [44]. The high level is a result from absence of the early immune response and rapid replication in gut-associated lymphoid tissue (GALT) and peripheral lymphoid tissue compartments [45-48]. The CD4 cells temporary decline in blood but partially recover following a rapid decline of plasma RNA levels and the emergence of immune response fighting the infection [30, 49]. Fiebig and colleagues has classified the acute and early stage of HIV-infection based on the presentation of different biomarkers [50].

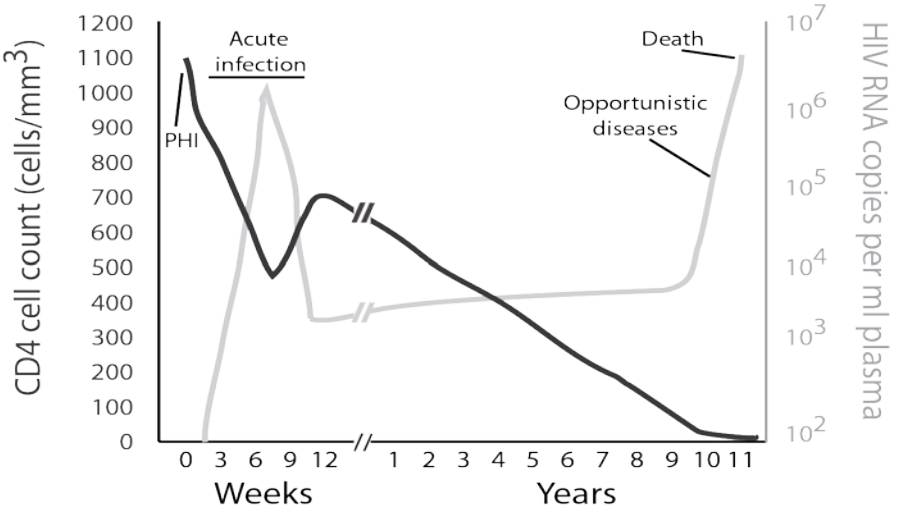

Figure 4. Typical course of HIV infection. Patterns of $CD4^+$ T-cell decline and virus load increase vary greatly between individuals.

The chronic phase (~1 - 20 years) of HIV is usually asymptomatic for the infected individual. During the chronic phase, the virus levels in plasma reaches a semi-steady state (the viral set-point) well below the levels during its peak in the acute phase (usually 1,000 - 100,000 RNA copies/µl). The plasma HIV RNA levels remain constant or slowly increasing whereas the CD4 cells slowly decrease [30].

AIDS is the end stage of the HIV infection and develops when the CD4 cells have declined so that immune system cannot control the HIV infection as well as other (opportunistic) infections and tumors. This occurs when CD4 counts have decreased to levels below 200 cells/µl, but it is not uncommon that early symptoms of immunodeficiency may appear already when CD4 counts are 200 - 500 cells/µl. During the AIDS stage, viremia steadily rises whilst the CD4 cell counts continue to decline. The infected individual may experience unusual opportunistic infections like pneumocystis jirovecii pneumonia, esophageal candidiasis and brain toxoplasmosis and/or rare malignancies like Kaposi´s sarcoma and Burkitt's lymphoma. The development from HIV-infection to AIDS takes on average 10 years [51] but varies between individuals.

There are infected individuals who can control the infection and remain asymptomatic despite the absence of ART. The virus is undetectable using standard assays but single viruses can be detected with special assays. These individuals are called long term non-progressors or elite controllers. The definitions of these two groups partly overlap but elite controllers are superior in controlling the infection.

1.3.2 Transmission

In 2012, 2.3 million new cases of HIV infection were estimated to have occurred globally [6]. HIV can be transmitted by sexual encounter (unprotected vaginal, anal and oral intercourse), but can also be vertically transmitted from mother to child or via contaminated blood or needles. Heterosexual transmission accounts for nearly 70 % of the new cases of HIV-1 infection worldwide [6].

In the absence of ART, the risk rate of penile-vaginal transmission of HIV-1 has been estimated between 1 in 2000 and 1 in 200. The probability for HIV transmission of unprotected anal intercourse is higher and range between 1 in 300 and 1 in 20 [49, 52-55]. The risk of HIV transmission is influenced by many factors. One of them is the viral load of the transmitting partner. A study in HIV-1 discordant couples has shown a 2.5-fold increase in transmission for every 10-fold increase observed in viral load [56, 57]. The clinical stage of infection in the transmitting partner is another factor. The risk of infection being transmitted from an individual with acute or early infection is higher than from one with an established infection due to the very high virus levels during this stage of the disease, but also because the infected person usually is unaware of his/her infection. In addition, co-infections may influence the risk rate, particularly infections causing genital inflammation ulcers in the genital tract [58]. The transmission event of HIV-1 involves a genetic bottleneck where one or a few virus particles establish the productive infection [59-61]. The number of transferred particles is dependent on the route of transmission [49, 62]. In 80 % of the heterosexual transmissions a single virus established the infection while the same number in injection-drug users and MSM is 40 % in both. CCR5 using viruses is found in most transmissions, but transmission of dual tropic CCR5/CXC4 using has been documented [62-64].

1.3.3 Prevention

The incidence of new infections in 2012 shows a 33 % decline compared to the 3.4 million in 2001 [6]. The decrease is largely due to ART. However, as noted above, over 2.3 million individuals were still infected during the year. Since no vaccine against HIV is available, the development of other prevention methods is continuously needed.

The use of cART has been shown to prevent sexual HIV transmission in several studies [65]. Results from the HPTN 052 study, where cART is used in combination with condoms and counseling in serodiscordant couples has in the published interim results shown a reduction in HIV transmissions by 96.4 % [66]. Male circumcision has in other studies been shown to reduce acquisition efficiency [67, 68]. The use of condom is always an important factor to avoid sexual transmission as well as treatment of other sexually transmitted disease if such is present. Mother to child transmission can be almost completely prevented if antiretroviral treatment is given to the mother and prophylaxis to the infant. Avoidance of breast feeding and in some cases, Caesarean section can further reduce the risk of mother to child transmission [69, 70]. Even though each of these and other prevention methods is helpful on their own, it is clear that a combination of intervention strategies must be used [71, 72]. The best solution would be an effective and safe HIV vaccine.

1.4 HIV-1 GENETIC VARIATION

1.4.1 Coreceptors

To infect a cell, the HIV-1 protein Env first binds to its primary receptor on the cell, the CD4, and then to a cellular coreceptor. The coreceptor used by HIV-1 is the C-C chemokine receptor type 5 (CCR5) and/or the C-X-C chemokine receptor type 4 (CXCR4). Viruses that use CCR5 are referred to as R5 viruses and viruses using CXCR4 are called X4 viruses. Some viruses are dual-tropic and use both coreceptors and they are referred to R5X4 virus [23]. Other coreceptors have been documented *in vitro*, but only CCR5 and CXCR4 are proven to be used *in vivo* [62].

The Env protein is divided into five conserved regions (C1-C5) interspersed with five variable regions (V1-V5). The gp120 coding domain of the *env* gene evolves faster (changing 1–2 % per year) than any other region of the genome [73]. The variable regions are presented on the surface of the protein and the principal determinant of coreceptors use is mainly located to the variable loop 3 (V3) [74], but the V1/V2, V4 and C4 regions have also been shown to affect the coreceptor binding [75-77]. Three amino acid changes, at positions 11, 24, and 25 of the 35-amino-acid-long V3 loop, are highly associated with the coreceptor switch [78, 79]. Positions outside the V3 loop have also been identified as statistically linked to changes within V3. Generally the genetic variation is greater after the switch, suggesting that substitutions are part of a more complex evolutionary pathway [80].

R5-using viruses are most often found to be the founder of a new infection, irrespective of the route of transmission but also X4-using and R5/X4-using virus have been detected in early infection. [62, 81-85]. The reason for the dominance of R5 virus is not fully understood. One theory is that the CCR5-using virus is preferred and selected for in a genetic bottleneck during transmission. A supporting fact that selection of R5-using viruses occur during transmission is found in humans who are homozygous defective for CCR5 expression. This defect is mediated by a deletion of 32 base pair in CCR5 (CCR5Δ32) causing a premature stop codon. Despite presence of functional X4-using virus the individuals who are homozygote for the deletion are highly protected from HIV-1 infection. Also individuals who are heterozygous seem to have some protection against the infection [86-88], but primarily show significantly slower rate of disease progression [89, 90]. The other theory suggests that virus type transmitted merely is a result of random selection. The dominance of R5-using virus is explained with the absence of X4-using virus in the transmitting partners. R5-viruses are most often the only virus present during major parts of the infection which by default results in the transmission of R5-using virus [91].

In 50-70 % of patients with untreated HIV infection X4 or X4R5 viruses emerge in the later stages of infection [92-96]. The cause of the coreceptor switch is not fully understood, but it is believed that the X4 viruses emerge from R5 viruses within an individual rather than are transmitted [97]. The coreceptor switch is associated with an accelerated decline of CD4 cells and a faster disease progression [97, 98]. It is not known if the switch to X4-using virus is a cause or/and a consequence of immunodeficiency [94, 99]. Longitudinal studies on a limited number of patients have shown the presence of minority X4-using viruses in samples obtained up to 12 months prior to the coreceptor switch [100].

Maraviroc was the first approved CCR5-antagonist [101]. Successful treatment has only been shown in patients with only CC5-tropic virus. Before initiating a treatment regimen containing maraviroc a HIV-1 tropism test should be performed to rule out the presence of X4 viruses [102].

1.4.2 Tropism prediction methods

The coreceptor use can be tested by phenotypic assays or and predicted bioinformatically from the sequence data (genotypic assay).

In the phenotypic tests, patient derived virus is tested for its' ability to replicate in specific cell lines expressing defined coreceptors. The MT-2 assay was the first widely used phenotypic test. In this assay peripheral blood mononuclear cells from a HIV

infected individual are co-cultivated together with MT-2 cells. If X4-using virus is present, they will infect the cells and form syncytia, while R5-using virus will not [103]. The disadvantages of this method are the lack of a negative control and, because the complete virus is used, the requirement for a biosafety level-3 facility. More recent phenotypic test uses parts or the entire *env* gene. The parts are amplified from plasma HIV RNA to generate recombinant virus or pseudovirions which in turn are used to infect human cell lines expressing CD4 and a coreceptor in cell cultures [104-106]. Both the virus and the cell line are usually specially engineered to allow high throughput, easy read-out and high reproducibility.

The Trofile phenotypic assay (Monogram Biosciences) [105] is the most widely used method to predict HIV coreceptor tropism in the US, while most of the screening in patients who are candidates for maraviroc therapy in Europe is performed by in-house genotypic tests [107, 108]. Genotypic assays are generally faster and less expensive compared to phenotypic assays.

Several algorithms to bioinformatically interpret the coreceptor use from the sequence data have been developed. The simplest method is the 11/25 charge rule. It only uses information of the charge of amino acids in positions 11 and 25 in the V3 loop to predict the virus tropism based on the finding that many X4 viruses have basic (positively charged) amino acids at one or both of these positions. The results show a moderate correlation with phenotypic tests [109]. PSSM and geno2pheno are more advanced prediction algorithms. Both algorithms use the amino acid sequence of entire V3 loop in the *env*-gene, and calculate scores with different methods. If the score in PSSM is below $-6{\cdot}96$ the sequence is considered R5, whereas sequences with s core above $-2{\cdot}88$ are predicted to be X4. In the geno2pheno the result of the interpretation is given as a quantitative value of the false positive rate (FPR). FPR is defined as the probability of classifying an R5 virus falsely as X4. Varying the FPR threshold value changes the sensitivity and specificity for X4 prediction. The genotypic tests have for a long time been based on Sanger population sequencing. One disadvantage with the population sequencing is the risk of minor variants present in less than 20 % of the population remain hidden. Such minority variants that have been shown to be of clinical relevance [109-112]. The majority of NGS-studies performed have used 454 sequencing to study coreceptors tropism but PacBio, Illumina and Ion Torrent have been demonstrated to predict minority X4 variants at similar levels [113].

1.5 ANTIRETROVIRAL THERAPY

1.5.1 History and current treatment

All steps of the virus replication cycle are potential targets for ART. Since viruses are obligatory intracellular parasites, they are completely dependent on the availability of suitable host cells. The processes targeted by ART must therefore differ from the host cell processes so that the ART primarily affects the viral replication as interference with host cell functions may lead to adverse side effects. Individuals with an untreated HIV-1 infection will in almost all cases develop AIDS which ultimately is followed by death, but the introduction of modern combination ART has transformed HIV infection into a treatable chronic disease [114]. Antiretroviral therapy suppresses the virus replication and thereby lowers the virus levels in the infected individual. In 1987, the first drug for HIV-1 infection treatment, azidothymidine (AZT), was introduced in the market followed by a few, similar, drugs during the early 1990's [115-117]. In 1996, the morbidity and mortality in AIDS dropped [118-120] dramatically due to the

development of new drugs and the introduction of new combination treatment methods. Since then, ART is given as a combination of at least three drugs simultaneously attacking different steps of the replication cycle [118, 119, 121, 122]. This treatment strategy is often referred to as highly active antiretroviral therapy (HAART) or combinational antiretroviral therapy (cART). To date, around 25 antiretroviral drugs have been approved for use in the treatment of HIV infection by the European medicine agency (EMA) in Europe and/or the Food and Drug Administration (FDA) in the United States [123, 124]. Through cART, it is possible to suppress the plasma HIV-1 viral load below detection limits of standard assays for quantification of plasma HIV-1 RNA (< 20-50 RNA copies/mL). There are six distinct classes of antiretroviral drugs but the majority of drugs are in three of the classes, nucleoside reverse transcriptase inhibitors (NRTIs), non-nucleoside reverse transcriptase inhibitors (NNRTIs) and protease inhibitors (PIs) (Table 1). Both NRTIs and NNRTIs target the HIV-specific enzyme reverse transcriptase and inhibit its function. NRTIs are compounds similar to and competing with the normal substrate of RT, i.e. nucleosides, but they are altered so that they lack a 3´hydroxyl group which leads to chain termination of the growing viral DNA chain [125-127]. NNRTIs are non-competitive and block the activity of reverse transcriptase by binding near to the active site of reverse transcriptase [127]. Protease is another of HIV-1's three essential enzymes. PIs resemble the normal peptide substrate of the protease and bind to the active site of enzyme and thereby inhibit the maturation of new viral particles, leaving them non-infective. Other drug classes are entry inhibitors, CCR5 antagonists, and fusion inhibitors.

In Sweden, treatment initiation is recommended when the CD4 count is < 500 cells/μl or if the patient experiences any of the following conditions regardless of CD4 count; AIDS diagnosis; some AIDS associated conditions; hepatitis B infection which demands treatment; non-HIV related cancer demanding cytostatic and/or radiation treatment; pregnancy; primary HIV-infection or a desire to minimize the transmission risk [19]. The first line treatment for previously untreated patients is a combination of two NRTIs and a PI, integrase inhibitor or NNRTI [19]. The first line treatment recommendations in the US are similar to the Swedish guidelines, but also include two NRTIs in combination with an integrase inhibitor. US treatment initiation is independent of the CD4 cell count and is recommended for all HIV-infected individuals to reduce the risk of disease progression [128].

First line treatment options in Sweden [19]:

- abacavir/lamivudine together with atazanavir/r
- abacavir/lamivudine together with darunavir/r
- abacavir/lamivudine or tenofovir/emtricitabin together with efavirenz
- abacavir/lamivudine or tenofovir/emtricitabin together with raltegravir

Table 1. ARV approved by FDA and EMA

Drug	Approved FDA/EMA
NRTIs	
abacavir (ABC)	1998/1999
didanosine (ddI)	1991[a]
emtricitabine(FTC)	2003/2003
lamivudine (3TC)	1995/1996
stavudine (d4T)	1994/1996b
tenofovir (TDF)	2001/2002
zalcitabine (ddC)	1992[a]
zidovudeine (AZT)	1987/1987
NNRTIs	
delavirdine (DLV)	1997/-
efavirenz (EFV)	1998/1999
etravirine (ETR)	2008/2008
nevirapine (NVP)	1996/1998
Rilpivirine	2011/2011
Pis	
atazanavir (ATV)	2003/2004
Darunavir	2006/2008
fosamprenavir (fAMP)	2003/2004
indinavir (IDV)	1996/1996
Lopinavir	2000/2001
nelfinavir (NFV)	1997/1998[c]
saquinavir (SQV)	1995/1996
tipranavir (TPV)	2005/2005
Fusion Inhibitor	
enfuvirtide (T-20)	2003/2003
Entry Inhibitor	
maraviroc (MVC)	2007/2007
HIV integrase strand transfer inhibitors	
raltegravir (RAL)	2007/2007
Dolutegravir	2013/2014

[a] the drug was withdrawn from the market by the manufacturer.
[b] not recommended by Swedish guidelines due to side effects.
[c] not recommended by Swedish guidelines due to low antiviral activity.

1.5.2 Treatment failure and drug resistance

Treatment failure is defined in three stages. 1) Virological failure occurs when plasma virus levels rebound or do not decrease sufficiently despite of cART. This might lead to 2) immunologic failure and 3) clinical failure.

Viral replication can be suppressed for decades when patients are treated under optimal conditions. Adherence is of greatest importance and without it, the patient risks virological treatment failure and development of drug resistance. Other factors such as poor drug tolerability and drug interactions between antiretrovirals (ARVs) and/or other medication may also lead to virologic failure and cause the evolution of drug resistance [129].

The high genetic variability in an HIV-infected patient creates a pool of genetically distinct HIV particles. In treatment-naïve patients, minority variants (virus variants that constitute less than approximately 20 % of the population in plasma) may contain low

levels of naturally occurring drug resistance mutations. When ART is initiated such variants with reduced susceptibility may be selected for and thereby contribute to treatment failure [29, 129].

Drug resistance mutations, especially those involved in development of PI resistance, are divided into primary and secondary mutations. Primary resistance mutations usually confer high level antiretroviral resistance, but are often also associated with a fitness cost. To compensate for the loss in fitness, secondary (compensatory) mutations may evolve. If successful cART is interrupted, the resistant virus usually is replaced by wild-type variants. The rebounding wild-type variants have been suggested to originate either from wild-type virus that had been archived in latently infected cells before start of therapy [130] or from continued evolution that leads to reversion of resistance mutations [131, 132].

Drug resistance is unequally prone to occur for different drugs and drug classes. For several NNRTIs and NRTIs a single mutation is enough to cause resistance. Hence, these drugs have a low genetic barrier. Other drugs have a higher genetic barrier, for example PIs, as several mutations are needed to cause high level resistance. Resistance to drugs with high genetic barrier usually requires suboptimal treatment during which the virus gets the chance to replicate during drug-selective pressure, which leads to *de novo* evolution of resistance mutations.

Drug resistant viruses can also be transmitted to newly infected individuals; this event is termed transmitted drug resistance (TDR). This is a clinical and epidemiological problem because it may contribute to failure of antiretroviral treatment. The prevalence varies geographically. In Sweden, 5.6 % of the newly diagnosed HIV-infections showed evidence of TDR [133], but most of these patients had low or moderate levels of resistance to one drug or drug class. In the US, the corresponding portion is 14.6 % [134] whilst the average in Europe is around 10 % [135].

For most of the drugs, the relevant resistance mutations and their impact on drug susceptibility is known. This makes it possible, and recommendable, to screen for the presence of drug resistant variants at diagnosis or before ART is initiated [19]. Resistance mutations may decrease the virus fitness. This is true for many of the drug classes, especially the NRTI lamivudine (3TC). For this reason, 3TC therapy is sometimes continued despite documented resistance to this drug. A risk if the replication is not completely suppressed by the other drugs used in the combination is that the virus might gain compensatory mutation that increase its fitness or accumulate more resistance mutations.

1.6 NEXT GENERATION SEQUENCING

1.6.1 History and current NGS-methods in short

Next-generation sequencing (NGS) has revolutionized the genomics research field. NGS is characterized by production of very large volumes of sequence data to a relatively low cost at a high speed. The automated Sanger sequencing [136] is considered a “first generation” DNA sequencing machine and new technologies following, with the 454-sequencer from Roche as the first in the market, are referred to as “the next generation” [137].

NGS is used to study whole genomes but it is also possible to study smaller, selected genomic regions more in depth. When long fragments, such as whole genomes are studied the common way is to fragmentize the DNA into small parts and sequence them, this is referred to as the shot-gun approach. After sequencing, reads must be assembled, either via multiple sequence alignment or to a reference sequence.

The choice of sequencing platform to some degree depends on the aim of the research project. There is a tradeoff between the amount of data, the read length, the accuracy of the generated data and the cost (Table 2). Generally, sequence platforms with high throughput and short reads like SOLiD and HiSeq 2000 are suitable for whole genome projects whilst in-depth studies of shorter regions benefit from longer reads such as the data from the GS-FLX Titanium (454 sequencing), Ion Torrent or pair-end sequencing on the MiSeq platform.

Table 2. Summary of current NGS technologies.

	454 GS-FLX Titanium/ 454 GS Junior	HiSeq 2500/ MiSeq	Ion Torrent (PGM)	RS II
Company	Roche	Illumina	Life technologies	Pacific bioscience
Amplification method	Emulsion PCR on beads	Bridge PCR in situ	Emulsion PCR on beads	No amplification is required
Principle (chemistry)	Synthesis (pyrosequencing)	Synthesis (reversible termination)	Synthesis (H+ detection)	Single molecule, real-time synthesis
Average read length (bp)	450/400	~2*150/ 2*300a	~400	4,200-8,500
Average yield/run (Gb)	0.45 /0.035	50-1000/ 0.3-15	1.2-2	0.02-0.08
Primary error and frequency reported (%)	Indels ~1	Substitutions ~0.32/0.1	Indels ~1	Indels ~13
Main advantage(s)	Long reads, maturity	Easy work flow, maturity	Low cost, fast run	Longest reads
Main disadvantage(s)	Homopolymer misreads, high cost per Mb	Shortest reads (HiSeq)	Homopolymer misreads	High error rate, expensive

1.6.2 PCR

Polymerase chain reaction (PCR) is a preparatory approach used to target and amplify selected regions of genetic material [138]. In the PCR process, a short synthetic oligonucleotide is designed to bind to the target DNA in the beginning of the fragment of interest and another one in the end of the same fragment. The two DNA complementary pieces of nucleotides are called primers, because they prime the reaction. The genetic material in between the two primers (the amplicon) is “cut out” and copied many times.

Primer design has always been important in project where PCR amplification and/or DNA sequencing is used, but with the NGS technology it has become even more crucial. The increased possibilities to study rare variants hidden in diverse populations, demand primers that are placed in conserved areas of the genetic material. This is especially challenging in RNA viruses and other divergent viruses. Primers that do not capture the full population diversity and thereby favor certain variants will cause a bias in the result. Other factors to consider when the primers are designed for NGS are the longer primers (gene specific primer together with unique sample tags and platform-specific adaptors) as well as the increased multiplexing (several samples in the same reaction). Both lead to a greater risk of primers and templates binding to themselves (forming hairpins) or to other primer/template present in the same reaction (dimerization).

1.6.3 454 sequencing methods-UDPS

454 sequencing was the first available NGS platform. The sequencing technology is based on a sequencing by synthesis chemistry called pyrosequencing [139]. The platform has many applications and one of them is targeted resequencing or ultra-deep pyrosequencing (UDPS) as it also is referred to. The methodology is described below and in Figure 5.
The library preparation is the first step of the process. It is initiated by targeting the region of interest and attachment of the 454-specific adaptors A and B. The double stranded DNA is separated into two single strands and each strand is attached to a DNA capture bead by binding to a complementary adaptor strand. A droplet is then formed around the bead by shaking a mixture of oil and water. Most droplets contain a single DNA fragment as well as many small enzyme beads. The droplet works as a mini-reactor and millions of immobilized DNA copies are produced in the emulsion PCR (emPCR). Each bead is then washed and placed on a PicoTiterPlate for sequencing. One bead is loaded into one well. Bases are flown sequentially over the plate, always in the same order (TACG). If one or several nucleotides of the type are complementary to the strand, they will be incorporated and a chemi-luminescent signal proportional to the number of nucleotides is produced. The light signal is recorded by a CCD camera and converted to bar graph of light intensities called a flowgram. Each well generates a flowgram and translated to a sequence (also referred to as a read) [140].

1.6.4 Possibilities of ultra-deep sequencing

UDPS, which also referred to as amplicon sequencing, is an application of the 454-platform. It has frequently been used to study viruses, in particular rapidly evolving RNA viruses, such as HIV and hepatitis C virus (HCV). During the last years, UDPS has been widely used and considered to be a valuable tool to study minority variants at frequency below the detection limit of standard genotyping assays. However, the development of other sequencing techniques and platforms has continued, and currently the 454-platform is being phased out in virology research to be replaced with other platforms with even greater potential. Ion Torrent and MiSeq are two of the newer platforms that are replacing the 454-platform in studies of TDR, coreceptor use, characterize within-host evolution and drug resistance [100, 141-144].

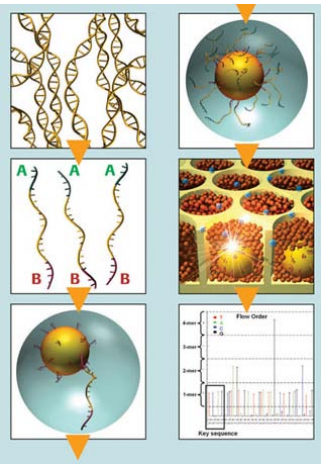

Figure 5. The 454 sequencing workflow.

1.6.5 454 sequencing limitations and overcoming errors

Compared to Sanger sequencing the NGS methods, and especially Roche-454 sequencing, are more error prone [145]. This is an obstacle, for example when the presence of drug resistance mutations in minority species in studied. It is of greatest importance to be able to distinguish a rare true biologic variant from a variant resulting from an artifact created somewhere in the cDNA synthesis, PCR or sequencing steps. Originally, Roche-454 error rates were estimated to 4 % for experimental samples, and 0.6 % for test fragments but subsequent versions of Roche-454 have greatly reduced these error rates [146]. Several strategies to identify, characterize and overcome these errors have been published. These bioinformatic strategies to obtain more reliable data differ. One approach is to filter sequence reads with low quality prior to or during alignment [147-149]. Another is to use statistical approaches where single nucleotide variation is detected and reconstructed. [150-153]. Both in-house software and public programs are used, each of the methods has its' specific pros and cons. Artificial recombination of templates created during the PCR also contribute to the error frequency and programs to bioinformatically identify these recombinants have been developed [150].

1.6.6 Molecular tagging – Primer IDs

Errors occur during PCR. PCR-free sequencing is rarely possible. Random sequence tags have been used to circumvent some of the remaining PCR artefacts [154, 155]. This method, where every individual molecule is tagged and resequenced was used in an HIV study by Jabara and colleagues [156]. The sequence tags were then referred to as Primer IDs. The Primer ID consisted of a stretch of randomized nucleotides (N's) in the primer used for cDNA synthesis. Using this approach, the sequence reads originating from the same template molecule can be identified and grouped according to their unique Primer ID. This makes it possible to construct a consensus sequence for template molecules that has been resequenced three times or more. The consensus sequence will be free from errors even if the single reads contain random PCR substitution errors and PCR recombination errors. The method requires high volumes of data since it is based on resequencing of the template molecules. It also needs a sequencing technique that produces long reads because a Primer ID of a certain length will be added to the amplicon. The length of the Primer ID is dependent on the number of cDNA template in the sample. The number of unique Primer IDs must be enough to label each template with a unique Primer ID.

2 AIMS

The specific aims of my thesis were:

I. To develop a software program that designs primers from a multiple alignment and are suitable for next generation sequencing.
II. To investigate the characteristics and sources of errors in data from ultra-deep pyrosequencing and to develop methods to reduce the error frequency.
III. To evaluate the quality and reproducibility of the UDPS technology in analysis of HIV-1 *pol*-gene variation.
IV. To investigate, by UDPS, the presence of drug resistance mutations in treatment naïve HIV-1 infected patients and the dynamic of drug resistance development and reversion during treatment initiation and discontinuation.
V. To investigate if CXCR4-using virus is present as a minority species already during primary HIV-1 infection in patients whose virus later switches to CXCR4-use.
VI. To study the utility of using an improved NGS methodology called Primer ID.

3 MATERIALS AND METHODS

3.1 MATERIALS

No human material was used in **Paper I**.

In **Paper II**, a SG3Δenv-plasmid was diluted to a single copy, amplified and sequenced in three separate runs on the Genome Sequencer FLX. The amplicon contained 167 nucleotides from the HIV-1 *pol* gene corresponding to the last nucleotide of amino acid 169, amino acids 170–224, and the first nucleotide from amino acid 225 as well as the sample tags and the 454-specific adaptors A and B. Sequence analyses were performed on the total dataset of 47,693 reads obtained from UDPS.

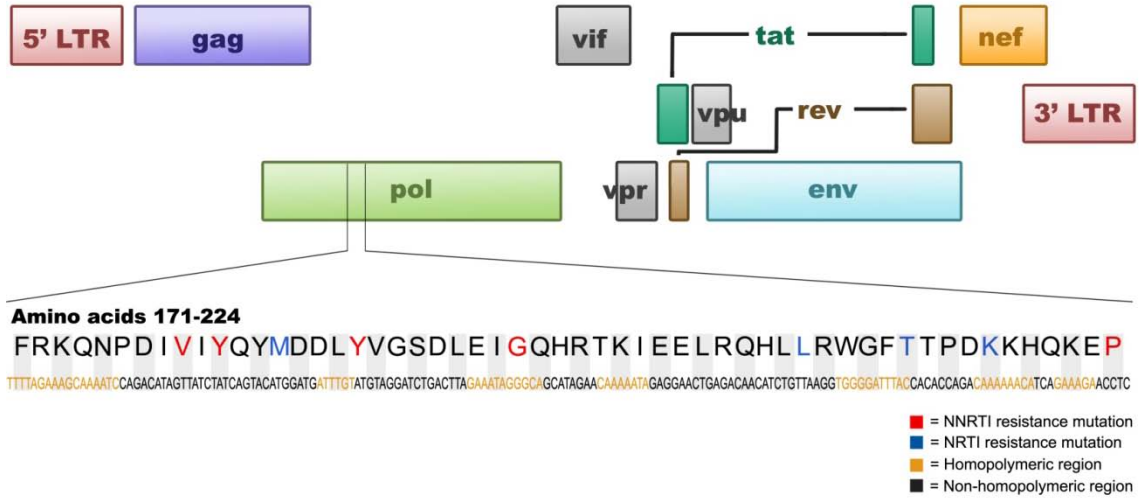

Figure 6. The HIV-1 genome organization and the sequence used in **Papers II**, **III**, **IV** and **VI**. Kindly provided by Anna Sahlberg.

In **Paper III**, four plasma samples (A-D) were used. Sample A and B were used to study repeatability, effects of sequence direction and the influence of primer-related selective amplification. These samples had approximately 1,050,000 and 1,600,000 HIV-1 RNA copies/ml, respectively. Plasma samples C and D were used to generate two molecular clones for studies of UDPS sensitivity and *in vitro* PCR recombination. These two clones were therefore chosen on the basis of sequence dissimilarity with the aim to maximize the number of informative sites. The sequence region was the same as in **Paper II** and the samples used were the same as described below in **Paper IV** (sample A, B, C and D correspond to sample 6.4, 2.5, 4.5, 3.5 in **Paper IV**).

In **Paper IV**, six to eight longitudinally obtained plasma samples from six patients were retrospectively investigated. All patients were infected with subtype B virus and had experienced virological treatment failure. The patient selection was based on the patients' treatment history and plasma viral load (ranging from 17,900–1,600,000 HIV-1 RNA copies/mL). All patients had started treatment before combination ART was used. Their exact treatment history varied but common for all patients was the use of 3TC, AZT and d4T. All patients, except one, were sampled before treatment was initiated and, all except one (not the same), underwent and were sampled during a subsequent treatment interruption. The sequence region was the same as in **Paper II** and **IV**.

In **Paper V**, four to nine longitudinally obtained plasma samples from each of three patients were retrospectively investigated. All patients were infected with subtype B virus and had a HIV population that switched coreceptor use from CCR5 to CXCR4. The information about the coreceptor use was based on the MT-2 assay, which had been performed when the samples were originally obtained. The MT-2 results had been stored in the database connected to the biobank. Patients 1 and 2 were sampled during PHI. Both patients were classified into Fiebig stage II based on a negative HIV antibody test and positive HIV antigen and HIV RNA tests. When the first sample was drawn from patient 3, he was classified to be in Fiebig stage IV–V based on a positive HIV ELISA antibody test and an incomplete Western blot profile that lacked a p31 band. For all three patients, the remaining samples were collected both before and after documented coreceptor switch.

In **Paper VI**, the SG3Δenv plasmid (same as in **Paper I**) was used as a control to investigate the accuracy of the Primer ID UDPS system. Plasma samples from three HIV-infected patients (A, B and C) were also investigated. The patients that were selected for evaluation of the Primer ID method were selected from a study on transmitted drug resistance in Sweden.

3.2 ETHICAL CONSIDERATION

For **Papers III**, **IV** and **V**, an ethics application was approved (Dnr 2008/122-31/2) by Regional Ethical Review Board in Stockholm, Sweden and for **Paper VI** an ethic application was approved (Dnr 2007/1533) by the same board.

All patients gave written or oral informed consent in accordance with the Declaration of Helsinki.

No patient material was used for **Papers I** and **II**, therefore no ethic application or approval was needed.

3.3 SEQUENCING

The sequence depth of UDPS primarily depends on the number of input molecules and the error frequency of the sequencing method. **In Paper IV**, the sequencing protocol was carefully optimized to maximize the number of HIV RNA molecules that were extracted, reverse transcribed, PCR amplified and subjected to UDPS. In-house HIV specific primers were used together with sample-specific tags to allow multiplexing during sequencing. The amplicon also contained 454 specific adaptors to allow UDPS.

The HIV RNA was extracted and purified. The amount of plasma used for extraction was adjusted according to the viral load of each sample. The number of viral templates (HIV-1 cDNA copy number) for each sample was quantified by limiting dilution PCR before UDPS so that the number of templates subjected to sequencing could be related to the number of UDPS sequences obtained. The protocol is presented in detail in **Paper IV** as well as in Figure 7 and was used in **Papers II-VI**.

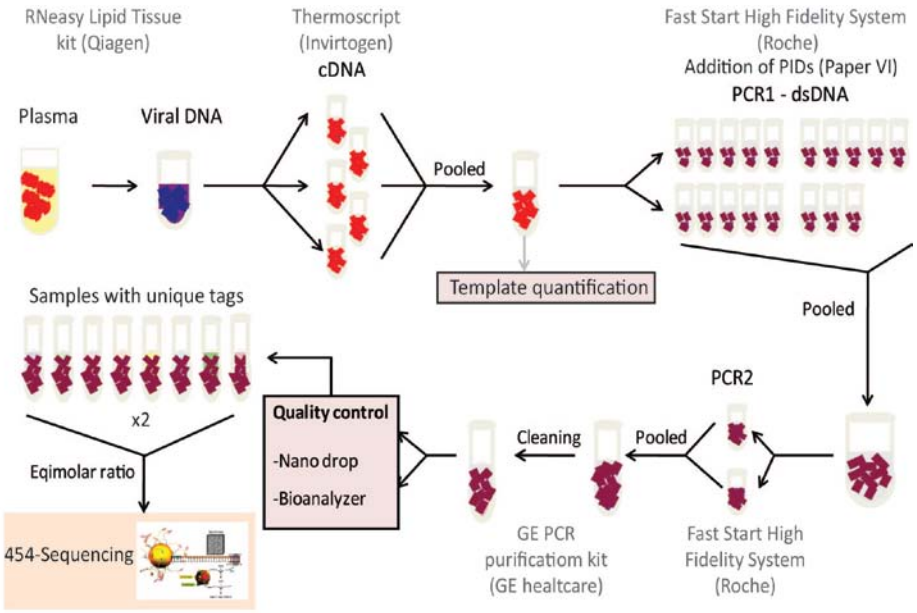

Figure 7. Schematic illustration of the experimental setup used in **Papers II-VI.**

3.3.1 Calculation of error frequencies

The Needleman-Wunsch algorithm was used to construct pairwise alignments between a reference sequence, a Sanger population sequence of the SG3Δenv plasmid, and UDPS reads. The identity score (the number of correctly aligned bases divided by the total number of bases) from the pairwise comparisons were added together and divided by the number of sequences.

We present different error frequencies derived from the same raw data in **Papers II** and **IV**. The different numbers are due to a difference in calculation. In **Paper IV**, missing nucleotides in reads that did cover the entire 167-basepair amplicon (short reads) were considered as sequencing errors and contributed to the average error frequency. In **Paper II**, we ignored such missing nucleotides in short reads which resulted in a lower error frequency. Other researchers have, in their papers, generally omitted how such missing data has been handled.

3.3.2 UDPS data filtering procedure

We designed a set of Perl scripts to filter UDPS data from reads that were likely to contain sequencing errors. Most other methods are based on correction of errors. Both approaches have their specific pros and cons. Filtering may lead to loss of data (reads), whereas correction algorithms may create artificial viral variants which were not present in the original sample.

Our data filtering strategy detected variation relative to the Sanger sequence of the SG3Δenv plasmid in the control experiments and a population Sanger sequence for each of the patient samples. Each filtering step divided the sequence reads into two files; one file with reads that passed the filtering step and another file with reads that were removed by the filtering because they had characteristics associated sequencing errors. Some or all filtering steps were use for **Paper II**, **III**, **IV** and **V**. In **Paper IV**, statistically derived cut-offs were applied to the cleaned data.

1) Identification of unique UDPS reads. To simplify the data handling and reduce the computational time, all sequences were collapsed to unique variants. The abundance (i.e. number of reads) of each unique variant was added to the sequence header of that variant; **2) Removal of low similarity reads.** The first filtering step removed reads with low similarity to a reference sequence, i.e. non-HIV sequences or HIV sequences with very low quality. We used the Needleman-Wunsch algorithm to construct pairwise alignments between a Sanger reference sequence, and the unique UDPS reads to obtain the similarity score. If the alignment identity score was below a user-defined threshold, the read was removed. In **Papers II**, **III** and **IV**, an 80 % similarity threshold was used. In **Paper V**, the corresponding threshold was 70 %, because the V3 region is more heterogeneous than the *pol* region; **3) Removal of reads with ambiguous base calls "N's".** The 454-software uses the character "N" to describe an ambiguous base call. Huse *et al*. showed that reads from the Genome Sequencer 20 (454 Life Sciences, Branford, CT) instrument containing N's have a higher error frequency than reads without ambiguous base calls [157]. Our data, that was generated using the GS-FLX instrument, also showed this and we therefore removed reads containing N. This was performed in **Papers II-V; 4) Removal of reads not covering the region of interest.** Reads that did not cover the entire region of interest (amino acids 180–220 in RT, position 3087 to 3206 in HxB2, GenBank accession number K03455) were removed in **Papers III** and **IV**. Remaining reads were imported into the GS amplicon software (Roche, Penzberg, Germany) and aligned; **5) Removal of reads with out-of-frame indels.** UDPS errors frequently involve indels, especially in homopolymeric regions [146]. Therefore, we identified reads with out-of-frame indels and longer (≥6 nucleotides) frame-shifted regions. This step retained reads with indels involving entire codons as well as reads with short frame-shifted regions (<6 nucleotides), which may represent functional HIV-1 variants. The latter reads were flagged to allow visual inspection, which was done in **Paper II**. In **Paper IV**, a slightly modified indel filtering was used. Reads with in-frame indels, ±3,6,9... nucleotides were retained while reads with out-of-frame indels were removed; **6) Removal of reads with stop codons.** UDPS data from coding regions that contain stop codons are likely to represent sequencing errors, or are otherwise evolutionary dead-ends. We would not apply this filter if we would have been interested in studying stop codons in UDPS data from clinical patient samples or if we would have studied non-coding regions; **7) Forward and reverse read comparisons.** The tally of each unique variant in forward and reverse reads was compared for all variants found in **Paper IV**. The abundance of a variant was set to the sum of the forward and reverse tallies unless the frequencies of the forward and reverse reads differed by more than a factor 10. If it did, we made the assumption that a systematic error had occurred during 454 sequencing and adjusted the frequency to the lower of the two estimates. If a variant was found to be absent in either forward or reverse direction it was discarded from further analyses; **8) Manual inspection.** The remaining alignments were manually inspected for any remaining sequencing errors in **Papers II-V**; **9) Cut-off values.** In **Papers III** and **IV**, variants were classified as high-confidence variants if their abundance exceeded a sample-specific cut-off value. The cut-off value was calculated using the overall average error frequency and the 95 % confidence interval from the SG3Δenv plasmid sequenced in three separate runs. In **Paper IV**, cut-off values were also derived for individual drug-resistance positions. For each individual nucleotide position the average error frequency at that site and its' 95 % confidence interval was obtained from the SG3Δenv plasmid sequenced in three separate runs. A Chi-square test with correction for continuity was used to evaluate if the frequencies of variants/drug resistance mutations were significantly higher than the observed experimental error. The variants/drug

resistance mutations with frequencies above the cut-offs were retained for further analysis.

3.4 MOLECULE TAGGING (PRIMER IDS)

The depth and accuracy of the sequencing analysis is strongly influenced by the frequency of introduction of experimental errors before and during sequencing. We have developed an NGS methodology that has the potential to generate NGS data with greatly reduced error frequency compared to standard NGS. The methodology was applied to UDPS on the 454 GS-FLX platform, but could also be used on other NGS platforms. The generation of sequence data and the bioinformatic pipeline to process the data is described below.

3.4.1 Experimental approach

RNA extraction, cDNA synthesis and semi-nested PCR amplification were performed according to the experimental protocol presented in **Paper IV** and Figure 7. The key feature of the method is the Primer ID, a unique sequence tag that labels each template molecule prior to PCR amplification, Figure 8. Our Primer ID consisted of 10 randomized nucleotides, which enables 1,048,576 unique combinations. The Primer ID was added to the HIV-specific reverse cDNA primer together with the 454 specific B adaptor. This primer was synthesized with uracils instead of thymidines, which allowed it to be degraded by uracil-DNA glycosylase and NaOH following cDNA synthesis. Sample tags were added to one of the forward PCR primers to allow multiplexing, just as in standard UDPS.

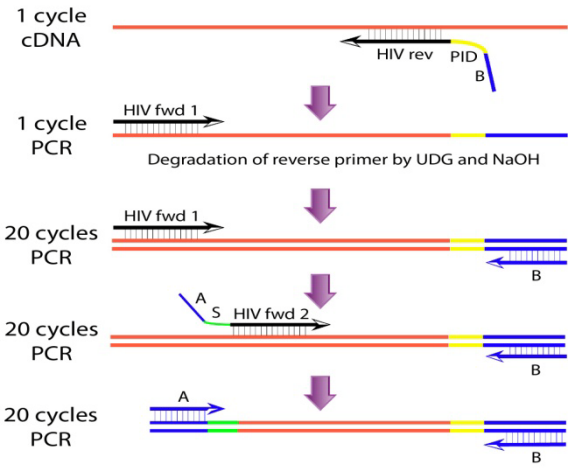

Figure 8. Schematic picture of the Primer ID process.

3.4.2 Bioinformatic approach

The sequences were first sorted by their sample tag. Sequences containing the same Primer ID originated from the same template molecule and were therefore sorted into groups. The sequences were multiply aligned using Muscle [158] and the alignments for each Primer ID were used to construct strict majority-rule consensus sequences if the group consisted of at least three sequences. These sequences were referred to as consensus template sequences because they should be an accurate reflection of the corresponding template sequence (HIV RNA molecule) in the patient sample, with the exception of errors that might have occurred during cDNA synthesis.

3.5 PROGRAMMING

Perl is a family of programming languages that have different built-in modules to make various tasks easier. BioPerl is a collection of Perl modules that facilitate the development of Perl scripts for bioinformatic applications, such as translation of nucleotides to amino acids and creating alignments. Perl 5 was used to handle data in **all Papers** and to clean data in **Papers II-V**. The programming language C was used for the computationally intensive dimerization risk estimation in PrimerDesign. Moose object-oriented programming in Perl has been used together with the Catalyst Model-View-Controller framework to construct the web interface for PrimerDesign in **Paper I**. In-house Perl scripts were first used, but later translated to Python scripts, to analyze the 454 sequence data in **Paper VI**.

3.6 PHYLOGENETIC ANALYSES

The evolutionary relationship of the sequence variants in **Paper V** was analyzed using maximum likelihood trees. jModelTest [159] was used to find the best-fit model of nucleotide substitution and PhyML 3.0 [160] was used to construct the trees. The maximum likelihood method was chosen before other phylogenetic methods because it is an accurate method that works well on rapidly evolving organisms like HIV [161].

3.7 TROPISM PREDICTION

In **Paper V**, the coreceptor tropism of the viruses was phenotypically tested using the MT-2 assay at the time of sampling. Genotypic coreceptor testing was performed on sequences from the V3-region with the prediction algorithms $PSSM_{x4/r5}$ and $geno2pheno_{[co\text{-}receptor]}$. Variants were considered to be X4-using viruses if the PSSM score was greater than - 2.88 and geno2pheno predicted an X4 phenotype using a FPR of 2.0 % or less. The PSSM score was chosen based on the European recommended guidelines [102]. Other cut-offs for $geno2pheno_{[co\text{-}receptor]}$ were also evaluated and the 2.0 % cut-off was chosen because it gave the highest agreement with PSSM predictions on our data.

3.8 STATISTICAL ANALYSES

Mann-Whitney U test was used to test whether error frequencies in homopolymeric- and non-homopolymeric regions were statistically significant (**Paper II**); Fisher exact test was used to test if the error frequencies of transition and transversion errors differed significantly (**Paper II**); Spearman rank correlation test was used to study the correlations of site-specific error frequencies between runs as well as between forward and reverse reads in the same run and their P-values were calculated using a z-test (**Paper II**). The repeatability of variant quantification between samples in **Paper III** was statistically tested using Bland-Altman analyses and plots.

4 RESULTS AND DISCUSSION

NGS, in general and UDPS in particular, introduced new possibilities of studying the genetics and evolution of HIV-1. In 2008, when I initiated my PhD-studies, the technique was relatively new. At that time, there was a widespread excitement concerning the use of the method, but it was also realized that the analysis of the large sequence datasets was challenging. Thus, there was also a great need to develop and fine tune the methodology and the downstream analyses. UDPS can be used to study diverse virus populations, but the utility of the method can be hampered by experimental errors occurring in library preparation, sequencing and analysis. This thesis includes six studies, **Papers I to VI**. The topics include characterization of the type and source of errors arising in UDPS, optimization of pre-UDPS protocols, evaluation of the UDPS system and development of novel bioinformatic methods to design primers and to reduce UDPS errors through a filtering strategy. We have also analyzed and evaluated the challenges of a new method to tag individual molecules in order to reduce sequencing errors. The results from these studies are presented and discussed below, based on the common themes across the **papers**.

4.1 PRIMER DESIGN

Paper I presents and discusses PrimerDesign. PrimerDesign is a computer program tailored to meet the requirements for primers in the settings of NGS and highly variable genetic targets. A vast number of commercial and non-commercial primer design programs were available before we initiated our work [162-169]. The large number reveals the importance and difficulty creating suitable primers. However, none of the existing programs fulfilled our requirements. In contrast to the available programs, PrimerDesign optimizes the construction of primer pairs with constraints including degenerate sites to maximize population coverage, matching of melting temperatures, optimizing de novo sequence length, finding optimal bio-barcodes to allow efficient downstream analysis, and minimizing risk of dimerization. PrimerDesign´s workflow follows a series of inter-connected steps. Each step is described in order as they appear:

1) **Determination of the primers target locations**. The user inputs a multiple alignment and states the region of interest (ROI) together with some other optional parameters. The primers generated are located outside the ROI but within a distance constrained by the sequencing length of the intended platform. The genetic complexity (i.e. the number of degenerated sites for each position) is calculated and the entropy for the region is estimated using Shannon entropy. All potential primers are thereafter listed in order of result from the entropy estimations; **2) Optimization of primer melting temperatures (Tm)**. Tm is estimated using the empirical nearest-neighbor model with respect to the possibility of degenerated sites [170, 171]. Should the difference in Tm between a forward and a reverse primer be within the maximum allowable limit, the primer-pair is included in the list of potential constructs; **3) Adding bio-barcodes and adaptors.** If desired, the program can be set to generate barcodes (also referred to as tags). These can be optimized by either a number of unique tags, a certain length of the tags or the edit distance (Levenshtein distance), i.e. the possible minimum number of nucleotide changes required to transform one tag to another is dependent on tag length. Adaptors can also be added by the user, either by choosing from existing platform specific adaptors or by manually creating new. The tags must not have repeated nucleotides at adjacent sites as this is known to cause misreads in UDPS. If adaptors are added, this is automatically controlled and constrained; **4) Estimation of dimerization risk**. When a primer binds to other primers or to

themselves, primer-dimers or hairpins may be created which may cause serious problems in the PCR. PrimerDesign analyzes the potential dimerization risk between all primer constructs (primer-tag-adaptor oligomers) that will be included in the same reaction. The dimerization risk is evaluated as a user-defined sliding window move along all potential primer-primer interactions and as a user-specified fraction of bonds in an interaction. Thus, each step includes user settings, which may be default values, as well as automatic parameters used within the algorithm. The final primers are presented in pairs.

One limitation of PrimerDesign is the restriction on tag generation. Currently, it allows generation of up to 200,000 tags and a tag length of 18 nucleotides. For long tags, the edit distance can be set up to 10. The tagging possibility is sufficient for most studies undertaken today, despite the restriction. An additional limitation is the risk that the algorithm becomes very computationally intensive, particularly if a large number of tags and/or a high complexity are included in the design.

PrimerDesign´s strength is its ability to use multiple alignments in combination with its comprehensive approach. The multiple alignments make it possible to avoid unnecessary target bias. This is done by finding primers located in as conserved regions as possible. The increasing usage of multiplexed NGS is supported by PrimerDesign's automatic and flexible tag and adaptor generation. The possibility to reduce the dimerization risks for the entire amplicon is to our knowledge a unique feature of PrimerDesign when compared to other similar programs.

Anterior primer design programs have focused on specific needs of certain experimental protocols and not NGS. In these programs, only a single sequence is generally used to design primers (e.g., Primer3 [169]) and, as discussed previously in this section, the risk of undetected genetic variability could result in amplification biases in genetically diverse target populations. There are programs that use multiple alignments (e.g., GeneFisher [165]) but lack the ability to simultaneously evaluate primers appropriately, for instance, by matching Tm's and dimerization risk.

The development of PrimerDesign was based on our experience with 454 ultra-deep sequencing of HIV-1, but the program is platform independent. Thus, the software can be used for primer and probe design for other NGS technologies which are preceded by a PCR step, e.g. IonTorrent, MiSeq/HiSeq and SOLiD as well as general PCR, real-time PCR and traditional Sanger sequencing protocols. Further, PrimerDesign can be used to design primers for all stages of variability in DNA-sequencing and not only for HIV sequencing.

The results reported in **Paper III** emphasize the importance of primer design. In that study, we evaluated two carefully designed sets of primers. They targeted the same *pol*-region and their ability to quantify viral variant abundance was investigated. Both sets of primers detected variants down to 0.2 % of the virus population. In one of the primer sets, which were also used for the patient samples in **Paper IV**, we estimated one variant to constitute 46 % of the population. In contrast, the same variant was only detected in 5.6 % of the reads with the alternative primers. We view this result as an indication of primer-related selective PCR amplification that occurred as a result of primer mismatch.

4.2 EVALUATION OF ULTRA-DEEP PYROSEQUENCING

We have used UDPS to detect minority variants containing drug resistance mutations (**Paper IV**), coreceptor usage (**Paper V**) and acute infection (**Paper V**). Others have shown that minority HIV resistance mutations, below the detection limit of population Sanger sequencing, may be of clinical relevance [172-176].

The resolution of our protocol and those reported in others studies is primarily determined by the number of input DNA templates, the error frequency of the method and the efficiency of data cleaning. Therefore we have focused on optimizing the experimental protocols in **Paper IV**, characterizing the type, frequency and source of errors and minimizing their impact in **Papers II** and **VI**.

4.2.1 Pre-UDPS experimental setup

RNA extraction, cDNA synthesis and PCR were all optimized for high recovery of templates in our pre-UDPS protocols. Quantification of the number of cDNA template was performed by a limiting dilution PCR using the very same PCR as that used for UDPS preparation. This template quantification showed that the number of cDNA molecules subjected to UDPS ranged from 2,300 to 570,000 in **Paper IV** and from 56 to 93,632 in **Paper V.** The UDPS generated from 3,827 to 41,490 reads per sample in **Paper IV** and 279 to 32,094 reads per sample in **Paper V**. We experienced low recovery in a few of the samples which could be due to long and suboptimal storage conditions. Most of the samples had been stored at -70℃ or -20℃ and some samples had been repeatedly freeze-thawn. Consequently, for some samples the UDPS reads exceeded the number of cDNA templates. In many preceding and simultaneous studies, a low number of viral templates were used as UDPS input and often not accurately quantified. This in combination with low number of reads resulted in higher detection limits for minority viral variants in these studies. Studies carried out today are generally more carefully designed. In these more recent studies, the number of input molecules is both quantified and high enough to benefit from the advantages of UDPS. However, some NGS studies would still have benefitted from the use of other methodologies such as single genome sequencing (SGS).

Overall, we sequenced a sufficient number of viral templates from the samples with sufficient depth to take advantage of the ability of UDPS to study minority HIV variants. However, in our studies we have definitely resampled the samples virus variants. Thus, it is important to remember that every sequence read does not correspond to one viral RNA template. Due to oversampling, the lower limit of detection of our UDPS studies were primarily limited by errors introduced during PCR and UDPS.

4.2.2 Characteristics and source of errors in raw UDPS data

In **Paper II**, an HIV-1 clone was diluted to a single copy, PCR amplified and ultra-deep sequenced in three separate runs. The sequenced region corresponds to a part of the *pol*-gene where many drug resistance mutations are found. The same region was used to evaluate the performance of UDPS in **Paper III** and studied in the patient samples in **Paper IV**. This region was also used to examine the challenges with Primer ID in **Paper VI**.

The sequence analysis presented in **Paper II** was performed on a complete dataset of 47,693 UDPS reads as well as separately for forward and reverse reads from each of the

three UDPS runs (the reads per sample range between 2,570 and 12,092). The average error frequency in our raw data was 0.30 %. UDPS-induced deletions in homopolymeric regions were the dominating error type. A substantial part of the deletions were only found in the reverse sequencing direction of the same run. This indicates that they were introduced during the UDPS. As anticipated, we found that homopolymeric regions had a higher average error frequency (0.59 % per nucleotide) compared to non-homopolymeric regions (0.12 % per nucleotide). This result is consistent with the findings of others published both before and after our study [146, 148, 157, 177, 178]. Despite this apparent difference in average error frequency, there was no statistically significant difference between homopolymeric and non-homopolymeric regions. This implies that a few single positions of the homopolymeric regions contributed to a substantial part of the elevated error frequency. This was confirmed when site-specific frequencies of deletion errors in homopolymeric regions ranged from 0.0021 % to 20.4 %. In fact, the site-specific error frequencies, particularly substitutions, were unevenly distributed across the region that was sequenced. Among the substitution errors, transitions were more common than transversions.

4.2.3 Filtering strategy

Based on our analysis of raw data in **Paper II** and **Paper IV**, as well as previous publications, we developed a set of scripts that filtered reads that were likely to contain sequencing errors. The backbone of the filtering strategy was to remove reads containing: 1) less than 80 % similarity to a user-defined reference sequence, 2) ambiguous nucleotide calls, 3) indels, and 4) stop codons. The steps are explained in detail above (Materials and Methods section). The filtering step where indels were removed had the most pronounced effect and reduced the average error frequency almost 5-fold from 0.28 % to 0.058 % per nucleotide. The cleaning procedure removed 31 % of the reads in the data for **Paper II**. Similar cleaning procedures for the data used in **Papers III** and **V** removed 20 % and 15 % respectively. In **Paper IV**, the data cleaning strategy was used together with cut-off values for high confidence variants which removed 30 % of the reads.

Other studies have used similar approaches and removed sequences associated with errors while other have reconstructed haplotypes e.g. ShoRAH. The best approach depends on the goal of the study. A limitation of the filtering strategy is the risk of removing true biological variants and that some remaining substitution errors may be interpreted as true variants. In addition, increasing read lengths may pose a problem with the filtering strategy since the probability for occurrence of a sequencing error increases which may lead to filtering of a large proportion of the reads. This risk is also dependent on the type of sequence (e.g. homopolymeric /non-homopolymeric regions). On the other hand, error correction leads to a risk of creating new variants or changing true low frequency variants.

4.2.4 Characteristics and source of errors in cleaned data

The filtering strategy, as presented above in the Material and Method section, was applied on the SG3Δenv-plasmid. The average error frequency per nucleotide for the six data sets in **Paper II** was reduced to 0.056 %. The error frequencies estimated for the cleaned data from the V3-region in **Paper V** was about the same for both the 454 GS FLX and the 454 GS FLX Junior Titanium platforms.

In **Paper II**, all except two reads with indel errors were removed. Interesting to note is that the cleaned average error frequency is about the same as the error frequency for

substitutions found in the raw data (0.057 %). The difference in error frequencies between homopolymeric and non-homopolymeric regions was almost completely removed in the cleaned data.

The average error frequency of transitions was 0.052 % per nucleotide and the corresponding number for transversions was 0.001 %, which is a 48-fold and significant difference. Site-specific error frequencies continued to vary across sites. Moderate, but significant, correlations in site-specific error frequencies were found when forward or reverse reads from three separate runs were compared (Spearman R= 0.31–0.65; p=0.001). Significant correlations were found between forward and reverse reads within runs (Spearman R= 0.33–0.60; p=0.001).

Altogether, this indicates that the PCR that preceded UDPS contributed to a substantial proportion of errors that remained in our cleaned UDPS data.

In **Paper III**, we showed that the *in vitro* recombination rate during PCR was low. Two clones that differed in 13 positions were mixed in 50:50 ratio before PCR amplification. The mixes were used in two experiments with 10,000 and 100,000 HIV DNA templates as input, respectively, and the estimated recombination rates were 0.29 % and 0.89 %. The majority of recombinant reads were single recombinants. The recombinant variants were found in low frequencies and below our limit of detection in **Paper IV**. The recombination rate in our control study (**Paper III**) was higher than the recombination rate (0.09 % - 0.11 %) estimated by Tsibris *et al*. [149] and lower than 1.9 % presented by Zagordi *et al*. [179]. The difference in *in vitro* recombination estimates may be a result of different mixture of clones, differences in amplicon length and differences in PCR amplification conditions. In our and collaborators´ recent, unpublished experiments, we observed that the PCR recombination rate could be greatly reduced if the number of PCR cycles was reduced from 60 to 30, i.e. by omitting the second, nested PCR.

Our PCR recombination studies were performed on DNA templates. This provides a limitation to our study as the first step in the PCR process, the cDNA synthesis where RNA it reverse transcribed into cDNA, is not included in our experiments. RTs are, as discussed in the introduction, error-prone enzymes and as a consequence our result may be an underestimation of true recombination. Fang and colleagues reported a 2.5 fold higher *in vitro* recombination in RT-PCR compared to DNA-PCR [180] while Metzner *et al*. did not find the RT step to particularly affect the recombination rate in a recent publication [177]. Our results should in our view be interpreted as showing that our UDPS method may be used to study genetic variants and mutational linkage at least over relatively short distances.

4.2.5 Using the information of error frequencies

The results from the optimized experimental protocols in **Paper IV** gave us the possibility to detect minority variants and low frequency mutations. In **Paper IV**, the number of templates that we obtained from a sample ranged from 2,300 to 570,000, this corresponds to a theoretical sequencing depth of 0.04 % (1/2,300). The corresponding numbers in **Paper V** ranged from 56 to 93,632 which is equivalent to a lowest theoretical sequencing depth of 1.8 % (1/56). However, the sequencing depth was much deeper for most samples. Thus, the sequencing depth for most samples was primarily dependent on the error frequency and not by the number of templates.

We carefully evaluated how to derive accurate statistical cut-off values in **Paper IV**. The use of cut-off values was supposed to allow us to better distinguish between rare, but genuine, sequence variants and single-site drug resistance mutations from sequencing artifacts. It was well known that homopolymeric regions posed a particular problem in pyrosequencing. When we started the study we believed that different cut-offs should be used for homopolymeric respective non-homopolymeric regions. However, since this error bias was removed by our *in-house* cleaning strategy we did not need to distinguish between homopolymeric and non-homopolymeric regions. Instead, we used the information about the variation in site-specific error rates and derived individual cut-offs for all drug resistance position of interest. The average cut-off value for drug resistance position was 0.05 %, ranging between 0.014 % and 0.29 %. Using the same method, the cut-offs for high confidence variants was estimated to 0.11 % (range 0.09 to 0.21 %). With the knowledge obtained in **Paper II**, we would today instead estimate the cut-offs in **Paper IV** (which chronologically was the first published paper) based on the differences in error frequency of transition vs. transversion instead of individual positions.

4.3 METHODS TO FURTHER REDUCE THE IMPACT OF ERRORS.

Paper VI was a continuation of our efforts to generate reliable sequence data to study minority variants and low frequency mutations in different settings, e.g. in TDR. Also after we have been able to reduce the average error frequency with data cleaning (**Paper II**, **Paper IV**), substitution errors created during PCR remained.

One possible method to further reduce the error frequency is to lower the number of PCR cycles. It is well known that the error frequency increases with an increased number of PCR cycles [181, 182]. Zagordi *et al*. and Shao *et al*. both reported a significant reduction in UDPS error frequencies in clones that were not PCR amplified [179, 183]. The same was shown by Di Giallonardo *et al*. in a recently published study. However, the effect that a reduced number of PCR cycles have on NGS error frequencies is not yet thoroughly evaluated. Besides, a certain number of PCR-cycles are needed to get the sufficient amount of material needed for the subsequent UDPS, but at the same time newer platforms and protocols such as the Nextera XL - Illumina pipeline allows NGS on tiny amounts (picograms) of DNA which means that PCR amplification cycles will be fewer.

Theoretically, the use of a polymerase with higher fidelity could produce sequences with lower error frequencies. However, many of the existing enzymes with very high fidelity seem to have a lower processivity, which leads to lower sensitivity in the PCR step. In addition, it has been reported that polymerase with high fidelity are more prone to introduce *in vitro* recombination [183].

Random barcodes, referred to as Primer IDs, can be used to circumvent some of the remaining PCR and sequencing errors. With this approach, every individual cDNA molecule is marked with a Primer ID before the PCR and UDPS step and then resequenced multiple times [154, 155]. This relatively new method was first presented in an HIV study by Jabara *et al*. [156]. Independently, we had started on a similar approach as Jabara and colleagues. Our method is described in detail in Materials and Methods. The method has potential to provide consensus sequences with highly reduced error rates compared to standard UDPS for every cDNA template sequenced. Another advantage of the Primer ID method is that the exact number of templates sequenced is directly determined. This removes both the need to quantify the number of

templates and the risk of overestimation of variants by repeated sampling of the same original variants. In **Paper VI**, we present some important challenges that we have identified when applying the Primer ID UDPS method to sequencing of a clone and three patient samples. The challenges we found may influence the outcome of Primer ID sequencing.

Our first observation was that we found a very low number of consensus sequences despite a high number of cDNA input molecules. In the clone control experiment, less than two percent of input templates were sequenced and the results were similar for the patient samples. Skewed resampling of Primer IDs was one reason for the low number of recovered templates was. Some templates were sequenced several thousand times and others only a few times. We found several templates sequenced less than three times which could therefore not be used to construct consensus sequences.

The second observation was that the number of sequence templates where overestimated due to PCR errors in Primer IDs. Theoretically, over a million unique Primer IDs were available (4^{10} combinations). Despite this, we identified several groups of Primer IDs that differed by only one or a few nucleotides. It is statistically highly unlikely that such groups of closely related Primers IDs would represent correctly labeled template molecules. Therefore, we believe that these closely related Primer IDs originated from the same original template molecule and were created when PCR errors were introduced in the Primer IDs during PCR and/or 454 sequencing. This primarily occurred in templates that had been resampled many times.

Finally, despite the use of Primer IDs, 21 of 35 (60 %) consensus template sequences from the clone were incorrect. The majority of the remaining errors were deletions in homopolymeric regions, mainly in a single position. The same problem was observed in **Paper II**. We believe that these errors had been introduced during UDPS because the mutations were present in the majority, but not all, reads from the template and because the errors were "typical" 454 homopolymeric errors.

We, and others, have shown good repeatability for standard UDPS (**Paper III**). This indicates that the Primer ID itself or the long cDNA primers are the source of the skewed result. Since the Primer IDs are random, known potential problems (e.g. homopolymeric stretches, dimerization) cannot be avoided, which may lead to differences in PCR efficiency between templates labeled with different PIDs. Even very small differences in PCR efficiency in every cycle may be a problem because the differences will be magnified during PCR cycling. Designed, instead of random, Primer IDs, is a possible way to proceed and something we are currently testing. This approach was used by Shiroguchi *et al.* in a related experimental setup [184]. PrimerDesign, presented in **Paper I**, could be used to design Primer IDs to avoid structural problems.

An additional potential source of problems is that the Primer ID tags are attached during, instead of before, cDNA synthesis. This leaves a window where errors may occur during cDNA synthesis and remain undetected. This may result in an overestimation of the viral quasispecies. In a recent study by Metzner and colleagues showed that the *in vitro RT* induced errors as well as recombination is a risk in the cDNA synthesis [177].

In conclusion, Primer ID has potential to generate accurate sequence data but it is important to acknowledge the challenges of the methodology which needs to be overcome for the technique to be a truly successful solution.

4.4 DETECTION AND IMPACT OF MINORITY VARIANTS IN HIV-1

The quasispecies of related, but distinct, HIV-1 variants existing in an individual include several minority variants. Some of these minority variants may be of clinical relevance. It has been shown that minority variants with resistance mutations can contribute to treatment failure in patients. Particularly, pre-existing minority variants with NNRTI resistance mutations have been shown to be of clinical significance [175, 185-187].

In **Paper III**, we show that minority variants can be detected by our UDPS system which is presented in **Paper IV**. In **Paper III**, we identified minority variants representing down to 0.05 % of a population in analysis from two clones mixed at ratios of 99.5:0.5 and 99.95:0.05. The proportions of identified minority variants were somewhat higher than expected in both experiments, i.e. 2.2 % and 0.31 % respectively. This might be due to stochastic effects or that minority strains were systematically overestimated. However, it cannot be excluded that the artificial mixtures contained slightly higher proportions of the minor virus variant than intended. The statistical cut-offs derived in **Paper IV** allowed us to consider minority viral variants constituting on average 0.11 % of the population as high confidence variants and single drug resistance mutation 0.05 % on average.

The repeatability for our system was evaluated in **Paper III** and in conclusion we found that repeatability was good. All viral variants representing 0.27 % or more of the population were found in repeated UDPS analyses of two patient plasma samples. A similar degree of consistency was observed between forward and reverse reads. Somewhat surprisingly, the detection of major and minor variants showed similar repeatability.

We have studied the importance of minority variants in two longitudinal patient studies. In **Paper IV**, we investigated quasispecies dynamics during suboptimal treatment, treatment failure and treatment interruption. In **Paper V**, the viral variants were studied from PHI until after coreceptor switch from CCR5-using to CXCR4-using viruses with the intention to establish whether CXCR4-using variants existed as a minority population during PHI and/or early infection. We also analyzed the number of variants that established the infection in each of the three patients.

4.4.1 Pre existing drug-resistance

Low but significant levels of the M184I, T215A and T215I (range 0.02 %-0.12 %) drug resistance mutations were found in the treatment-naïve patients studied in **Paper IV**. This was rather expected since these resistance mutations only require a single mutation from “wild-type”. With the high genetic diversity of HIV-1, these mutations are expected to arise spontaneously every day. We looked for, but did not identify, significant pre-existence of the major drug resistance mutations M184V, Y181C, Y188C. All of which also require a single mutation. The absence could be explained by an increased fitness cost for the virus or the higher limit of detection in the position of interest (e.g. 0.15 % for M184V compared to 0.07 % for M184I). As discussed above in connection with **Paper II**, transitions and transversions have vastly different error frequencies. This knowledge could have been used in **Paper IV** if we would have had the results at that time. However, both the M184V and M184I are transitions and suggesting that the difference in their detection is not explained by this. The cause may instead be a context dependent error that we have not observed, but may also be due to

biologic reasons. The M184I mutation is known to be a transient state in the development of M184V during (3TC) treatment failure. There exist several theories as to why M184I tends to develop first only to replace by M184V. Our data suggest that a higher frequency of pre-existing M184I, may originate from HIV-1s higher A->G mutation frequency. T215Y/F was also screened for but not found. This is however an amino acid change that requires two nucleotide substitutions and is therefore less likely to occur.

4.4.2 Dynamics of HIV-1 quasispecies

The sensitivity of the experimental method and data filtering allowed us to study virus variants present in >0.11 % of the population. The sequencing length also allowed us to study linkage between mutations and to track individual variants through time. Five of the six patients studied in **Paper IV** were sampled before initiation of ART. As expected, different wild-type virus variants coexisted in the pre-treatment samples. The virus population remained heterogeneous after initiation of treatment. During treatment failure, diversity gradually decreased concurrent with the evolution of specific, linked drug resistance mutations. Particularly, M184V-T215Y variants were found. This suggests that they were more fit during selective pressure from the treatment regimen, which consisted of 3TC, d4T and ddI. Wild-type variants were only detected in one patient during treatment failure, suggesting that they have very low fitness in the presence of ART and that the contribution is very low from the latent reservoirs, where wild-type virus should have be archived. The drug resistant variants were replaced by wild-type soon after removal of ART, which is a proof that such virus was indeed archived. This was not unexpected since drug resistant variants, especially many resistance mutations are known to have reduced fitness in the absence of drugs [130, 188]. Thus, it is very likely that wild-type virus variants were present at levels below our detection limit during treatment failure.

4.4.3 Transmitted virus and coreceptor switch during PHI

Samples from the three patients taken during PHI were analyzed in **Paper V**. The samples were used to study the evolutionary relationships in the virus population. Two of the patients appeared to have one founder variant. The largest minority variant represented 0.19 % or less of the population in these two patients. These variants are likely to have evolved from the founder virus after transmission because the genetic diversity was stochastically distributed with any single variant carrying at most one or two mutations relative to the founder virus. In one of the patients, two or three viruses established the infection. The three major variants made out 47 %, 38 % and 10 % of the virus population, respectively. The second and third most common variants differed from the first variant by a minimum of four nucleotides, which makes it highly unlikely that they evolved after transmission. As expected, all three patients had a low genetic diversity during PHI.

Our result is consistent with recent observations that only one or a few viruses establish the infection following transmission to a new host. Studies have suggested that the proportion of infections that are founded by a single variant differs according to route of transmission, with IDUs and MSM more often having two or more founding variants than heterosexually infected patients [49, 62]. Our patients were MSM, which means that it is expected to observe transmission of more than one variant in approximately 60 % of cases. However, since we only studied three patients, no firm conclusions can be drawn from our data [49].

UDPS analysis did not show any indication on presence of X4-using minority variants in any of the three patients during PHI or prior coreceptor switch as detected by the MT-2 assay. However, the presence of X4 for a shorter time period before coreceptor switch cannot be excluded since the samples obtained prior to switch were drawn a minimum of 17 months before the switch. Bunnik and colleagues reported that X4 variants usually evolved gradually during a 12-month period prior to overt coreceptor switch [100]. In agreement with this, our phylogenetic analysis indicated that the X4 populations originated from R5 variants that evolved after the last R5-only sample was obtained. This strengthens the theory that a one or a few, primarily R5-using viruses are the predecessors of the X4 population. However, our result does not completely rule out the possibility that minority X4 variants transmitted and remained present at levels below the detection limit of our UDPS assay until overt coreceptor switch. The three individuals studied in **Paper V** had an atypical course of infection with a rapidly progressing immunodeficiency. This is consistent with the observed coreceptor switch, but again it should be stressed that the patients were too few draw any general conclusions.

5 CONCLUSIONS AND FUTURE PERSPECTIVE

HIV-1 is a virus with a very variable genome. It has the ability to adapt to changes in the environment and thereby escape both immune pressure and suboptimal ART. NGS, and especially UDPS, has enabled deep sequencing studies with unprecedented resolution, but the technology is more error prone than traditional sequencing.

The comprehensive work to identify, characterize and reduce errors as well as investigate the UDPS performance performed in **Papers II**, **III** and **IV** has allowed more accurate interpretations of the biological findings in **Papers IV** and **V**. It also encouraged us to develop the novel software in **Paper I** and the new method developed and evaluated in **Paper VI.** In **Paper I**, we developed a novel computer program, named PrimerDesign. It is tailored to designs primers from a multiple alignment and is suitable for all types of NGS that is preceded by amplification. The algorithm was successfully used in studies of HIV-1 and should be equally useful for designing primers targeting other organisms independent of the level of genetic variation.

The NGS technology has enabled the entire HIV-genome to be deep sequenced. In an article by Henn *et al*. [142], the HIV genome was amplified in overlapping regions but with varied coverage. A possible future development for PrimerDesign is to extend the algorithm to find several, compatible, primer pairs in a longer region, e.g. the complete genome. This would save laboratory time as samples may be multiplexed and allow even deeper sequencing. In an ongoing project, PrimerDesign was used to design primers for effective and high-coverage Illumina sequencing of entire HIV-genomes.

We optimized the pre-UDPS protocol in **Paper IV** and investigated the characteristics and sources of errors that occurred when UDPS was used to sequence a fragment of the HIV-1 *pol* gene in **Paper II**. Our in-house data cleaning software removed UDPS-introduced indel errors in homopolymeric regions. The remaining errors were primarily substitution errors that were introduced in the PCR that preceded UDPS. Transitions were significantly more frequent than transversions, which will limit detection of minor variants and mutations in HIV-1 as well as other species. Importantly, this problem is applicable to all NGS platforms where sequencing is preceded by PCR. We further evaluated the quality and reproducibility of the UDPS technology in **Paper III.** We concluded that repeatability was good, both for majority and minority variants. In our experimental settings, *in vitro* recombination and sequencing directions posed a minor problem, but still needs to be considered especially for minor viral variants and studies of linkage between mutations. The design of primers is of particular importance in UDPS to avoid selective amplification which may skew the result of frequency estimations.

Because the NGS technologies are evolving very rapidly, the 454 sequencing approach that we have used in this thesis is not the method we would have used if we had started the projects today. Instead, we would probably have decided to use Illumina or possibly Ion Torrent. Illumina has the advantage of a lower error frequency, higher throughput and an easier workflow, but shorter read length. However, the read length has increased and recent pair-end sequencing protocols on the MiSeq Illumina platform has a read length of approximately 600 bases, which is sufficient for many applications. Ion Torrent generates relatively long reads, but makes the same type of errors (indels) as 454. Both are cost efficient and generates substantially more data compared to 454. The Pacific Biosciences´ platform is also an interesting platform, which offers very long read lengths, but unfortunately it also has a quite high error frequency.

One limitation to cross field work, even within the same field, is the use of different nomenclature. With the increasing development speed, I think it is of particular interest to have a joint language to make analysis simpler. One example is when a script is used to parse a sequence file created by someone else. If no standardized method to name the sequences has been used, it results in problem to create automated tools for simple analysis, which leads to time consuming manual work. Another example is the method we refer to as Primer ID. At least three different names in two different research areas are used for the molecule tagging approach. As a contrast, the quality score associated with every sequenced nucleotide for Sanger sequencing and 454 sequencing are both called Phred but are not equivalent.

Minority variants and drug resistance mutations were studied in **Papers IV** and **V**. We examined the presence of pre-existing drug resistance mutations in treatment-naïve HIV-1 patients and found very low levels of M184I, T215A and T215I, but no presence of M184V, Y181C, Y188C or T215Y/F. This indicated that the natural occurrence of these mutations was very low, i.e. below our detection limit. When patients experienced treatment failure almost 100 % of the wild-type virus was replaced with drug sensitive variants and when therapy was interrupted, 100 % of the drug resistance variants were replaced with wild-type. The quasispecies in patients followed from PHI to a coreceptor switch were investigated in **Paper V**. We did not find any X4-using virus present as a minority species during PHI. The results indicate that the X4 population most probably stepwise evolved *de novo* from the R5 populations in each of the three patients.

Minority drug resistance mutation and minority variants of the virus coreceptor tropism have both been shown to play an important role in successful ART. Already today, detection and quantification of drug resistance is recommended for treatment initialization and the standard care for patients failing therapy and requiring new cART. I believe that we will see an increased use of NGS sequencing instruments in both routine and research laboratories, which will be very beneficial. Hospitals and research laboratories working with sequencing will have their own bench top sequencer within a couple of years and whole genome sequencing will be performed on more or less a daily basis. This scenario could benefit patients by providing additional possibilities and accuracy in personalized treatment. As a consequence, the cost for resistance testing and other sequencing will temporarily increase. Bioinformatic expertise will become even more needed to interpret and handle the data generated. The rapid development of NGS technology will require continuous development of new methods to adjust and take advantage of newer NGS platform, just as I have done in this project.

Successful treatment with the CCR5-antagonist maraviroc is dependent of the presence of solely R5-using virus in the patient. I would recommend more studies of transmission pairs to further evaluate whether R5-using virus is selected for during transmission or not. I would not be surprised if the results show that the likelihood of R5 or X4 transmission is proportional to their abundance in the donor. This would of course support the use of coreceptor tropism prediction before treatment initiation, but also already at the time of diagnosis. The coreceptor use might affect when treatment should be initiated since X4-using virus is associated with a faster disease progression.

The Primer ID methodology has the potential to provide highly accurate deep sequencing. We identified three major challenges (**Paper VI**); a skewed resampling of Primer IDs, low recovery of templates and erratic consensus sequences. These

problems can lead to an underestimation of the diversity of the quasispecies as well as skewed or incorrect results if they are not detected. As many of our other findings, our results concerning the Primer ID approach is not limited to HIV or virology. We are currently evaluating the Primer ID methodology on other NGS platforms with promising results.

In the future, all parts of the sequencing process will be further optimized, from the pre-sequencing experimental protocols, via sequencing platforms, to the data interpretation. Every time consuming step will be considered a bottleneck in an otherwise streamlined process. Read lengths will increase. Already today, the RS II from Pacific Bioscience generates reads with an average read length of 4,200 to 8,500 base pairs and the longest reads cover over 30,000 base pairs. 454 (Roche) recently presented a new improved chemistry, GS FLX + which is said to have the capability to generate reads up to 1,000 bases. The error frequencies will be reduced. The Nextera XL - Illumina pipeline allows NGS to start from tiny amounts of DNA which reduces the PCR cycles needed and thereby reduces the introduction of the PCR errors. The lowered error frequencies will not only depend on sequencing free from errors but, just as we have attempted to reduce error frequency in **Paper VI**, other methods to circumvent the errors will be developed and improved. Pacific Bioscience´s sequencing libraries are made from circular DNA molecules with adapters (hairpin loops) ligated at both ends of the DNA insert, the raw sequence reads often contain multiple determinations of the DNA insert sequence, separated by the adapter sequences. This allows a user to extract the consensus sequence, but full potential of the longer read length is still blunted by artificial recombination occurring in the PCR that precedes sequencing. Sequence data will be generated faster and cheaper. I believe that the big challenge in the future is to efficiently carry out data analysis and store the enormous amount of data. It will be even more crucial to develop pipelines where as little manual work as possible is required.

The possibilities of data storage are rapidly developing but the costs for archiving data can however be considerable. Storage must be done in an efficient way for two main reasons. Larger collaboration is often needed in these types of analysis and data must be possible to send between people and locations. It is also important for the transparency that others are able to access data after publication. Today, Sanger sequences and to some extent NGS data is stored in large public databases, but other solutions are required for the growing amounts of NGS data. Another problem with both transparency and comparison between studies is the lack of standardized methods to state which methods are being used in the particular experiment. Conclusions from experiments are being drawn after numerous steps data cleaning, normalization of data, the use of cut-off values etc., sometimes without being fully declared.

Many of the applications that are being developed, including our methods and software, reach further than to HIV and virology. Genomics research in general would gain from more cross-field collaboration and interaction.

In conclusion, we have developed and used new NGS and bioinformatic methods to study genetic variation and evolution in HIV-1. We showed that UDPS can be used to gain new insights in HIV evolution and to detect minority drug resistance mutations as well as minority variants.

I believe that we only have seen the beginning of the sequencing revolution.

6 ACKNOWLEDGEMENTS

Many people have contributed and deserve to be acknowledged for the making of this thesis. During my years as a PhD-student I have learned more than I ever hoped for and met some wonderful persons. I would like to take the opportunity to mention some of you and to say “Thank you”, to all of you.

Jan Albert, my supervisor. Janne, jag kunde inte ha önskat mig en bättre handledare. Tack för att jag har fått vara en del av din forskningsgrupp. Du har alltid haft tid och en öppen dörr, även när du har varit upptagen. Du är en fantastik lärare och chef. Jag beundrar din breda kunskap och är glad för ditt positiva synsätt på resultat. Jag är tacksam för att du hela tiden har uppmutrat mig till att utvecklas genom att prova nya saker och besöka nya platser.

Thomas Leitner, my co-supervisor. Tack för att du välkomnade mig till din grupp på LANL och får att du alltid får mig att känna mig som att jag kan saker. Att diskutera vetenskap och andra livsviktiga frågor med dig har varit både underhållande och mycket lärorikt.

Björn Andersson, my co-supervisor. Det var du som först välkomnade mig till KI. Tack för att jag fick vara en del av din grupp i början och för alla fortsatta givande samtal.

Richard Neher, Thank you for great scientific collaboration and good times after work. I’m so glad I got the opportunity to meet you. I have felt very welcome in both Santa Barbara and Teubingen.

Sven Britton, Ghana var fantastiskt, du gjorde det till en exceptionellt lärorik resa med din antusiasm och förmåga att engagera.

Benita Zweygberg Wirgart, som både välkomnade oss till mikrobiologen och var en exemplarisk mikrobiologiexaminator.

Collaborators and co-authors, **Göran Bratt**, **Bette Korber**, **Mohan Krishnamoorthy**, **Gayathri Athreya**, **Will Fischer**, **Peter Hraber**, **Cheryl Gleasner** and **Lance Green**. It has been a pleasure working with you all.

Colleagues and friends at KI/KS/SMI/LANL
Thank you/Tack till: Charlotte Hedskog, för alla gemensamma projekt som inte skulle kunna ha genomförts i närheten av lika bra utan dig, våra roliga resor, givande samtal och för att du har blivit min fina vän. **Lina Thebo** för all labhjälp och för att du förgyller mina dagar på kontoret. **Mattias Mild**, för din positiva energi och bra projektsamarbeten. **Ewa Ericsson** för att du har visat mig hur labarbete ska gå till och alltid är hjälpsam.

Lina Odevall, för alla äventyr och inspirerande samtal. Det finns ingen som jag hellre delar skrivbord med än dig, min vän! **Susanne von Stockenström**, för många härliga och trevliga samtal på kontoret samt roliga upptåg i Seattle. **Viktor Dahl** för roliga och givande diskussioner. **Sarah Palmer** and **Bates Gill** for being so including, crazy and wonderful. **Helena Skar** för att du är så inspirerande. **Alexander Hiddini**, för dina

svåra frågor som har tvingat mig att tänka efter. **Salma Nowroozalizadeh**, för fina samtal och för att du fortsätter att hålla ihop oss. **Joakim Esbjörnsson**, för att du alltid är så hjälpsam och positiv. **Wendy Murillo**, for always spreading happiness. **Leda Parham**, **Carina Perez**, **Dace Balode**, **Melissa Norström**, **Marcus Buggert** and **Irina Maljkovic Berry**, for all nice discussions. You have been the best roommates. **Ellen Sherwood**, för all hjälp när jag först kom till KI. **Åsa Onshagen** som pratade och skrattade sig igenom ett halvår av projektarbete och blev min fina vän. **Marianne Jansson**, **Annika Karlsson**, **Kajsa Apetina** and **Maria Axelsson** för hjälp med prover, material, glada tillrop och trevliga samtal. **Lisbeth Löfstrand** för ovärderlig administrativ hjälp.

Tack till PhD Club, **Therese Högfeldt** och **Cecilia Jädert**. Det har varit ett nöje att arbeta tillsammans med så drivna och inspirerande tjejer som er och övriga medlemmar.

Tack till **alla mina fantastiska vänner** utanför min akademiska värld. Vi har pratat oss igenom trevliga middagar och gått på långa och korta promenader som ger mig positiv energi. Ni har också gett mig värdefulla, praktiska förslag. Tack **Sabina Hjerppe** och **Anna Dovärn** för er ovärderliga vänskap och uppmuntran längs vägen. Ett särskilt tack till **Anna Sahlberg** som förutom att ha varit en underbar vän också har har hjälpt mig med bilder. Jag är så glad över att du, och din familj, så länge har varit en så fin del av mitt liv.

Stora familjen Brodin, **Ernst**, **Kinna**, **Jojjo**. **Anna** och Lagercrantz, **Svetlana**, **Karolina**, **Marcus**, **Alexander** och **Victor**. Ernst, tack för en bra introduktion till KI som jag aldrig skulle ha fått utan dig. Tack till er alla för uppmuntran och framför allt för att jag har fått en så extrafamilj.

Mamma, **Matilda** och **Emelie** med fina familjer. Tack för all uppmuntran och för att ni alltid bara är ett telefonsamtal bort.

Pappa – Tack för allt stöd, för alla heja-på samtal, för att du alltid har trott på mig och fått mig att känna mig att att jag kan göra precis vad jag vill.

Min bästaste, underbara familj. **Kristofer**, jag är så glad över att ha dig vid min sida. Jag skulle aldrig ha gjort den här resan utan din uppmuntran. Tack för all värdefull hjälp under vägen och med avhandlingen. **Theodor**, världens finaste, finaste lille kille. Du får mig att vilja göra allting lite bättre. Tack för all kärlek från er båda. Nu fortsätter vi vår färd framåt med nya familjeprojekt.

7 REFERENCES

1. Kaposi's sarcoma and Pneumocystis pneumonia among homosexual men--New York City and California. *MMWR Morb Mortal Wkly Rep* 1981,**30**:305-308.
2. Greene WC. A history of AIDS: looking back to see ahead. *Eur J Immunol* 2007,**37 Suppl 1**:S94-102.
3. Barre-Sinoussi F, Chermann JC, Rey F, Nugeyre MT, Chamaret S, Gruest J, *et al.* Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science* 1983,**220**:868-871.
4. Gallo RC, Salahuddin SZ, Popovic M, Shearer GM, Kaplan M, Haynes BF, *et al.* Frequent detection and isolation of cytopathic retroviruses (HTLV-III) from patients with AIDS and at risk for AIDS. *Science* 1984,**224**:500-503.
5. Popovic M, Sarngadharan MG, Read E, Gallo RC. Detection, isolation, and continuous production of cytopathic retroviruses (HTLV-III) from patients with AIDS and pre-AIDS. *Science* 1984,**224**:497-500.
6. UNAIDS. UNAIDS Report on the Global AIDS Epidemic- 2013. In; 2013.
7. Sharp PM, Hahn BH. The evolution of HIV-1 and the origin of AIDS. *Philos Trans R Soc Lond B Biol Sci* 2010,**365**:2487-2494.
8. Andersson S, Norrgren H, da Silva Z, Biague A, Bamba S, Kwok S, *et al.* Plasma viral load in HIV-1 and HIV-2 singly and dually infected individuals in Guinea-Bissau, West Africa: significantly lower plasma virus set point in HIV-2 infection than in HIV-1 infection. *Arch Intern Med* 2000,**160**:3286-3293.
9. Marlink R, Kanki P, Thior I, Travers K, Eisen G, Siby T, *et al.* Reduced rate of disease development after HIV-2 infection as compared to HIV-1. *Science* 1994,**265**:1587-1590.
10. Hahn BH, Shaw GM, De Cock KM, Sharp PM. AIDS as a zoonosis: scientific and public health implications. *Science* 2000,**287**:607-614.
11. Worobey M, Gemmel M, Teuwen DE, Haselkorn T, Kunstman K, Bunce M, *et al.* Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature* 2008,**455**:661-664.
12. Taylor BS, Hammer SM. The challenge of HIV-1 subtype diversity. *N Engl J Med* 2008,**359**:1965-1966.
13. Peeters M, Jung M, Ayouba A. The origin and molecular epidemiology of HIV. *Expert Rev Anti Infect Ther* 2013,**11**:885-896.
14. De Leys R, Vanderborght B, Vanden Haesevelde M, Heyndrickx L, van Geel A, Wauters C, *et al.* Isolation and partial characterization of an unusual human immunodeficiency retrovirus from two persons of west-central African origin. *J Virol* 1990,**64**:1207-1216.
15. Gurtler LG, Hauser PH, Eberle J, von Brunn A, Knapp S, Zekeng L, *et al.* A new subtype of human immunodeficiency virus type 1 (MVP-5180) from Cameroon. *J Virol* 1994,**68**:1581-1585.
16. Simon F, Mauclere P, Roques P, Loussert-Ajaka I, Muller-Trutwin MC, Saragosti S, *et al.* Identification of a new human immunodeficiency virus type 1 distinct from group M and group O. *Nat Med* 1998,**4**:1032-1037.
17. Vallari A, Bodelle P, Ngansop C, Makamche F, Ndembi N, Mbanya D, *et al.* Four new HIV-1 group N isolates from Cameroon: Prevalence continues to be low. *AIDS Res Hum Retroviruses* 2010,**26**:109-115.
18. Plantier JC, Leoz M, Dickerson JE, De Oliveira F, Cordonnier F, Lemee V, *et al.* A new human immunodeficiency virus derived from gorillas. *Nat Med* 2009,**15**:871-872.

19. RAV RfAT. Antiretroviral behandling av HIV-infektion 2013, uppdaterad version 2014-02-10. In; 2013.
20. Folkhälsomyndigheten. HIV-infektion-epidemiska årsrapport 2012-2013. In; 2013.
21. Maddon PJ, Dalgleish AG, McDougal JS, Clapham PR, Weiss RA, Axel R. The T4 gene encodes the AIDS virus receptor and is expressed in the immune system and the brain. *Cell* 1986,**47**:333-348.
22. McDougal JS, Kennedy MS, Sligh JM, Cort SP, Mawle A, Nicholson JK. Binding of HTLV-III/LAV to T4+ T cells by a complex of the 110K viral protein and the T4 molecule. *Science* 1986,**231**:382-385.
23. Berger EA, Doms RW, Fenyo EM, Korber BT, Littman DR, Moore JP, *et al.* A new classification for HIV-1. *Nature* 1998,**391**:240.
24. Wilen CB, Tilton JC, Doms RW. HIV: cell binding and entry. *Cold Spring Harb Perspect Med* 2012,**2**.
25. Duffy S, Shackelton LA, Holmes EC. Rates of evolutionary change in viruses: patterns and determinants. *Nat Rev Genet* 2008,**9**:267-276.
26. Korber B, Gaschen B, Yusim K, Thakallapally R, Kesmir C, Detours V. Evolutionary and immunological implications of contemporary HIV-1 variation. *Br Med Bull* 2001,**58**:19-42.
27. Lauring AS, Andino R. Quasispecies theory and the behavior of RNA viruses. *PLoS Pathog* 2010,**6**:e1001005.
28. Eigen M, Schuster P. The hypercycle. A principle of natural self-organization. Part A: Emergence of the hypercycle. *Naturwissenschaften* 1977,**64**:541-565.
29. Meyerhans A, Cheynier R, Albert J, Seth M, Kwok S, Sninsky J, *et al.* Temporal fluctuations in HIV quasispecies in vivo are not reflected by sequential HIV isolations. *Cell* 1989,**58**:901-910.
30. Coffin J, Swanstrom R. HIV pathogenesis: dynamics and genetics of viral populations and infected cells. *Cold Spring Harb Perspect Med* 2013,**3**:a012526.
31. Wei X, Ghosh SK, Taylor ME, Johnson VA, Emini EA, Deutsch P, *et al.* Viral dynamics in human immunodeficiency virus type 1 infection. *Nature* 1995,**373**:117-122.
32. Perelson AS, Neumann AU, Markowitz M, Leonard JM, Ho DD. HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. *Science* 1996,**271**:1582-1586.
33. Ho DD, Neumann AU, Perelson AS, Chen W, Leonard JM, Markowitz M. Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. *Nature* 1995,**373**:123-126.
34. Preston BD. Reverse transcriptase fidelity and HIV-1 variation. *Science* 1997,**275**:228-229; author reply 230-221.
35. Mansky LM, Temin HM. Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. *J Virol* 1995,**69**:5087-5094.
36. O'Neil PK, Sun G, Yu H, Ron Y, Dougherty JP, Preston BD. Mutational analysis of HIV-1 long terminal repeats to explore the relative contribution of reverse transcriptase and RNA polymerase II to viral mutagenesis. *J Biol Chem* 2002,**277**:38053-38061.
37. Abram ME, Ferris AL, Shao W, Alvord WG, Hughes SH. Nature, position, and frequency of mutations made in a single cycle of HIV-1 replication. *J Virol* 2010,**84**:9864-9878.

38. Shriner D, Shankarappa R, Jensen MA, Nickle DC, Mittler JE, Margolick JB, *et al.* Influence of random genetic drift on human immunodeficiency virus type 1 env evolution during chronic infection. *Genetics* 2004,**166**:1155-1164.
39. Svarovskaia ES, Cheslock SR, Zhang WH, Hu WS, Pathak VK. Retroviral mutation rates and reverse transcriptase fidelity. *Front Biosci* 2003,**8**:d117-134.
40. Neher RA, Leitner T. Recombination rate and selection strength in HIV intra-patient evolution. *PLoS Comput Biol* 2010,**6**:e1000660.
41. Kimura M. Change of gene frequencies by natural selection under population number regulation. *Proc Natl Acad Sci U S A* 1978,**75**:1934-1937.
42. Tindall B, Cooper DA. Primary HIV infection: host responses and intervention strategies. *AIDS* 1991,**5**:1-14.
43. Schacker T, Collier AC, Hughes J, Shea T, Corey L. Clinical and epidemiologic features of primary HIV infection. *Ann Intern Med* 1996,**125**:257-264.
44. Ribeiro RM, Qin L, Chavez LL, Li D, Self SG, Perelson AS. Estimation of the initial viral growth rate and basic reproductive number during acute HIV-1 infection. *J Virol* 2010,**84**:6096-6102.
45. Veazey RS, DeMaria M, Chalifoux LV, Shvetz DE, Pauley DR, Knight HL, *et al.* Gastrointestinal tract as a major site of CD4+ T cell depletion and viral replication in SIV infection. *Science* 1998,**280**:427-431.
46. Guadalupe M, Reay E, Sankaran S, Prindiville T, Flamm J, McNeil A, *et al.* Severe CD4+ T-cell depletion in gut lymphoid tissue during primary human immunodeficiency virus type 1 infection and substantial delay in restoration following highly active antiretroviral therapy. *J Virol* 2003,**77**:11708-11717.
47. Brenchley JM, Schacker TW, Ruff LE, Price DA, Taylor JH, Beilman GJ, *et al.* CD4+ T cell depletion during all stages of HIV disease occurs predominantly in the gastrointestinal tract. *J Exp Med* 2004,**200**:749-759.
48. Haase AT. Targeting early infection to prevent HIV-1 mucosal transmission. *Nature* 2010,**464**:217-223.
49. Shaw GM, Hunter E. HIV transmission. *Cold Spring Harb Perspect Med* 2012,**2**.
50. Fiebig EW, Wright DJ, Rawal BD, Garrett PE, Schumacher RT, Peddada L, *et al.* Dynamics of HIV viremia and antibody seroconversion in plasma donors: implications for diagnosis and staging of primary HIV infection. *AIDS* 2003,**17**:1871-1879.
51. Lifson AR, Buchbinder SP, Sheppard HW, Mawle AC, Wilber JC, Stanley M, *et al.* Long-term human immunodeficiency virus infection in asymptomatic homosexual and bisexual men with normal CD4+ lymphocyte counts: immunologic and virologic characteristics. *J Infect Dis* 1991,**163**:959-965.
52. Hladik F, McElrath MJ. Setting the stage: host invasion by HIV. *Nat Rev Immunol* 2008,**8**:447-457.
53. Powers KA, Poole C, Pettifor AE, Cohen MS. Rethinking the heterosexual infectivity of HIV-1: a systematic review and meta-analysis. *Lancet Infect Dis* 2008,**8**:553-563.
54. Boily MC, Baggaley RF, Wang L, Masse B, White RG, Hayes RJ, *et al.* Heterosexual risk of HIV-1 infection per sexual act: systematic review and meta-analysis of observational studies. *Lancet Infect Dis* 2009,**9**:118-129.
55. Muessig KE, Smith MK, Powers KA, Lo YR, Burns DN, Grulich AE, *et al.* Does ART prevent HIV transmission among MSM? *AIDS* 2012,**26**:2267-2273.
56. Quinn TC, Wawer MJ, Sewankambo N, Serwadda D, Li C, Wabwire-Mangen F, *et al.* Viral load and heterosexual transmission of human immunodeficiency virus type 1. Rakai Project Study Group. *N Engl J Med* 2000,**342**:921-929.

57. Fideli US, Allen SA, Musonda R, Trask S, Hahn BH, Weiss H, *et al.* Virologic and immunologic determinants of heterosexual transmission of human immunodeficiency virus type 1 in Africa. *AIDS Res Hum Retroviruses* 2001,**17**:901-910.
58. Galvin SR, Cohen MS. The role of sexually transmitted diseases in HIV transmission. *Nat Rev Microbiol* 2004,**2**:33-42.
59. Wolfs TF, Zwart G, Bakker M, Goudsmit J. HIV-1 genomic RNA diversification following sexual and parenteral virus transmission. *Virology* 1992,**189**:103-110.
60. Wolinsky SM, Wike CM, Korber BT, Hutto C, Parks WP, Rosenblum LL, *et al.* Selective transmission of human immunodeficiency virus type-1 variants from mothers to infants. *Science* 1992,**255**:1134-1137.
61. Ping LH, Joseph SB, Anderson JA, Abrahams MR, Salazar-Gonzalez JF, Kincer LP, *et al.* Comparison of viral Env proteins from acute and chronic infections with subtype C human immunodeficiency virus type 1 identifies differences in glycosylation and CCR5 utilization and suggests a new strategy for immunogen design. *J Virol* 2013,**87**:7218-7233.
62. Keele BF, Giorgi EE, Salazar-Gonzalez JF, Decker JM, Pham KT, Salazar MG, *et al.* Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proc Natl Acad Sci U S A* 2008,**105**:7552-7557.
63. Li M, Gao F, Mascola JR, Stamatatos L, Polonis VR, Koutsoukos M, *et al.* Human immunodeficiency virus type 1 env clones from acute and early subtype B infections for standardized assessments of vaccine-elicited neutralizing antibodies. *J Virol* 2005,**79**:10108-10125.
64. Salazar-Gonzalez JF, Salazar MG, Keele BF, Learn GH, Giorgi EE, Li H, *et al.* Genetic identity, biological phenotype, and evolutionary pathways of transmitted/founder viruses in acute and early HIV-1 infection. *J Exp Med* 2009,**206**:1273-1289.
65. Cohen MS, Smith MK, Muessig KE, Hallett TB, Powers KA, Kashuba AD. Antiretroviral treatment of HIV-1 prevents transmission of HIV-1: where do we go from here? *Lancet* 2013,**382**:1515-1524.
66. Cohen MS, Chen YQ, McCauley M, Gamble T, Hosseinipour MC, Kumarasamy N, *et al.* Prevention of HIV-1 infection with early antiretroviral therapy. *N Engl J Med* 2011,**365**:493-505.
67. Gray R, Kigozi G, Kong X, Ssempiija V, Makumbi F, Wattya S, *et al.* The effectiveness of male circumcision for HIV prevention and effects on risk behaviors in a posttrial follow-up study. *AIDS* 2012,**26**:609-615.
68. Quinn TC. Circumcision and HIV transmission. *Curr Opin Infect Dis* 2007,**20**:33-38.
69. Medrano J, Soriano V. [Mother-to-child transmission of HIV infection in the era of highly active antiretroviral therapy]. *Med Clin (Barc)* 2009,**132**:505-506.
70. Kourtis AP, Lee FK, Abrams EJ, Jamieson DJ, Bulterys M. Mother-to-child transmission of HIV-1: timing and implications for prevention. *Lancet Infect Dis* 2006,**6**:726-732.
71. McNairy ML, Cohen M, El-Sadr WM. Antiretroviral therapy for prevention is a combination strategy. *Curr HIV/AIDS Rep* 2013,**10**:152-158.
72. Coates TJ, Richter L, Caceres C. Behavioural strategies to reduce HIV transmission: how to make them work better. *Lancet* 2008,**372**:669-684.
73. Leitner T, Kumar S, Albert J. Tempo and mode of nucleotide substitutions in gag and env gene fragments in human immunodeficiency virus type 1 populations with a known transmission history. *J Virol* 1997,**71**:4761-4770.

74. Bagnarelli P, Mazzola F, Menzo S, Montroni M, Butini L, Clementi M. Host-specific modulation of the selective constraints driving human immunodeficiency virus type 1 env gene evolution. *J Virol* 1999,**73**:3764-3777.
75. Carrillo A, Ratner L. Cooperative effects of the human immunodeficiency virus type 1 envelope variable loops V1 and V3 in mediating infectivity for T cells. *J Virol* 1996,**70**:1310-1316.
76. Jansson M, Backstrom E, Scarlatti G, Bjorndal A, Matsuda S, Rossi P, *et al.* Length variation of glycoprotein 120 V2 region in relation to biological phenotypes and coreceptor usage of primary HIV type 1 isolates. *AIDS Res Hum Retroviruses* 2001,**17**:1405-1414.
77. Labrosse B, Treboute C, Brelot A, Alizon M. Cooperation of the V1/V2 and V3 domains of human immunodeficiency virus type 1 gp120 for interaction with the CXCR4 receptor. *J Virol* 2001,**75**:5457-5464.
78. de Jong JJ, Goudsmit J, Keulen W, Klaver B, Krone W, Tersmette M, *et al.* Human immunodeficiency virus type 1 clones chimeric for the envelope V3 domain differ in syncytium formation and replication capacity. *J Virol* 1992,**66**:757-765.
79. Milich L, Margolin BH, Swanstrom R. Patterns of amino acid variability in NSI-like and SI-like V3 sequences and a linked change in the CD4-binding domain of the HIV-1 Env protein. *Virology* 1997,**239**:108-118.
80. Arrildt KT, Joseph SB, Swanstrom R. The HIV-1 env protein: a coat of many colors. *Curr HIV/AIDS Rep* 2012,**9**:52-63.
81. Chalmet K, Dauwe K, Foquet L, Baatz F, Seguin-Devaux C, Van Der Gucht B, *et al.* Presence of CXCR4-using HIV-1 in patients with recently diagnosed infection: correlates and evidence for transmission. *J Infect Dis* 2012,**205**:174-184.
82. Raymond S, Delobel P, Mavigner M, Cazabat M, Encinas S, Souyris C, *et al.* CXCR4-using viruses in plasma and peripheral blood mononuclear cells during primary HIV-1 infection and impact on disease progression. *AIDS* 2010,**24**:2305-2312.
83. Fiore JR, Bjorndal A, Peipke KA, Di Stefano M, Angarano G, Pastore G, *et al.* The biological phenotype of HIV-1 is usually retained during and after sexual transmission. *Virology* 1994,**204**:297-303.
84. Zhu T, Mo H, Wang N, Nam DS, Cao Y, Koup RA, *et al.* Genotypic and phenotypic characterization of HIV-1 patients with primary infection. *Science* 1993,**261**:1179-1181.
85. Abbate I, Vlassi C, Rozera G, Bruselles A, Bartolini B, Giombini E, *et al.* Detection of quasispecies variants predicted to use CXCR4 by ultra-deep pyrosequencing during early HIV infection. *AIDS* 2011,**25**:611-617.
86. Dean M, Carrington M, Winkler C, Huttley GA, Smith MW, Allikmets R, *et al.* Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene. Hemophilia Growth and Development Study, Multicenter AIDS Cohort Study, Multicenter Hemophilia Cohort Study, San Francisco City Cohort, ALIVE Study. *Science* 1996,**273**:1856-1862.
87. Samson M, Libert F, Doranz BJ, Rucker J, Liesnard C, Farber CM, *et al.* Resistance to HIV-1 infection in caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. *Nature* 1996,**382**:722-725.
88. Liu R, Paxton WA, Choe S, Ceradini D, Martin SR, Horuk R, *et al.* Homozygous defect in HIV-1 coreceptor accounts for resistance of some multiply-exposed individuals to HIV-1 infection. *Cell* 1996,**86**:367-377.

89. Hutter G, Nowak D, Mossner M, Ganepola S, Mussig A, Allers K, *et al.* Long-term control of HIV by CCR5 Delta32/Delta32 stem-cell transplantation. *N Engl J Med* 2009,**360**:692-698.
90. Piacentini L, Biasin M, Fenizia C, Clerici M. Genetic correlates of protection against HIV infection: the ally within. *J Intern Med* 2009,**265**:110-124.
91. Hedskog C, Mild M, Albert J. Transmission of the X4 phenotype of HIV-1: is there evidence against the "random transmission" hypothesis? *J Infect Dis* 2012,**205**:163-165.
92. Koot M, Vos AH, Keet RP, de Goede RE, Dercksen MW, Terpstra FG, *et al.* HIV-1 biological phenotype in long-term infected individuals evaluated with an MT-2 cocultivation assay. *AIDS* 1992,**6**:49-54.
93. Karlsson A, Parsmyr K, Sandstrom E, Fenyo EM, Albert J. MT-2 cell tropism as prognostic marker for disease progression in human immunodeficiency virus type 1 infection. *J Clin Microbiol* 1994,**32**:364-370.
94. Schuitemaker H, van 't Wout AB, Lusso P. Clinical significance of HIV-1 coreceptor usage. *J Transl Med* 2011,**9 Suppl 1**:S5.
95. Esbjornsson J, Mansson F, Martinez-Arias W, Vincic E, Biague AJ, da Silva ZJ, *et al.* Frequent CXCR4 tropism of HIV-1 subtype A and CRF02_AG during late-stage disease--indication of an evolving epidemic in West Africa. *Retrovirology* 2010,**7**:23.
96. Bratt G, Karlsson A, Leandersson AC, Albert J, Wahren B, Sandstrom E. Treatment history and baseline viral load, but not viral tropism or CCR-5 genotype, influence prolonged antiviral efficacy of highly active antiretroviral treatment. *AIDS* 1998,**12**:2193-2202.
97. Verhofstede C, Nijhuis M, Vandekerckhove L. Correlation of coreceptor usage and disease progression. *Curr Opin HIV AIDS* 2012,**7**:432-439.
98. Connor RI, Sheridan KE, Ceradini D, Choe S, Landau NR. Change in coreceptor use correlates with disease progression in HIV-1--infected individuals. *J Exp Med* 1997,**185**:621-628.
99. Moore JP, Kitchen SG, Pugach P, Zack JA. The CCR5 and CXCR4 coreceptors--central to understanding the transmission and pathogenesis of human immunodeficiency virus type 1 infection. *AIDS Res Hum Retroviruses* 2004,**20**:111-126.
100. Bunnik EM, Swenson LC, Edo-Matas D, Huang W, Dong W, Frantzell A, *et al.* Detection of inferred CCR5- and CXCR4-using HIV-1 variants and evolutionary intermediates using ultra-deep pyrosequencing. *PLoS Pathog* 2011,**7**:e1002106.
101. Dorr P, Westby M, Dobbs S, Griffin P, Irvine B, Macartney M, *et al.* Maraviroc (UK-427,857), a potent, orally bioavailable, and selective small-molecule inhibitor of chemokine receptor CCR5 with broad-spectrum anti-human immunodeficiency virus type 1 activity. *Antimicrob Agents Chemother* 2005,**49**:4721-4732.
102. Vandekerckhove LP, Wensing AM, Kaiser R, Brun-Vezinet F, Clotet B, De Luca A, *et al.* European guidelines on the clinical management of HIV-1 tropism testing. *Lancet Infect Dis* 2011,**11**:394-407.
103. Kootstra NA, Schuitemaker H. Determination of cell tropism of HIV-1. *Methods Mol Biol* 2005,**304**:317-325.
104. Trouplin V, Salvatori F, Cappello F, Obry V, Brelot A, Heveker N, *et al.* Determination of coreceptor usage of human immunodeficiency virus type 1 from patient plasma samples by using a recombinant phenotypic assay. *J Virol* 2001,**75**:251-259.

105. Whitcomb JM, Huang W, Fransen S, Limoli K, Toma J, Wrin T, *et al.* Development and characterization of a novel single-cycle recombinant-virus assay to determine human immunodeficiency virus type 1 coreceptor tropism. *Antimicrob Agents Chemother* 2007,**51**:566-575.
106. Raymond S, Delobel P, Mavigner M, Cazabat M, Souyris C, Encinas S, *et al.* Development and performance of a new recombinant virus phenotypic entry assay to determine HIV-1 coreceptor usage. *J Clin Virol* 2010,**47**:126-130.
107. Obermeier M, Symons J, Wensing AM. HIV population genotypic tropism testing and its clinical significance. *Curr Opin HIV AIDS* 2012,**7**:470-477.
108. Poveda E, Alcami J, Paredes R, Cordoba J, Gutierrez F, Llibre JM, *et al.* Genotypic determination of HIV tropism - clinical and methodological recommendations to guide the therapeutic use of CCR5 antagonists. *AIDS Rev* 2010,**12**:135-148.
109. Low AJ, Dong W, Chan D, Sing T, Swanstrom R, Jensen M, *et al.* Current V3 genotyping algorithms are inadequate for predicting X4 co-receptor usage in clinical isolates. *AIDS* 2007,**21**:F17-24.
110. Swenson LC, Moores A, Low AJ, Thielen A, Dong W, Woods C, *et al.* Improved detection of CXCR4-using HIV by V3 genotyping: application of population-based and "deep" sequencing to plasma RNA and proviral DNA. *J Acquir Immune Defic Syndr* 2010,**54**:506-510.
111. Palmer S, Kearney M, Maldarelli F, Halvas EK, Bixby CJ, Bazmi H, *et al.* Multiple, linked human immunodeficiency virus type 1 drug resistance mutations in treatment-experienced patients are missed by standard genotype analysis. *J Clin Microbiol* 2005,**43**:406-413.
112. Poveda E, Seclen E, Gonzalez Mdel M, Garcia F, Chueca N, Aguilera A, *et al.* Design and validation of new genotypic tools for easy and reliable estimation of HIV tropism before using CCR5 antagonists. *J Antimicrob Chemother* 2009,**63**:1006-1010.
113. Archer J, Weber J, Henry K, Winner D, Gibson R, Lee L, *et al.* Use of four next-generation sequencing platforms to determine HIV-1 coreceptor tropism. *PLoS One* 2012,**7**:e49602.
114. Ray M, Logan R, Sterne JA, Hernandez-Diaz S, Robins JM, Sabin C, *et al.* The effect of combined antiretroviral therapy on the overall mortality of HIV-infected individuals. *AIDS* 2010,**24**:123-137.
115. Arts EJ, Hazuda DJ. HIV-1 antiretroviral drug therapy. *Cold Spring Harb Perspect Med* 2012,**2**:a007161.
116. Gershon D. Green light for ddI. *Nature* 1991,**353**:589.
117. Young FE. The role of the FDA in the effort against AIDS. *Public Health Rep* 1988,**103**:242-245.
118. Palella FJ, Jr., Delaney KM, Moorman AC, Loveless MO, Fuhrer J, Satten GA, *et al.* Declining morbidity and mortality among patients with advanced human immunodeficiency virus infection. HIV Outpatient Study Investigators. *N Engl J Med* 1998,**338**:853-860.
119. Staszewski S, Miller V, Rehmet S, Stark T, De Cree J, De Brabander M, *et al.* Virological and immunological analysis of a triple combination pilot study with loviride, lamivudine and zidovudine in HIV-1-infected patients. *AIDS* 1996,**10**:F1-7.
120. Palella FJ, Jr., Baker RK, Moorman AC, Chmiel JS, Wood KC, Brooks JT, *et al.* Mortality in the highly active antiretroviral therapy era: changing causes of death and disease in the HIV outpatient study. *J Acquir Immune Defic Syndr* 2006,**43**:27-34.

121. Collier AC, Coombs RW, Schoenfeld DA, Bassett RL, Timpone J, Baruch A, *et al.* Treatment of human immunodeficiency virus infection with saquinavir, zidovudine, and zalcitabine. AIDS Clinical Trials Group. *N Engl J Med* 1996,**334**:1011-1017.
122. D'Aquila RT, Hughes MD, Johnson VA, Fischl MA, Sommadossi JP, Liou SH, *et al.* Nevirapine, zidovudine, and didanosine compared with zidovudine and didanosine in patients with HIV-1 infection. A randomized, double-blind, placebo-controlled trial. National Institute of Allergy and Infectious Diseases AIDS Clinical Trials Group Protocol 241 Investigators. *Ann Intern Med* 1996,**124**:1019-1030.
123. EMA. European public assessment reports. In; 2014.
124. FDA. Antiretroviral drugs used in the treatment of HIV infection. In; 2014.
125. Furman PA, Barry DW. Spectrum of antiviral activity and mechanism of action of zidovudine. An overview. *Am J Med* 1988,**85**:176-181.
126. Cheng YC, Dutschman GE, Bastow KF, Sarngadharan MG, Ting RY. Human immunodeficiency virus reverse transcriptase. General properties and its interactions with nucleoside triphosphate analogs. *J Biol Chem* 1987,**262**:2187-2189.
127. Richman DD. HIV chemotherapy. *Nature* 2001,**410**:995-1001.
128. AIDSinfo. Adult and adolescent ARV guidelines. In. Edited by Services DoHaH; 2014.
129. Coffin JM. HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy. *Science* 1995,**267**:483-489.
130. Joos B, Fischer M, Kuster H, Pillai SK, Wong JK, Boni J, *et al.* HIV rebounds from latently infected cells, rather than from continuing low-level replication. *Proc Natl Acad Sci U S A* 2008,**105**:16725-16730.
131. Kitchen CM, Lu J, Suchard MA, Hoh R, Martin JN, Kuritzkes DR, *et al.* Continued evolution in gp41 after interruption of enfuvirtide in subjects with advanced HIV type 1 disease. *AIDS Res Hum Retroviruses* 2006,**22**:1260-1266.
132. Paredes R, Sagar M, Marconi VC, Hoh R, Martin JN, Parkin NT, *et al.* In vivo fitness cost of the M184V mutation in multidrug-resistant human immunodeficiency virus type 1 in the absence of lamivudine. *J Virol* 2009,**83**:2038-2043.
133. Karlsson A, Bjorkman P, Bratt G, Ekvall H, Gisslen M, Sonnerborg A, *et al.* Low prevalence of transmitted drug resistance in patients newly diagnosed with HIV-1 infection in Sweden 2003-2010. *PLoS One* 2012,**7**:e33484.
134. Wheeler WH, Ziebell RA, Zabina H, Pieniazek D, Prejean J, Bodnar UR, *et al.* Prevalence of transmitted drug resistance associated mutations and HIV-1 subtypes in new HIV-1 diagnoses, U.S.-2006. *AIDS* 2010,**24**:1203-1212.
135. Vercauteren J, Wensing AM, van de Vijver DA, Albert J, Balotta C, Hamouda O, *et al.* Transmission of drug-resistant HIV-1 is stabilizing in Europe. *J Infect Dis* 2009,**200**:1503-1508.
136. Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol* 1975,**94**:441-448.
137. Metzker ML. Sequencing technologies - the next generation. *Nat Rev Genet* 2010,**11**:31-46.
138. Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, Song X, *et al.* Direct selection of human genomic loci by microarray hybridization. *Nat Methods* 2007,**4**:903-905.
139. Ronaghi M, Uhlen M, Nyren P. A sequencing method based on real-time pyrophosphate. *Science* 1998,**281**:363, 365.
140. How-genome-sequencing-is-done-FINAL.pdf. In: 454 Life Sciences.

141. Bull RA, Luciani F, McElroy K, Gaudieri S, Pham ST, Chopra A, *et al.* Sequential bottlenecks drive viral evolution in early acute hepatitis C virus infection. *PLoS Pathog* 2011,**7**:e1002243.
142. Henn MR, Boutwell CL, Charlebois P, Lennon NJ, Power KA, Macalalad AR, *et al.* Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. *PLoS Pathog* 2012,**8**:e1002529.
143. Bar KJ, Li H, Chamberland A, Tremblay C, Routy JP, Grayson T, *et al.* Wide variation in the multiplicity of HIV-1 infection among injection drug users. *J Virol* 2010,**84**:6241-6247.
144. Poon AF, Swenson LC, Bunnik EM, Edo-Matas D, Schuitemaker H, van 't Wout AB, *et al.* Reconstructing the dynamics of HIV evolution within hosts from serial deep sequence data. *PLoS Comput Biol* 2012,**8**:e1002753.
145. McElroy K, Thomas T, Luciani F. Deep sequencing of evolving pathogen populations: applications, errors, and bioinformatic solutions. *Microb Inform Exp* 2014,**4**:1.
146. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 2005,**437**:376-380.
147. Jerome M, Noirot C, Klopp C. Assessment of replicate bias in 454 pyrosequencing and a multi-purpose read-filtering tool. *BMC Res Notes* 2011,**4**:149.
148. Rozera G, Abbate I, Bruselles A, Vlassi C, D'Offizi G, Narciso P, *et al.* Massively parallel pyrosequencing highlights minority variants in the HIV-1 env quasispecies deriving from lymphomonocyte sub-populations. *Retrovirology* 2009,**6**:15.
149. Tsibris AM, Korber B, Arnaout R, Russ C, Lo CC, Leitner T, *et al.* Quantitative deep sequencing reveals dynamic HIV-1 escape and large population shifts during CCR5 antagonist therapy in vivo. *PLoS One* 2009,**4**:e5683.
150. Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ. Removing noise from pyrosequenced amplicons. *BMC Bioinformatics* 2011,**12**:38.
151. Zagordi O, Geyrhofer L, Roth V, Beerenwinkel N. Deep sequencing of a genetically heterogeneous sample: local haplotype reconstruction and read error correction. *J Comput Biol* 2010,**17**:417-428.
152. McElroy K, Zagordi O, Bull R, Luciani F, Beerenwinkel N. Accurate single nucleotide variant detection in viral populations by combining probabilistic clustering with a statistical test of strand bias. *BMC Genomics* 2013,**14**:501.
153. Macalalad AR, Zody MC, Charlebois P, Lennon NJ, Newman RM, Malboeuf CM, *et al.* Highly sensitive and specific detection of rare variants in mixed viral populations from massively parallel sequence data. *PLoS Comput Biol* 2012,**8**:e1002417.
154. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B. Detection and quantification of rare mutations with massively parallel sequencing. *Proc Natl Acad Sci U S A* 2011,**108**:9530-9535.
155. Fu GK, Hu J, Wang PH, Fodor SP. Counting individual DNA molecules by the stochastic attachment of diverse labels. *Proc Natl Acad Sci U S A* 2011,**108**:9026-9031.
156. Jabara CB, Jones CD, Roach J, Anderson JA, Swanstrom R. Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. *Proc Natl Acad Sci U S A* 2011,**108**:20166-20171.
157. Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol* 2007,**8**:R143.

158. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004,**32**:1792-1797.
159. Posada D. jModelTest: phylogenetic model averaging. *Mol Biol Evol* 2008,**25**:1253-1256.
160. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* 2010,**59**:307-321.
161. Guindon S. Bayesian estimation of divergence times from large sequence alignments. *Mol Biol Evol* 2010,**27**:1768-1781.
162. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, *et al.* Primer3--new capabilities and interfaces. *Nucleic Acids Res* 2012,**40**:e115.
163. Vallone PM, Butler JM. AutoDimer: a screening tool for primer-dimer and hairpin structures. *Biotechniques* 2004,**37**:226-231.
164. Kalendar R, Lee D, Schulman AH. Java web tools for PCR, in silico PCR, and oligonucleotide assembly and analysis. *Genomics* 2011,**98**:137-144.
165. Giegerich R, Meyer F, Schleiermacher C. GeneFisher--software support for the detection of postulated genes. *Proc Int Conf Intell Syst Mol Biol* 1996,**4**:68-77.
166. Fredslund J, Madsen LH, Hougaard BK, Nielsen AM, Bertioli D, Sandal N, *et al.* A general pipeline for the development of anchor markers for comparative genomics in plants. *BMC Genomics* 2006,**7**:207.
167. Weckx S, De Rijk P, Van Broeckhoven C, Del-Favero J. SNPbox: a modular software package for large-scale primer design. *Bioinformatics* 2005,**21**:385-387.
168. Rouillard JM, Zuker M, Gulari E. OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. *Nucleic Acids Res* 2003,**31**:3057-3062.
169. Rozen S, Skaletsky H. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* 2000,**132**:365-386.
170. Allawi HT, SantaLucia J, Jr. Nearest neighbor thermodynamic parameters for internal G.A mismatches in DNA. *Biochemistry* 1998,**37**:2170-2179.
171. SantaLucia J, Jr., Allawi HT, Seneviratne PA. Improved nearest-neighbor parameters for predicting DNA duplex stability. *Biochemistry* 1996,**35**:3555-3562.
172. Varghese V, Shahriar R, Rhee SY, Liu T, Simen BB, Egholm M, *et al.* 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* 2009,**52**:309-315.
173. Palmer S, Boltz V, Martinson N, Maldarelli F, Gray G, McIntyre J, *et al.* 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* 2006,**103**:7094-7099.
174. Wang C, Mitsuya Y, Gharizadeh B, Ronaghi M, Shafer RW. Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. *Genome Res* 2007,**17**:1195-1201.
175. Li JZ, Paredes R, Ribaudo HJ, Svarovskaia ES, Metzner KJ, Kozal MJ, *et al.* Low-frequency HIV-1 drug resistance mutations and risk of NNRTI-based antiretroviral treatment failure: a systematic review and pooled analysis. *JAMA* 2011,**305**:1327-1335.
176. Liang B, Luo M, Scott-Herridge J, Semeniuk C, Mendoza M, Capina R, *et al.* 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* 2011,**6**:e26745.
177. Di Giallonardo F, Zagordi O, Duport Y, Leemann C, Joos B, Kunzli-Gontarczyk M, *et al.* Next-generation sequencing of HIV-1 RNA genomes: determination of error rates and minimizing artificial recombination. *PLoS One* 2013,**8**:e74249.
178. Gilles A, Meglecz E, Pech N, Ferreira S, Malausa T, Martin JF. Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. *BMC Genomics* 2011,**12**:245.
179. Zagordi O, Klein R, Daumer M, Beerenwinkel N. Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. *Nucleic Acids Res* 2010,**38**:7400-7409.
180. Fang G, Zhu G, Burger H, Keithly JS, Weiser B. Minimizing DNA recombination during long RT-PCR. *J Virol Methods* 1998,**76**:139-148.
181. Eckert KA, Kunkel TA. DNA polymerase fidelity and the polymerase chain reaction. *PCR Methods Appl* 1991,**1**:17-24.
182. Wu JY, Jiang XT, Jiang YX, Lu SY, Zou F, Zhou HW. Effects of polymerase, template dilution and cycle number on PCR based 16 S rRNA diversity analysis using the deep sequencing method. *BMC Microbiol* 2010,**10**:255.
183. Shao W, Boltz VF, Spindler JE, Kearney MF, Maldarelli F, Mellors JW, *et al.* Analysis of 454 sequencing error rate, error sources, and artifact recombination for detection of Low-frequency drug resistance mutations in HIV-1 DNA. *Retrovirology* 2013,**10**:18.
184. Shiroguchi K, Jia TZ, Sims PA, Xie XS. Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. *Proc Natl Acad Sci U S A* 2012,**109**:1347-1352.
185. Johnson JA, Li JF, Wei X, Lipscomb J, Irlbeck D, Craig C, *et al.* Minority HIV-1 drug resistance mutations are present in antiretroviral treatment-naive populations and associate with reduced treatment efficacy. *PLoS Med* 2008,**5**:e158.
186. Metzner KJ, Giulieri SG, Knoepfel SA, Rauch P, Burgisser P, Yerly S, *et al.* Minority quasispecies of drug-resistant HIV-1 that lead to early therapy failure in treatment-naive and -adherent patients. *Clin Infect Dis* 2009,**48**:239-247.
187. Simen BB, Simons JF, Hullsiek KH, Novak RM, Macarthur RD, Baxter JD, *et al.* Low-abundance drug-resistant viral variants in chronically HIV-infected, antiretroviral treatment-naive patients significantly impact treatment outcomes. *J Infect Dis* 2009,**199**:693-701.
188. Deeks SG. Treatment of antiretroviral-drug-resistant HIV-1 infection. *Lancet* 2003,**362**:2002-2011.