KATHOLIEKE UNIVERSITEIT LEUVEN
FACULTEIT GENEESKUNDE
BIOMEDISCHE WETENSCHAPPEN

Vergelijking van verschillende staalvoorbereidingsprotocollen voor volledige genoomanalyse van HIV met behulp van nieuwe sequeneringstechnologieën

[Evaluation of next generation sequencing protocols for HIV complete genome sequencing]

Masterproef voorgedragen
tot het behalen van de
graad van Master in de
Biomedische Wetenschappen
door Nídia TROVÃO

Promotor: Prof. Dr. Philippe LEMEY
Coordinator: Ir. Bram VRANCKEN
Faculteit Geneeskunde
Department Microbiologie en Immunologie
Laboratorium Klinische en Epidemiologische Virologie

Tweede Promotor: Prof. Dr. João PIEDADE
Universidade Nova de Lisboa
Instituto de Higiene e Medicina Tropical
Departament of Virology

Leuven, 2011

Acknowledgements

To my promoter and coordinator, Prof. Dr. Philippe Lemey and Ir. Bram Vrancken. Thank you for believing in me and giving me the oportunity to grow as a researcher and as a person over this year. I am very greatful for you support and guidance, that made this manuscript most of what it is.

I would like to express my sincere gratitude to Prof. Dr. Anne-Mieke Vandamme for giving me the opportunity to work and develop my master thesis at an institute as prestigious as the REGA Institute for Medical Research.

It is a pleasure to thank Prof. Dr. João Piedade, Prof. Dr. Maria Amélia Grácio, Prof. Dr. Miguel Viveiros, Prof. Dr. Virgílio do Rosário and Prof. Dr. Zulmira Maria de Araújo, who made this thesis possible by supporting me not only academically but also with their friendship.

To my colleagues from AV's laboratory who made this learning period a friendly experience. Thank you Lien Kerremans and Yoeri Schrooten for your enthusiam and support in the laboratory.

I wish to thank my friends and family for providing me a loving environment and for always being there for me.

Muito obrigado Mamã e Papá por me amarem e ensinarem, sempre em apoio incondicional.

Gonçalinho, muito obrigado por todo o teu apoio, as noites passadas sem dormir, fazendo da tua felicidade o alcançar deste meu objectivo. Obrigado por tudo o que és para mim.

Although nature commences with reason and ends in experience,
it is necessary for us to do the opposite, that is to commence with
experience and from this to proceed to investigate the reason.
Leonardo da Vinci

Summary

HIV is a retrovirus that gave rise to a worldwide epidemic after its successful zoonotic transmission in the first half of the twentieth century. Current therapy, referred to as Highly Active AntiRetroviral Therapy (HAART), can significantly delay disease progression. However, despite more than 25 years of intensive research there is still no cure available.

All available antiretroviral drugs are faced with the insurmountable challenge posed by the high evolutionary potential of HIV. This implies that regardless the administered drug cocktail, drug resistance can and will develop. To manage these negative effects, patients should be screened on a regular basis in order to detect the development of drug resistance in an early phase, so the therapy regimen can be timely adjusted. Importantly, both drug resistant variants that have evolved *de novo* or were acquired through transmission can negatively impact on therapy outcome. Thus, also therapy-naive patients should be screened before therapy onset.

This screening usually involves genotyping of the viral population through the direct sequencing of the RT-PCR products. Unfortunately, this approach does not allow the reliable detection of viral variants present in less then at about 20%-25% of the population. The association of such minor variants harboring drug resistance mutations with therapy failure fueled investigations to exploit the recently developed Roche® 454 NGS platform in an attempt to gain a more accurate in-depth view of the viral population. These inquiries are characterized by two major drawbacks: their focus on limited genomic regions and the need for large amounts of input material characteristic for the proprietary Roche® 454 fragmentation approach.

As part of a lager project on the comparison of currently available sample preprocessing protocols for complete genome sequencing of clinical HIV plasma and PBMC samples, and the identification of the most suitable viral reservoir for resistance testing in newly infected patients as a secondary objective, this thesis focuses on the corresponding practical aspects of pre-processing prior to sequence data generation. Because of time restrictions, a complete data-analysis with respect tot the research question of this larger project falls outside the scope of this thesis.

Specifically, all wet-lab procedures for both the sequence-specific and random priming amplification strategies were carried out. For the former, we generated 6 overlapping amplicons to cover the entire HIV-1 genome. After equimolar pooling of all amplicons for each sample, we performed two enzymatic fragmentation methods. These will be compared to conventional mechanical 454 shearing. The successful sequencing of one sample and the completion of all sample pre-processing procedures is promising for further applications but a comprehensive evaluation of the sequence data to be generated is necessary to make an informed choice among the different approaches.

Samenvatting

Meer dan 25 jaar na de ontdekking van HIV leven volgens de laatste cijfers van de Wereld Gezondheidsorganisatie ongeveer 33 miljoen mensen met het aidsvirus. Gelukkig kan sinds de goedkeuring van het eerste antiretrovirale middel in 1987 de ziekteprogressie significant vertraagd worden, hetgeen geleid heeft tot een sterk gedaalde morbiditeit. De keerzijde van de medaille is, althans tot een geneeskrachtige therapie ontwikkeld wordt, de levenslange behandeling met dikwijls vele ongewenste neveneffecten. Het werd al snel duidelijk dat therapiefalen ten gevolge van residuele virale replicatie, hetgeen de selectie van varianten met mutaties die het virus minder gevoelig maken voor de toegediende drugs, onontkoombaar is. Zelfs het gelijktijdig inhiberen van verschillende stappen in de virale replicatiecyclus, hetgeen momenteel mogelijk is dankzij de 25 verschillende beschikbare antiretrovirale middelen, blijkt niet te volstaan om de ontwikkeling van resistentie te voorkomen. Bijgevolg zijn we genoodzaakt om diagnostische tests toe te passen die de ontwikkeling van resistentie aan de huidige combinatietherapieën in een zo vroeg mogelijk stadium kunnen opsporen. Dit laatste is belangrijk omdat een tijdige aanpassing van het therapieregime de verdere ontwikkeling van resistentie gedeeltelijk kan vermijden.

Vanwege de relatief lagere kost, de eenvoudigheid en snelheid van toepassing wordt voor deze resistentietest standaard gekozen voor het genotyperen van de virale populatie. Meestal gebeurt dit door de volgorde van de nucleotiden van de RT-PCR producten te bepalen met behulp van de door Sanger en Coulson ontwikkelde methode. Ondanks de grote vooruitgang op het gebied van therapiemanagment die met deze techniek geboekt werd, komt therapiefalen nog dikwijls voor. Dit kan deels verklaard worden door de hoge detectielimiet van de standaard genotyperingstesten: virale varianten met een aandeel van minder dan 20%-25% van de virale populatie kunnen op deze manier niet betrouwbaar gedetecteerd worden. De associatie tussen zulke laagfrequente varianten met resistentiemutaties en therapiefalen heeft geleid tot de toepassing van de recent door Roche gecommercialiseerde nieuwe sequeneringstechnologie, algemeen gekend onder de naam ”454 pyrosequencing”, om een meer accuraat dieptezicht in de complexe virale populaties te bekomen.

Deze studies lijden echter onder twee grote nadelen. Door de nood aan grote hoeveelheden startmateriaal voor het standaard Roche® 454 fragmentatieprotocol beperken veel studies zich tot stalen met een hoge virale lading, hetgeen de bruikbaarheid van deze benadering sterk beperkt. Dit kan omzeild worden door het genereren en sequencen van een set overlappende amplicons van geschikte lengte. Van de noodzakelijke amplificaties kan dan gebruik gemaakt worden om de noodzakelijke adaptors en, indien gewenst, unieke labels in te bouwen. Echter, door het veelvuldige gebruik van sequentiespecifieke primers kan deze amplicon-gebaseerde methode de representatieve weerspiegeling van de eigenlijke virale populatie niet garanderen. Bovendien ligt bij deze studies de focus op beperkte genomische regio’s, hetgeen een alomvattende analyse van virale evolutie verhindert.

Als onderdeel van een groter project waarbij verschillende staalvoorbereidingstechnieken met elkaar vergeleken worden qua performantie op klinische plasma en PBMC stalen van HIV patiënten, en waarbij

het identificeren van het meest geschikte virale reservoir voor resistentietesting bij nieuw geïnfecteerde patienten een secundair objectief is, legt deze thesis zich toe op de praktische aspecten van de staalvoorbereiding die voorafgaat aan het genereren van de sequentiedata. Meer specifiek is het doel van deze thesis na te gaan welke staalvoorbereidingsmethode het meest geschikt is voor een zo omvattend mogelijke analyse van volledige HIV genomen uit te voeren.

Hiervoor wordt een sequentie-onafhankelijke amplificatiemethode (WT-Ovation^{TM} RNA-Seq System, NuGEN$^{®}$), die resulteert in amplificatieproducten met een geschikte lengte om te sequencen, vergeleken met een aantal fragmentatiemethoden die een voorafgaande primer-gebaseerde amplificatie vereisen. Om bij deze laatste de kans op een vertekend beeld van de virale populatie te minimalizeren wordt gekozen voor een minimale primerset. Hiervoor baseren we ons op generische PCR-protocollen voor HIV-1 groep M (verantwoordelijk voor de wereldwijde epidemie) die door onze collega's van het Aids Referentie Laboratorium (UZ Leuven/KU Leuven) ontwikkeld werden, hetgeen ons verzekert van de breedst mogelijke toepasbaarheid. De fragmentatiemethoden die vergeleken worden vertegenwoordigen het volledige beschikbare spectrum aan vereist startmateriaal, gaande van laag (Nextera^{TM}) over middelmatig (NEBNext® dsDNA Fragmentase^{TM}) tot hoog (standaard Roche$^{®}$ 454 fragmentatie).

Een overzicht van de experimentele setup kan teruggevonden worden in Figuur 3.1. Belangrijk is dat van elke stap een aantal herhalingen gebeuren, die, vooraleer verder te gaan met de procedure, samengevoegd worden. Hierdoor wordt de impact van stochastische invloeden tijdens elke stap van het proces geminimalizeerd met als neveneffect een substantiële verhoging van de werklast. Alle nodige amplificaties, opzuiveringen en fragmentaties met de verschillende methoden werden uitgevoerd. Met een makkelijk op te volgen en eenduidig experiment voor ogen werd ervoor gekozen om alle stalen tijdens dezelfde ”454-run” te sequencen. Ten gevolge van tijdslimieten heeft deze strategie ertoe geleid dat we van slechts 1 staal sequentiedata kunnen rapporteren (zie hieronder).

Binnen de sequentiespecifieke arm van het onderzoek duiden de eerste resultaten al op het potentiële voordeel van de lage inputvereiste van het Nextera^{TM} protocol ten opzichte van de andere fragmentatiemethoden. Doordat de tweede ronde van amplificaties hierbij overbodig is, zelfs voor het staal met de laagste virale lading waarvoor plasma beschikbaar was, wordt ook de bijhorende kans op een vertekend beeld van de virale populatie geminimalizeerd. Van de 11 stalen die met het Nextera^{TM} protocol gefragmenteerd werden, kon het Genomics Core centrum (UZ Leuven/KU Leuven) al een analyse van de fragmentgrootte-verdeling uitvoeren. Deze lijken gemiddeld genomen te groot voor een efficiënte emPCR, maar dit patroon kan te wijten kan zijn aan niet-covalente bindingen tussen de adaptorsequenties. Hoewel dit volgens de fabrikant geen contraindicatie is voor een succesvolle emPCR, werd uit voorzichtigheid geopteerd voor een test-run van 1 staal op 1/16e regio van de picotiterplaat. Hierbij werden twee condities uitgetest: een emPCR met 0,15cpb (copies per bead) en 0,30cpb. Uit een initiële analyse blijkt dat voor beide condities het aantal reads binnen de te verwachten grenzen ligt, maar dat er een iets grotere dan verwachtte proportie kortere reads is. Welke de oorzaak hiervan is, vraagt een gedetailleerde analyse die buiten het kader van deze thesis valt.

Het overzicht van de ruwe sequentiedata die bekomen werden (Figuur 4.5) illustreert zowel de mogelijkheid om een goed dieptezicht over de volledige lengte van het genoom te bekomen, als de grote hoeveelheden data waarmee dit gepaard gaat. Aangezien dit laatste diepgaande analyses zeer tijdrovend maakt, werd bij wijze van voorbeeld 1 regio volledig geanalyseerd (Figuur 4.4).

In conclusie kunnen we stellen dat het succesvol sequencen van 1 staal en het afwerken van alle staalvoorbereidingen voor dit onderzoek veelbelovend is voor toekomstige toepassingen, maar dat een uitgebreide analyse van de -deels nog te genereren- data nodig is om een onderbouwde keuze tussen de verschillende fragmentatiebenaderingen te maken.

Contents

Abbreviations

3TC - lamivudine
ABC - abacavir
AIDS - acquired immunodeficiency syndrome
APV - amprenavir
ART - antiretroviral therapy
ATP - adenosine triphosphate
ATV - atazanavir
AZT - zidovudine
bp - base pair
Ca - capsid
CCR5 - chemokine (C-C motif) receptor 5
CD4 - cluster of differentiation 4
cDNA - complementary deoxyribonucleic acid
CRF - circulating recombinant forms
CXCR4 - chemokine (C-X-C motif) receptor 4
d4T - stavudine
ddC - zalcitabine
ddI - didanosine
didNTPs - dideoxy nucleoside triphophates
DLV - delavirdine
DNA - deoxyribonucleic acid
dNTP - deoxyribonucleotide triphosphate
DRM - drug resistance mutations
DRV - darunavir
dsDNA - double stranded deoxyribonucleic acid
EFV - efavirenz
EI - entry inhibitors
emPCR - emulsion polymerase chain reaction
Env - viral envelope
ETR - etravirine
fAPV - fosamprenavir
FTC - emtricitabine
Gag - group specific antigen
HAART - highly active antiretroviral therapy
HIV - human immunodeficiency virus
IDV - indinavir
IκB - inhibitors of NF-κB
In - integrase
INI - integrase inhibitors
kb - kilobases
LPV/r - lopinavir
LTR - long terminal repeats
Ma - matrix protein
MCV - maraviroc
NC - nucleocapsid protein
Nef - negative regulatory factor
NFV - nelfinavir
NF-κB - nuclear factor kappa-light-chain-enhancer of activated B cells
NGS - next generation sequencing
NNRTI - non-nucleoside reverse transcriptase inhibitors
NRTI - nucleoside / nucleotide reverse transcriptase inhibitors
NVP - nevirapine
PASS - parallel allelespecific sequencing
PBMC - peripheral blood mononuclear cells
PCR - polymerase chain reaction
PI - protease inhibitors
Pol - polymerase
PPi - pyrophosphate
Pr - protease
RAL - raltegravir
Rev - regulator of viral expression
RNA - ribonucleic acid
RT - reverse transcriptase
RT-PCR - reverse transcription followed by polymerase chain reaction
RTV - ritonavir
SIV - simian immunodeficiency virus
SQV - saquinavir
ssDNA - single stranded deoxyribonucleic acid
T-20 - enfuvirtide
Tat - trans-activator of transcription
TDF - tenofovir
TDR - transmission of drug resistance
TF - transcription factors
TPV - tripanavir
Vif - virulence factor
Vpr - viral protein R
Vpu - viral protein U

List of Figures

Chapter 1

Introduction

1.1 HIV/AIDS

On June 5 1981, a new clinical syndrome was reported among homosexual men by the Center for Disease Control. Infected patients had very low numbers of CD4$^+$ T-cells and were susceptible for infections and pathologies associated with immunodeficiency, such as *Pneumocystis carninii* pneumonia, candidiasis and Kaposi's sarcoma [12, 13].

A virus discovered two years later by Françoise Barré-Sinoussi and her colleagues at the Institut Pasteur [14], was soon recognized as the causative pathogen of the Acquired Immunodeficiency Syndrome (AIDS). In 1986 a consensus name was agreed upon and the virus became known as the Human Immunodeficiency Virus (HIV) [15]. That same year, a close relative was discovered and classified as HIV-2 [16], while the former virus was referred to as HIV-1.

The disease course can be divided in three phases, defined by the CD4-count[1] and the viral load[2] [17]. The acute phase is characterized by a high viral load and a depletion of CD4$^+$ T-cell pool, usually accompanied by flu-like symptoms and/or rash. A viral set point is subsequently reached, which reflects the balance between viral replication and control by the host immune system. This relatively stable viral load can be maintained for longer periods of time, on average 10 years, but eventually this chronic and latent phase comes to an end due to the gradual loss of CD4$^+$ T-cells (Figure 1.1). In the final symptomatic stage of the disease course, the host becomes susceptible for opportunistic infections and after AIDS has developed, death occurs usually within a year [18].

Fortunately, from 1987 on, antiretroviral drugs became available [19]. Although they cannot completely abolish viral replication, the disease onset can be significantly delayed and life expectancies of well-treated HIV-infected people nowadays are almost as high as those of uninfected people [20].

The HIV life cycle includes a step that converts the RNA into a proviral DNA genome. The virally encoded reverse transcriptase responsible for this task lacks proofreading activity and is the main contributor to the high error rate associated with the RNA to DNA transition, estimated at about 10^{-4} to 10^{-5} errors per base per cycle [21]. In combination with a rapid turnover and a large ”replication space”, this provides the HIV virus with an enormous potential to generate diversity. Within a host, the HIV population consequently exists as a swarm of closely related variants, often referred to as the ”quasispecies”. On

[1] number of CD4$^+$ T-cells per mL of blood

[2] number of HIV-1 RNA copies per mL plasma

a global scale this diversity is reflected in types, groups and numerous subtypes as well as Circulating Recombinant Forms (CRFs) [22].

The three most prevalent subtypes are A, B and C, whereby the latter accounts for almost 50% of all HIV-1 infections worldwide and more than 80% of all HIV-1 infections in southern Africa and India. According to the latest estimates of the World Health Organization and UNAIDS, 33.3 million people were living with HIV/AIDS at the end of 2009.

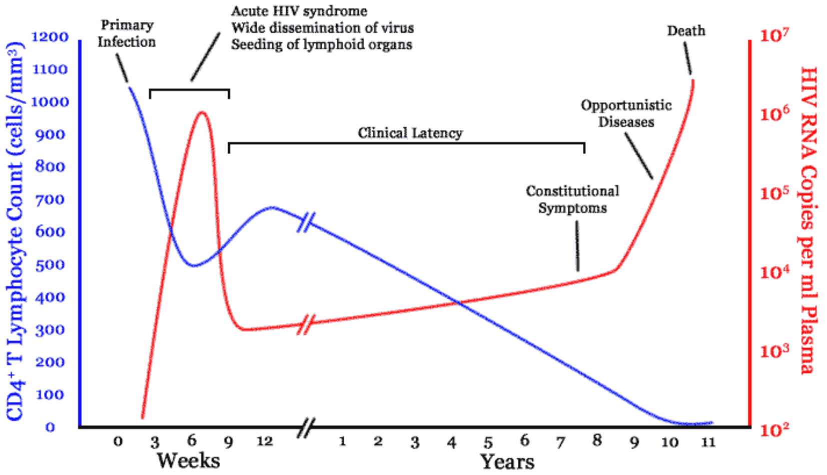


Figure 1.1: Natural course of HIV-1 infection (red - RNA copies per mL plasma; $blue$ - number of CD4$^+$ cells per mL). Adapted from [1]

That same year, an estimated 2.6 million persons became newly infected and 1.8 million died of AIDS, including 2.5 million children [23].

Despite of being pandemic, HIV/AIDS is characterized by large geographic prevalence disparities, with the highest burden carried by Sub-Sahara African countries. In this region an estimated 22.4 million people are living with HIV/AIDS - around two thirds of the global total [24].

1.2 Origins and classification of HIV

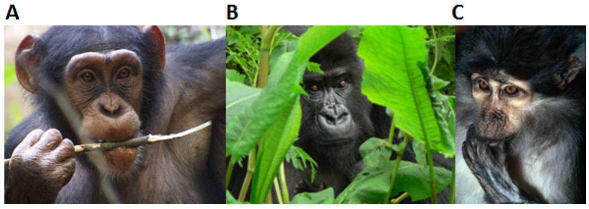


Figure 1.2: (A) *Pan troglodytes troglodytes*, (B) *Gorilla gorilla gorilla* and (C) *Cercocebus atys*. Adapted from [2, 3]

The most compelling evidence that points in the direction of a closely related virus in simians as the progenitor of HIV can be summarized in two arguments. First, HIV lineages are phylogenetically interspersed with Simian Immunodeficiency Viruses (SIV): HIV-1 groups appear most closely related to either SIVcpz or SIVgor, respectively found in the chimpanzee subspecies *Pan troglodytes troglodytes* and in *Gorilla gorilla gorilla*. Likewise, HIV-2 is most closely related to SIVsm, found in sooty mangabeys (*Cercocebus atys*) [25]. Importantly, since the human viruses are closer with their animal counterpart, such a tree topology indicates multiple cross-species transmissions [26]. Second, the most diverse population of viruses is expected to circulate in regions where they have been around for the longest period of time. This is indeed the case for the geographical patterns of HIV diversity, which point to the regions where *Pan troglodytes troglodytes*, *Gorilla gorilla gorilla* and *Cercocebus atys* live, Central and Western Africa respectively, as the cradle of HIV [27] (Figure 1.2). In both regions humans are exposed to simian viruses in multiple ways (through hunting and keeping pets), which implicates blood-blood contact as a plausible route for crossing the species barrier [28].

At least four such transmissions from chimpanzees and gorillas to humans occurred, each giving rise to one of the clades HIV-1 is categorized into: group M (main), N (not M, not O), O (outlier) and P [29, 30].

These transmissions had very different outcomes: group P likely represents a dead-end transmission, groups N and O remain confined to relatively few individuals [31, 32, 33, 34, 35], whereas group M strains have spread throughout the world and represent more than 90% of HIV-1 infections, and can be further organized into 9 subtypes and numerous Circulating Recombinant Form (CRFs) [36]. The latter reflects the high genetic variability, which, together with founder effects, accounts for the geographical linkage of subtypes [37].

Based on molecular clock analysis calibrated using historical samples, the emergence of HIV/AIDS has been dated back to the beginning of the 20th century [38, 39]. A plausible explanation on the geographical origins of HIV/AIDS is that it originates from Cameroon [40] and reached Kinshasa at a time when circumstantial factors where permissive for its initial spread [26].

1.3 Viral structure and genome

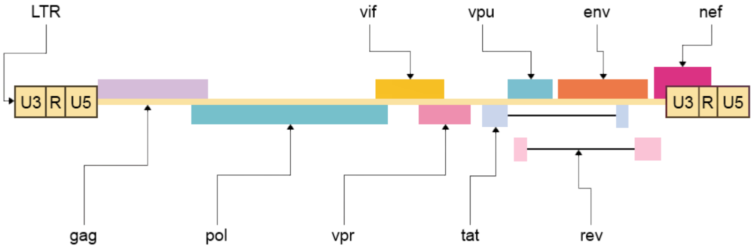


Figure 1.3: Schematic organization of the HIV-1 genome. Adapted from [4]

In terms of taxonomy, HIV belongs to the *lentivirus* genus of the *retroviridae* family [41]. The viral genome consists of two positive stranded RNA molecules of ≈9.7 kilobases (kb), each one with a 5' CAP end and a 3' poly(A) tail as to mimic the eukaryotic mRNAs. The general structure follows that of the other retroviruses, with three main and several accessory genes. These are encoded in 9 open reading frames (Figure 1.3) and produce 15 proteins as a result of differential splicing and post-translational cleaving.

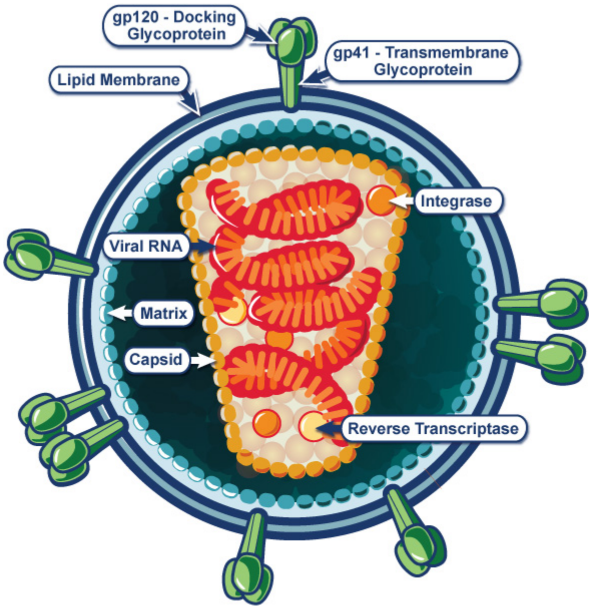


Figure 1.4: Schematic representation of the HIV-1 viral structure [5]

From the *gag* region, four proteins are produced via precursor polyprotein (p55) cleavage by HIV protease: Matrix (Ma or p17), which surrounds the Capsid (Ca or p24), that in turn encloses the Nucleocapsid (Nc or p7), and p6. Processing of the *pol*-encoded polyprotein precursor results in Reverse Transcriptase (RT or p66/p51) , Protease (Pr or p11) and Integrase (IN or p31). The *env* gene codes for the precursor polyprotein gp160, that is cleaved by cellular protease giving rise to the Surface (Su or gp120) and Transmembrane (Tm or gp41) glycoproteins, which anchor to and protrude from the viral lipid membrane (Figure 1.4).

In addition to the structural and enzymatic proteins HIV also possesses regulatory and accessory genes: *tat*, *rev*, *nef*, *vif*, *vpr* and *vpu* (Figure 1.3). Their main function is the regulation of HIV transcription and the modulation of the host cell machinery to favor its own replication cycle (Figures 1.5).

The nine genes are flanked by two repetitive regions called non-coding long terminal repeats (LTRs) which function as promoters [42]. These LTRs are divided into the U3, R, and U5 regions, where by the U3 region can be further divided into the modulatory, enhancer and promoter regions [43].

1.4 Replication cycle

The HIV-1 replication cycle is divided in two phases.The early phase encompasses stages from cell attachment up to the integration of the viral cDNA, whereas the late phase refers to the expression of viral genes and subsequent release and maturation of progeny virions (Figure 1.5).

1.4.1 Early phase

Cellular entry is mediated by the envelope proteins, which are inserted in the virion's membrane as trimers of gp41/gp120. The entry process starts with the recognition of a CD4 molecule by the gp120 trimer. This binding induces a conformational change, which allows neighboring regions to interact with a co-receptor adjacent to the CD4 molecule in the cell membrane [44], usually the chemokine receptors CCR5 or CXCR4.

The conformational changes in gp41 induce a trimer-of-hairpins to bring virion and cell membrane together, allowing fusion [45]. After its release into the cytoplasm, the RNA genome is converted to a dsDNA molecule by the virally encoded reverse transcriptase during transport to the nucleus. This complex process includes two template switches [46, 47], and thus contributes to the ability of HIV to generate diversity through recombination [22]. The integrase mediated insertion of the viral genome upon nuclear entry concludes the events of the early phase [48].

1.4.2 Late phase

Once integrated, the HIV proviral DNA will proceed to the late phase of its replication cycle. As part of the cellular DNA, the provirus is transcribed by the cellular machinery. All transcripts are generated from a single promoter in the 5' LTR to which cellular transcription factors (TFs) will bind, as well as viral proteins, like Tat [49]. A key cellular TF is nuclear factor kappa-light-chain-enhancer of activated B cells (NF-κB), which binds two adjacent sites in the U3 region. In a positive feed-back loop, viral stimulation of the T cell receptor triggers the liberation of the inactive form of NF-κB from its cytoplasmic inhibitor IκB. This in turn enables the translocation of NF-κB to the nucleus where it induces the expression of a series of T-cell activation-specific genes, concomitantly activating HIV-1 transcription [43].

The viral transcripts are 5' capped, 3' polyadenylated and can be divided in three classes: single spliced mRNAs ($\approx 4kb$) which encode Env, Vif, Vpu and Vpr, multiple spliced mRNAs ($\approx 2kb$) which are translated into Rev, Tat and Nef and genomic unspliced mRNA ($\approx 9kb$), which will be inserted in assembling virions and serves as the template for the Gag and Gag-Pol polyprotein precursors [50, 51].

This range of transcripts with intronic sequences are produced by suboptimal splicing sites [52], and can be exported to the cytoplasm by Rev [53]. This is followed by translation at the free ribosomes or at endoplasmatic reticulum associated ribosomes, for Env [54].

Virion assembly starts with a complex interplay between Gag and Env and is crucially dependent on p6, which is required for the separation of the virion envelope from the host plasma membrane [55]. During or shortly after release, maturation is achieved by protease mediated cleavage of the Gag and Pol polyproteins in order to produce the necessary proteins. Concomitantly, the immature spherical particles become mature, infectious conical capsids [56].

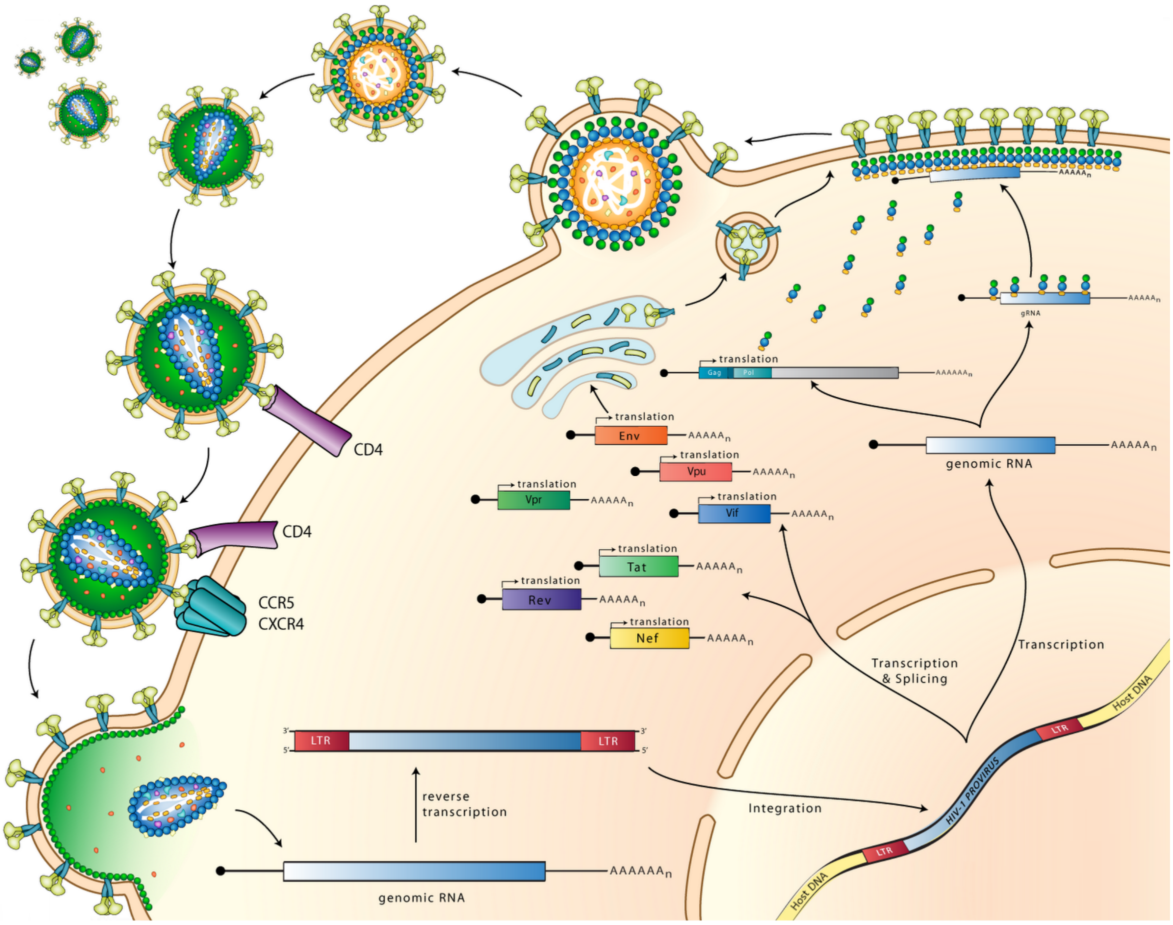


Figure 1.5: Schematic representation of the HIV-1 replication cycle. Adapted from [6]. See text for details.

1.5 Next generation sequencing platforms - upscaling the amount of sequence information

The past decades of genomic research have been characterized by an increasing demand of more affordable and higher throughput sequencing. Whereas the initial upscaling through parallelization led to the 96-well automated capillary electrophoresis systems, the explosion in sequencing capacity only started with the commercialization of the first so-called next generation sequencing (NGS) platforms in 2005 [57]. Here, the miniaturization of the sequencing process in combination with an *in vitro* cloning step led to the ability of massive parallel sequencing of *in vitro* clonally amplified fragments.

From the three major players that dominate this field (see Table 1.1), the Roche® 454's GS FLX™ system (further referred as ”454”) is by far the most exploited in HIV research, in part because it was the first available NGS system. The main reason however is that the long read length permits more efficient data recovery through mapping procedures [7], which is why we implemented this system in our research.

Because the 454 platform allows the cost-effective analysis of, currently, about 1 million of individual sequences at a time, it can be applied to upscale the amount of sequence information in two dimensions (Figure 1.6). By aiming at a high coverage one can achieve an accurate in-depth view of the viral population composition, whereas at the same time, focus can be expanded from limited genomic markers towards whole genome studies.

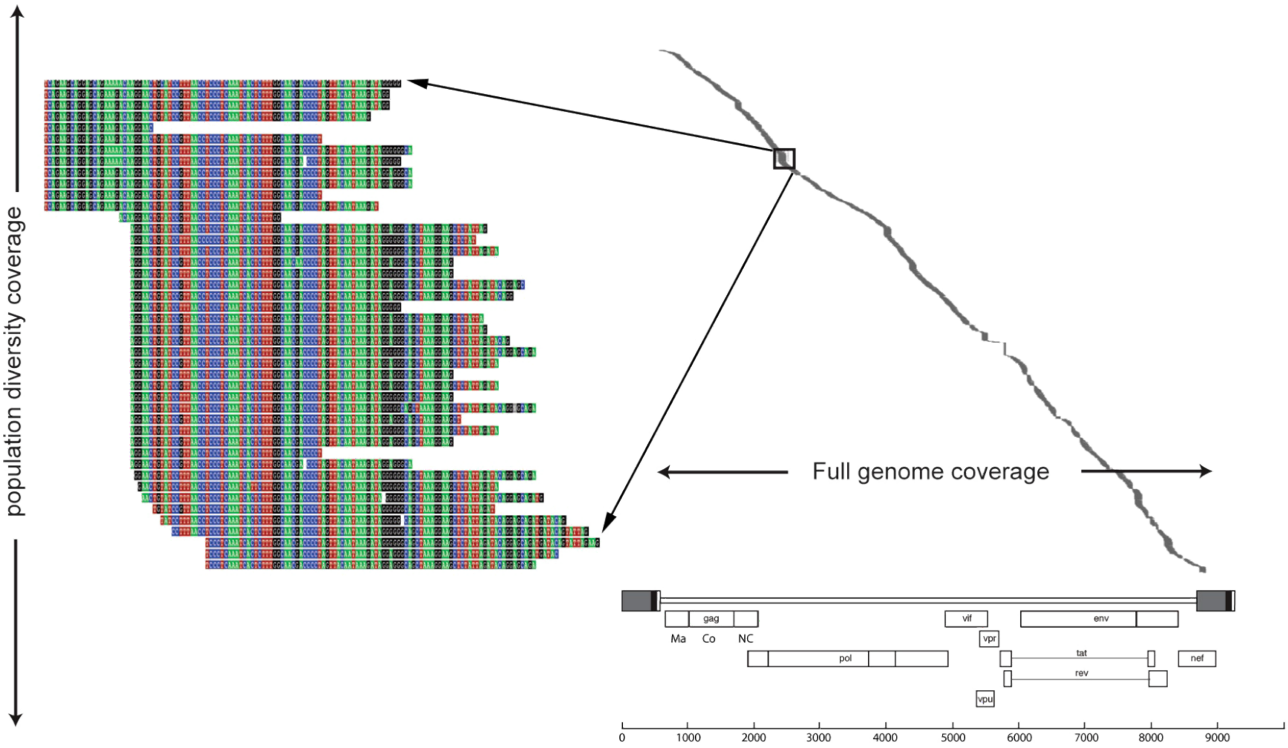
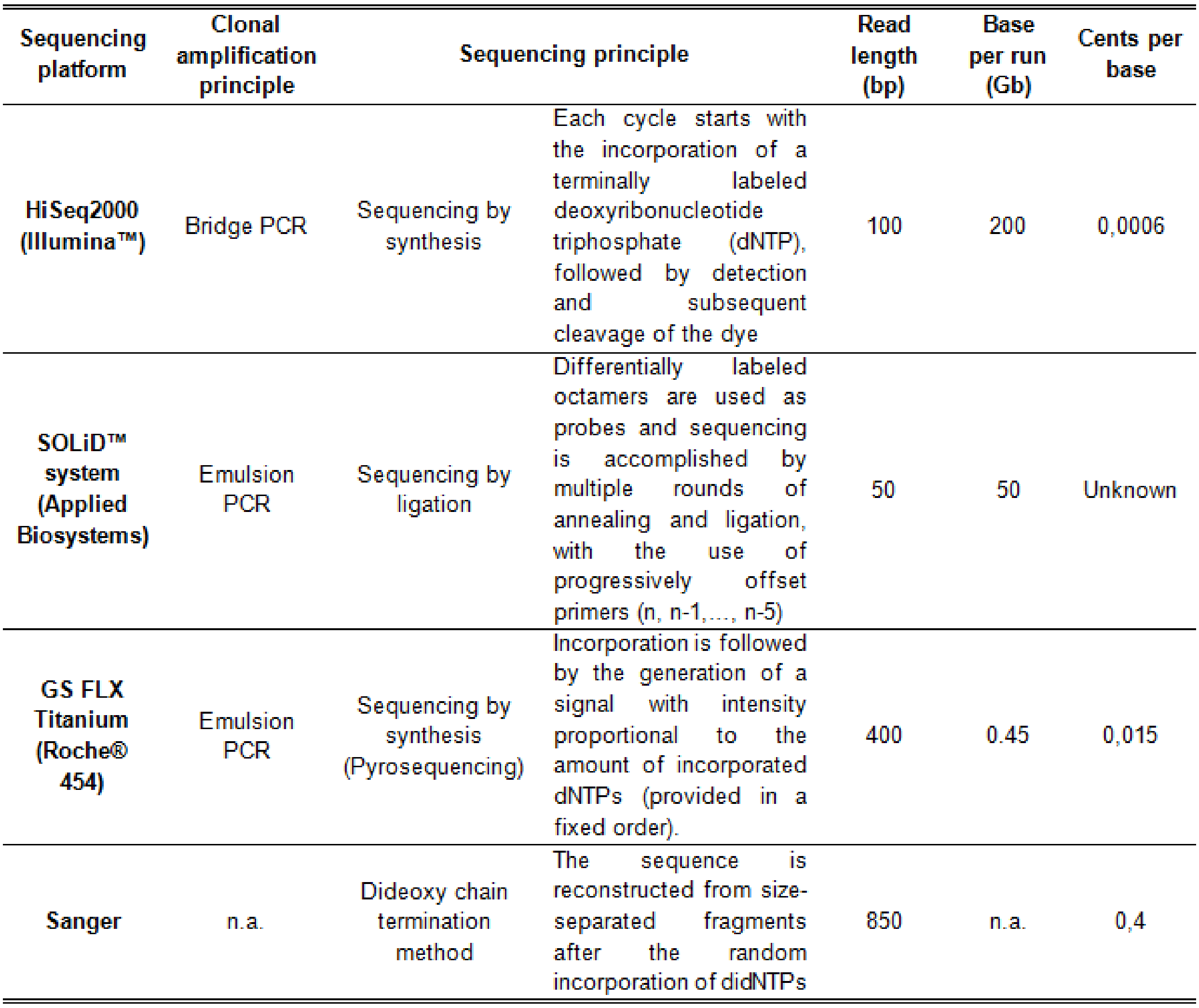


Figure 1.6: Illustration of both dimensions of sequence upscaling. The rectangle marks an area in the HIV-1 genome alignment, for which the aligned reads and the obtained coverage are presented in more detail. Results obtained from a pilot GS20 run [7]

The 454 system was readily adopted for the detection of low frequency variants in within-host populations in order to investigate the potential clinical importance of minor drug resistant variants [58]. Such research crucially depends on a high coverage in order to obtain the desired resolution, which is why such procedures are often referred to as ”deep” sequencing.

This dimension of upscaling has received most attention but analysis remained confined to specific portions of the genome, notably those most often targeted by therapy (PR, RT and V3-loop of Env) [59, 60]. Obviously, by ignoring the genomic context to a large extent, this focus on short regions poses limits to our understanding of viral evolution. For example, the comprehensive study of HIV evolution in response to the cellular immune response requires we take the whole viral proteome into account. Also, drugs that target different steps in the viral replication cycle are becoming available and, as they are encoded by different genes, genotyping will have to be extended towards all possible targets. A third advantage of the second dimension of upscaling lies in the applications in epidemiological surveys because, by aiming at a minimal coverage, the complete genome of many samples can be simultaneously sequenced.

Table 1.1 - Comparison of next generation sequencing platforms. Adapted from [61, 62]



Sequencing platform	Clonal amplification principle	Sequencing principle		Read length (bp)	Base per run (Gb)	Cents per base
HiSeq2000 (Illumina™)	Bridge PCR	Sequencing by synthesis	Each cycle starts with the incorporation of a terminally labeled deoxyribonucleotide triphosphate (dNTP), followed by detection and subsequent cleavage of the dye	100	200	0,0006
SOLiD™ system (Applied Biosystems)	Emulsion PCR	Sequencing by ligation	Differentially labeled octamers are used as probes and sequencing is accomplished by multiple rounds of annealing and ligation, with the use of progressively offset primers (n, n-1,..., n-5)	50	50	Unknown
GS FLX Titanium (Roche® 454)	Emulsion PCR	Sequencing by synthesis (Pyrosequencing)	Incorporation is followed by the generation of a signal with intensity proportional to the amount of incorporated dNTPs (provided in a fixed order).	400	0.45	0,015
Sanger	n.a.	Dideoxy chain termination method	The sequence is reconstructed from size-separated fragments after the random incorporation of didNTPs	850	n.a.	0,4

1.5.1 The Roche® 454's GS FLX^{TM} Titanium sequencing procedure

An important feature of the 454 platform library preparation, which it shares with the other currently prevailing NGS platforms, is that it is essentially aimed at the creation of fragments of suitable length flanked with different upstream and downstream adaptors A and B (Figure 1.7A). One of these adaptors is used in a next step to bind to a bead coated with complementary oligomers (Figure 1.7B), under conditions whereby on average only one DNA fragment is bound to each bead. These beads are then subjected to an oil-in-water emulsion PCR, which serves to amplify the signal strength (Figure 1.7C). At the same time, the emulsion PCR effectively ensures the massive parallel *in vitro* clonal amplification of the bead-bound fragments. This, combined with the parallel sequencing using about 1 million beads on the PicoTiterPlate^{TM}, is responsible for the enormous upscaling of the sequencing throughput. The PicoTiterPlate^{TM} (Figure 1.7D-F), in the current configuration of 454 platform, contains about 3.4 million wells. The dimensions of the latter are such that these can accommodate exactly one fragment-bound bead, as well as several smaller beads. The smaller beads serve two purposes: they keep the "big" bead in place and anchor the enzymes necessary for the actual sequencing reaction (Figure 1.7G-H). The actual sequencing process starts by delivering the dNTPs in a fixed and known order to the beads on the PicoTiterPlate^{TM}. Each time a dNTP is incorporated in the growing strands on a bead, the released pyrophosphate is used by ATP sulfurylase to release ATP. This in turn drives the conversion of luciferine to oxyluciferine by luciferase, which is a luminiscent reaction (Figure 1.7I). The equimolarity of this chain of reactions ensures, over a limited range, a linear relation between the number of incorporated dNTPs and the signal intensity. All signals are detected by the instrument's sensor, and because each light flash can be traced back to a specific dNTP-flow, the sequence can be reconstructed from the obtained flowgram (Figure 1.7J) [63, 64].

1.6 Front-end amplification protocols

The need for large quantities of input material for the standard library preparation protocol of 454 platform necessitates a prior amplification step when low input samples are to be used.

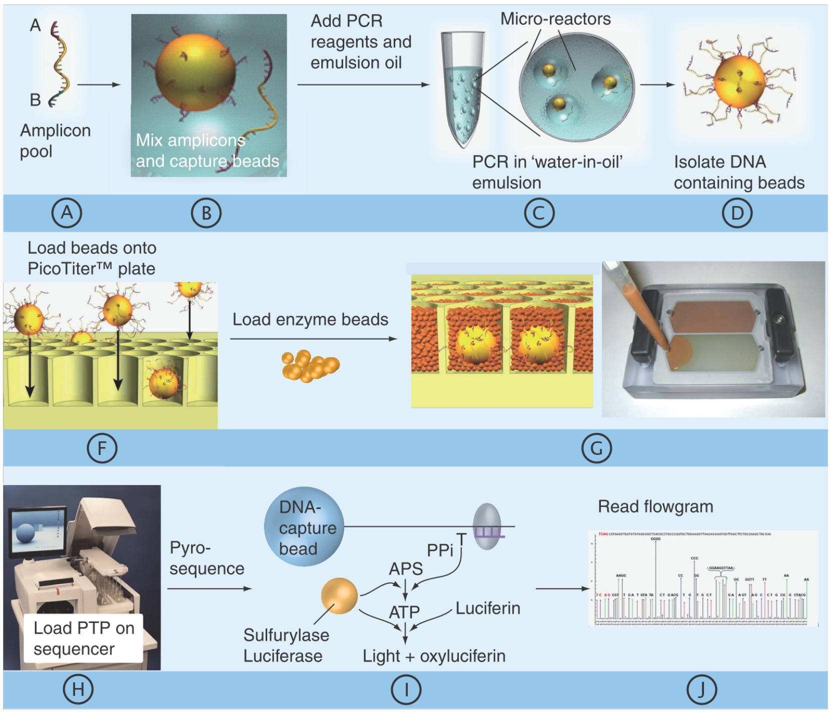


Figure 1.7: The 454 library preparation and pyrosequencing procedure. Adapted from [8]

However, even a nested PCR approach whereby many copies of a large genomic region are generated does not necessarily guarantee the yield required for the standard library preparation protocol (5μg) [64]. Therefore, many of the initial studies who took this approach were confined to high-input samples [58], which limits the application scope since clinical plasma samples may have lower numbers of viral RNA.

The latter downside might be, at least partially, overcome by the recently commercialized enzymatic random fragmentation method (NEBNext® dsDNA Fragmentase^{TM}) that has a minimum input requirement 1μg [9].

Alternatively, a set of overlapping amplicons of suitable length can be created. Here, the necessary amplification steps can be used to incorporate the platform-specific adaptors, which in effect replaces the inefficient standard fragmentation procedure. A useful extension of the latter methodology was reported by Hoffmann *et al.* [59], who added unique labels to the primers, which makes it possible to afterwards attribute each read to its original sample.

Unfortunately, such target specific primer-based technique does not guarantee the proportional representation of templates and may introduce systematic as well as stochastic biases [65, 66, 67]. Also, extending the amplicon-sequencing approach towards full genome sequencing becomes impractically labor intensive and costly, in particular because replicates are advised in order to minimize the potential impact of stochastic influences [68, 60].

An elegant solution to this problem was described by Bimber *et al.* [69]. By creating a small set of large overlapping amplicons that span the entire genome, followed by fragmentation using an efficient transposon based method, they circumvent the need for multiple PCRs. However, because this approach still requires a sequence-specific amplification, it focuses on minimizing the possibility of biases by limiting the number of primers. Alternatively, samples may be accommodated for NGS platform sequencing through bias-free random amplification strategies, as recently demonstrated by Willerth *et al.* [70]. However, the performance of the latter on clinical plasma samples has not yet been evaluated.

1.7 Antiretroviral therapy

The discovery that zidovudine, an analogue of thymidine, was able to suppress HIV-1 replication in cell-culture, spurred hopes for a treatment of HIV infection [71]. In 1987 it became the first antiretroviral agent approved for the treatment of HIV/AIDS [19]. Currently, 25 drugs have been approved by the US Food and Drug Administration for treatment of HIV-1 infection. These can be classified into 5 classes according to their mode of action (Table 1.2).

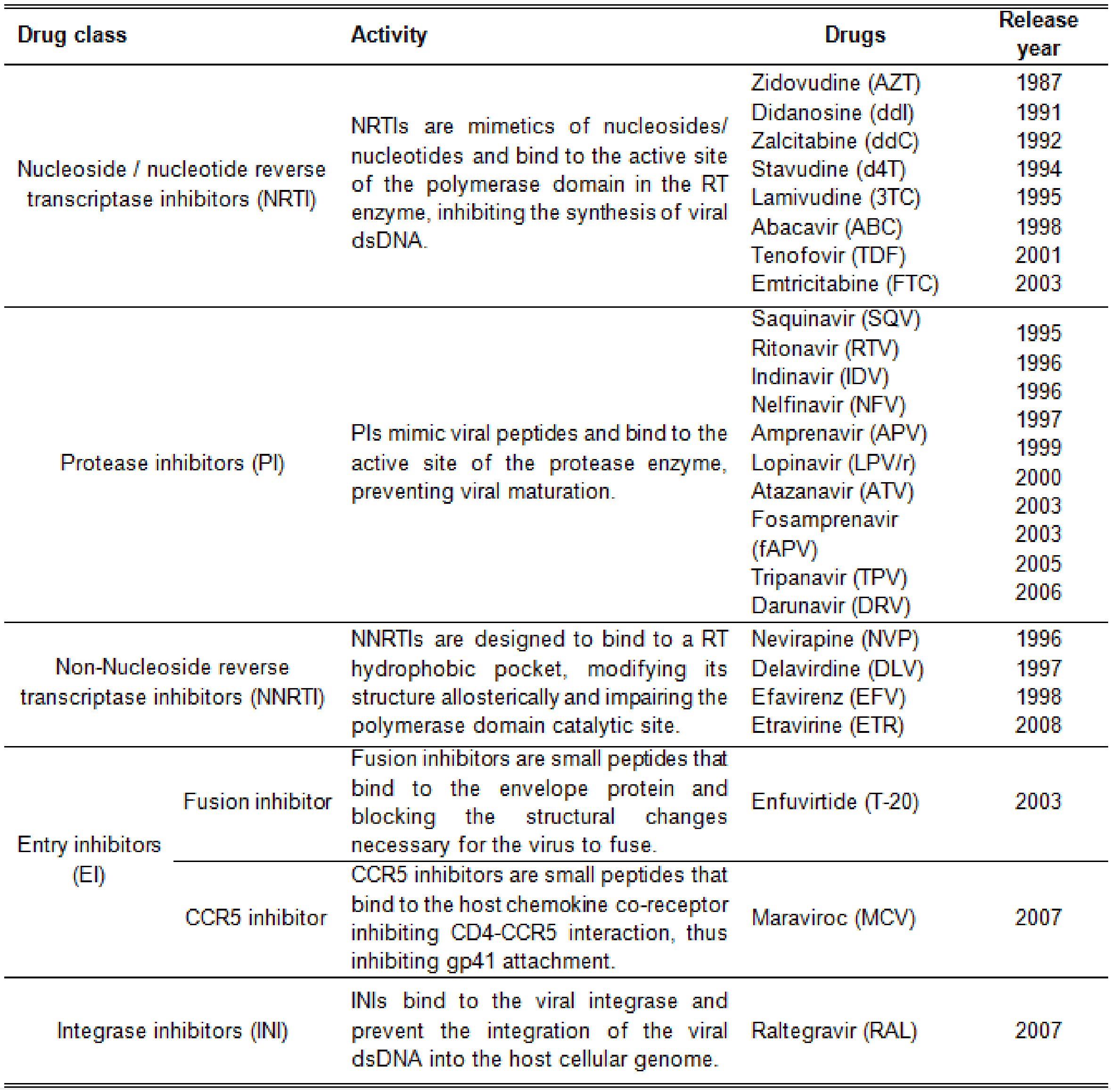
Thanks to antiretroviral therapy (ART), HIV infection can now be managed as a chronic disease with significantly improved life quality and expectancy [72]. However, single and dual therapy proved not to be efficient enough to control virus replication for prolonged periods of time. Therefore current standard therapy is a combination of 3 different drugs that simultaneously target at least 2 steps in the viral replication cycle, a strategy commonly referred to as Highly Active AntiRetroviral Therapy (HAART) [73].

Unfortunately, complete suppression of viral replication remains, even under the most potent regimens, unattainable. This residual replication, often a result of adherence problems, inevitably leads to the selection of variants with drug resistance mutations (DRM) and is associated with therapy failure.

Such drug resistance mutations may already be fixed in the viral population before (salvage) therapy start and can have a detrimental effect on therapy-efficacy. For this reason resistance testing has become a routine clinical tool [74, 75]. Genotyping is the preferred method for predicting the response to the next therapy because of its lower cost, less complicated technique and faster turnaround time compared to phenotyping assays [76]. The standard procedure involves the direct sequencing of the RT-PCR products by the dideoxy-chain termination method, a method often referred to as population or bulk sequencing. A major limitation of such approaches is that mutations present in less than at about 20% to 25% of the viral population cannot be reliably detected [7].

This limitation, whereby potentially clinically significant low-level drug resistant variants are often overlooked, can be overcome through deep sequencing with the 454 system (see above), which can reveal an accurate and quantitative view of the genetic diversity [7].

Table 1.2 - Summary of antiretroviral drug classes for treatment of HIV-1 infection. Adapted from [77]



Drug class		Activity	Drugs	Release year
Nucleoside / nucleotide reverse transcriptase inhibitors (NRTI)		NRTIs are mimetics of nucleosides/ nucleotides and bind to the active site of the polymerase domain in the RT enzyme, inhibiting the synthesis of viral dsDNA.	Zidovudine (AZT) Didanosine (ddI) Zalcitabine (ddC) Stavudine (d4T) Lamivudine (3TC) Abacavir (ABC) Tenofovir (TDF) Emtricitabine (FTC)	1987 1991 1992 1994 1995 1998 2001 2003
Protease inhibitors (PI)		PIs mimic viral peptides and bind to the active site of the protease enzyme, preventing viral maturation.	Saquinavir (SQV) Ritonavir (RTV) Indinavir (IDV) Nelfinavir (NFV) Amprenavir (APV) Lopinavir (LPV/r) Atazanavir (ATV) Fosamprenavir (fAPV) Tripanavir (TPV) Darunavir (DRV)	1995 1996 1996 1997 1999 2000 2003 2003 2005 2006
Non-Nucleoside reverse transcriptase inhibitors (NNRTI)		NNRTIs are designed to bind to a RT hydrophobic pocket, modifying its structure allosterically and impairing the polymerase domain catalytic site.	Nevirapine (NVP) Delavirdine (DLV) Efavirenz (EFV) Etravirine (ETR)	1996 1997 1998 2008
Entry inhibitors (EI)	Fusion inhibitor	Fusion inhibitors are small peptides that bind to the envelope protein and blocking the structural changes necessary for the virus to fuse.	Enfuvirtide (T-20)	2003
	CCR5 inhibitor	CCR5 inhibitors are small peptides that bind to the host chemokine co-receptor inhibiting CD4-CCR5 interaction, thus inhibiting gp41 attachment.	Maraviroc (MCV)	2007
Integrase inhibitors (INI)		INIs bind to the viral integrase and prevent the integration of the viral dsDNA into the host cellular genome.	Raltegravir (RAL)	2007

1.8 Viral reservoirs at primary HIV-1 infection

When cells become infected with HIV this is usually followed by the hijacking of the cellular machinery and the production of many offspring virions (see above). Sometimes, however, the viral genome is inserted in CD4$^+$ T-cells that, instead of being activated, return to the resting memory state. This process leads to the formation of a latent reservoir that not only prevents the eradication of HIV -proviruses are no target of current ART- but also serves as a memory wherein all diversity ever generated is present. Interestingly, the dynamics of he proviral reservoir are such that it's fueled primarily by the transmitted variant or variants [78, 79]. Because DRMs usually confer a fitness cost, viral variants that harbor such mutations will tend to either revert to wild-type in the absence of drug selective pressure, or persist as minor variants under the detection limit of conventional genotyping assays, impeding their detection in plasma. However, the long half-life of the resting CD4$^+$ T-cells and the relatively slow genetic turnover of the PBMC-reservoir [80], which ensures we can find transmitted DRMs in PBMCs long after the initial infection, make PBMCs are a plausible candidate for TDR detection [81].

Chapter 2

Aims

We propose to comprehensively evaluate a number of sample pre-processing methods, for in-depth complete genome characterization of clinical HIV-1 samples, on the 454 platform. We aim to test both sequence-specific and random-priming amplification strategies. For the former, we adopt a strategy similar to Bimber *et al.* [69] and generate 6 overlapping amplicons to cover the entire HIV-1 genome.

In a next step, we compare two enzymatic fragmentation methods, NEBNext® dsDNA Fragmentase™ and Nextera™ transposon-based technology, to conventional mechanical 454 shearing. These represent the state-of-the-art fragmentation approaches, and range from high (standard Roche® 454 shearing) to moderate (NEBNext® dsDNA Fragmentase™) and even low (Nextera™) input requirements. The fragmentation procedures will also be compared with a sequence-independent amplification method that has proven useful for complete HIV sequencing on the Illumina™ platform [70]. We aim to study how they impact coverage distribution, both in terms of uniform coverage of the complete genome and depth of quasispecies variation, and to determine which approach allows for the most comprehensive evolutionary characterization of HIV genomes in clinical samples. Although we aim to complete all pre-processing procedures on our samples as part of this thesis, we anticipate that the generation and comprehensive analysis of the sequence data is beyond the scope of the current project due to time restrictions.

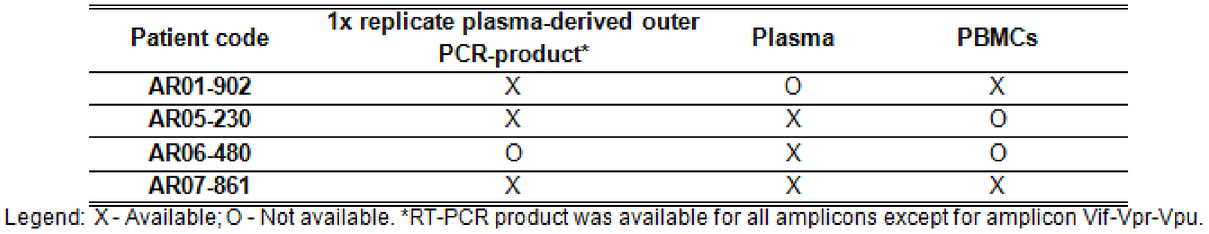
We apply the proposed sample pre-processing protocols in parallel to six pre-therapy samples originating from four patients that constitute a small HIV-1 subtype B transmission chain. By investigating both plasma and PBMC populations, we hope to elucidate which viral compartment may be best suited to detect transmitted drug resistance at a pre-treatment stage.

Chapter 3

Materials and Methods

In this project, we used the clinical samples available from patients that constitute a small HIV-1 subtype B transmission chain [82]. An overview of the samples at our disposal can be found in Table 3.1.

Table 3.1 - Overview of the samples available for each patient.



Patient code	1x replicate plasma-derived outer PCR-product*	Plasma	PBMCs
AR01-902	X	O	X
AR05-230	X	X	O
AR06-480	O	X	O
AR07-861	X	X	X

Legend: X - Available; O - Not available. *RT-PCR product was available for all amplicons except for amplicon Vif-Vpr-Vpu.

A general overview of the experimental setup, which includes the pre-processing procedures we compared and to which we refer below, can be found in Figure 3.1.

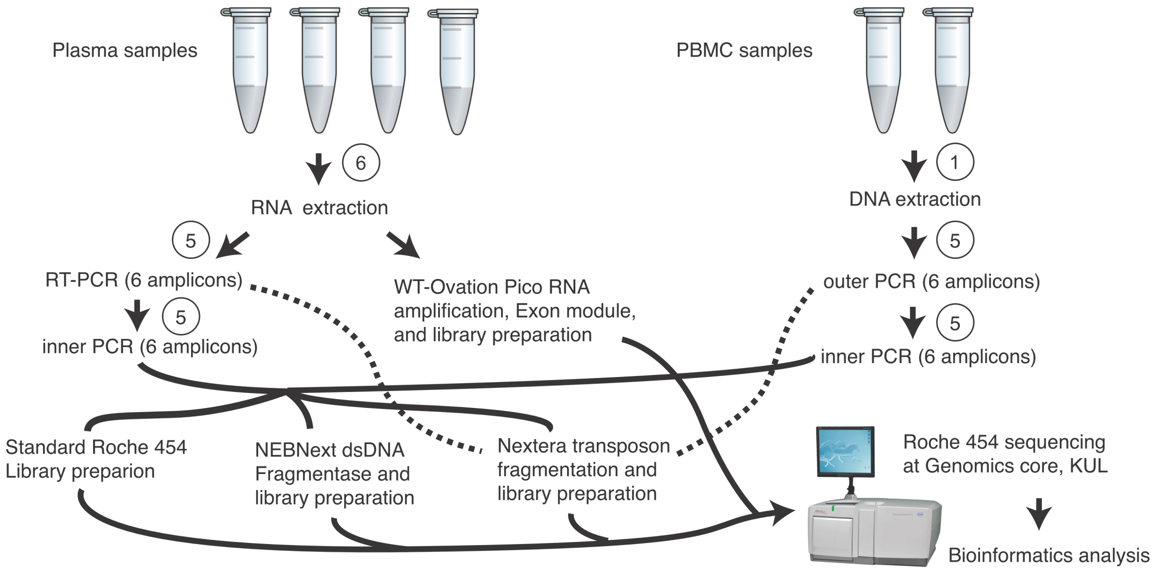

Plasma samples
PBMC samples
6
RNA extraction
1
DNA extraction
5
5
RT-PCR (6 amplicons)
WT-Ovation Pico RNA amplification, Exon module, and library preparation
outer PCR (6 amplicons)
5
5
inner PCR (6 amplicons)
inner PCR (6 amplicons)
Standard Roche 454 Library preparion
NEBNext dsDNA Fragmentase and library preparation
Nextera transposon fragmentation and library preparation
Roche 454 sequencing at Genomics core, KUL
Bioinformatics analysis

Figure 3.1: Schematic representation of the experimental setup. The number of replicates are indicated in circles. The dotted lines indicate potential steps without resorting to inner PCR amplification.

3.1 Sequence specific amplification

Analogous to Bimber *et al.*, we generated a set of 6 overlapping amplicons that span the viral genome for both the viral and proviral samples (Figure 3.2).

We extend on the applicability of this approach by ensuring generic detection capability for HIV-1 group M (responsible for the worldwide epidemic), for which we have evaluated (RT-)PCR protocols that are routinely used by our collaborators at the AIDS Reference Laboratory (University Hospitals Leuven).

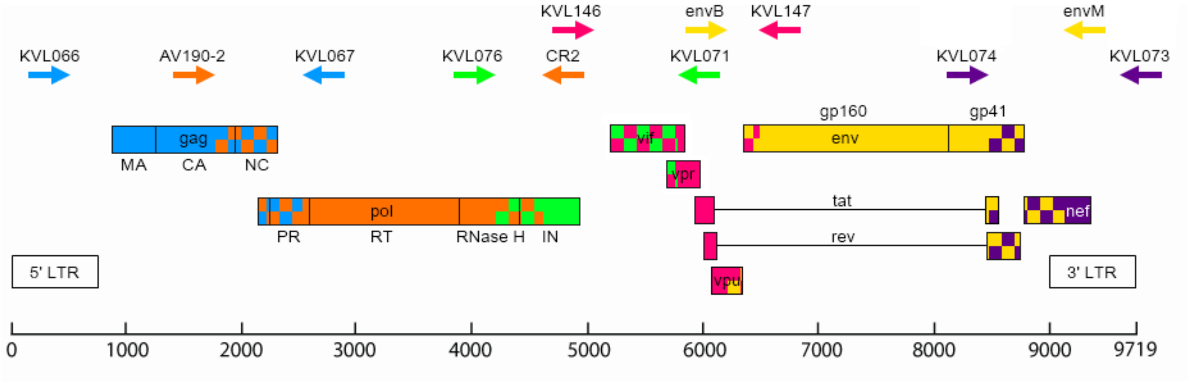


Figure 3.2: Overview of the sequence specific amplification strategy. The primer pairs targeting a genomic region are indicated with the same color. The scale bar represents the genomic coordinates according to HXB2.

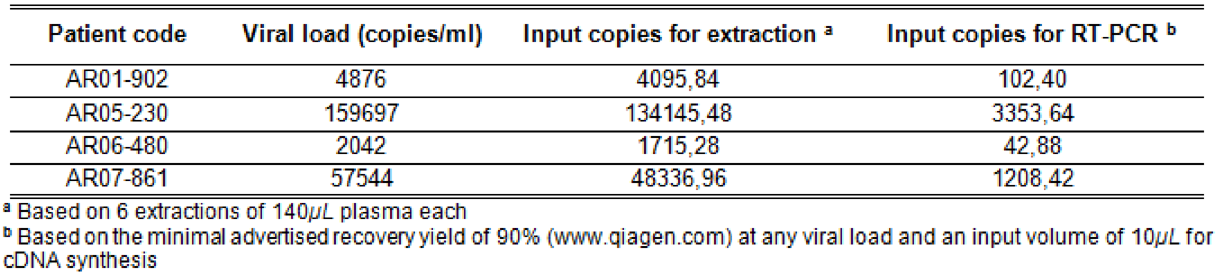
3.1.1 RNA extraction

The ability to detect variants at their true proportions depends on the number of viral templates available for cDNA synthesis [83], which in turn is determined by the number of RNA molecules in the original sample and potential losses during the upstream procedures.

According to the results of Poon *et al.*[68], sampling variation during the RNA extraction represents the largest source of error in the sample pre-processing procedure in terms of obtaining a correct view on the actual population composition. For this reason, they advise a minimum of 2 replicate extractions as a standard for the use of NGS in genotypic resistance testing. We opted for a 6 replicate extraction because this number provided us with the volume required for the replicate RT-PCRs. Concomitantly, this ensures a more representative view on the true population composition. HIV RNA was extracted from 140μL plasma with the QIAgen Viral RNA mini kit (Qiagen®) according to the manufacturer's instructions, and eluted in 60μL.

The efficiency of the RNA-extraction has, together with the viral load of the samples, a large impact on the interpretation of the results because they determine the number of independent viral RNA molecules that are available for RT-PCR amplification. To illustrate this, we present an overview of the samples' viral load in relation to the maximum number of different input molecules in Table 3.2. Besides *de facto* defining the biologically meaningful detection limit, this also impacts on the reproducibility of the results. This has been statistically analyzed by Stenman *et al.* [84] who, like Karrer *et al.* [65], found that the lower the abundance of any template, the less likely its true abundance will be reflected in the amplified library. The source of this error lies in the distribution of, in this case, the viral RNA templates in the RNA-extract, from which a small volume is used as input for the (RT-)PCR. This sampling effect can be countered by striving for a maximal concentration of the viral RNA, which can be achieved by aiming for an as high as possible input volume for the RNA extraction procedure. Alternatively, the chosen strategy of pooling 6 replicate extractions -which only increases the chance that all variants are represented in their true proportions in the RNA extract, but does not increase the concentration of the viral RNA templates- afterwards averages out this potential source of error. At the same time, this pooling strategy minimizes the impact of stochastic fluctuations due to PCR drift [68, 60].

Table 3.2 - Viral load and the expected number of independent input molecules for extraction and cDNA synthesis



Patient code	Viral load (copies/ml)	Input copies for extraction [a]	Input copies for RT-PCR [b]
AR01-902	4876	4095,84	102,40
AR05-230	159697	134145,48	3353,64
AR06-480	2042	1715,28	42,88
AR07-861	57544	48336,96	1208,42

[a] Based on 6 extractions of 140μL plasma each

[b] Based on the minimal advertised recovery yield of 90% (www.qiagen.com) at any viral load and an input volume of 10μL for cDNA synthesis

3.1.2 DNA extraction

Before proceeding to the actual DNA extraction, the number of available cells was determined as advised in Appendix B of the QIAamp DNA Blood Mini Kit (Qiagen®) in order to assure that the maximum input number of cells was not exceeded.

This revealed we were limited to a 1x replicate DNA extraction from 1.000.000 PBMCs from sample AR01-902 and 3.000.000 cells from sample AR07-861.

Estimates of a proviral load between 1 provirus per 10^5 and 10^3 PBMCs [85] lead us, analogues to the reasoning above, to expect 0,5 to 50 independent copies in each outer PCR for sample AR01-902 and 1,5 to 150 for sample AR07-861[1]. To ensure the purification of RNA-free DNA, the optional addition of RNaseA (Qiagen®) was included in the protocol. All steps were performed according to the manufacturer's protocol.

3.1.3 PCR

For both training purposes and because of the changes we have introduced to the (RT-)PCR protocols (see Addendum and also below: the final extension time was omitted), we tested their performance on RNA extract derived form laboratory strains. Also, the (RT-)PCR protocols which we based ourselves upon were optimized for the amplification of plasma-derived viruses. Before proceeding with the amplification of the patients'

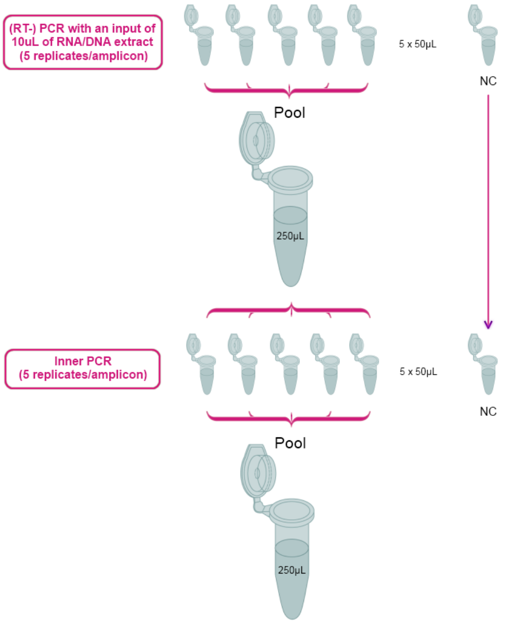

(RT-) PCR with an input of 10uL of RNA/DNA extract (5 replicates/amplicon)
5 x 50µL
NC
Pool
250µL
Inner PCR (5 replicates/amplicon)
5 x 50µL
NC
Pool
250µL

Figure 3.3: Schematic overview of the amplification protocol for both plasma and PBMC samples with the sequence specific amplification approach.

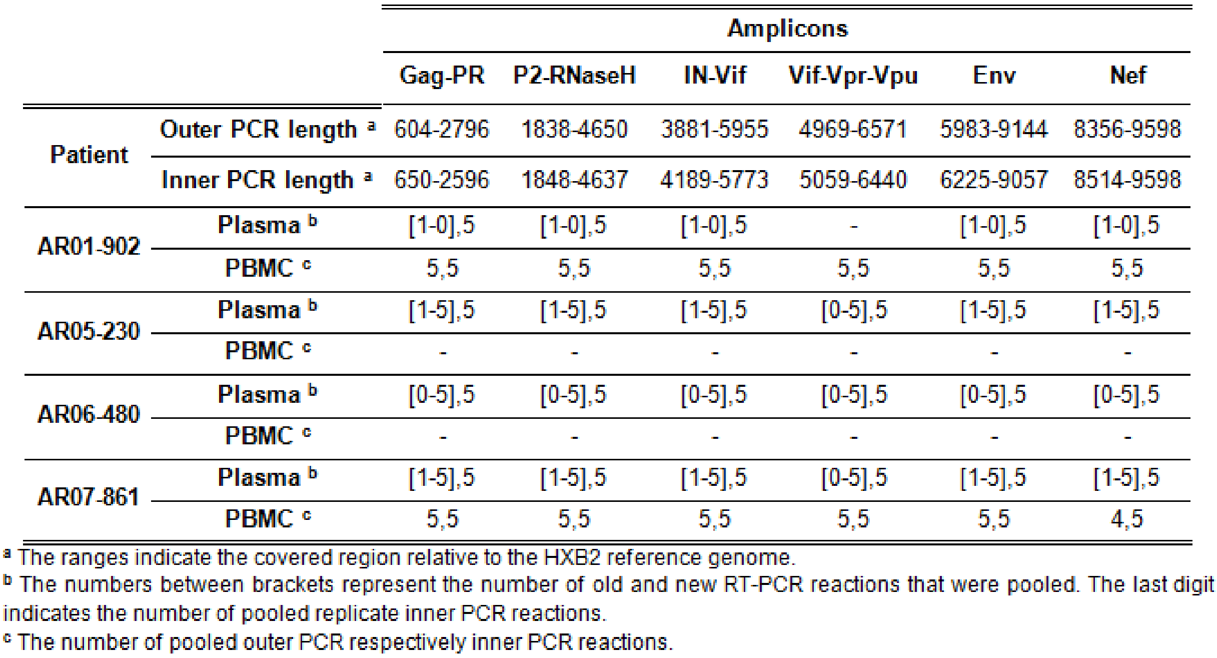
[1] Based on an extraction efficiency of 100% for the used elution volume of 200μL (www.qiagen.com) and an input volume of 10μL for the outer PCR

PBMC-derived proviral DNA, we first evaluated their performance with exclusion of the RT-step on cell culture derived provirus.

Briefly, viral RNA was transcribed to cDNA and amplified in a one-step RT-PCR reaction, as 6 amplicons (Figure 3.2). 5 replicate RT-PCR products of each amplicon were pooled with the already available outer PCR product (see Table 3.2 and Table 3.3). The pooled outer PCR product was used as input for a 5 replicate inner PCR for each amplicon, except for patient AR01. For the latter, the absence of stored plasma required us to exclusively rely on the available 1 replicate outer PCR product for all amplifications. The products of the inner nested-PCR were pooled per amplicon for each patient (Figure 3.3).

Conditions for the amplification of proviral DNA were identical, except for the omission of the RT-step. The use of RNaseA at the extraction step combined with the omission of the RT-step in the hot-started cycling program should in effect result in a traditional outer PCR. Also, the obtained extraction volume did not permit 5 replicates of the outer PCR with the usual 10μL extract as input. Instead, to arrive at the same number of replicates, 6μL of DNA extract was complemented with 4μL of H2O except for Nef. Here we added 10μL of the DNA extract because the test with the 6μL-approach yielded no positive result as determined on a 1% agarose-gel electrophoresis. Because of this only 4 replicate outer PCR reactions were possible for Nef (Table 3.3).

Table 3.3 - Overview of the generated amplicons and the number of replicate (RT-)PCR reactions per patient and per sample.



		Amplicons					
		Gag-PR	P2-RNaseH	IN-Vif	Vif-Vpr-Vpu	Env	Nef
Patient	Outer PCR length [a]	604-2796	1838-4650	3881-5955	4969-6571	5983-9144	8356-9598
	Inner PCR length [a]	650-2596	1848-4637	4189-5773	5059-6440	6225-9057	8514-9598
AR01-902	Plasma [b]	[1-0],5	[1-0],5	[1-0],5	-	[1-0],5	[1-0],5
	PBMC [c]	5,5	5,5	5,5	5,5	5,5	5,5
AR05-230	Plasma [b]	[1-5],5	[1-5],5	[1-5],5	[0-5],5	[1-5],5	[1-5],5
	PBMC [c]	-	-	-	-	-	-
AR06-480	Plasma [b]	[0-5],5	[0-5],5	[0-5],5	[0-5],5	[0-5],5	[0-5],5
	PBMC [c]	-	-	-	-	-	-
AR07-861	Plasma [b]	[1-5],5	[1-5],5	[1-5],5	[0-5],5	[1-5],5	[1-5],5
	PBMC [c]	5,5	5,5	5,5	5,5	5,5	4,5

[a] The ranges indicate the covered region relative to the HXB2 reference genome.
[b] The numbers between brackets represent the number of old and new RT-PCR reactions that were pooled. The last digit indicates the number of pooled replicate inner PCR reactions.
[c] The number of pooled outer PCR respectively inner PCR reactions.

To minimize the amount of PCR-induced recombination, which leads to loss of linkage and thus precludes haplotype reconstruction (which falls outside the scope of this thesis), the final extension step was excluded from all employed cycling programs [86]. To quantify the effect of excluding the final extension time on PCR-induced recombination, we amplified the gp160 region from a 50%-50% mixture of the RNA extract of laboratory strains *IIIb* and *pNL4.3* with the same amplification protocol as for the clinical plasma samples. We aimed to compare the crossover patterns with those obtained from amplification of the same mixture by the same protocol wherein a final extension time of 10 minutes at 68 degrees Celsius was added to the cycling programs.

All amplification reactions were performed in a Biometra T3000 thermal cycler. Primers were synthesized by Eurogentec® (Belgium). An overview table with al (RT-)PCRs mixes, the cycling programs and a list of primers can be found in Addendum.

3.1.4 Purification and quantification

Before the plasma and PBMC samples were submitted to the various shotgun fragmentation methods, the pooled PCR products were purified and quantified, after which all amplicons were equimolarly mixed per sample.

Since the fragmentation methods we compared represent the state-of-the-art approaches and range from high (standard 454 shearing) to moderate (NEBNext® dsDNA Fragmentase^{TM}) and even low (Nextera^{TM}) input requirements, both the outer and inner PCR products were purified and quantified. This allowed us to check whether the second-round PCR amplification can be avoided for the moderate and low-input fragmentation approach, which can further reduce potential biases. All quantifications were done with the Quant-iT dsDNA HS Assay kit or Quant-iT dsDNA BR Assay Kit from Invitrogen^{TM} according to the manufacturers' instructions. Samples were measured 3 times and the average of these measurements was used. To equimolarly mix the amplicons, following formula was used: $Molecules/\mu l = \frac{(Sampleconc.;ng/\mu l)\times(6.022\times10^{23})}{(328.3\times10^{9})\times(avg.fragmentlength;nt)}$.

The pooled outer and inner PCR products were divided over two separate experiments in our protocol (Figure 3.4). One part was purified with DNA Clean&Concentrator^{TM} (Zymo Research), and after quantification and equimolar pooling these samples served as input for the Transposome^{TM} based fragmentation (Nextera^{TM}). The remainder of the product was purified with the illustra® GFX^{TM} PCR DNA and Gel Band Purification kit from GE Healthcare Life Sciences. After quantification and equimolar pooling, these samples served as input for the general 454 shearing method and for the NEBNext® dsDNA Fragmentase^{TM}.

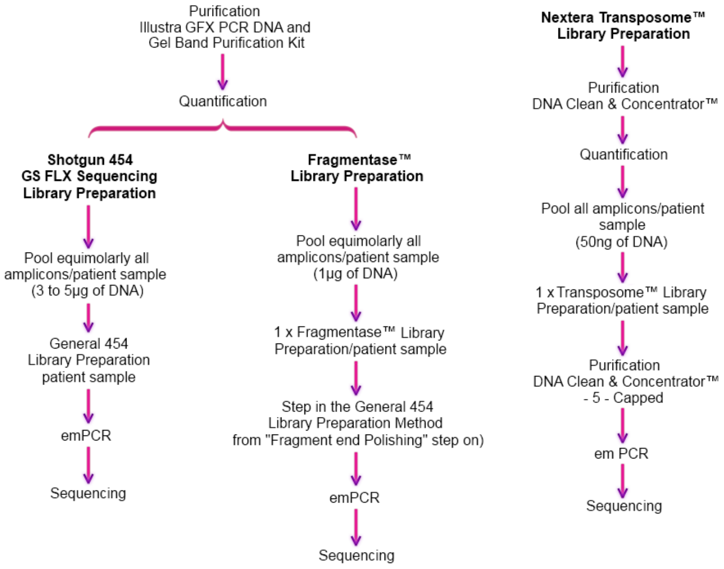


Figure 3.4: Experimental setup of the applied shearing methods.

3.1.5 Shearing methods

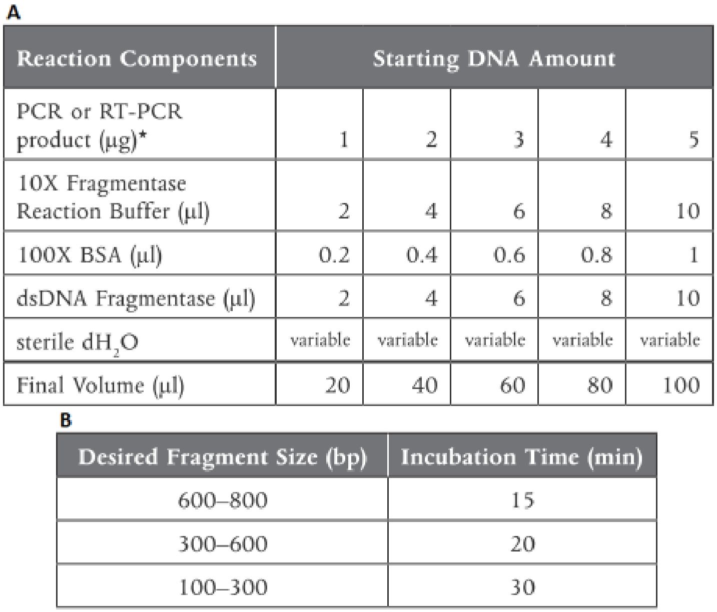
In this section we provide a brief overview of the conventional mechanical Roche 454 shearing and the two enzymatic fragmentation methods we compared.

3.1.5.1 Standard Roche® 454's GS FLX Titanium$^{\text{TM}}$ shearing method

Due to the losses that occur during the different steps of the process, a high amount of input DNA is required. The procedure starts with mechanical fragmentation through nebulization. Because this results in a rather broad fragment length distribution, this step is followed by a size selection. During this step the fragments with lengths that give rise to the best sequencing results are selected. More specifically, the size range of 500-800bp is selected because these fragments amplify well during the emulsion PCR and take advantage of the possible long read lengths. Next, the fragment ends are polished and the platform-specific adaptors are ligated to the ends. The remaining steps of this library preparation procedure are aimed at selectively removing those fragments that have the same adaptor at both opposing ends [64]. Fragmentation through mechanical shearing is performed by Genomics Core (University Hospitals Leuven/K.U. Leuven), who will incorporate different barcodes for each sample if required.

3.1.5.2 NEBNext$^{\text{TM}}$ dsDNA Fragmentase$^{\text{TM}}$

This protocol is based on the activity of two enzymes, a *V. vulnificus* nuclease that generates random nicks, blended with a modified T7 endonuclease that recognizes the nicks and cleaves the opposite strand. Here, size distribution is a time-dependent variable, as well as dependent on input size (Figure 3.5). All reactions with the NEBNext® dsDNA Fragmentase$^{\text{TM}}$ were performed according to the manufacturer's instructions [9]. The chosen incubation time at 37 degrees Celsius was 20 minutes because this generates fragments with a peak in size distribution of 300-600bp, which are well suited for the 454 platform.



A

Reaction Components	Starting DNA Amount				
PCR or RT-PCR product (µg)*	1	2	3	4	5
10X Fragmentase Reaction Buffer (µl)	2	4	6	8	10
100X BSA (µl)	0.2	0.4	0.6	0.8	1
dsDNA Fragmentase (µl)	2	4	6	8	10
sterile dH$_2$O	variable	variable	variable	variable	variable
Final Volume (µl)	20	40	60	80	100

B

Desired Fragment Size (bp)	Incubation Time (min)
600–800	15
300–600	20
100–300	30

Figure 3.5: (A) Reaction conditions and (B) times for (RT-)PCR. Adapted from [9].

From here on, we resort to the standard library preparation method, and all further reactions are performed by Genomics Core (University Hospitals Leuven/K.U. Leuven), who will incorporated different barcodes for each sample if required.

3.1.5.3 Nextera$^{\text{TM}}$ technology

This shearing method is based on transposon mediated insertion of the adapter sequences and thus enables to directly obtain a library in finished form (Figure 3.6). The efficiency of this reaction, and the introduction of the adapter sequences simultaneously with fragmentation, ensure both a low input requirement, as little as 50ng, and a library that is ready for downstream sequencing.

All reactions with the Nextera^{TM} DNA Sample Prep Kit (454 compatible) were performed according to the manufacturer's protocol (Epicentre Biotechnologies). During the addition of the emPCR compatible sites, each sample was barcoded with a unique tag (Table 3.4).

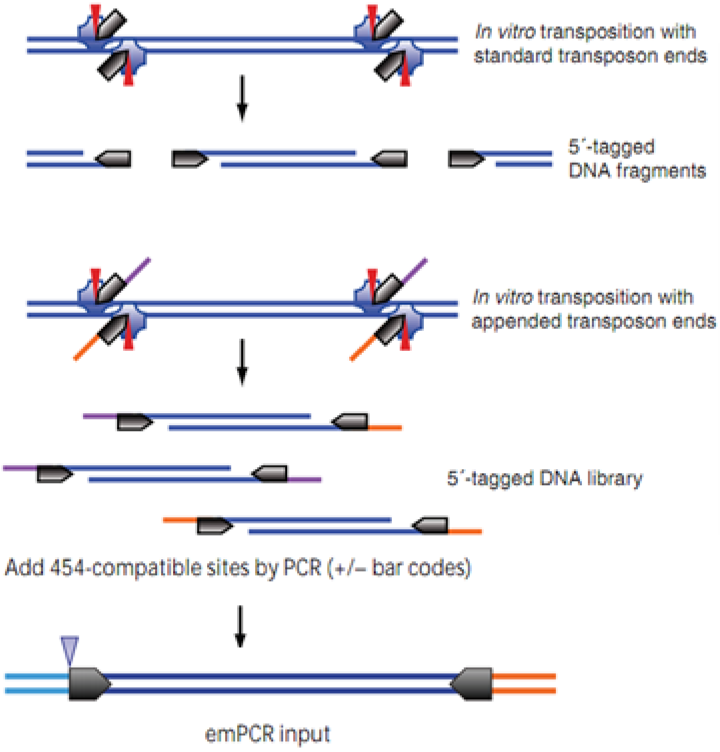

In vitro transposition with standard transposon ends
5´-tagged DNA fragments
In vitro transposition with appended transposon ends
5´-tagged DNA library
Add 454-compatible sites by PCR (+/– bar codes)
emPCR input

Figure 3.6: Nextera^{TM} technology protocol. Adapted from [10]

Table 3.4 - Overview of the samples' barcodes

PBCMs samples	**Outer - PCR**	**Patient**	**MID code**	**MID sequence**
		AR01	MID1	5'-CCATCTCATCCCTGCGTGTCTCCGACTCAG**ACGAGTGCGT**AGATGTGTATAAGAGACAG-3'
		AR07	MID3	5'-CCATCTCATCCCTGCGTGTCTCCGACTCAG**AGACGCACTC**AGATGTGTATAAGAGACAG-3'
	Inner - PCR	**Patient**	**MID code**	**MID sequence**
		AR01	MID2	5'-CCATCTCATCCCTGCGTGTCTCCGACTCAG**ACGCTCGACA**AGATGTGTATAAGAGACAG-3'
		AR07	MID4	5'-CCATCTCATCCCTGCGTGTCTCCGACTCAG**AGCACTGTAG**AGATGTGTATAAGAGACAG-3'
Plasma samples	**Outer - PCR**	**Patient**	**MID code**	**MID sequence**
		AR01	---	---
		AR05	MID6	5'-CCATCTCATCCCTGCGTGTCTCCGACTCAG**ATATCGCGAG**AGATGTGTATAAGAGACAG-3'
		AR06	MID8	5'-CCATCTCATCCCTGCGTGTCTCCGACTCAG**CTCGCGTGTC**AGATGTGTATAAGAGACAG-3'
		AR07	MID11	5'-CCATCTCATCCCTGCGTGTCTCCGACTCAG**TGATACGTCT**AGATGTGTATAAGAGACAG-3'
	Inner - PCR	**Patient**	**MID code**	**MID sequence**
		AR01	MID5	5'-CCATCTCATCCCTGCGTGTCTCCGACTCAG**ATCAGACACG**AGATGTGTATAAGAGACAG-3'
		AR05	MID7	5'-CCATCTCATCCCTGCGTGTCTCCGACTCAG**CGTGTCTCTA**AGATGTGTATAAGAGACAG-3'
		AR06	MID10	5'-CCATCTCATCCCTGCGTGTCTCCGACTCAG**TCTCTATGCG**AGATGTGTATAAGAGACAG-3'
		AR07	MID13	5'-CCATCTCATCCCTGCGTGTCTCCGACTCAG**CATAGTAGTG**AGATGTGTATAAGAGACAG-3'

3.2 Sequence independent amplification

New 6 replicate RNA extractions from the plasma samples were done with the same methods as described before. The pooled RNA extract was used for full genome amplification using WT-Ovation$^{\text{TM}}$ RNA-Seq System from NuGEN$^{®}$ (Figure 3.7). The heart of the system is a random reverse transcription with hybrid random hexamer and oligo(dT) primers tailed with a 5' RNA portion, which is a unique sequence that is not complementary to the RNA sample and does not hybridize to it [87], followed by an amplification step with primers directed at the previously incorporated tails.

The resulting product has a median length of 200bp. Given the median the read length of the 454's GS FLX$^{\text{TM}}$ system with Titanium chemistry is 400bp, the use of this kit will probably result in a loss of potential output.

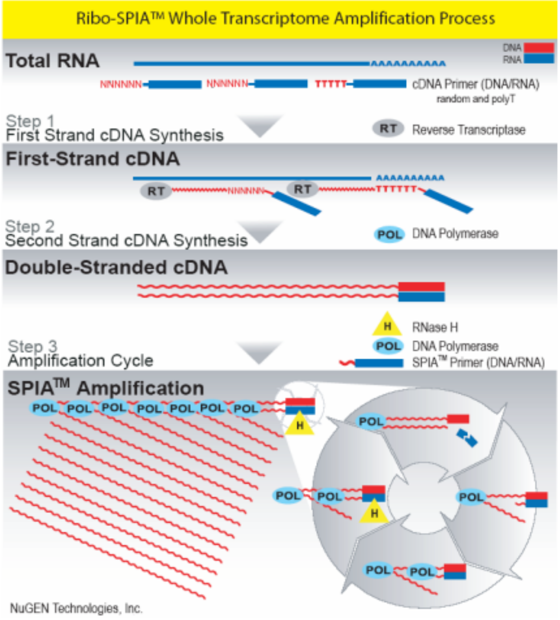


Figure 3.7: RNA amplification process used by WT-Ovation$^{\text{TM}}$ RNA-Seq System. Adapted from [11].

3.3 Sequencing

Sequencing using the Roche$^{®}$ 454's GS FLX$^{\text{TM}}$ platform with Titanium chemistry is performed at Genomics Core (University Hospitals Leuven/K.U. Leuven). We refer to the introduction for background on 454 sequencing.

3.3.1 Assay-related error

One of the objectives of this study, elucidating which viral compartment is best suited for the detection of pre-therapy drug resistance, critically depends on the ability to accurately detect DRMs, even when these are present at low frequencies. Hence, it is important to accurately distinguish true variants from experimentally introduced errors.

Errors that result from the sequencing process itself, as well as those introduced during the sample preprocessing stadia invariably impact on the sensitivity. For this reason, we opted for high-fidelity enzymes for the sequence specific amplification approach. Importantly, the extent to which errors from both sources impact on the detection limit differs. Substitution errors introduced during the PCR amplifications can be modeled as random events which happen at relatively low rates, and are therefore not expected to systematically inflate the frequencies at which low-level DRMs are detected. For example, an error during the first round of amplification in a PCR with one template molecule as input will be passed on in 25% of the progeny molecules. Such a mutation event in the second round will affect 12,5% of the progeny and so forth. In order to arrive at the same level of amplification-introduced errors when a PCR starts with 100 input templates, that same position will have to be independently substituted in all input molecules during that particular round of the cycling program or another extremely unlikely combination of substitution events has to take place. Because we use Platinum Taq High Fidelity (Invitrogen^{TM}) for the outer PCR, which has an error rate of at about 10^{-6}, an impact on the more prevalent DRMs becomes highly unlikely. Expressed in numbers for the outer PCR in which we generate the longest amplicon (gp160, 2888bp), 1000 input templates equal 2888000nt. Such an error rate can thus be expected to result in at about 3 errors during the first round. Assuming these occur on different positions, which is the most likely scenario, this theoretically limits the sensitivity to 0,025%. In the event 3 mutations independently occur on the same position, this still only affects sequencing depths $<$0,1%, which is not meaningful to strive for at the given input of 1000 molecules.

Nonetheless, when looking at very rare minor variants, these errors may become important, especially because we have not taken into account the error rate during the RT-step in the above reasoning, which can be considerable higher in comparison with that of the high fidelity polymerases [83].

In contrast, errors that occur during the sequencing process itself occur at higher frequencies (Q20[2] for the first 400bp [88]), particularly at homopolymer runs [83], and therefore will be of greater importance. To statistically account for these non-random distributed [57] errors, we adopted the approach developed by Wang *et al.* [58]. This procedure relies on the error pattern found in the pyro-reads derived from a clonally amplified sequence. For this, the sequence of the clone as determined with the standard chain-termination technology is used as a gold standard, and all observed differences with the 454 reads from the same clone are considered a 454 sequencing error. Specifically, errors in the homopolymer and heteropolymer regions can be modeled according to a Poisson-distribution, and we only consider those variants that could not be explained by Poisson error(P $<$0,001) and thus are highly unlikely to result from a sequencing artifact.

To obtain this clonal sequence, we performed a limiting dilution PCR whereby we amplified the gp160 region of two laboratory strains, *IIIb* and *pNL*4.3. The PCR product derived from the dilution whereby less than one-third of the reactions gave a positive result was considered to be derived from one template [89] and will be sequenced with both Sanger and 454 technology.

3.4 Sequencing and data analysis

Samples were sequenced with the GS FLX^{TM} Sequencer with Titanium chemistry by the Genomics Core centre (University Hospitals Leuven/K.U. Leuven) and the results were provided to us as Standard Flowgram Files. For the data-analysis we rely on a combination of freely available software because this allows us to address the specific characteristics of pyro-reads derived from highly complex populations such as HIV during the read mapping. Sequence data were extracted and converted into FASTA-format with a freeware python-script, sff_extract.py [90].

Prior to the read-mapping, reads with an exact match to both the barcode and, for the Nextera-fragmented samples, the transposon end sequence are extracted with Segminator [91] and included in the next stages of the analysis. Because the 5' end of the reads is slightly more error-prone and sequencing errors tend to occur on a disproportionately small number of reads [92], this effectively functions as an heuristic filter that increases the overall quality of the remaining reads. Conveniently, the non-viral "search string" sequences are clipped from the reads during the extraction process.

The Segminator-program [91] plays a central role in the read mapping stage of our data-analysis because

[2] Q20 is a *Phred*-equivalent expression. It implies an accuracy of 99%

the implemented innovative mapping algorithm -based on the cover density of so-called k-mers across the reference template- ensures an unbiased and minimal data-loss during these initial stages. The authors noted that the use of a closely related reference sequence significantly reduces data-losses [91]. The absence of a data-specific consensus sequence required us to adopt a two-staged mapping process: during the first step reads are mapped against a distantly related reference genome, in our case we used the HXB2 reference genome (accession number K03455). The resulting pairwise alignment was manually edited and served to construct a data-specific consensus sequence. During the manual editing, all reads that were either very short ($<30nt$) or contained ambiguity characters (N) were removed [92]. The second step involves mapping the sample reads against the newly created data-specific consensus sequence. The resulting pairwise alignment is again manually edited and can, depending on the sample under investigation, be used to screen for DRMs or to determine the appropriate cut-off above which the observed frequency of a mutation is highly unlikely to result from sequencing error (see above). For the former we make use of the publicly available, curated online Stanford Drug Resistance Database [93], which provides a comprehensive list of all resistance mutations in the gene regions targeted by ART.

Chapter 4

Results

Within the time frame of this thesis, we completed the RNA/DNA extraction, evaluation and application of the amplification steps, fragmentation using three different approaches for the available clinical samples, as well as the sequence independent amplification procedure. Sequencing of these products is currently ongoing. Below, we report the results of the amplification step, the quantification of the DNA yield before applying the general 454 shearing, NEBNext® dsDNA Fragmentase^{TM} and Nextera^{TM} technologies, and the fragment size distribution for the latter. In addition, we present the analysis of the sequence data generated for a single sample as an illustration of the work that needs to be completed.

4.1 Amplifications

Having evaluated and adapted the PCR protocols for our purposes, we performed the 6 replicate RNA extraction. Only one replicate DNA extraction of the PBMC samples was done. Afterwards, the pooled extracts were subjected to a 5 replicates (RT-)PCR, of which the products were subsequently pooled per amplicon. The pooled outer-PCR products were used as input for a 5 replicate inner-PCR for each amplicon.

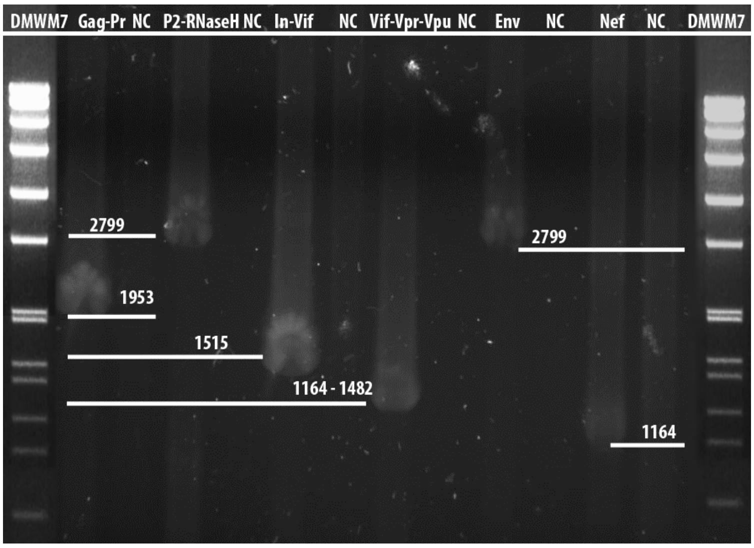


Figure 4.1: Nested-PCR of all amplicons for a PBMC sample of patient AR01-902. (NC) Negative control; (DMWM) DNA molecular weight marker

The resulting products of the inner nested-PCR were pooled per amplicon for each patient, and the amplification results were evaluated on a 1% agarose gel electrophoresis (Figure 4.1).

The amplified products were consistent with the expected length of the targeted amplicons. Also, negative controls were always included in order to detect potential contamination, and always resulted in negative results.

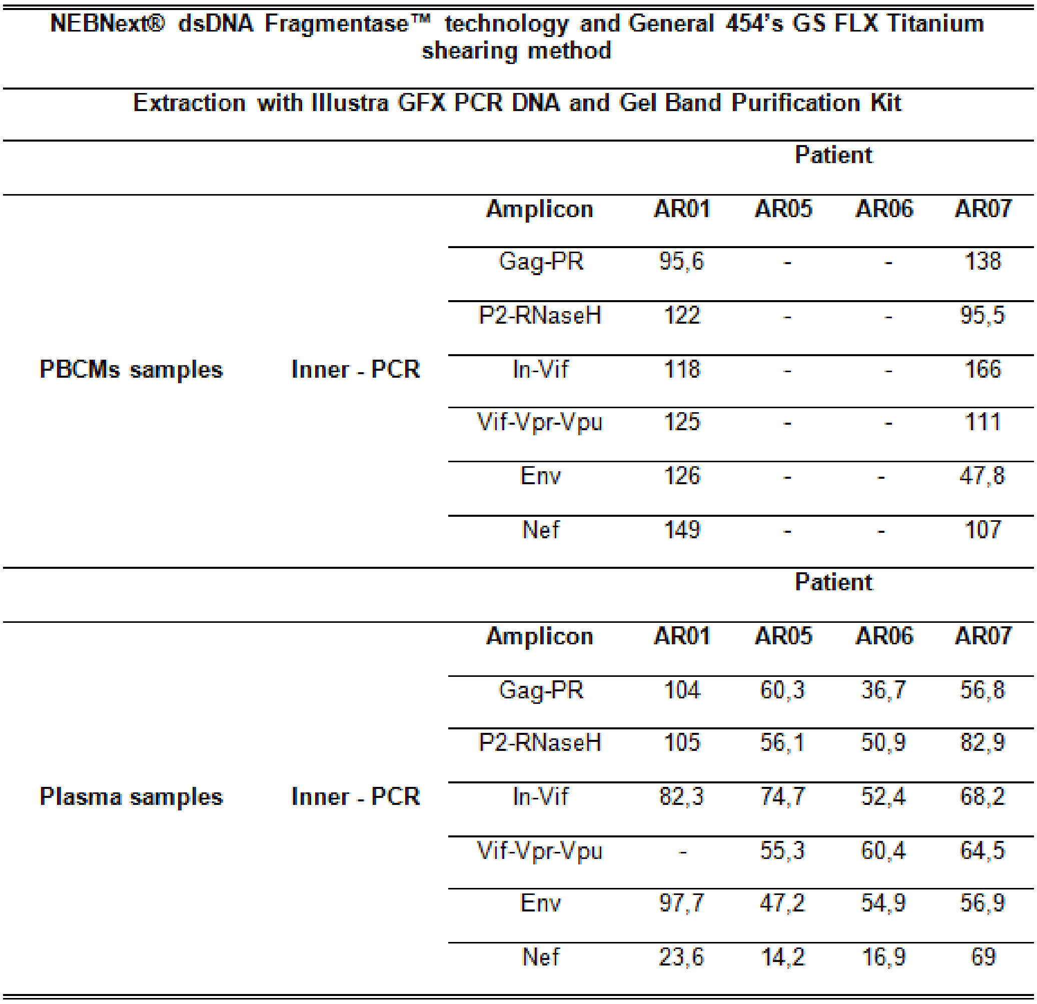
4.2 Quantification

The quantification results $(ng/\mu L)$ of all samples purified with illustra® GFX^{TM} PCR DNA and Gel Band Purification kit (GE Healthcare Life Sciences) with an input volume of 100μL and eluted in 25μL according to the manufacturer's instructions, are presented in Table 4.1. These samples were further subjected to both NEBNext® dsDNA Fragmentase^{TM} and the general 454 shearing method.

We also present the quantification results $(ng/\mu L)$ of all samples, purified with DNA Clean&Concentrator^{TM} (Zymo Research) with 100μL of input volume and eluted in 10μL according to the manufacturer's instructions, in Table 4.2. These samples were further subjected to Nextera^{TM} technology. Because of the lower elution volume, the DNA concentrations are on average higher compared to the results obtained by the illustra® GFX^{TM} PCR DNA and Gel Band Purification procedure.

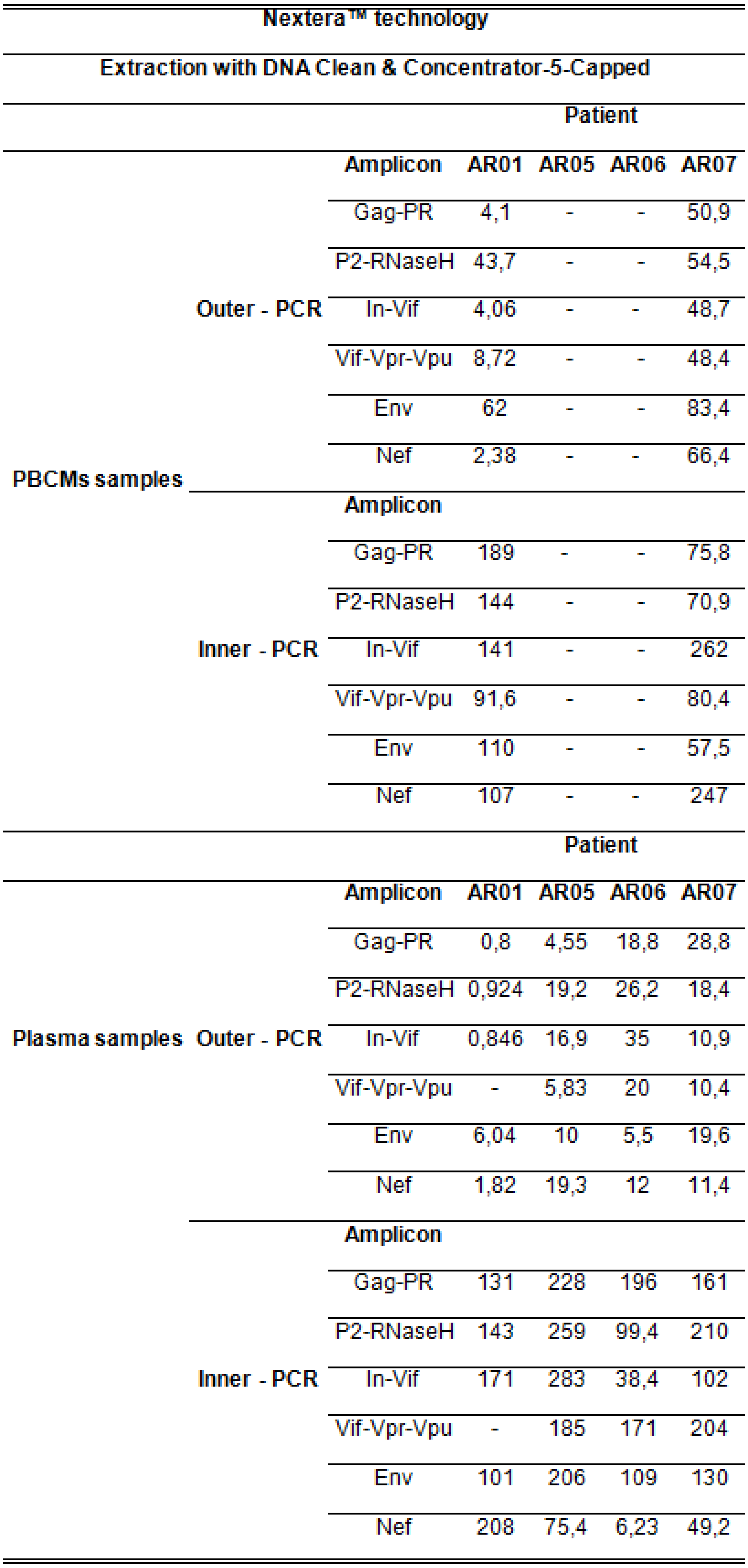
Interestingly, the outer-PCRs of both plasma and PBMC samples did not contain enough material to arrive at the required amount of input material for the NEBNext® dsDNA Fragmentase^{TM} and general 454 shearing method. For this reason and taking into account the minimal input DNA requirement for each fragmentation method, the outer-PCRs were only used for the Nextera^{TM} technology, with the exception of the plasma outer-PCR product of patient AR01-902 that proved to have resulted in a too low yield.

Table 4.1 - DNA concentration $(ng/\mu L)$ of inner PCRs for both PBMCs and plasma samples



NEBNext® dsDNA Fragmentase™ technology and General 454's GS FLX Titanium shearing method						
Extraction with Illustra GFX PCR DNA and Gel Band Purification Kit						
			Patient			
		Amplicon	**AR01**	**AR05**	**AR06**	**AR07**
PBCMs samples	**Inner - PCR**	Gag-PR	95,6	-	-	138
		P2-RNaseH	122	-	-	95,5
		In-Vif	118	-	-	166
		Vif-Vpr-Vpu	125	-	-	111
		Env	126	-	-	47,8
		Nef	149	-	-	107
			Patient			
		Amplicon	**AR01**	**AR05**	**AR06**	**AR07**
Plasma samples	**Inner - PCR**	Gag-PR	104	60,3	36,7	56,8
		P2-RNaseH	105	56,1	50,9	82,9
		In-Vif	82,3	74,7	52,4	68,2
		Vif-Vpr-Vpu	-	55,3	60,4	64,5
		Env	97,7	47,2	54,9	56,9
		Nef	23,6	14,2	16,9	69

Table 4.2 - DNA concentration ($ng/\mu L$) of outer and inner PCRs for both PBMCs and plasma samples



Nextera™ technology						
Extraction with DNA Clean & Concentrator-5-Capped						
			Patient			
		Amplicon	AR01	AR05	AR06	AR07
PBCMs samples	Outer - PCR	Gag-PR	4,1	-	-	50,9
		P2-RNaseH	43,7	-	-	54,5
		In-Vif	4,06	-	-	48,7
		Vif-Vpr-Vpu	8,72	-	-	48,4
		Env	62	-	-	83,4
		Nef	2,38	-	-	66,4
	Inner - PCR	Amplicon				
		Gag-PR	189	-	-	75,8
		P2-RNaseH	144	-	-	70,9
		In-Vif	141	-	-	262
		Vif-Vpr-Vpu	91,6	-	-	80,4
		Env	110	-	-	57,5
		Nef	107	-	-	247
			Patient			
		Amplicon	AR01	AR05	AR06	AR07
Plasma samples	Outer - PCR	Gag-PR	0,8	4,55	18,8	28,8
		P2-RNaseH	0,924	19,2	26,2	18,4
		In-Vif	0,846	16,9	35	10,9
		Vif-Vpr-Vpu	-	5,83	20	10,4
		Env	6,04	10	5,5	19,6
		Nef	1,82	19,3	12	11,4
	Inner - PCR	Amplicon				
		Gag-PR	131	228	196	161
		P2-RNaseH	143	259	99,4	210
		In-Vif	171	283	38,4	102
		Vif-Vpr-Vpu	-	185	171	204
		Env	101	206	109	130
		Nef	208	75,4	6,23	49,2

4.3 Sample pre-processing methods

All inner-PCRs were subjected to both NEBNext® dsDNA Fragmentase™ and the standard 454 shearing method. All outer- and inner-PCRs were subjected to Nextera™ and for those, we report the results on the fragment size distribution below.

4.3.1 Fragment size distribution

To avoid and inefficient and relatively expensive sequencing run, it is standard protocol to examine the fragment size distribution prior to 454 sequencing to assure that the length of the fragments is in the appropriate range.

According to the distributions obtained with a Bioanalyzer (Agilent Technologies) by Genomics Core (University Hospitals Leuven/K.U. Leuven) (Figure 4.2), the fragment size distributions peak at lengths >1000bp. We further comment on these findings in the Discussion.

4.3.2 Sequence analysis

The results on the fragment size distribution led us to generate reads using two emPCR conditions (0,15 and 0,30 copies per bead (cpb))(see Discussion) for one sample only, each run on 1/16th region of the PicoTiterPlate™. We used the Nextera-fragmented, PBMC-derived inner-PCR product of patient AR01 as a test case. We will further refer to the reads of both emPCR conditions as AR01-015 and AR01-030, referring to the emPCR condition with 0,15cpb and 0,30cpb respectively.

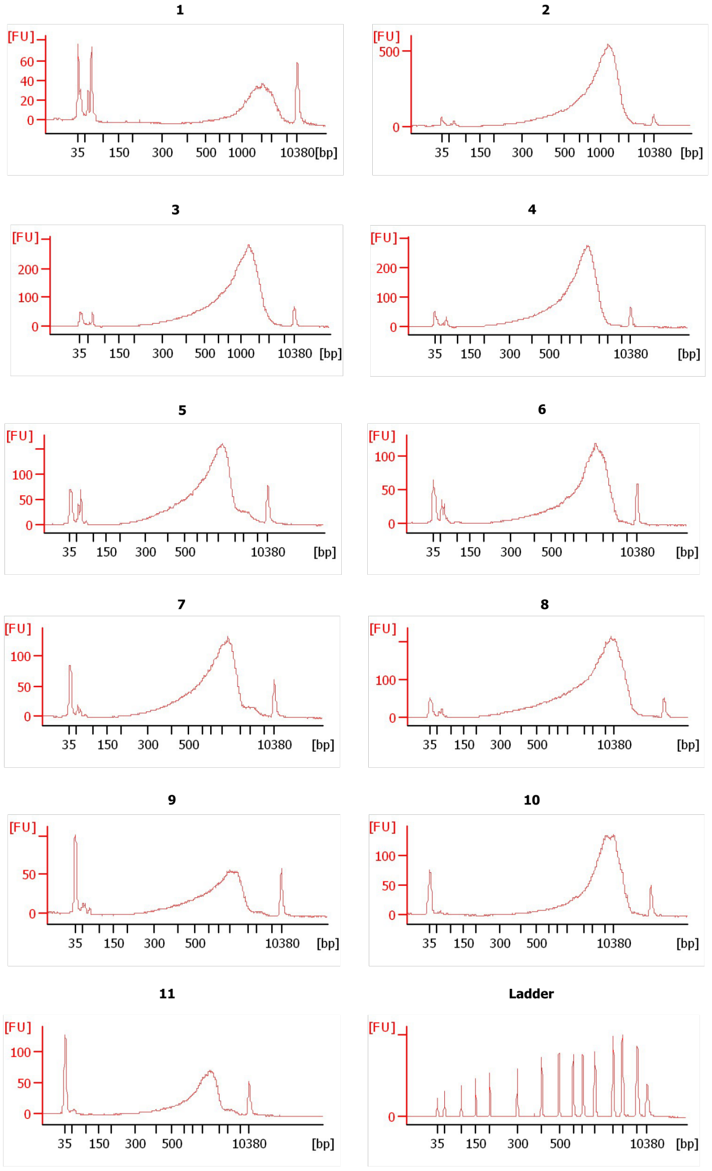


Figure 4.2: Fragment size distribution for samples subjected to Nextera™ technology. (FU) Fluorescent units; (1) AR01 Outer-PCR PBMC sample; (2) AR01 Inner-PCR PBMC sample; (3) AR07 Outer-PCR PBMC sample; (4) AR07 Inner-PCR PBMC sample; (5) AR01 Inner-PCR plasma sample; (6) AR05 Outer-PCR plasma sample; (7) AR05 Inner-PCR plasma sample; (8) AR06 Outer-PCR plasma sample; (9) AR06 Inner-PCR plasma sample; (10) AR07 Outer-PCR plasma sample; (11) AR07- Inner-PCR plasma sample

The initial quality control revealed that the number of filter-passed reads was within the expected range: 31.787 reads for AR01-015 and 33.609 reads for AR01-030. The median read length, 314nt for AR01-015 and 328nt for AR01-030, however, appears to be shorter than expected. Also, the profiles of read length versus the number of reads do not seem to be affected to a large extent by the different emPCR conditions (Figure 4.3).

Next, reads without an exact match to both the barcode (MID2, see Table 3.4) and the 19bp Transposome^{TM} end sequence (AGATGTGTATAAGAGACAG) were discarded. This filtering step excluded 2.900 reads of sample AR01-015 and 2.887 reads of sample AR01-030 from further analysis.

Because the adopted two-stage mapping strategy is very time consuming, we have limited ourselves to a complete mapping process of the protease region, part of the Gag-Pr amplicon, of sample AR01-015. The extracted reads served as the basis for the multiple alignment, for which we constructed the coverage profile after editing (Figure 4.4). This illustrates the homogenizing effect of removing short reads on the coverage profile.

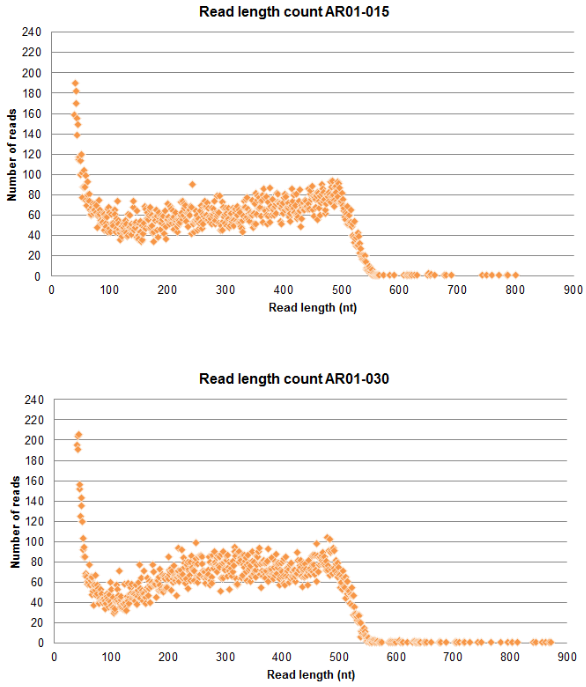


Figure 4.3: Read length count for AR01-015 and AR01-030

We screened the resulting alignment for the presence of DRMs. An overview of the DRMs and their frequency can be found in Table 4.3. Also indicated in this table is the presence or absence of the DRMs in the protease gene sequence of the corresponding plasma sample from patient AR01-902 that was previously obtained using standard genotyping (population sequencing). These preliminary results seem to corroborate earlier findings that most, if not all, highly prevalent DRMs are also detected by standard bulk sequencing approaches but that low to moderate frequent DRMs often go undetected [7] Indeed, all high-frequency DRMs were also detected with the bulk sequencing approach, but, notwithstanding I50V represented 7,42% of the 454 reads, it was not present in the Sanger sequence. Of note, one DRM (E35D) was only found in the plasma-derived Sanger sequence and not in the pyro-reads from the PBMC sample. Because the proviral reservoir has a slow genetic turnover [80], this might be explained by the recent appearance and fixation of this DRM.

For all other amplicons of both samples we confined the analysis to the first stage of the mapping procedure for illustration purposes. The residue frequencies were subsequently extracted to construct the corresponding coverage profiles (Figure 4.5). These profiles demonstrate that in general, a good in-depth view over the complete length of the genome can be obtained.

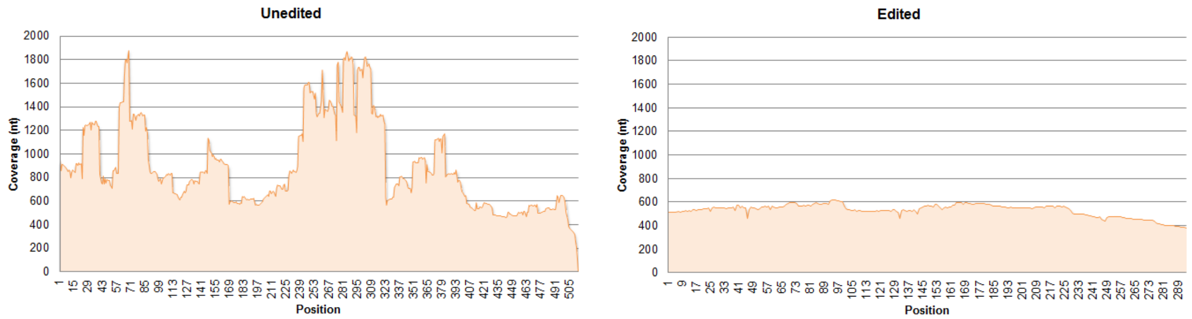
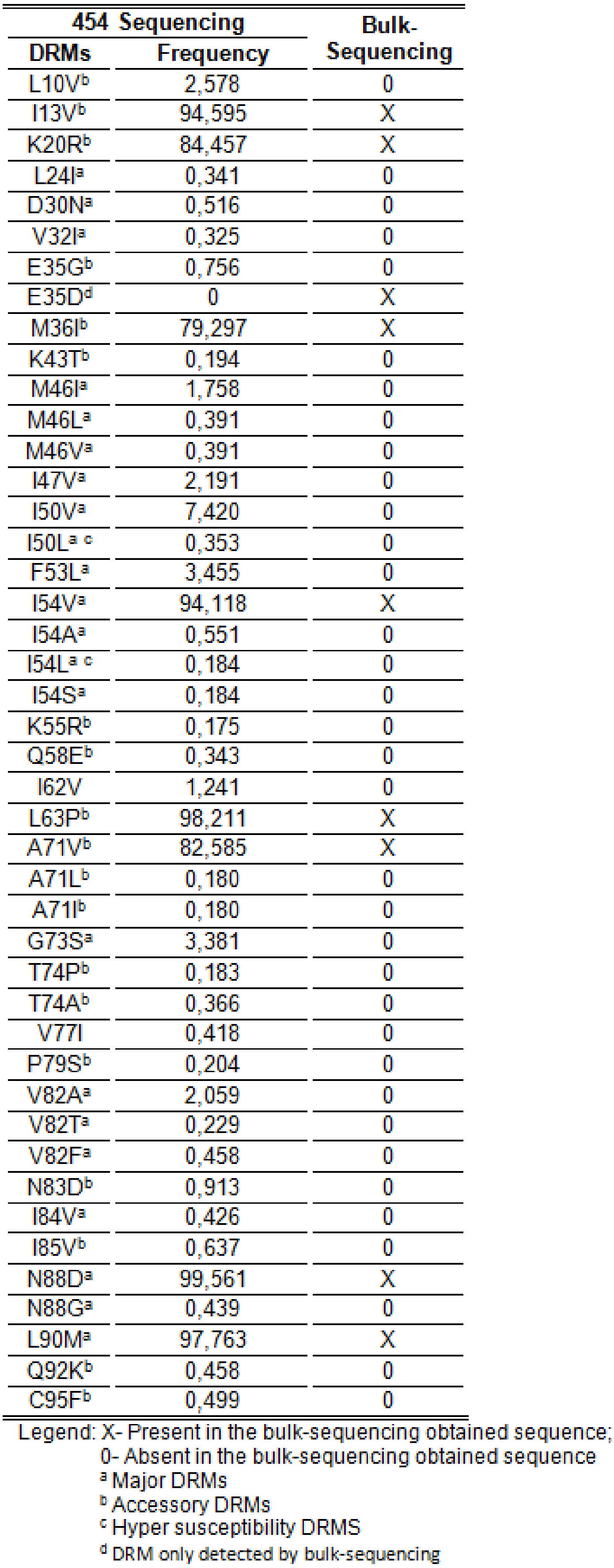


Figure 4.4: Comparison between Pr-region unedited and edited coverage profiles against data-specific consensus

Table 4.3 - DRMs frequencies (%) based on Stanford Drug Resistance Database



454 Sequencing		Bulk-Sequencing
DRMs	Frequency	
L10V[b]	2,578	0
I13V[b]	94,595	X
K20R[b]	84,457	X
L24I[a]	0,341	0
D30N[a]	0,516	0
V32I[a]	0,325	0
E35G[b]	0,756	0
E35D[d]	0	X
M36I[b]	79,297	X
K43T[b]	0,194	0
M46I[a]	1,758	0
M46L[a]	0,391	0
M46V[a]	0,391	0
I47V[a]	2,191	0
I50V[a]	7,420	0
I50L[a] [c]	0,353	0
F53L[a]	3,455	0
I54V[a]	94,118	X
I54A[a]	0,551	0
I54L[a] [c]	0,184	0
I54S[a]	0,184	0
K55R[b]	0,175	0
Q58E[b]	0,343	0
I62V	1,241	0
L63P[b]	98,211	X
A71V[b]	82,585	X
A71L[b]	0,180	0
A71I[b]	0,180	0
G73S[a]	3,381	0
T74P[b]	0,183	0
T74A[b]	0,366	0
V77I	0,418	0
P79S[b]	0,204	0
V82A[a]	2,059	0
V82T[a]	0,229	0
V82F[a]	0,458	0
N83D[b]	0,913	0
I84V[a]	0,426	0
I85V[b]	0,637	0
N88D[a]	99,561	X
N88G[a]	0,439	0
L90M[a]	97,763	X
Q92K[b]	0,458	0
C95F[b]	0,499	0

Legend: X- Present in the bulk-sequencing obtained sequence;
0- Absent in the bulk-sequencing obtained sequence
[a] Major DRMs
[b] Accessory DRMs
[c] Hyper susceptibility DRMS
[d] DRM only detected by bulk-sequencing

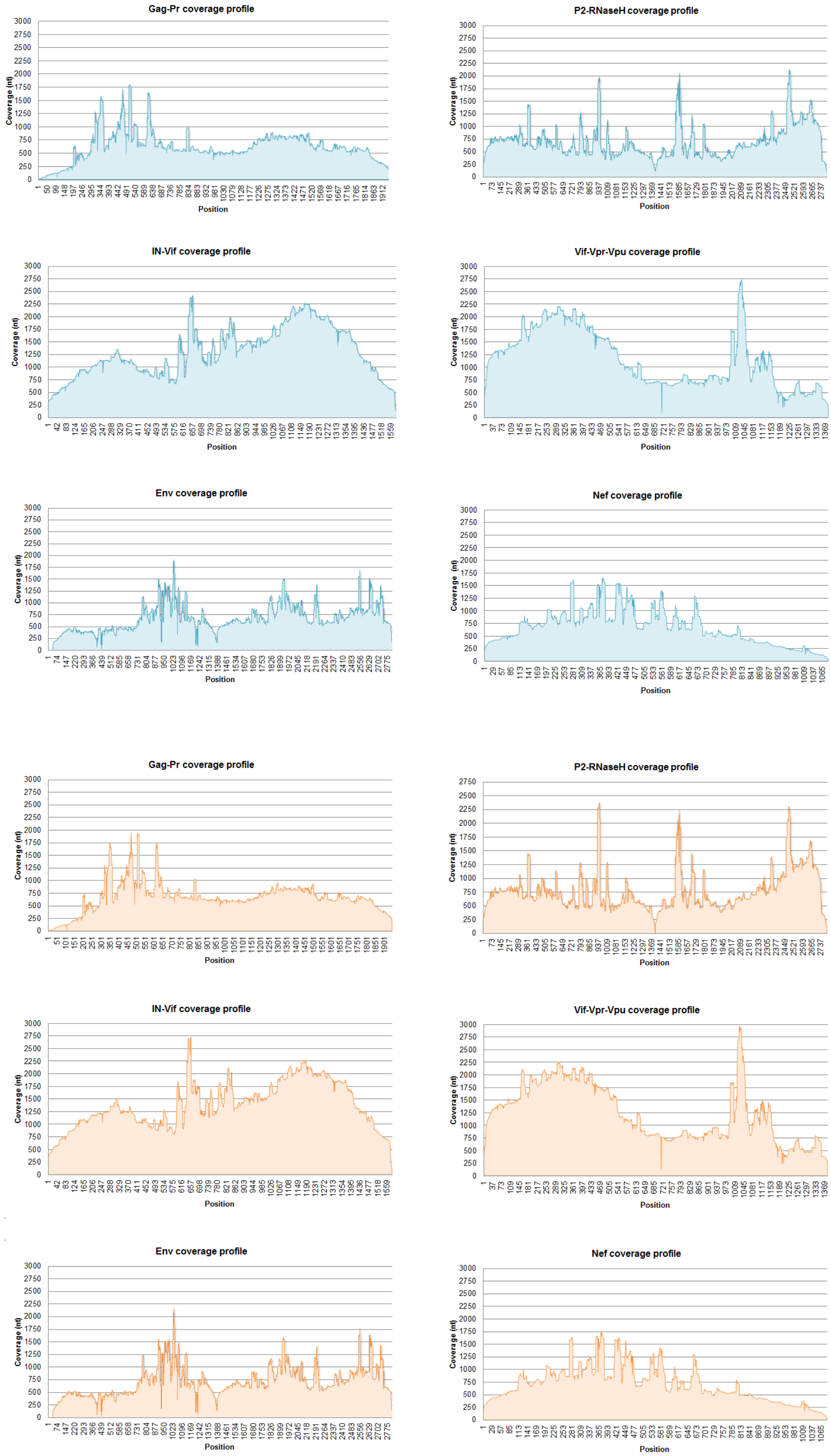


Figure 4.5: Coverage profiles for AR01-015 ($blue$) and AR01-030 ($orange$) amplicons.

Chapter 5

Discussion

Next generation sequencing methods provide a unique opportunity to study the composition of complex viral populations such as HIV. Amongst the NGS platforms, we opted for the Roche® 454's GS FLX^{TM} system because it excels in read length, which greatly facilitates the read alignment and assembly, the first step in many analysis. We ambitiously set out to compare several sample pre-processing protocols and coupled this to a research question that involves the comparison of plasma and PBMC viral reservoirs in order to investigate which one would be suitable to use for genotyping in a pre-therapy setting. Due to time constraints, only part of these goals could be accomplished in this thesis: all samples are amplified, purified and fragmented by the diverse array of methodologies we selected but only one sample could be sequenced. We choose to complete these procedures prior to sequencing so that all the products could then be simultaneously sequenced using a single 454 run.

Our first results already highlight the potential advantage of the 50ng input requirement of the Nextera^{TM} sample preparation protocol over the alternative pre-processing methods. The benefit of such a low input requirement is that the second-round PCR amplification could potentially be avoided, which eliminates potential biases that result from second-round amplification. A potential downside of this method is the expected drop in sequence coverage starting at about 125bp from the end of the DNA (both ends) because the Transposomes^{TM} do not attach to the ends of dsDNA. In our experimental setup with relatively small overlapping amplicons, this may influence the coverage distribution. However, based on the coverage-plots of the raw data (Figure 4.5), this feature does not appear to seriously impact on the detection limit.

Both NEBNext® dsDNA Fragmentase^{TM} and general Roche® 454 shearing, with a respective input requirement of 1μg and 5μg, could only be performed on the inner PCR products of all samples. As this is the case for both samples with high and low (pro-)viral loads, and additional amplification step may always be required irrespective of the (pro-)viral load. This is not only time-consuming and labor intensive, but it also brings about potential systematic and stochastic errors that may distort the relative frequencies in which the viral variants are present. For example, Mild *et al.* [94] recently illustrated that primer selection may impact on the detection and quantification of low-level variants. It is of interest to note that even with this low input requirement, the outer-PCR product of patient AR01 could not be fragmented with this approach. Should the Nextera^{TM} approach prove to be the preferred method, this may constitute an additional argument in favor for standardizing the use of replicate (RT-)PCRs followed by the pooling of the products, in studies where NGS platforms are used to interrogate complex (viral) populations.

Of all samples, the Genomics Core center (University Hospitals Leuven/K.U. Leuven) was able to perform an analysis of the fragment size distribution of the 11 samples that were fragmented with the Nextera^{TM} approach. The peaks of the fragment size distribution are centered around $1000bp$. In addition, the distribution appears to be skewed, which makes that for most samples more than 50% of the fragments are shifted to higher/lower molecular weights ($>1000bp$). Such high molecular weights are indicative for an inefficient emPCR (see above), but according to the manufacturer such a pattern can also be caused by non-covalent bonds between the adaptor sequences and should thus not be considered as a contra-indication for a successful emPCR. A cautionary approach led us to opt for a test-run of one sample on 1/16th region of the PTP whereby two emPCR conditions were tested (0,15cpb and 0,30cpb). The number of reads generated for both tests (31.787 and 33.609) lies well within the expected range (25.000 - 37.500)[1], which therefore seems to substantiate the manufacturer's claim. However, the median read length is shorter then expected (Figure 4.3), which can be due to a too long average fragment length- a feature we cannot exclude based on the available data. Therefore, these results should be compared with those obtained from the Nextera^{TM} fragmentation of the same sample, which achieves adaptor ligation and library enrichment with less PCR cycles.

The coverage plots of the raw sequence data (Figure 4.5) illustrate both the possibility to obtain a good in-depth view over the complete length of the genome with this method, as well as the large amounts of data this entails. The latter, in turn, warrants that any profound data-analysis will be very time-consuming and constitutes the reason why only one region was looked at in greater detail. Analysis of the coverage plots in Figure 4.4 revealed that removing short sequences, which are more likely to not uniquely map on one region of the genome, has a profound impact on the "coverage peaks". More specifically, by lowering the coverage, a higher frequency at which a DRM needs to be detected at that position before it is accepted as truly present, is required. Consequently, the false-discovery rate will be reduced.

The false-discovery rate is important with respect to the low frequencies at which most DRMs were detected, even often at levels below 1% (see Table 4.3). Because no statistical analysis to asses the appropriate frequency cut-off to distinguish sequencing noise from true variants has been carried out, we must stress it is likely that most of these extremely rare DRMs are the result of sequencing error. However, the detection of variants at levels as low as 0,175% for K55R supports the applicability of this NGS platform for future studies aimed at gaining a more profound insight into the clinical importance of minority DRMs. With respect to the latter, the different genetic barrier to resistance of the diverse therapy regimens implies that the clinical relevance of minority DRMs may vary, which in turn implicates various prevalence cut-offs may need to be determined [95]. As opposed to PIs, NNRTIs have a relatively low genetic barrier to resistance. It is therefore not surprising that several studies point to a relatively high sensitivity of NNRTI-based regimens to minority DRMs [96, 97, 98].

The fragmentation procedures were also compared with a sequence-independent amplification method, the WT-Ovation^{TM} RNA-Seq System from NuGEN®, which has a minimum input 500pg of RNA. This input requirement can be an obstacle, in particular for low viral load samples, because RNA extraction procedures are not very efficient and because the major composition of the extracted product is carrier RNA. For this reason, accurate direct quantification of target viral RNA for RNA-Seq input with traditional methods is not possible. The main advantage of this technology is that its starts from the extracted RNA from the samples, thus avoiding primer-specific amplification and the inherent risk of a result that is not representative of the actual variant proportions in the sample. The sequence-independent methodology therefore holds better potential to uncover the true scope of the population, although not necessarily in

[1] 1.000.000 reads / 16 = 62.500 reads per 1/16th PTP. Accounting for a gasket-induced output loss of 40%-60%, the read number is expected to lie in the range of 25.000 - 37.500

the most cost efficient way.

An important research question that we would eventually like to address with our sequence data involves transmitted drug resistance. Drug resistant variants can not only be generated *de novo*, but they may also evolve from resistant virus acquired through transmission. In both cases, they may lie at the basis of a suboptimal response to regimens [99]. Furthermore, it is currently assumed that drug-resistant variants persist life-long as archived provirus in the PBMCs, which implies that TDR might exert a long-term impact on responses to antiretroviral therapy [100]. Consequently, TDR poses a challenge with potentially important clinical and public health implications, which led to the implementation of genotypic resistance testing in therapy-naive patients in international guidelines [74].

However, transmitted drug resistant variants tend to revert to wild-type in the absence of the selective pressure of therapy as a result of the fitness cost that is often associated with drug resistance mutations, the so-called reversion phenomenon [101, 102]. Alternatively, they may persist as minor variants under the detection limit of standard sequencing approaches. Because such conventional ”bulk sequencing” based on genotyping assays are most oftenly used to assess TDR, the prevalence estimates of approximately 10% in Europe and the USA [101, 103] are prone to underestimation [104]. Here, the 454 platform, through its massive parallel sequencing ability, can analyze large numbers of specimens simultaneously, and therefore presents the potential for a more accurate and cost-competitive assessment of the true scope of TDR on a population scale [105]. Given the increasing number of people receiving antiretroviral treatment, estimated at about 5.2 million in 2009 [106], and the associated high dropout rate [107], such accurate population scale screenings may become an important tool to comprehensively characterize the epidemiology of TDR. More specifically, implementation of such routine screening, especially in resource-limited regions that currently experience a scale-up in the availability of therapy, may inform on the optimization of the implemented therapy-policy [108]. A more precise evaluation of TDR may also serve a public health purpose by aiding in the identification of transmission networks and populations on which to focus preventive measures [109].

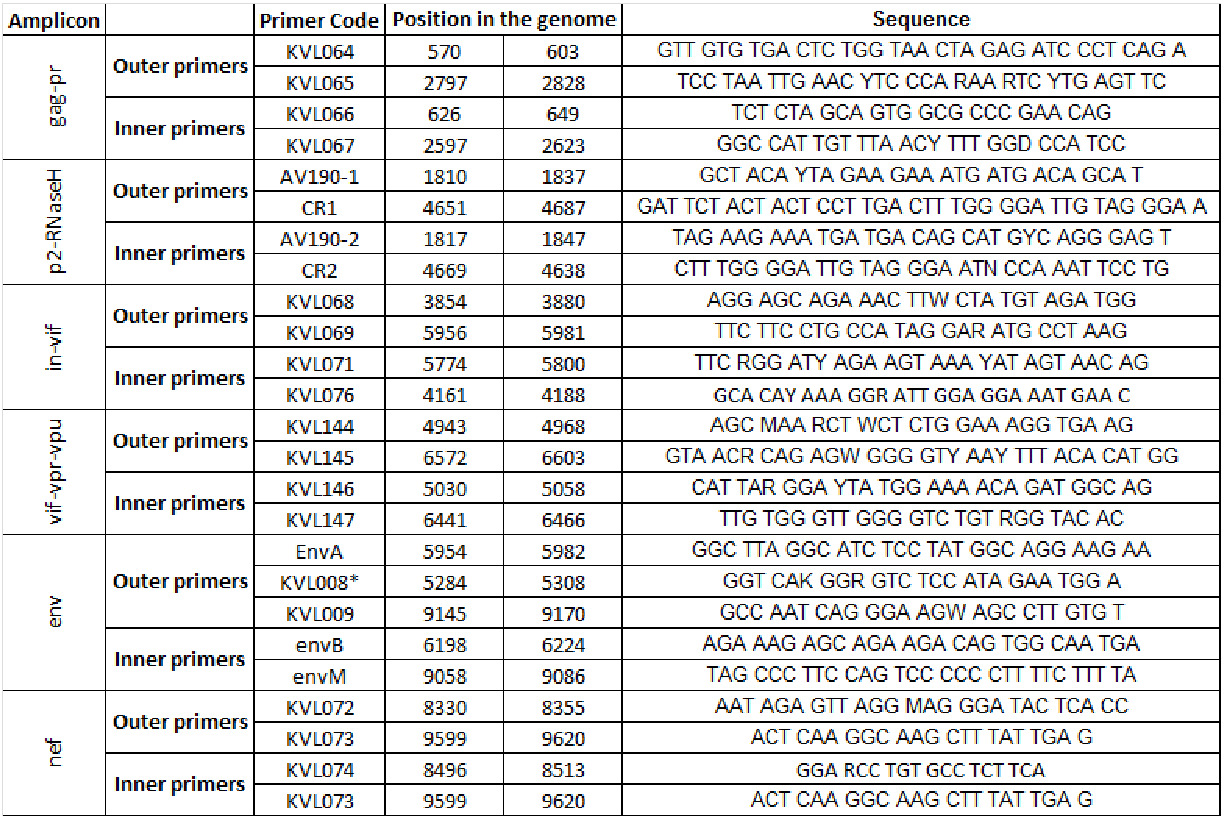
In addition to the choice between the tools at hand, a second possible variable that may impact the assessment of TDR prevalence, and therefore the clinician's choice for an appropriate first-line therapy regimen, lies in the viral reservoir under investigation. Whereas plasma virus generally represents recently produced virions, the proviral DNA in PBMCs may serve as an archive of genetic diversity. When cells become infected with HIV this is usually followed by the hijacking of the cellular machinery and the production of many offspring virions (see above).

Sometimes, however, the viral genome is inserted in CD4$^+$ T-cells that, instead of being activated return to the resting memory state [110]. The latter process leads to the formation of a latent reservoir that not only prevents the eradication of HIV - proviruses are no target of current ART- but also serves as a memory wherein all diversity ever generated is present. The dynamics of of the proviral reservoir are such that it's fueled primarily by the transmitted variant(s) [78, 79]. The combination of the former with the relatively slow genetic turnover of the PBMC reservoir [80], makes it possible to find transmitted DRMs in PBMCs long after the initial infection. The long half-life of resting CD4$^+$ T-cells, make PBMCs a plausible candidate for a more sensitive TDR detection [81]. More specifically, chances are that DRMs ”hidden” by the reversion phenomenon can still be detected in the proviral reservoir. Here it is of interest to note that all 5 DRMs present at frequencies between 2% and 79,3% are only found in the PBMC sample of patient AR01-902 and not in the plasma-derived Sanger sequence (Table 4.3). Thus, although statistical support needs to be evaluated and a comparison with the 454 data of the corresponding plasma sample needs to be performed, the preliminary analysis of DRMs in the examined PBMC sample of patient AR01-902 hint at the possibility of more sensitive TDR detection in PBMCs. However, there was one mutation in the plasma population that was not detected in our PBMC-derived reads, which could suggest that the proviral reservoir is not always superior.

In conclusion, the successful sequencing of one sample and the completion of all preparatory work for this research offers hope for useful applications in the near future. A comprehensive analysis of the sequence data that is currently generated is however required in order to make an informed choice between the different sample preparation methods available. Finally, we are aware that our choice for a pre-processing procedure may only be a current snapshot that should evolve with the swift evolution of state-of-the-art sequencing technologies.

Addendum

Table 1 – Overview of all (RT-)PCRs primers.



Amplicon		Primer Code	Position in the genome		Sequence
gag-pr	**Outer primers**	KVL064	570	603	GTT GTG TGA CTC TGG TAA CTA GAG ATC CCT CAG A
		KVL065	2797	2828	TCC TAA TTG AAC YTC CCA RAA RTC YTG AGT TC
	Inner primers	KVL066	626	649	TCT CTA GCA GTG GCG CCC GAA CAG
		KVL067	2597	2623	GGC CAT TGT TTA ACY TTT GGD CCA TCC
p2-RNaseH	**Outer primers**	AV190-1	1810	1837	GCT ACA YTA GAA GAA ATG ATG ACA GCA T
		CR1	4651	4687	GAT TCT ACT ACT CCT TGA CTT TGG GGA TTG TAG GGA A
	Inner primers	AV190-2	1817	1847	TAG AAG AAA TGA TGA CAG CAT GYC AGG GAG T
		CR2	4669	4638	CTT TGG GGA TTG TAG GGA ATN CCA AAT TCC TG
in-vif	**Outer primers**	KVL068	3854	3880	AGG AGC AGA AAC TTW CTA TGT AGA TGG
		KVL069	5956	5981	TTC TTC CTG CCA TAG GAR ATG CCT AAG
	Inner primers	KVL071	5774	5800	TTC RGG ATY AGA AGT AAA YAT AGT AAC AG
		KVL076	4161	4188	GCA CAY AAA GGR ATT GGA GGA AAT GAA C
vif-vpr-vpu	**Outer primers**	KVL144	4943	4968	AGC MAA RCT WCT CTG GAA AGG TGA AG
		KVL145	6572	6603	GTA ACR CAG AGW GGG GTY AAY TTT ACA CAT GG
	Inner primers	KVL146	5030	5058	CAT TAR GGA YTA TGG AAA ACA GAT GGC AG
		KVL147	6441	6466	TTG TGG GTT GGG GTC TGT RGG TAC AC
env	**Outer primers**	EnvA	5954	5982	GGC TTA GGC ATC TCC TAT GGC AGG AAG AA
		KVL008*	5284	5308	GGT CAK GGR GTC TCC ATA GAA TGG A
		KVL009	9145	9170	GCC AAT CAG GGA AGW AGC CTT GTG T
	Inner primers	envB	6198	6224	AGA AAG AGC AGA AGA CAG TGG CAA TGA
		envM	9058	9086	TAG CCC TTC CAG TCC CCC CTT TTC TTT TA
nef	**Outer primers**	KVL072	8330	8355	AAT AGA GTT AGG MAG GGA TAC TCA CC
		KVL073	9599	9620	ACT CAA GGC AAG CTT TAT TGA G
	Inner primers	KVL074	8496	8513	GGA RCC TGT GCC TCT TCA
		KVL073	9599	9620	ACT CAA GGC AAG CTT TAT TGA G

* PBMCs sense outer primer

Table 2 - Overview of all (RT-) PCRs mixes and conditions.

	Mixes	Amplicons																	
		Gag-PR			P2-RNaseH			In-Vif			Vif-Vpr-Vpu			Env			Nef		
			Conditions			Conditions			Conditions			Conditions			Conditions			Conditions	
RT-PCR			RNA incubation			RNA incubation			RNA incubation			RNA incubation			RNA incubation			RNA incubation	
	Reaction Mix	1x	65°C 30"		1x	65°C 30"		1x	65°C 30"		1x	65°C 30"		1x	65°C 30"		1x	65°C 30"	
	Sense primer	0,2µM	55°C 5'		0,2µM	55°C 5'		0,2µM	55°C 5'		0,2µM	55°C 5'		0,2µM	55°C 5'		0,2µM	55°C 5'	
	Antisense primer	0,2µM	everse Transcriptio		0,2µM	everse Transcriptio		0,2µM	everse Transcriptio		0,2µM	everse Transcripti		0,2µM	everse Transcripti		0,2µM	everse Transcriptic	
	MgSO4	0,8mM	55°C 30'		1,3mM	55°C 30'		1,05mM	55°C 30'		1mM	50°C 30'		0mM	55°C 30'		1,05mM	55°C 30'	
	Superscript III RT / Plat HF	1µL	PCR cycling profile		1µL	PCR cycling profile		1µL	PCR cycling profile		1µL	PCR cycling profile		1µL	PCR cycling profile		1µL	PCR cycling profile	
	RNA Protector	10U	94°C 2'		10U	94°C 2'		10U	94°C 2'		10U	94°C 2'		10U	94°C 1'30"		10U	94°C 2'	
	RNA extract	10µL	94°C 15"	40x	10µL	94°C 15"	40x	10µL	94°C 15"	40x	10µL	94°C 15"	40x	10µL	94°C 15"	40x	10µL	94°C 15"	40x
	H2O	Add until final	57°C 30"		Add until final	61°C 30"		Add until final	53°C 30"		Add until	54°C 30"		Add until	55°C 30"		Add until	56°C 30"	
		volume of 50µL	68°C 2'		volume of 50µL	68°C 3'		volume of 50µL	68°C 2'30"		final volume	68°C 1'30"		final volume	68°C 3' 30"		final volume	68°C 1'30"	
			4°C infinite			4°C infinite			4°C infinite			4°C infinite			4°C infinite			4°C infinite	
		For amplification of PBMCs samples, identical master mixes as for the plasma samples' reverse transcription and amplification were used, with the substitution of RNA protector by the same volume of water. The use of RNAseA at the extraction step and the omission of the RT-step in the hot-started cycling program results in a traditional outer PCR.																	
Nested-PCR	**Mix 1**																		
	dNTPs	200µM	Conditions		0,2 mM	Conditions		200µM	Conditions		0,4mM	Conditions		0,2mM	Conditions		0,4mM	Conditions	
	Sense primer	0,4µM	PCR cycling profile		0,5µM	PCR cycling profile		0,4µM	PCR cycling profile		0,8µM	PCR cycling profile		0,2µM	PCR cycling profile		0,8µM	PCR cycling profile	
	Antisense primer	0,4µM	95°C 2'		0,5µM	94°C 2'		0,4µM	95°C 15"	10x	0,8µM	94°C 2'		0,2µM	94°C 1'30"		0,8µM	95°C 2'	
	Outer PCR product	5uL	95°C 15"	10x	2µL	94°C 15"	30x	5uL	55°C 30"		5uL	94°C 15"	40x	1uL	94°C 15"	5x	5uL	95°C 15"	10x
	H20	Add until final	58°C 30"		Add until final	59°C 30"		Add until final	68°C 2'30"		Add until	56°C 30"		Add until	55°C 30"		Add until	53°C 30"	
		volume of 25µL	68°C 2'30"		volume of 25µL	68°C 3'		volume of 25µL	95°C 15"	30x	final volume	68°C 1'30"		final volume	68°C 2'45"		final volume	72°C 1'	
	Mix 2		95°C 15"	30x		4°C infinite			55°C 30"			4°C infinite			94°C 15"	30x		95°C 15"	30x
	Expand buffer	1x	58°C 30"		1x			1x	68°C 2'30" + 5"/cycl		2x			1x	55°C 30"		2x	53°C 30"	
	MgCl2	2mM	68°C 2'30" + 5"/cycle		3,5mM			2mM	4°C infinite		2,4mM			2mM	68°C 3' + 1"/cycle		8mM	72°C 1'*	
	Expand HF enzyme	2,625U	4°C infinite		2,6U			2,625U			2,625U			1U	4°C infinite		2,625U	4°C infinite	
	H2O	Add until final			Add until final			Add until final			Add until			Add until			Add until		
		volume of 25µL			volume of 25µL			volume of 25µL			final volume			final volume			final volume		

Bibliography

[1] Fauci AS, Pantaleo G, Stanley S, Weissman D. Immunopathogenic mechanisms of HIV infection. Ann Intern Med. 1996 Apr;124(7):654–663.

[2] Trivedi B. The primate connection. Nature. 2010 Jul;466(7304):S5. Available from: `http://dx.doi.org/10.1038/nature09236`.

[3] Various. United Nations Environment Programme. wwwuneporg;.

[4] Various. HIV-1 Gene Map. Los Alamos: National Laboratory - HIV Sequence Database Website. 2010;(Jan). Available from: `http://www.hiv.lanl.gov/content/sequence/HIV/MAP/landmark.html`.

[5] Various. Biology of HIV. web-books http://wwwweb-bookscom/eLibrary/ON/B0/B22/05MHIVhtml;.

[6] Ramdohr P. http://wwwflickrcom/photos/opiado/4191226676/. 2009 December;.

[7] Vrancken B, Lequime S, Theys K, Lemey P. Covering all bases in HIV research: unveiling a hidden world of viral evolution. AIDS Rev. 2010;12(2):89–102.

[8] Gega A, Kozal MJ. New technology to detect low-level drug-resistant HIV variants. Future Virology. 2011;6(1):17–26.

[9] NEBNextTM dsDNA FragmentaseTM Manual. New England Biolabs;.

[10] Fraz Syed NC Haiying Grunenwald. Optimized library preparation method for next-generation sequencing - Epicentre Biotechnologies. Nature Methods. 2009;6.

[11] NuGEN. Ovation RNA-Seq System User Guide. NuGEN Technologies, Inc.; 2009.

[12] for Disease Control (CDC) C. Kaposi's sarcoma and Pneumocystis pneumonia among homosexual men–New York City and California. MMWR Morb Mortal Wkly Rep. 1981 Jul;30(25):305–308.

[13] Gottlieb MS, Schroff R, Schanker HM, Weisman JD, Fan PT, Wolf RA, et al. Pneumocystis carinii pneumonia and mucosal candidiasis in previously healthy homosexual men: evidence of a new acquired cellular immunodeficiency. N Engl J Med. 1981 Dec;305(24):1425–1431. Available from: `http://dx.doi.org/10.1056/NEJM198112103052401`.

[14] 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 May;220(4599):868–871.

[15] Coffin J, Haase A, Levy JA, Montagnier L, Oroszlan S, Teich N, et al. Human immunodeficiency viruses. Science. 1986 May;232(4751):697.

[16] Clavel F, Guetard D, Brun-Vezinet F, Chamaret S, Rey MA, Santos-Ferreira MO, et al. Isolation of a new human retrovirus from West African patients with AIDS. Science. 1986 Jul;233(4761):343–346.

[17] Fauci AS. Immunopathogenesis of HIV infection. J Acquir Immune Defic Syndr. 1993 Jun;6(6):655–662.

[18] Morgan D, Mahe C, Mayanja B, Okongo JM, Lubega R, Whitworth JAG. HIV-1 infection in rural Africa: is there a difference in median time to AIDS and survival compared with that in industrialized countries? AIDS. 2002 Mar;16(4):597–603.

[19] Broder S. The development of antiretroviral therapy and its impact on the HIV-1/AIDS pandemic. Antiviral Res. 2010 Jan;85(1):1–18. Available from: `http://dx.doi.org/10.1016/j.antiviral.2009.10.002`.

[20] Gulick RM. Antiretroviral treatment 2010: progress and controversies. J Acquir Immune Defic Syndr. 2010 Dec;55 Suppl 1:S43–S48. Available from: `http://dx.doi.org/10.1097/QAI.0b013e3181f9c09e`.

[21] 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 Oct;84(19):9864–9878. Available from: `http://dx.doi.org/10.1128/JVI.00915-10`.

[22] Tebit DM, Arts EJ. Tracking a century of global expansion and evolution of HIV to drive understanding and to combat disease. Lancet Infect Dis. 2011 Jan;11(1):45–56. Available from: `http://dx.doi.org/10.1016/S1473-3099(10)70186-9`.

[23] Various. Towards Universal access: Scaling up priority HIV/AIDS interventions in the health sector. World Health Organization. 2010;p. 1–150.

[24] Various. Global Report: UNAIDS Report on the Global AIDS Epidemic. World Health Organization. 2010;p. 1–364.

[25] Lemey P, Rambaut A, Pybus OG. HIV evolutionary dynamics within and among hosts. AIDS Rev. 2006;8(3):125–140.

[26] de Sousa JD, Muller V, Lemey P, Vandamme AM. High GUD incidence in the early 20 century created a particularly permissive time window for the origin and initial spread of epidemic HIV strains. PLoS One. 2010;5(4):e9936. Available from: `http://dx.doi.org/10.1371/journal.pone.0009936`.

[27] McGrath KM, Hoffman NG, Resch W, Nelson JA, Swanstrom R. Using HIV-1 sequence variability to explore virus biology. Virus Res. 2001 Aug;76(2):137–160.

[28] Wolfe ND, Switzer WM, Carr JK, Bhullar VB, Shanmugam V, Tamoufe U, et al. Naturally acquired simian retrovirus infections in central African hunters. Lancet. 2004 Mar;363(9413):932–937. Available from: http://dx.doi.org/10.1016/S0140-6736(04)15787-5.

[29] Rambaut A, Posada D, Crandall KA, Holmes EC. The causes and consequences of HIV evolution. Nat Rev Genet. 2004 Jan;5(1):52–61. Available from: http://dx.doi.org/10.1038/nrg1246.

[30] Heuverswyn FV, Li Y, Neel C, Bailes E, Keele BF, Liu W, et al. Human immunodeficiency viruses: SIV infection in wild gorillas. Nature. 2006 Nov;444(7116):164. Available from: http://dx.doi.org/10.1038/444164a.

[31] Simon F, Mauclere P, Roques P, Loussert-Ajaka I, Muller-Trutwin M, Saragosti S. Identification of a new human immunodeficiency virus type 1 distinct from group M and group O. Nature. 1998;4:1032–1037.

[32] Guertler 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 Mar;68(3):1581–1585.

[33] Haesevelde MV, Decourt JL, Leys RJD, Vanderborght B, van der Groen G, van Heuverswijn H, et al. Genomic cloning and complete sequence analysis of a highly divergent African human immunodeficiency virus isolate. J Virol. 1994 Mar;68(3):1586–1596.

[34] Charneau P, Borman AM, Quillent C, Guetard D, Chamaret S, Cohen J, et al. Isolation and envelope sequence of a highly divergent HIV-1 isolate: definition of a new HIV-1 group. Virology. 1994 Nov;205(1):247–253. Available from: http://dx.doi.org/10.1006/viro.1994.1640.

[35] Loussert-Ajaka I, Chaix ML, Korber B, Letourneur F, Gomas E, Allen E, et al. Variability of human immunodeficiency virus type 1 group O strains isolated from Cameroonian patients living in France. J Virol. 1995 Sep;69(9):5640–5649.

[36] Various. Overview of the subtypes of primate immunodeficiency viruses. Los Alamos: National Laboratory - HIV Sequence Database Website. 2010;(Jan). Available from: http://www.hiv.lanl.gov/content/sequence/HIV/MAP/landmark.html.

[37] Tebit DM, Nankya I, Arts EJ, Gao Y. HIV diversity, recombination and disease progression: how does fitness ”fit” into the puzzle? AIDS Rev. 2007;9(2):75–87.

[38] Korber B, Muldoon M, Theiler J, Gao F, Gupta R, Lapedes A, et al. Timing the ancestor of the HIV-1 pandemic strains. Science. 2000 Jun;288(5472):1789–1796.

[39] 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 Oct;455(7213):661–664. Available from: http://dx.doi.org/10.1038/nature07390.

[40] Keele BF, Heuverswyn FV, Li Y, Bailes E, Takehisa J, Santiago ML, et al. Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. Science. 2006 Jul;313(5786):523–526. Available from: http://dx.doi.org/10.1126/science.1126531.

[41] Fields BN, Knipe DM, Howley PM, Griffin DE. Fields virology. 4th ed. Philadelphia: Lippincott Williams & Wilkins; 2001. Available from: http://www.loc.gov/catdir/enhancements/fy0653/00067800-d.html.

[42] Dong C, Kwas C, Wu L. Transcriptional restriction of human immunodeficiency virus type 1 gene expression in undifferentiated primary monocytes. J Virol. 2009 Apr;83(8):3518–3527. Available from: http://dx.doi.org/10.1128/JVI.02665-08.

[43] Kilareski EM, Shah S, Nonnemacher MR, Wigdahl B. Regulation of HIV-1 transcription in cells of the monocyte-macrophage lineage. Retrovirology. 2009;6:118. Available from: http://dx.doi.org/10.1186/1742-4690-6-118.

[44] Nisole S, Saib A. Early steps of retrovirus replicative cycle. Retrovirology. 2004;1:9. Available from: http://dx.doi.org/10.1186/1742-4690-1-9.

[45] Pan C, Liu S, Jiang S. HIV-1 gp41 fusion intermediate: a target for HIV therapeutics. J Formos Med Assoc. 2010 Feb;109(2):94–105. Available from: http://dx.doi.org/10.1016/S0929-6646(10)60029-0.

[46] Basu VP, Song M, Gao L, Rigby ST, Hanson MN, Bambara RA. Strand transfer events during HIV-1 reverse transcription. Virus Res. 2008 Jun;134(1-2):19–38. Available from: http://dx.doi.org/10.1016/j.virusres.2007.12.017.

[47] Fitzgerald ME, Drohat AC. Structural studies of RNA/DNA polypurine tracts. Chem Biol. 2008 Mar;15(3):203–204. Available from: http://dx.doi.org/10.1016/j.chembiol.2008.03.001.

[48] McColl DJ, Chen X. Strand transfer inhibitors of HIV-1 integrase: bringing IN a new era of antiretroviral therapy. Antiviral Res. 2010 Jan;85(1):101–118. Available from: http://dx.doi.org/10.1016/j.antiviral.2009.11.004.

[49] Karn J. Tat, a novel regulator of HIV transcription and latency. MRC Laboratory of Molecular Biology, Cambridge;.

[50] Balvay L, Rifo RS, Ricci EP, Decimo D, Ohlmann T. Structural and functional diversity of viral IRESes. Biochim Biophys Acta. 2009;1789(9-10):542–557. Available from: http://dx.doi.org/10.1016/j.bbagrm.2009.07.005.

[51] Butsch M, Boris-Lawrie K. Destiny of unspliced retroviral RNA: ribosome and/or virion? J Virol. 2002 Apr;76(7):3089–3094.

[52] Tazi J, Bakkour N, Marchand V, Ayadi L, Aboufirassi A, Branlant C. Alternative splicing: regulation of HIV-1 multiplication as a target for therapeutic action. FEBS J. 2010 Feb;277(4):867–876. Available from: http://dx.doi.org/10.1111/j.1742-4658.2009.07522.x.

[53] Guttler T, Madl T, Neumann P, Deichsel D, Corsini L, Monecke T, et al. NES consensus redefined by structures of PKI-type and Rev-type nuclear export signals bound to CRM1. Nat Struct Mol Biol. 2010 Nov;17(11):1367–1376. Available from: http://dx.doi.org/10.1038/nsmb.1931.

[54] Raska M, Novak J. Involvement of envelope-glycoprotein glycans in HIV-1 biology and infection. Arch Immunol Ther Exp (Warsz). 2010 Jun;58(3):191–208. Available from: http://dx.doi.org/10.1007/s00005-010-0072-3.

[55] VerPlank L, Bouamr F, LaGrassa TJ, Agresta B, Kikonyogo A, Leis J, et al. Tsg101, a homologue of ubiquitin-conjugating (E2) enzymes, binds the L domain in HIV type 1 Pr55(Gag). Proc Natl Acad Sci U S A. 2001 Jul;98(14):7724–7729. Available from: http://dx.doi.org/10.1073/pnas.131059198.

[56] Hill M, Tachedjian G, Mak J. The packaging and maturation of the HIV-1 Pol proteins. Curr HIV Res. 2005 Jan;3(1):73–85.

[57] Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005 Sep;437(7057):376–380. Available from: http://dx.doi.org/10.1038/nature03959.

[58] 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 Aug;17(8):1195–1201. Available from: http://dx.doi.org/10.1101/gr.6468307.

[59] Hoffmann C, Minkah N, Leipzig J, Wang G, Arens MQ, Tebas P, et al. DNA bar coding and pyrosequencing to identify rare HIV drug resistance mutations. Nucleic Acids Res. 2007;35(13):e91. Available from: http://dx.doi.org/10.1093/nar/gkm435.

[60] Vandenbroucke I, Marck HV, Mostmans W, Eygen VV, Rondelez E, Thys K, et al. HIV-1 V3 envelope deep sequencing for clinical plasma specimens failing in phenotypic tropism assays. AIDS Res Ther. 2010;7:4. Available from: http://dx.doi.org/10.1186/1742-6405-7-4.

[61] Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010 Jan;11(1):31–46. Available from: http://dx.doi.org/10.1038/nrg2626.

[62] Bushman FD, Hoffmann C, Ronen K, Malani N, Minkah N, Rose HM, et al. Massively parallel pyrosequencing in HIV research. AIDS. 2008 Jul;22(12):1411–1415. Available from: http://dx.doi.org/10.1097/QAD.0b013e3282fc972e.

[63] Droege M, Hill B. The Genome Sequencer FLX System–longer reads, more applications, straight forward bioinformatics and more complete data sets. J Biotechnol. 2008 Aug;136(1-2):3–10. Available from: http://dx.doi.org/10.1016/j.jbiotec.2008.03.021.

[64] Roche. GS FLX Titanium General Library Preparation Method Manual. Roche; 2009.

[65] Karrer EE, Lincoln JE, Hogenhout S, Bennett AB, Bostock RM, Martineau B, et al. In situ isolation of mRNA from individual plant cells: creation of cell-specific cDNA libraries. Proc Natl Acad Sci U S A. 1995 Apr;92(9):3814–3818.

[66] Polz MF, Cavanaugh CM. Bias in template-to-product ratios in multitemplate PCR. Appl Environ Microbiol. 1998 Oct;64(10):3724–3730.

[67] Bracho MA, Garcia-Robles I, Jimenez N, Torres-Puente M, Moya A, Gonzalez-Candelas F. Effect of oligonucleotide primers in determining viral variability within hosts. Virol J. 2004;1:13. Available from: http://dx.doi.org/10.1186/1743-422X-1-13.

[68] Poon AFY, Swenson LC, Dong WWY, Deng W, Pond SLK, Brumme ZL, et al. Phylogenetic analysis of population-based and deep sequencing data to identify coevolving sites in the nef gene of HIV-1. Mol Biol Evol. 2010 Apr;27(4):819–832. Available from: http://dx.doi.org/10.1093/molbev/msp289.

[69] Bimber BN, Dudley DM, Lauck M, Becker EA, Chin EN, Lank SM, et al. Whole-genome characterization of human and simian immunodeficiency virus intrahost diversity by ultradeep pyrosequencing. J Virol. 2010 Nov;84(22):12087–12092. Available from: http://dx.doi.org/10.1128/JVI.01378-10.

[70] Willerth SM, Pedro HAM, Pachter L, Humeau LM, Arkin AP, Schaffer DV. Development of a low bias method for characterizing viral populations using next generation sequencing technology. PLoS One. 2010;5(10):e13564. Available from: http://dx.doi.org/10.1371/journal.pone.0013564.

[71] Mitsuya H, Weinhold KJ, Furman PA, Clair MHS, Lehrman SN, Gallo RC, et al. 3'-Azido-3'-deoxythymidine (BW A509U): an antiviral agent that inhibits the infectivity and cytopathic effect of human T-lymphotropic virus type III/lymphadenopathy-associated virus in vitro. Proc Natl Acad Sci U S A. 1985 Oct;82(20):7096–7100.

[72] Gardner EM, Burman WJ, Steiner JF, Anderson PL, Bangsberg DR. Antiretroviral medication adherence and the development of class-specific antiretroviral resistance. AIDS. 2009 Jun;23(9):1035–1046. Available from: http://dx.doi.org/10.1097/QAD.0b013e32832ba8ec.

[73] Volberding PA, Deeks SG. Antiretroviral therapy and management of HIV infection. Lancet. 2010 Jul;376(9734):49–62. Available from: http://dx.doi.org/10.1016/S0140-6736(10)60676-9.

[74] Vandamme AM, Sonnerborg A, Ait-Khaled M, Albert J, Asjo B, Bacheler L, et al. Updated European recommendations for the clinical use of HIV drug resistance testing. Antivir Ther. 2004 Dec;9(6):829–848.

[75] Hirsch MS, Gunthard HF, Schapiro JM, Brun-Vezinet F, Clotet B, Hammer SM, et al. Antiretroviral drug resistance testing in adult HIV-1 infection: 2008 recommendations of an International AIDS Society-USA panel. Top HIV Med. 2008;16(3):266–285.

[76] Geretti AM. Antiretroviral resistance in clinical practice. London, editor. Mediscript Ltd.; 2006.

[77] Various. Guidelines for the Use of Antiretroviral Agents in HIV-1 Infected Adults and Adolescents - HIV and Its Treatment - Approved Medications to Treat HIV Infection. US Department of Health and Human Services. December 2009;.

[78] Masquelier B, Taieb A, Reigadas S, Marchou B, Cheneau C, Spire B, et al. Cellular HIV-1 DNA quantification and short-term and long-term response to antiretroviral therapy. J Antimicrob Chemother. 2011 Jul;66(7):1582–1589. Available from: http://dx.doi.org/10.1093/jac/dkr153.

[79] d'Ettorre G, Zaffiri L, Ceccarelli G, Mastroianni CM, Vullo V. The role of HIV-DNA testing in clinical practice. New Microbiol. 2010 Jan;33(1):1–11.

[80] Bi X, Gatanaga H, Ida S, Tsuchiya K, Matsuoka-Aizawa S, Kimura S, et al. Emergence of protease inhibitor resistance-associated mutations in plasma HIV-1 precedes that in proviruses of peripheral blood mononuclear cells by more than a year. J Acquir Immune Defic Syndr. 2003 Sep;34(1):1–6.

[81] Ghosn J, Pellegrin I, Goujard C, Deveau C, Viard JP, Galimand J, et al. HIV-1 resistant strains acquired at the time of primary infection massively fuel the cellular reservoir and persist for lengthy periods of time. AIDS. 2006 Jan;20(2):159–170.

[82] Van Laethem YPKNRMWEVAM Kristel;Schrooten. Transmission cluster of dual-class resistant HIV-1 in untreated patients. International BioInformatics Workshop on Virus Evolution and Molecular Epidemiology. 2007;.

[83] Hedskog C, Mild M, Jernberg J, Sherwood E, Bratt G, Leitner T, et al. Dynamics of HIV-1 quasispecies during antiviral treatment dissected using ultra-deep pyrosequencing. PLoS One. 2010;5(7):e11345. Available from: http://dx.doi.org/10.1371/journal.pone.0011345.

[84] Stenman J, Orpana A. Accuracy in amplification. Nat Biotechnol. 2001 Nov;19(11):1011–1012. Available from: http://dx.doi.org/10.1038/nbt1101-1011b.

[85] Liu SL, Rodrigo AG, Shankarappa R, Learn GH, Hsu L, Davidov O, et al. HIV quasispecies and resampling. Science. 1996 Jul;273(5274):415–416.

[86] Shao W. Sequencing Error Assessment and Reduced PCR-based Recombination for HIV-1 Drug Resistance Mutation Linkage Determination by 454 Pyrosequencing;.

[87] Kurn N, Chen P, Heath JD, Kopf-Sill A, Stephens KM, Wang S. Novel isothermal, linear nucleic acid amplification systems for highly multiplexed applications. Clin Chem. 2005 Oct;51(10):1973–1981. Available from: http://dx.doi.org/10.1373/clinchem.2005.053694.

[88] Roche. 454 Life Sciences, a Roche Company; 2011. www.454.com.

[89] Varghese V, Wang E, Babrzadeh F, Bachmann MH, Shahriar R, Liu T, et al. Nucleic acid template and the risk of a PCR-Induced HIV-1 drug resistance mutation. PLoS One. 2010;5(6):e10992. Available from: http://dx.doi.org/10.1371/journal.pone.0010992.

[90] at COMAV B. COMAV's bioinformatics & genomics - sffextract; 2011. http://bioinf.comav.upv.es/index.html.

[91] Archer J, Rambaut A, Taillon BE, Harrigan PR, Lewis M, Robertson DL. The evolutionary analysis of emerging low frequency HIV-1 CXCR4 using variants through time–an ultra-deep approach. PLoS Comput Biol. 2010;6(12):e1001022. Available from: http://dx.doi.org/10.1371/journal.pcbi.1001022.

[92] Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM. Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol. 2007;8(7):R143. Available from: http://dx.doi.org/10.1186/gb-2007-8-7-r143.

[93] University S. HIV Drug Resistance Database - Stanford University; 2011. http://hivdb.stanford.edu/.

[94] 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. Available from: http://dx.doi.org/10.1371/journal.pone.0022741.

[95] Paredes R, Clotet B. Clinical management of HIV-1 resistance. Antiviral Res. 2010 Jan;85(1):245–265. Available from: http://dx.doi.org/10.1016/j.antiviral.2009.09.015.

[96] 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 Mar;199(5):693–701.

[97] Metzner KJ, Rauch P, von Wyl V, Leemann C, Grube C, Kuster H, et al. Efficient suppression of minority drug-resistant HIV type 1 (HIV-1) variants present at primary HIV-1 infection by ritonavir-boosted protease inhibitor-containing antiretroviral therapy. J Infect Dis. 2010 Apr;201(7):1063–1071. Available from: http://dx.doi.org/10.1086/651136.

[98] Halvas EK, Wiegand A, Boltz VF, Kearney M, Nissley D, Wantman M, et al. Low frequency nonnucleoside reverse-transcriptase inhibitor-resistant variants contribute to failure of efavirenz-containing regimens in treatment- experienced patients. J Infect Dis. 2010 Mar;201(5):672–680. Available from: http://dx.doi.org/10.1086/650542.

[99] Deeks SG. Transmitted minority drug-resistant HIV variants: a new epidemic? PLoS Med. 2008 Jul;5(7):e164. Available from: http://dx.doi.org/10.1371/journal.pmed.0050164.

[100] Booth CL, Geretti AM. Prevalence and determinants of transmitted antiretroviral drug resistance in HIV-1 infection. J Antimicrob Chemother. 2007 Jun;59(6):1047–1056. Available from: http://dx.doi.org/10.1093/jac/dkm082.

[101] programme SPREAD. Transmission of drug-resistant HIV-1 in Europe remains limited to single classes. AIDS. 2008 Mar;22(5):625–635.

[102] Vercauteren J, Wensing AMJ, van de Vijver DAMC, Albert J, Balotta C, Hamouda O, et al. Transmission of drug-resistant HIV-1 is stabilizing in Europe. J Infect Dis. 2009 Nov;200(10):1503–1508. Available from: http://dx.doi.org/10.1086/644505.

[103] Weinstock HS, Zaidi I, Heneine W, Bennett D, Garcia-Lerma JG, Douglas JM, et al. The epidemiology of antiretroviral drug resistance among drug-naive HIV-1-infected persons in 10 US cities. J Infect Dis. 2004 Jun;189(12):2174–2180. Available from: http://dx.doi.org/10.1086/420789.

[104] Jakobsen MR, Tolstrup M, Sogaard OS, Jorgensen LB, Gorry PR, Laursen A, et al. Transmission of HIV-1 drug-resistant variants: prevalence and effect on treatment outcome. Clin Infect Dis. 2010 Feb;50(4):566–573. Available from: http://dx.doi.org/10.1086/650001.

[105] Ji H, Masse N, Tyler S, Liang B, Li Y, Merks H, et al. HIV drug resistance surveillance using pooled pyrosequencing. PLoS One. 2010;5(2):e9263. Available from: http://dx.doi.org/10.1371/journal.pone.0009263.

[106] Conference XIA. Advancing evidence and equity: Report on the XVIII international AIDS conference. International AIDS Society; 2010.

[107] Mills EJ, Nachega JB, Bangsberg DR, Singh S, Rachlis B, Wu P, et al. Adherence to HAART: a systematic review of developed and developing nation patient-reported barriers and facilitators. PLoS Med. 2006 Nov;3(11):e438. Available from: http://dx.doi.org/10.1371/journal.pmed.0030438.

[108] Jordan MR. Assessments of HIV drug resistance mutations in resource-limited settings. Clin Infect Dis. 2011 Apr;52(8):1058–1060. Available from: http://dx.doi.org/10.1093/cid/cir093.

[109] Johnson JA, Geretti AM. Low-frequency HIV-1 drug resistance mutations can be clinically significant but must be interpreted with caution. J Antimicrob Chemother. 2010 Jul;65(7):1322–1326. Available from: http://dx.doi.org/10.1093/jac/dkq139.

[110] Duverger A, Jones J, May J, Bibollet-Ruche F, Wagner FA, Cron RQ, et al. Determinants of the establishment of human immunodeficiency virus type 1 latency. J Virol. 2009 Apr;83(7):3078–3093. Available from: http://dx.doi.org/10.1128/JVI.02058-08.